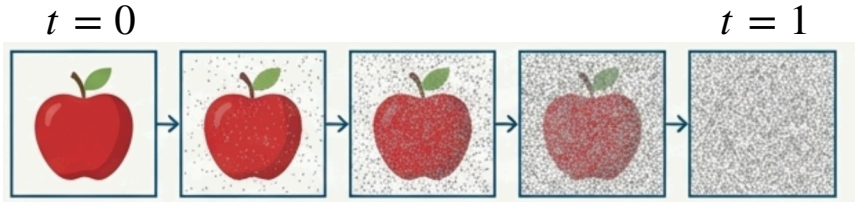
Demystifying Diffusion Objectives

Reweighted Losses are Better Variational Bounds

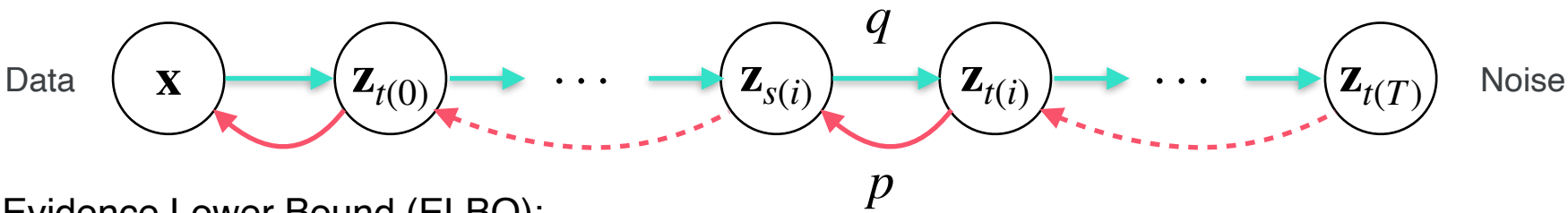
Jiaxin Shi
Work done @Google DeepMind
2025/12/6

jiaxins.io

Diffusion models



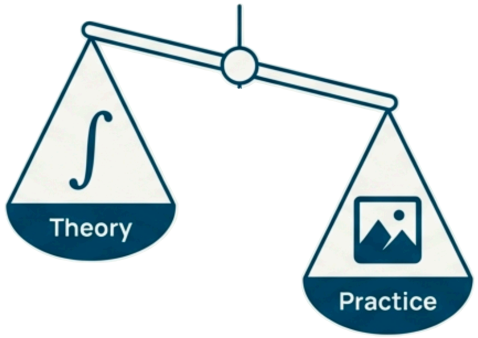
Breaking the time into T intervals $[s(i), t(i)]$, $t(i) = i/T$, $s(i) = (i-1)/T$, we get a deep hierarchical VAE:



Evidence Lower Bound (ELBO):

$$\log p(\mathbf{x}) \geq \mathbb{E}_{q(\mathbf{z}_{t(0)}|\mathbf{x})}[\log p(\mathbf{x}\,|\,\mathbf{z}_{t(0)})] - \mathrm{KL}(q(\mathbf{z}_{t(T)}\,|\,\mathbf{x})\|p(\mathbf{z}_{t(T)})) - \sum_{i=2}^{T} \mathbb{E}_{q(\mathbf{z}_{t(i)}|\mathbf{x})}\Big[\mathrm{KL}(q(\mathbf{z}_{s(i)}\,|\,\mathbf{z}_{t(i)}, \mathbf{x})\|p(\mathbf{z}_{s(i)}\,|\,\mathbf{z}_{t(i)}))\Big]$$

Diffusion loss



Negative ELBO (cont-time limit)

Derived from maximizing the log-likelihood of data.

$$\mathscr{L}_\infty(\mathbf{x}) = \frac{1}{2}\mathbb{E}_{t\sim U[0,1],\boldsymbol{\epsilon}\sim\mathcal{N}(0,\mathbf{I})}\left[\lambda'(t)\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t)\|^2\right]$$

Reweighted Loss

Practically a "reweighted" version is used

$$\mathscr{L}_\infty(\mathbf{x}) = \frac{1}{2}\mathbb{E}_{t\sim U[0,1],\boldsymbol{\epsilon}\sim\mathcal{N}(0,\mathbf{I})}\left[\tilde{w}(t)\lambda'(t)\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t)\|^2\right]$$

The popular "simple" objective sets $\tilde{w}(t) = 1/\lambda'(t)$. This change has a massive positive impact on image sample quality.

Weighting functions used in continuous diffusion

Based on Kingma & Gao (2023)

Name	Parameterization	$\lambda(t)$	$\hat{w}(\lambda)$	$\tilde{w}(t)$
EDM	mean prediction	$F^{-1}_{\mathcal{N}(2.4,2.4^2)}(1-t)$	$p_{\mathcal{N}(2.4,2.4^2)}(\lambda)\frac{e^{-\lambda}+0.5^2}{0.5^2}$	$w(\lambda(t))$
IDDPM	ϵ prediction	$-2\log\tan(\frac{\pi}{2}t)$	$\operatorname{sech}(\frac{\lambda}{2})$	$2\sin(\frac{\pi}{2}t)\cos(\frac{\pi}{2}t)$
Sigmoid	ϵ prediction	$-2\log\tan(\frac{\pi}{2}t)$	$\operatorname{sigmoid}(-\lambda+k)$	$\frac{1}{1+e^{-k}\tan(\frac{\pi}{2}t)^{-2}}$
FM	velocity prediction	$2\log\frac{1-t}{t}$	$e^{-\frac{\lambda}{2}}$	$\frac{t}{1-t}$

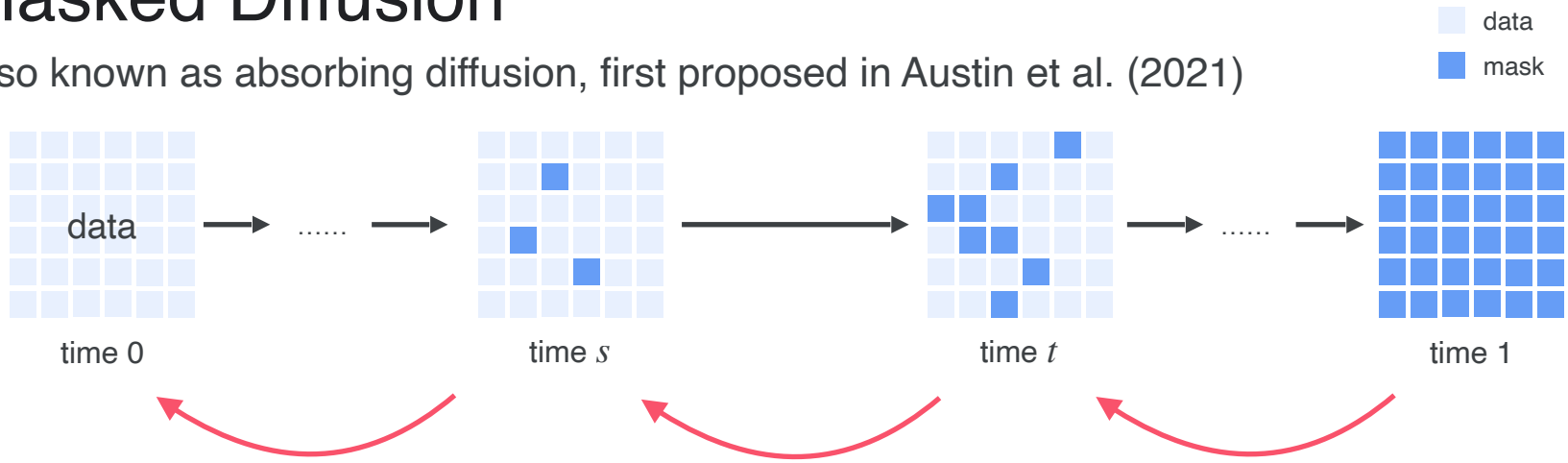
It's less well-understood why the reweighted loss has better sample quality

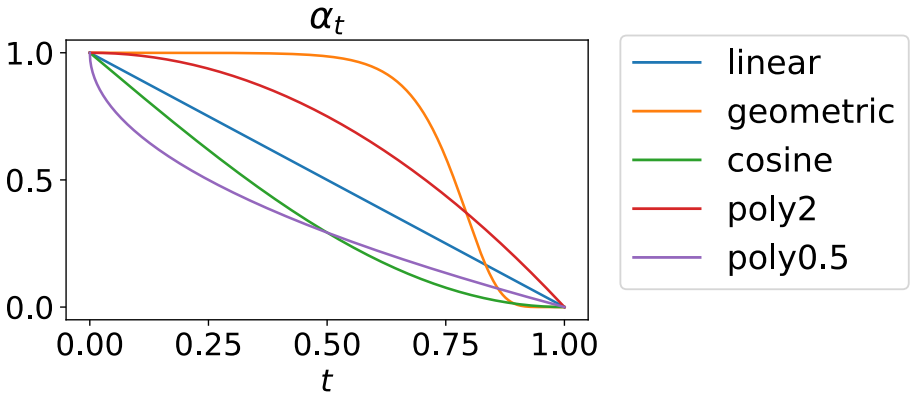
- Kingma & Gao (2023) interpreted the loss as the integral of ELBOs under Gaussian noise-augmented data, thus prioritizing perceptually important low-frequency signals.
- It's unclear how to generalize this beyond standard continuous diffusion

Kingma, D., & Gao, R. (2023). Understanding diffusion objectives as the ELBO with simple data augmentation.

Masked Diffusion

Also known as absorbing diffusion, first proposed in Austin et al. (2021)





Masking schedule α_t: The expected proportion of unmasked elements at t

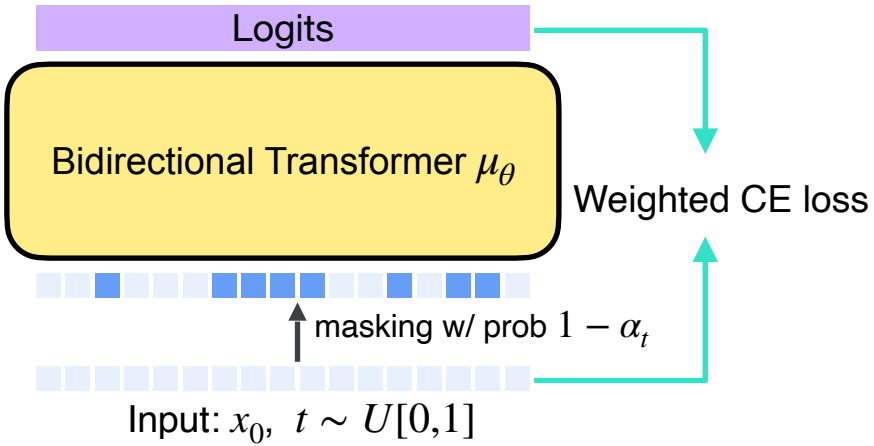
Austin et al. (2021). Structured denoising diffusion models in discrete state-spaces.

Simplified Masked Diffusion Models (MD4)

Continuous-time ELBO

$$\log p_\theta(x_0) \geq -\int_0^1 \frac{\alpha_t'}{1-\alpha_t} \mathbb{E}_{q(x_t|x_0)} \left[\sum_{n:x_t^{(n)}=m} (x_0^{(n)})^\top \log \mu_\theta^{(n)}(x_t, t) \right] \mathrm{d}t.$$

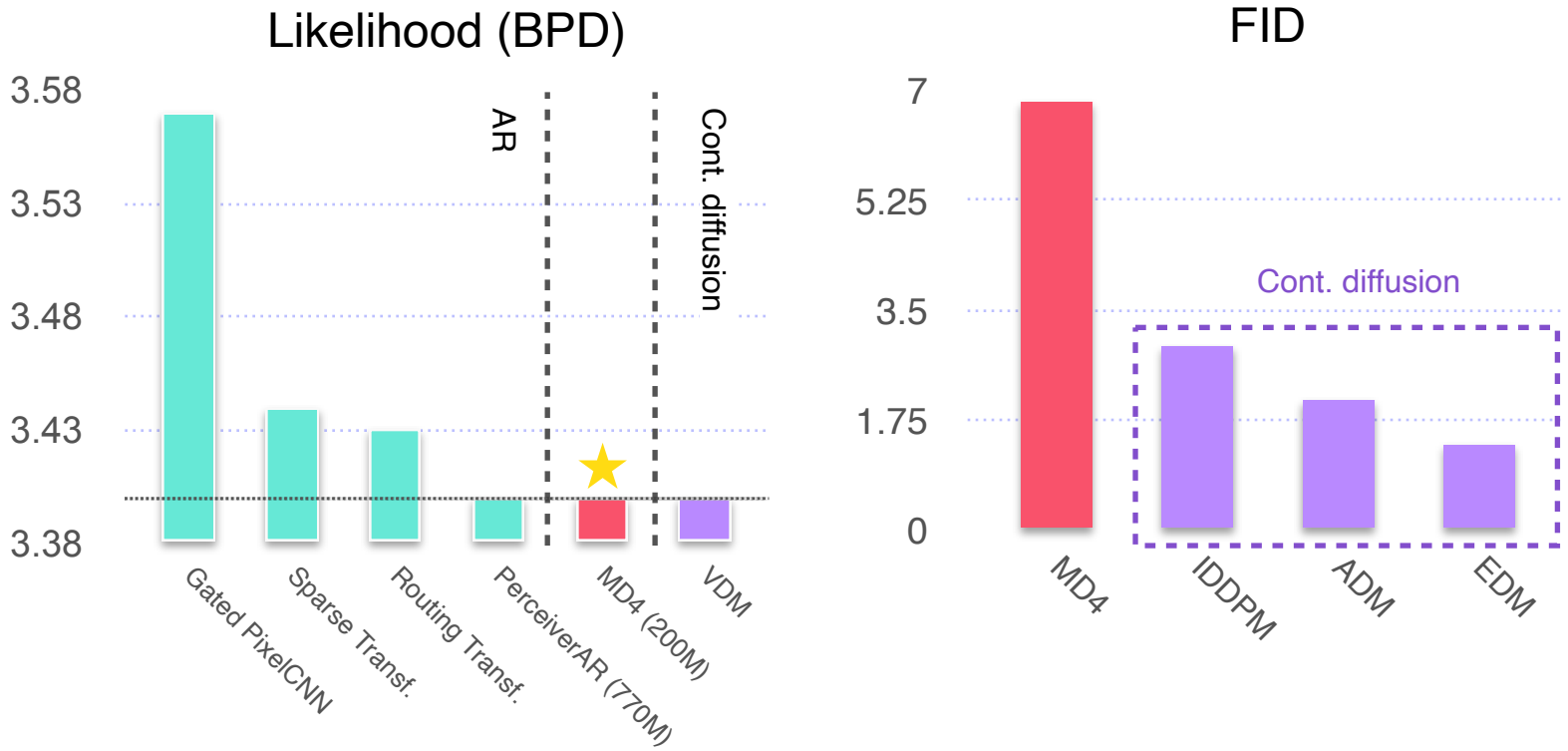
- Maximum likelihood is as simple as training an ensemble of BERTs
- Many popular diffusion LLMs are now based on masked diffusion and this objective
- Promise for universal multimodal generation



Shi et al. (2024). Simplified and Generalized Masked Diffusion for Discrete Data.
Concurrent work RADD and MDLM (Ou et al., 2024; Sahoo et al., 2024) studied similar loss for language

Can masked diffusion match continuous diffusion in image generation?

- MD4 reports likelihood comparable to strong continuous diffusion (VDM)
- **Challenge:** sample quality (as measured by FID) is still behind



Demystifying Diffusion Objectives: Reweighted Losses are Better Variational Bounds

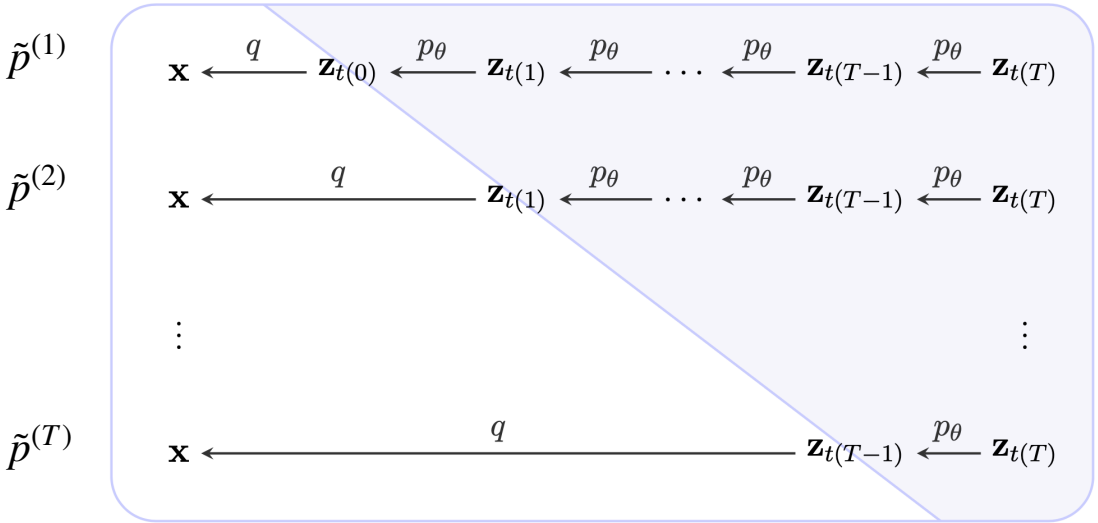
Jiaxin Shi[1] and Michalis K. Titsias[1]

[1]Google DeepMind
{jiaxins,mtitsias}@google.com

arxiv.org/abs/2511.19664

Diffusion models with “optimal decoders”

- Consider a diffusion model with a modified generative process that, for the last i reverse steps, does not use our learned denoiser.
- Instead, it uses the ground truth reverse transition $q(\mathbf{x} \mid \mathbf{z}_{t(i)}) \propto q(\mathbf{z}_{t(i)} \mid \mathbf{x}) q(\mathbf{x})$
- Define a sequence of these models $\tilde{p}^{(i)}$ for $i = 1$ to T



Theorem 1: More "optimal" reverse steps lead to better ELBOs

Let $\mathcal{L}^{(i)}(\mathbf{x})$ be the ELBO of the model on the i-th row

$$\mathcal{L}^{(i)}(\mathbf{x}) \triangleq \mathbb{E}_{q(\mathbf{z}_{s(i)}|\mathbf{x})}[\log q(\mathbf{x}|\mathbf{z}_{s(i)})] - \mathrm{KL}(q(\mathbf{z}_{t(T)}|\mathbf{x})\|p(\mathbf{z}_{t(T)})) - \sum_{j=i}^{T} \mathbb{E}_{q(\mathbf{z}_{t(i)}|\mathbf{x})}\left[\mathrm{KL}(q(\mathbf{z}_{s(i)}|\mathbf{z}_{t(i)}, \mathbf{x})\|p(\mathbf{z}_{s(i)}|\mathbf{z}_{t(i)}))\right]$$

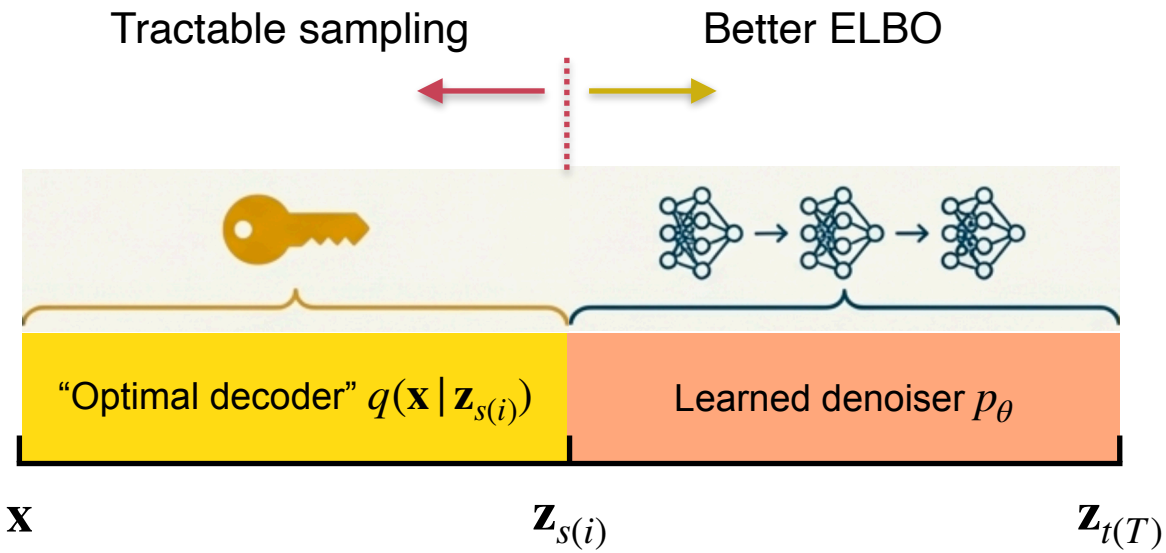
From top to bottom, the ELBO improves

$$\mathbb{E}_{q(\mathbf{x})}[\mathcal{L}^{(i+1)}(\mathbf{x})] \geq \mathbb{E}_{q(\mathbf{x})}[\mathcal{L}^{(i)}(\mathbf{x})]$$

Since $\mathrm{KL}(q(\mathbf{x})\|\tilde{p}^{(i)}(\mathbf{x})) = -\mathbb{E}_{q(\mathbf{x})}[\log \tilde{p}^{(i)}(\mathbf{x})] + C \leq -\mathbb{E}_{q(\mathbf{x})}[\mathcal{L}^{(i)}(\mathbf{x})] + C$, this implies that incorporating an additional optimal reverse step results in a smaller upper bound on data-model KL divergence

Although the decoder term in tractable to compute, it is constant wrt. θ. We can still train the denoiser using the improved ELBO.

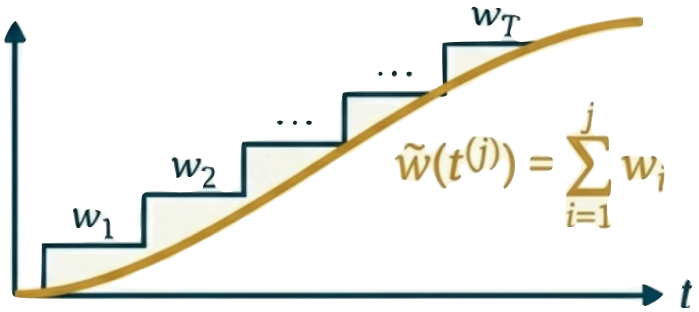
Tradeoff between better ELBO & tractable sampling



Can we construct an objective that leverages the improved ELBOs while at the same time explicitly trains the denoiser at all noise levels?

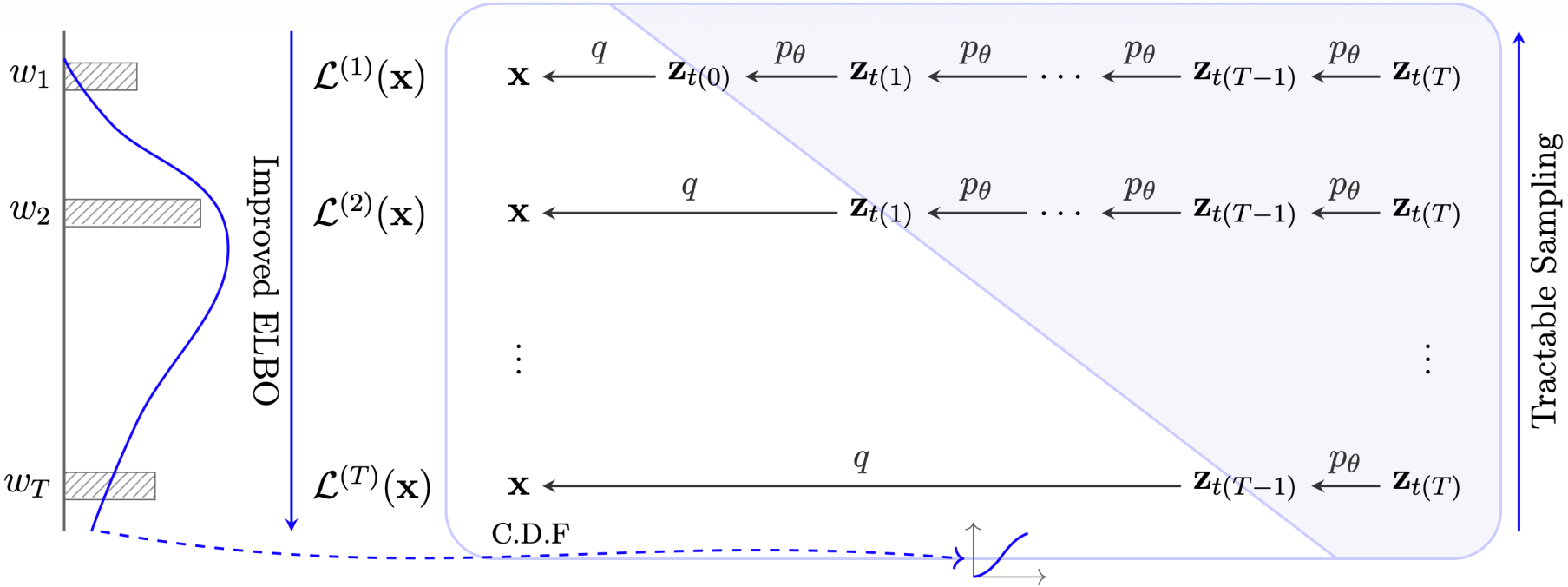
Theorem 2: Common diffusion objectives are a weighted sum of the improved ELBOs

$$\lim_{T\to\infty}\sum_{i=1}^{T} w_i \mathcal{L}^{(i)}(\mathbf{x}) = \mathcal{L}^{\tilde{w}}(\mathbf{x}) + C$$



Conclusion: Reweighted losses work better because they provide improved training signals for denoisers at high-noise level. This also implies $\tilde{w}(t)$ must be monotonic increasing, which aligns with the conclusion of Kingma & Gao (2023).

Reweighted Losses are Better Variational Bounds



Diffusion objectives: $\mathcal{L}^{\tilde{w}}(\mathbf{x}) = \lim_{T\to\infty} \sum_{i=1}^{T} w_i \mathcal{L}^{(i)}(\mathbf{x}) = \int_0^1 \tilde{w}(t) \mathbb{E}_{q(\mathbf{z}_t|\mathbf{x})} \left[L_{\text{denoise}}(\mathbf{z}_t, \mathbf{x}, t) \right] dt + \text{C}$

Reweighted loss for masked diffusion

- The framework is general and can be applied to any diffusion processes
- Repeating the Theorem 2 derivation for masked diffusion yields:

$$\mathscr{L}^{\tilde{w}}(\mathbf{x}) = -\int_0^1 \tilde{w}(t)\frac{\alpha_t'}{1-\alpha_t}\mathbb{E}_{q(\mathbf{z}_t|\mathbf{x})}\left[\delta_{\mathbf{z}_t,m}\mathbf{x}^\top \log \mu_\theta(\mathbf{z}_t)\right] dt$$

How to choose $\tilde{w}(t)$?

- Migrate the common continuous diffusion weightings
- Match in log-SNR (λ) space instead of t space to maintain schedule invariance

Name	$\lambda(t)$	$\hat{w}(\lambda)$	$\tilde{w}(t)$
EDM	$\log \frac{\alpha_t}{1-\alpha_t}$	$p_{\mathcal{N}(2.4,2.4^2)}(\lambda)\frac{e^{-\lambda}+0.5^2}{0.5^2}$	$w(\lambda(t))$
IDDPM		$\operatorname{sech}(\frac{\lambda}{2})$	$2\sqrt{\alpha_t(1-\alpha_t)}$
Sigmoid		$\operatorname{sigmoid}(-\lambda+k)$	$\frac{1-\alpha_t}{1-(1-e^{-k})\alpha_t}$
FM		$e^{-\frac{\lambda}{2}}$	$\sqrt{\frac{1-\alpha_t}{\alpha_t}}$

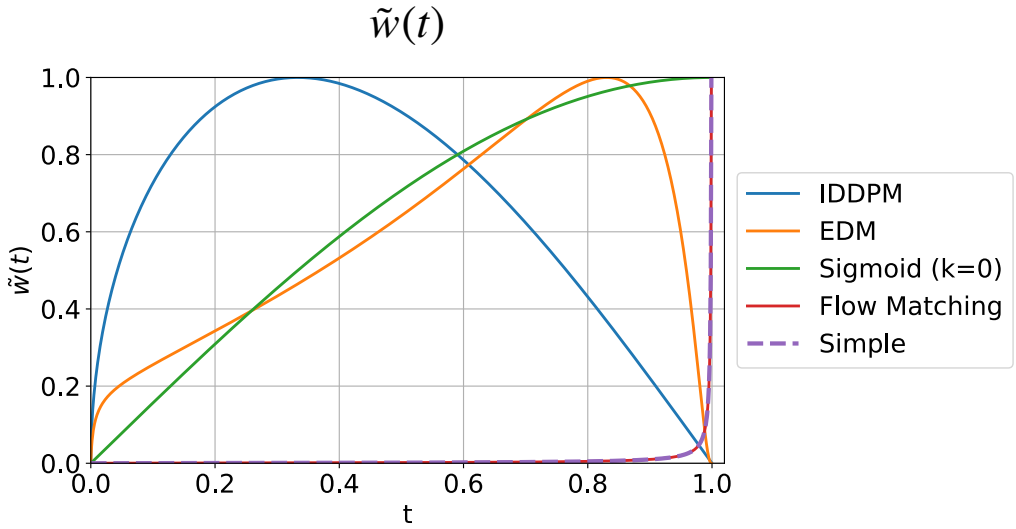
The “simple” weighting

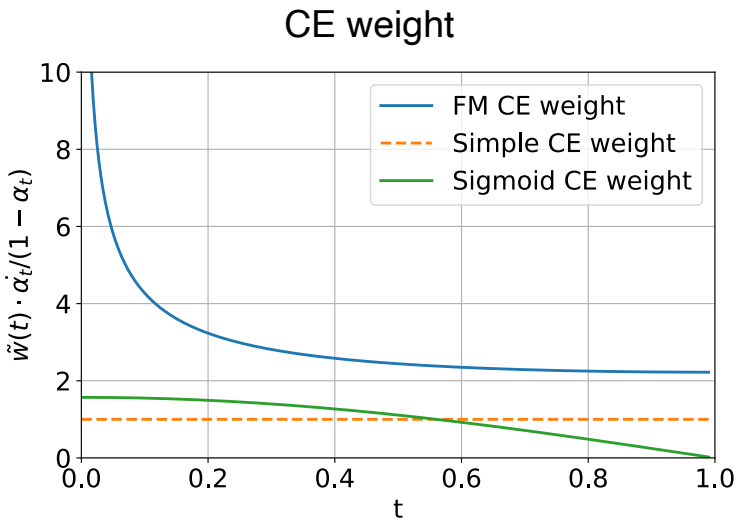
- Masked image models (MaskGIT/MAR): a class of generative models closely related to masked diffusion
- Loss computation in these models: sum all mask-prediction losses in the batch and normalize by #masks
- Under CLT (large minibatch), this equals using a constant CE weighting, or equivalently, $\tilde{w}(t) = \frac{1-\alpha_t}{\alpha_t'}$

$$\mathscr{L}^{\tilde{w}}(\mathbf{x}) = -\int_0^1 \mathbb{E}_{q(\mathbf{z}_t|\mathbf{x})}\left[\delta_{\mathbf{z}_t,m}\mathbf{x}^\top \log \mu_\theta(\mathbf{z}_t)\right] dt$$

Chang et al. (2022). MaskGIT: Masked generative image transformer.
Li et al. (2024). Autoregressive image generation without vector quantization.

Weighting functions for masked diffusion models

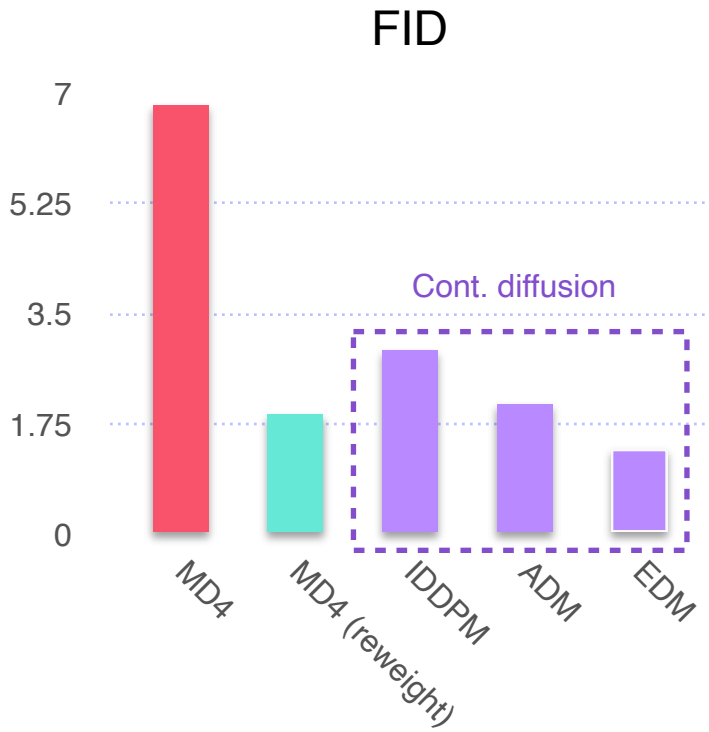




Impact of different weighting functions

Class-conditional generation of ImageNet 64x64 (pixels)

Method	#Params	FID (↓)	IS (↑)
Gaussian Diffusion			
IDDPM (Nichol and Dhariwal, 2021)		2.92	
ADM (Dhariwal and Nichol, 2021)	296M	2.07	
EDM (Karras et al., 2022)	296M	**1.36**	
VDM++ (Kingma and Gao, 2023)	296M	1.43	63.7
Masked Image Models			
MAR (Li et al., 2025)	479M	2.93	
FractalMAR (Li et al., 2025)		2.72	
Masked Diffusion			
MD4 (ELBO)	204M	6.84	30.3
Weighting:			
- IDDPM (non-monotonic)	204M	11.14	22.9
- EDM (nearly-monotonic)	204M	4.42	37.3
- Sigmoid ($k = 0$)	204M	3.91	40.1
- FM	204M	3.43	43.3
- Simple	204M	**2.96**	**46.7**
- Simple	324M	**1.92**	**57.9**



Sample (324M Params, FID 1.92)

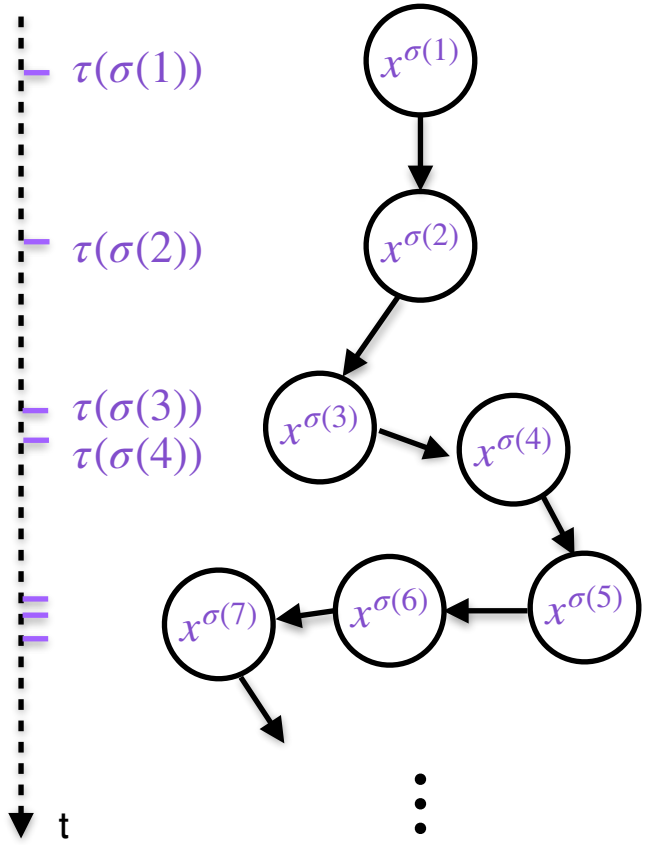
Final thoughts: Diffusion is really not so different from AR

- Masked diffusion is random-order autoregression
- With loss-reweighting, a pixel-space autoregression can give diffusion-level sample quality

Figure: Nano Banana

Thanks

Masked Diffusion as Any-Order AR



Consider the continuous-time reverse process

- $\tau(n)$: unmasking time of the n-th component
- there exists an ordering of the unmasking times

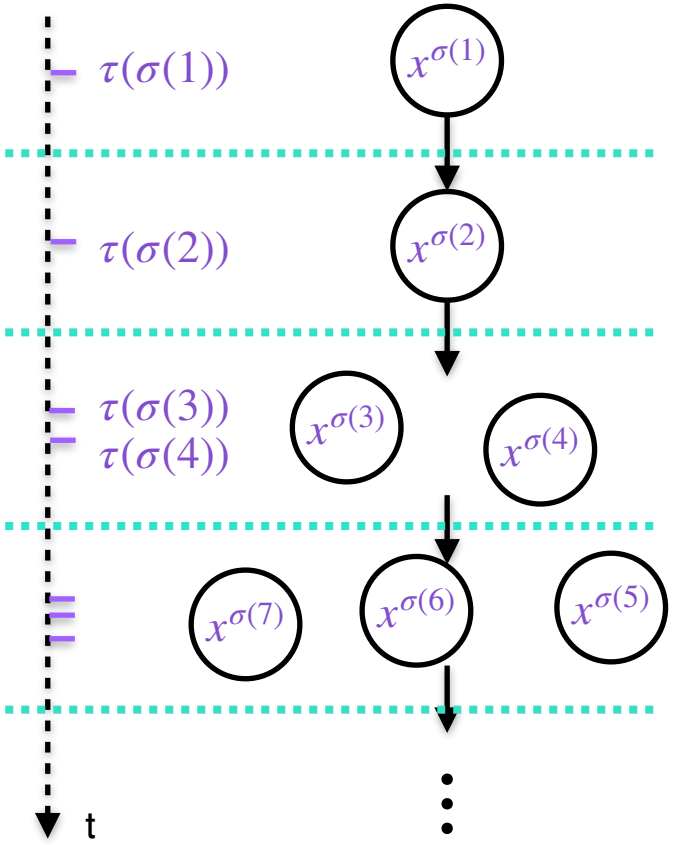
CDF of jump times:

$$P(\tau(n) \leq t) = P(x_t^{(n)} = m) = 1 - \alpha_t$$

- When α_t is global, all $\tau(n)$s share the same distribution, the unmasking order is uniform across all permutations
- The equivalence is no longer true for generalized masked diffusion (Shi et al. 2024)

Uria, B. et al. (2014). A deep and tractable density estimator.
Hoogeboom et al. (2021). Autoregressive diffusion models.

Masked Diffusion as Any-Order AR



Masking schedule: a new degree of freedom in any-order AR

- In continuous time (training), all choices of α_t lead to the same any-order AR model
- In discretized time (sampling), shape of α_t determines the parallel sampling bandwidth