KRAFTON NVIDIA

ICLR International Conference On Learning Representations

DiTTo-TTS

Diffusion **T**ransformers for Scalable **T**ext-**t**o-**S**peech without Domain-Specific Factors

ICLR 2025

Keon Lee[1], *Dong Won Kim*[1], *Jaehyeon Kim*[2*], *Seungjun Chung*[1], *Jaewoong Cho*[1]

[1]KRAFTON, [2]NVIDIA

[*]This work was done when Jaehyeon Kim was at KRAFTON.

ditto-tts.github.io

Contents

1. DiTTo-TTS: DiT to TTS
2. Related Work
3. Research Question
4. Method
5. Experiment

DiTTo-TTS: DiT to TTS

Latent Diffusion based TTS

DiTTo-TTS: DiT to TTS

Latent Diffusion based TTS w/o phoneme & duration

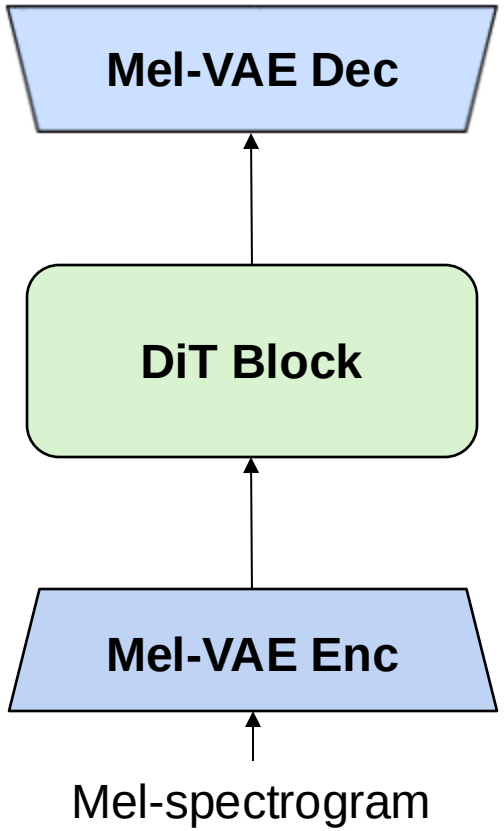
DiTTo-TTS: DiT to TTS

Latent Diffusion based TTS w/o phoneme & duration

TTS-Specific Factors !

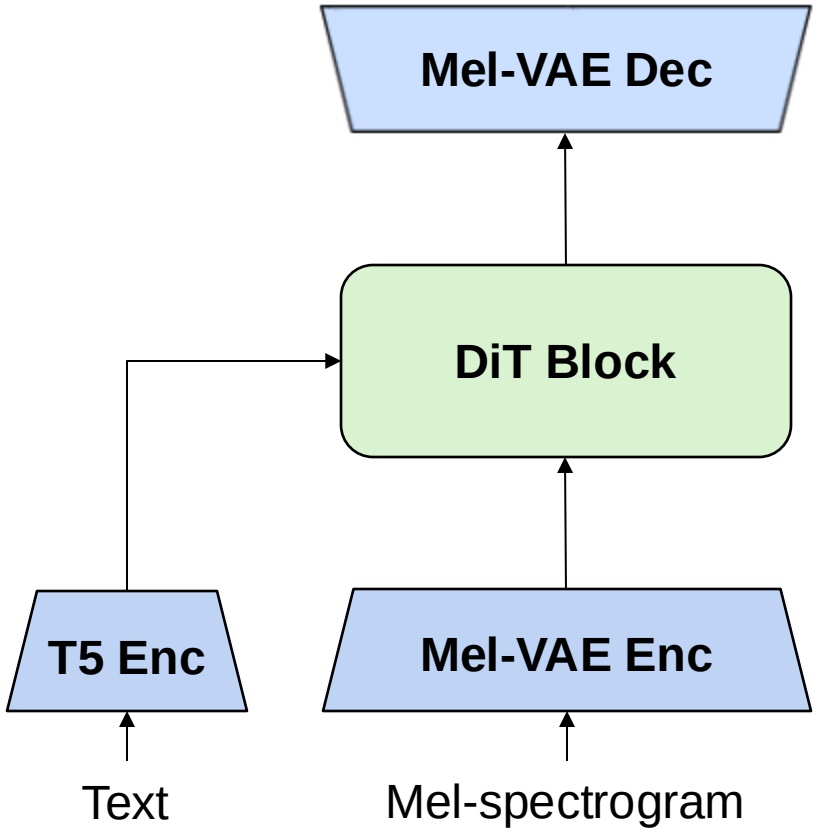
DiTTo-TTS: DiT to TTS

Latent Diffusion based TTS w/o phoneme & duration



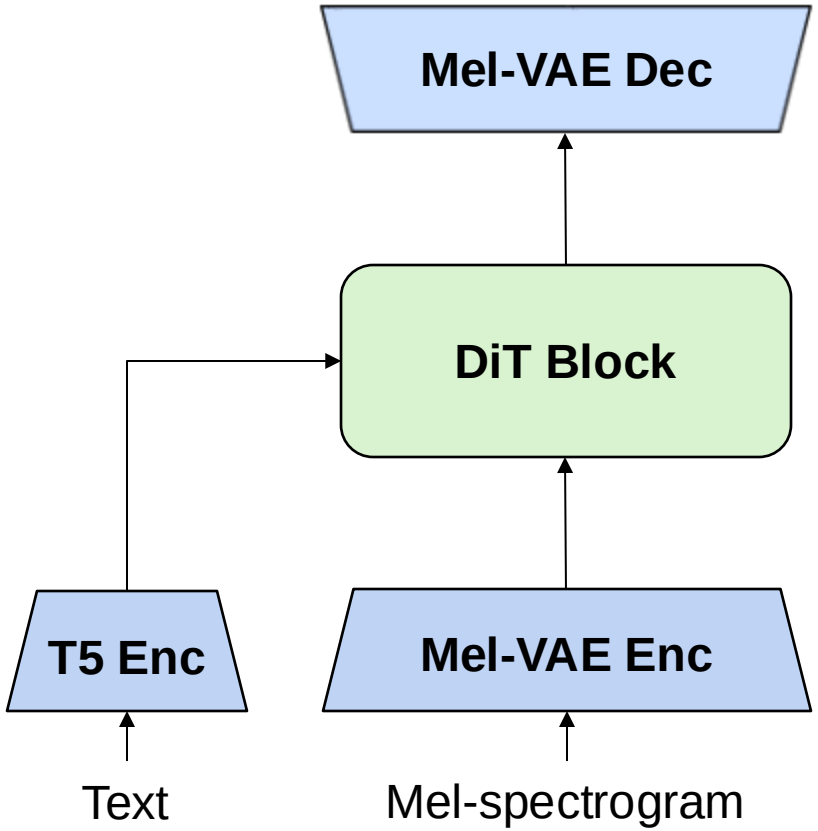
DiTTo-TTS: DiT to TTS

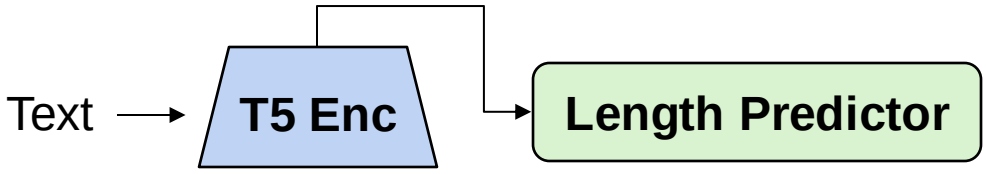
Latent Diffusion based TTS w/o phoneme & duration



DiTTo-TTS: DiT to TTS

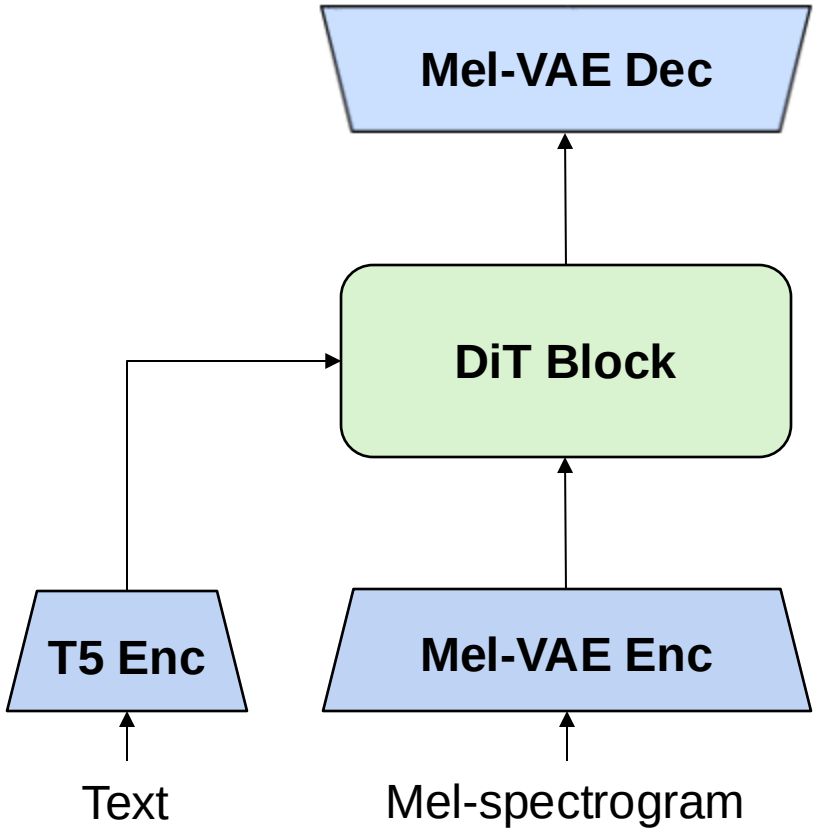
Latent Diffusion based TTS w/o phoneme & duration

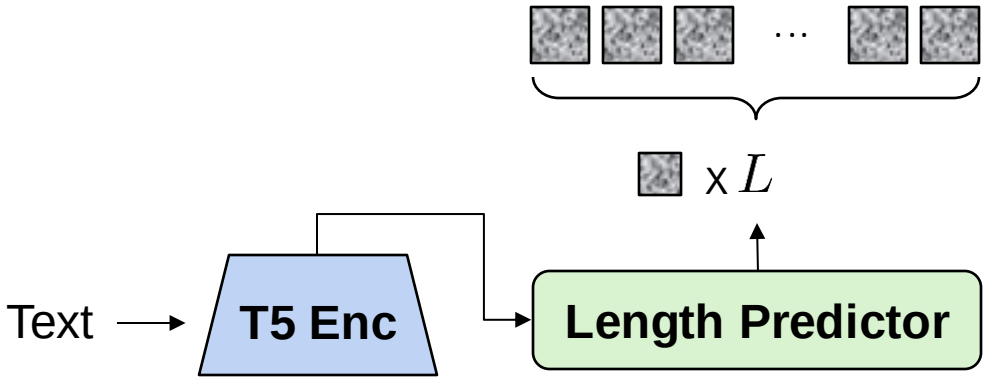




DiTTo-TTS: DiT to TTS

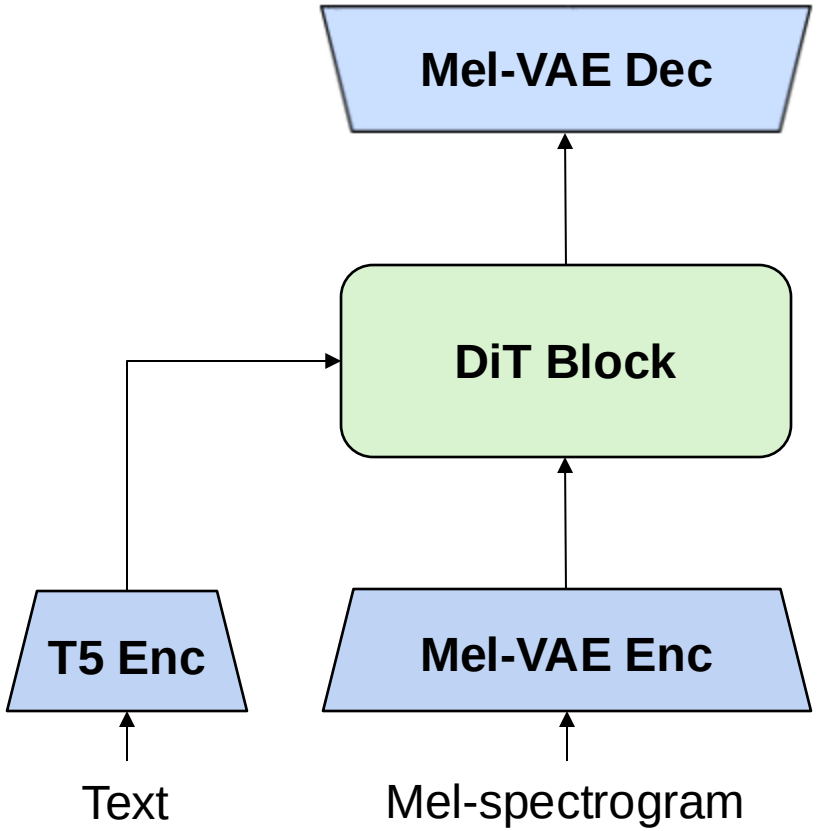
Latent Diffusion based TTS w/o phoneme & duration

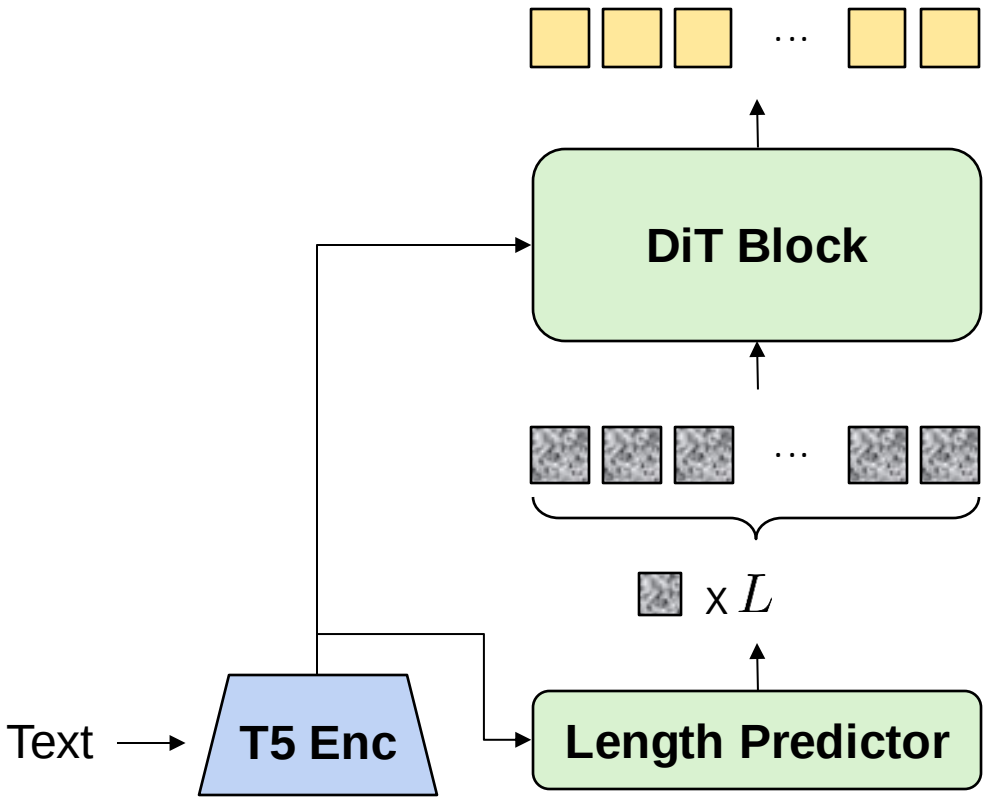




DiTTo-TTS: DiT to TTS

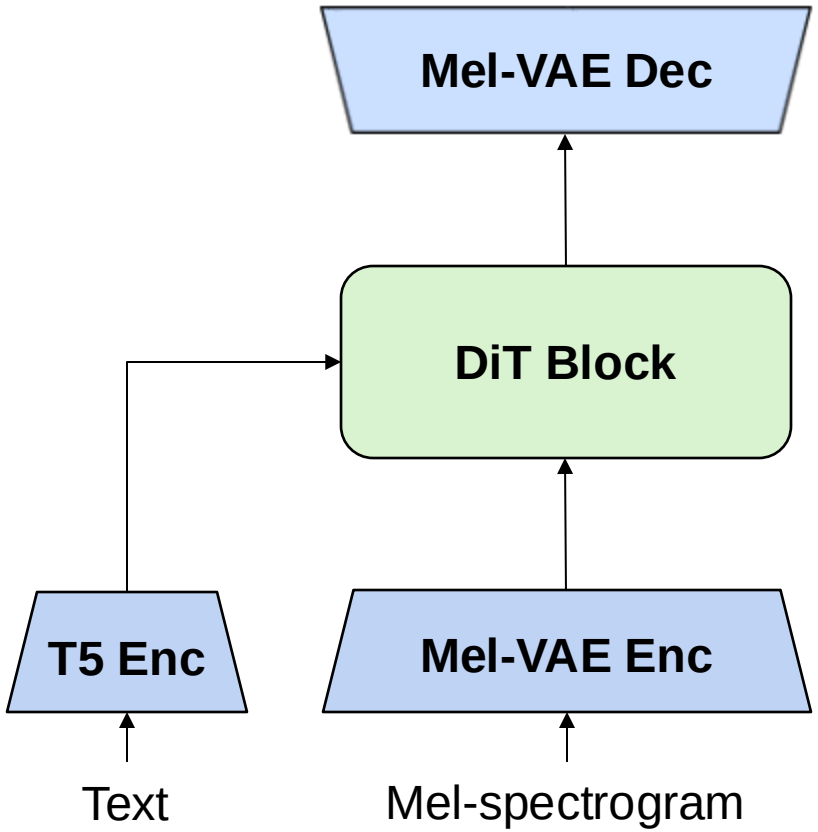
Latent Diffusion based TTS w/o phoneme & duration

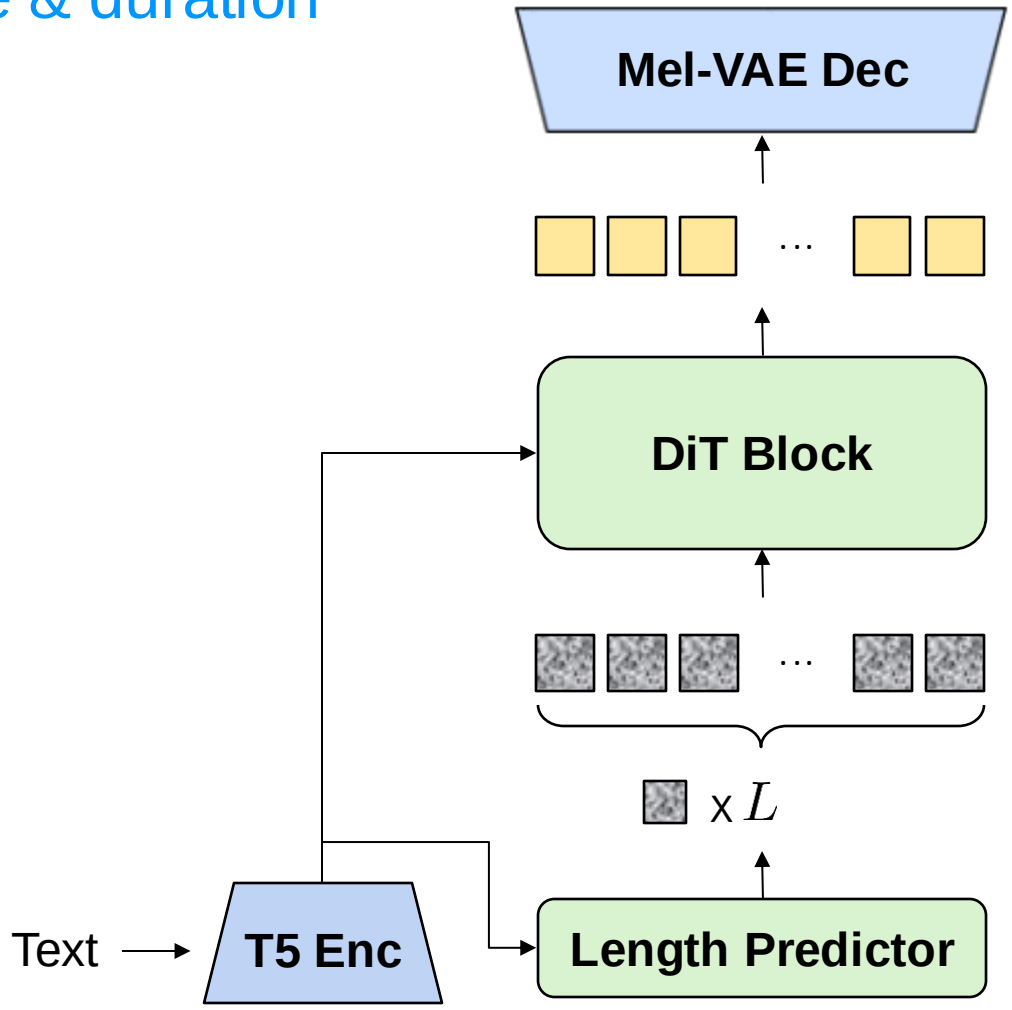




DiTTo-TTS: DiT to TTS

Latent Diffusion based TTS w/o phoneme & duration





Related Work (1/2)

Non-AR high-fidelity and controllable zero-shot TTS

Shen, K., Ju, Z., Tan, X., Liu, Y., Leng, Y., He, L., ... & Bian, J. (2023). Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. arXiv preprint arXiv:2304.09116.
Ju, Z., Wang, Y., Shen, K., Tan, X., Xin, D., Yang, D., ... & Zhao, S. (2024). NaturalSpeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. arXiv preprint arXiv:2403.03100.

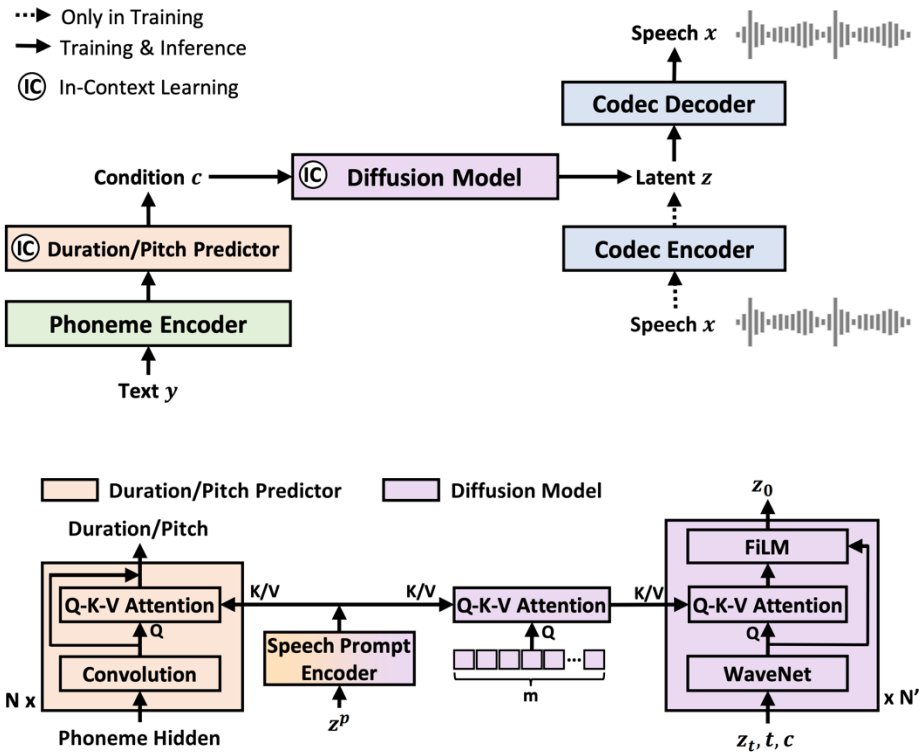
Related Work (1/2)

Non-AR high-fidelity and controllable zero-shot TTS w/ phoneme, duration, pitch …

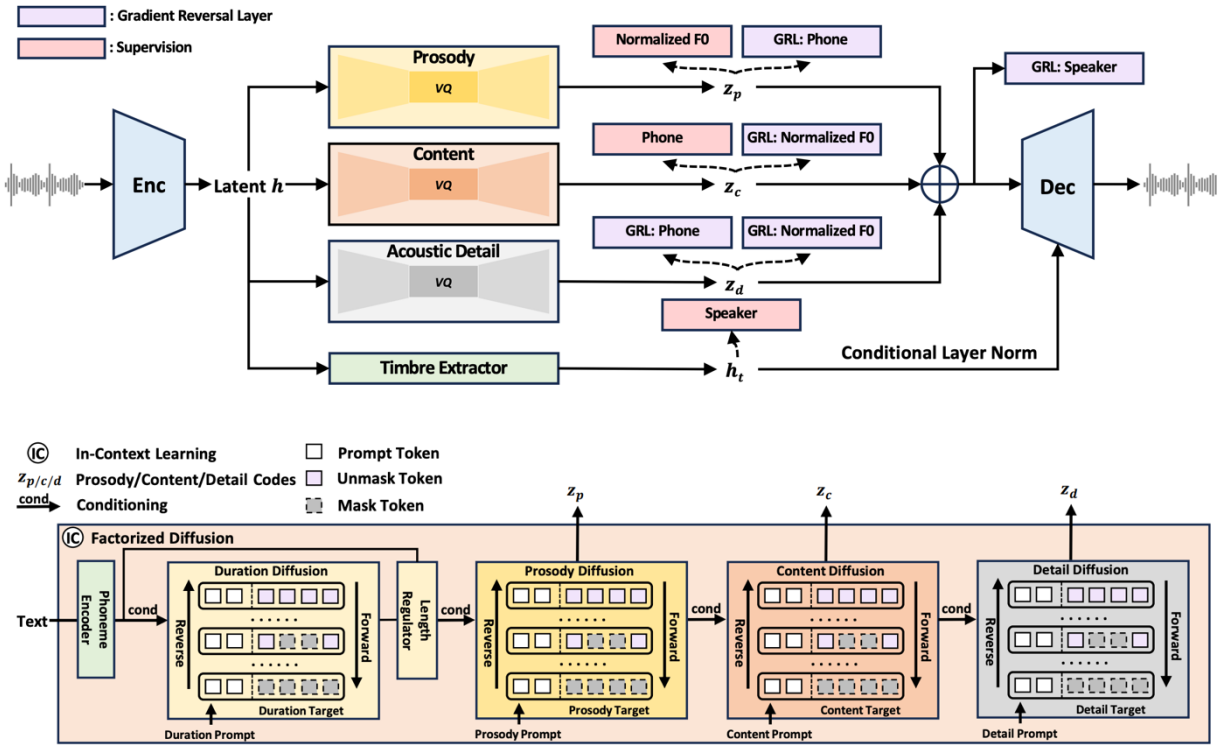
Shen, K., Ju, Z., Tan, X., Liu, Y., Leng, Y., He, L., ... & Bian, J. (2023). Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. arXiv preprint arXiv:2304.09116.
Ju, Z., Wang, Y., Shen, K., Tan, X., Xin, D., Yang, D., ... & Zhao, S. (2024). NaturalSpeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. arXiv preprint arXiv:2403.03100.

Related Work (1/2)

Non-AR high-fidelity and controllable zero-shot TTS w/ phoneme, duration, pitch …



NaturalSpeech 2



NaturalSpeech 3

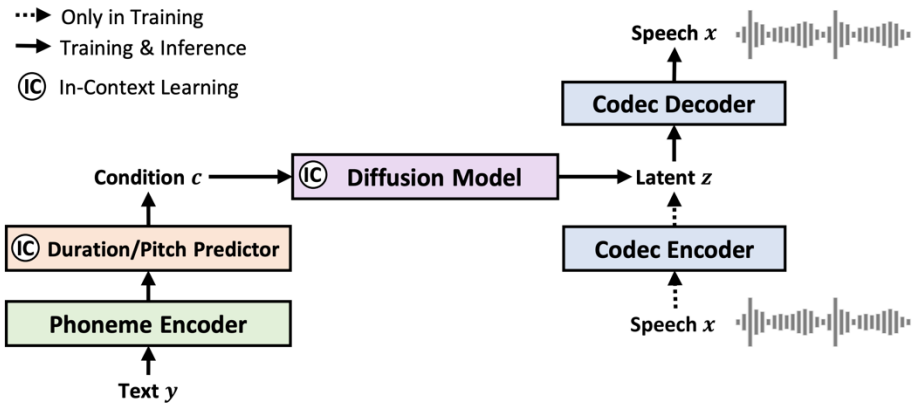
Shen, K., Ju, Z., Tan, X., Liu, Y., Leng, Y., He, L., ... & Bian, J. (2023). Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. arXiv preprint arXiv:2304.09116.

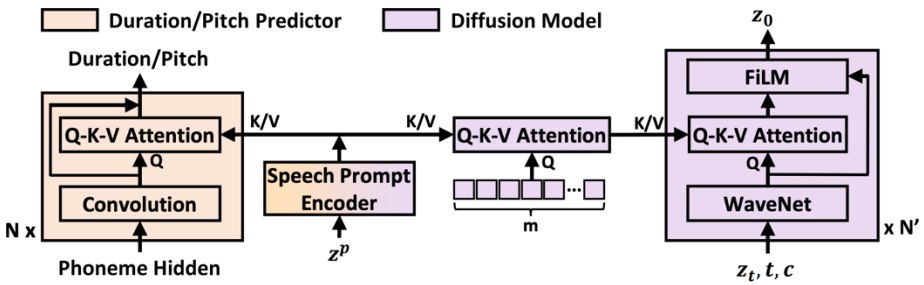
Ju, Z., Wang, Y., Shen, K., Tan, X., Xin, D., Yang, D., ... & Zhao, S. (2024). NaturalSpeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. arXiv preprint arXiv:2403.03100.

Related Work (1/2)

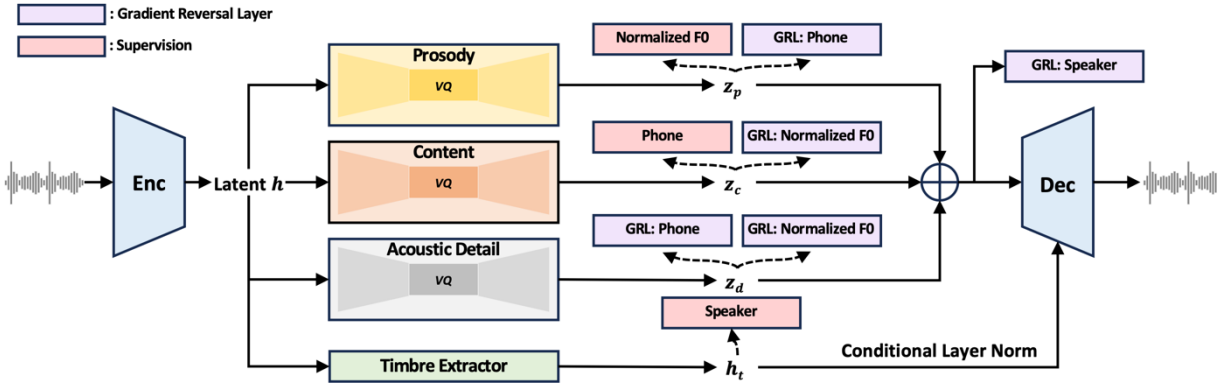
Non-AR high-fidelity and controllable zero-shot TTS w/ phoneme, duration, pitch …

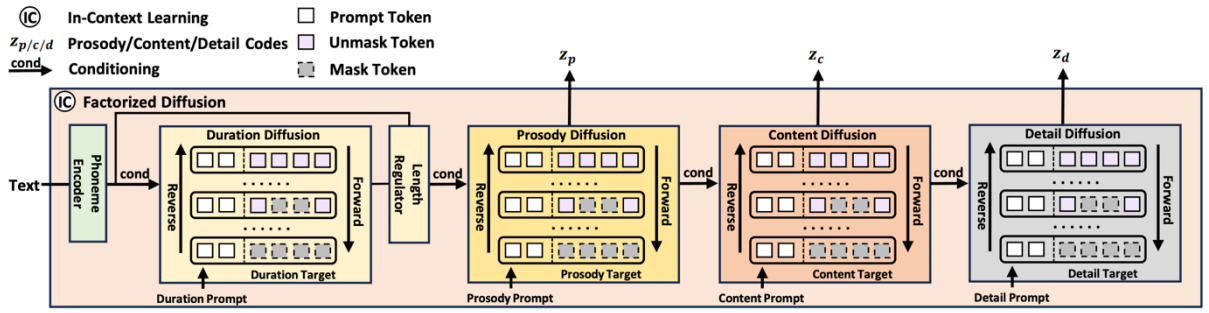
This requires a complex architecture and data preprocessing !





NaturalSpeech 2





NaturalSpeech 3

Shen, K., Ju, Z., Tan, X., Liu, Y., Leng, Y., He, L., ... & Bian, J. (2023). Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. arXiv preprint arXiv:2304.09116.

Ju, Z., Wang, Y., Shen, K., Tan, X., Xin, D., Yang, D., ... & Zhao, S. (2024). NaturalSpeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. arXiv preprint arXiv:2403.03100.

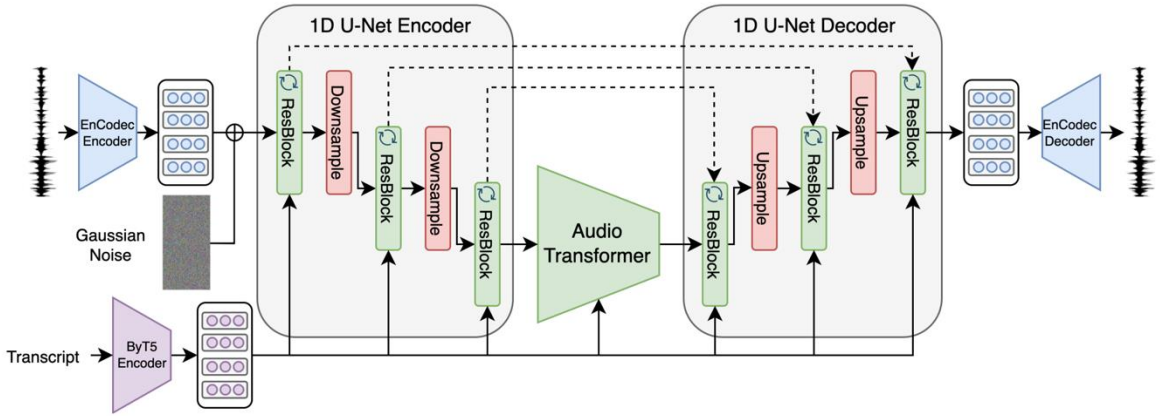
Related Work (2/2)

Simple diffusion based TTS w/o phoneme & duration

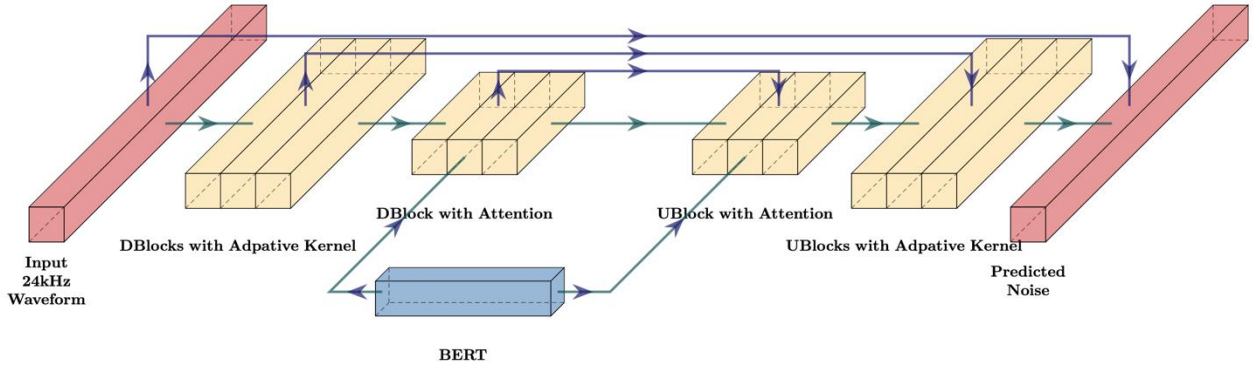
Lovelace, J., Ray, S., Kim, K., Weinberger, K. Q., & Wu, F. (2023). Simple-TTS: End-to-End Text-to-Speech Synthesis with Latent Diffusion.
Gao, Y., Morioka, N., Zhang, Y., & Chen, N. (2023, December). E3 TTS: Easy End-to-End Diffusion-Based Text To Speech. In 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) (pp. 1-8). IEEE.

Related Work (2/2)

Simple diffusion based TTS w/o phoneme & duration



Simple-TTS



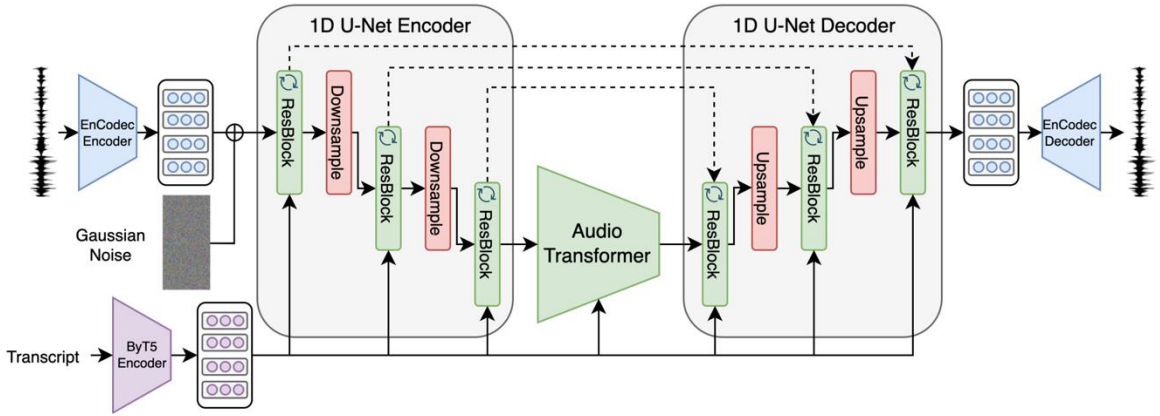
E3 TTS

Lovelace, J., Ray, S., Kim, K., Weinberger, K. Q., & Wu, F. (2023). Simple-TTS: End-to-End Text-to-Speech Synthesis with Latent Diffusion.
Gao, Y., Morioka, N., Zhang, Y., & Chen, N. (2023, December). E3 TTS: Easy End-to-End Diffusion-Based Text To Speech. In 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) (pp. 1-8). IEEE.

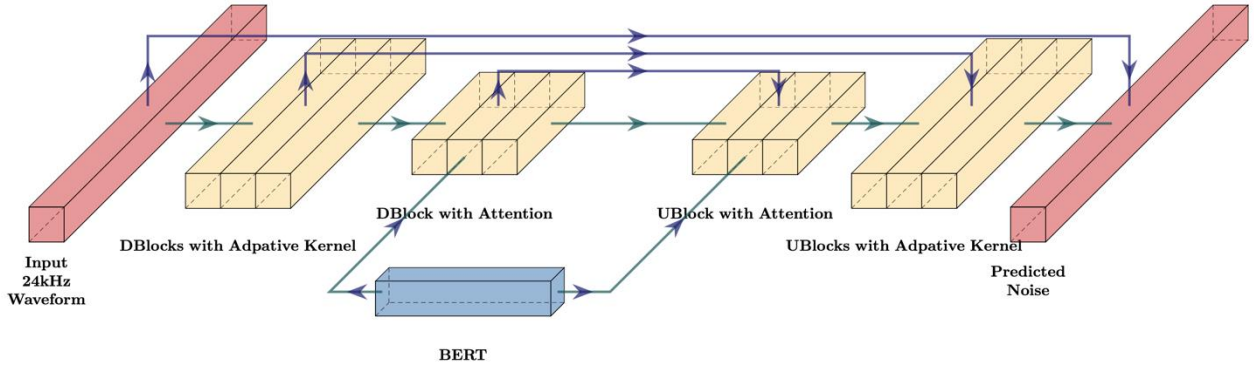
Related Work (2/2)

Simple diffusion based TTS w/o phoneme & duration

1. Their performance remains suboptimal.



Simple-TTS



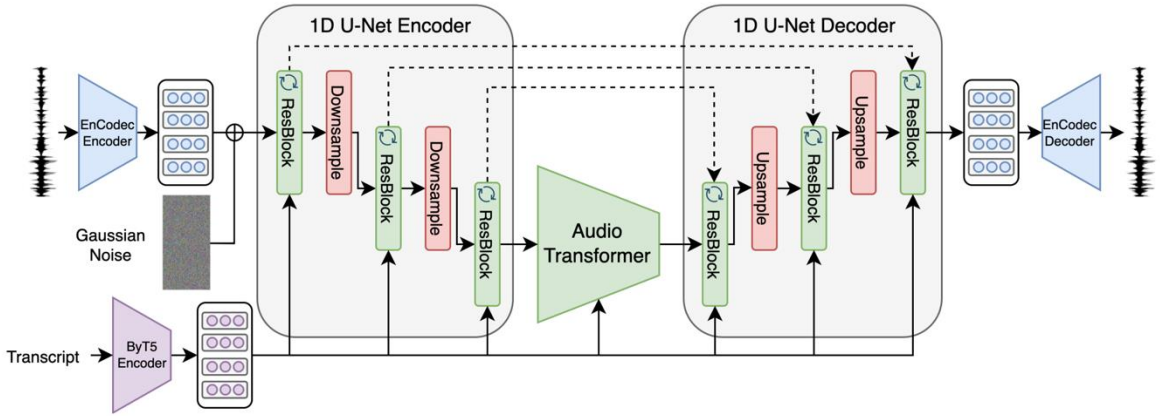
E3 TTS

Lovelace, J., Ray, S., Kim, K., Weinberger, K. Q., & Wu, F. (2023). Simple-TTS: End-to-End Text-to-Speech Synthesis with Latent Diffusion.
Gao, Y., Morioka, N., Zhang, Y., & Chen, N. (2023, December). E3 TTS: Easy End-to-End Diffusion-Based Text To Speech. In 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) (pp. 1-8). IEEE.

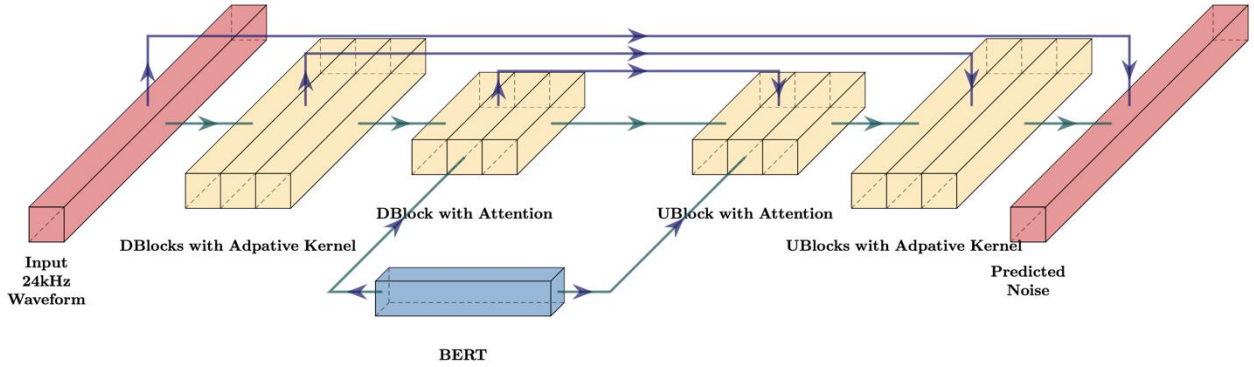
Related Work (2/2)

Simple diffusion based TTS w/o phoneme & duration

1. Their performance remains suboptimal.
2. They impose a fixed length on the target audio.



Simple-TTS



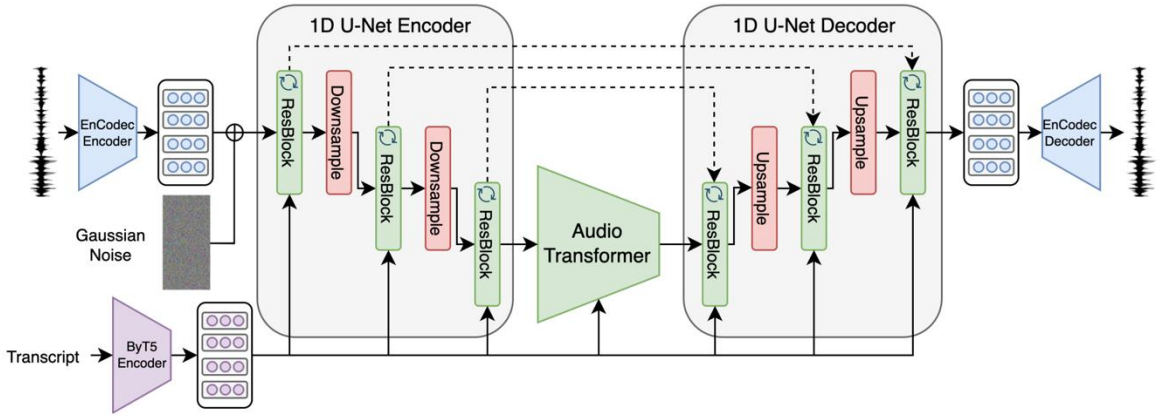
E3 TTS

Lovelace, J., Ray, S., Kim, K., Weinberger, K. Q., & Wu, F. (2023). Simple-TTS: End-to-End Text-to-Speech Synthesis with Latent Diffusion.
Gao, Y., Morioka, N., Zhang, Y., & Chen, N. (2023, December). E3 TTS: Easy End-to-End Diffusion-Based Text To Speech. In 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) (pp. 1-8). IEEE.

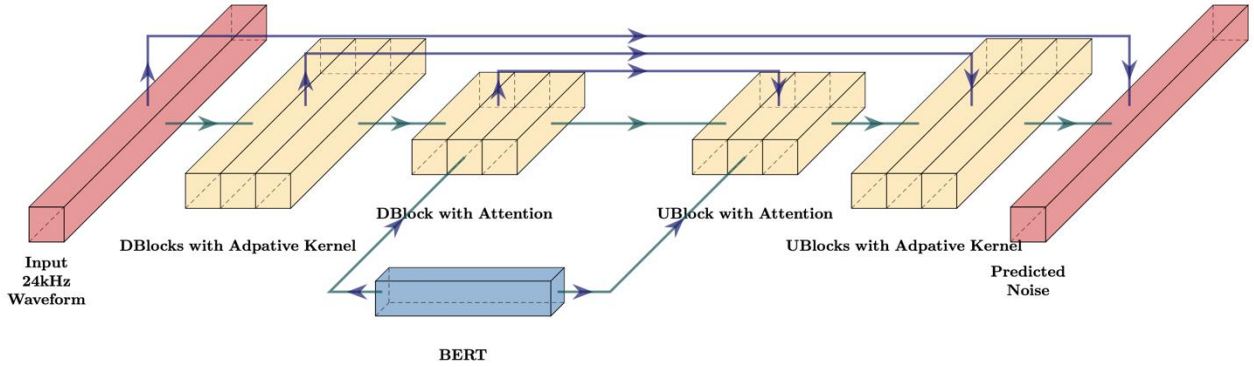
Related Work (2/2)

Simple diffusion based TTS w/o phoneme & duration

1. Their performance remains suboptimal.
2. They impose a fixed length on the target audio.
3. They are not yet large-scale and handle only English.



Simple-TTS



E3 TTS

Lovelace, J., Ray, S., Kim, K., Weinberger, K. Q., & Wu, F. (2023). Simple-TTS: End-to-End Text-to-Speech Synthesis with Latent Diffusion.
Gao, Y., Morioka, N., Zhang, Y., & Chen, N. (2023, December). E3 TTS: Easy End-to-End Diffusion-Based Text To Speech. In 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) (pp. 1-8). IEEE.

Research Question

RQ1.

RQ2.

Research Question

RQ1. Can LDM achieve state-of-the-art performance in text-to-speech tasks at scale **without relying on domain-specific factors**, similar to successes in other domains?

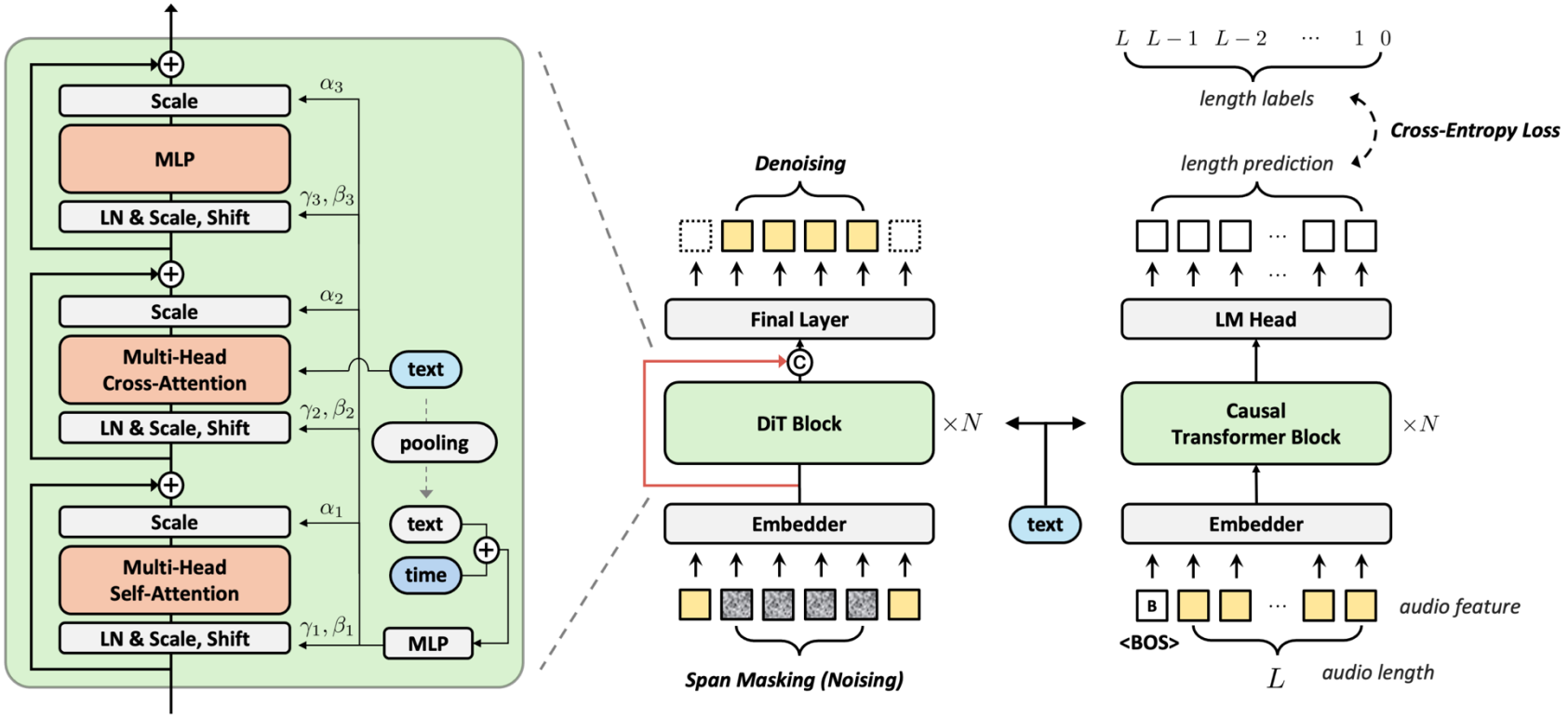
RQ2.

Research Question

RQ1. Can LDM achieve state-of-the-art performance in text-to-speech tasks at scale **without relying on domain-specific factors**, similar to successes in other domains?

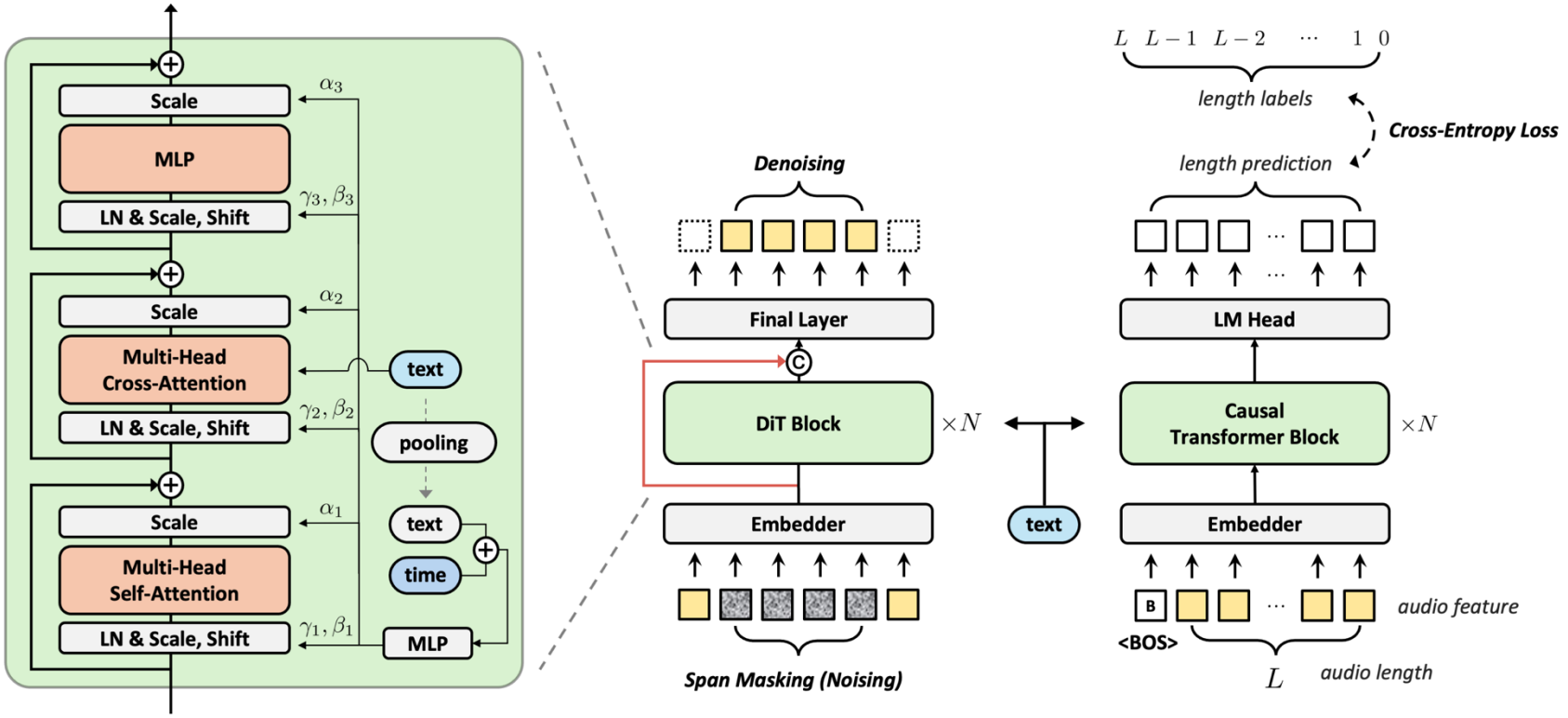
RQ2. If LDM can achieve this, **what are the primary aspects** that enable its success?

Method



$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{t\sim\mathcal{U}(1,T),\boldsymbol{\epsilon}\sim\mathcal{N}(0,I)}\left[\|\boldsymbol{m}\odot(\boldsymbol{v}^{(t)}-\boldsymbol{v}_\theta(\boldsymbol{z}^{(t)}_{mask},\boldsymbol{m},\boldsymbol{z}_{text},t))\|^2\right]$$

Method

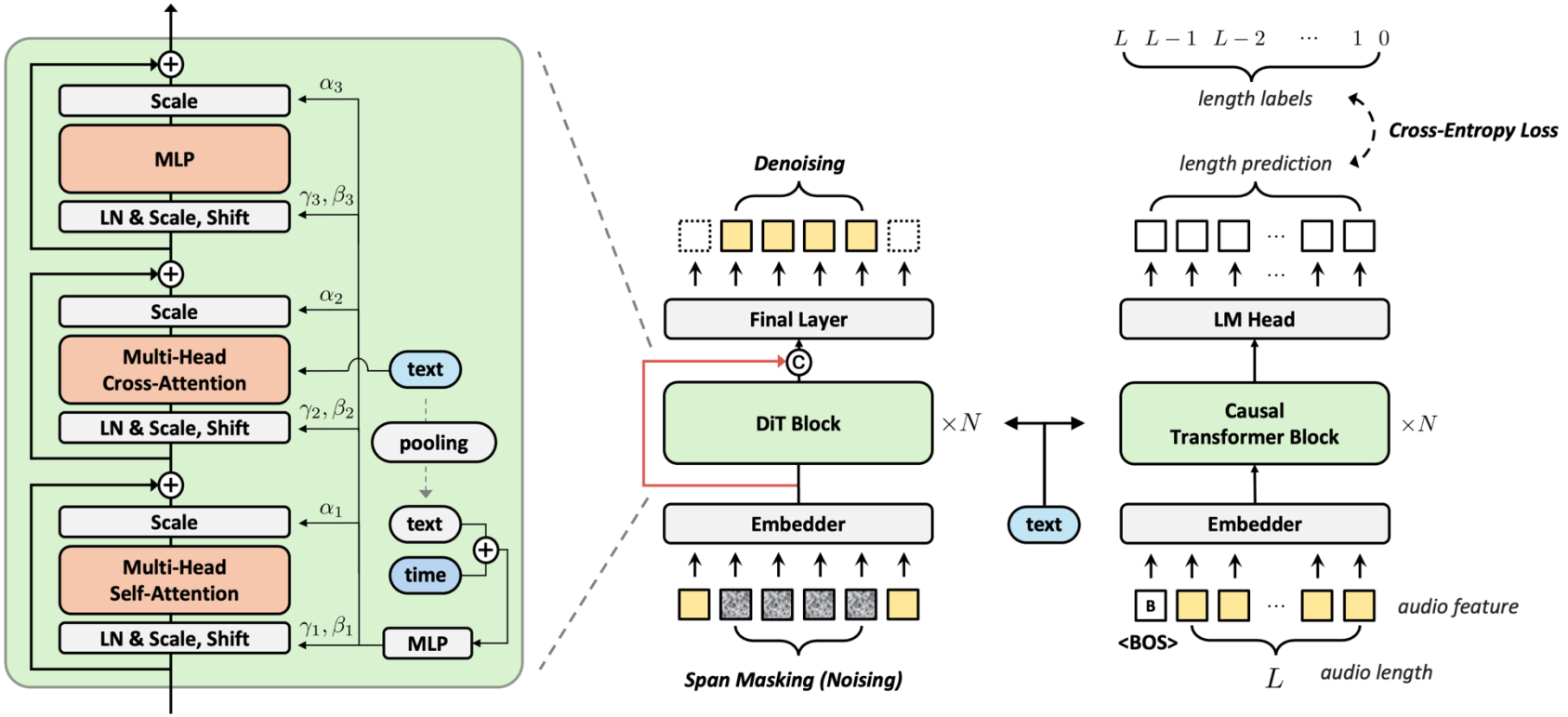


gaussian noise

$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{t\sim\mathcal{U}(1,T),\boldsymbol{\epsilon}\sim\mathcal{N}(0,I)}\left[\|\boldsymbol{m}\odot(\boldsymbol{v}^{(t)} - \boldsymbol{v}_\theta(\boldsymbol{z}^{(t)}_{mask}, \boldsymbol{m}, \boldsymbol{z}_{text}, t))\|^2\right]$$

diffusion timestep

Method



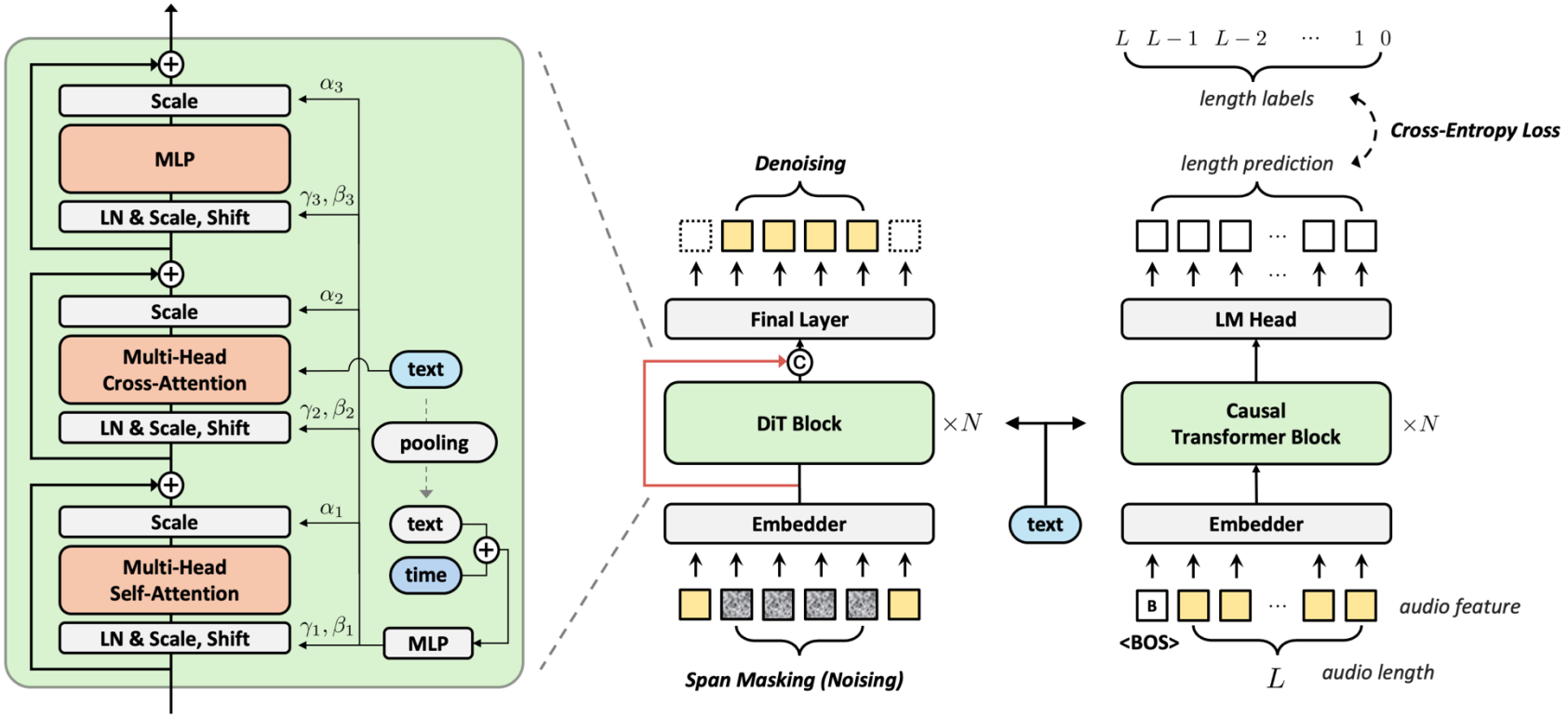
gaussian noise

velocity for v-prediction

$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{t\sim\mathcal{U}(1,T),\boldsymbol{\epsilon}\sim\mathcal{N}(0,I)}\left[\|\boldsymbol{m}\odot(\boldsymbol{v}^{(t)} - \boldsymbol{v}_\theta(\boldsymbol{z}^{(t)}_{mask}, \boldsymbol{m}, \boldsymbol{z}_{text}, t))\|^2\right]$$

diffusion timestep

Method

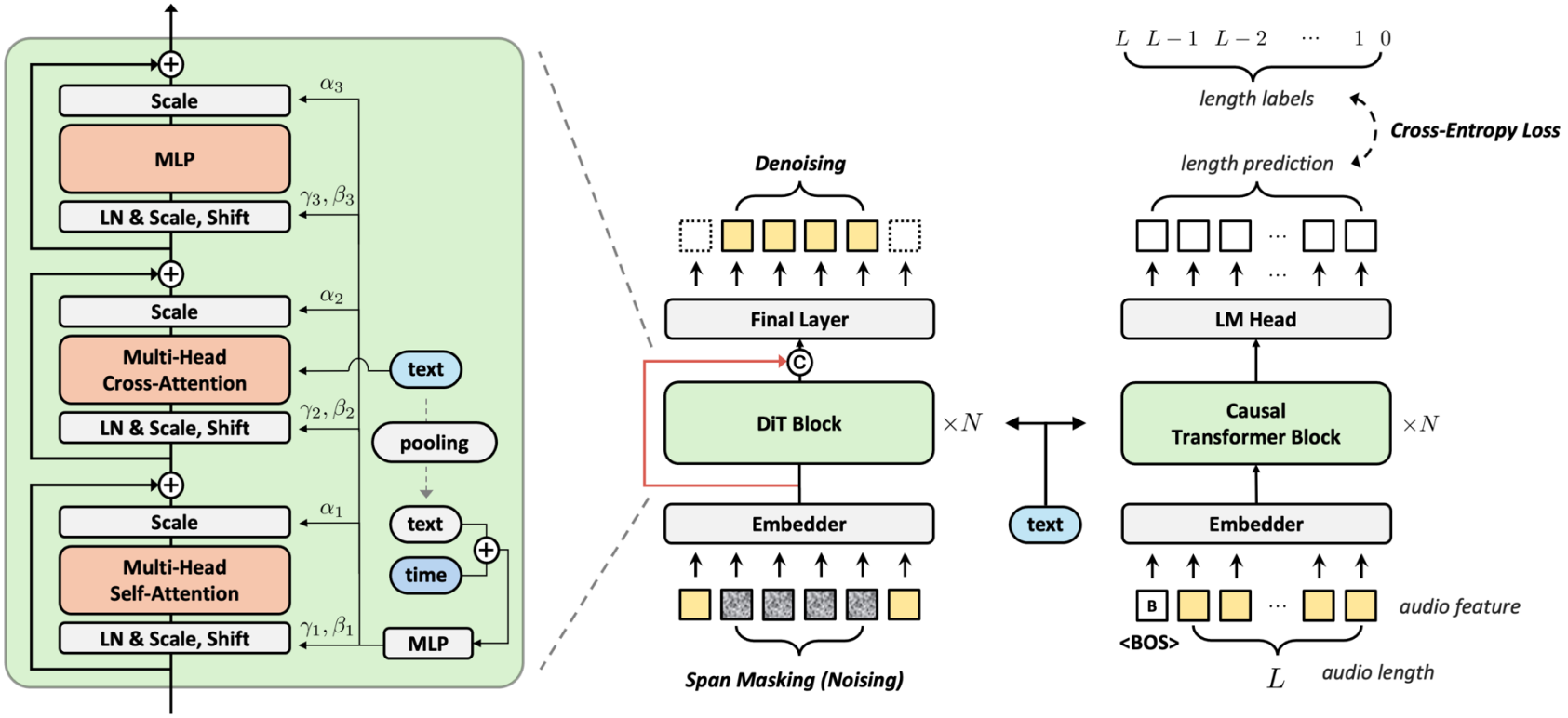


gaussian noise | velocity | text latent

$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{t\sim\mathcal{U}(1,T),\boldsymbol{\epsilon}\sim\mathcal{N}(0,I)}\left[\|\boldsymbol{m}\odot(\boldsymbol{v}^{(t)} - \boldsymbol{v}_\theta(\boldsymbol{z}^{(t)}_{mask}, \boldsymbol{m}, \boldsymbol{z}_{text}, t))\|^2\right]$$

diffusion timestep | masked speech latent

Method



$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{t\sim\mathcal{U}(1,T),\boldsymbol{\epsilon}\sim\mathcal{N}(0,I)}\left[\|\boldsymbol{m}\odot(\boldsymbol{v}^{(t)}-\boldsymbol{v}_\theta(\boldsymbol{z}^{(t)}_{mask},\boldsymbol{m},\boldsymbol{z}_{text},t))\|^2\right]$$

gaussian noise · velocity · text latent · diffusion timestep · mask · masked speech latent

RQ1. We Achieve SOTA without TTS-Specific Factors

Model	Objective Metrics					
	WER ↓	CER ↓	SIM-o ↑	SIM-r ↑	Inference Time ↓	#Param.
Ground Truth	2.15	0.61	0.7395	-	n/a	n/a
YourTTS	7.57	3.06	0.3928	-	-	-
VALL-E	3.8*	-	0.452*	0.508*	∼6.2s*	302M
Voicebox	2.0*	-	0.593*	0.616*	∼6.4s* (64 NFE)	364M
CLaM-TTS	2.36*	0.79*	0.4767*	0.5128*	4.15s*	584M
Simple-TTS	3.86	2.24	0.4413	0.4668	17.897s (250 NFE)	243M
DiTTo-en-S	2.01	0.60	0.4544	0.4935	**0.884s**	42M
DiTTo-en-B	1.87	0.52	0.5535	0.5855	0.903s	152M
DiTTo-en-L	1.85	0.50	0.5596	0.5913	1.479s	508M
DiTTo-en-XL	**1.78**	**0.48**	0.5773	0.6075	1.616s	740M
DiTTo-en-XL†	1.80	0.48	**0.6051**	**0.6283**	-	740M

Table 1. English-only *continuation* Task

Model	WER ↓	CER ↓	SIM-o ↑	SIM-r ↑
YourTTS	7.92 (7.7*)	3.18	0.3755 (0.337*)	-
VALL-E	5.9*	-	-	0.580*
SPEAR-TTS	-	1.92*	-	0.560*
Voicebox	**1.9***	-	**0.662***	**0.681***
CLaM-TTS	5.11*	2.87*	0.4951*	0.5382*
Simple-TTS	4.09 (3.4*)	2.11	0.5026	0.5305 (0.514*)
DiTTo-en-S	3.07	1.08	0.4984	0.5373
DiTTo-en-B	2.74	0.98	0.5977	0.6281
DiTTo-en-L	2.69	0.91	0.6050	0.6355
DiTTo-en-XL	2.56	**0.89**	0.6270	0.6554
DiTTo-en-XL†	2.64	0.94	0.6538	0.6752

Table 2. English-only *cross-sentence* Task

Model	SMOS	CMOS
Simple-TTS	2.15±0.19	-1.64
CLaM-en	3.42±0.16	-0.52
DiTTo-en-XL	**3.91±0.16**	**0.00**
Ground Truth (*recon*)	4.07±0.14	+0.11
Ground Truth	4.08±0.14	+0.13

Table 3. MOS on *cross-sentence* Task

Model	SMOS	CMOS
Voicebox (Le et al., 2023)	2.87±0.45	**0.14**
DiTTo-en	**3.57±0.39**	0.0
VALL-E (Wang et al., 2023)	3.50±0.46	-0.94
DiTTo-en	**3.90±0.38**	**0.0**
MegaTTS (Jiang et al., 2023)	3.76±0.44	-0.31
DiTTo-en	**3.82±0.32**	**0.0**
E3-TTS (Gao et al., 2023)	4.25±0.19	-0.30
DiTTo-en	**4.28±0.32**	**0.0**
NaturalSpeech 2 (Shen et al., 2024)	3.99±0.37	-0.29
DiTTo-en	**4.01±0.35**	**0.0**
NaturalSpeech 3 (Ju et al., 2024)	**4.42±0.28**	-0.27
DiTTo-en	3.92±0.27	**0.0**

Table 12. MOS on demo samples from various SOTA with DiTTo-en (XL)

RQ1. We Achieve SOTA without TTS-Specific Factors

Model	Objective Metrics					
	WER ↓	CER ↓	SIM-o ↑	SIM-r ↑	Inference Time ↓	#Param.
Ground Truth	2.15	0.61	0.7395	-	n/a	n/a
YourTTS	7.57	3.06	0.3928	-	-	-
VALL-E	3.8*	-	0.452*	0.508*	∼6.2s*	302M
Voicebox	2.0*	-	0.593*	0.616*	∼6.4s* (64 NFE)	364M
CLaM-TTS	2.36*	0.79*	0.4767*	0.5128*	4.15s*	584M
Simple-TTS	3.86	2.24	0.4413	0.4668	17.897s (250 NFE)	243M
DiTTo-en-S	2.01	0.60	0.4544	0.4935	**0.884s**	42M
DiTTo-en-B	1.87	0.52	0.5535	0.5855	0.903s	152M
DiTTo-en-L	1.85	0.50	0.5596	0.5913	1.479s	508M
DiTTo-en-XL	**1.78**	**0.48**	0.5773	0.6075	1.616s	740M
DiTTo-en-XL†	1.80	0.48	**0.6051**	**0.6283**	-	740M

Table 1. English-only *continuation* Task

Model	WER ↓	CER ↓	SIM-o ↑	SIM-r ↑
YourTTS	7.92 (7.7*)	3.18	0.3755 (0.337*)	-
VALL-E	5.9*	-	-	0.580*
SPEAR-TTS	-	1.92*	-	0.560*
Voicebox	**1.9***	-	**0.662***	**0.681***
CLaM-TTS	5.11*	2.87*	0.4951*	0.5382*
Simple-TTS	4.09 (3.4*)	2.11	0.5026	0.5305 (0.514*)
DiTTo-en-S	3.07	1.08	0.4984	0.5373
DiTTo-en-B	2.74	0.98	0.5977	0.6281
DiTTo-en-L	2.69	0.91	0.6050	0.6355
DiTTo-en-XL	2.56	**0.89**	0.6270	0.6554
DiTTo-en-XL†	2.64	0.94	0.6538	0.6752

Table 2. English-only *cross-sentence* Task

Model	SMOS	CMOS
Simple-TTS	2.15±0.19	-1.64
CLaM-en	3.42±0.16	-0.52
DiTTo-en-XL	**3.91±0.16**	**0.00**
Ground Truth (*recon*)	4.07±0.14	+0.11
Ground Truth	4.08±0.14	+0.13

Table 3. MOS on *cross-sentence* Task

Model	SMOS	CMOS
Voicebox (Le et al., 2023)	2.87±0.45	**0.14**
DiTTo-en	**3.57±0.39**	0.0
VALL-E (Wang et al., 2023)	3.50±0.46	-0.94
DiTTo-en	**3.90±0.38**	**0.0**
MegaTTS (Jiang et al., 2023)	3.76±0.44	-0.31
DiTTo-en	**3.82±0.32**	**0.0**
E3-TTS (Gao et al., 2023)	4.25±0.19	-0.30
DiTTo-en	**4.28±0.32**	**0.0**
NaturalSpeech 2 (Shen et al., 2024)	3.99±0.37	-0.29
DiTTo-en	**4.01±0.35**	**0.0**
NaturalSpeech 3 (Ju et al., 2024)	**4.42±0.28**	-0.27
DiTTo-en	3.92±0.27	**0.0**

Table 12. MOS on demo samples from various SOTA with DiTTo-en (XL)

RQ1. We Achieve SOTA without TTS-Specific Factors

Model	Objective Metrics					
	WER ↓	CER ↓	SIM-o ↑	SIM-r ↑	Inference Time ↓	#Param.
Ground Truth	2.15	0.61	0.7395	-	n/a	n/a
YourTTS	7.57	3.06	0.3928	-	-	-
VALL-E	3.8*	-	0.452*	0.508*	∼6.2s*	302M
Voicebox	2.0*	-	0.593*	0.616*	∼6.4s* (64 NFE)	364M
CLaM-TTS	2.36*	0.79*	0.4767*	0.5128*	4.15s*	584M
Simple-TTS	3.86	2.24	0.4413	0.4668	17.897s (250 NFE)	243M
DiTTo-en-S	2.01	0.60	0.4544	0.4935	**0.884s**	42M
DiTTo-en-B	1.87	0.52	0.5535	0.5855	0.903s	152M
DiTTo-en-L	1.85	0.50	0.5596	0.5913	1.479s	508M
DiTTo-en-XL	**1.78**	**0.48**	0.5773	0.6075	1.616s	740M
DiTTo-en-XL†	1.80	0.48	**0.6051**	**0.6283**	-	740M

Table 1. English-only *continuation* Task

Model	WER ↓	CER ↓	SIM-o ↑	SIM-r ↑
YourTTS	7.92 (7.7*)	3.18	0.3755 (0.337*)	-
VALL-E	5.9*	-	-	0.580*
SPEAR-TTS	-	1.92*	-	0.560*
Voicebox	**1.9***	-	**0.662***	**0.681***
CLaM-TTS	5.11*	2.87*	0.4951*	0.5382*
Simple-TTS	4.09 (3.4*)	2.11	0.5026	0.5305 (0.514*)
DiTTo-en-S	3.07	1.08	0.4984	0.5373
DiTTo-en-B	2.74	0.98	0.5977	0.6281
DiTTo-en-L	2.69	0.91	0.6050	0.6355
DiTTo-en-XL	2.56	**0.89**	0.6270	0.6554
DiTTo-en-XL†	2.64	0.94	0.6538	0.6752

Table 2. English-only *cross-sentence* Task

Model	SMOS	CMOS
Simple-TTS	2.15±0.19	-1.64
CLaM-en	3.42±0.16	-0.52
DiTTo-en-XL	**3.91±0.16**	**0.00**
Ground Truth (*recon*)	4.07±0.14	+0.11
Ground Truth	4.08±0.14	+0.13

Table 3. MOS on *cross-sentence* Task

Model	SMOS	CMOS
Voicebox (Le et al., 2023)	2.87±0.45	**0.14**
DiTTo-en	**3.57±0.39**	0.0
VALL-E (Wang et al., 2023)	3.50±0.46	-0.94
DiTTo-en	**3.90±0.38**	**0.0**
MegaTTS (Jiang et al., 2023)	3.76±0.44	-0.31
DiTTo-en	**3.82±0.32**	**0.0**
E3-TTS (Gao et al., 2023)	4.25±0.19	-0.30
DiTTo-en	**4.28±0.32**	**0.0**
NaturalSpeech 2 (Shen et al., 2024)	3.99±0.37	-0.29
DiTTo-en	**4.01±0.35**	**0.0**
NaturalSpeech 3 (Ju et al., 2024)	**4.42±0.28**	-0.27
DiTTo-en	3.92±0.27	**0.0**

Table 12. MOS on demo samples from various SOTA with DiTTo-en (XL)

RQ1. We Achieve SOTA without TTS-Specific Factors

Model	Objective Metrics					
	WER ↓	CER ↓	SIM-o ↑	SIM-r ↑	Inference Time ↓	#Param.
Ground Truth	2.15	0.61	0.7395	-	n/a	n/a
YourTTS	7.57	3.06	0.3928	-	-	-
VALL-E	3.8*	-	0.452*	0.508*	∼6.2s*	302M
Voicebox	2.0*	-	0.593*	0.616*	∼6.4s* (64 NFE)	364M
CLaM-TTS	2.36*	0.79*	0.4767*	0.5128*	4.15s*	584M
Simple-TTS	3.86	2.24	0.4413	0.4668	17.897s (250 NFE)	243M
DiTTo-en-S	2.01	0.60	0.4544	0.4935	**0.884s**	42M
DiTTo-en-B	1.87	0.52	0.5535	0.5855	0.903s	152M
DiTTo-en-L	1.85	0.50	0.5596	0.5913	1.479s	508M
DiTTo-en-XL	**1.78**	**0.48**	0.5773	0.6075	1.616s	740M
DiTTo-en-XL†	1.80	0.48	**0.6051**	**0.6283**	-	740M

Table 1. English-only *continuation* Task

Model	WER ↓	CER ↓	SIM-o ↑	SIM-r ↑
YourTTS	7.92 (7.7*)	3.18	0.3755 (0.337*)	-
VALL-E	5.9*	-	-	0.580*
SPEAR-TTS	-	1.92*	-	0.560*
Voicebox	**1.9***	-	**0.662***	**0.681***
CLaM-TTS	5.11*	2.87*	0.4951*	0.5382*
Simple-TTS	4.09 (3.4*)	2.11	0.5026	0.5305 (0.514*)
DiTTo-en-S	3.07	1.08	0.4984	0.5373
DiTTo-en-B	2.74	0.98	0.5977	0.6281
DiTTo-en-L	2.69	0.91	0.6050	0.6355
DiTTo-en-XL	2.56	**0.89**	0.6270	0.6554
DiTTo-en-XL†	2.64	0.94	0.6538	0.6752

Table 2. English-only *cross-sentence* Task

Model	SMOS	CMOS
Simple-TTS	2.15±0.19	-1.64
CLaM-en	3.42±0.16	-0.52
DiTTo-en-XL	**3.91±0.16**	**0.00**
Ground Truth (*recon*)	4.07±0.14	+0.11
Ground Truth	4.08±0.14	+0.13

Table 3. MOS on *cross-sentence* Task

Model	SMOS	CMOS
Voicebox (Le et al., 2023)	2.87±0.45	**0.14**
DiTTo-en	**3.57±0.39**	0.0
VALL-E (Wang et al., 2023)	3.50±0.46	-0.94
DiTTo-en	**3.90±0.38**	**0.0**
MegaTTS (Jiang et al., 2023)	3.76±0.44	-0.31
DiTTo-en	**3.82±0.32**	**0.0**
E3-TTS (Gao et al., 2023)	4.25±0.19	-0.30
DiTTo-en	**4.28±0.32**	**0.0**
NaturalSpeech 2 (Shen et al., 2024)	3.99±0.37	-0.29
DiTTo-en	**4.01±0.35**	**0.0**
NaturalSpeech 3 (Ju et al., 2024)	**4.42±0.28**	-0.27
DiTTo-en	3.92±0.27	**0.0**

Table 12. MOS on demo samples from various SOTA with DiTTo-en (XL)

RQ2. Key Aspects Behind The Scene (1/3)

DiT Fits Better Than U-Net

Model	WER ↓	SIM-r ↑	Inference Time ↓
U-Net	3.70	0.3890	1.328s
Flat-U-Net	2.97	0.5471	1.310s
DiTTo-*mls*	**2.93**	**0.5877**	**0.903s**

Table 4. English-only *cross-sentence* Task

RQ2. Key Aspects Behind The Scene (1/3)

DiT Fits Better Than U-Net

Model	WER ↓	SIM-r ↑	Inference Time ↓
U-Net	3.70	0.3890	1.328s
Flat-U-Net	2.97	0.5471	1.310s
DiTTo-*mls*	**2.93**	**0.5877**	**0.903s**

Table 4. English-only *cross-sentence* Task

1. DiT offers higher accuracy and better captures the speaker's characteristics.

RQ2. Key Aspects Behind The Scene (1/3)

DiT Fits Better Than U-Net

Model	WER ↓	SIM-r ↑	Inference Time ↓
U-Net	3.70	0.3890	1.328s
Flat-U-Net	2.97	0.5471	1.310s
DiTTo-*mls*	**2.93**	**0.5877**	**0.903s**

Table 4. English-only *cross-sentence* Task

1. DiT offers higher accuracy and better captures the speaker's characteristics.

2. DiT is faster.

RQ2. Key Aspects Behind The Scene (2/3)

Variable Length Modeling Outperforms Fixed Length Modeling

Model	WER ↓	SIM-r ↑	Inference Time ↓
fixed-length-full	8.89	0.4078	1.254s
fixed-length	6.81	0.4385	1.265s
SLP-CE	<u>5.58</u>	**0.4961**	0.948s
SLP-Regression	**5.36**	<u>0.4636</u>	**0.930s**

Table 5. Speech length modeling

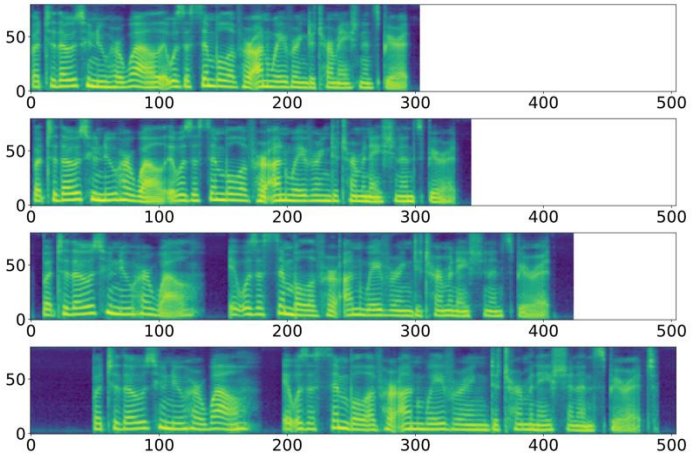


Figure 2. Speech rate controllability

RQ2. Key Aspects Behind The Scene (2/3)

Variable Length Modeling Outperforms Fixed Length Modeling

	Model	WER ↓	SIM-r ↑	Inference Time ↓
Simple-TTS, E3 TTS	*fixed-length-full*	8.89	0.4078	1.254s
	fixed-length	6.81	0.4385	1.265s
Ours	*SLP-CE*	<u>5.58</u>	**0.4961**	0.948s
	SLP-Regression	**5.36**	<u>0.4636</u>	**0.930s**

Table 5. Speech length modeling

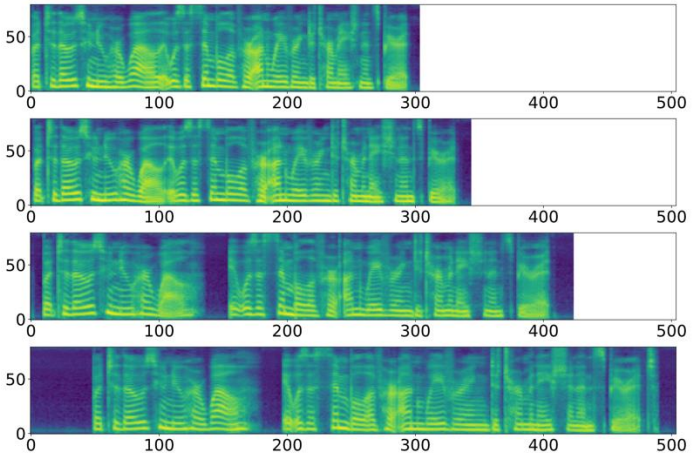


Figure 2. Speech rate controllability

1. Variable length delivers better quality and faster performance.

RQ2. Key Aspects Behind The Scene (2/3)

Variable Length Modeling Outperforms Fixed Length Modeling

	Model	WER ↓	SIM-r ↑	Inference Time ↓
Simple-TTS, E3 TTS	*fixed-length-full*	8.89	0.4078	1.254s
	fixed-length	6.81	0.4385	1.265s
Ours	*SLP-CE*	<u>5.58</u>	**0.4961**	0.948s
	SLP-Regression	**5.36**	<u>0.4636</u>	**0.930s**

Table 5. Speech length modeling

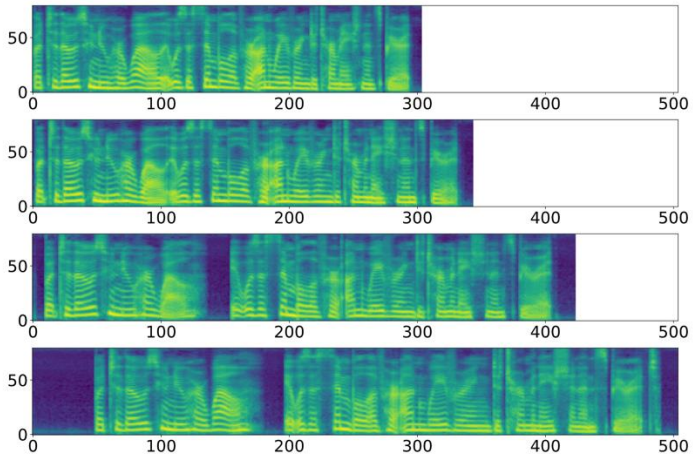


Figure 2. Speech rate controllability

1. Variable length delivers better quality and faster performance.
2. It also allows for precise control over the speech rate.

RQ2. Key Aspects Behind The Scene (3/3)

Aligned Text-Speech Embeddings Improve Performances

Text Encoder	Neural Audio Codec	WER ↓	CER ↓	SIM-o ↑	SIM-r ↑
U (ByT5-base)	*U* (Mel-VAE)	6.22	3.82	**0.5482**	0.5945
J (SpeechT5)	*U* (Mel-VAE)	3.07	1.15	0.5423	0.5858
U (ByT5-base)	*J* (Mel-VAE++)	3.11	1.17	0.5323	0.5965
J (SpeechT5)	*J* (Mel-VAE++)	**2.99**	**1.06**	0.5364	**0.5982**

Table 6. English-only *cross-sentence* Task for Mel-VAE++
(U refers to **U**nimodal and J refers to **J**ointly-trained)

RQ2. Key Aspects Behind The Scene (3/3)

Aligned Text-Speech Embeddings Improve Performances

Text Encoder	Neural Audio Codec	WER ↓	CER ↓	SIM-o ↑	SIM-r ↑
U (ByT5-base)	*U* (Mel-VAE)	6.22	3.82	**0.5482**	0.5945
J (SpeechT5)	*U* (Mel-VAE)	3.07	1.15	0.5423	0.5858
U (ByT5-base)	*J* (Mel-VAE++)	3.11	1.17	0.5323	0.5965
J (SpeechT5)	*J* (Mel-VAE++)	**2.99**	**1.06**	0.5364	**0.5982**

Table 6. English-only *cross-sentence* Task for Mel-VAE++
(U refers to **U**nimodal and J refers to **J**ointly-trained)

$$\mathcal{L}(\psi) = \mathcal{L}_{NAC}(\psi) + \lambda \mathcal{L}_{LM}(\psi), \quad \mathcal{L}_{LM}(\psi) = -\log p_{\phi}(\boldsymbol{x}|f(\boldsymbol{z}_{speech}))$$

RQ2. Key Aspects Behind The Scene (3/3)

Aligned Text-Speech Embeddings Improve Performances

Text Encoder	Neural Audio Codec	WER ↓	CER ↓	SIM-o ↑	SIM-r ↑
U (ByT5-base)	*U* (Mel-VAE)	6.22	3.82	**0.5482**	0.5945
J (SpeechT5)	*U* (Mel-VAE)	3.07	1.15	0.5423	0.5858
U (ByT5-base)	*J* (Mel-VAE++)	3.11	1.17	0.5323	0.5965
J (SpeechT5)	*J* (Mel-VAE++)	**2.99**	**1.06**	0.5364	**0.5982**

Table 6. English-only *cross-sentence* Task for Mel-VAE++
(U refers to **U**nimodal and J refers to **J**ointly-trained)

1. Jointly trained and aligned text-speech embeddings perform better.

RQ2. Key Aspects Behind The Scene (3/3)

Aligned Text-Speech Embeddings Improve Performances

Text Encoder	Neural Audio Codec	WER ↓	CER ↓	SIM-o ↑	SIM-r ↑
U (ByT5-base)	*U* (Mel-VAE)	6.22	3.82	**0.5482**	0.5945
J (SpeechT5)	*U* (Mel-VAE)	3.07	1.15	0.5423	0.5858
U (ByT5-base)	*J* (Mel-VAE++)	3.11	1.17	0.5323	0.5965
J (SpeechT5)	*J* (Mel-VAE++)	**2.99**	**1.06**	0.5364	**0.5982**

Table 6. English-only *cross-sentence* Task for Mel-VAE++
(U refers to **U**nimodal and J refers to **J**ointly-trained)

1. Jointly trained and aligned text-speech embeddings perform better.

2. Fine-tuning the speech latent is sufficient, obviating the need for expensive LLM encoder fine-tuning.

Ablation Study

Model	WER ↓	SIM-o ↑	SIM-r ↑	Inference Time ↓	Codec PESQ ↑	Codec ViSQOL ↑
DiTTo-*local-adaln*	3.38	0.5263	0.5673	0.937s	2.95	4.66
DiTTo-*uvit-skip*	3.17	0.5456	0.5848	0.940s		
DiTTo-*no-skip*	3.30	0.5304	0.5727	0.905s		
DiTTo-*no-pooled-text*	3.00	0.5410	0.5791	0.912s		
DiTTo-*no-rvq-decoding*	2.97	0.5468	**0.5883**	**0.894s**		
DiTTo-*mls* (from Table 4)	**2.93**	0.5467	0.5877	0.903s		
DiTTo-*encodec*	4.19	0.5105	0.5460	n/a	2.59	4.26
DiTTo-*dac-24k*	7.21	**0.5478**	0.5545	n/a	**4.37**	**4.91**
DiTTo-*dac-44k*	14.58	0.5391	0.5597	n/a	3.74	4.85

Table 7. English-only *cross-sentence* Task

Ablation Study

Model	WER ↓	SIM-o ↑	SIM-r ↑	Inference Time ↓	Codec PESQ ↑	Codec ViSQOL ↑
DiTTo-*local-adaln*	3.38	0.5263	0.5673	0.937s	2.95	4.66
DiTTo-*uvit-skip*	3.17	0.5456	0.5848	0.940s		
DiTTo-*no-skip*	3.30	0.5304	0.5727	0.905s		
DiTTo-*no-pooled-text*	3.00	0.5410	0.5791	0.912s		
DiTTo-*no-rvq-decoding*	2.97	0.5468	**0.5883**	**0.894s**		
DiTTo-*mls* (from Table 4)	**2.93**	0.5467	0.5877	0.903s		
DiTTo-*encodec*	4.19	0.5105	0.5460	n/a	2.59	4.26
DiTTo-*dac-24k*	7.21	**0.5478**	0.5545	n/a	**4.37**	**4.91**
DiTTo-*dac-44k*	14.58	0.5391	0.5597	n/a	3.74	4.85

Table 7. English-only *cross-sentence* Task

1. Extensive model architecture searches validate the effectiveness of DiTTo.

Ablation Study

Model	WER ↓	SIM-o ↑	SIM-r ↑	Inference Time ↓	Codec PESQ ↑	Codec ViSQOL ↑
DiTTo-*local-adaln*	3.38	0.5263	0.5673	0.937s		
DiTTo-*uvit-skip*	3.17	0.5456	0.5848	0.940s		
DiTTo-*no-skip*	3.30	0.5304	0.5727	0.905s	2.95	4.66
DiTTo-*no-pooled-text*	3.00	0.5410	0.5791	0.912s		
DiTTo-*no-rvq-decoding*	2.97	0.5468	**0.5883**	**0.894s**		
DiTTo-*mls* (from Table 4)	**2.93**	0.5467	0.5877	0.903s		
DiTTo-*encodec*	4.19	0.5105	0.5460	n/a	2.59	4.26
DiTTo-*dac-24k*	7.21	**0.5478**	0.5545	n/a	**4.37**	**4.91**
DiTTo-*dac-44k*	14.58	0.5391	0.5597	n/a	3.74	4.85

Table 7. English-only *cross-sentence* Task

1. Extensive model architecture searches validate the effectiveness of DiTTo.

2. The compactness of the time-domain for speech latent is essential.

Thank you ☺

ditto-tts.github.io