Diffusion-Based Voice Conversion with **FAST** Maximum Likelihood Sampling Scheme

ICLR 2022

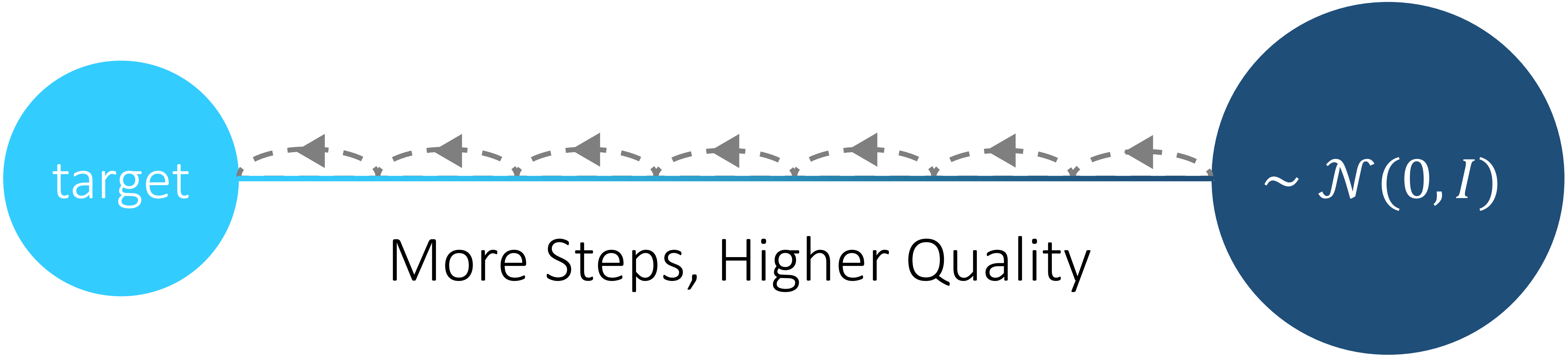
Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, Mikhail Sergeevich Kudinov, Jiansheng Wei

Why Use Diffusion Models?

- More Stable than GAN
- Higher Quality than VAE
- Easier to Design than Flow Models

But **Slower** than Them

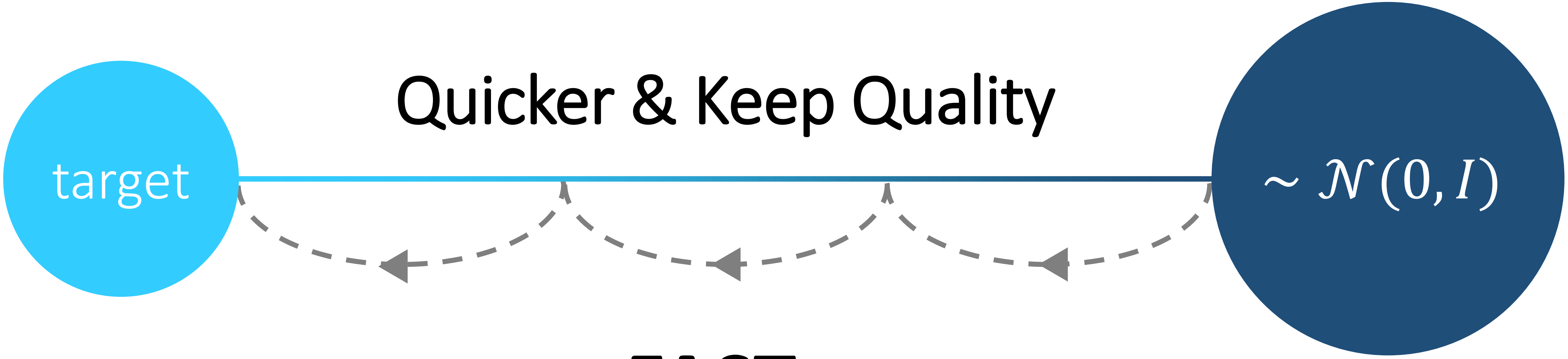
Diffusion Steps



CIFAR10 (32×32)					CelebA (64×64)				
10	20	50	100	1000	10	20	50	100	1000
367.43	133.37	32.72	9.99	**3.17**	299.71	183.83	71.71	45.20	**3.26**

CIFAR10 and CelebA image generation measured in FID.
Image source: Song et al., 2020

Speed Up



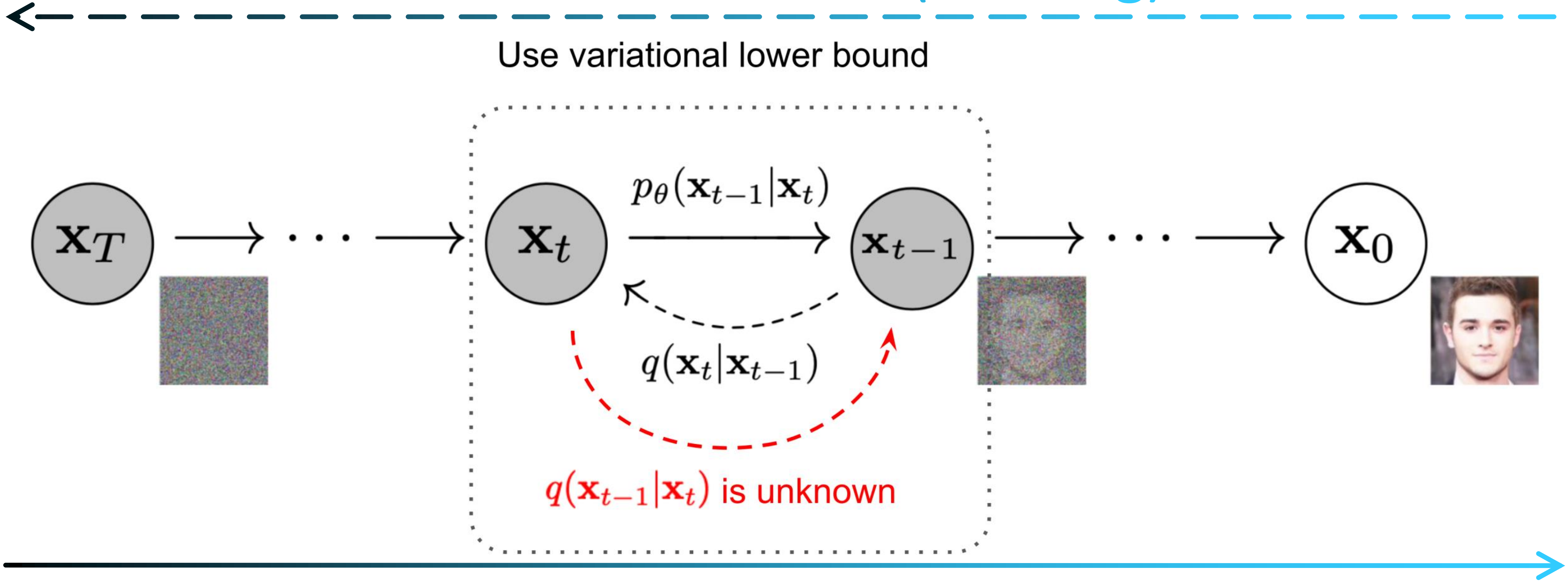
Propose a **FAST** Sampling Scheme

+

Average Voice Encoder

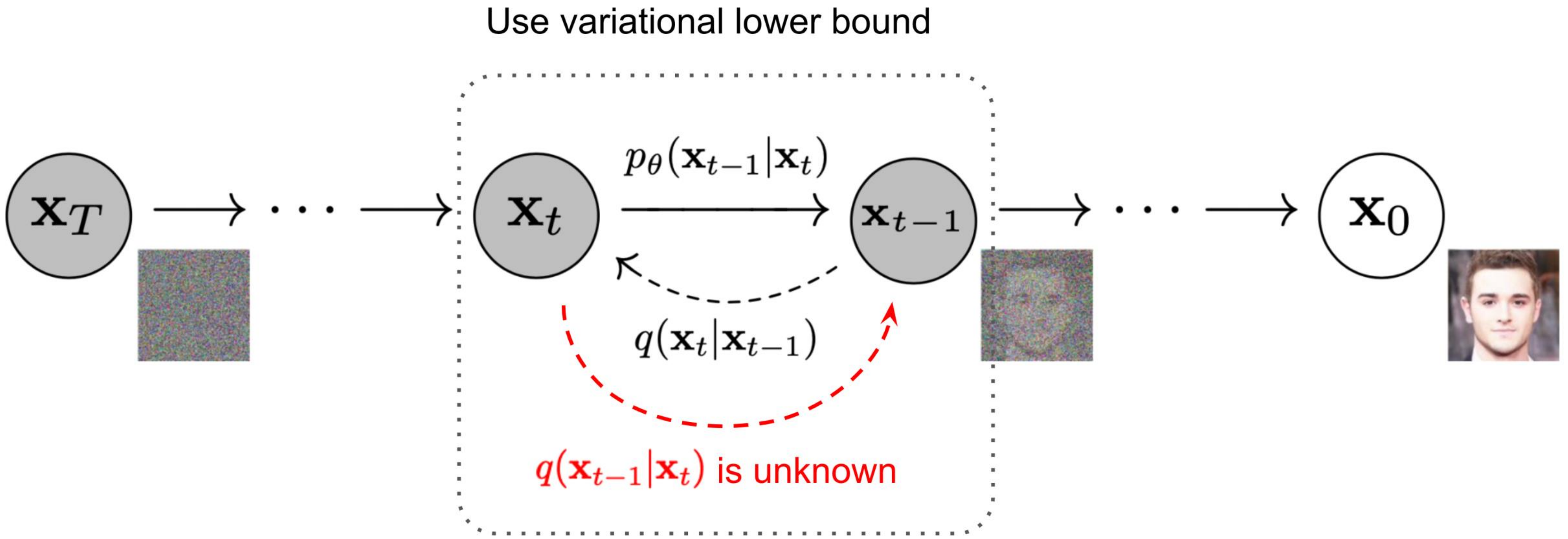
= SOTA Any to Any VC

Forward Diffusion (Training)



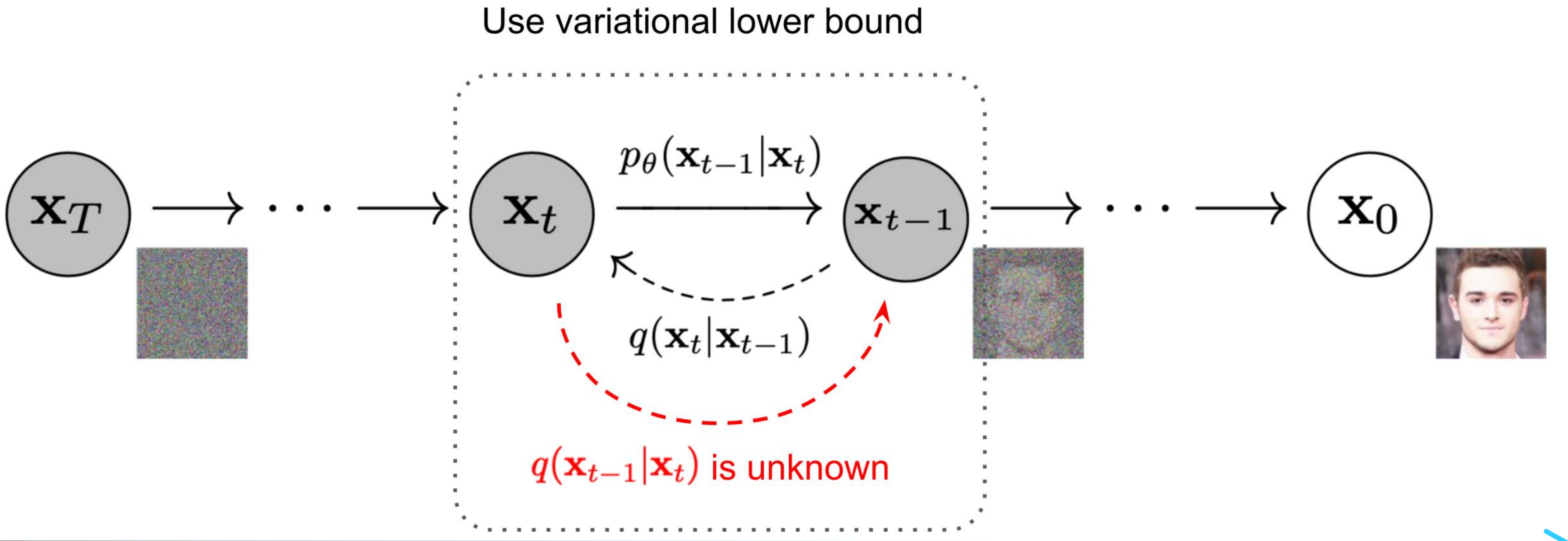
Reverse Diffusion (Inference)

Forward Diffusion (Training)



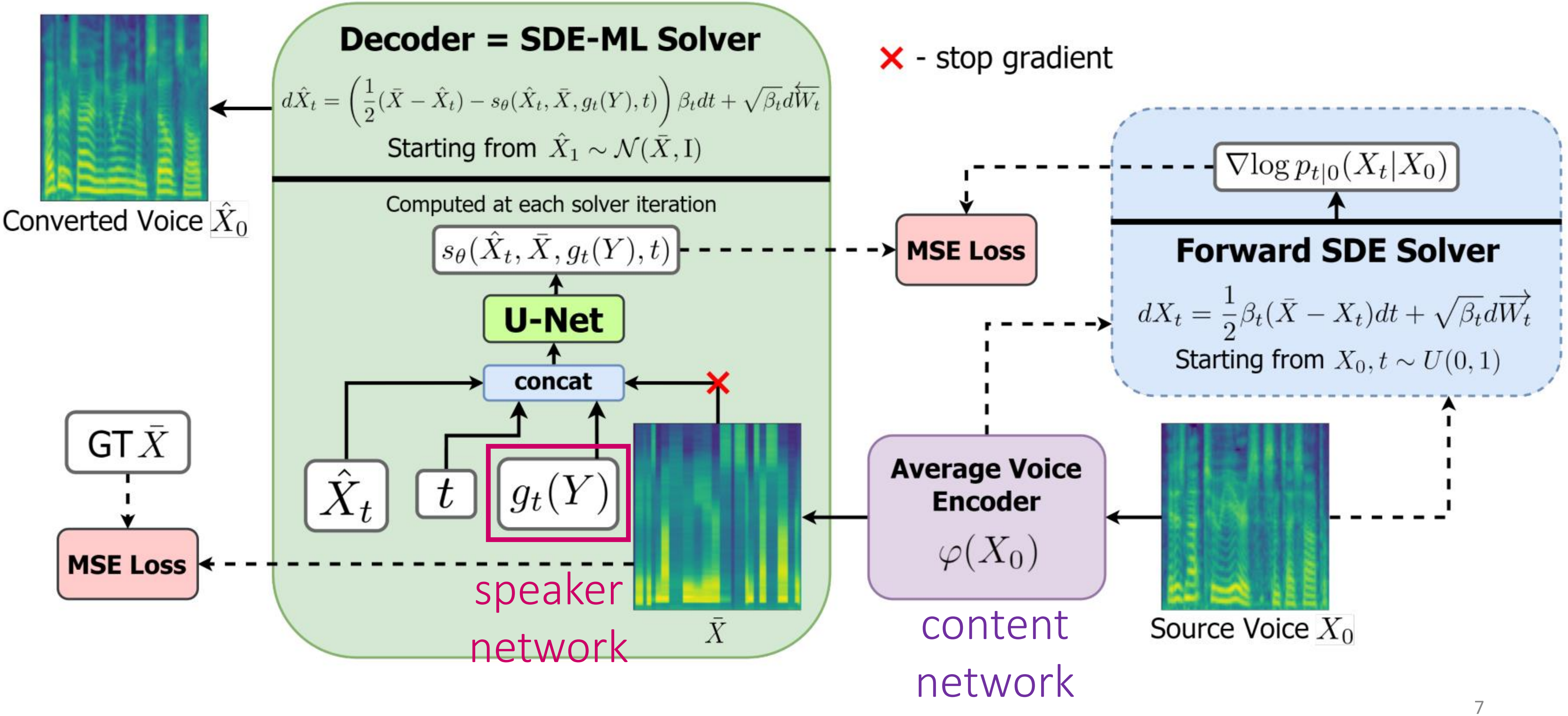
Mix X_0 with noise to get X_t, and train the model to estimate noise.

Reverse the diffusion process and sampling from it, you can generate real samples from Gaussian noise.

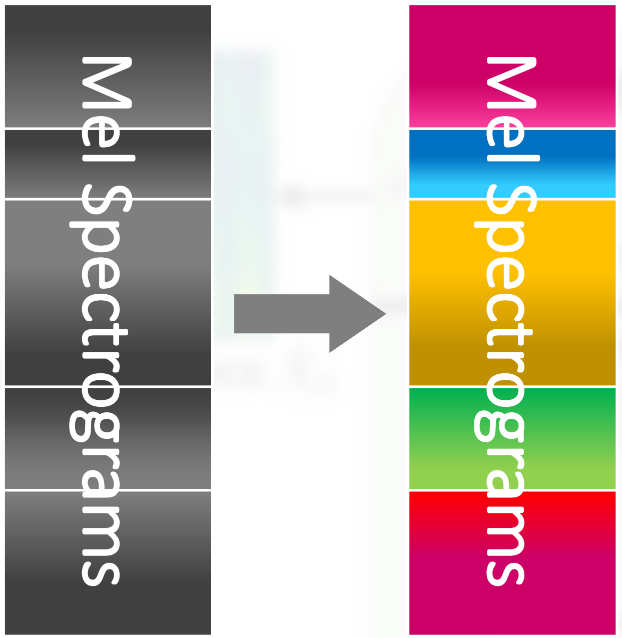


Reverse Diffusion (Inference)

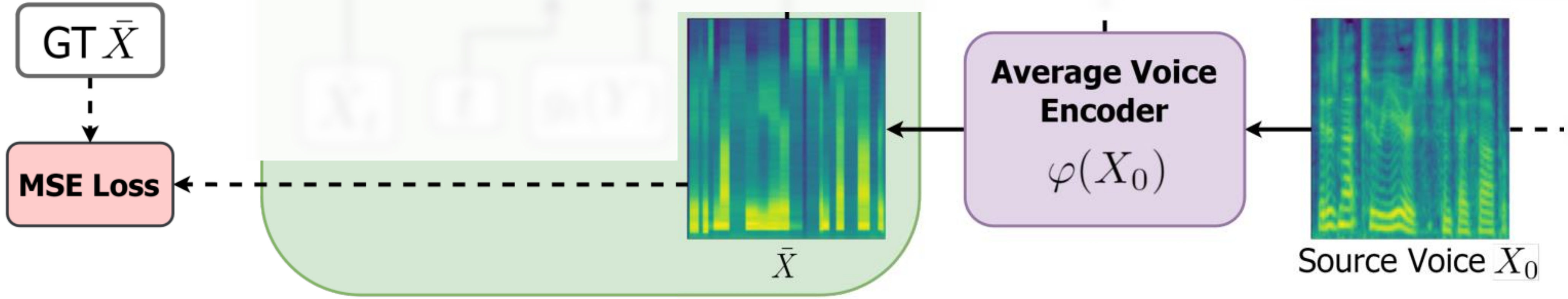
Voice Conversion Diffusion Model



Average Voice Encoder

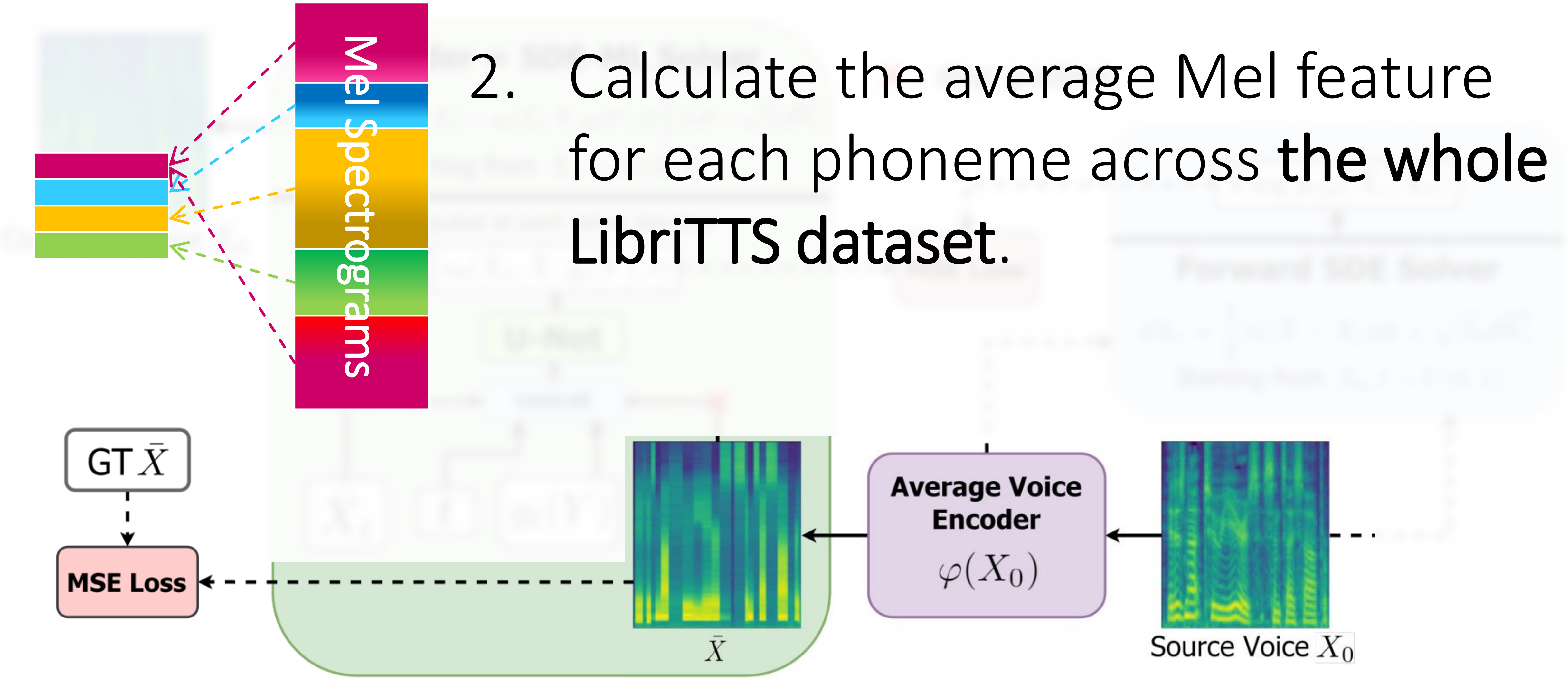


1. Apply Montreal Forced Aligner to align speech frames with phonemes.

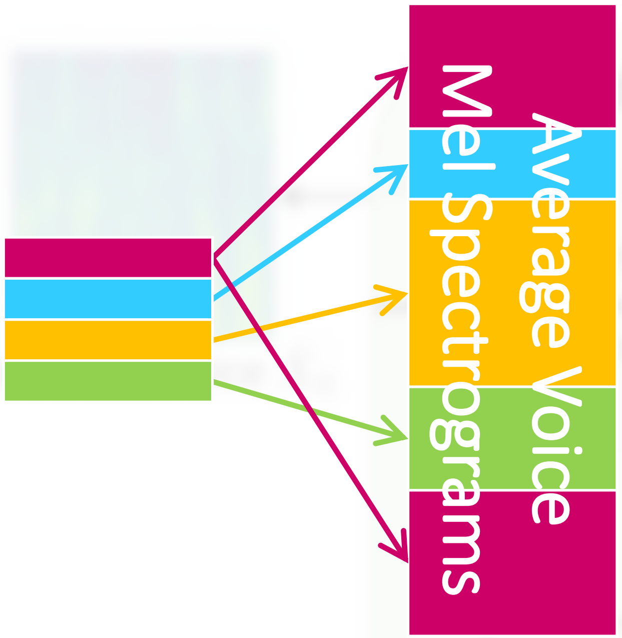


Average Voice Encoder

2. Calculate the average Mel feature for each phoneme across **the whole LibriTTS dataset**.

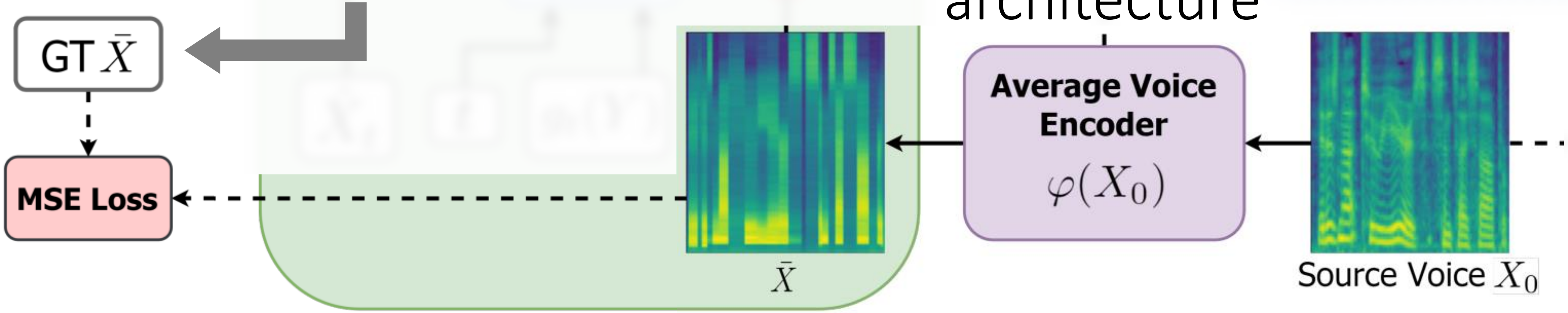


Average Voice Encoder

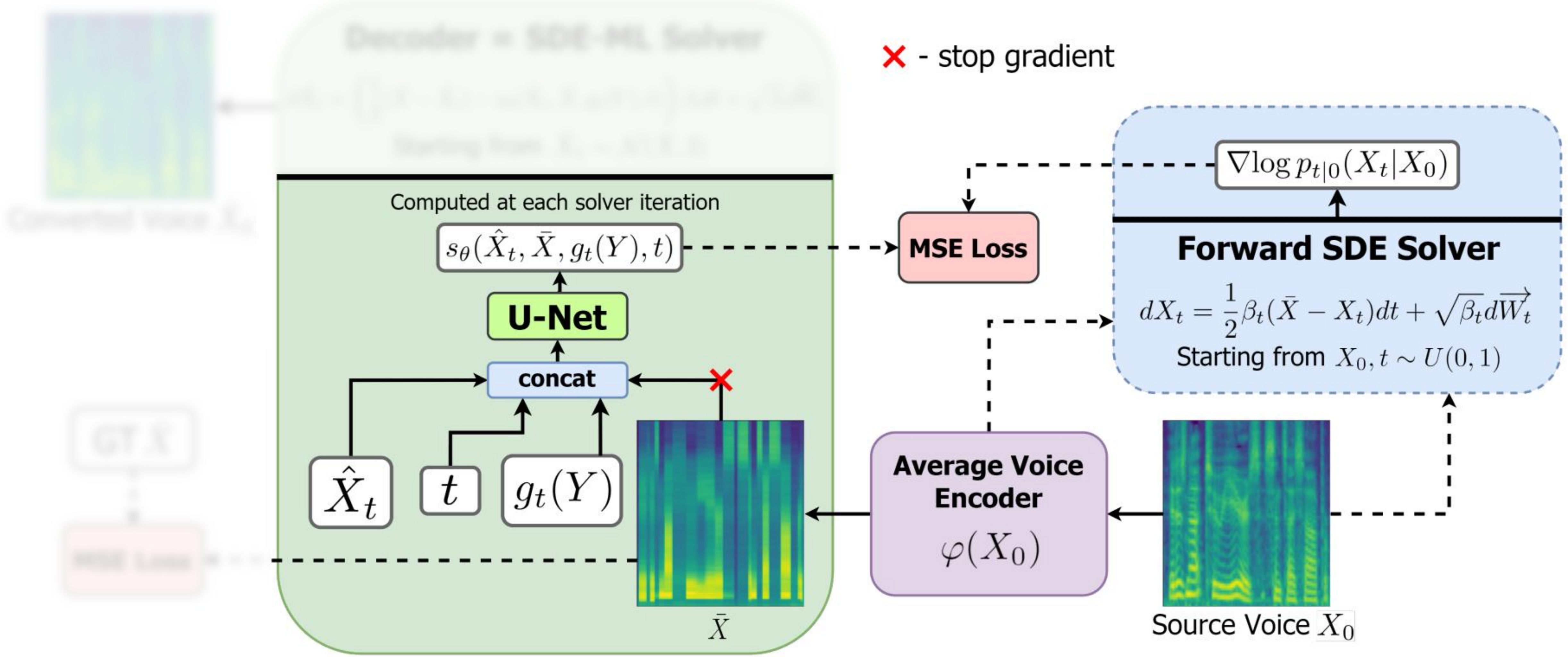


3. Encoder is trained with "Average Voice" Mel Spectrogram as the target.

Transformer-based architecture



Forward Diffusion



Forward Diffusion: Sample X_t

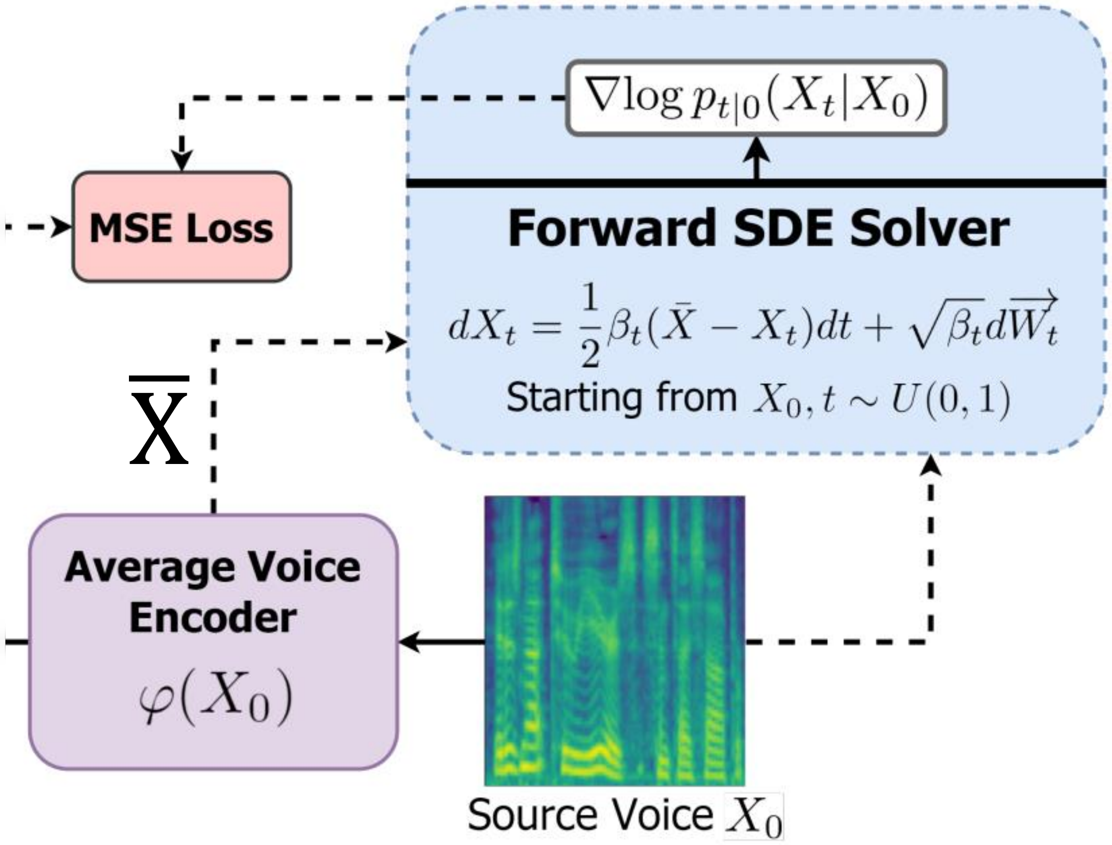
$$X_t = \gamma_{0,t} X_0 + \left(1 - \gamma_{0,t}\right)\bar{X} + \nabla \log p_{t|0}\left(X_t | X_0\right)\sqrt{1 - \gamma_{0,t}^2}$$

$$\nabla \log p_{t|0}(X_t|X_0) \sim \mathcal{N}(0, \mathrm{I})$$

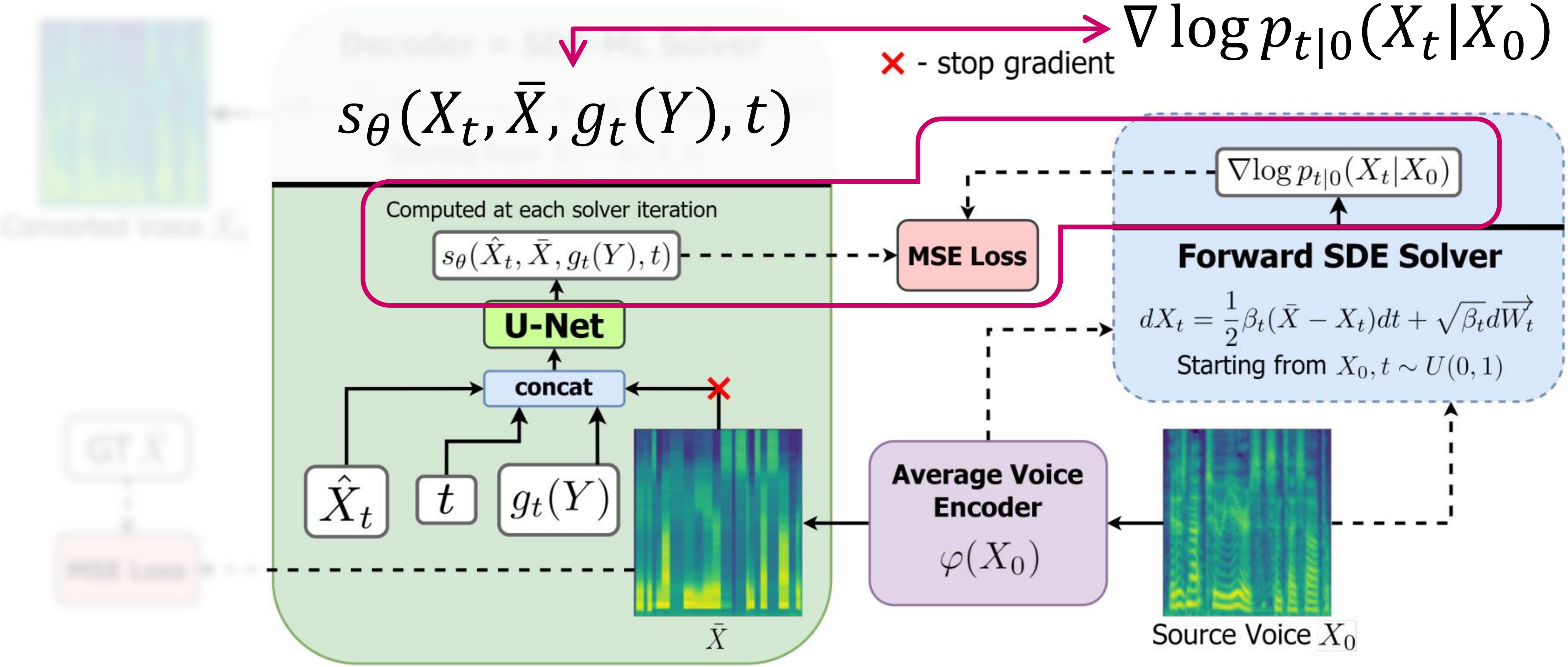
$$\gamma_{s,t} = e^{-\frac{1}{2}\int_s^t \beta_u du}$$

$$\beta_t = \beta_0 + t(\beta_1 - \beta_0)$$

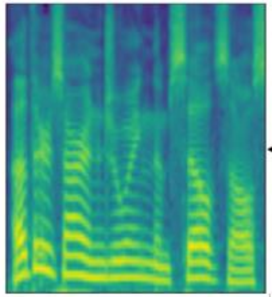
Hyper parameters



Forward Diffusion: Loss



Reverse Diffusion: Euler-Maruyama



Converted Voice $\hat{X}_0$

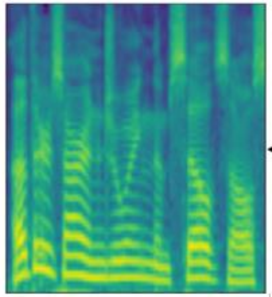
Decoder = SDE-ML Solver

$$d\hat{X}_t = \left(\frac{1}{2}(\bar{X} - \hat{X}_t) - s_\theta(\hat{X}_t, \bar{X}, g_t(Y), t)\right)\beta_t dt + \sqrt{\beta_t}\, d\overleftarrow{W}_t$$

Starting from $\hat{X}_1 \sim \mathcal{N}(\bar{X}, \mathrm{I})$

$$\hat{X}_{t-h} = \hat{X}_t + \overset{=\sqrt{\beta_t h}}{\hat{\sigma}_{t,h}} \nabla \log p_{t|0}(X_t|X_0) + \underbrace{\beta_t h}_{\text{step size}}\left(\left(\frac{1}{2} + \hat{\omega}_{t,h}\right)(\hat{X}_t - \bar{X}) + (1 + \hat{\omega}_{t,h})s_\theta(X_t, \bar{X}, g_t(Y), t)\right)$$

Reverse Diffusion: Maximum Likelihood



Converted Voice $\hat{X}_0$

Decoder = SDE-ML Solver

$$d\hat{X}_t = \left(\frac{1}{2}(\bar{X} - \hat{X}_t) - s_\theta(\hat{X}_t, \bar{X}, g_t(Y), t)\right)\beta_t dt + \sqrt{\beta_t}\, d\overleftarrow{W}_t$$

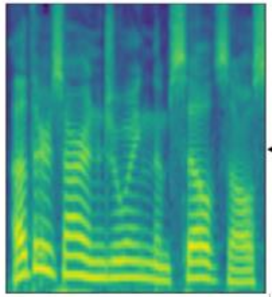
Starting from $\hat{X}_1 \sim \mathcal{N}(\bar{X}, \mathrm{I})$

By Theorem 1.

$$\hat{\sigma}_{t,h} = \sigma^*_{t,h},\ \hat{\omega}_{t,h} = \omega^*_{t,h},\ \hat{\kappa}_{t,h} = \kappa^*_{t,h}$$

$$\hat{X}_{t-h} = \hat{X}_t + \hat{\sigma}_{t,h} \nabla \log p_{t|0}(X_t|X_0) + \underbrace{\beta_t h}_{\text{step size}}\left(\left(\frac{1}{2} + \hat{\omega}_{t,h}\right)(\hat{X}_t - \bar{X}) + (1 + \hat{\kappa}_{t,h}) s_\theta(X_t, \bar{X}, g_t(Y), t)\right)$$

Reverse Diffusion: Maximum Likelihood



Converted Voice $\hat{X}_0$

Decoder = SDE-ML Solver

$$d\hat{X}_t = \left(\frac{1}{2}(\bar{X} - \hat{X}_t) - s_\theta(\hat{X}_t, \bar{X}, g_t(Y), t)\right)\beta_t dt + \sqrt{\beta_t} d\overleftarrow{W}_t$$

Starting from $\hat{X}_1 \sim \mathcal{N}(\bar{X}, \mathrm{I})$

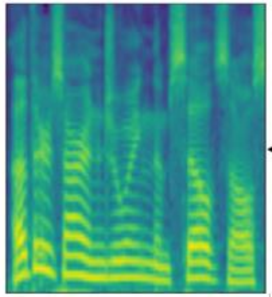
By Theorem 1.

$$\hat{\sigma}_{t,h} = \sigma^*_{t,h}, \ \hat{\omega}_{t,h} = \omega^*_{t,h}, \ \hat{\kappa}_{t,h} = \kappa^*_{t,h}$$

$$\kappa^*_{t,h} = \frac{\nu_{t-h,t}(1-\gamma^2_{0,t})}{\gamma_{0,t}\beta_t h} - 1, \quad \omega^*_{t,h} = \frac{\mu_{t-h,t} - 1}{\beta_t h} + \frac{1+\kappa^*_{t,h}}{1-\gamma^2_{0,t}} - \frac{1}{2},$$

$$(\sigma^*_{t,h})^2 = \sigma^2_{t-h,t} + \frac{1}{n}\nu^2_{t-h,t}\mathbb{E}_{X_t}\left[\mathrm{Tr}\left(\mathrm{Var}\left(X_0|X_t\right)\right)\right],$$

Reverse Diffusion: Maximum Likelihood



Converted Voice $\hat{X}_0$

Decoder = SDE-ML Solver

$$d\hat{X}_t = \left(\frac{1}{2}(\bar{X} - \hat{X}_t) - s_\theta(\hat{X}_t, \bar{X}, g_t(Y), t)\right)\beta_t dt + \sqrt{\beta_t}\, d\overleftarrow{W}_t$$

Starting from $\hat{X}_1 \sim \mathcal{N}(\bar{X}, \mathrm{I})$

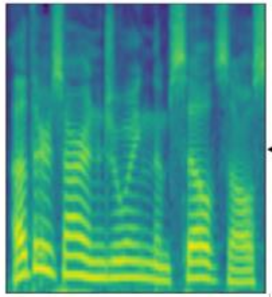
$$\mu_{s,t} = \gamma_{s,t}\frac{1-\gamma_{0,s}^2}{1-\gamma_{0,t}^2}, \quad \nu_{s,t} = \gamma_{0,s}\frac{1-\gamma_{s,t}^2}{1-\gamma_{0,t}^2}, \quad \sigma_{s,t}^2 = \frac{(1-\gamma_{0,s}^2)(1-\gamma_{s,t}^2)}{1-\gamma_{0,t}^2},$$

$$\kappa_{t,h}^* = \frac{\nu_{t-h,t}(1-\gamma_{0,t}^2)}{\gamma_{0,t}\beta_t h} - 1, \quad \omega_{t,h}^* = \frac{\mu_{t-h,t}-1}{\beta_t h} + \frac{1+\kappa_{t,h}^*}{1-\gamma_{0,t}^2} - \frac{1}{2},$$

$$(\sigma_{t,h}^*)^2 = \sigma_{t-h,t}^2 \cancel{+ \frac{1}{n}\nu_{t-h,t}^2 \mathbb{E}_{X_t}\left[\mathrm{Tr}\left(\mathrm{Var}\left(X_0|X_t\right)\right)\right]},$$

Without in source code?

Reverse Diffusion: Maximum Likelihood



Converted Voice $\hat{X}_0$

Decoder = SDE-ML Solver

$$d\hat{X}_t = \left(\frac{1}{2}(\bar{X} - \hat{X}_t) - s_\theta(\hat{X}_t, \bar{X}, g_t(Y), t)\right)\beta_t dt + \sqrt{\beta_t} d\overleftarrow{W}_t$$

Starting from $\hat{X}_1 \sim \mathcal{N}(\bar{X}, \mathrm{I})$

$\hat{X}_1 \sim \mathcal{N}(\bar{X}, \mathrm{I})$

for $i = 0$ **to** $N - 1$ **do**

$\quad t \leftarrow 1 - i \times h$

$\quad X'_{t-h} \leftarrow \hat{X}_t + \beta_t h\left(\left(\frac{1}{2} + \hat{\omega}_{t,h}\right)(\hat{X}_t - \bar{X}) + (1 + \hat{\kappa}_{t,h}) s_\theta(X_t, \bar{X}, g_t(Y), t)\right)$

$\quad \nabla \log p_{t|0}(X_t|X_0) \sim \mathcal{N}(0, I)$

$\quad \hat{X}_{t-h} \leftarrow X'_{t-h} + \hat{\sigma}_{t,h} \nabla \log p_{t|0}(X_t|X_0)$

return $\hat{X}_0$

Speaker Conditional Analysis

Input types for speaker conditioning $g_t(Y)$ compared in terms of speaker similarity.

	Diff-LibriTTS			*Diff-VCTK*		
	d-only	*wodyn*	*whole*	*d-only*	*wodyn*	*whole*
Most similar	27.0%	**38.0%**	34.1%	27.2%	**46.7%**	23.6%
Least similar	**28.9%**	29.3%	38.5%	25.3%	**23.9%**	48.6%

- d-only: Y = target Mel-spectrogram Y_0
- wodyn: Y = **Noisy** target Mel-spectrogram Y_t
- whole: Y = $\{\mathrm{Y}_\mathrm{t}, \mathrm{Y}_{0.5/15}, \mathrm{Y}_{1.5/15}, \ldots \mathrm{Y}_{14.5/15}\}$, channel = 16

Any-to-Any Voice Conversion

	VCTK test (9 speakers, 54 pairs)		Whole test (25 speakers, 350 pairs)	
	Naturalness	Similarity	Naturalness	Similarity
AGAIN-VC	1.98 ± 0.05	1.97 ± 0.08	1.87 ± 0.03	1.75 ± 0.04
FragmentVC	2.20 ± 0.06	2.45 ± 0.09	1.91 ± 0.03	1.93 ± 0.04
VQMIVC	2.89 ± 0.06	2.60 ± 0.10	2.48 ± 0.04	1.95 ± 0.04
Diff-VCTK-ML-6	$\mathbf{3.73 \pm 0.06}$	3.47 ± 0.09	3.39 ± 0.04	$\mathbf{2.69 \pm 0.05}$
Diff-VCTK-ML-30	$\mathbf{3.73 \pm 0.06}$	$\mathbf{3.57 \pm 0.09}$	$\mathbf{3.44 \pm 0.04}$	$\mathbf{2.71 \pm 0.05}$
Ground truth	4.55 ± 0.05	4.52 ± 0.07	4.55 ± 0.05	4.52 ± 0.07

Conv Auto Encoder
Attention-based
Vector Quantization

Train on VCTK, **100** speakers

All subjective human evaluation was carried out on Amazon Mechanical Turk.

Any-to-Any Voice Conversion

	VCTK test (9 speakers, 54 pairs)		Whole test (25 speakers, 350 pairs)		
	Naturalness	Similarity	Naturalness	Similarity	
AGAIN-VC	1.98 ± 0.05	1.97 ± 0.08	1.87 ± 0.03	1.75 ± 0.04	Conv Auto Encoder
FragmentVC	2.20 ± 0.06	2.45 ± 0.09	1.91 ± 0.03	1.93 ± 0.04	Attention-based
VQMIVC	2.89 ± 0.06	2.60 ± 0.10	2.48 ± 0.04	1.95 ± 0.04	Vector Quantization
Diff-VCTK-ML-6	$\mathbf{3.73 \pm 0.06}$	3.47 ± 0.09	3.39 ± 0.04	$\mathbf{2.69 \pm 0.05}$	
Diff-VCTK-ML-30	$\mathbf{3.73 \pm 0.06}$	$\mathbf{3.57 \pm 0.09}$	$\mathbf{3.44 \pm 0.04}$	$\mathbf{2.71 \pm 0.05}$	
Ground truth	4.55 ± 0.05	4.52 ± 0.07	4.55 ± 0.05	4.52 ± 0.07	

Real-Time Factor on GPU (unknow model)

- 6 step: around **0.1**
- 30 step: around 0.5

Any-to-Any Voice Conversion

Train on LibriTTS

approximately **1100** speakers.

	VCTK test (9 speakers, 54 pairs)		Whole test (25 speakers, 350 pairs)	
	Naturalness	Similarity	Naturalness	Similarity
Diff-LibriTTS-EM-6	1.68 ± 0.06	1.53 ± 0.07	1.57 ± 0.02	1.47 ± 0.03
Diff-LibriTTS-PF-6	3.11 ± 0.07	2.58 ± 0.11	2.99 ± 0.03	2.50 ± 0.04
Diff-LibriTTS-ML-6	3.84 ± 0.08	3.08 ± 0.11	3.80 ± 0.03	3.27 ± 0.05
Diff-LibriTTS-ML-30	$\mathbf{3.96 \pm 0.08}$	3.23 ± 0.11	$\mathbf{4.02 \pm 0.03}$	$\mathbf{3.39 \pm 0.05}$
BNE-PPG-VC	$\mathbf{3.95 \pm 0.08}$	$\mathbf{3.27 \pm 0.12}$	3.83 ± 0.03	3.03 ± 0.05

BEN-PPG-VC: combining a bottleneck feature extractor obtained from a phoneme recognizer with a seq2seq-based synthesis module.

Any-to-Any Voice Conversion

The proposed maximum likelihood (ML) sampling scheme over other sampling methods for a small number of inference steps.

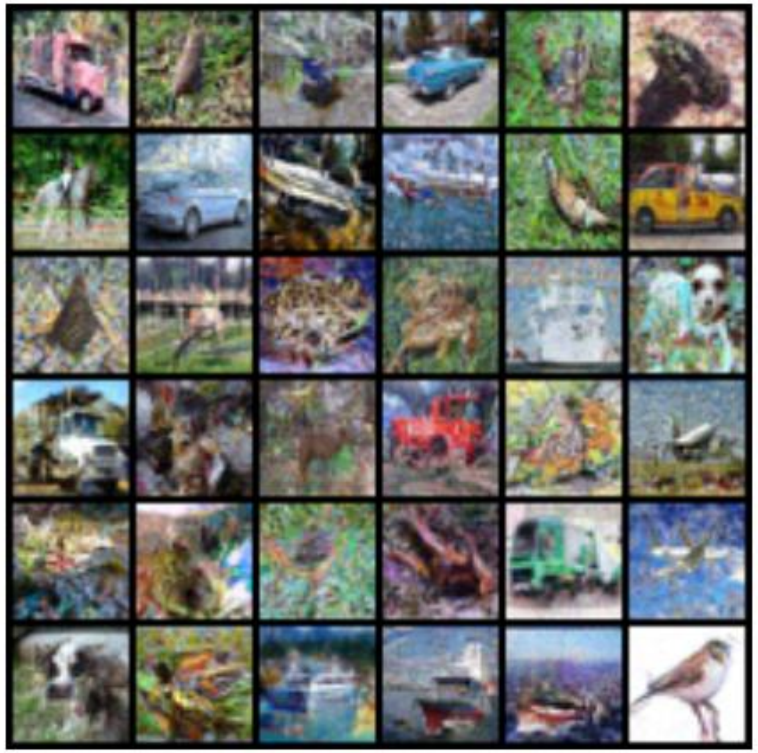
	VCTK test (9 speakers, 54 pairs)		Whole test (25 speakers, 350 pairs)	
	Naturalness	Similarity	Naturalness	Similarity
Diff-LibriTTS-EM-6	1.68 ± 0.06	1.53 ± 0.07	1.57 ± 0.02	1.47 ± 0.03
Diff-LibriTTS-PF-6	3.11 ± 0.07	2.58 ± 0.11	2.99 ± 0.03	2.50 ± 0.04
Diff-LibriTTS-ML-6	3.84 ± 0.08	3.08 ± 0.11	3.80 ± 0.03	3.27 ± 0.05
Diff-LibriTTS-ML-30	$\mathbf{3.96 \pm 0.08}$	3.23 ± 0.11	$\mathbf{4.02 \pm 0.03}$	$\mathbf{3.39 \pm 0.05}$
BNE-PPG-VC	$\mathbf{3.95 \pm 0.08}$	$\mathbf{3.27 \pm 0.12}$	3.83 ± 0.03	3.03 ± 0.05

BEN-PPG-VC: combining a bottleneck feature extractor obtained from a phoneme recognizer with a seq2seq-based synthesis module.

Maximum Likelihood Sampling

Euler-Maruyama

Probability Flow



Maximum Likelihood

CIFAR-10 images randomly sampled from VP DPM by running **10 reverse diffusion steps**.

Conclusion

- Average Voice Encoder
  a new disentanglement method.
- Diffusion-based Decoder
  achieve good results both in terms of similarity and naturalness.
- Novel Sampling Scheme
  High-quality results in just a few steps.