A Practical Guide to

Diffusion Language Models

Himanshu Sahni
06/08/2025

1. Why?

Language Models

The

quick

brown

fox

jumped

over

the

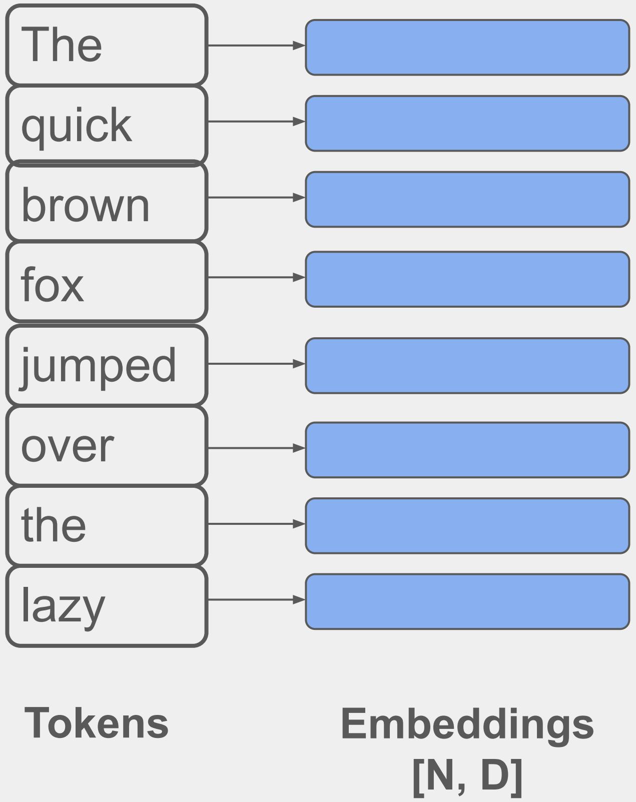
lazy

Language Models

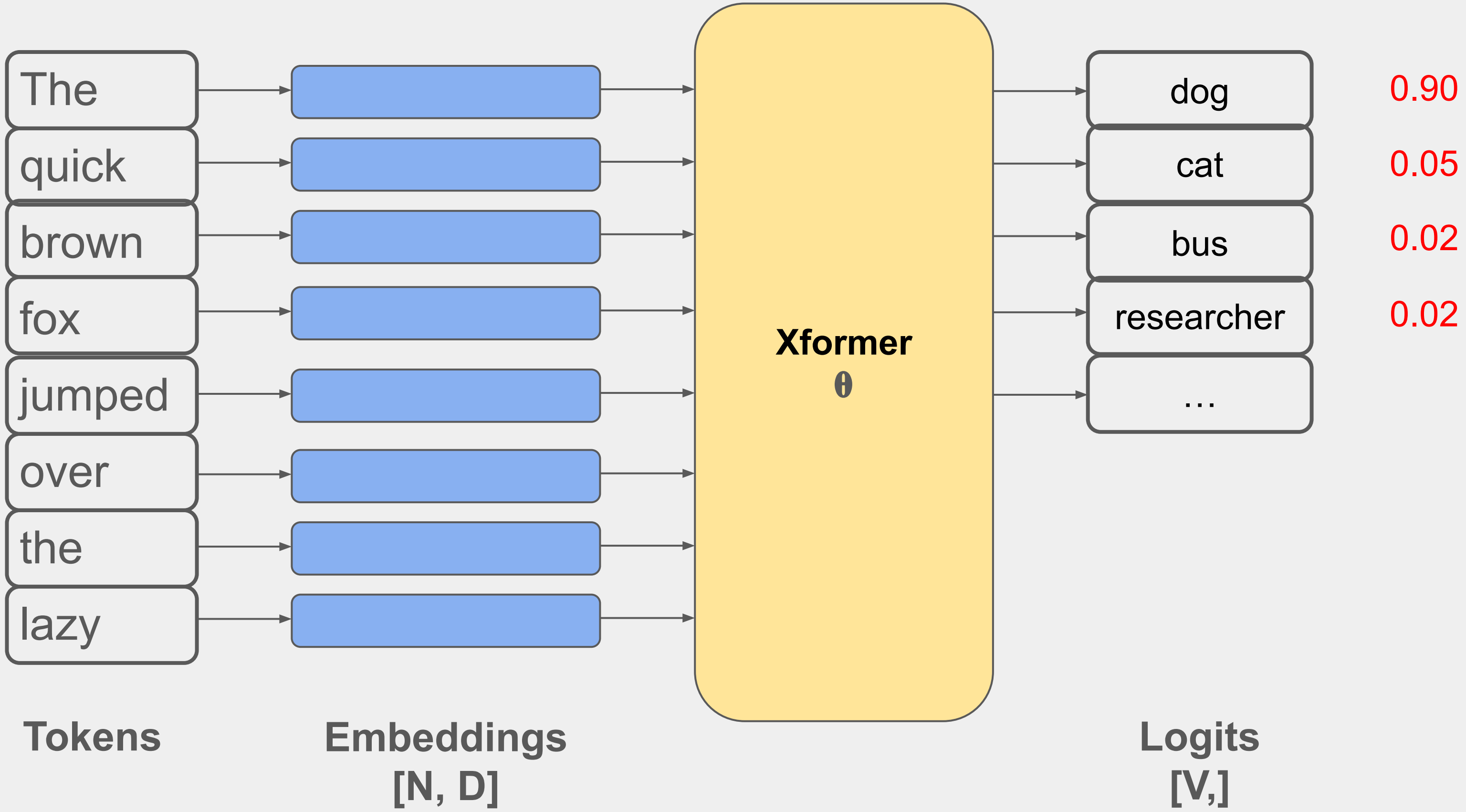
The
quick
brown
fox
jumped
over
the
lazy

Tokens

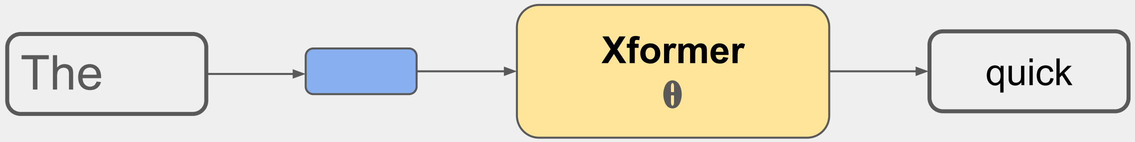
Language Models



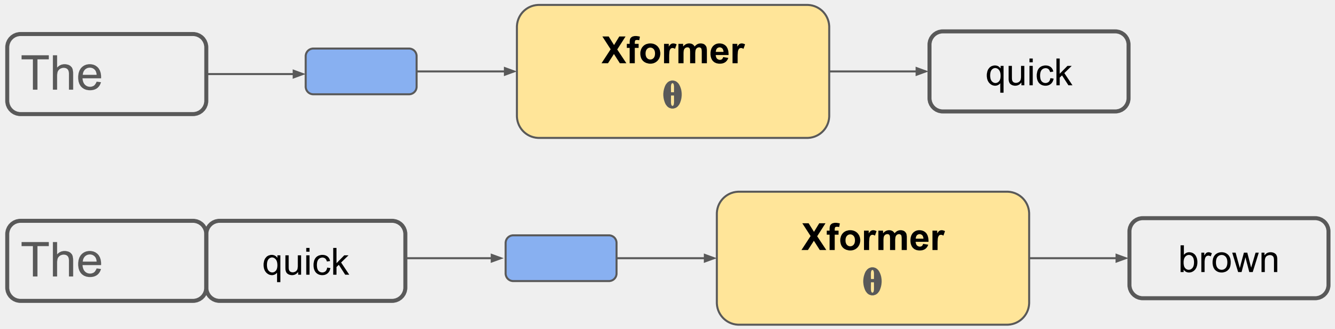
Language Models



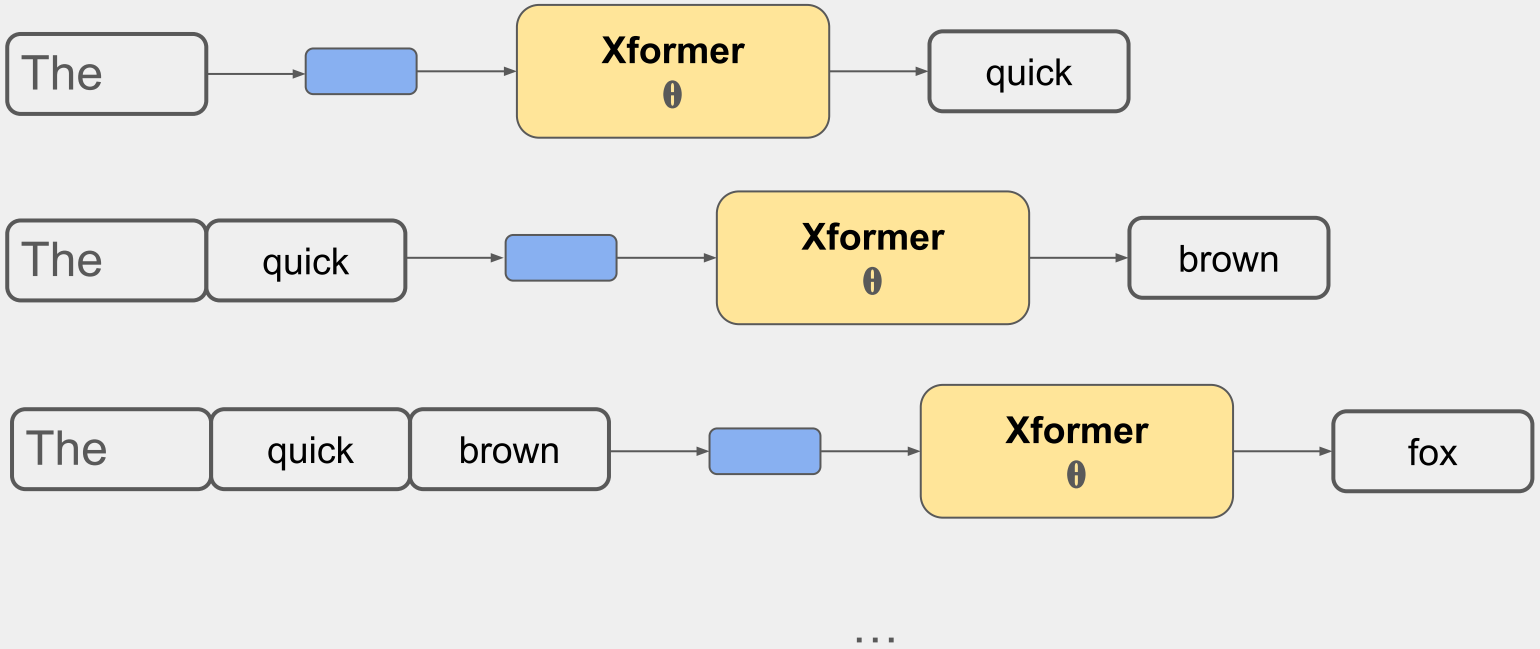
Why Diffusion Language Models?



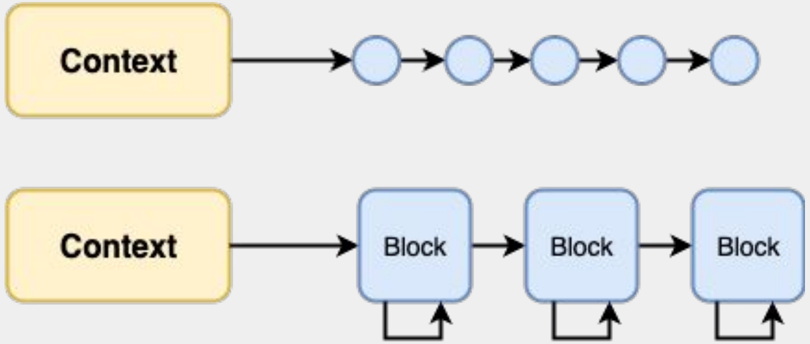
Why Diffusion Language Models?



Why Diffusion Language Models?



Diffusion vs. Autoregressive Language Models



1 token at a time :(

n tokens at a time :)

Diffusion vs. Autoregressive Language Models

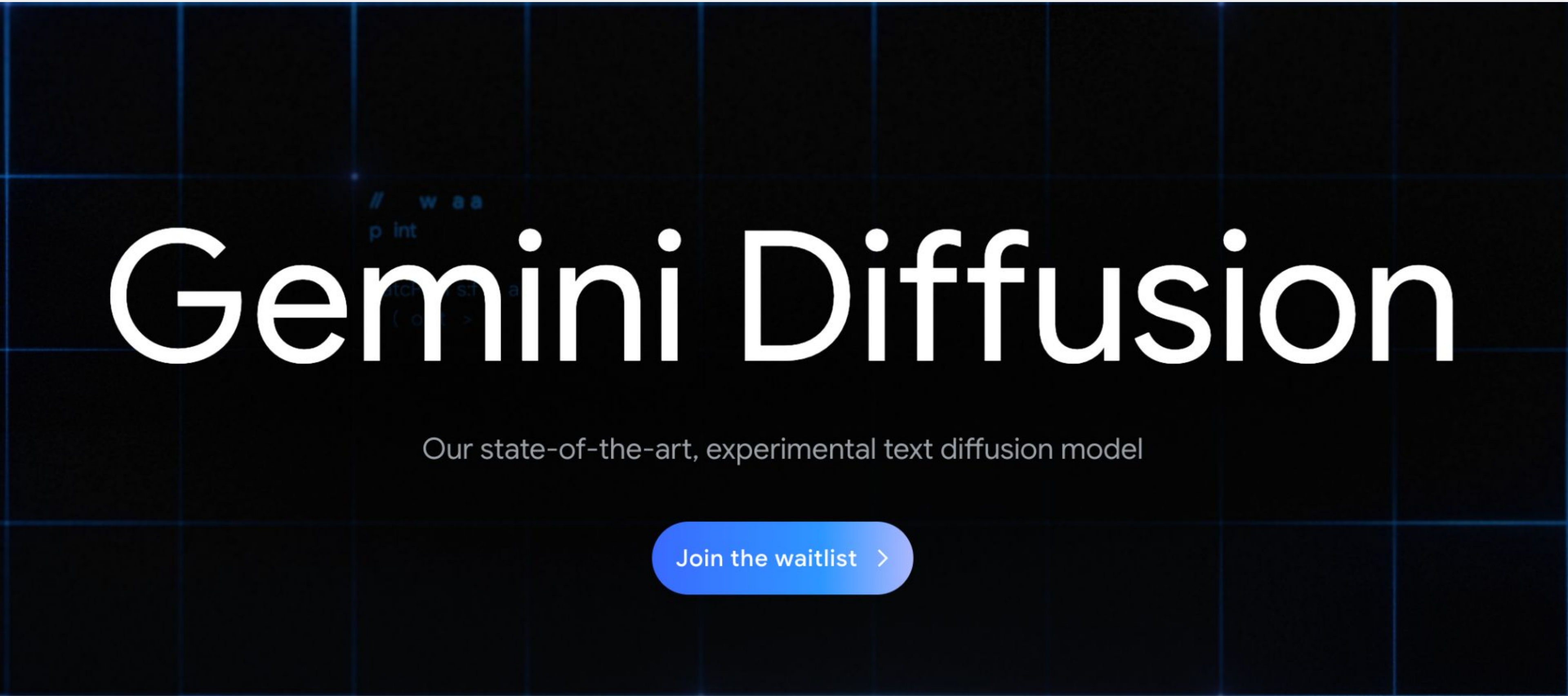
	Autoregression	Diffusion
High Quality	✓	✓
Arbitrary Length	✓	✓
KV Caching	✓	✓
Parallel Decoding	❌	✓

Diffusion vs. Autoregressive Language Models

	Autoregression	Diffusion
High Quality	✓	✓
Arbitrary Length	✓	✓
KV Caching	✓	✓
Parallel Decoding	❌	✓
Self-correction	❌	✓
Unify multimodality	❌	✓

2. The hype

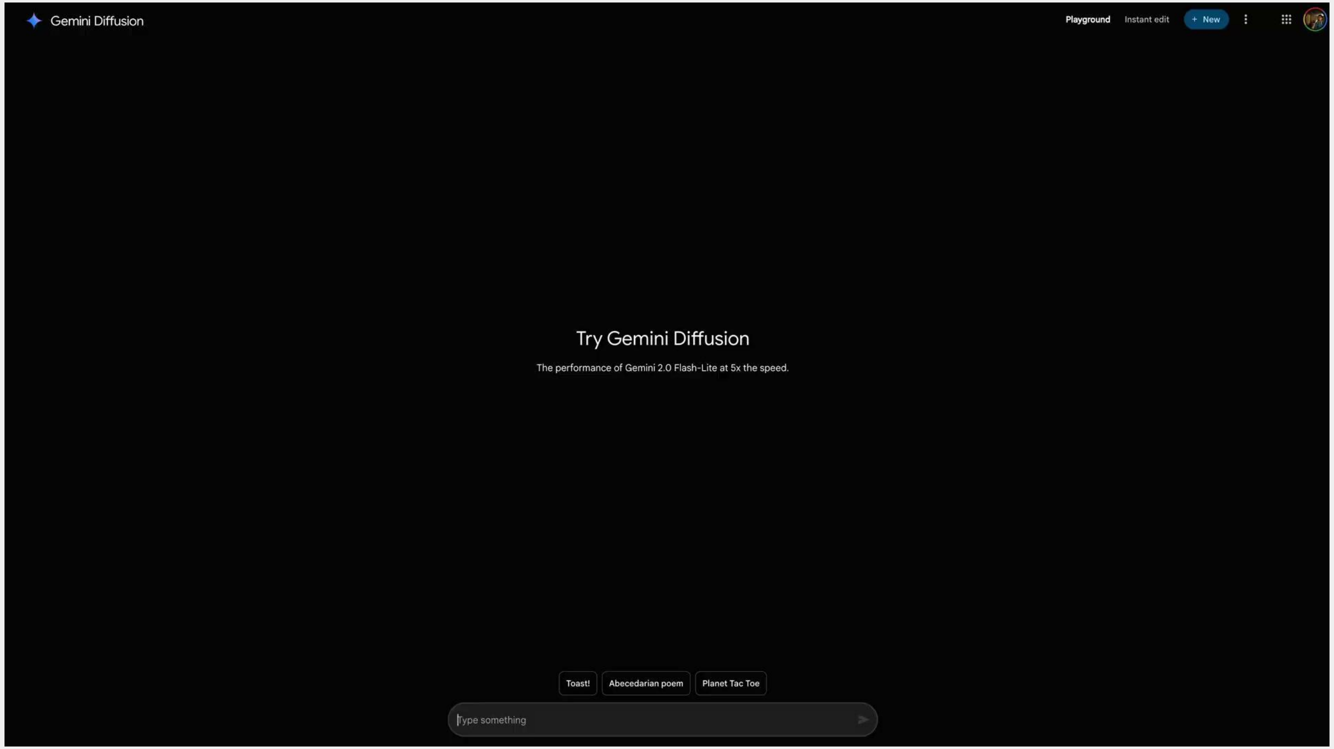
The hype



Gemini Diffusion

Our state-of-the-art, experimental text diffusion model

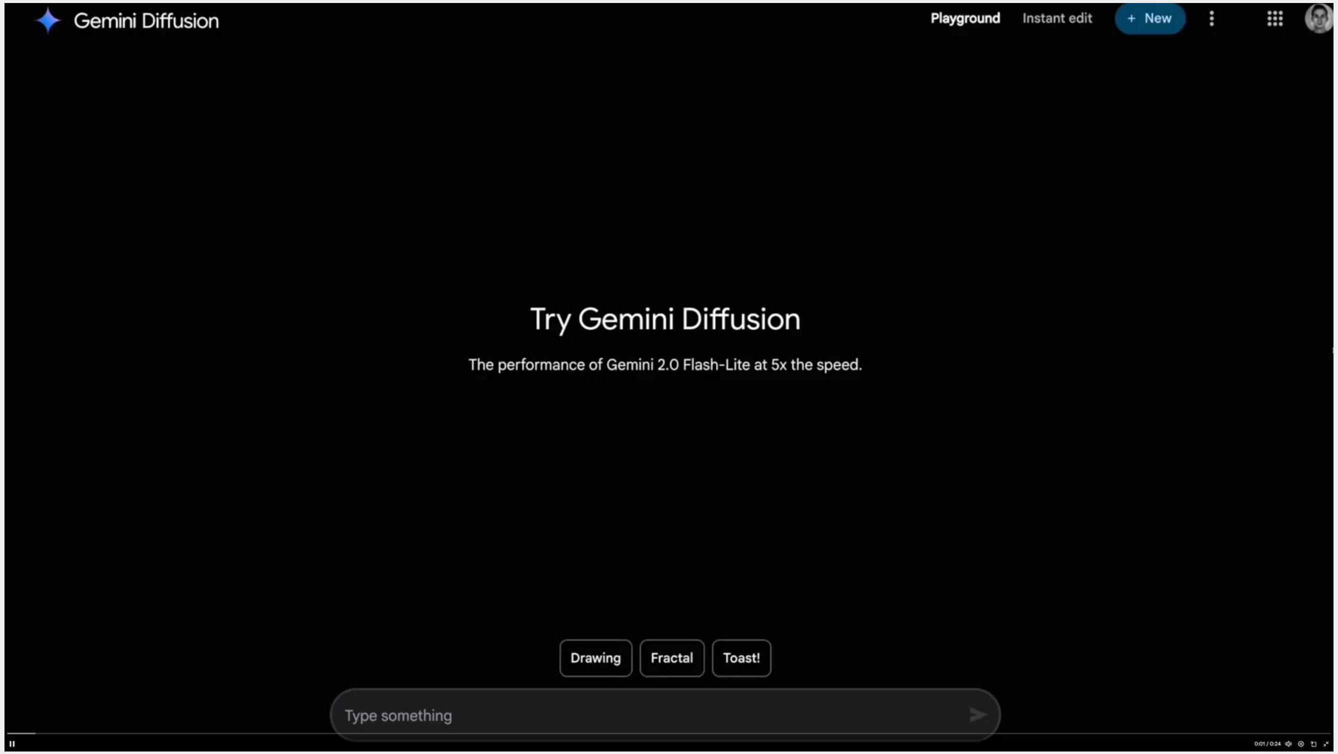
The hype



Try Gemini Diffusion

The performance of Gemini 2.0 Flash-Lite at 5x the speed.

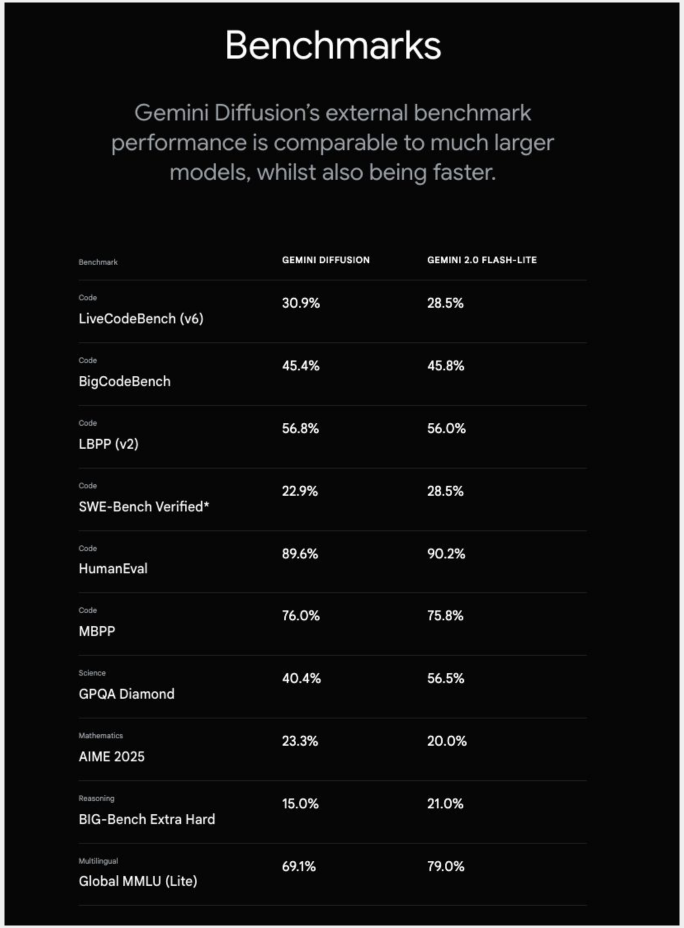
The hype



Try Gemini Diffusion

The performance of Gemini 2.0 Flash-Lite at 5x the speed.

The hype



Benchmarks

Gemini Diffusion's external benchmark performance is comparable to much larger models, whilst also being faster.

Benchmark	GEMINI DIFFUSION	GEMINI 2.0 FLASH-LITE
Code LiveCodeBench (v6)	30.9%	28.5%
Code BigCodeBench	45.4%	45.8%
Code LBPP (v2)	56.8%	56.0%
Code SWE-Bench Verified*	22.9%	28.5%
Code HumanEval	89.6%	90.2%
Code MBPP	76.0%	75.8%
Science GPQA Diamond	40.4%	56.5%
Mathematics AIME 2025	23.3%	20.0%
Reasoning BIG-Bench Extra Hard	15.0%	21.0%
Multilingual Global MMLU (Lite)	69.1%	79.0%

The hype

ByteDance | Seed

inception

Machine Learning Research

The hype

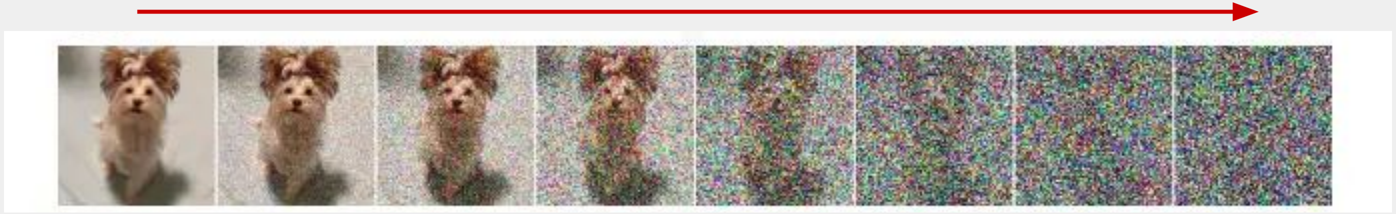
Nie et al., "Large Language Diffusion Models," arXiv:2502.09992, 2025.

3. How?

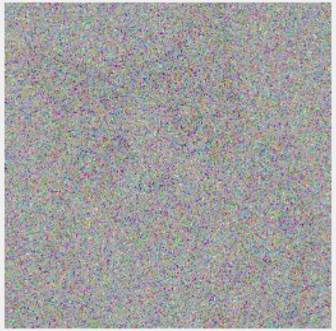
3. How?
(Part 1)

Image Diffusion

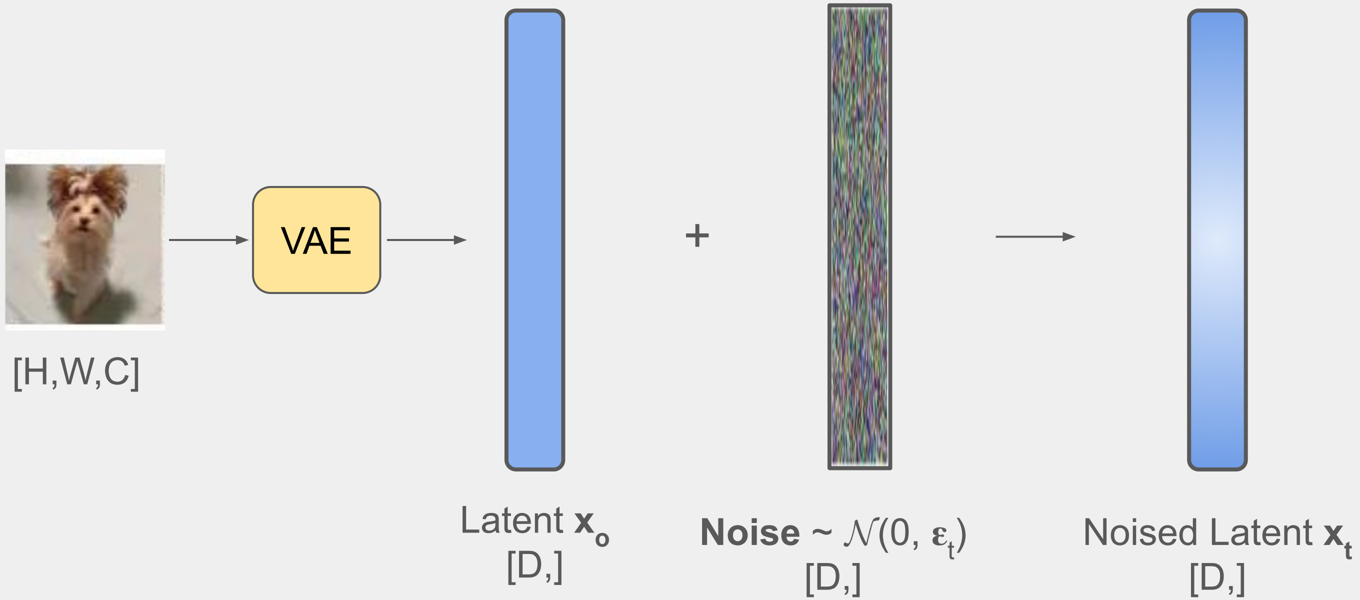
Forward Process



Reverse Process

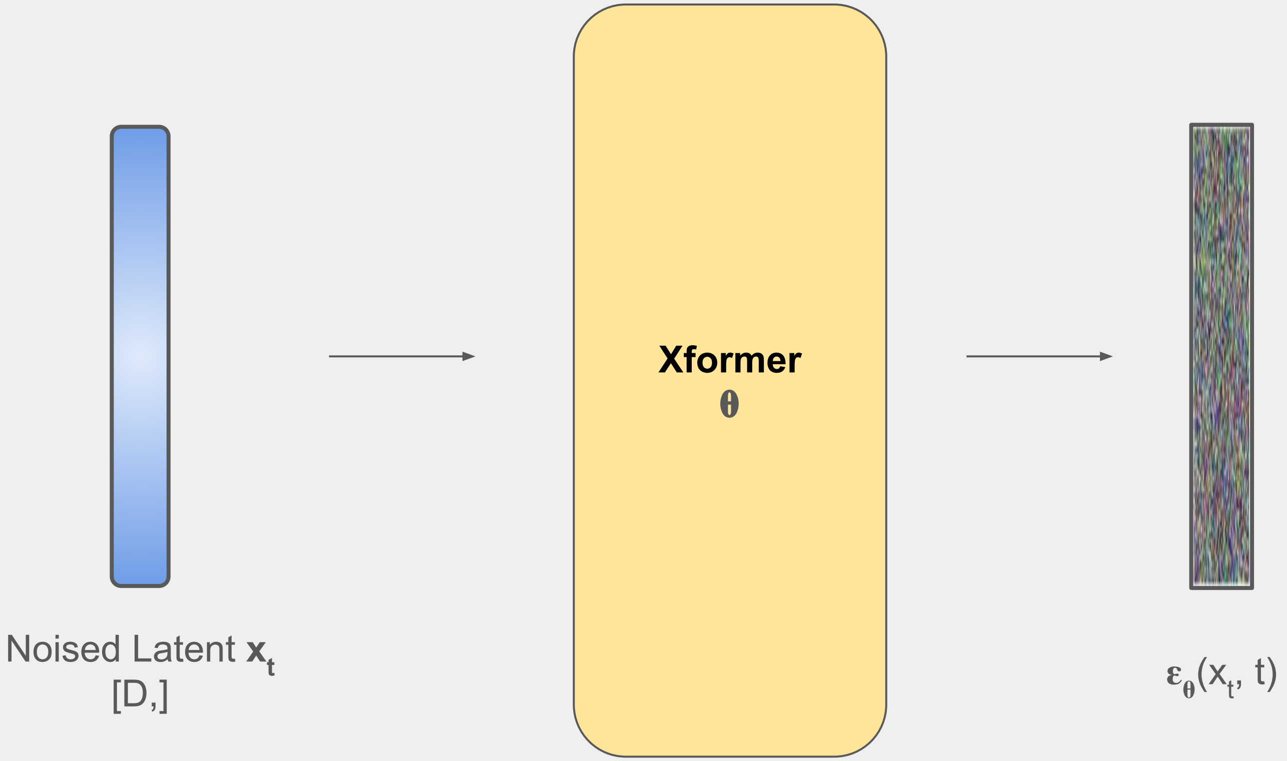


Latent Diffusion


[H,W,C]
VAE
Latent $\mathbf{x}_o$
[D,]
+
Noise ~ $\mathcal{N}(0, \varepsilon_t)$
[D,]
Noised Latent $\mathbf{x}_t$
[D,]

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1 - \bar{\alpha}_t}\, \varepsilon_t$$

Latent Diffusion



Latent Diffusion

Network prediction

Actual Noise

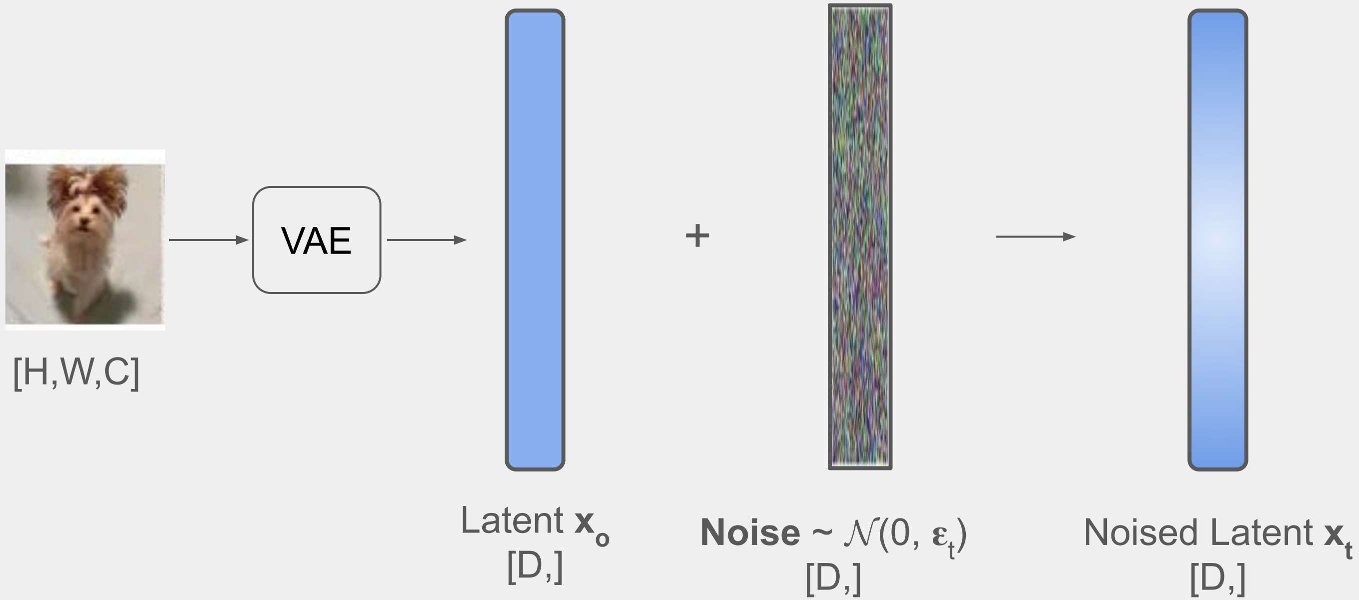
Train 🔥

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{x_0,\varepsilon,t}\left[\left\|\varepsilon_\theta(x_t,t) - \varepsilon\right\|^2\right]$$

Sample 🌊

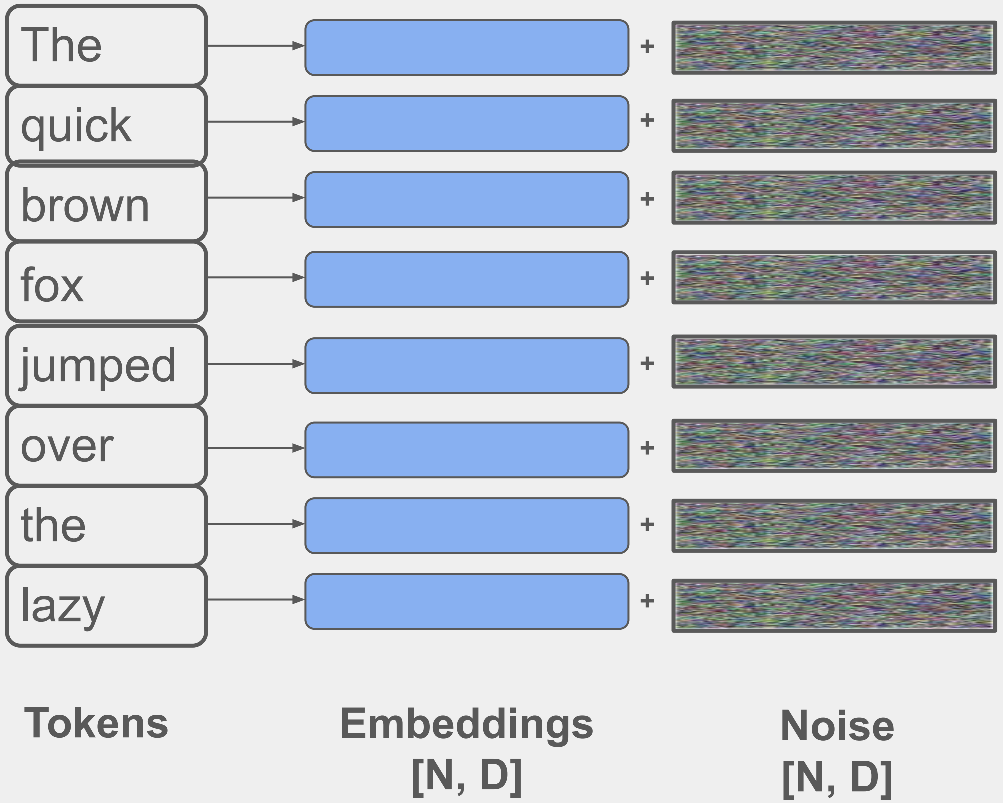
$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_\theta(x_t,t)\right) + \beta_t\varepsilon$$

Latent Diffusion

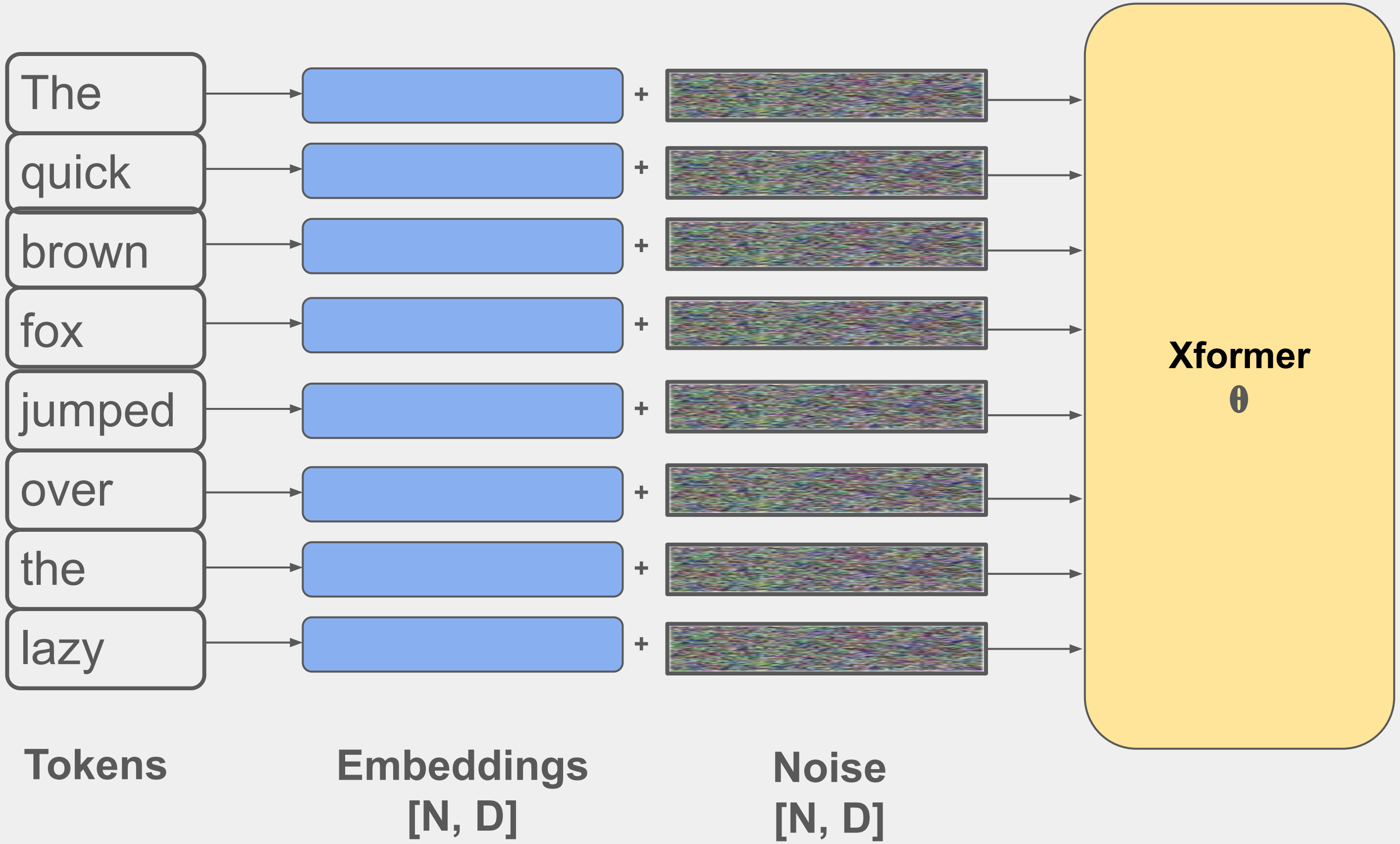

[H,W,C]
VAE
Latent $\mathbf{x}_0$
[D,]
+
Noise ~ $\mathcal{N}(0, \varepsilon_t)$
[D,]
Noised Latent $\mathbf{x}_t$
[D,]

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \varepsilon_t$$

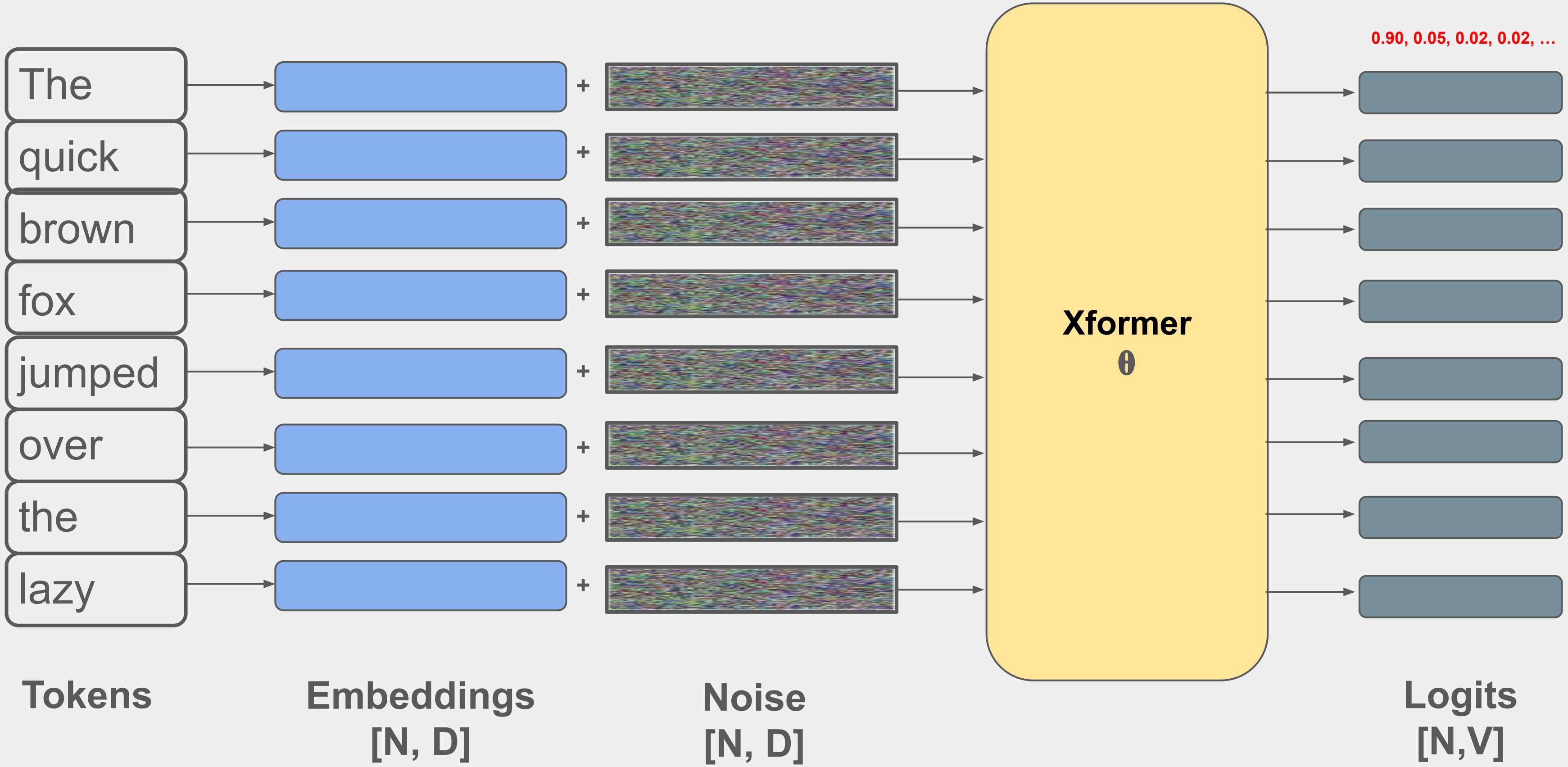
Latent Diffusion Language Models



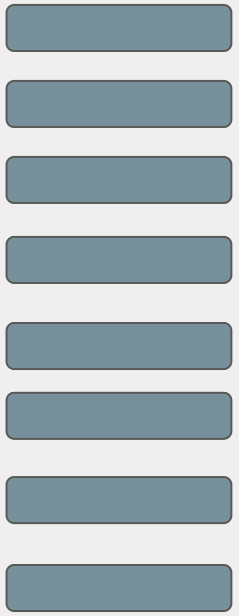
Latent Diffusion Language Models



Latent Diffusion Language Models



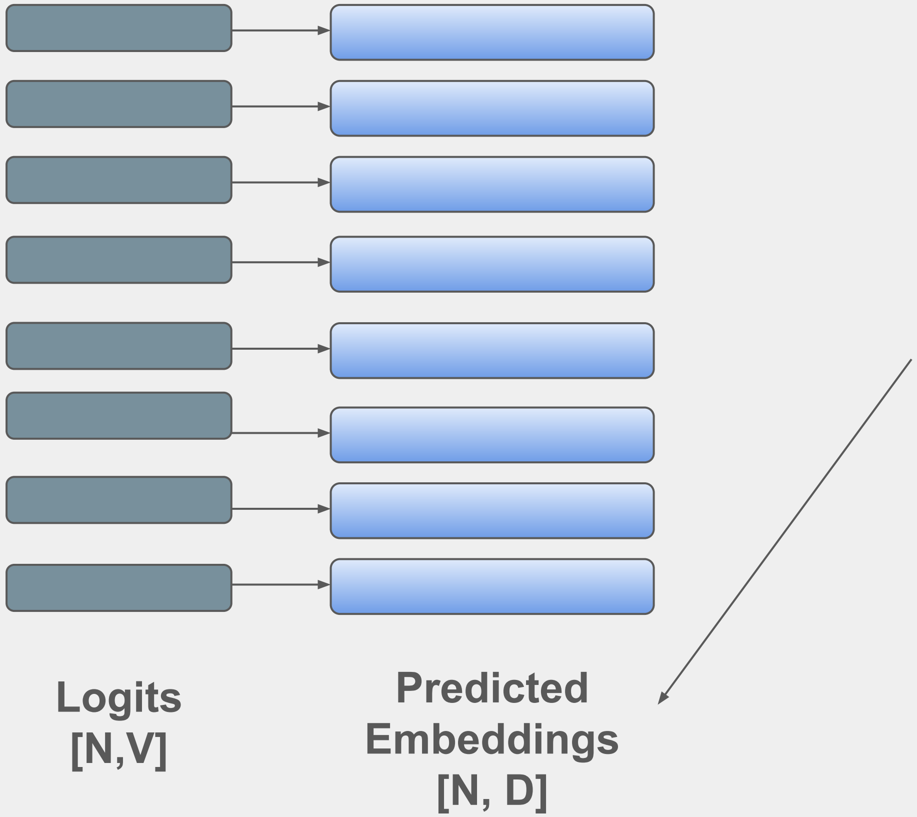
Latent Diffusion Language Models



For each logit, compute:

$$\mathbb{E}_{p(x_0|x,t)}[x_0] = \sum_{i=1}^{V} p(x_0 = e_i|x,t) \cdot e_i$$

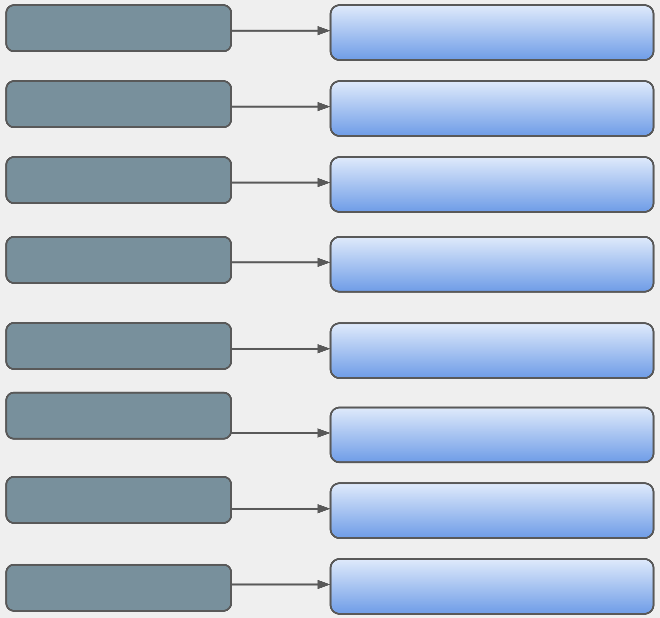
Latent Diffusion Language Models



$$\mathbb{E}_{p(x_0|x,t)}[x_0] = \sum_{i=1}^{V} p(x_0 = e_i|x,t) \cdot e_i$$

Let’s say we’re good at predicting these

Latent Diffusion Language Models



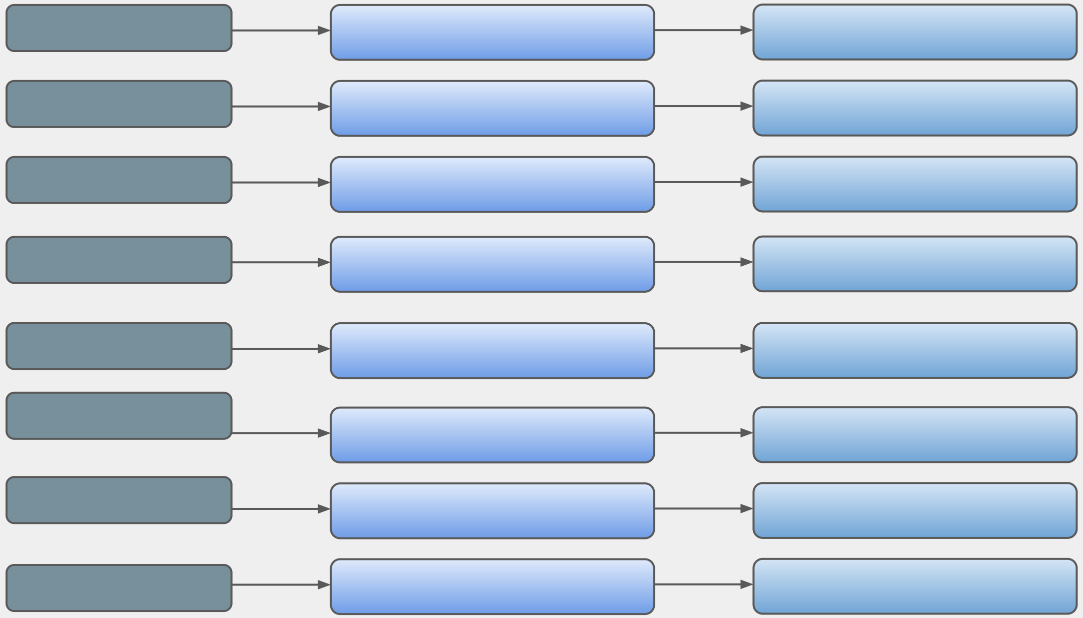
$$\nabla_x \log p_t(x) = s(x,t)$$
$$s(x,t) \approx s(x,t \mid x_0) = \frac{1}{t^2}(x_0 - x)$$

$$\hat{\mathbf{s}}(\mathbf{x},t) = \frac{\mathbb{E}_{p(x_0|\mathbf{x},t)}[\mathbf{x}_0] - \mathbf{x}}{t^2}$$

Dieleman et al., "Continuous Diffusion for Categorical Data," arXiv:2211.15089, 2022.

Latent Diffusion Language Models

$$\hat{\mathbf{s}}(\mathbf{x}, t) = \frac{\mathbb{E}_{p(x_0|\mathbf{x},t)}[\mathbf{x}_0] - \mathbf{x}}{t^2}$$

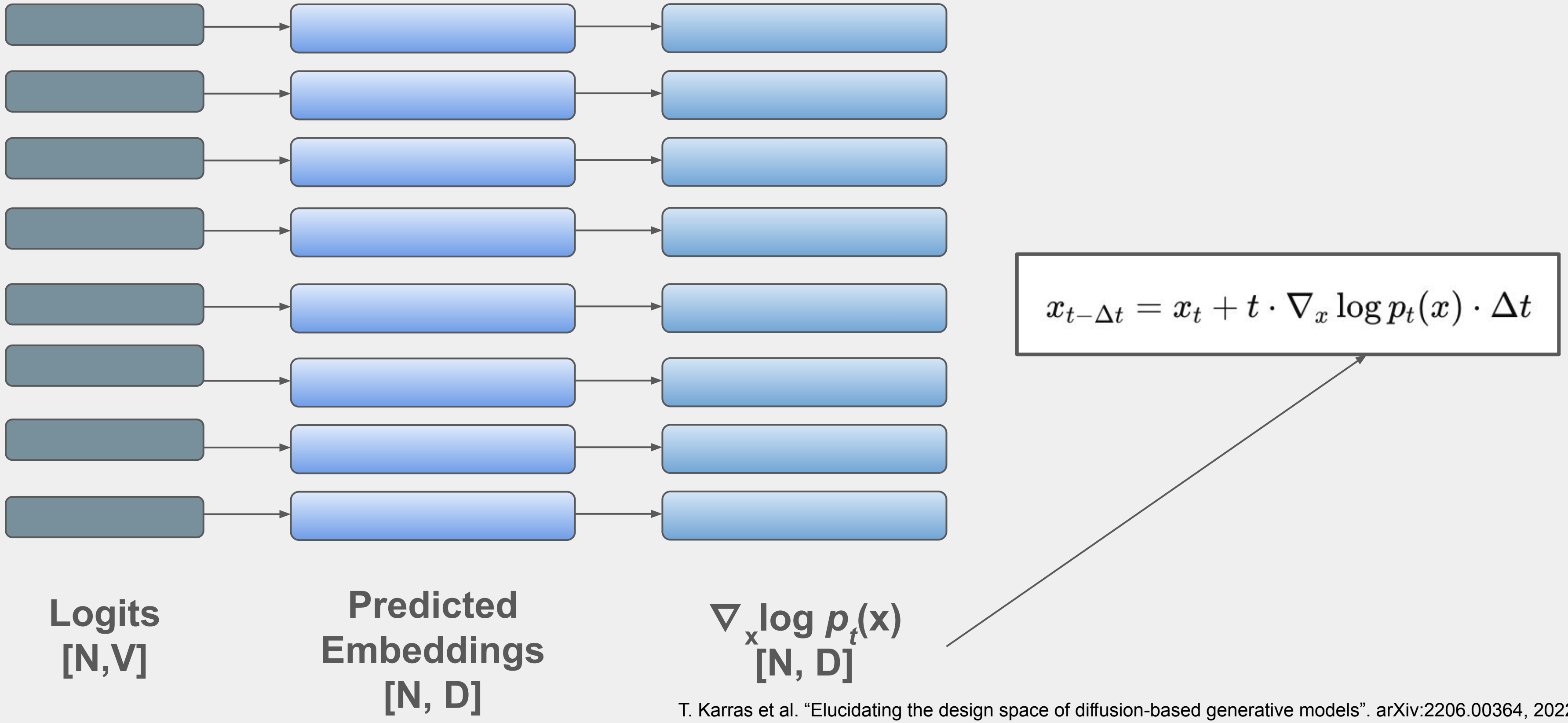


Logits
[N,V]

Predicted
Embeddings
[N, D]

$\nabla_x \log p_t(x)$
[N, D]

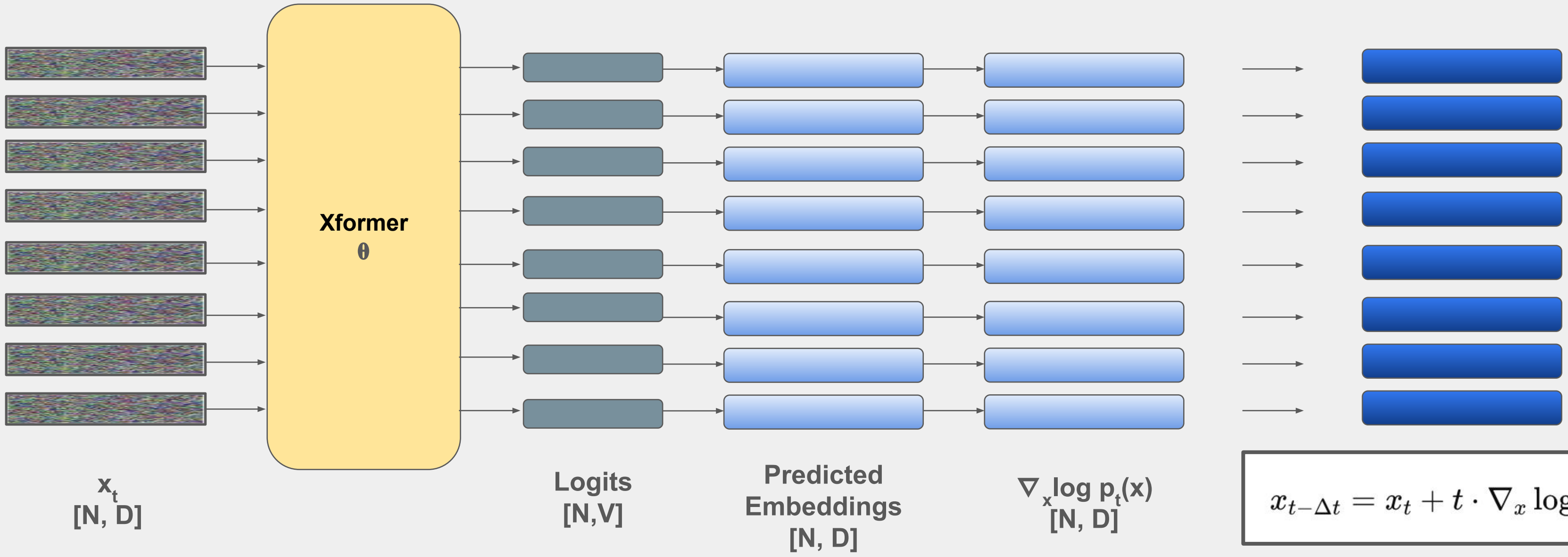
Latent Diffusion Language Models



$$x_{t-\Delta t} = x_t + t \cdot \nabla_x \log p_t(x) \cdot \Delta t$$

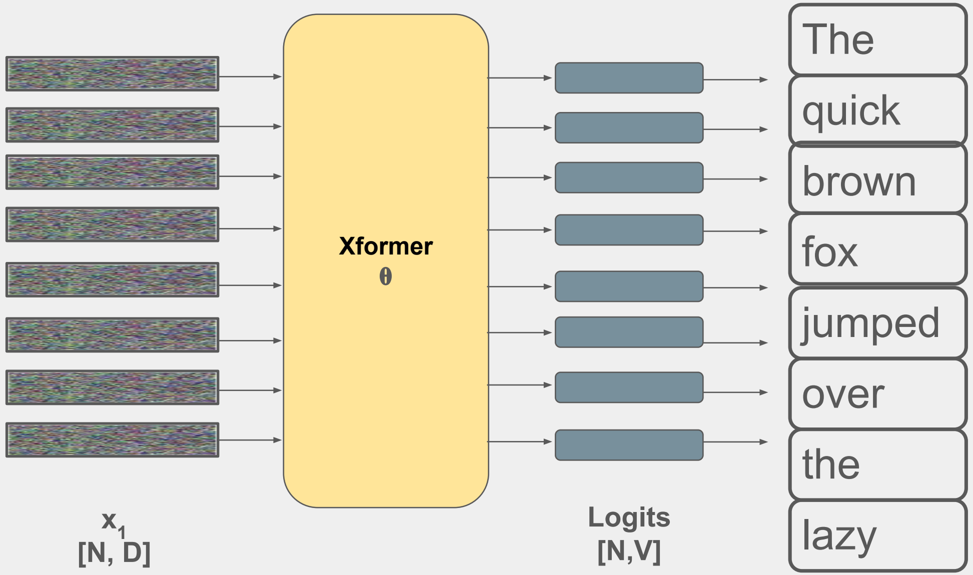
T. Karras et al. “Elucidating the design space of diffusion-based generative models”. arXiv:2206.00364, 2022.

Latent Diffusion Language Models - Sampling

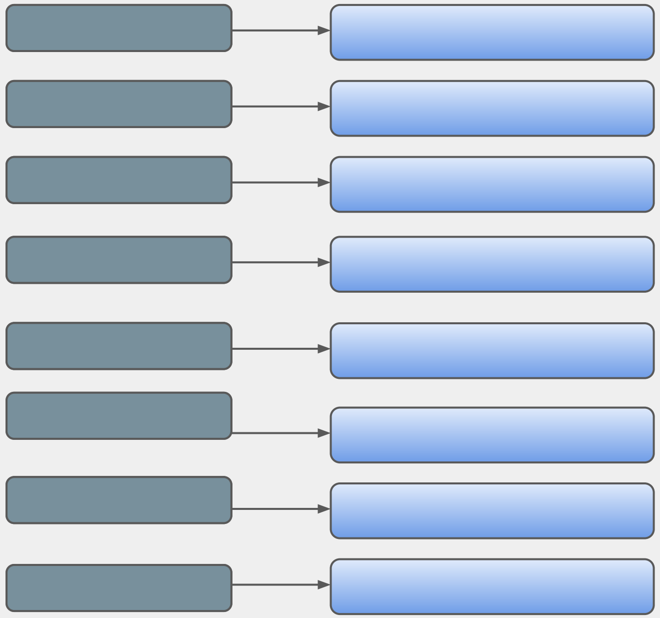


$$x_{t-\Delta t} = x_t + t \cdot \nabla_x \log p_t(x) \cdot \Delta t$$

Latent Diffusion Language Models - Sampling

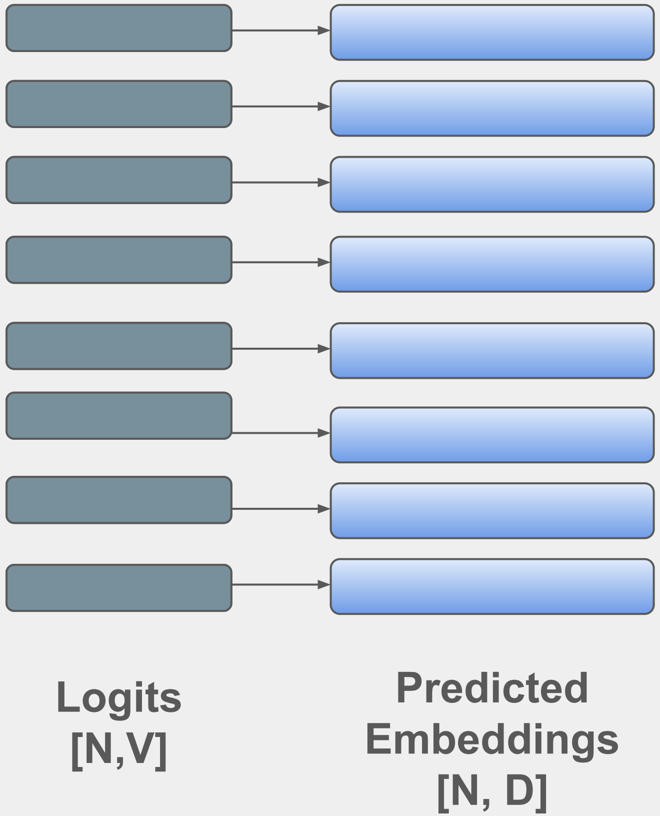


Latent Diffusion Language Models - Training



How do we get good at predicting x_0?

Latent Diffusion Language Models - Training



$$\left|\left|\mathbb{E}_{p_\theta(x_0|x,t)}[x_0] - x_0\right|\right|^2$$

Latent Diffusion Language Models - Training

$$\boxed{\mathcal{L}_{\text{mse}} = \left\| \sum_i p_\theta(i \mid x, t) \cdot e_i - e_{x_0} \right\|^2}$$

Dieleman et al., "Continuous Diffusion for Categorical Data," arXiv:2211.15089, 2022.

Latent Diffusion Language Models - Training

Turns out, we can also train the embeddings!

$$\boxed{\mathcal{L}_{\text{mse}} = \left\| \sum_i p_\theta(i \mid x, t) \cdot e_i - e_{x_0} \right\|^2}$$

Latent Diffusion Language Models - Training

Turns out, we can also train the embeddings!

$$\mathcal{L}_{\text{mse}} = \left\| \sum_i p_\theta(i \mid x, t) \cdot e_i - e_{x_0} \right\|^2$$

Not so fast!

Latent Diffusion Language Models - Training

Turns out, we can also train the embeddings!

$$\mathcal{L}_{\text{mse}} = \left\| \sum_i p_\theta(i \mid x, t) \cdot e_i - e_{x_0} \right\|^2$$

Not so fast!

The model can just learn to make all the embeddings the same!

Dieleman et al., "Continuous Diffusion for Categorical Data," arXiv:2211.15089, 2022.

Latent Diffusion Language Models - Training

$$\boxed{\mathcal{L}_{\text{mse}} = \left\| \sum_i p_\theta(i \mid x, t) \cdot e_i - e_{x_0} \right\|^2}$$

Turns out, we can also train the embeddings!

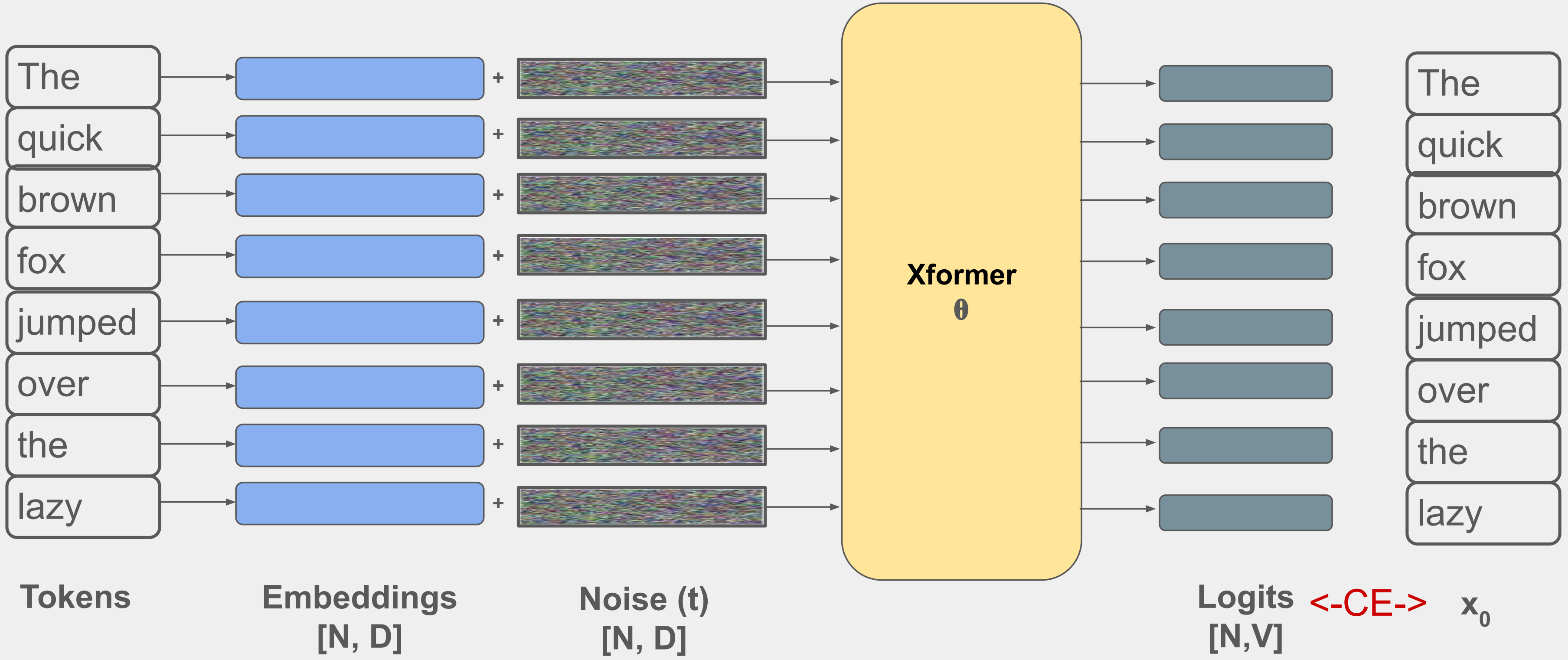
Not so fast!

The model can just learn to make all the embeddings the same!

Also, MSE is not a good loss for MSE data.

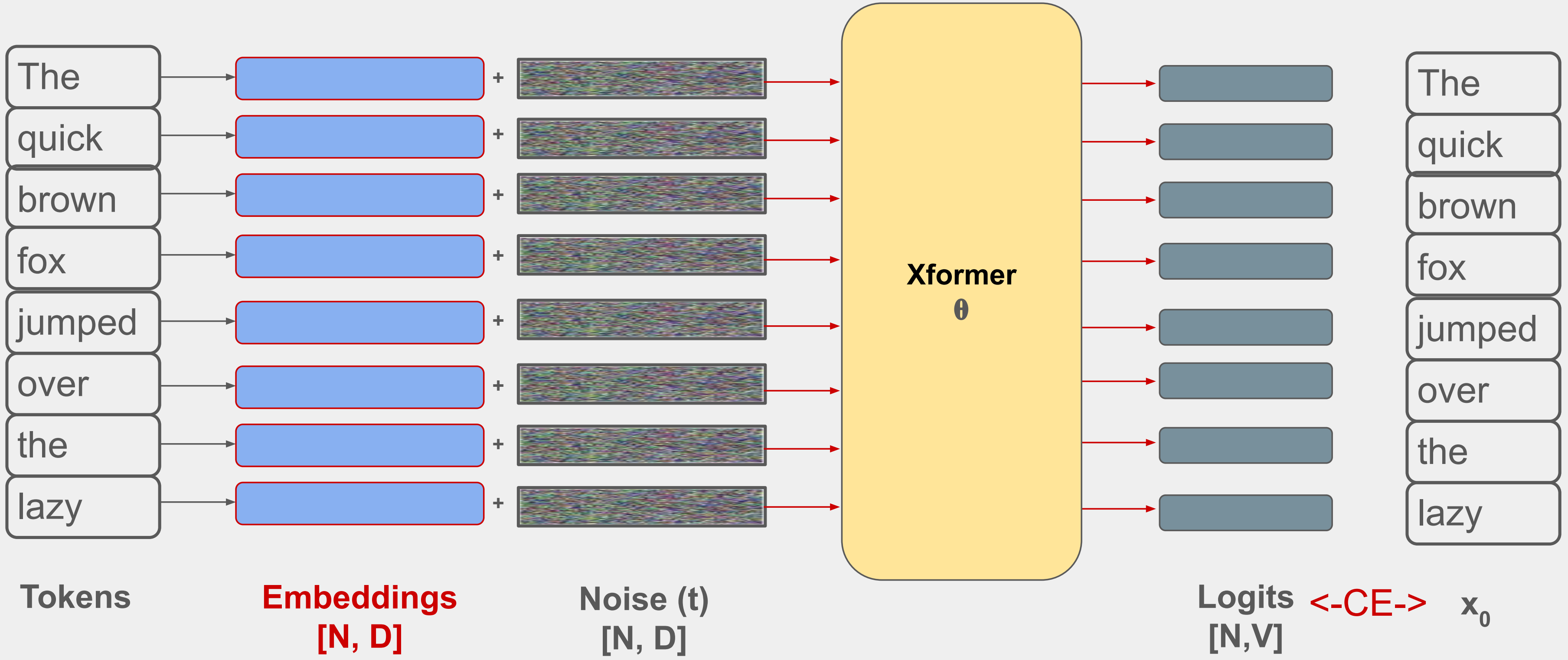
Dieleman et al., "Continuous Diffusion for Categorical Data," arXiv:2211.15089, 2022.

Latent Diffusion Language Models - Training



$$\boxed{\mathcal{L} = -\log p_\theta(x_0|x,t).}$$

Latent Diffusion Language Models - Training



$$\mathcal{L} = -\log p_\theta(x_0 | x, t).$$

Latent Diffusion Language Models - Training

Turns out, we can also train the embeddings!

Not so fast!

Latent Diffusion Language Models - Training

Turns out, we can also train the embeddings!

Not so fast!

The model can just learn to push embeddings to the extreme to make them easy to predict from noise!!*

3. How?

(Part Deux)

Latent Diffusion Language Models - Training

The model can just learn to push embeddings to the extreme to make them easy to predict from noise!!

- At scale, ^^ turns out to be a huge problem!
- The higher dimensional latent you use, the bigger the problem!
- How to learn embeddings suited for diffusion - high frequency to low frequency with gaussian noise?

ooof!

Discrete Diffusion Language Models

Continuous noising process	**Discrete noising process**
Embed tokens to latent space	Operate on tokens directly
Gaussian Noise added to embeddings	Flip tokens to other tokens in the vocabulary or MASK token
Uncertainty maintained until last step	Commit to tokens at each step
Difficulty scaling beyond 1B	Straightforward to scale to 8B

Discrete noising process

The	cat	ate	the	dog	food	t = 0
The	cat	ate	ઢ	dog	food	t = 0.1
The	吙	ate	ઢ	dog	Hi	t = 0.2
			…			
?	吙	З	ઢ	ফ	Hi	t = 1.0

Forward Process

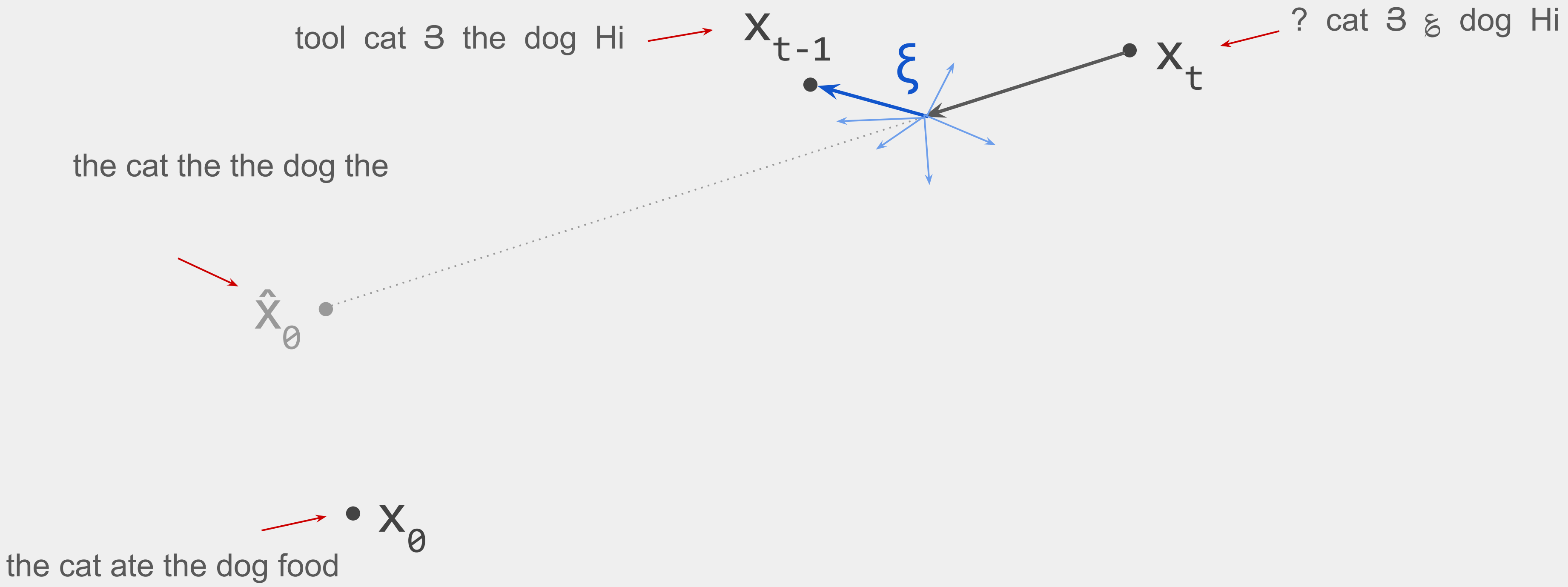
Discrete noising process

$$\text{The}\ \boxed{\text{[M]}}\ \text{ate}\ \boxed{\text{[M]}}\ \text{dog}\ \boxed{\text{[M]}}\ \xrightarrow{p_\theta(x_t,\, t)}\ \text{The}\ \boxed{\text{cat}}\ \text{ate}\ \boxed{\text{the}}\ \text{dog}\ \boxed{\text{food}}$$

$$-\mathbb{E}_{t,p_0,r_0,r_t}\left[\frac{1}{t}\sum_{i=1}^{L'}\mathbf{1}[r_t^i = \mathbf{M}]\log p_\theta(r_0^i|p_0,r_t)\right],$$

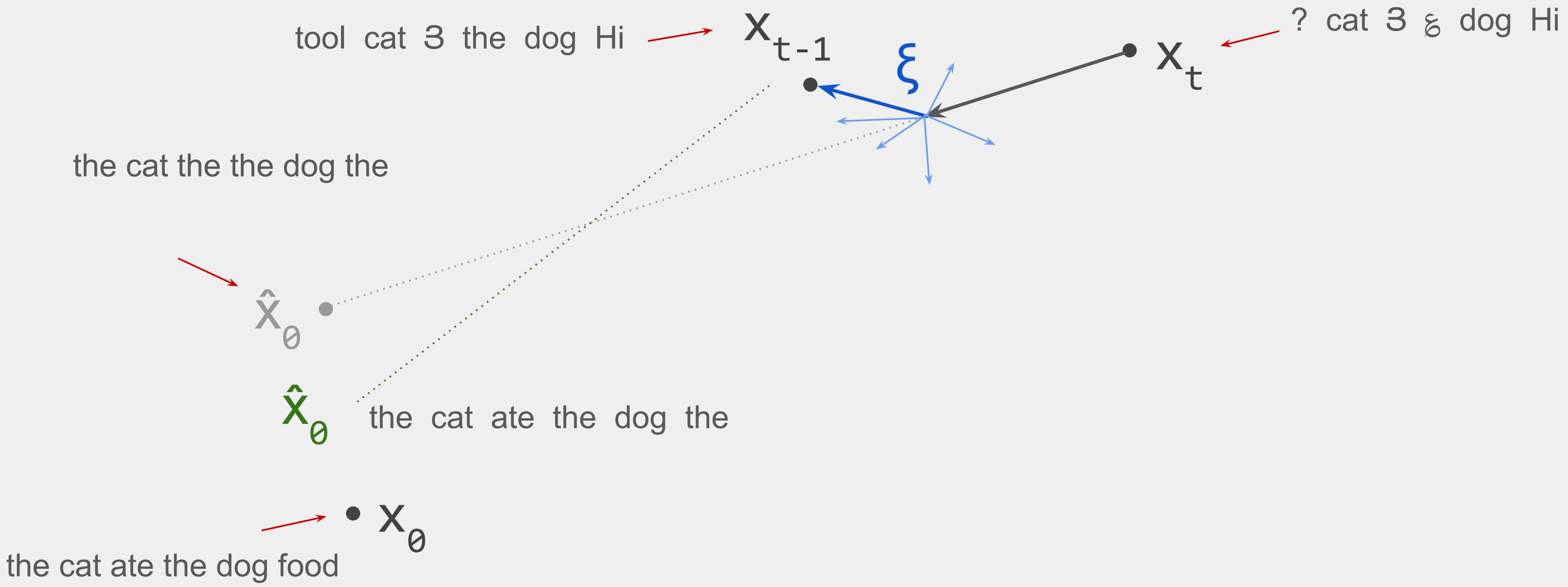
Nie et al., “Large Language Diffusion Models,” arXiv:2502.09992, 2025.

Discrete diffusion for text - Inference



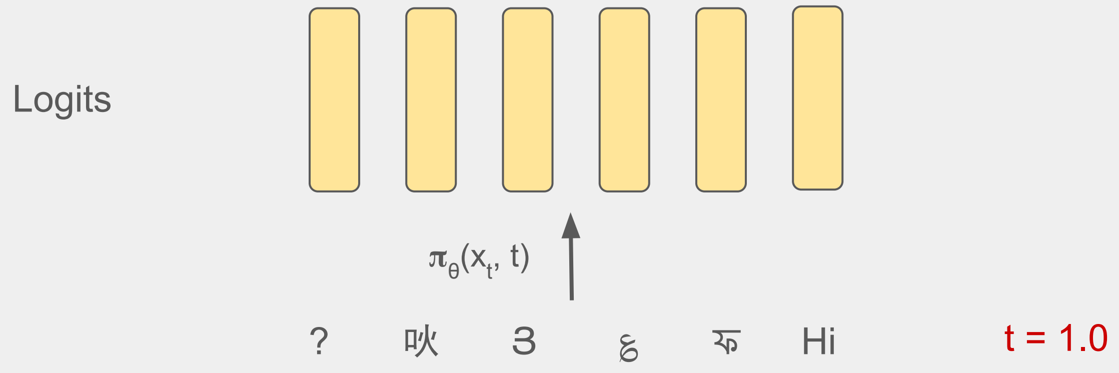
Slide credit: Sander Dieleman

Discrete diffusion for text - Inference

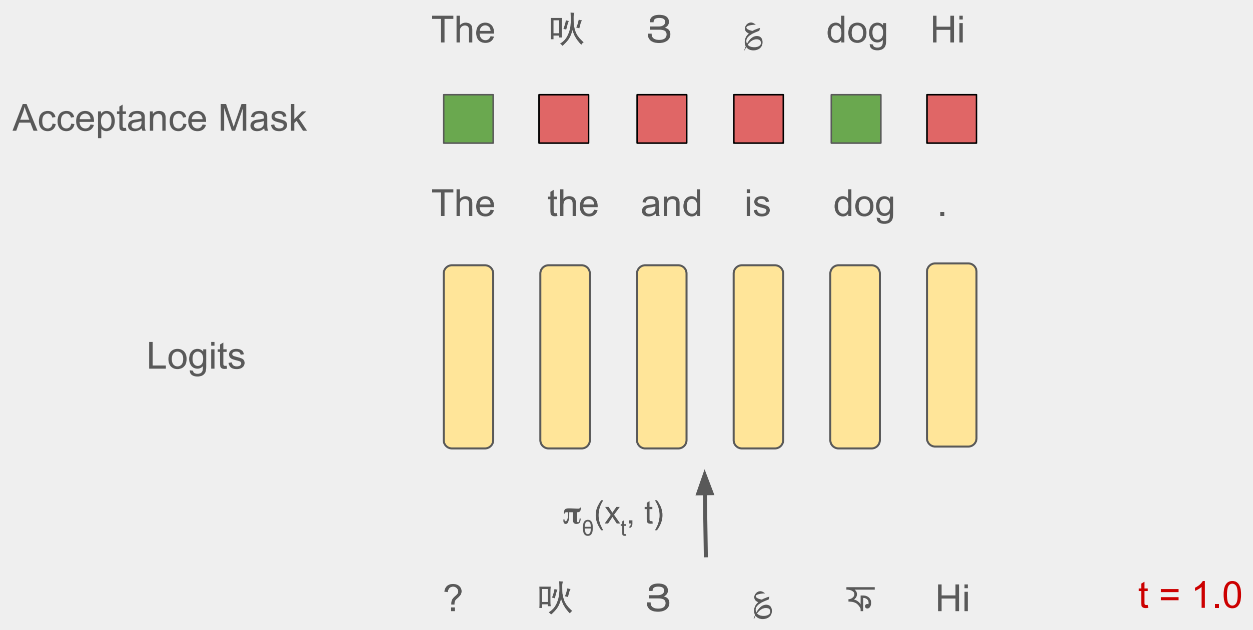


Slide credit: Sander Dieleman

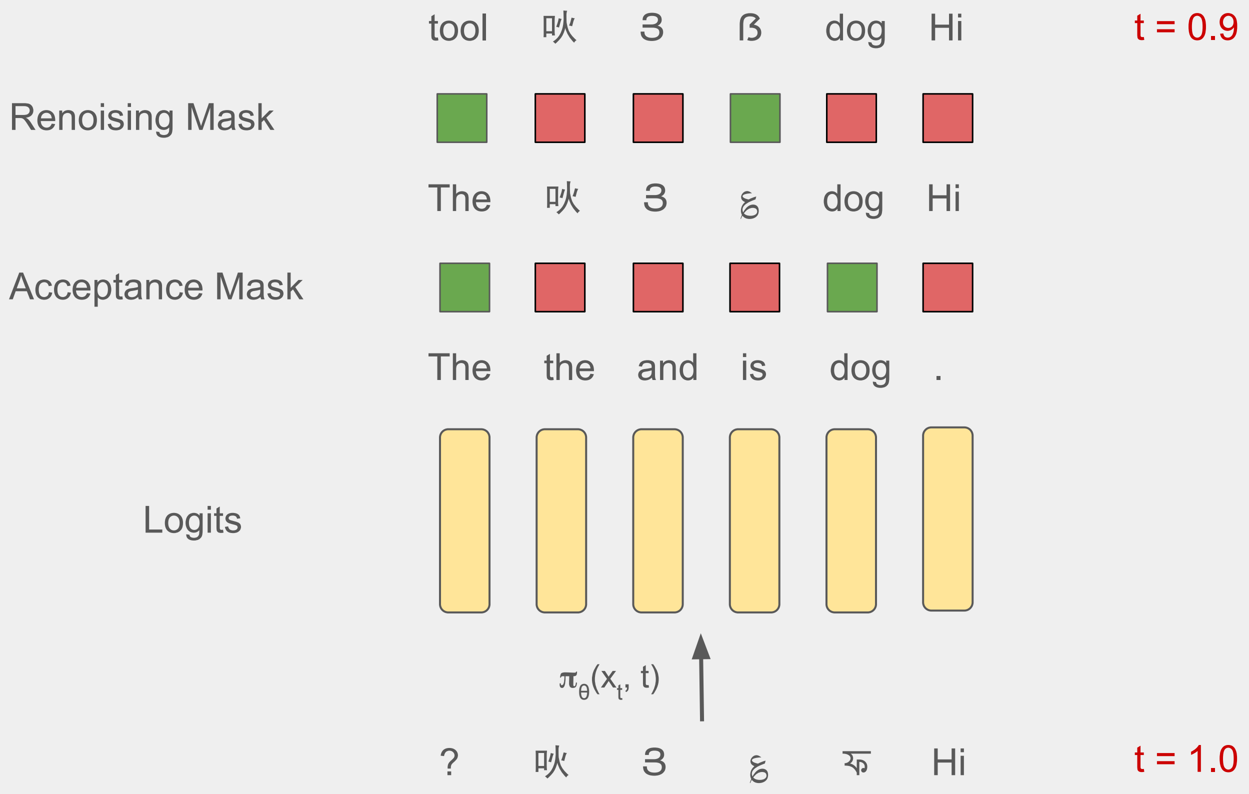
Discrete diffusion for text - Sampler



Discrete diffusion for text - Sampler



Discrete diffusion for text - Sampler



Discrete diffusion for text - Sampler

Many design choices and hypers

- Sampling temperature - $\mathbf{T}(t_n)$
- Acceptance probability profile - $\boldsymbol{p}_{\boldsymbol{a}}(t_n, t_{n-1})$
- Renoising probability profile - $\boldsymbol{p}_{\boldsymbol{r}}(t_{n-1})$
- Number of denoising steps - **N**

Discrete Diffusion Language Models are SOTA

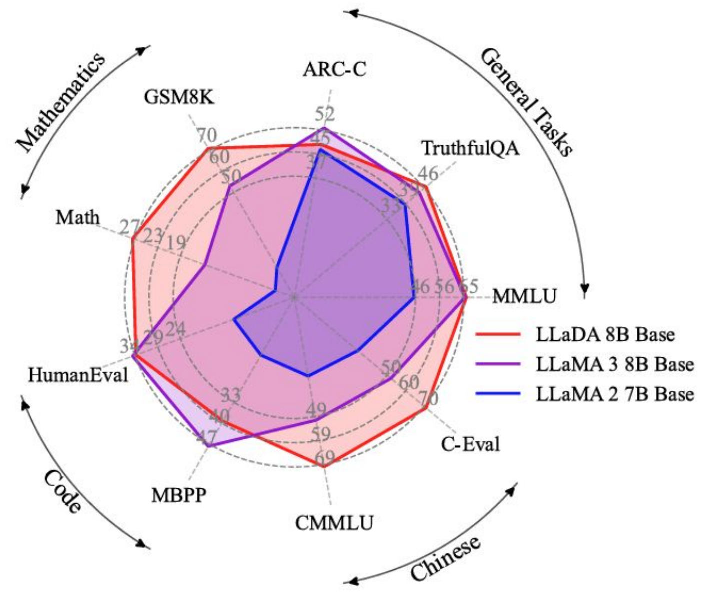
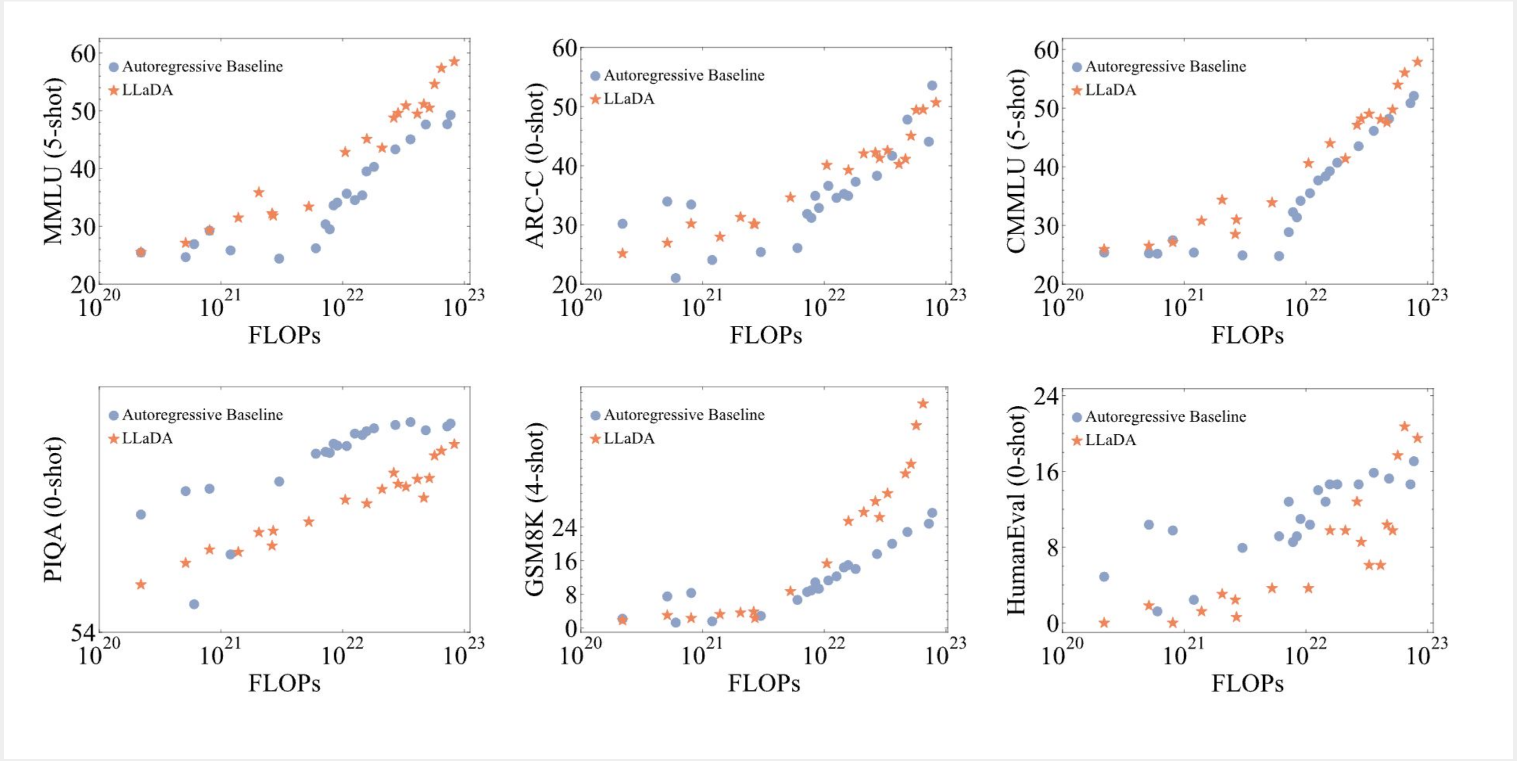


Figure 1. **Zero/Few-Shot Benchmarks.** We scale LLaDA to an unprecedented size of 8B parameters from scratch, achieving competitive performance with strong LLMs (Dubey et al., 2024).

Nie et al., "Large Language Diffusion Models," arXiv:2502.09992, 2025.

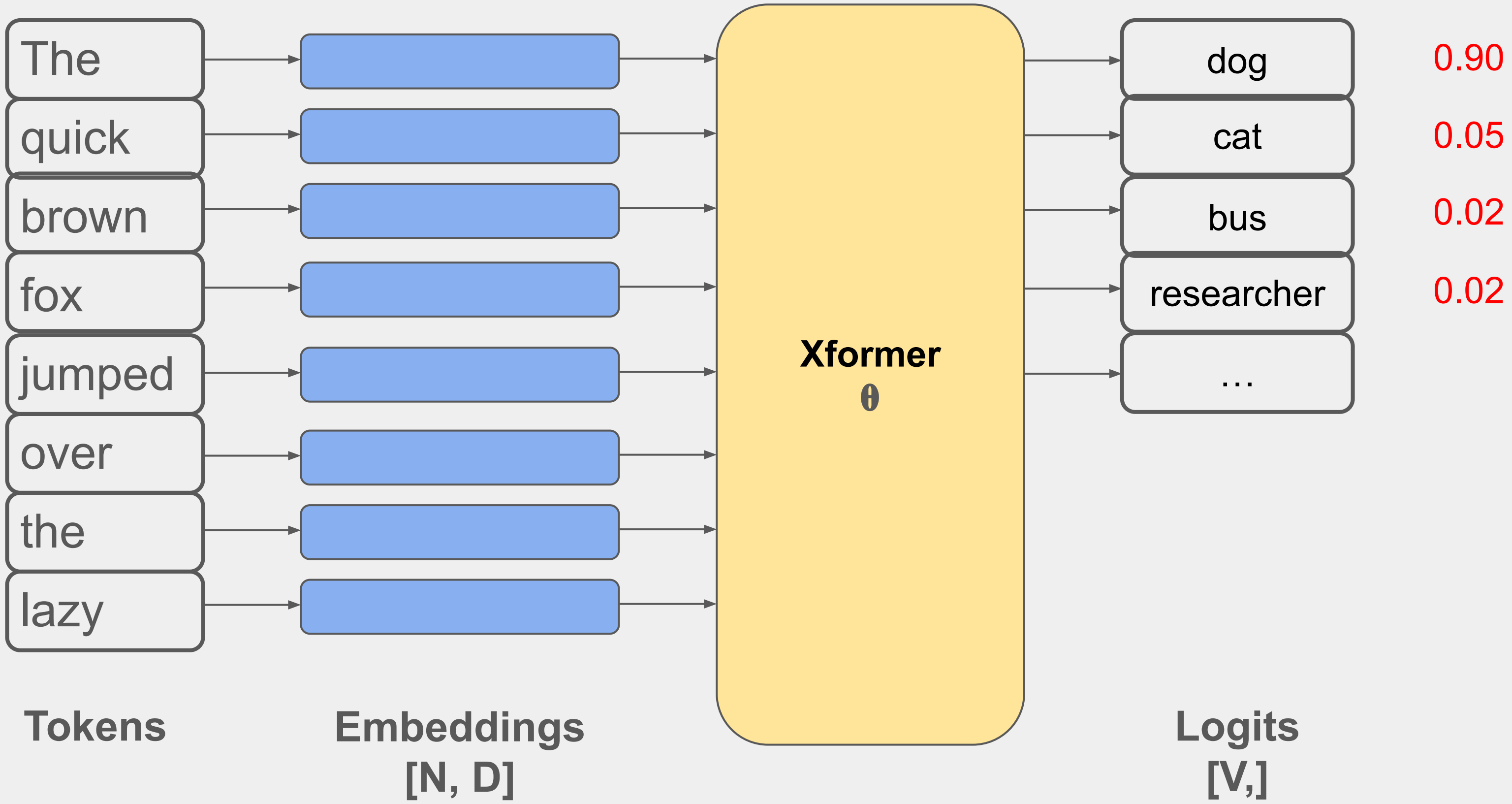
Discrete Diffusion Language Models are SOTA



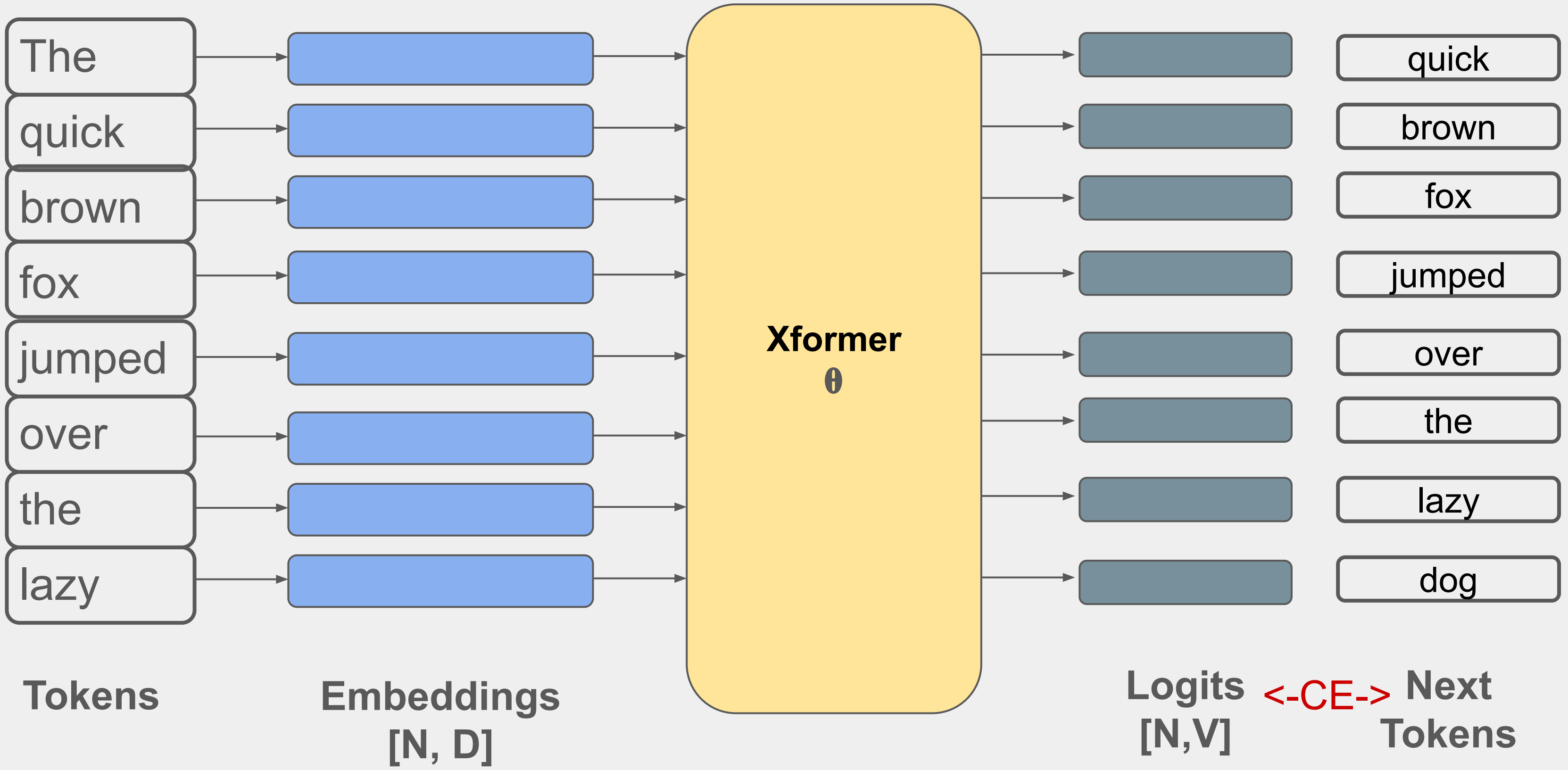
Nie et al., "Large Language Diffusion Models," arXiv:2502.09992, 2025.

3. How?
(Part tres)

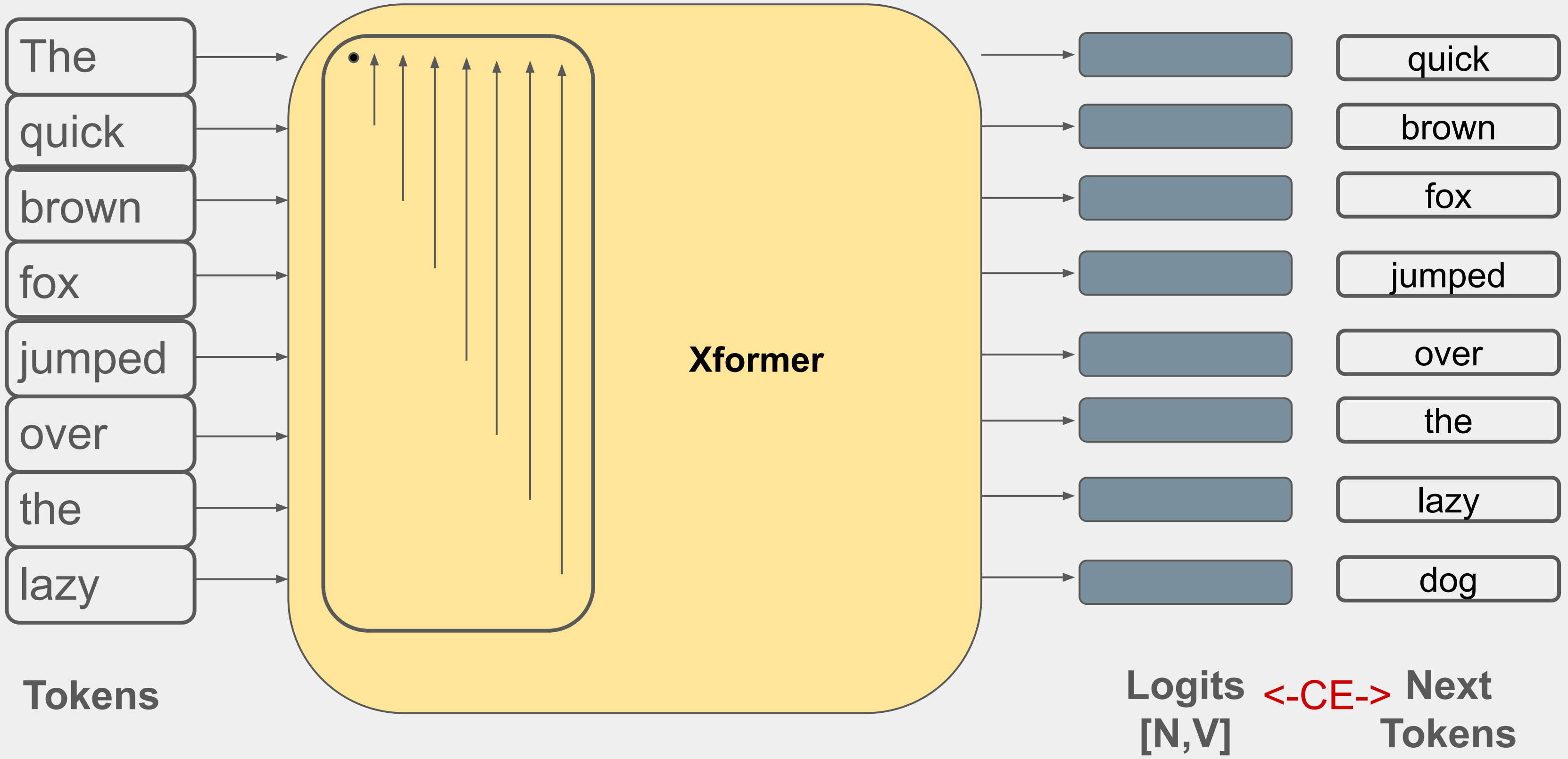
Language Models



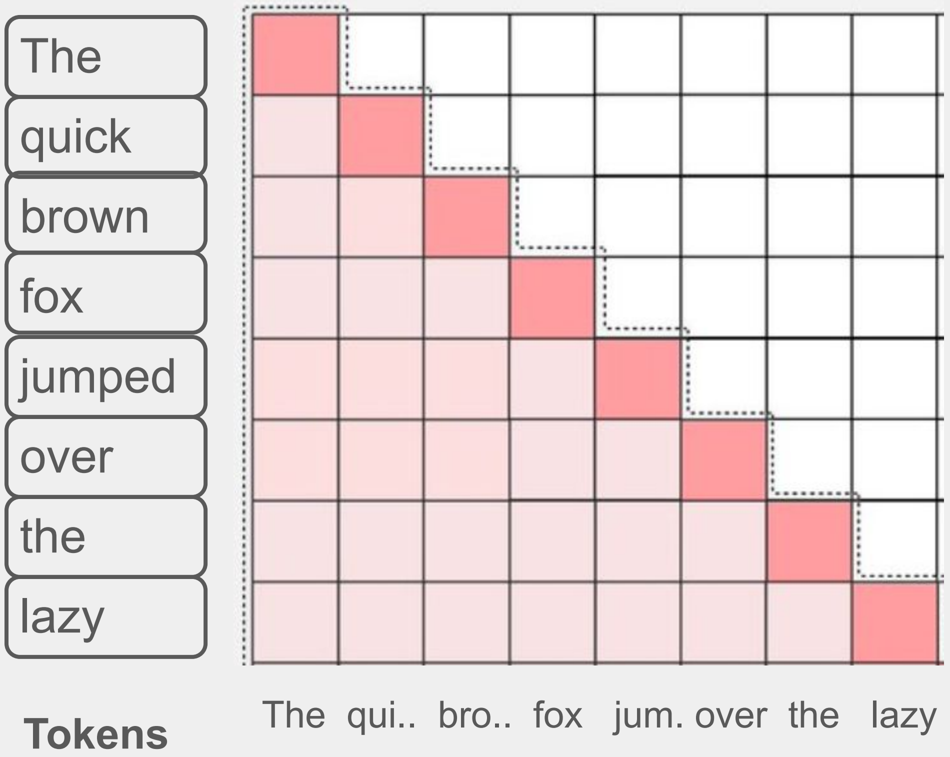
Language Models - Efficient Training



Language Models - Efficient Training

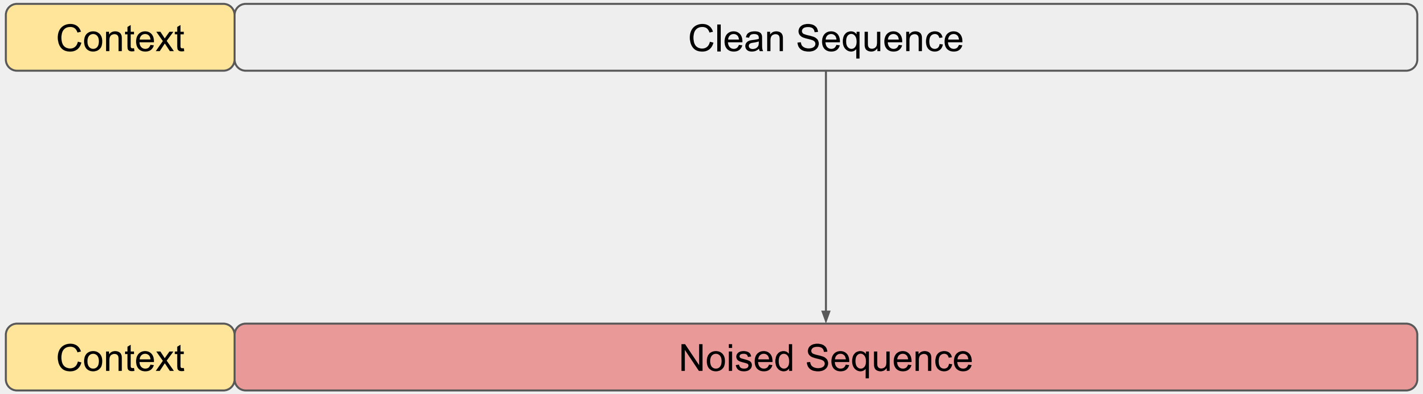


Language Models - Efficient Training

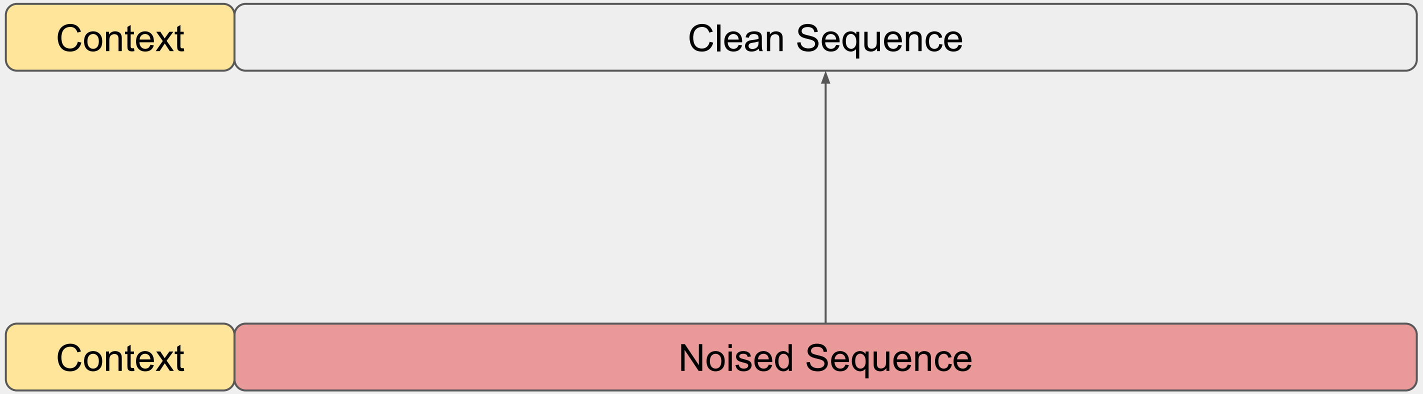


No for loops!

Discrete diffusion for text - Training



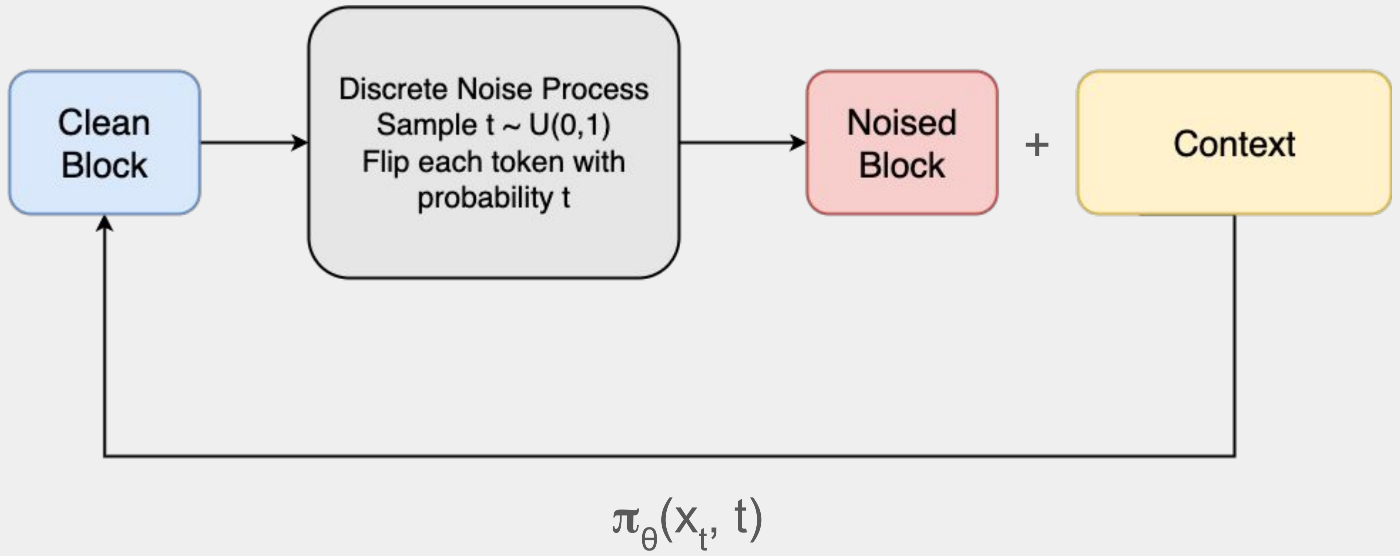
Discrete diffusion for text



[16384,] tokens

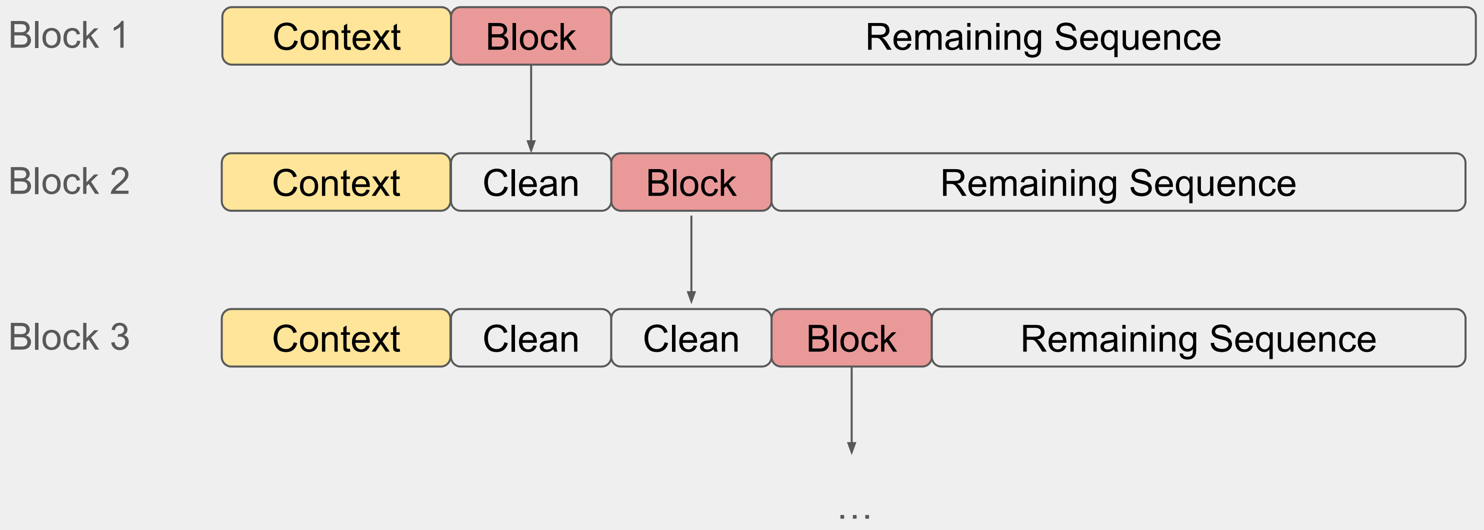
Full attention over all tokens can be very expensive!

Discrete diffusion for text - Training

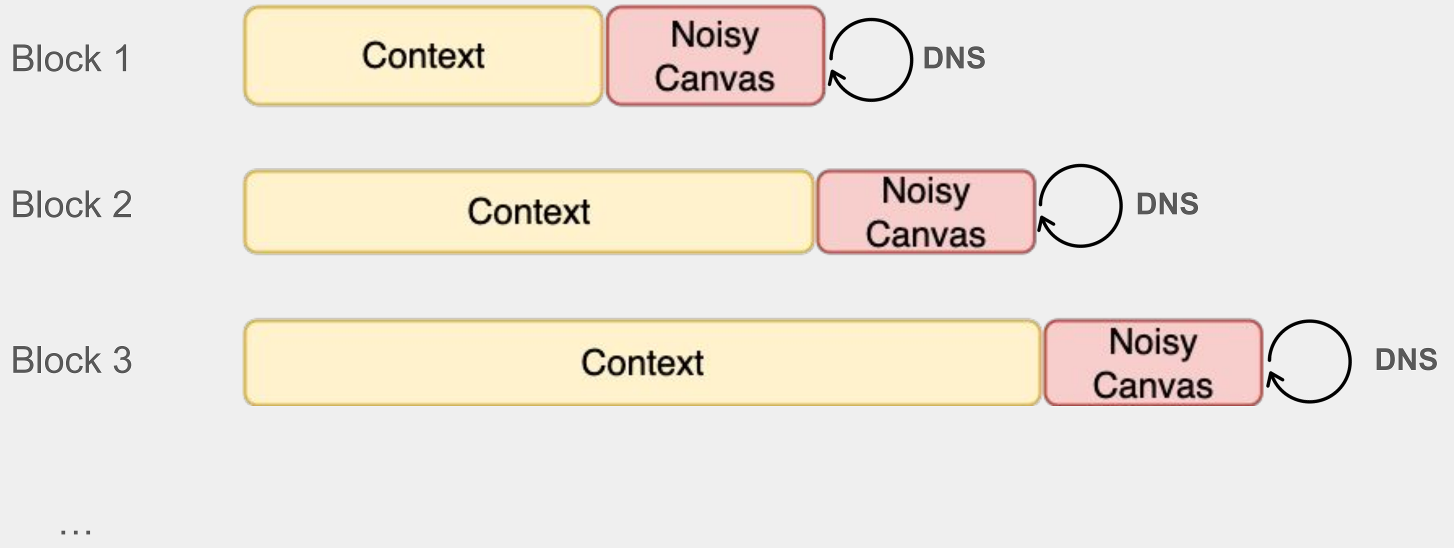


Discrete diffusion for text - Training

for block in blocks:



Discrete diffusion for text - Inference

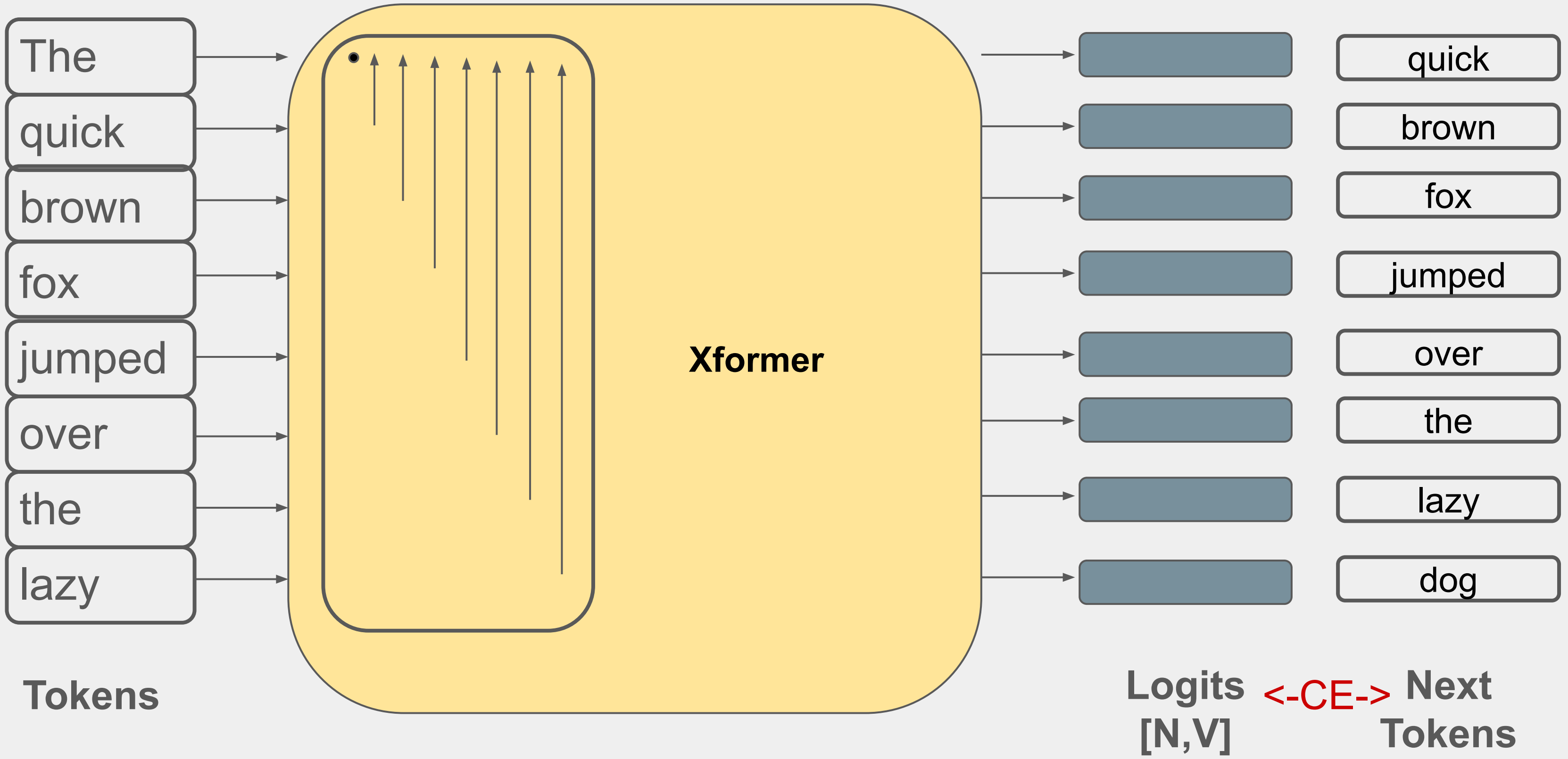


Enables KV caching between blocks by using block causal attention pattern!

Discrete diffusion for text - Training

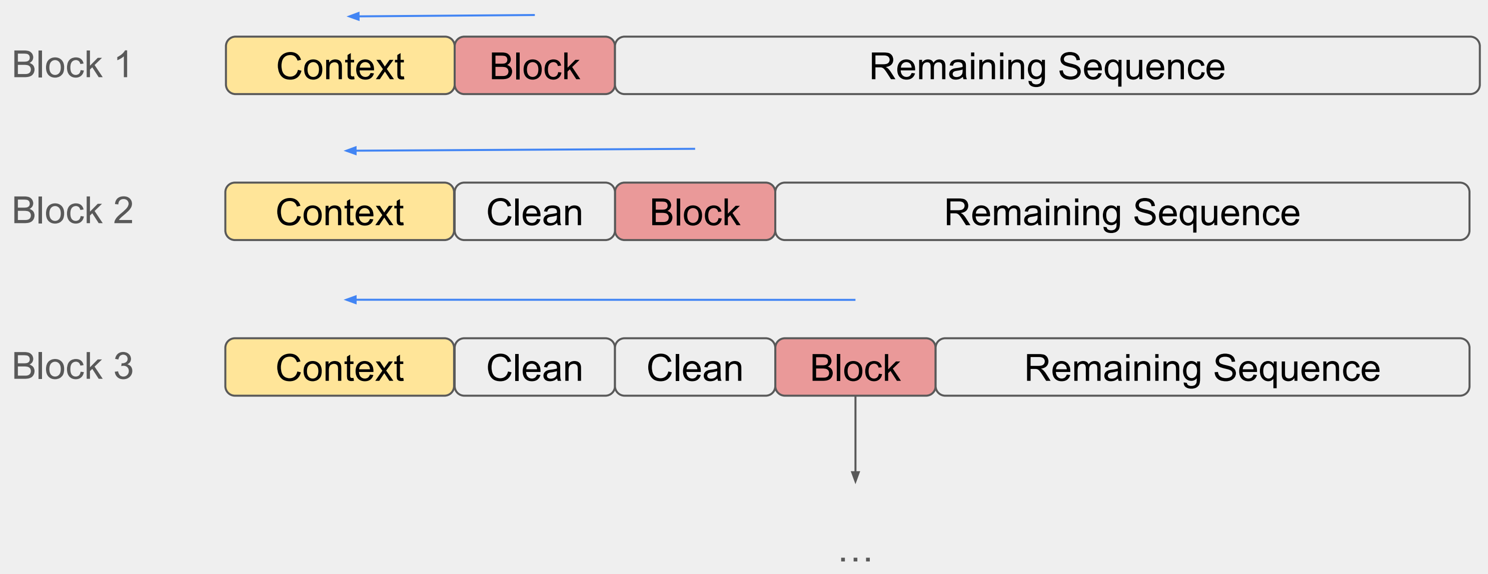
Can we do better?

Language Models - Efficient Training



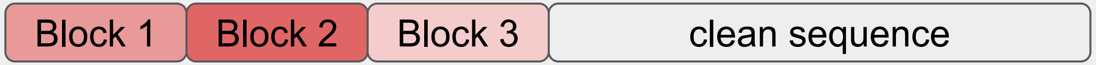
Discrete diffusion for text - Training

for block in blocks:



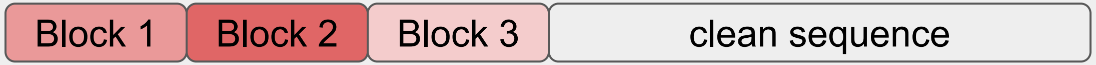
Discrete diffusion for text - Efficient Training

Concatenate all noised blocks with the clean sequence!



Arriola et al., "Block Diffusion," arXiv:2503.09573, 2025.

Discrete diffusion for text - Efficient Training



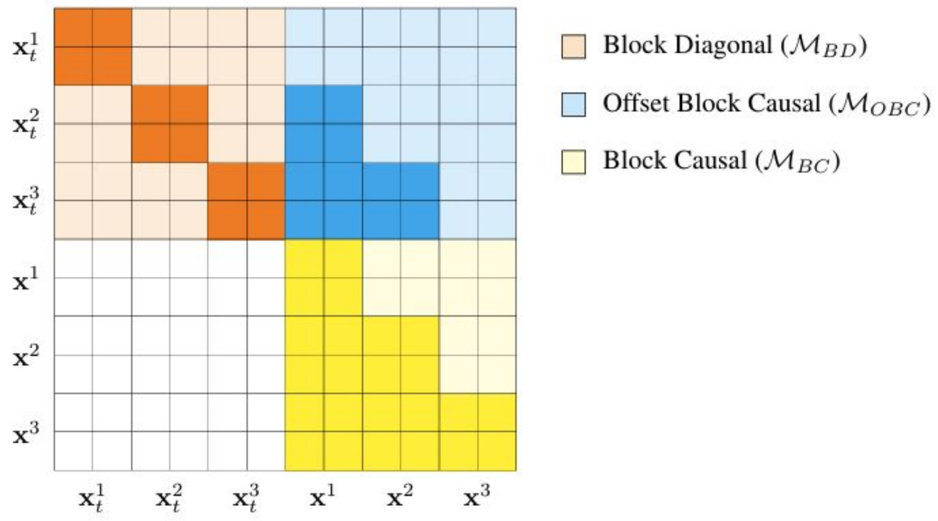
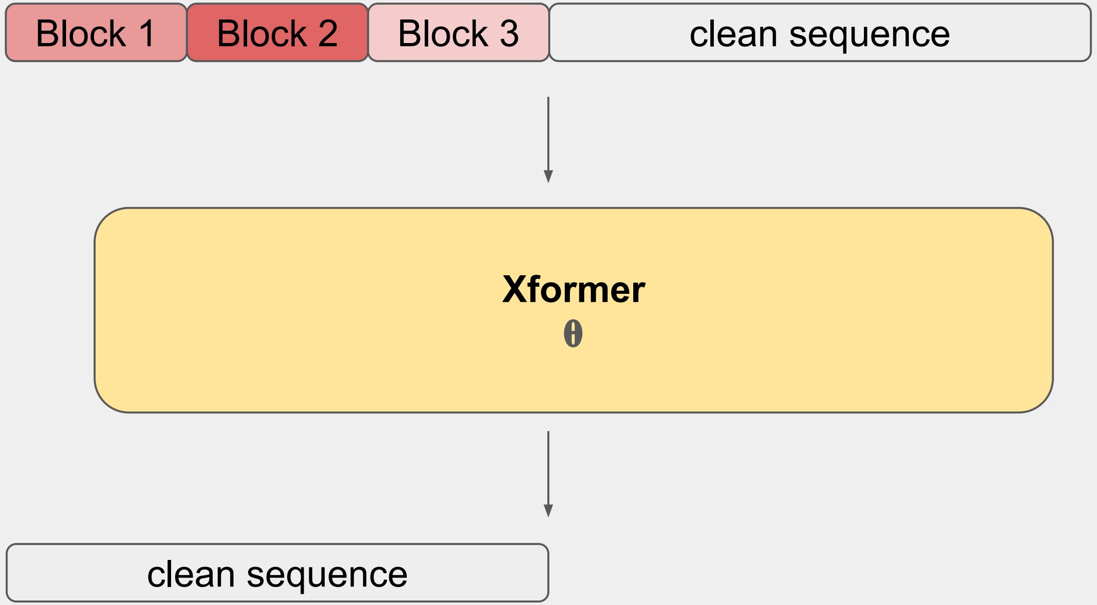

$\mathbf{x}_t^1$
$\mathbf{x}_t^2$
$\mathbf{x}_t^3$
$\mathbf{x}^1$
$\mathbf{x}^2$
$\mathbf{x}^3$
Block Diagonal ($\mathcal{M}_{BD}$)
Offset Block Causal ($\mathcal{M}_{OBC}$)
Block Causal ($\mathcal{M}_{BC}$)

Figure 3: Example of Specialized Attention Mask

Arriola et al., "Block Diffusion," arXiv:2503.09573, 2025.

Discrete diffusion for text - Efficient Training



Arriola et al., "Block Diffusion," arXiv:2503.09573, 2025.

Discrete diffusion for text - Inference

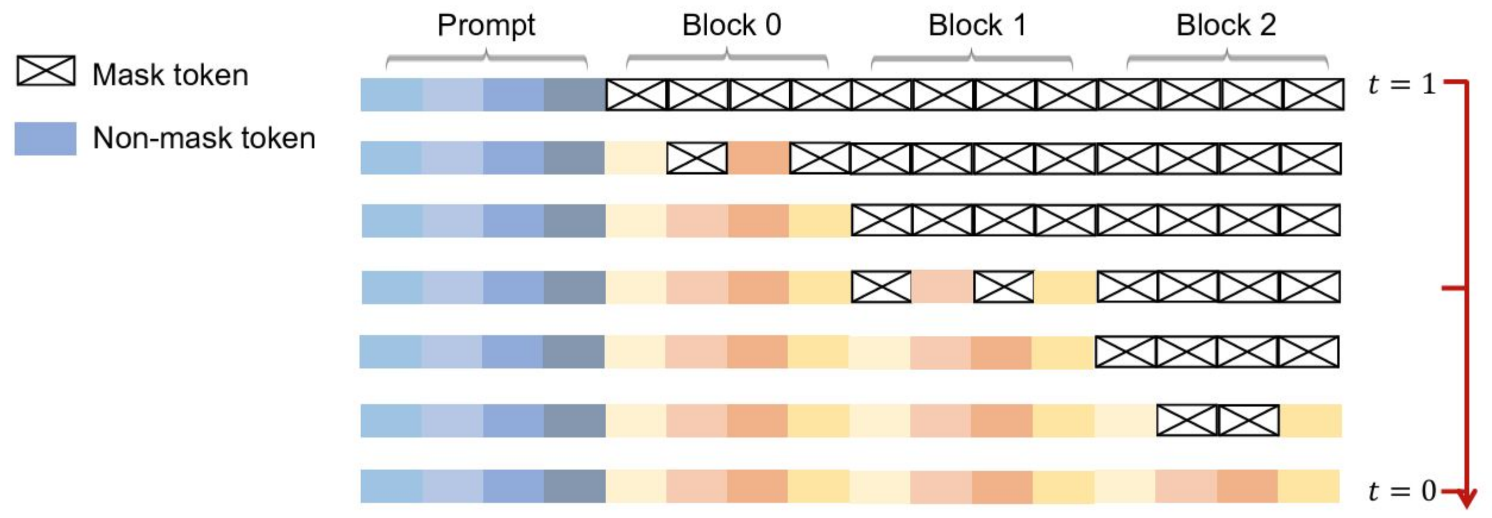


Figure 4. **A Conceptual Overview of the Semi-autoregressive Sampling.**

Nie et al., "Large Language Diffusion Models," arXiv:2502.09992, 2025.

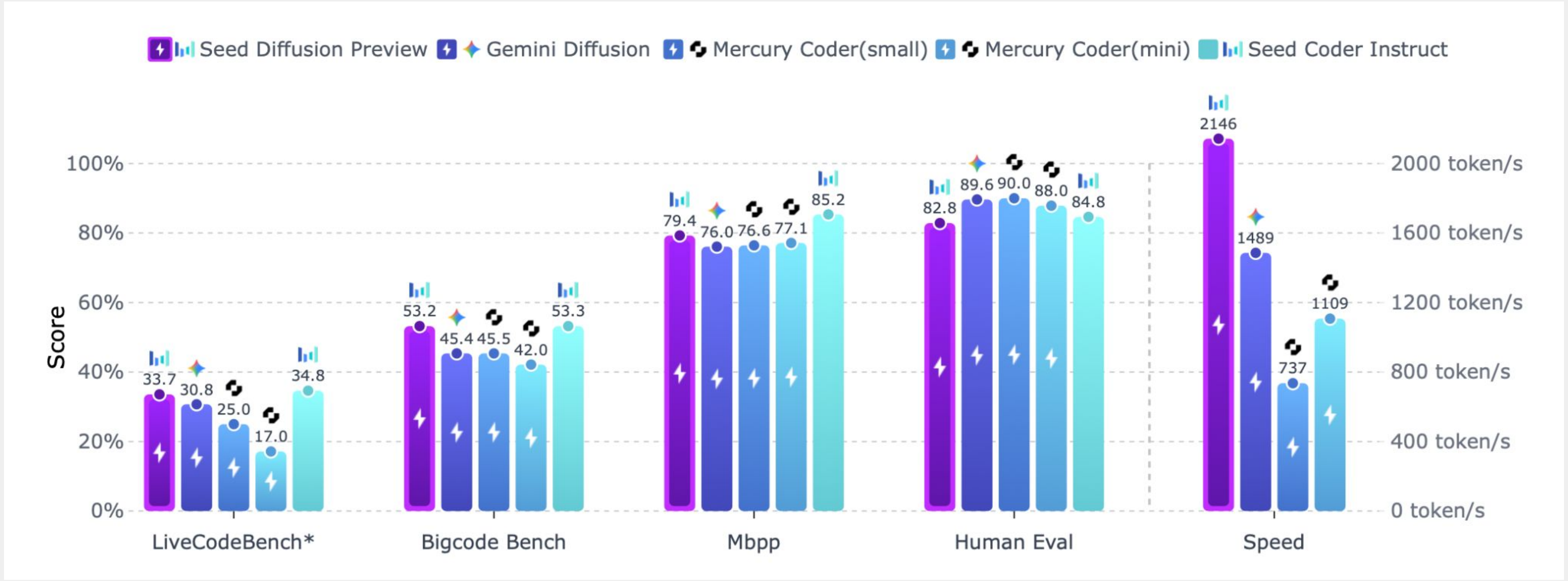
4. What’s Next?

Diffusion Language Models

Recent, fast Moving Field

- Dieleman et al., “Continuous Diffusion for Categorical Data - Dec. 2022
- Nie et al., “Large Language Diffusion Models - Feb 2025
- Gong et al., “Scaling Diffusion Language Models” - April 2025
- Arriola et al., “Block Diffusion” - May 2025
- Google Deepmind, “Gemini Diffusion” - May 2025
- ByteDance, “Seed Diffusion” - August 2025

Diffusion Language Models



ByteDance, “Seed Diffusion”, arXiv: 2508.02193v1

Diffusion Language Models

Lots of ideas in the air:

- Edit-based Noise Process
- On-policy Diffusion Learning
- Distillation
- Adaptive Computation
- Combining Continuous and Discrete
- Post-training and RL

…

Diffusion Language Models

Twitter: @sahnihim

Gmail: himsahni@google.com