# IDLM: Inverse-distilled Diffusion Language Models

Accepted to ICML 2026

David Li[*; 1], Nikita Gushchin[*; 2,3], Dmitry Abulkhanov, Eric Moulines[1,4], Ivan Oseledets[2,3], Maxim Panov[1], Alexander Korotin[2,3]

1 MBZUAI, Abu Dhabi, UAE

2 Applied AI Institute, Moscow, Russia

3 AXXX, Moscow, Russia

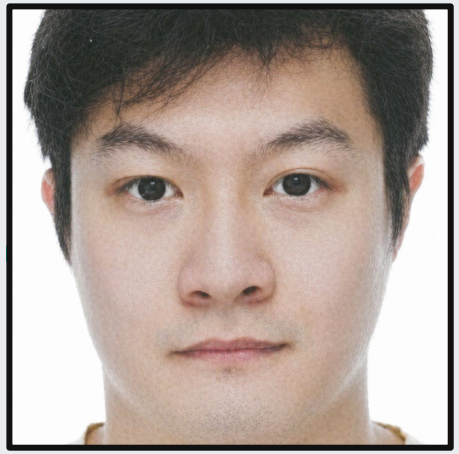
4 EPITA, Paris, France

6 BRAIn Lab, Moscow, Russia

* Equal contribution


Bio

# Team



David Li*

Nikita Gushchin*

Dmitry Abulkhanov

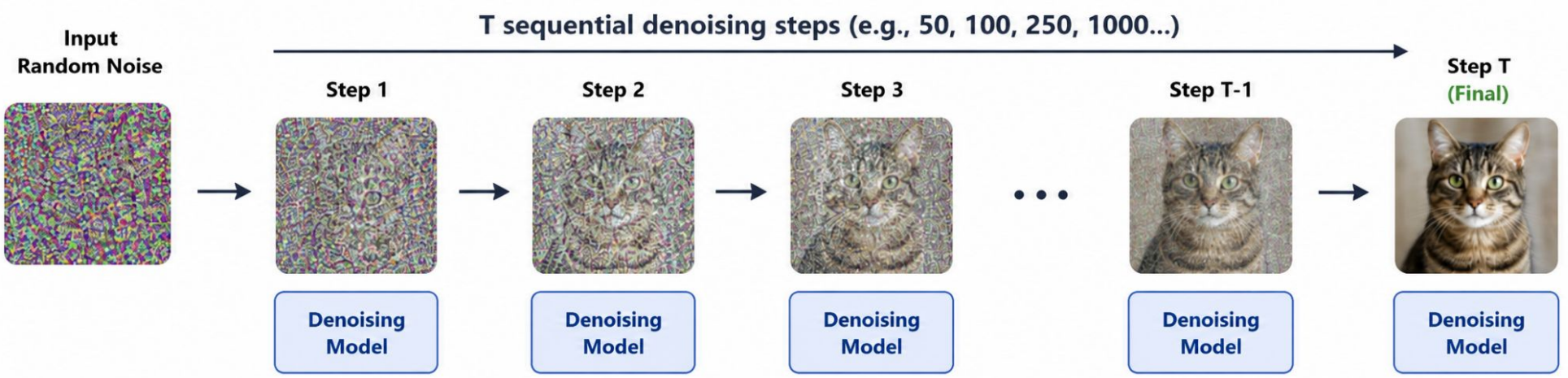
Eric Moulines

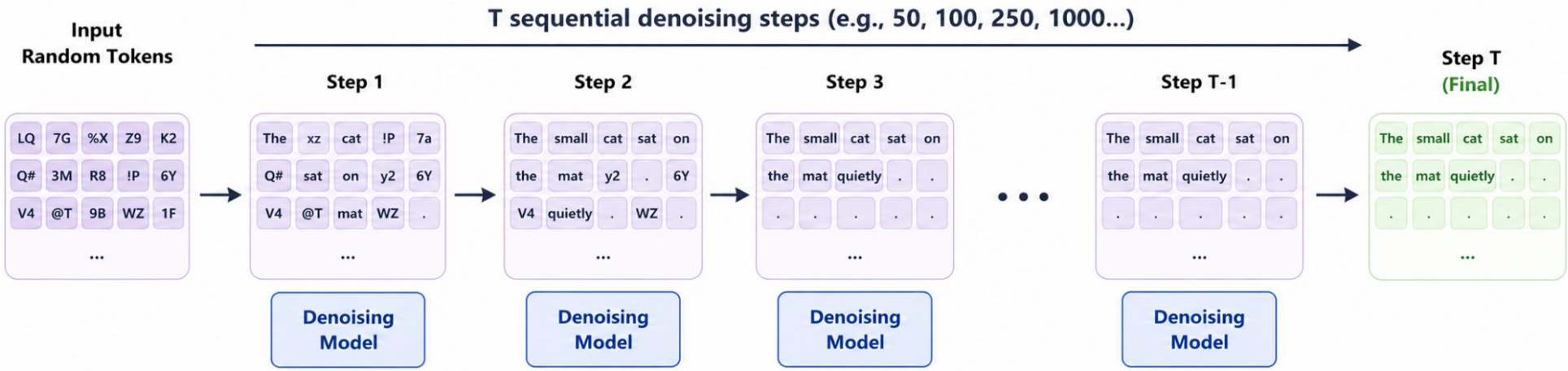
Ivan Oseledets

Maxim Panov

Alexander Korotin

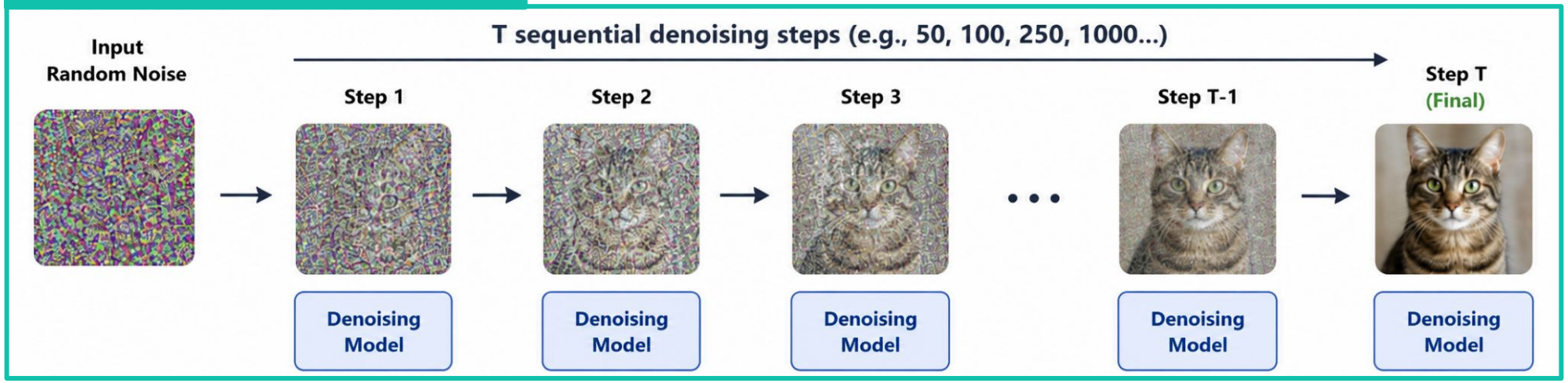
# Challenge: Slow Inference



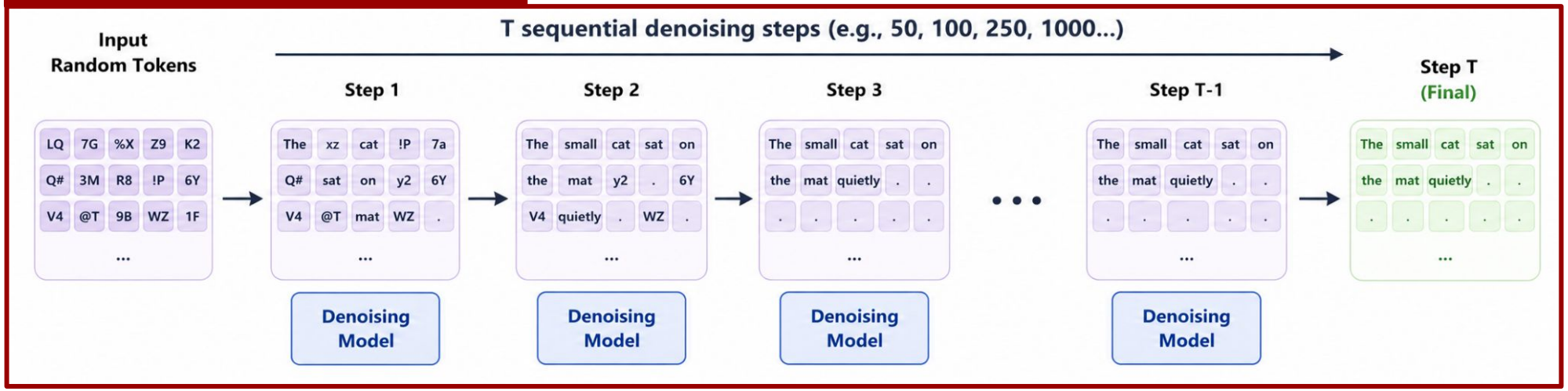



# Challenge: Slow Inference

## Solved via distillation

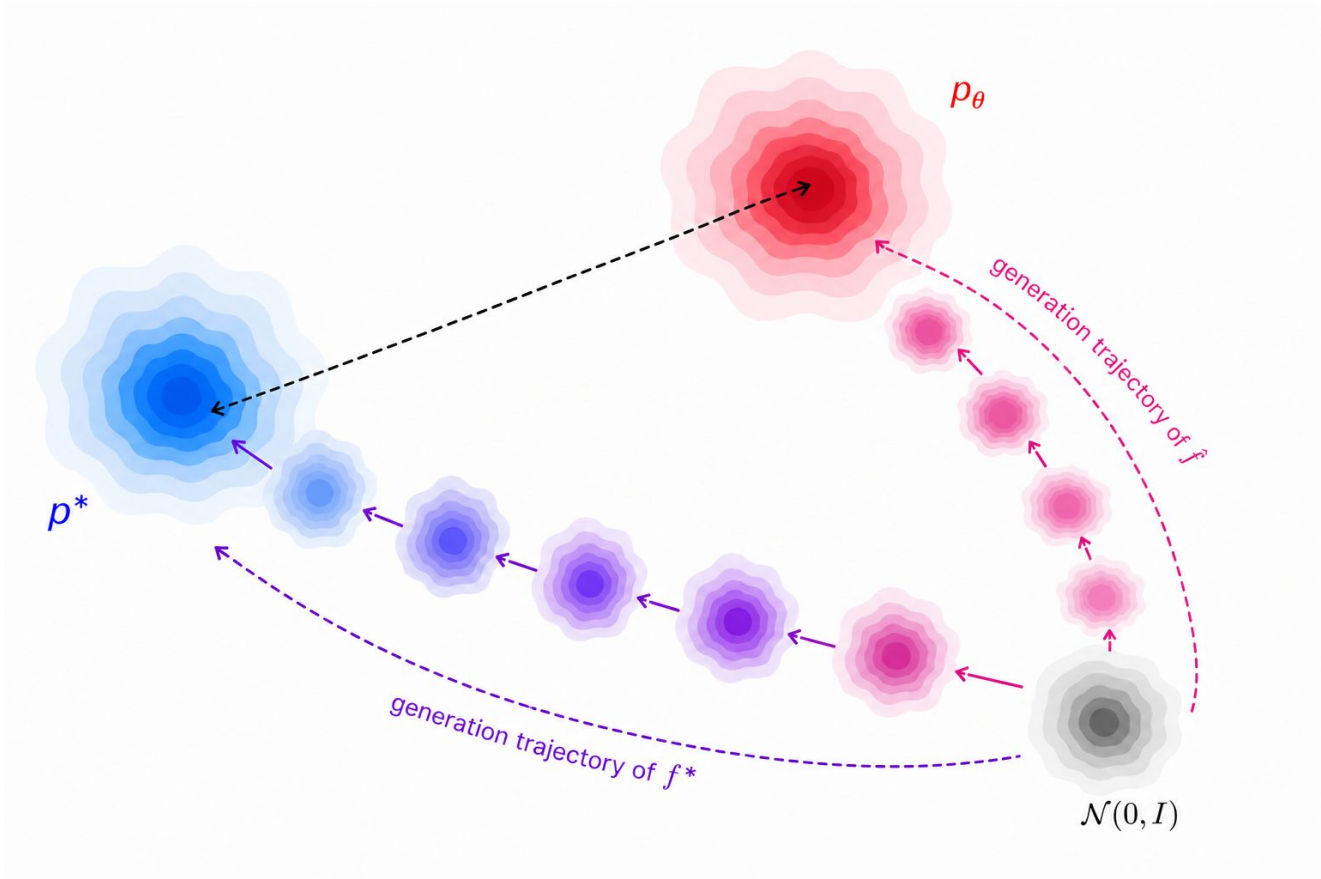



## Not exhaustively explored



# Inverse Distillation loss function

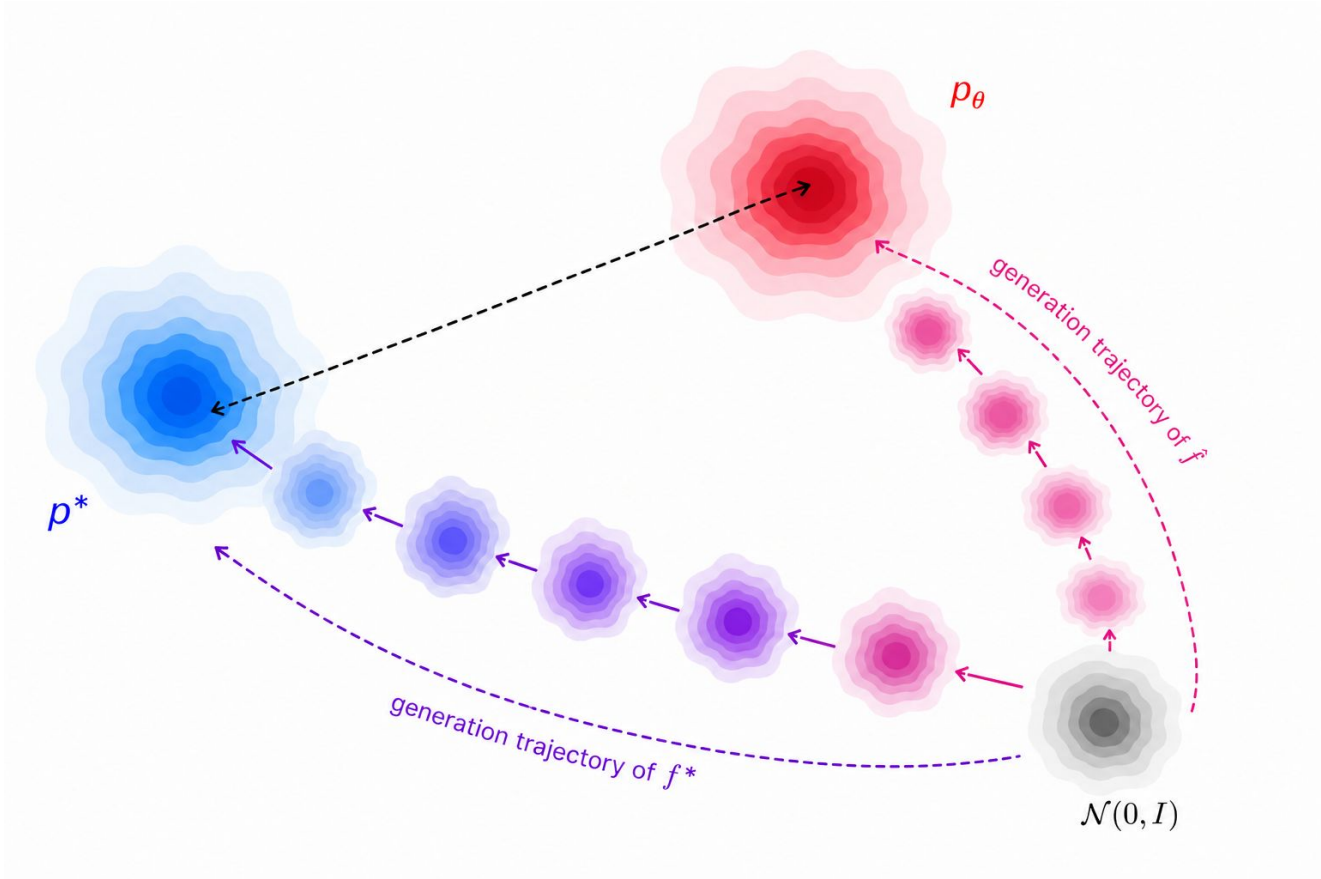
$$\mathcal{L}_{\text{Inverse}}(\theta) = \mathcal{L}(f^*, p_\theta) - \min_{\hat{f}} \mathcal{L}(\hat{f}, p_\theta)$$



# Inverse Distillation loss function

$$\mathcal{L}_{\text{Inverse}}(\theta) = \mathcal{L}(f^*, p_\theta) - \min_{\hat{f}} \mathcal{L}(\hat{f}, p_\theta)$$

**Train "fake" model**



# Contributions of IDLM

**We want to extend Inverse Distillation for Discrete Diffusion Models**

$$\mathcal{L}_{\text{IDLM}}(\theta) = \mathcal{L}_{\text{discr.}}(f^*, p_\theta) - \min_{\widehat{f}} \mathcal{L}_{\text{discr.}}(\widehat{f}, p_\theta)$$

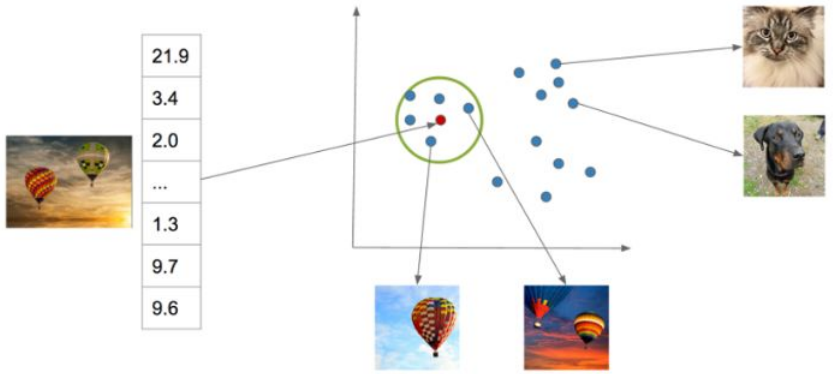
**Possible problems:**

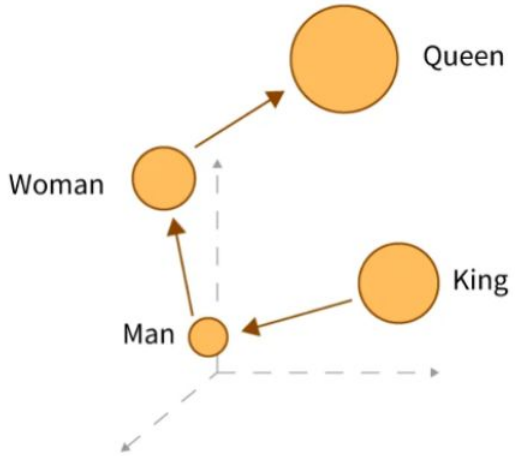
- Theoretical groundness:

???

$$p_\theta = p^* \iff \mathcal{L}_{\text{IDLM}}(\theta) = 0$$

- Non-smooth discrete domain (will clarify later):



# Theoretical Contribution

**Theorem**

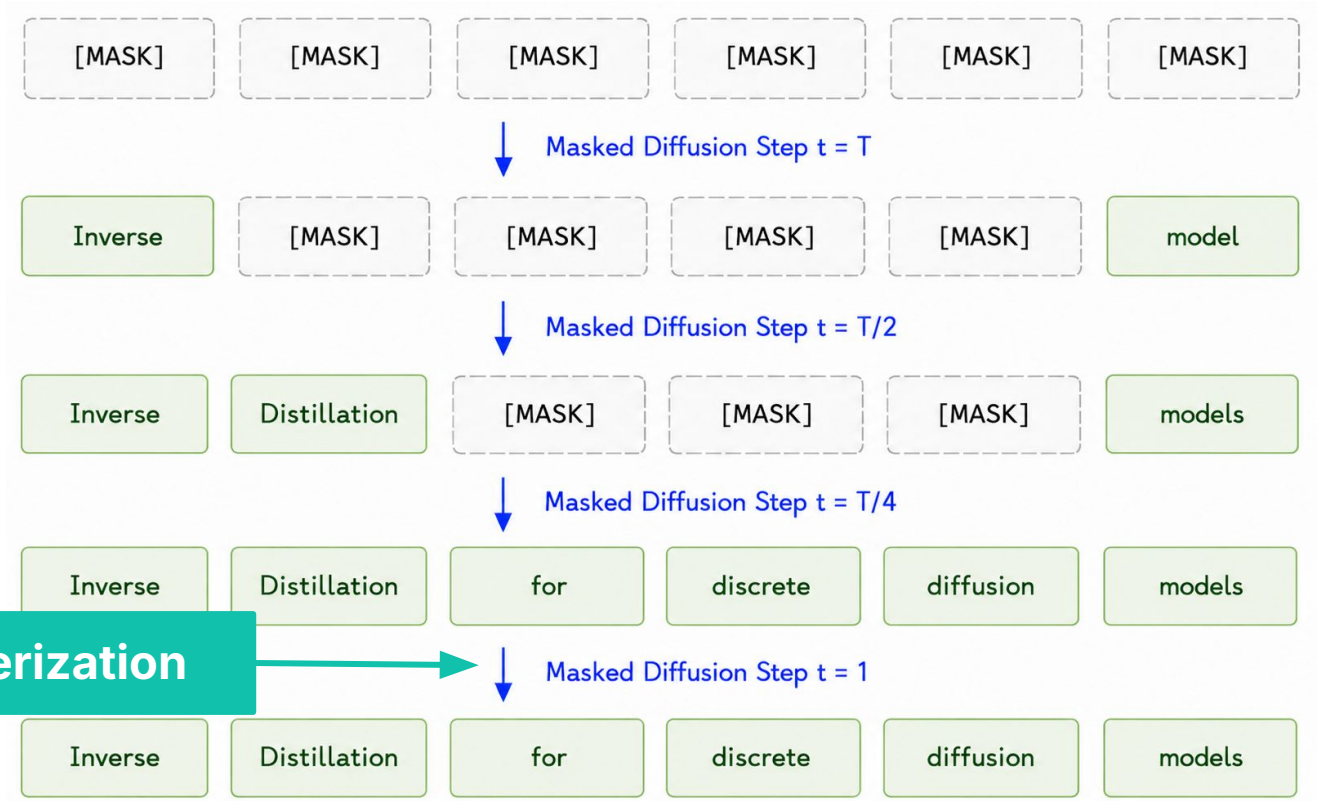
For SEDD, MDLM, and Duo (in the $\tau \to 0^+$ limit), the IDLM objective satisfies

$$\mathcal{L}_{\mathrm{IDLM}}(\theta) \;\geq\; \mathcal{D}_{\mathrm{KL}}(p^{\theta} \,\|\, p^{*}) \;\geq\; 0,$$

with equality iff $p^{\theta} = p^{*}$.

# Intuition of Masked Diffusion (Teacher model)

$$\mathcal{L}_{\text{discr.}}(f, p) = -\mathbb{E}_{t, x_0 \sim p(x_0), x_t \sim q(x_t|x_0)} \langle x_0, \log f(x_t, t) \rangle$$




## Intuition

Unmasking occurs only **once** during generation, so the transition timestep does not matter. We only need to **move** to the correct token.

# Practical Contribution: IDLM-MDLM

**Masked Diffusion**

$$\mathcal{L}_{\text{IDLM}}(\theta) = -\mathbb{E}_{t, x_0 \sim p_\theta(x_0), x_t \sim q(x_t|x_0)} \langle x_0, \log f^*(x_t, t) - \log \widehat{f}(x_t, t) \rangle$$

# Practical Contribution: IDLM-MDLM

## Masked Diffusion

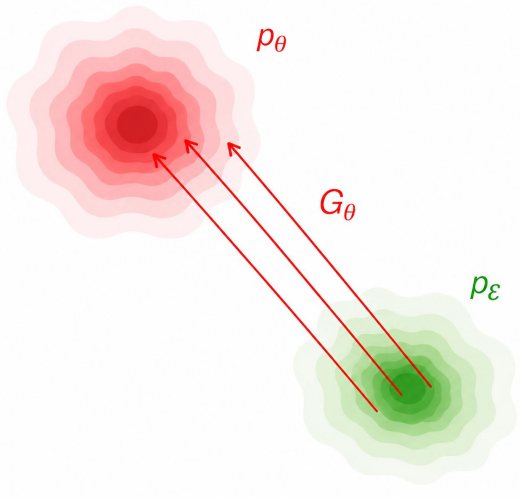
$$\mathcal{L}_{\text{IDLM}}(\theta) = -\mathbb{E}_{t, x_0 \sim p_\theta(x_0), x_t \sim q(x_t|x_0)} \langle x_0, \log f^*(x_t, t) - \log \hat{f}(x_t, t) \rangle$$

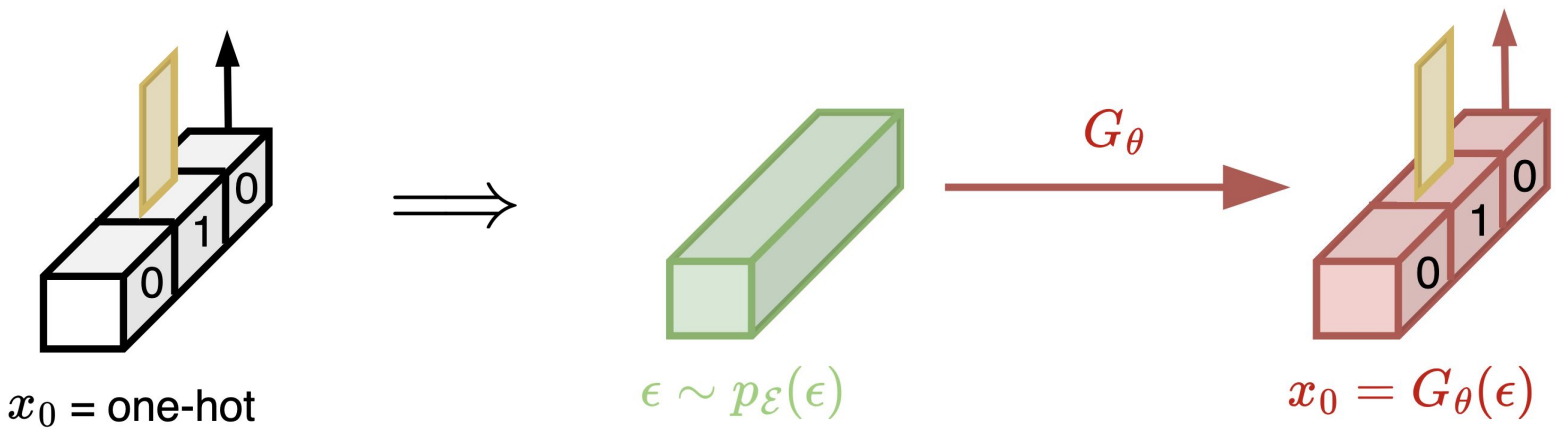
1 2

**How to take gradient through it?**

## 1 Reparametrization

$$\mathcal{L}_{\text{IDLM}}(\theta) = -\mathbb{E}_{t, \epsilon \sim p_{\mathcal{E}}(\epsilon), x_t \sim q(x_t|G_\theta(\epsilon))} \langle G_\theta(\epsilon), \log f^*(x_t, t) - \log \hat{f}(x_t, t) \rangle$$



# Practical Contribution: IDLM-MDLM

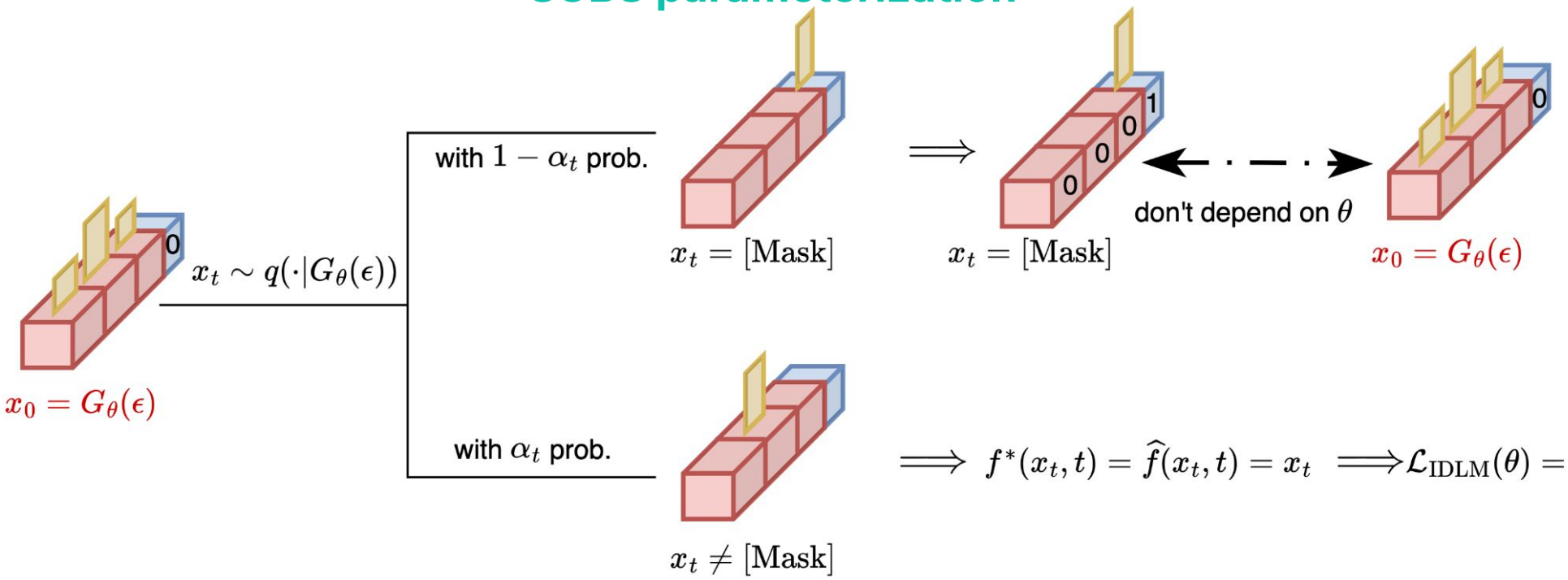
## Masked Diffusion

$$\mathcal{L}_{\text{IDLM}}(\theta) = -\mathbb{E}_{t,\epsilon\sim p_{\mathcal{E}}(\epsilon),x_t\sim q(x_t|G_\theta(\epsilon))}\langle G_\theta(\epsilon), \log f^*(x_t,t) - \log \widehat{f}(x_t,t)\rangle$$

2

**How to take gradient through it?**

## SUBS parameterization



# Practical Contribution: IDLM-MDLM

**Masked Diffusion**

$$\mathcal{L}_{\text{IDLM}}(\theta) = -\mathbb{E}_{t,\epsilon\sim p_{\mathcal{E}}(\epsilon),x_t\sim q(x_t|G_\theta(\epsilon))}\langle G_\theta(\epsilon), \log f^*(x_t,t) - \log \widehat{f}(x_t,t)\rangle$$

**SUBS parameterization**

$$\mathcal{L}_{\text{IDLM}}(\theta) = -\mathbb{E}_{t,\epsilon\sim p_{\mathcal{E}}(\epsilon),x_t\sim q(x_t|G_\theta(\epsilon))}\left\langle G_\theta(\epsilon), \text{sg}\left(\log f^*(x_t,t) - \log \widehat{f}(x_t,t)\right)\right\rangle$$

# Multistep Distillation

## Challenges in one-step generation

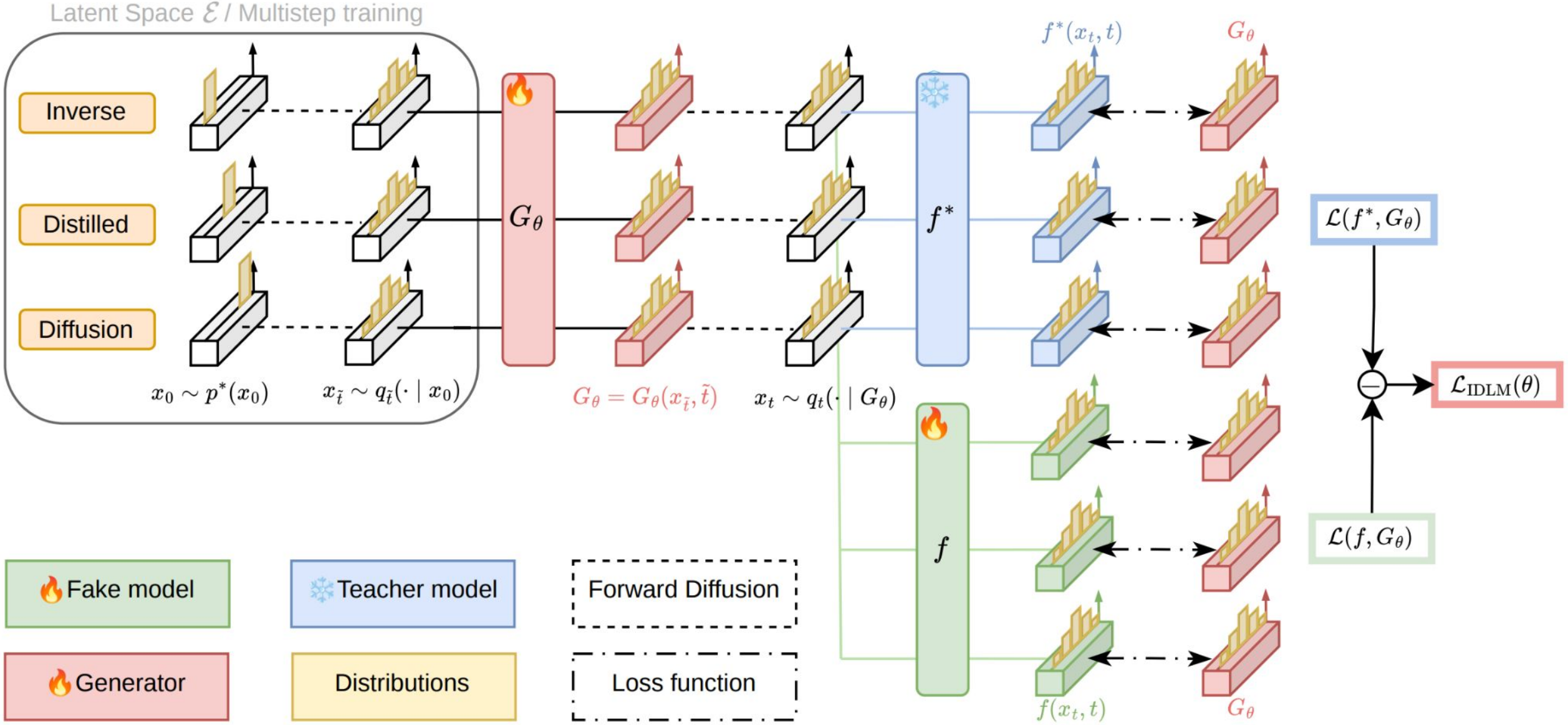
Single-step text generator distillation remains **challenging**, likely because each latent must collapse to a delta distribution, causing instability.

**Sampling**

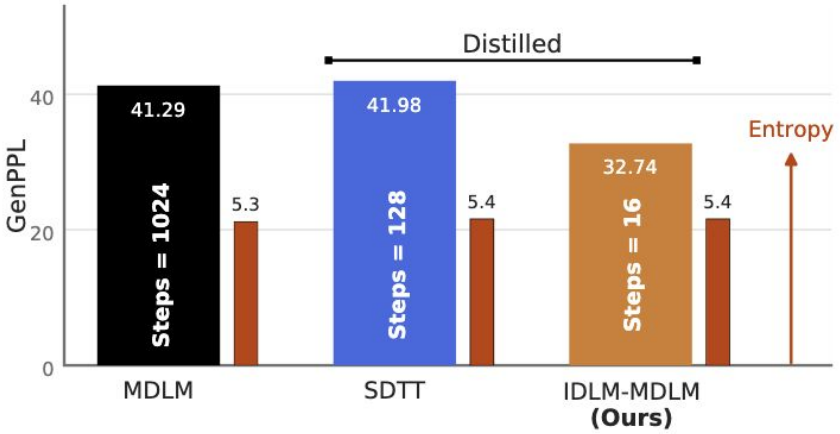
Training the model for few-step generation allows us to use the same **sampling procedure as the teacher model**

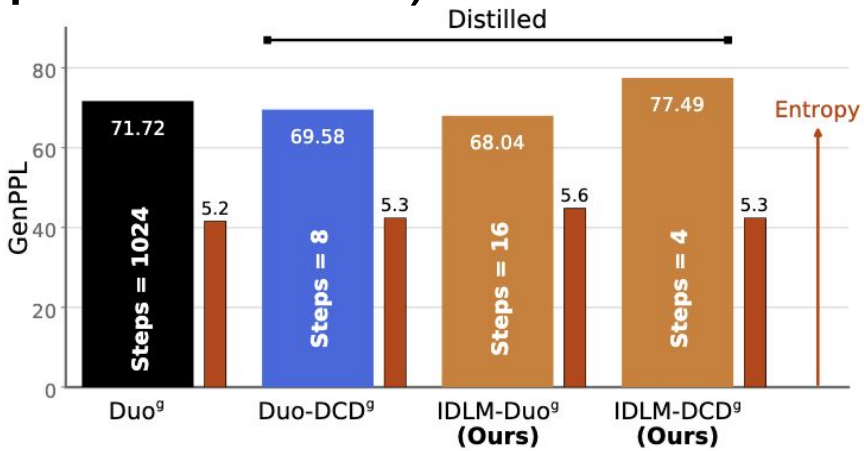
$$q(\mathbf{z}_s|\mathbf{z}_t,\mathbf{x}) = \mathrm{Cat}\left(\mathbf{z}_s; \frac{[\alpha_{t|s}\mathbf{z}_t + (1-\alpha_{t|s})\mathbf{1}\boldsymbol{\pi}^\top \mathbf{z}_t] \odot [\alpha_s \mathbf{x} + (1-\alpha_s)\boldsymbol{\pi}]}{\alpha_t \mathbf{z}_t^\top \mathbf{x} + (1-\alpha_t)\mathbf{z}_t^\top \boldsymbol{\pi}}\right)$$

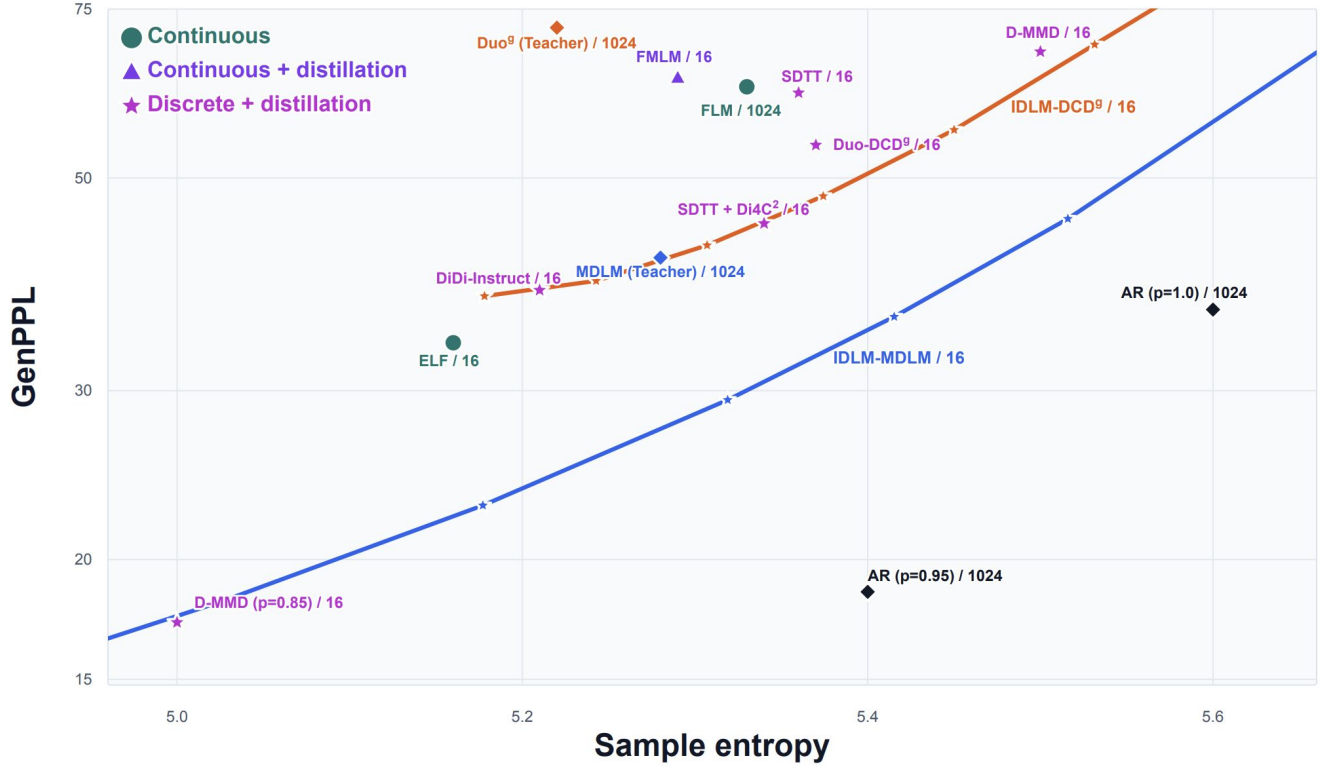
# Putting everything together



# Experiments: Unconditional generation (OpenWebText)



# Experiments: Conditional generation (TyniGSM)

| Model | Steps | Accuracy (%) |
|---|---|---|
| *Autoregressive* | | |
| Sample | 512 | 53.9 |
| Greedy | | **63.3** |
| *Diffusion* | | |
| MDLM (Teacher) (Sahoo et al., 2024) | 1024 | **18.0** |
| Duo (Teacher) (Sahoo et al., 2025) | | 17.2 |
| FLM (Lee et al., 2026) | | 0.3 |
| S-FLM (Deschenaux & Gulcehre, 2026) | | **18.0** |
| *Distillation* | | |
| IDLM-MDLM (**Ours**) | 128 | 19.86 |
| IDLM-Duo (**Ours**) | | **21.38** |
| IDLM-MDLM (**Ours**) | 64 | 14.94 |
| IDLM-Duo (**Ours**) | | **19.03** |
| IDLM-MDLM (**Ours**) | 32 | 12.81 |
| IDLM-Duo (**Ours**) | | **15.39** |

# Links


Bio