Simple Guidance Mechanisms for Discrete Diffusion Models


Yair Schiff*


Subham Sahoo*


Hao Phung*


Guanghan Wang*


Sam Boshar


Hugo Dalla-torre

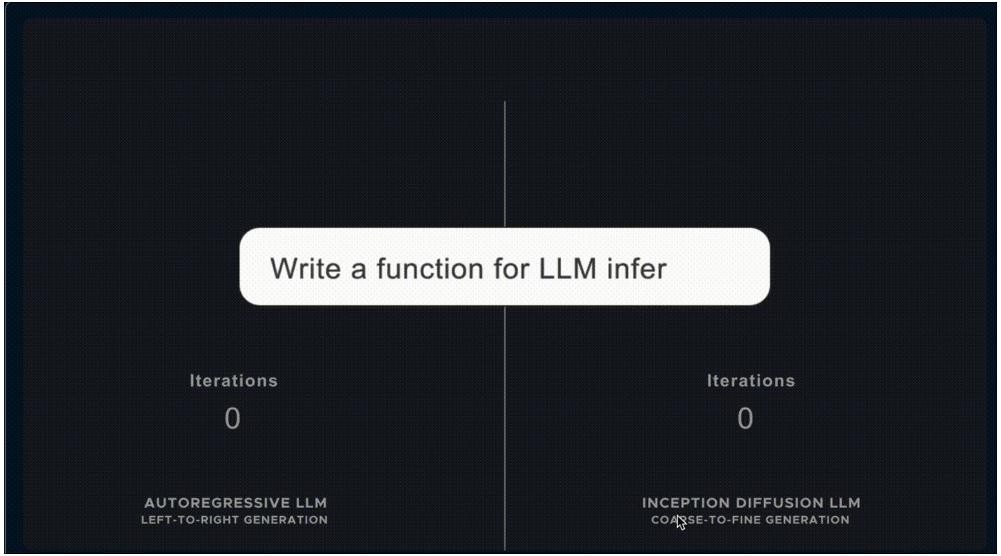

Bernardo P de Almeida

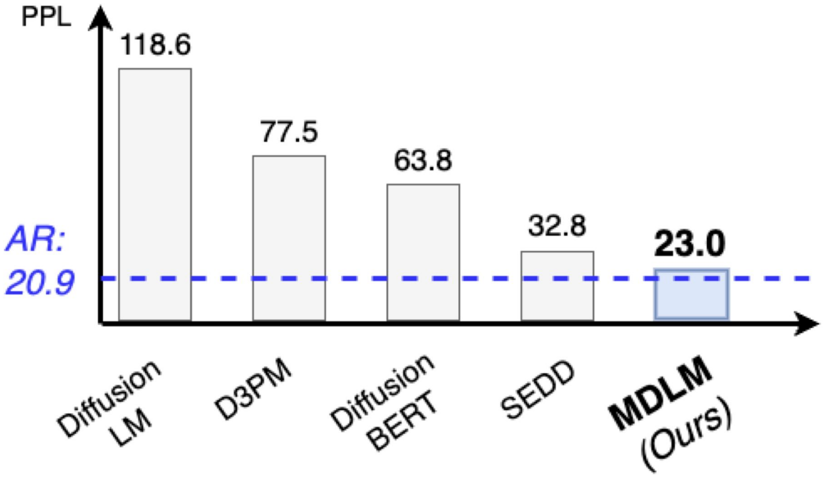

Alexander Rush


Thomas Pierrot

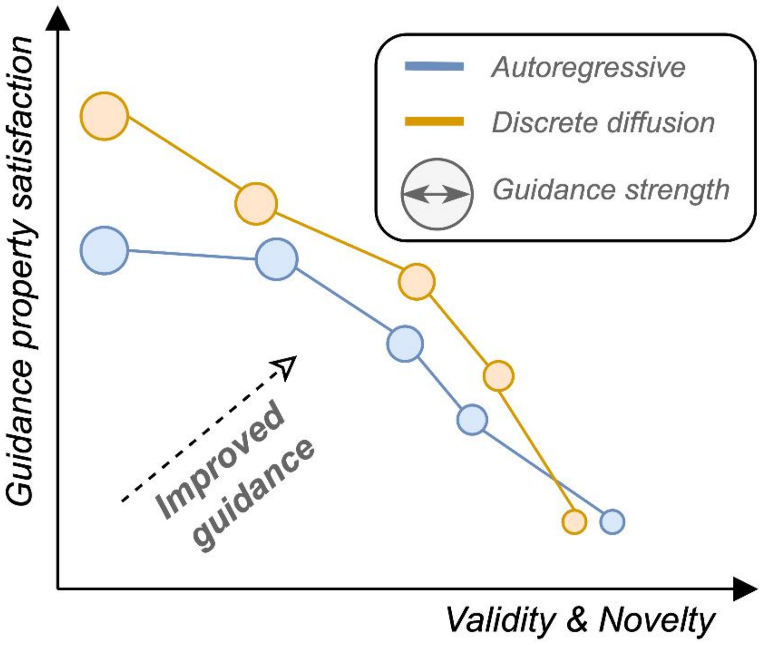

Volodymyr Kuleshov

*Equal contribution





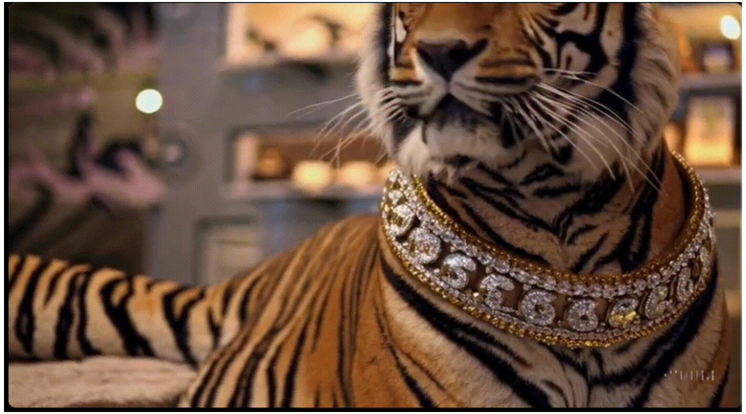
Sahoo, Subham, et al. "Simple and effective masked diffusion language models." Advances in Neural Information Processing Systems 37 (2024): 130136-130184.



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

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

An ink sketch style illustration of a small hedgehog holding a piece of watermelon with its tiny paws, taking little bites with its eyes closed in delight.



The 'bling zoo' shop in new york city is both a jewelry store and zoo. sabertooth tigers with diamond and gold adornments…

Image credit: (Left) https://openai.com/index/dall-e-3/; (Right) https://openai.com/sora/

Notation

$$s, t \text{ timesteps}, s < t \in [0, 1]$$

$$\mathbf{z}_t \text{ latent vector}$$

$$\text{Continuous: } \mathbf{z}_t \in \mathbb{R}^d$$

$$\text{Discrete: } \mathbf{z}_t^{(1:L)} \in |\mathcal{V}|^L \text{ (vocab } \mathcal{V}, \text{seq len } L)$$

Guidance Background

Sample from **conditional**, **tempered** distribution

$$\mathbf{z}_s \sim p^{\boxed{\gamma}}(\mathbf{z}_s \mid \mathbf{z}_\mathbf{t}, \boxed{y})$$

Sample from **conditional**, tempered distribution

$$\mathbf{z}_s \sim p\left(\mathbf{z}_s \mid \mathbf{z}_t, y\right)$$

$$\downarrow$$

$$p\left(\mathbf{z}_s \mid \mathbf{z}_t, y\right) \propto p\left(y \mid \mathbf{z}_s, \mathbf{z}_t\right) \cdot p\left(\mathbf{z}_s \mid \mathbf{z}_t\right)$$

Sample from conditional, **tempered** distribution

$$\mathbf{z}_s \sim p^{\gamma}(\mathbf{z}_s \mid \mathbf{z_t}, y)$$

$$\downarrow$$

$$p^{\gamma}(\mathbf{z}_s \mid \mathbf{z}_t, y) \propto p(y \mid \mathbf{z}_s, \mathbf{z}_t)^{\gamma} \cdot p(\mathbf{z}_s \mid \mathbf{z}_t)$$

Continuous: Follow log-prob gradient with respect to input

$$p^\gamma(\mathbf{z}_s \mid \mathbf{z}_t, y) \propto p(y \mid \mathbf{z}_s, \mathbf{z}_t)^\gamma \cdot p(\mathbf{z}_s \mid \mathbf{z}_t)$$

$$\downarrow$$

$$\nabla_{\mathbf{z}_s} \log p^\gamma(\mathbf{z}_s \mid \mathbf{z}_t, y) = \gamma \nabla_{\mathbf{z}_s} \log p(y \mid \mathbf{z}_s, \mathbf{z}_t) + \nabla_{\mathbf{z}_s} \log p(\mathbf{z}_s \mid \mathbf{z}_t)$$

Continuous: Follow log-prob gradient with respect to input

$$p^{\gamma}(\mathbf{z}_s \mid \mathbf{z}_t, y) \propto p(y \mid \mathbf{z}_s, \mathbf{z}_t)^{\gamma} \cdot p(\mathbf{z}_s \mid \mathbf{z}_t)$$

$$\downarrow$$

$$\nabla_{\mathbf{z}_s} \log p^{\gamma}(\mathbf{z}_s \mid \mathbf{z}_t, y) = \gamma \nabla_{\mathbf{z}_s} \log \boxed{p(y \mid \mathbf{z}_s, \mathbf{z}_t)} + \nabla_{\mathbf{z}_s} \log \boxed{p(\mathbf{z}_s \mid \mathbf{z}_t)}$$

External classifier

Diffusion model

Classifier-based guidance

Continuous: Follow log-prob gradient with respect to input

$$\nabla_{\mathbf{z}_s} \log p^{\gamma}(\mathbf{z}_s \mid \mathbf{z}_t, y) = \gamma \nabla_{\mathbf{z}_s} \log p(y \mid \mathbf{z}_s, \mathbf{z}_t) + \nabla_{\mathbf{z}_s} \log p(\mathbf{z}_s \mid \mathbf{z}_t)$$

$$\nabla_{\mathbf{z}_s} \log p(\mathbf{z}_s \mid \mathbf{z}_t, y) - \nabla_{\mathbf{z}_s} \log p(\mathbf{z}_s \mid \mathbf{z}_t)$$

Continuous: Follow log-prob gradient with respect to input

$$\nabla_{\mathbf{z}_s} \log p^{\gamma}(\mathbf{z}_s \mid \mathbf{z}_t, y) = \gamma \nabla_{\mathbf{z}_s} \log p(y \mid \mathbf{z}_s, \mathbf{z}_t) + \nabla_{\mathbf{z}_s} \log p(\mathbf{z}_s \mid \mathbf{z}_t)$$

$$\downarrow$$

$$\nabla_{\mathbf{z}_s} \log p^{\gamma}(\mathbf{z}_s \mid \mathbf{z}_t, y) = \gamma \nabla_{\mathbf{z}_s} \log p(\mathbf{z}_s \mid \mathbf{z}_t, y) + (1-\gamma) \nabla_{\mathbf{z}_s} \log p(\mathbf{z}_s \mid \mathbf{z}_t)$$

Continuous: Follow log-prob gradient with respect to input

$$\nabla_{\mathbf{z}_s} \log p^{\gamma}(\mathbf{z}_s \mid \mathbf{z}_t, y) = \gamma \nabla_{\mathbf{z}_s} \log p(y \mid \mathbf{z}_s, \mathbf{z}_t) + \nabla_{\mathbf{z}_s} \log p(\mathbf{z}_s \mid \mathbf{z}_t)$$

$$\downarrow$$

$$\nabla_{\mathbf{z}_s} \log p^{\gamma}(\mathbf{z}_s \mid \mathbf{z}_t, y) = \gamma \nabla_{\mathbf{z}_s} \log \boxed{p(\mathbf{z}_s \mid \mathbf{z}_t, y)} + (1-\gamma) \nabla_{\mathbf{z}_s} \log \boxed{p(\mathbf{z}_s \mid \mathbf{z}_t)}$$

Classifier-<u>free</u> guidance

Conditional diffusion model

Unconditional diffusion model

Discrete Guidance

Discrete:

$$\nabla_{\mathbf{z}_S}$$

Discrete: Back to the drawing board

$$\log p^{\gamma}(\mathbf{z}_s \mid \mathbf{z}_t, y) = \gamma \log p(y \mid \mathbf{z}_s, \mathbf{z}_t) + \log p(\mathbf{z}_s \mid \mathbf{z}_t) + const.$$

Discrete: Preview of where we'll end up

$$p^{\gamma}(\mathbf{z}_s \mid \mathbf{z}_t, y) \propto$$
$$(p_\theta) \cdot (\text{some other scaling distribution})$$

Discrete Classifier-Free Guidance

Discrete: Classifier-free guidance

$$\log p^{\gamma}(\mathbf{z}_s \mid \mathbf{z}_t, y) = \boxed{\gamma \log p(y \mid \mathbf{z}_s, \mathbf{z}_t)} + \log p(\mathbf{z}_s \mid \mathbf{z}_t) + const.$$

$$\downarrow$$

$$\gamma \log p(\mathbf{z}_s \mid \mathbf{z}_t, y) - \gamma \log p(\mathbf{z}_s \mid \mathbf{z}_t) + \underbrace{\gamma \log p(\mathbf{y} \mid \mathbf{z}_t)}_{\text{constant}}$$

Discrete: Classifier-free guidance

$$\log p^{\gamma}(\mathbf{z}_s \mid \mathbf{z}_t, y) = \gamma \log p(y \mid \mathbf{z}_s, \mathbf{z}_t) + \log p(\mathbf{z}_s \mid \mathbf{z}_t) + const.$$

$$\downarrow$$

$$\log p^{\gamma}(\mathbf{z}_s \mid \mathbf{z}_t, y) = \gamma \log \boxed{p(\mathbf{z}_s \mid \mathbf{z}_t, y)} + (1-\gamma) \log \boxed{p(\mathbf{z}_s \mid \mathbf{z}_t)} + const.$$

Conditional diffusion model

Unconditional diffusion model

Discrete Classifier-Based Guidance

Discrete: Classifier-based guidance

$$p^{\gamma}(\mathbf{z}_s \mid \mathbf{z}_t, y) \propto p(y \mid \mathbf{z}_s, \mathbf{z}_t)^{\gamma} \cdot p(\mathbf{z}_s \mid \mathbf{z}_t)$$

Discrete: Classifier-based guidance

$$p^{\gamma}(\mathbf{z}_s \mid \mathbf{z}_t, y) \propto p(y \mid \mathbf{z}_s, \mathbf{z}_t)^{\gamma} \cdot p(\mathbf{z}_s \mid \mathbf{z}_t)$$

In practice, we model **sequences of text**, not just individual tokens

$$\mathbf{z} \to \mathbf{z}^{(1:L)}$$

$$p(y \mid \mathbf{z}_s, \mathbf{z}_t) \rightarrow p(y \mid \mathbf{z}_s^{(1:L)}, \mathbf{z}_t^{(1:L)})$$

$$\mathcal{O}(|V|^L)$$ terms

Assume $p^\gamma(\mathbf{z}_s^{(1:L)} \mid \mathbf{z}_t^{(1:L)}, y) = \prod_{\ell=1}^{L} p^\gamma(\mathbf{z}_s^{(\ell)} \mid \mathbf{z}_t^{(1:L)}, y)$

We will model this per token distribution instead

$$p\left(\mathbf{z}_s^{(\ell)} \mid \mathbf{z}_t^{(1:L)}, y\right) \propto p\left(y \mid \mathbf{z}_s^{(\ell)}, \mathbf{z}_t^{(1:L)}\right) \cdot p\left(\mathbf{z}_s^{(\ell)} \mid \mathbf{z}_t^{(1:L)}\right)$$

$$p(\mathbf{z}_s^{(\ell)} \mid \mathbf{z}_t^{(1:L)}, y) \propto p(y \mid \mathbf{z}_s^{(\ell)}, \mathbf{z}_t^{(1:L)}) \cdot \boxed{p(\mathbf{z}_s^{(\ell)} \mid \mathbf{z}_t^{(1:L)})}$$

Model with
discrete diffusion

$$p_\theta$$

$$p(\mathbf{z}_s^{(\ell)} \mid \mathbf{z}_t^{(1:L)}, y) \propto \boxed{p(y \mid \mathbf{z}_s^{(\ell)}, \mathbf{z}_t^{(1:L)})} \cdot p(\mathbf{z}_s^{(\ell)} \mid \mathbf{z}_t^{(1:L)})$$

Train classifier on noisy sequences:

$$p_\phi(\mathbf{z}_t^{(1:L)})$$

Evaluate it on sequences of the form:

$$p_\phi(y \mid [\mathbf{z}_t^{(1:\ell-1)}, \mathbf{z}_s^{(\ell)}, \mathbf{z}_t^{(\ell+1:L)}])$$

Putting it all together

Discrete Classifier-free guidance

$$p^{\gamma}(\mathbf{z}_s^{(\ell)} \mid \mathbf{z}_t^{(1:L)}, y) \propto p(\mathbf{z}_s^{(\ell)} \mid \mathbf{z}_t^{(1:L)}, y)^{\gamma} \cdot p(\mathbf{z}_s^{(\ell)} \mid \mathbf{z}_t^{(1:L)})^{(1-\gamma)}$$

Discrete Classifier-based guidance

$$p^{\gamma}(\mathbf{z}_s^{(\ell)} \mid \mathbf{z}_t^{(1:L)}, y) \propto p_{\phi}(y \mid [\mathbf{z}_t^{(1:\ell-1)}, \mathbf{z}_s^{(\ell)}, \mathbf{z}_t^{(\ell+1:L)}])^{\gamma} \cdot p(\mathbf{z}_s^{(\ell)} \mid \mathbf{z}_t^{(1:L)})$$

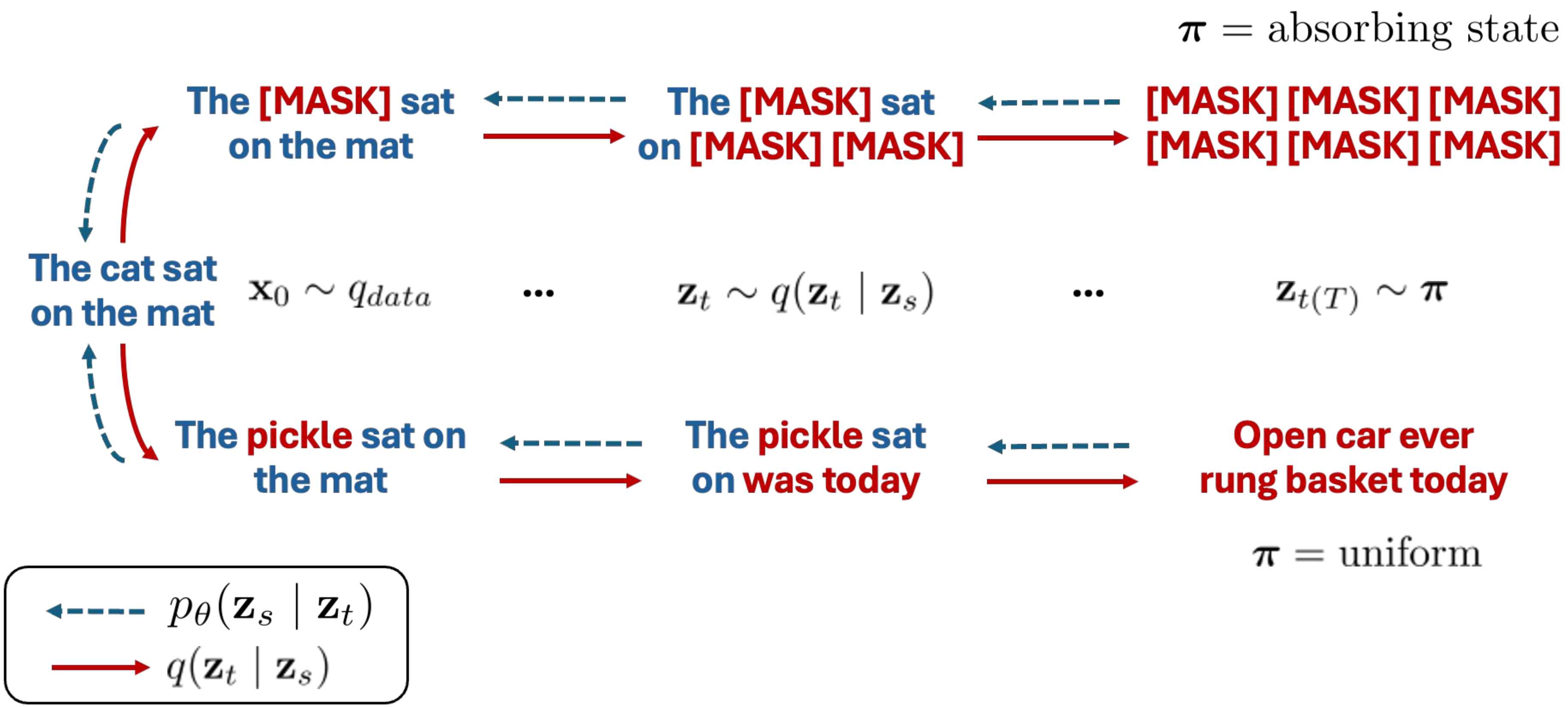
Uniform Diffusion Language Models

For Masked Diffusion Models, generated tokens are ‘fixed’

$$q(\mathbf{z}_s \mid \mathbf{z}_t, \mathbf{x}) = \begin{cases} \mathrm{Cat}(\mathbf{z}_s; \mathbf{z}_t), & \mathbf{z}_t \neq \mathbf{m} \\ \mathrm{Cat}\left(\mathbf{z}_s; \frac{\alpha_s - \alpha_t}{1-\alpha_t}\mathbf{x} + \frac{1-\alpha_s}{1-\alpha_t}\mathbf{m}\right), & \mathbf{z_t} = \mathbf{m} \end{cases}$$

Unmasked tokens are ‘locked-in’ (even if incorrect!)

$$p_\theta(\mathbf{z}_s \mid \mathbf{z}_t) = \begin{cases} \mathrm{Cat}(\mathbf{z}_s; \mathbf{z}_t), & \mathbf{z}_t \neq \mathbf{m} \\ \mathrm{Cat}\left(\mathbf{z}; \frac{\alpha_s - \alpha_t}{1-\alpha_t}\mathbf{x}_\theta + \frac{1-\alpha_s}{1-\alpha_t}\mathbf{m}\right), & \mathbf{z}_t = \mathbf{m} \end{cases}$$



$\boldsymbol{\pi}$ = absorbing state
The [MASK] sat on the mat
The [MASK] sat on [MASK] [MASK]
[MASK] [MASK] [MASK] [MASK] [MASK] [MASK]
The cat sat on the mat
$\mathbf{x}_0 \sim q_{data}$
...
$\mathbf{z}_t \sim q(\mathbf{z}_t \mid \mathbf{z}_s)$
...
$\mathbf{z}_{t(T)} \sim \boldsymbol{\pi}$
The pickle sat on the mat
The pickle sat on was today
Open car ever rung basket today
$\boldsymbol{\pi}$ = uniform
$p_\theta(\mathbf{z}_s \mid \mathbf{z}_t)$
$q(\mathbf{z}_t \mid \mathbf{z}_s)$

Experiments

Table 1: UDLM performs best with smaller vocabs. Perplexity (↓) on various datasets. Best values are **bolded.** * indicates values reported from early stopping on the validation set; otherwise validation performance at the end of training is used. †From Sahoo et al. (2024a). $From Lou et al. (2023).

	|Vocab.|	AR	MDLM	UDLM
Species10	12	**2.88**	$3.17_{\leq}$	$3.15_{\leq}$
QM9*	40	2.19	$2.12_{\leq}$	$\mathbf{2.02}_{\leq}$
CIFAR10	256	-	$\mathbf{9.14}_{\leq}$	$11.21_{\leq}$
text8	35	**2.35**$	$2.62_{\leq}$	$2.71_{\leq}$
Amazon*	30,522	**21.67**	$24.93_{\leq}$	$27.27_{\leq}$
LM1B	30,522	**22.32**†	$27.04^{\dagger}_{\leq}$	$31.28_{\leq}$

Table 2: UDLM outperforms other uniform discrete diffusion on text8. Best value is **bolded** & best uniform diffusion value is underlined. †From Lou et al. (2023). *From Shi et al. (2024).

Method	BPC (↓)
Discrete Uniform Diffusion	
D3PM Uniform† (Austin et al., 2021)	$1.61_{\leq}$
SEDD Uniform† (Lou et al., 2023)	$1.47_{\leq}$
UDLM (*Ours*)	$\underline{1.44}_{\leq}$

Table 3: UDLM outperforms other uniform discrete diffusion on LM1B. Best value is **bolded** & best discrete uniform diffusion value is underlined. †From Sahoo et al. (2024a). *From Lou et al. (2023). $From Austin et al. (2021).

Method	PPL (↓)
Discrete Uniform Diffusion	
D3PM Uniform$ (Austin et al., 2021)	$137.9_{\leq}$
SEDD Uniform* (Lou et al., 2023)	$40.25_{\leq}$
UDLM (*Ours*)	$\underline{31.28}_{\leq}$

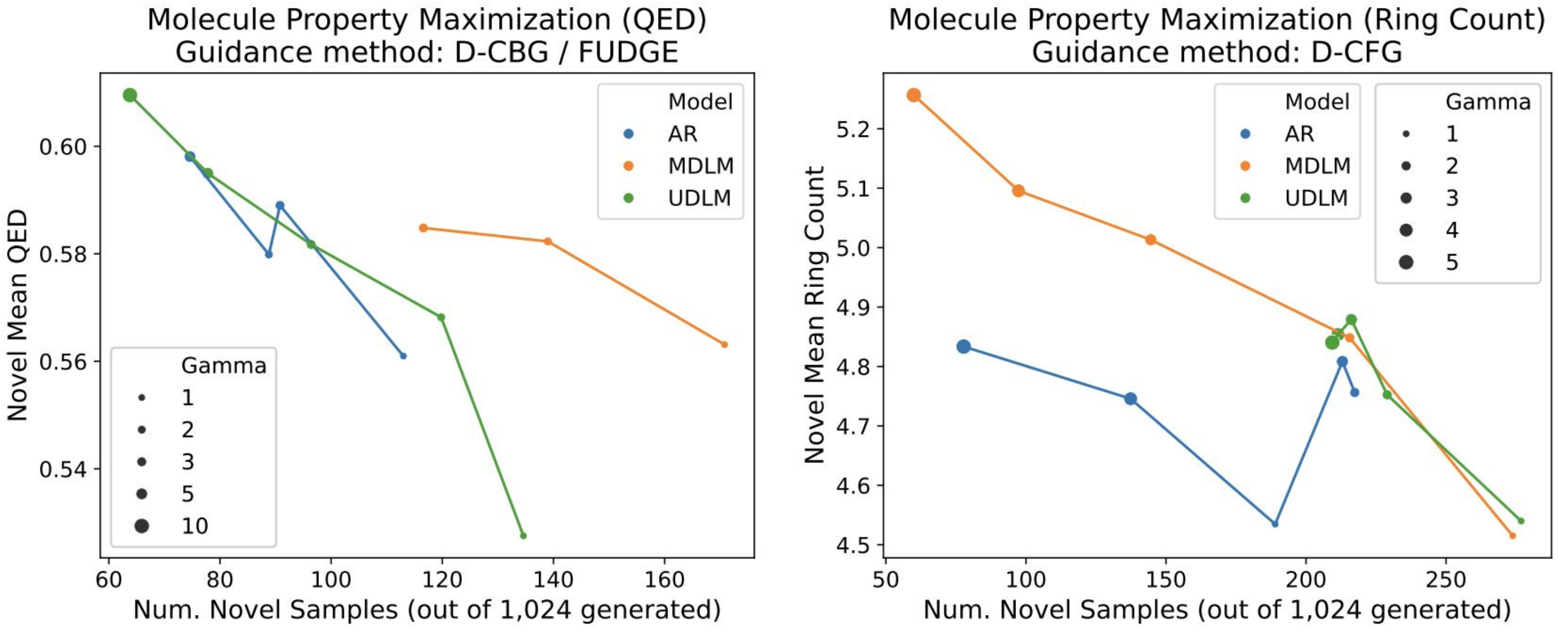
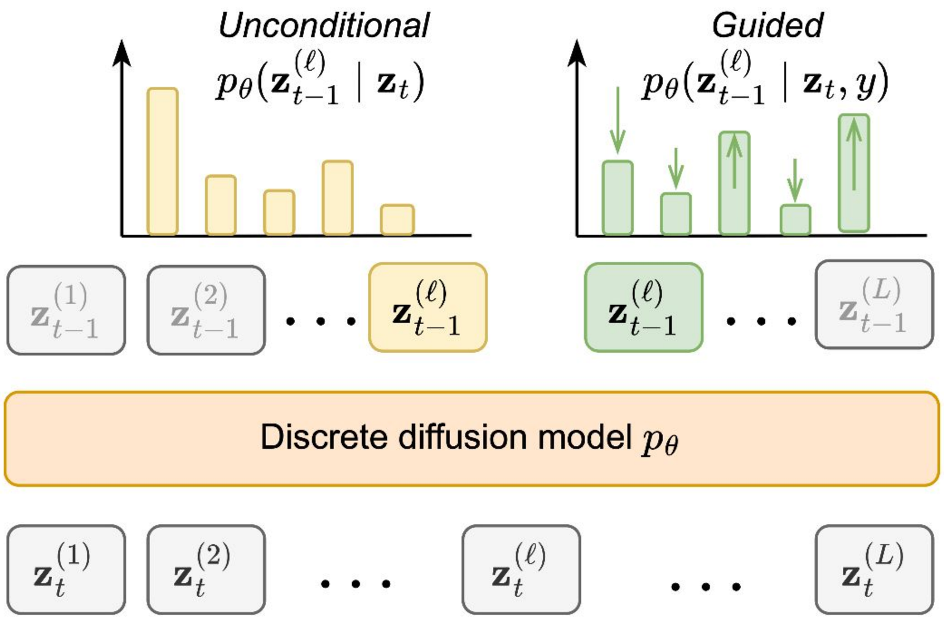
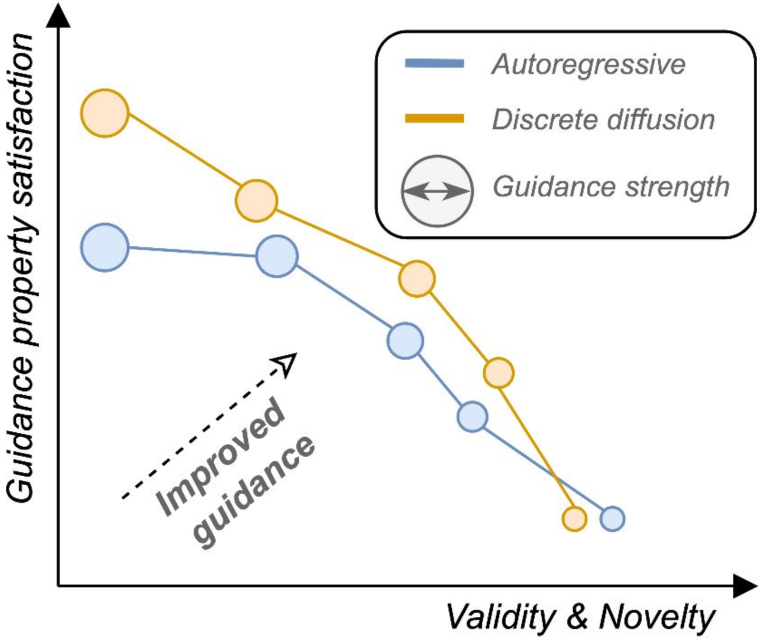


Figure 3: Diffusion models extend the steer-ability Pareto frontier. *(Left)* D-CBG outperforms FUDGE classifier guidance when maximizing drug-likeness (QED). *(Right)* D-CFG with diffusion better trades-off novel generation and ring-count maximization compared to AR.

Conclusion





https://arxiv.org/abs/2412.10193

https://github.com/kuleshov-group/discrete-diffusion-guidance

Thank you!

Thursday April 24

Hall 3 + Hall 2B

Poster session 1