Guided TTS 2

A Diffusion Model for High-quality **Adaptive** Text-to-Speech with **Untranscribed** Data

arXiv preprint, 2022

Sungwon Kim, Heeseung Kim, Sungroh Yoon
Seoul National University

Text-to-Speech

In recent years, **single-speaker TTS methods** have been able to synthesize high-quality speech.

Issue

- The direct usage of **Untranscribed Data** remains a challenge.
- Requires sufficient amounts of **Large-Scale Data** for the **Target Speaker**.

Adaptive TTS

Adaptive TTS models aim to generate high-quality speech for the target speaker given a **limited amount of reference data**.

- by Zero-shot or Fine-tune

Issue

- When the amount of the reference speech is only around 10 seconds, the sample quality and speaker similarity are poor.

Diffusion Models

Recent diffusion models show impressive results in **class-conditional** and **text-conditional** image generation tasks via diffusion guidance methods.

- e.g., DALL· E 2, Midjourney, Stable Diffusion

Advantage

- High-Quality.
- Stable.
- Easy to Guided.

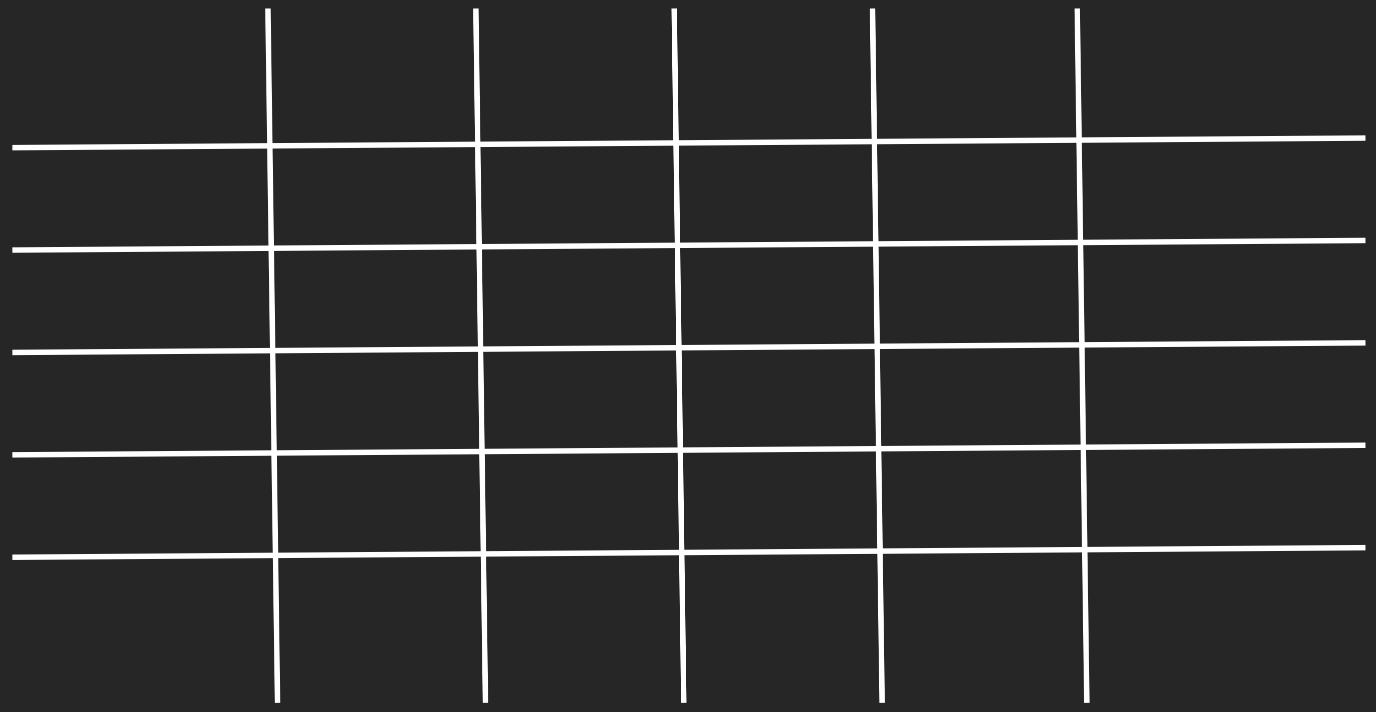
Guided TTS

Taking advantage of the easy-to-guide nature of the Diffusion models, Guided TTS successfully constructed a single-speaker TTS that effectively utilizes untranscribed data.

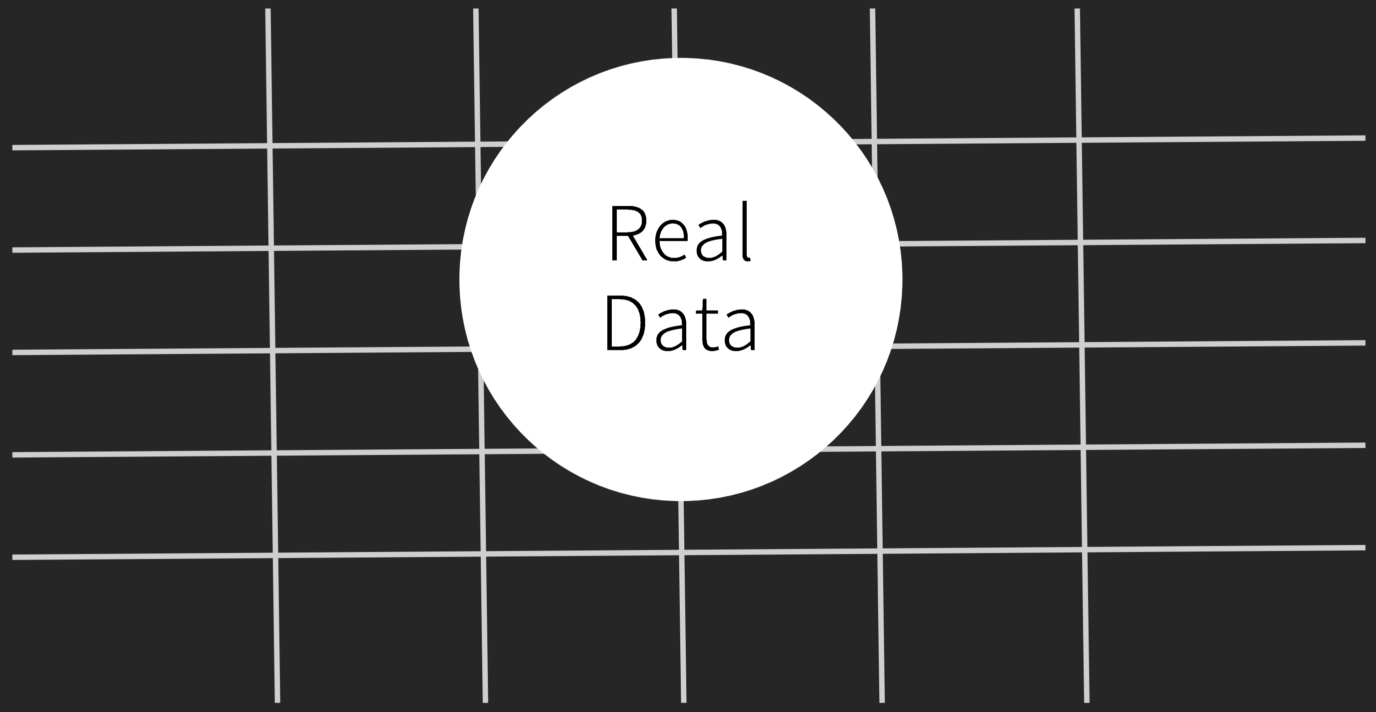
Guided TTS 2

- Generate high-quality and high-similar speech with only 10 seconds of reference material.

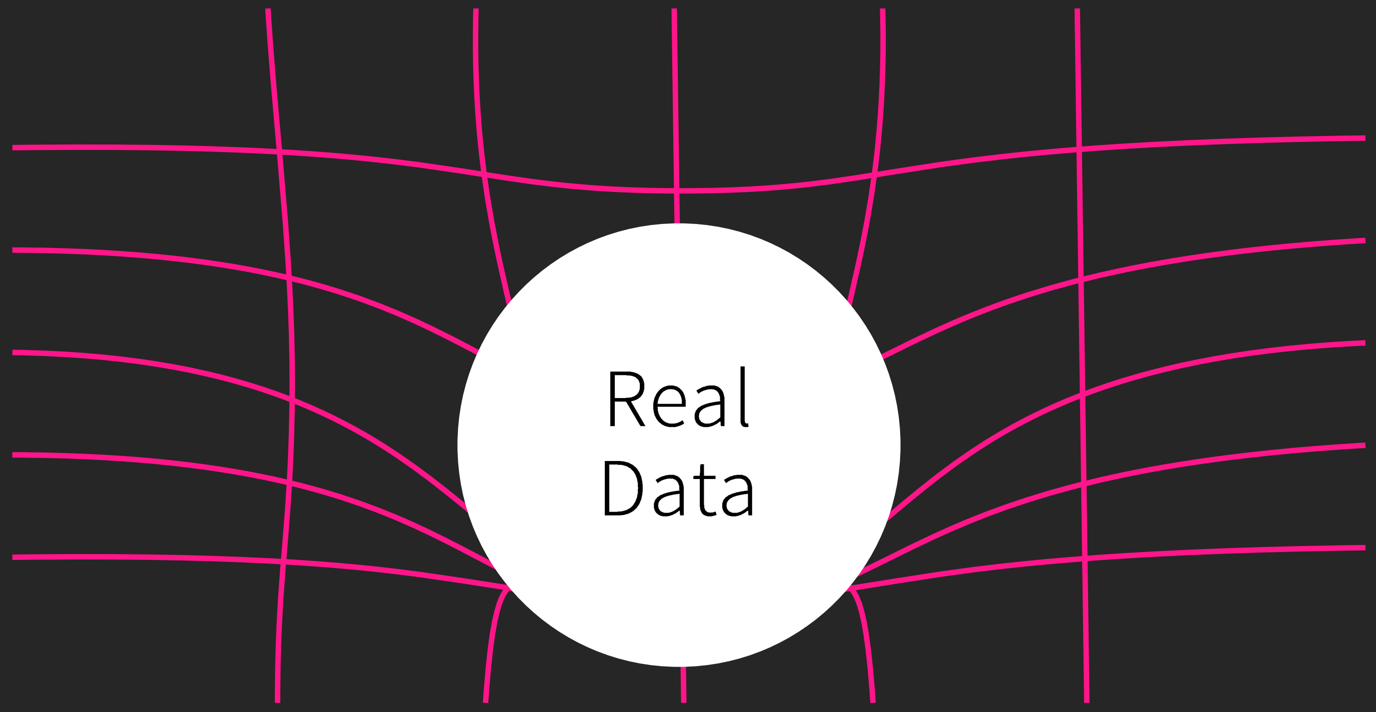
Diffusion Models



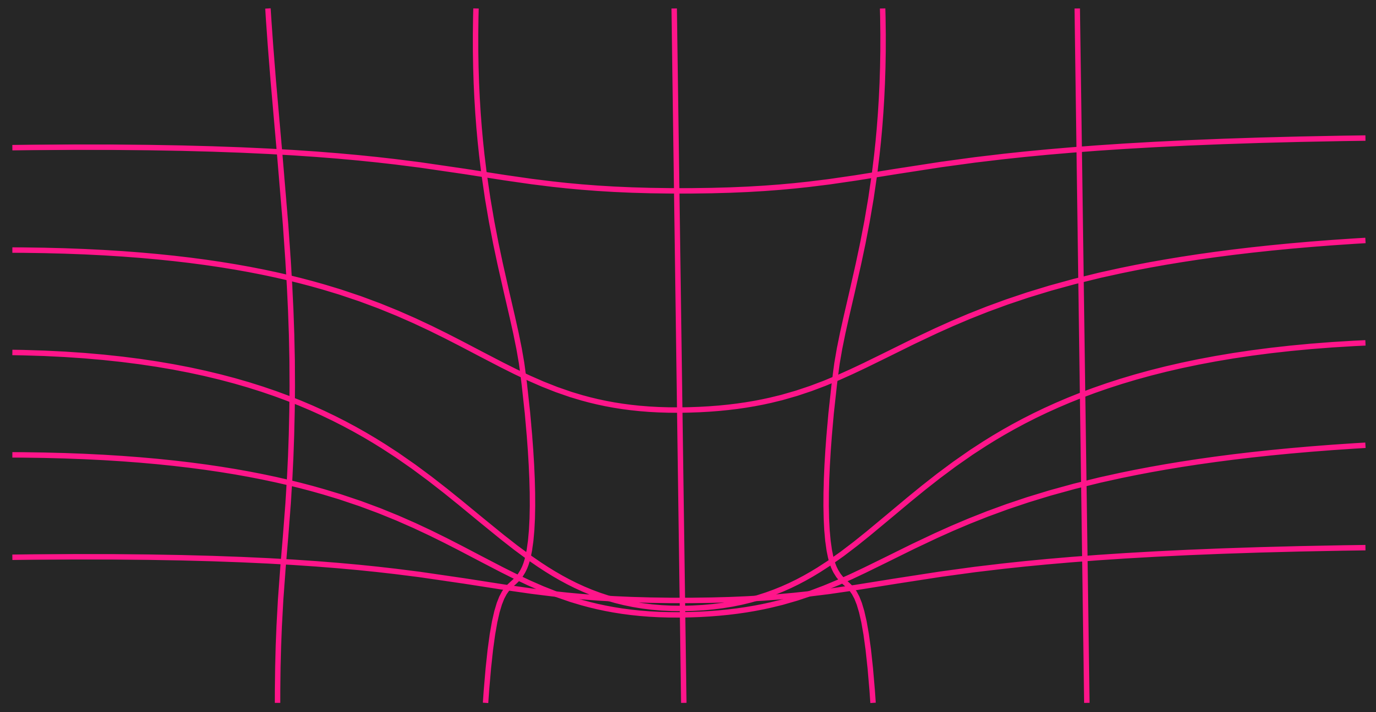
Diffusion Models



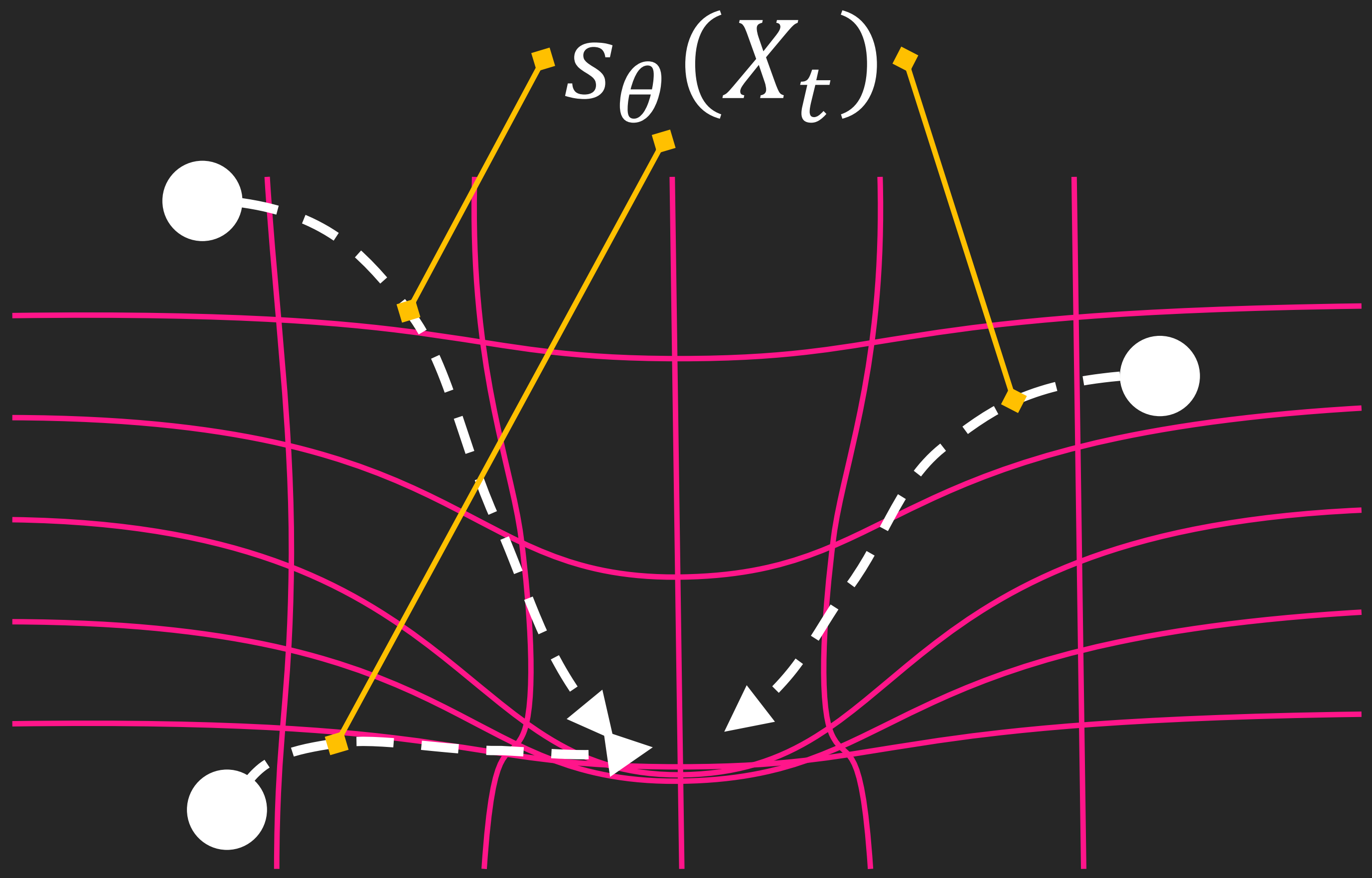
Diffusion Models



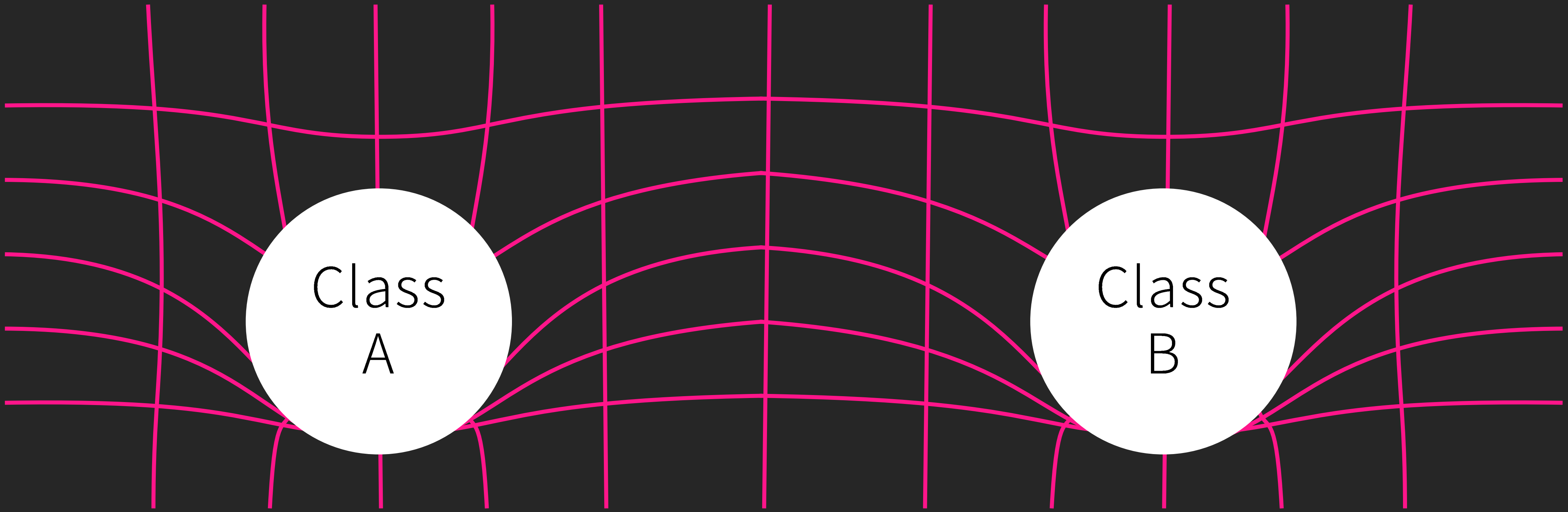
Diffusion Models



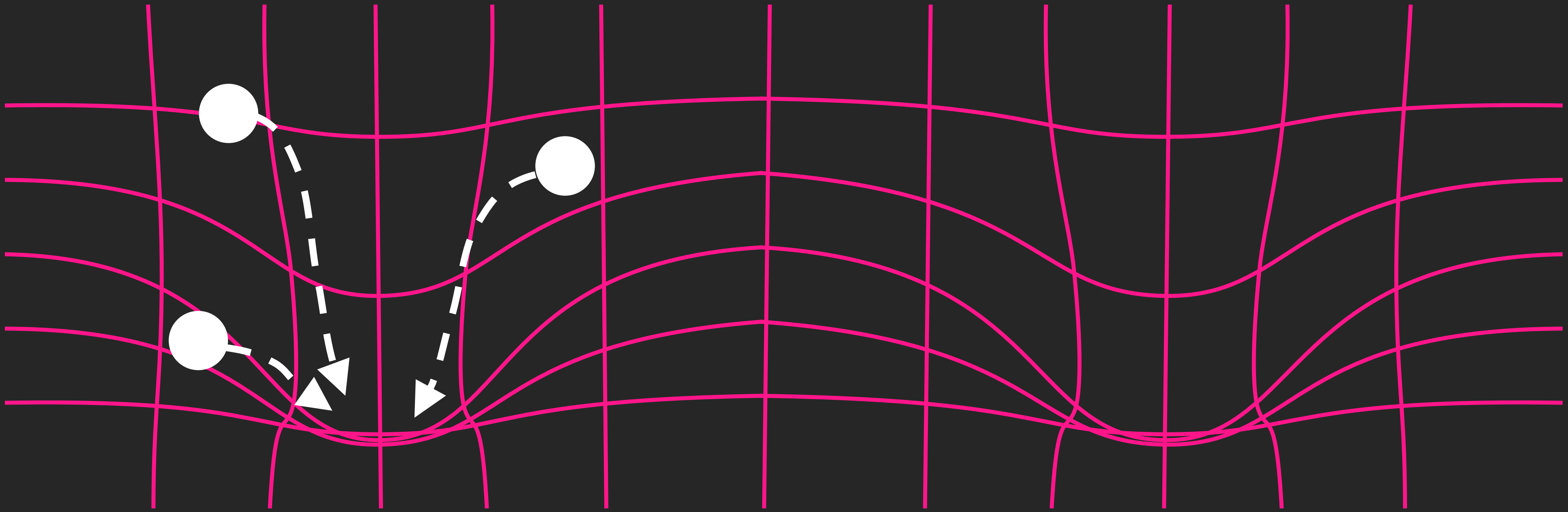
Diffusion Models



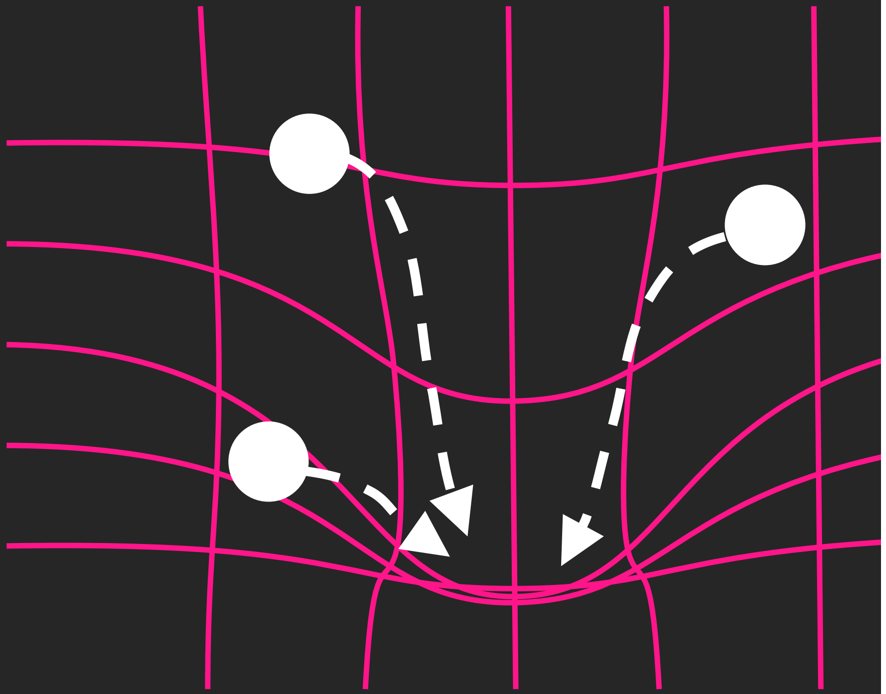
How to Guide?

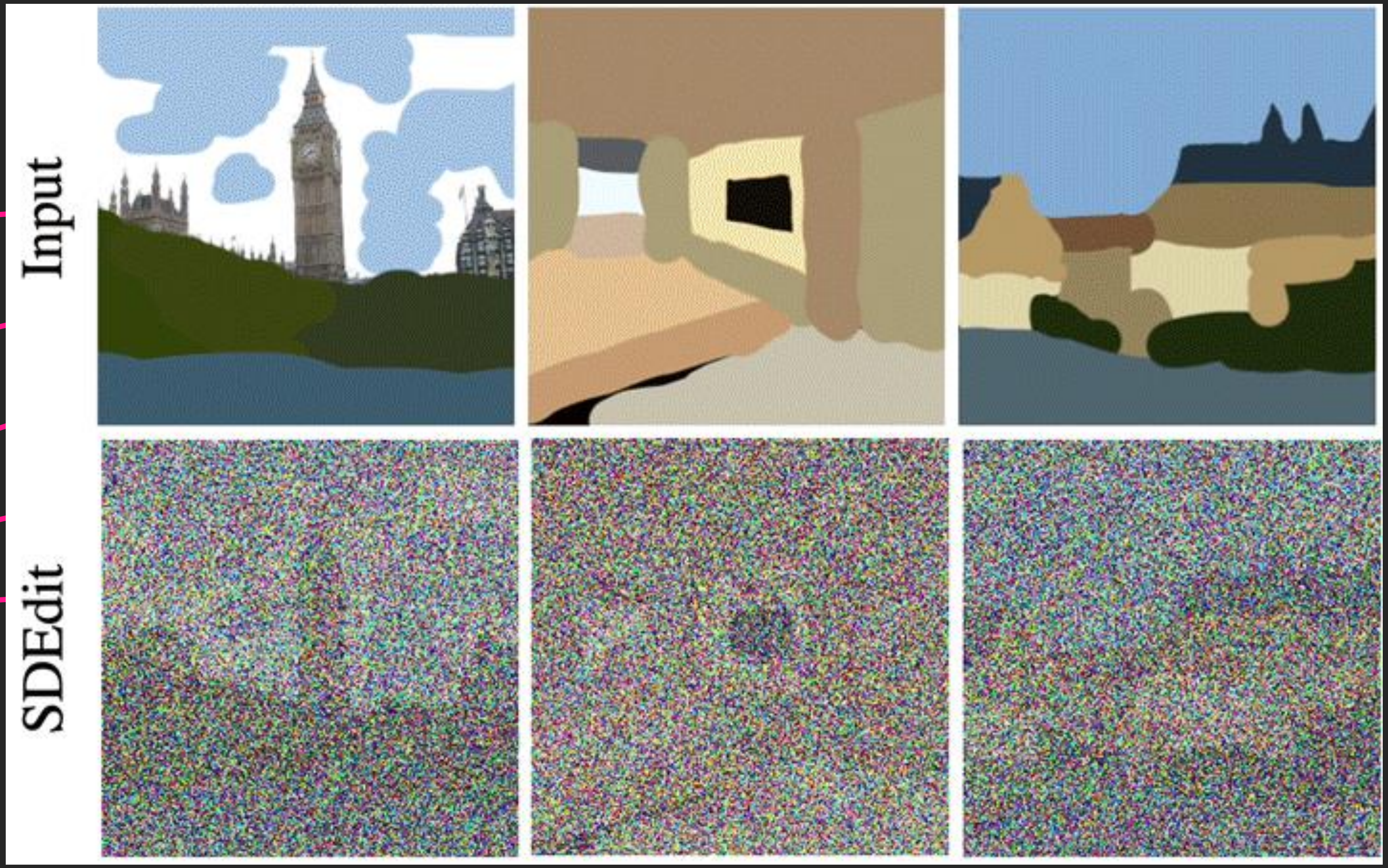


Hijack the Initial State

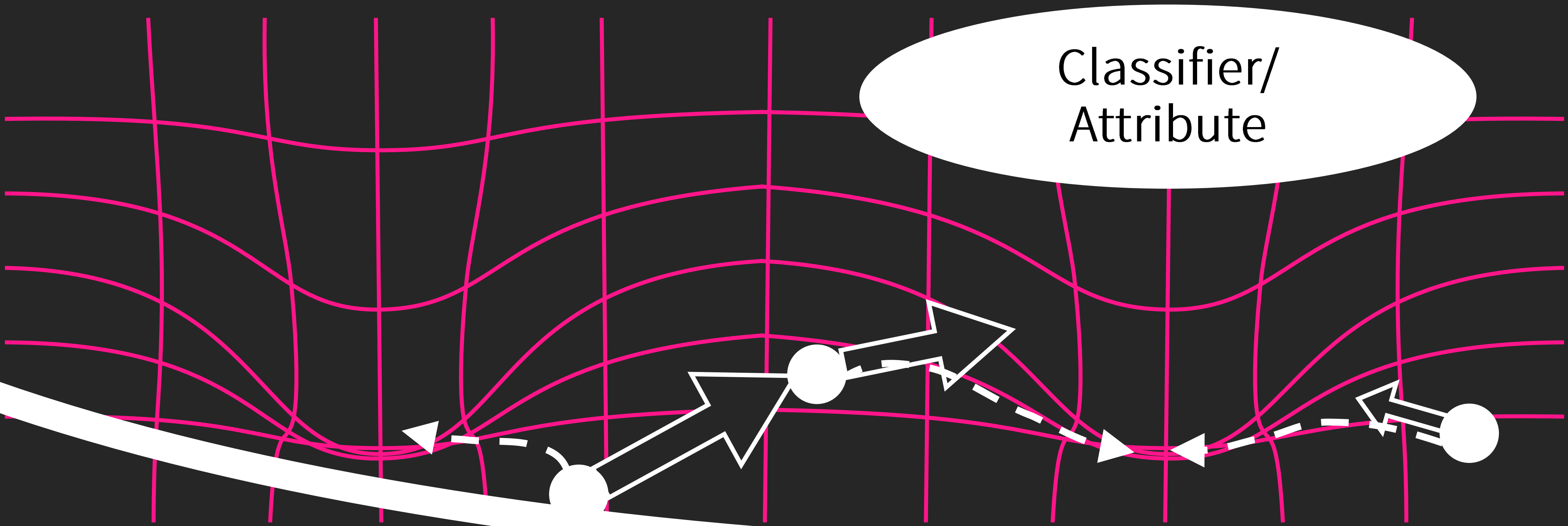


Hijack the Initial State

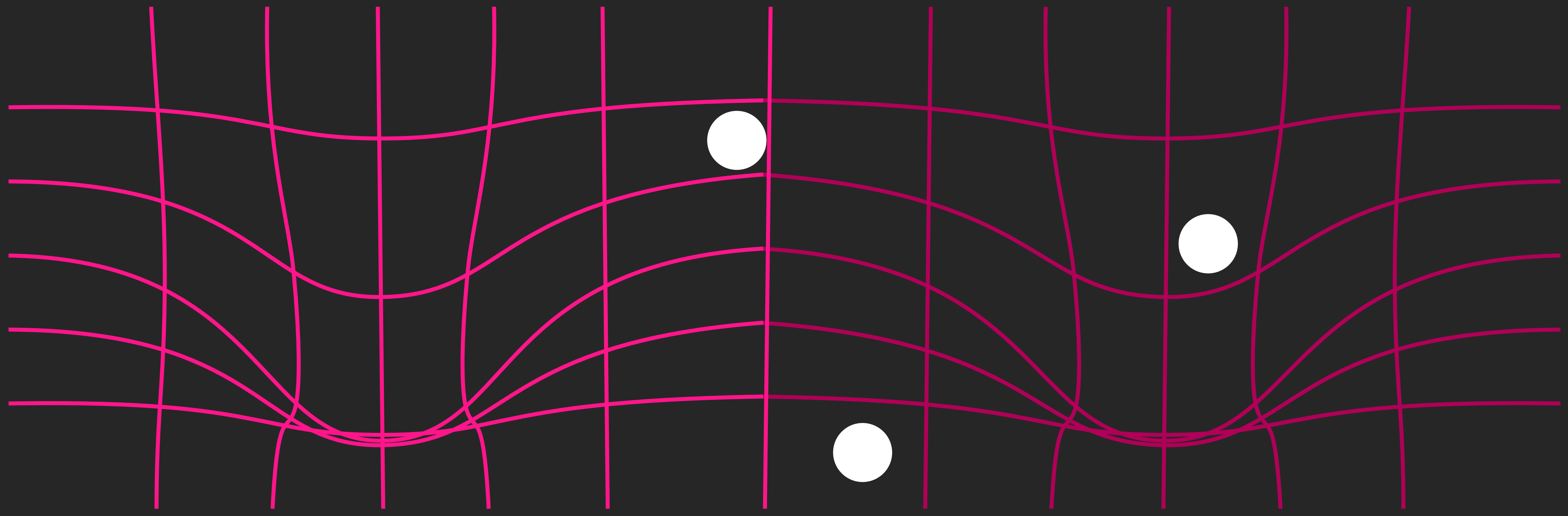




Classifier Guidance

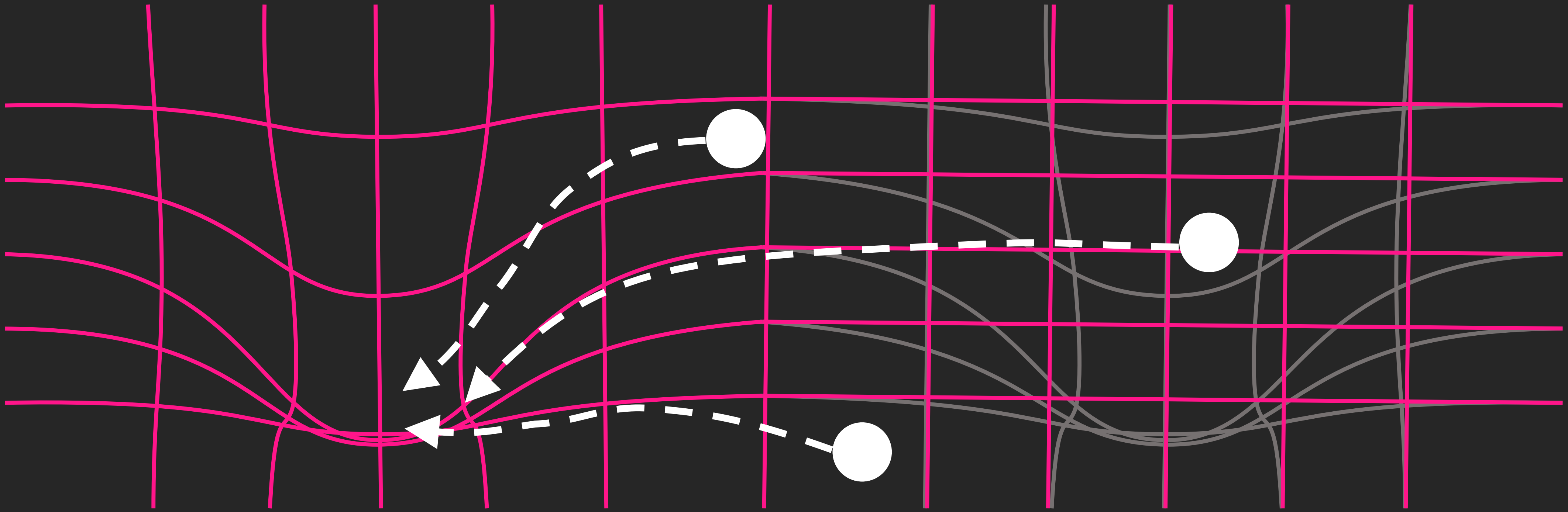


Classifier-free Guidance



$$s_\theta(X_t)$$

Classifier-free Guidance



$$s_\theta(X_t|c = A)$$

Guided TTS 2

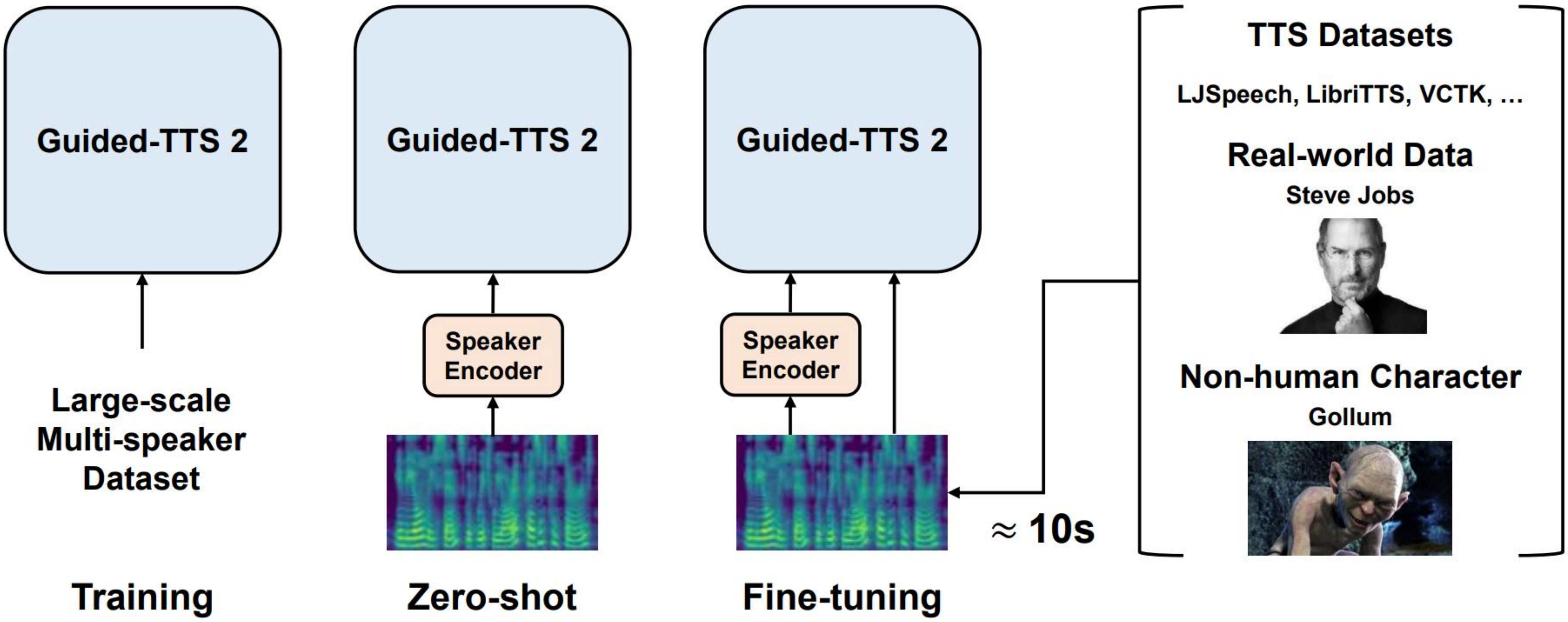


Figure 1: The overview of Guided-TTS 2.

Guided TTS 2

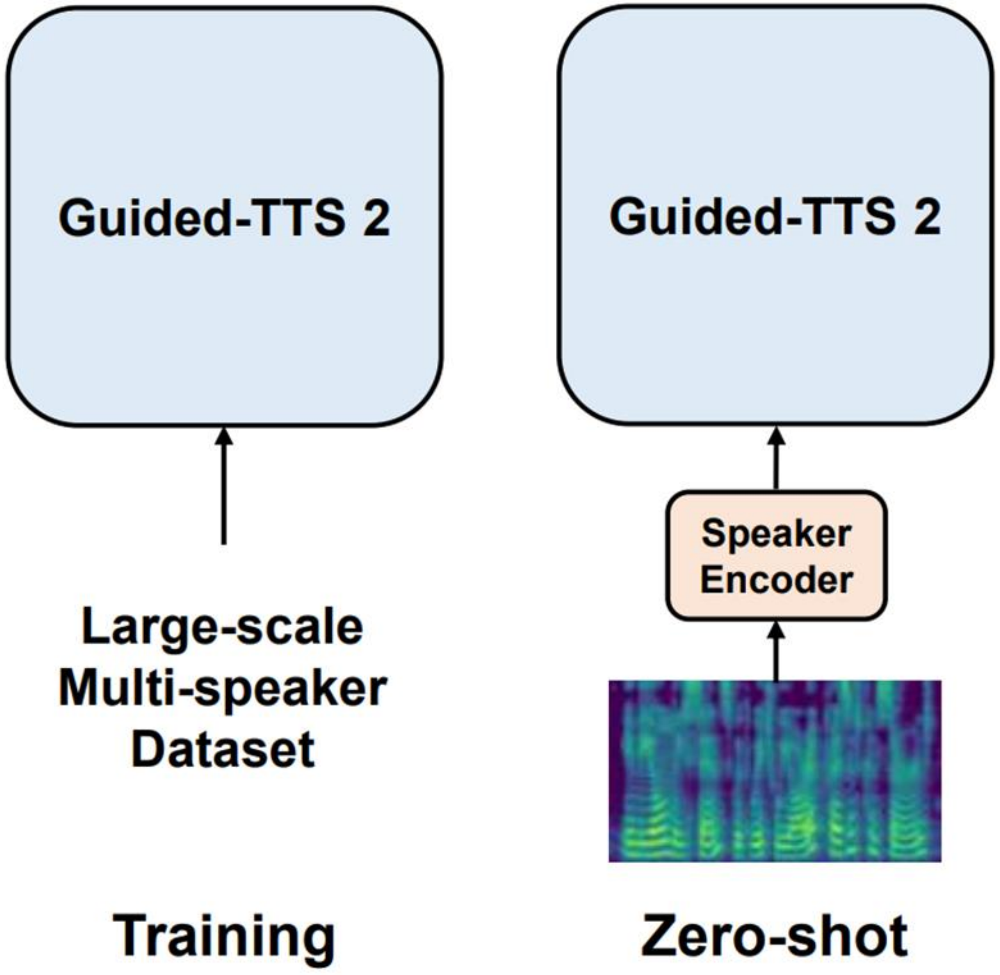


Figure 1: The

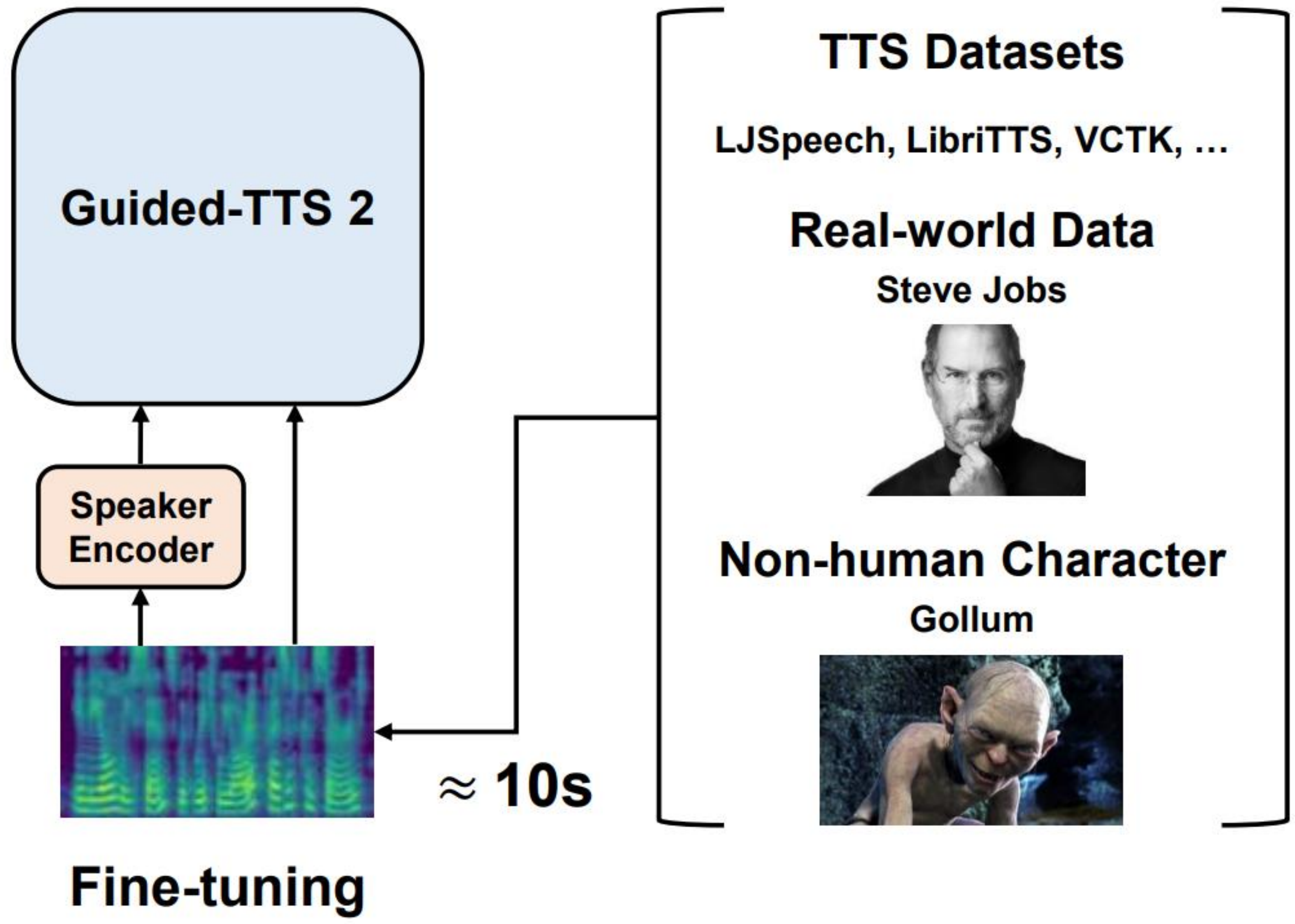
Zero-shot

- Speaker-conditional DDPM
- Guided by speaker embedding.
- Train a Speaker Encoder with GE2E loss.

Guided TTS 2

Fine-tune

- Reference data: 10s
- Learning rate: 2e-5.
  lower than the pre-training learning rate 1e-4.
- Iterations: 500.
- Training Time: 40s on a NVIDIA RTX 8000 GPU.



overview of Guided-TTS 2.

Guided TTS 2

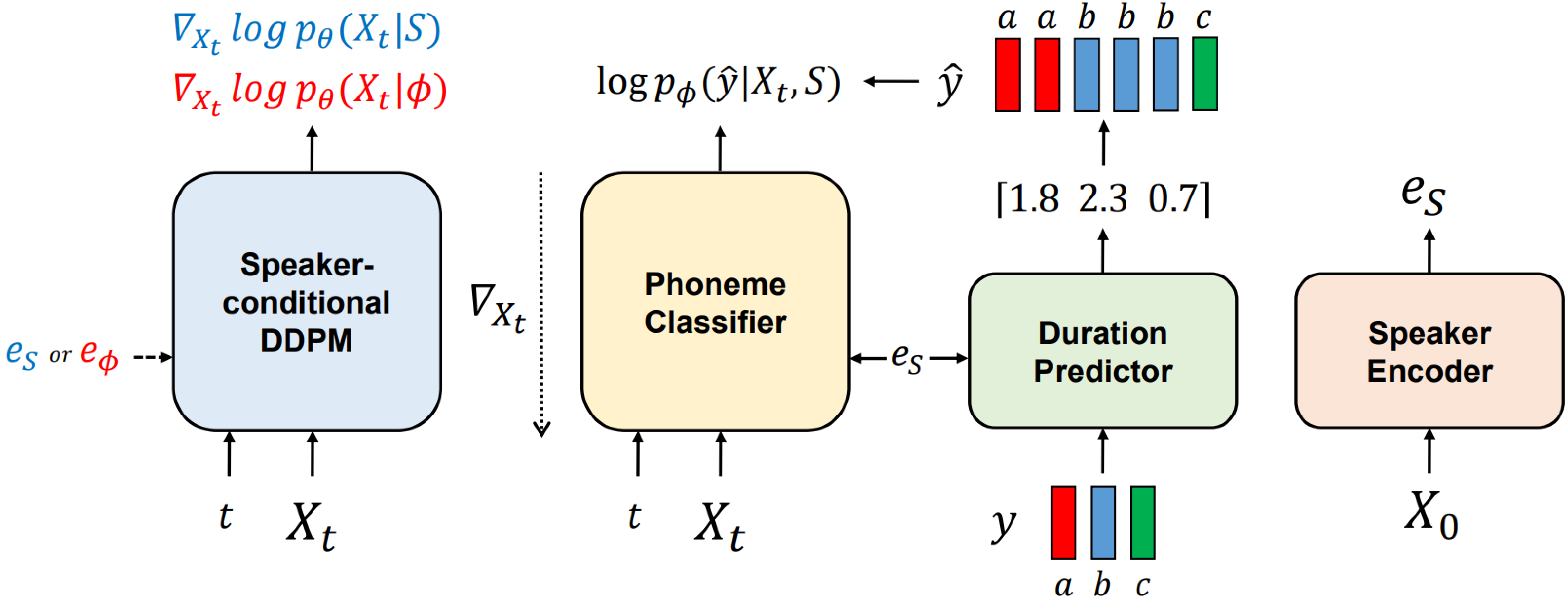


Figure 2: The overall components of Guided-TTS 2.

Guided TTS 2: Training

Training Algorithm

repeat until converged:

$$X_0 \sim \mathbf{X}$$

$$t \sim \mathcal{U}\left(\frac{1}{N}, 1\right)$$

$$\epsilon \sim \mathcal{N}(0, I)$$

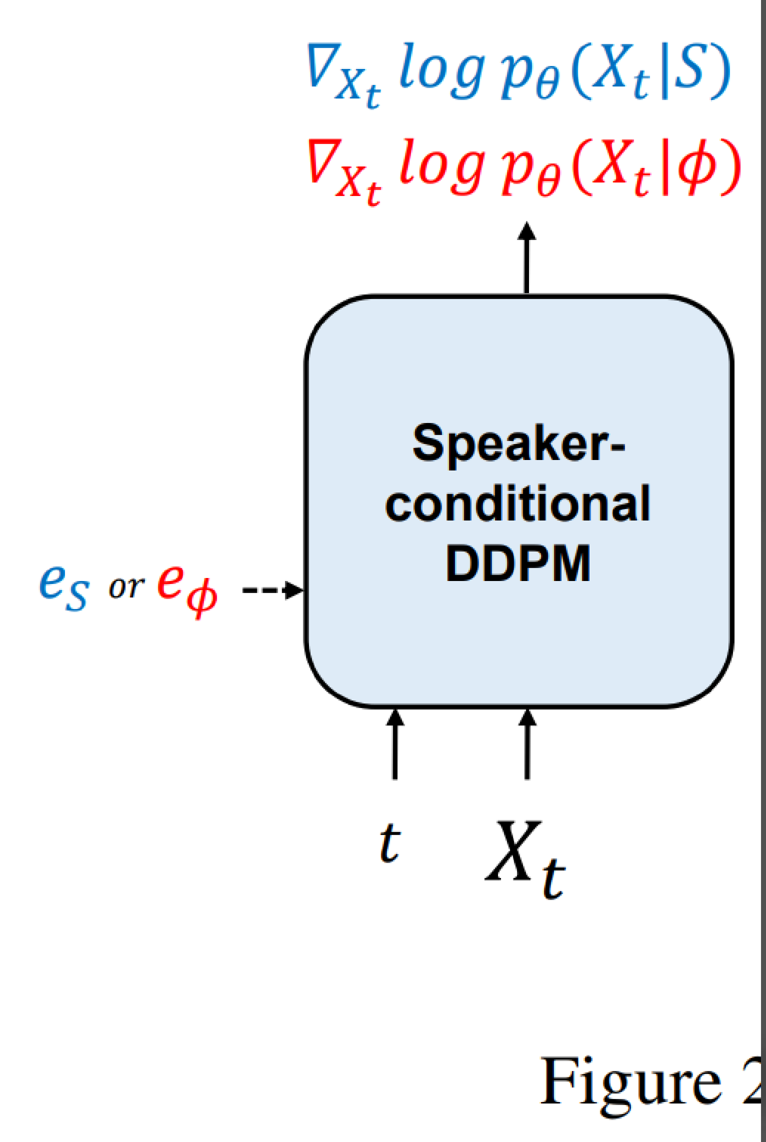
$$X_t \leftarrow \sqrt{I - \lambda(t)}X_0 + \sqrt{\lambda(t)}\,\epsilon$$

$$\nabla_\theta \mathbb{E}_{t,X_0,\epsilon_t}\left[\left\| s_\theta(X_t | S = e_S \; or \; e_\emptyset) + \frac{\epsilon}{\sqrt{\lambda(t)}} \right\|_2^2\right]$$

Hyperparameter

- $N = 50$
- $\lambda(t) = I - e^{-\int_0^t \beta_s \, ds}$
- $\beta_t = \beta_0 + (\beta_T - \beta_0) \cdot t$
  - $\beta_0 = 0.05, \beta_T = 20$
- Dropout rate of condition: 50%

Speaker-conditional Guidance

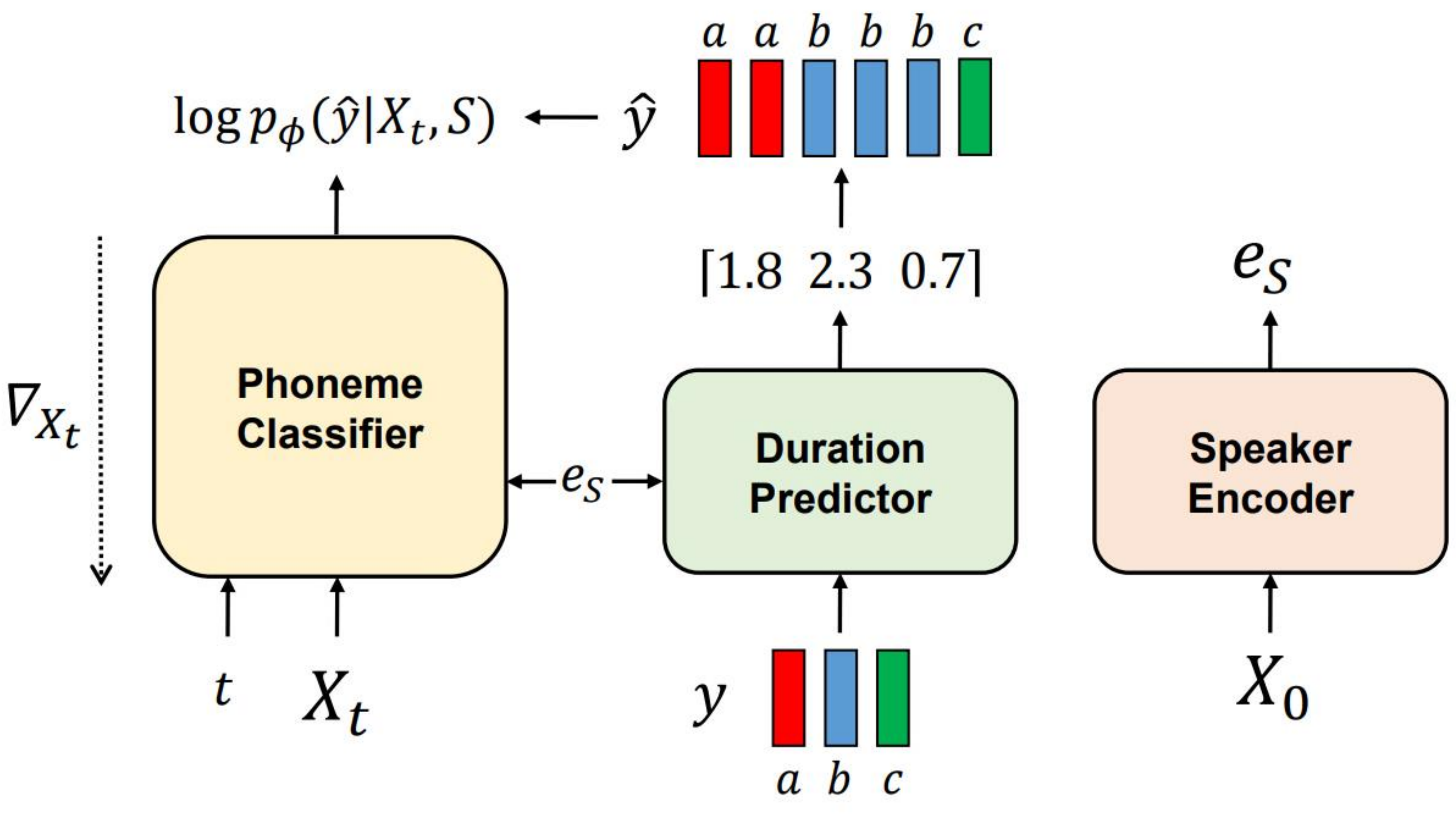


$$\hat{s}_\theta(X_t|\hat{S}) = s_\theta(X_t|\hat{S}) + \gamma_S \cdot \left(s_\theta(X_t|\hat{S}) - s_\theta(X_t|\emptyset)\right)$$

Proposed by Classifier-Free Diffusion Guidance

$$\hat{s}_\theta(X_t|\hat{y},\hat{S}) = \hat{s}_\theta(X_t|\hat{S}) + \gamma_T \cdot \nabla_{X_t} log\, p_\varphi(\hat{y}|X_t,\hat{S})$$

Framewise phoneme Classifier Guidance



2: The overall components of Guided-TTS 2.

Framewise Phoneme Classifier Guidance

This is about **70 times** larger than the norm of the classifier gradient near X_0

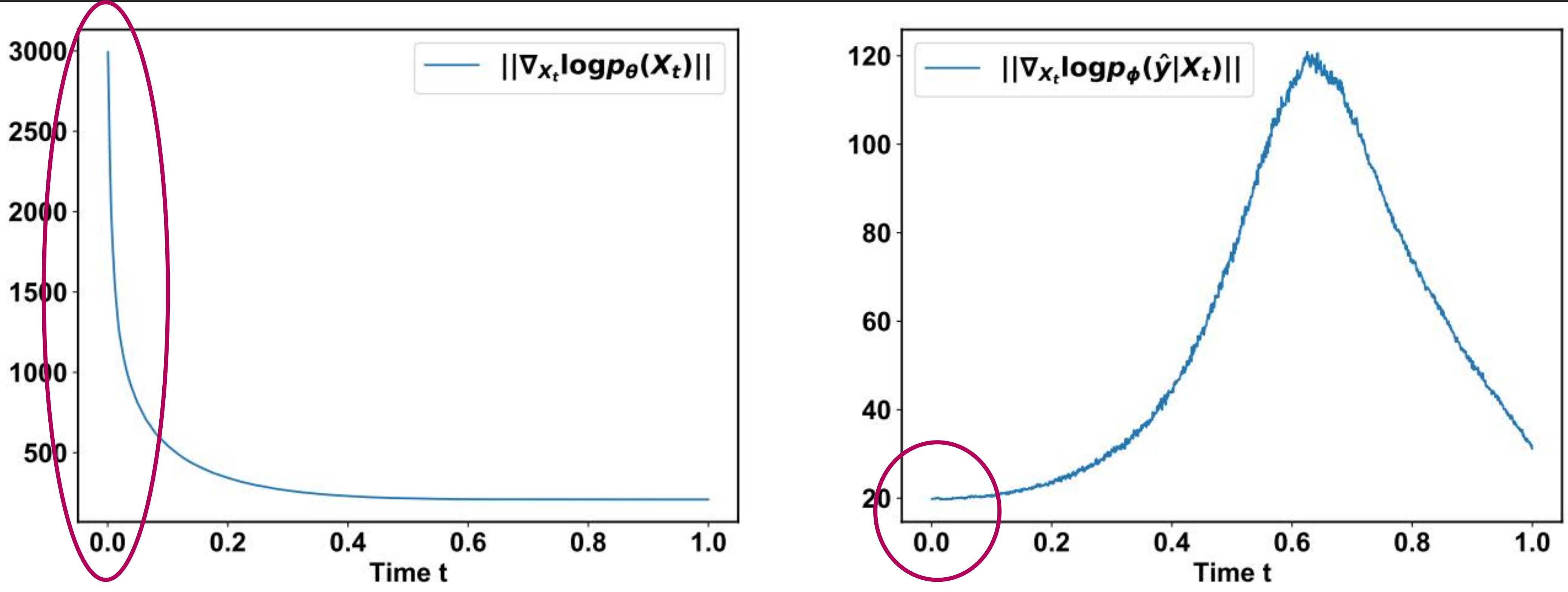
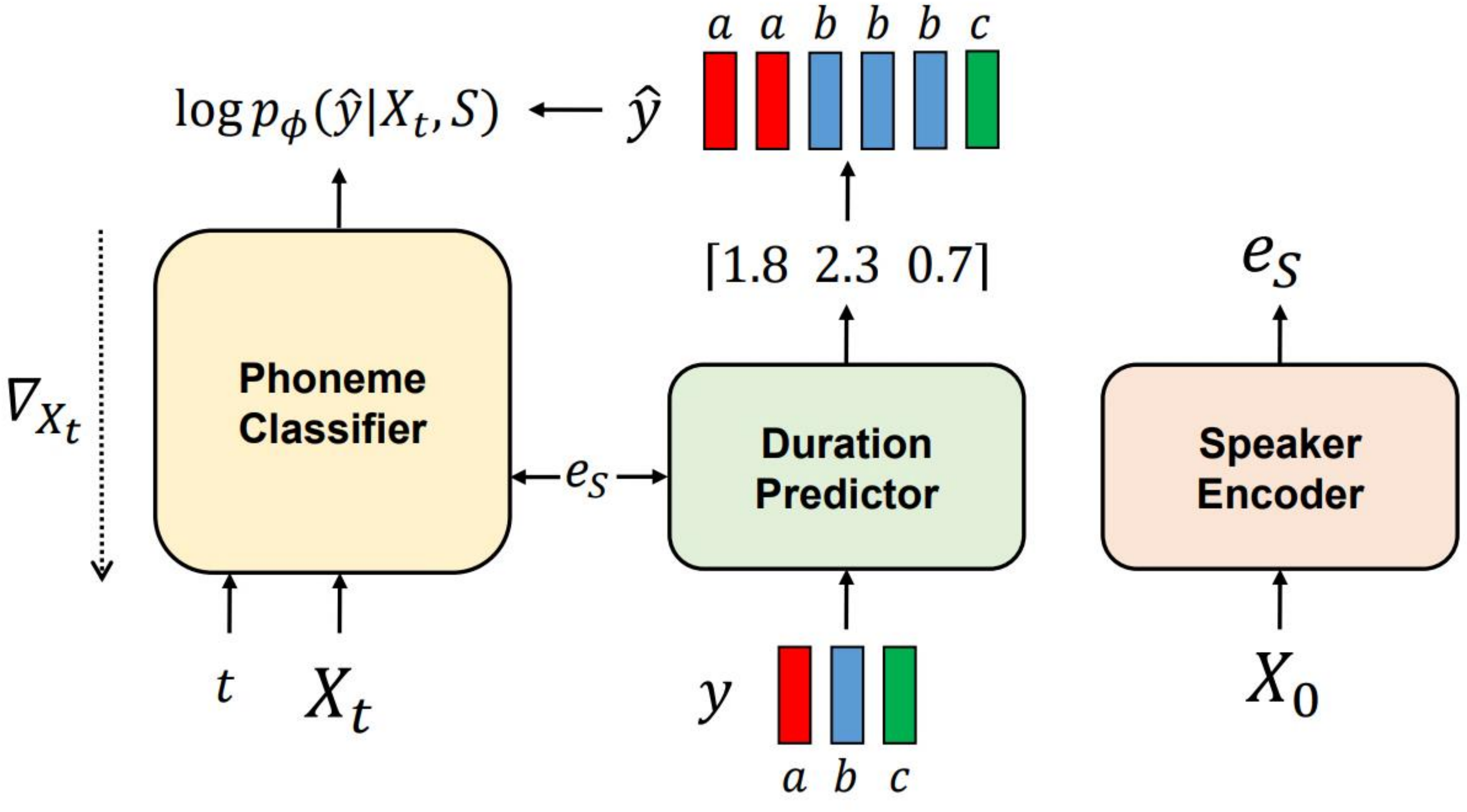


Figure 4: The norm of the unconditional score and the classifier gradient for each timestep t. (Left) The norm of the unconditional score (Right) The norm of the classifier gradient.

$$\hat{s}_\theta(X_t|\hat{y},\hat{S}) = \hat{s}_\theta(X_t|\hat{S}) + \gamma_T \cdot \frac{\|\hat{s}_\theta(X_t|\hat{S})\|}{\|\nabla_{X_t} log\, p_\varphi(\hat{y}|X_t,\hat{S})\|} \cdot \nabla_{X_t} log\, p_\varphi(\hat{y}|X_t,\hat{S})$$

Norm-based Guidance

The amount of scaling is proportional to the norm of the score.



2: The overall components of Guided-TTS 2.

Guided TTS 2: Sampling

Sampling Algorithm

$\hat{y}$: framewise phoneme label, τ: temperature

$\hat{S}$: target speaker condition, θ: parameter of DDPM

$X_1 \sim \mathcal{N}(0, \tau^{-1} I)$

for t in $\left\{1, \ldots \frac{2}{N}, \frac{1}{N}\right\}$:

$z_t \sim \mathcal{N}(0, \tau^{-1} I)$

$$X_{t-\frac{1}{N}} \leftarrow X_t + \frac{\beta_t}{N}\left(\frac{1}{2} X_t + \hat{s}_\theta\left(X_t \middle| \hat{y}, \hat{S}\right)\right) + \sqrt{\frac{\beta_t}{N}} z_t$$

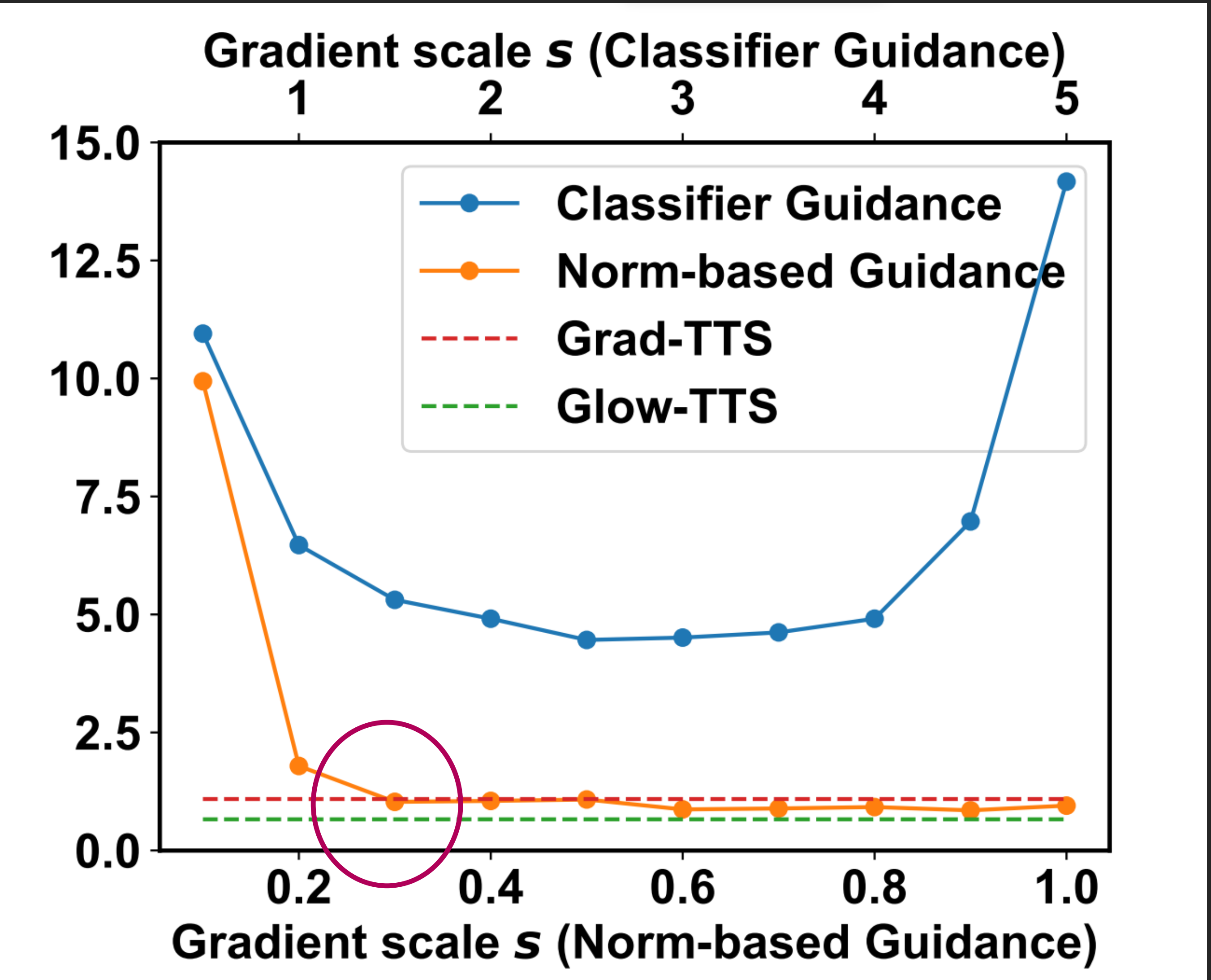
return X_0

Hyperparameter

- $\boldsymbol{\tau} = 1.5$

Experiments

CER of Norm-based Guidance (Guided-TTS 1)



- Classifier Guidance has more mispronunciations.
  - $\hat{s}_\theta(X_t|\hat{y},\hat{S}) = \hat{s}_\theta(X_t|\hat{S}) + \gamma_T \cdot \nabla_{X_t} \log p_\varphi(\hat{y}|X_t,\hat{S})$
- Norm-based Guidance has higher quality.
- $\gamma_T = 0.3$

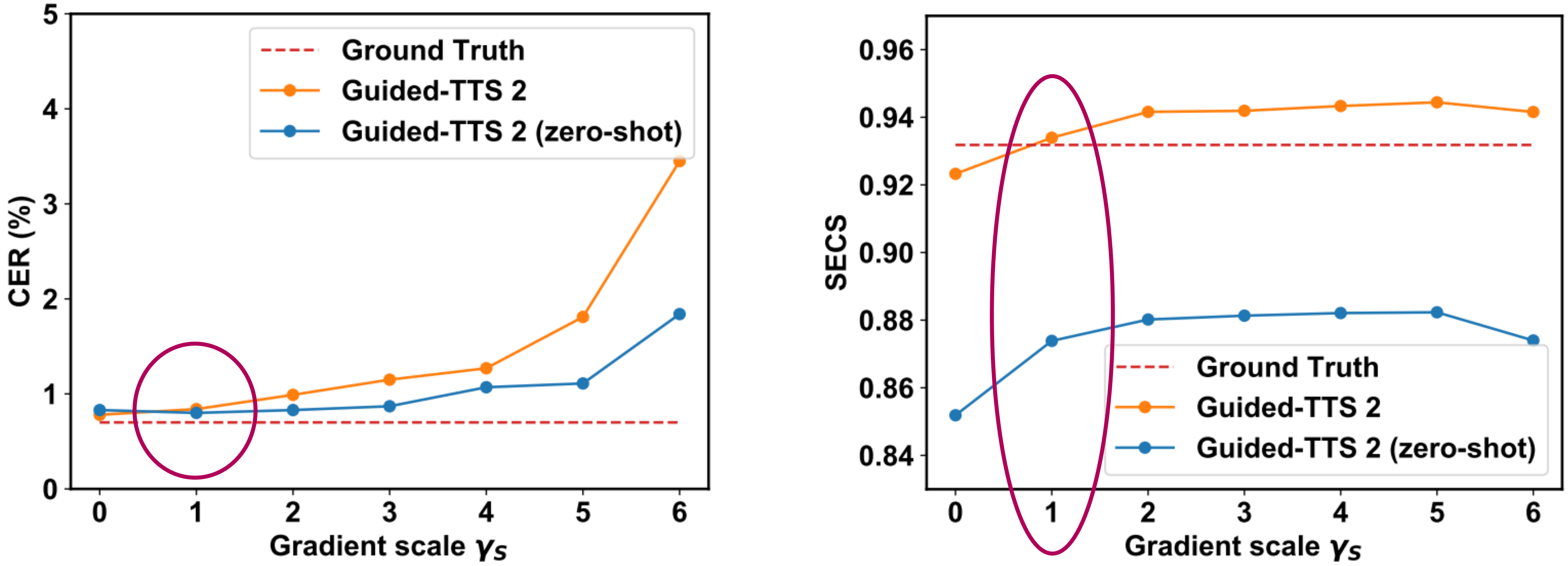


Figure 3: CER and SECS of Guided-TTS 2 according to speaker gradient scales. We refer to the fine-tuning setting of our model as Guided-TTS 2 and the zero-shot setting as Guided-TTS 2 (zero-shot).

$$\hat{s}_\theta(X_t|\hat{S}) = s_\theta(X_t|\hat{S}) + \gamma_S \cdot \left(s_\theta(X_t|\hat{S}) - s_\theta(X_t|\emptyset)\right)$$

- $\gamma_S = 1$

Experiments

Experiments

- Fine-tune has a higher **Similarity.**
- Zero-shot has a higher **Quality.**

Comparison with single-speaker TTS methods trained with LJSpeech.

Method	5-scale MOS	CER(%)	5-scale SMOS
Ground Truth	4.45 ± 0.05	0.64	3.85 ± 0.08
Mel + HiFi-GAN (Kong et al. (2020))	4.24 ± 0.08	0.86	3.80 ± 0.08
Grad-TTS (Popov et al. (2021))	4.22 ± 0.08	0.98	3.67 ± 0.09
Guided-TTS (Kim et al. (2021a))	4.17 ± 0.09	1.23	3.63 ± 0.09
Guided-TTS 2 (LT+LL fine-tune)	4.21 ± 0.09	1.12	3.69 ± 0.09
Guided-TTS 2 (LT+LL zero-shot)	4.23 ± 0.09	0.89	3.51 ± 0.08
Guided-TTS 2 (LT fine-tune)	4.22 ± 0.09	1.16	3.74 ± 0.09
Guided-TTS 2 (LT zero-shot)	4.16 ± 0.09	1.03	3.47 ± 0.09

Experiments

Comparison with adaptive TTS methods.

Dataset	Method	5-scale MOS	CER(%)	5-scale SMOS
LibriTTS Dataset	Ground Truth	4.52 ± 0.05	0.70	3.91 ± 0.07
	Mel + HiFi-GAN (Kong et al. (2020))	4.28 ± 0.08	0.75	3.86 ± 0.08
	Guided-TTS 2 (fine-tune)	4.20 ± 0.08	0.84	3.70 ± 0.09
	Guided-TTS 2 (zero-shot)	4.25 ± 0.09	0.80	3.51 ± 0.10
	YourTTS (Casanova et al. (2021))	4.02 ± 0.10	2.38	3.30 ± 0.10
	Meta-StyleSpeech (Min et al. (2021))	3.98 ± 0.11	1.52	3.42 ± 0.09
VCTK Dataset	Ground Truth	4.45 ± 0.05	2.40	3.71 ± 0.07
	Mel + HiFi-GAN (Kong et al. (2020))	4.21 ± 0.08	2.81	3.72 ± 0.07
	Guided-TTS 2 (fine-tune)	4.11 ± 0.09	1.49	3.57 ± 0.10
	Guided-TTS 2 (zero-shot)	4.23 ± 0.09	0.81	3.39 ± 0.09
	YourTTS (Casanova et al. (2021))	3.94 ± 0.10	2.36	3.19 ± 0.09
	Meta-StyleSpeech (Min et al. (2021))	3.65 ± 0.13	1.84	3.26 ± 0.10

Experiments: Fine-tune

Method	Length(seconds)	CER(%)	SECS
Guided-TTS 2	3	2.44	0.925
	5	1.67	0.930
	10	1.12	0.929
	30	0.98	0.932
	60	1.14	0.931

Method	Iterations	Optimizer	CER(%)	SECS
Guided-TTS 2	0	-	0.80	0.873
	50	Initialize	0.82	0.908
	200	Initialize	0.88	0.929
	500	Initialize	0.84	0.937
	2000	Initialize	1.49	0.945
	500	Load	1.39	0.925

Conclusion

This study proposes a novel Diffusion-based adaptive TTS method that generates speech quality comparable to the results of single-speaker TTS with only a few references.

- Adaptive TTS method.
- Hight-quality results.
- Only a few references are required.
- Effectively utilizes untranscribed data.

Demo: Steve Jobs

This audio was generated by a text-to-speech model for Steve Jobs. We use ten second untranscribed speech from Steve Jobs' Stanford Commencement Address.

- reference
- fine-tune
- zero-shot

Demo: Gollum

This audio was generated by a text-to-speech model for Gollum (魔戒 咕嚕), which can adapt to non-human characters using untranscribed data.

- reference
- fine-tune
- zero-shot

Appendix: Trainset

Speaker Encoder	**Speaker dependent phoneme classifier & duration predictor**	**Speaker conditional DDPM**
VoxCeleb2 • 6,112 speaker • Over 1M utterance	LibriSpeech • 2,484 speaker • Approximately 1,000 hours	LibriTTS • 2,456 speaker • 585 hours
		Libri-Light (unlab-600 + unlab-6k) • 2,231 speaker • Approximately 6,300 hours