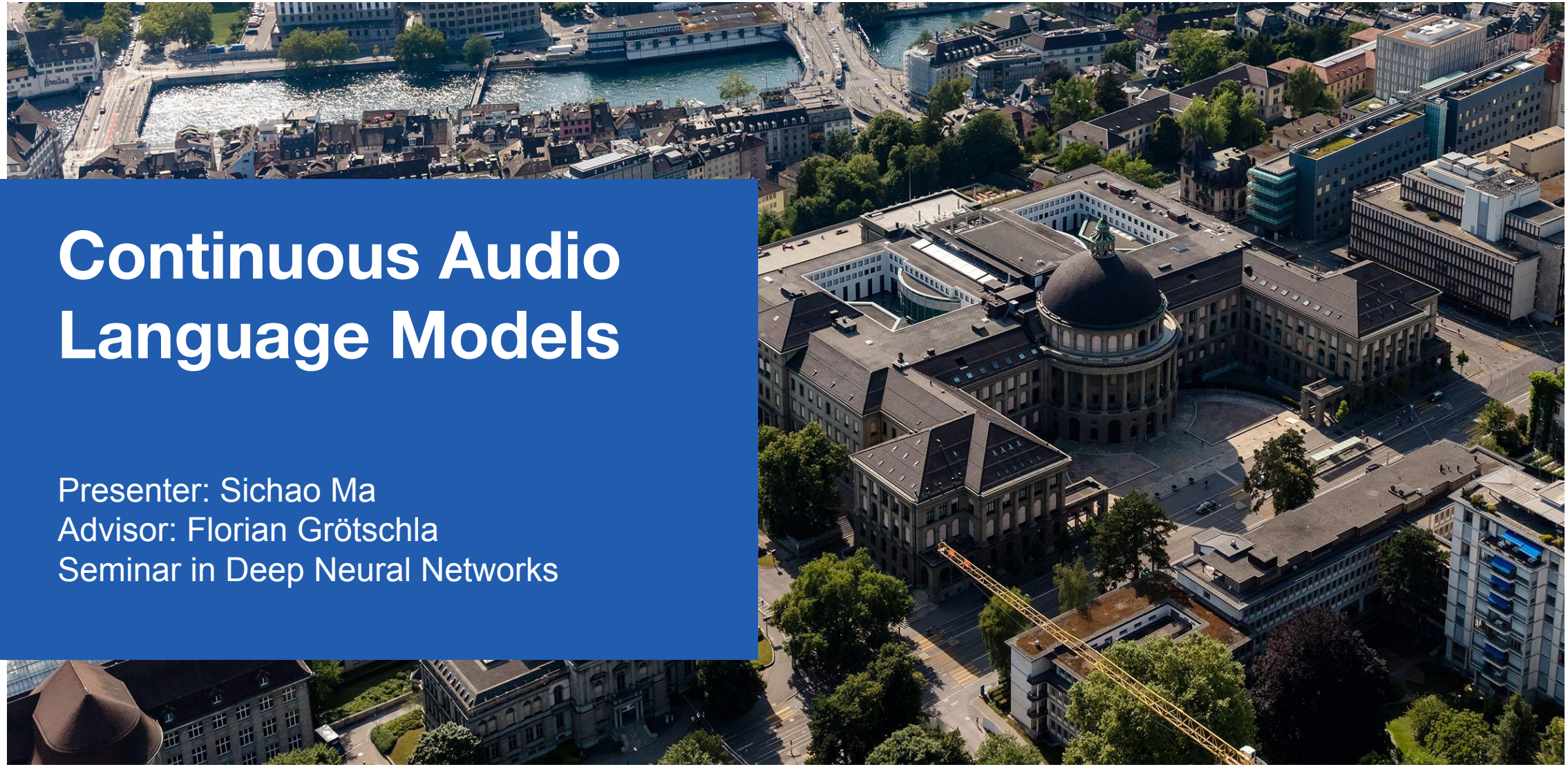
ETH zürich



Continuous Audio Language Models

Presenter: Sichao Ma
Advisor: Florian Grötschla
Seminar in Deep Neural Networks

CONTINUOUS AUDIO LANGUAGE MODELS

Simon Rouard*
Kyutai
UMR STMS
IRCAM-CNRS Sorbonne Univ.
simon@kyutai.org

Manu Orsini*
Kyutai
manu@kyutai.org

Axel Roebel
UMR STMS
IRCAM-CNRS Sorbonne Univ.

Neil Zeghidour
Kyutai

Alexandre Défossez
Kyutai
alex@kyutai.org

- What is **continuous** model?
- Why not using **continuous** model before?
- Why using **continuous** model now?

The Story from Discrete to Continuous

Continuous Domain

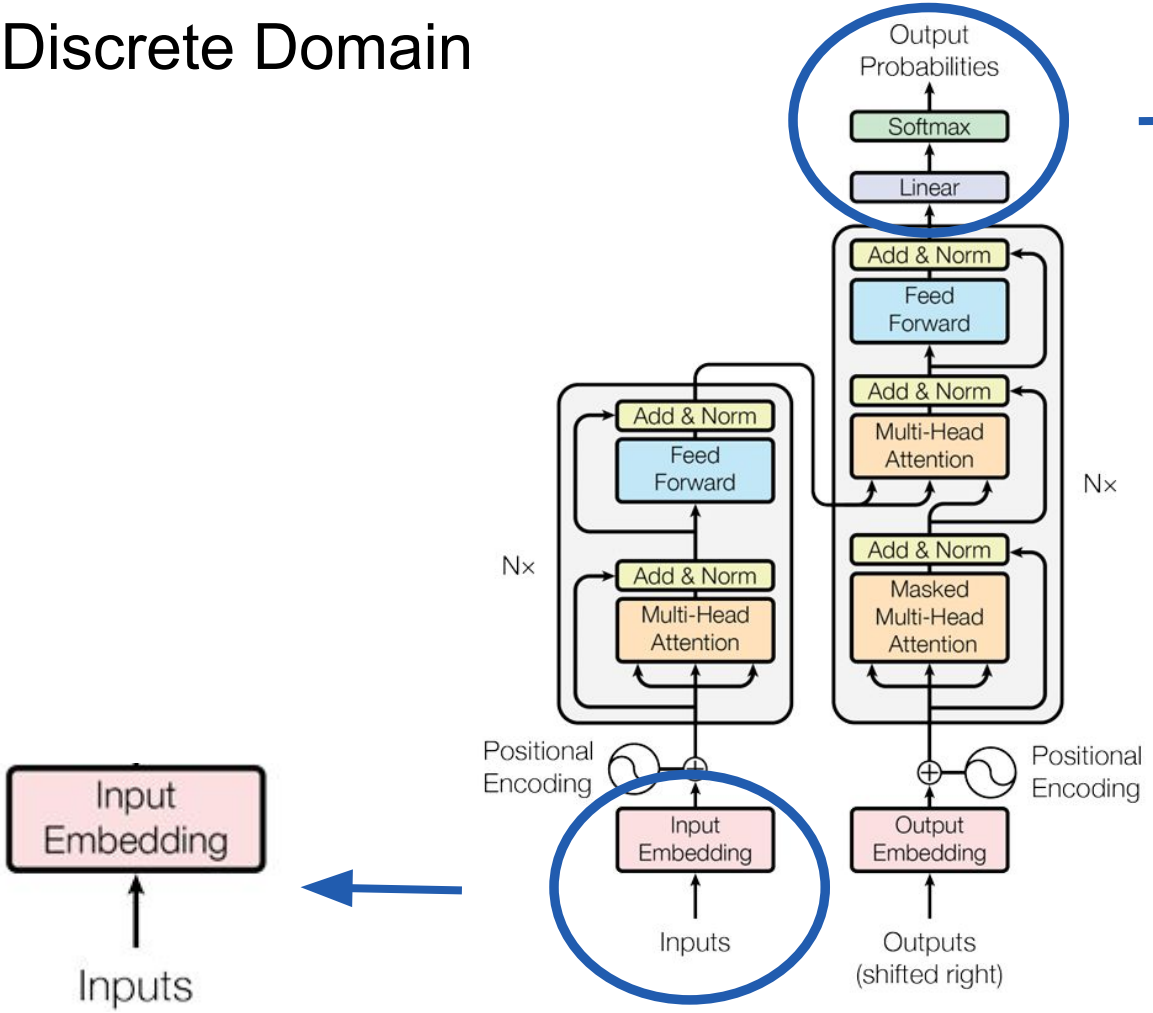
Discrete Domain

Hello World!

你好，世界！

The Story from Discrete to Continuous

Discrete Domain



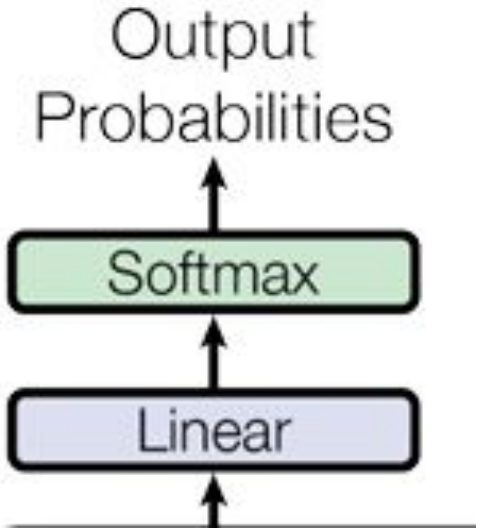


Figure 1: The Transformer - model architecture.

Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

The Story from Discrete to Continuous

VQ-VAE

Continuous latent space -> finite codebook

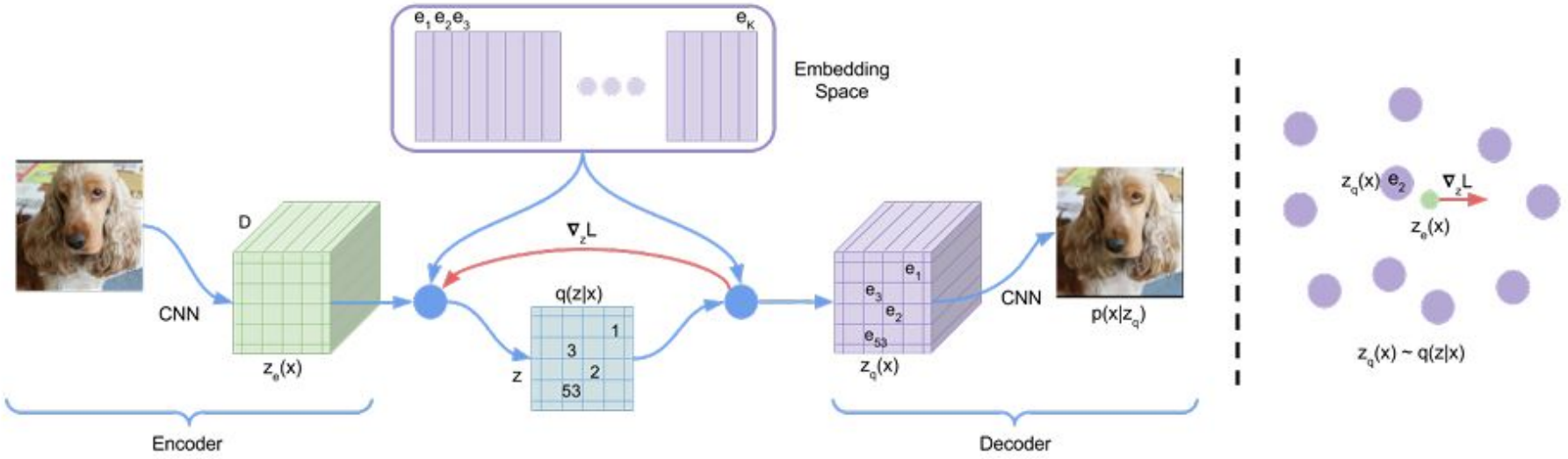
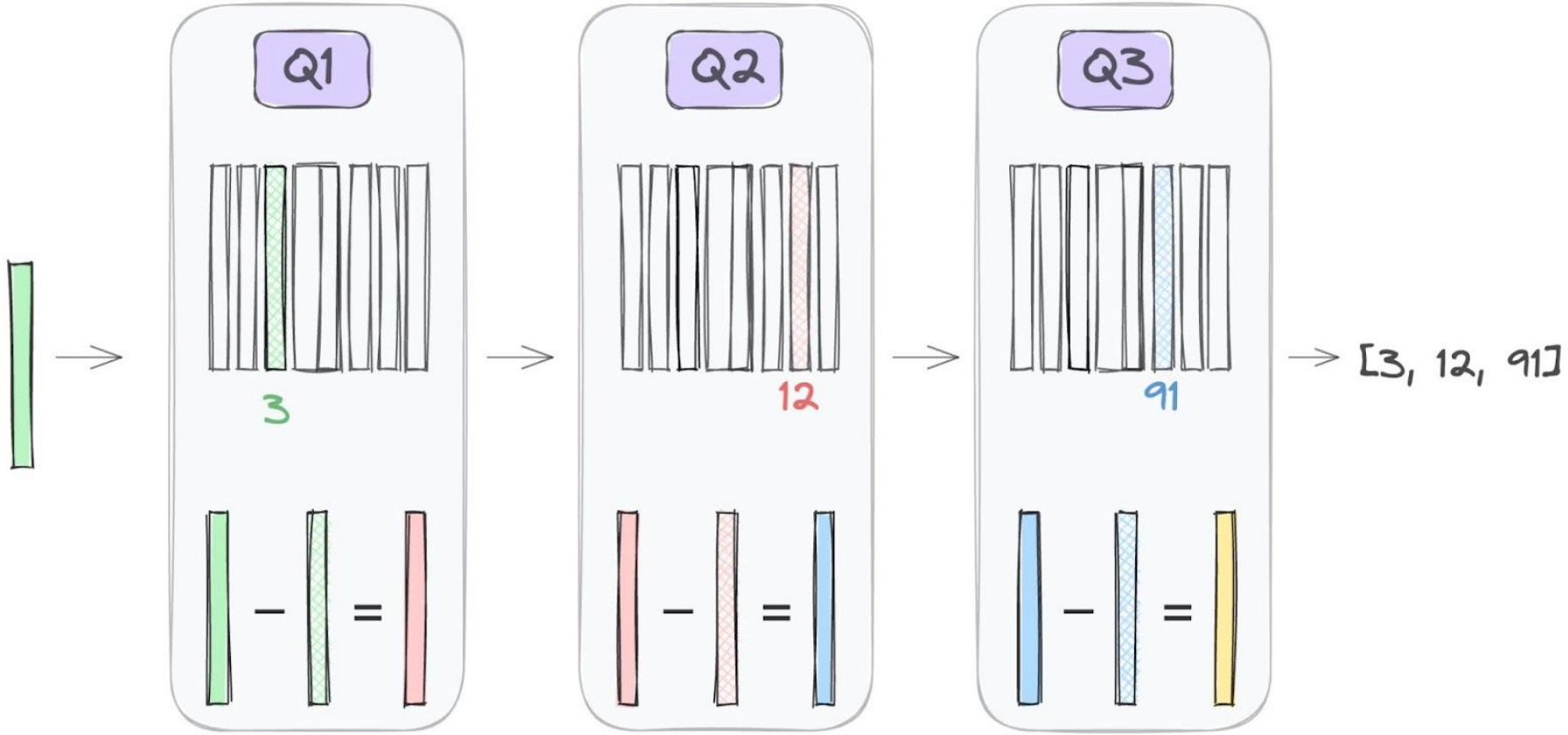


Figure 1: Left: A figure describing the VQ-VAE. Right: Visualisation of the embedding space. The output of the encoder $z(x)$ is mapped to the nearest point e_2. The gradient $\nabla_z L$ (in red) will push the encoder to change its output, which could alter the configuration in the next forward pass.

Van Den Oord, Aaron, and Oriol Vinyals. "Neural discrete representation learning." *Advances in neural information processing systems* 30 (2017).

The Story from Discrete to Continuous

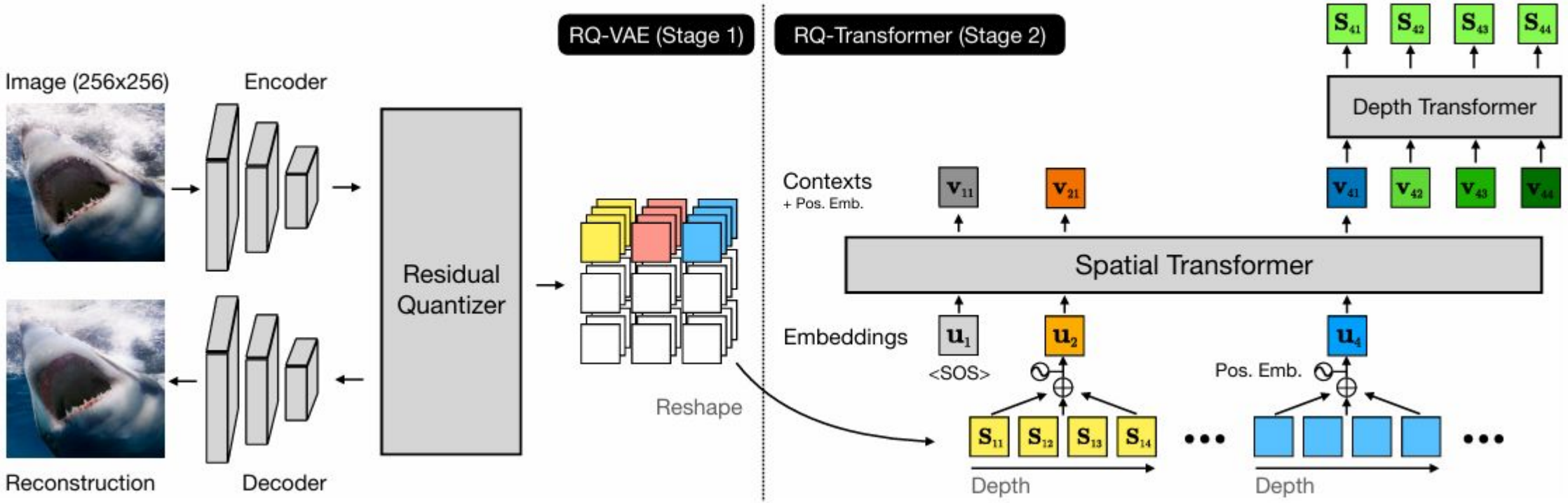
Residual Vector Quantization (RVQ)



source: https://www.assemblyai.com/blog/what-is-residual-vector-quantization

The Story from Discrete to Continuous

RQ-Transformer



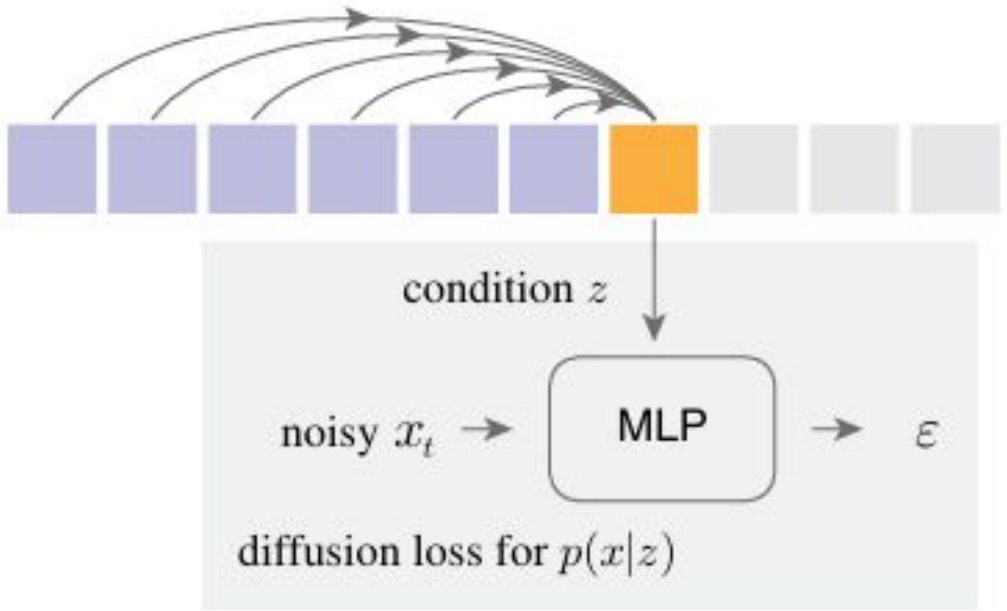
Lee, Doyup, et al. "Autoregressive image generation using residual quantization." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

The Story from Discrete to Continuous

Higher fidelity requires less compression.

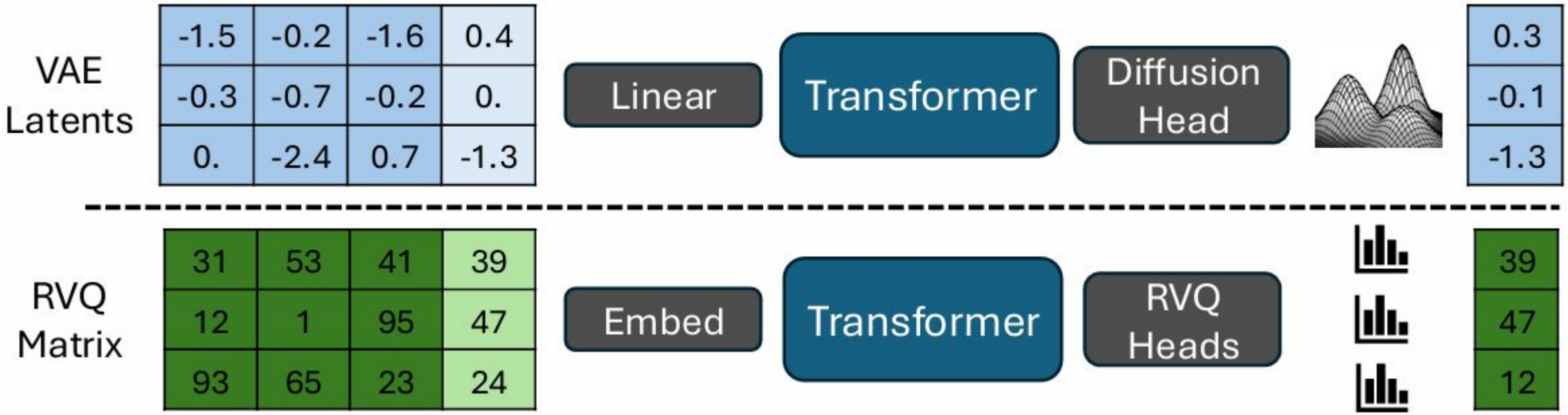
The Story from Discrete to Continuous

masked autoregressive (MAR) model



Li, Tianhong, et al. "Autoregressive image generation without vector quantization." *Advances in Neural Information Processing Systems* 37 (2024): 56424-56445.

The Story from Discrete to Continuous

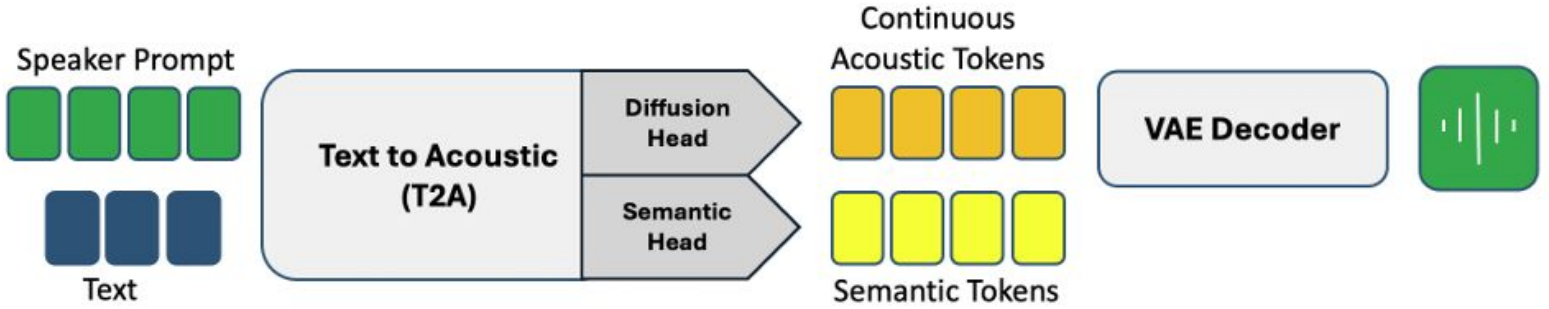


Turetzky, Arnon, et al. "Speech Synthesis From Continuous Features Using Per-Token Latent Diffusion." *arXiv preprint arXiv:2410.16048* (2024).

Audio Language Model with MAR

SALAD

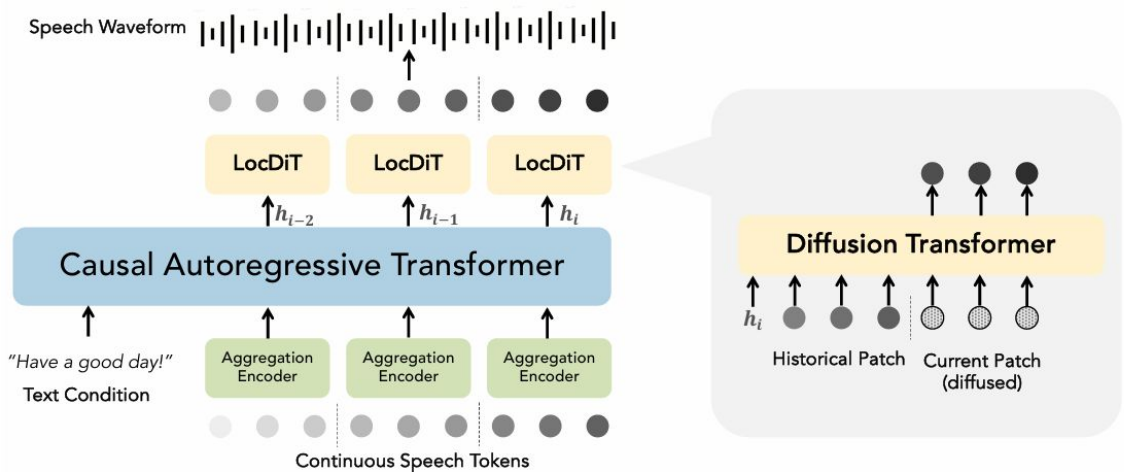
text-to-speech



Turetzky, Arnon, et al. "Speech Synthesis From Continuous Features Using Per-Token Latent Diffusion." *arXiv preprint arXiv:2410.16048* (2024).

DiTAR

text-to-speech



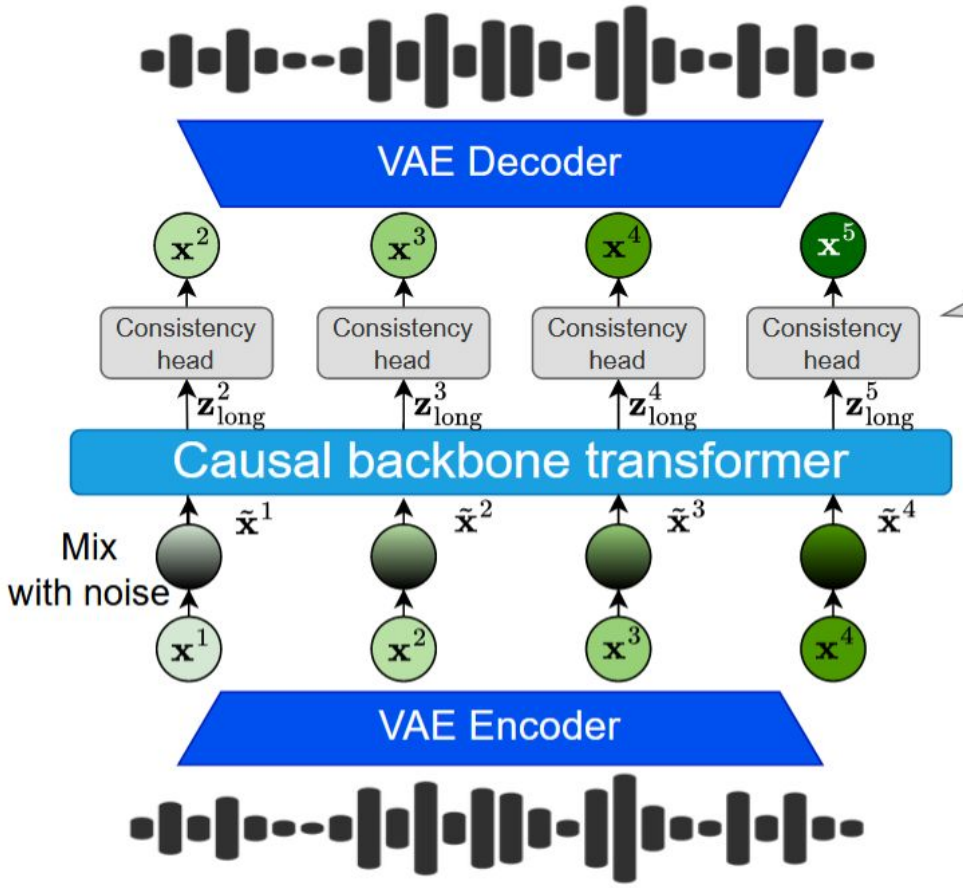
Jia, Dongya, et al. "Ditar: Diffusion transformer autoregressive modeling for speech generation." *arXiv preprint arXiv:2502.03930* (2025).

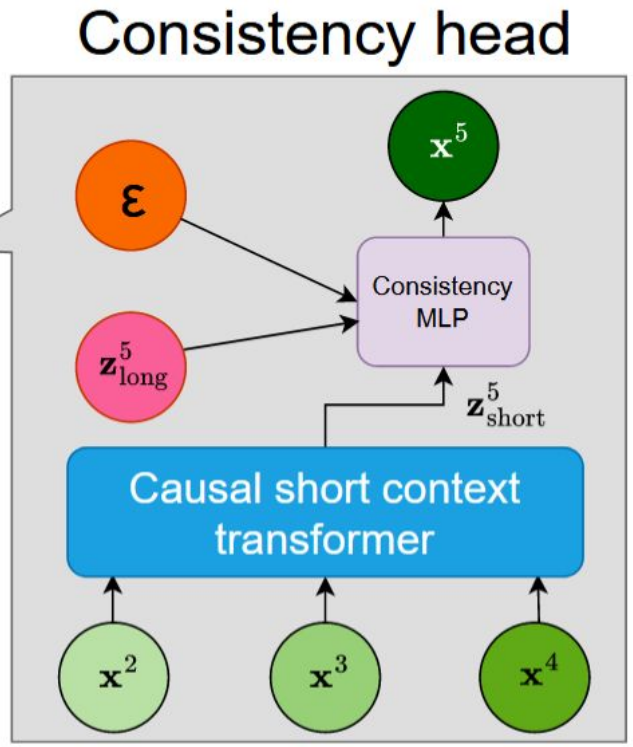
CONTINUOUS AUDIO LANGUAGE MODELS

CALM

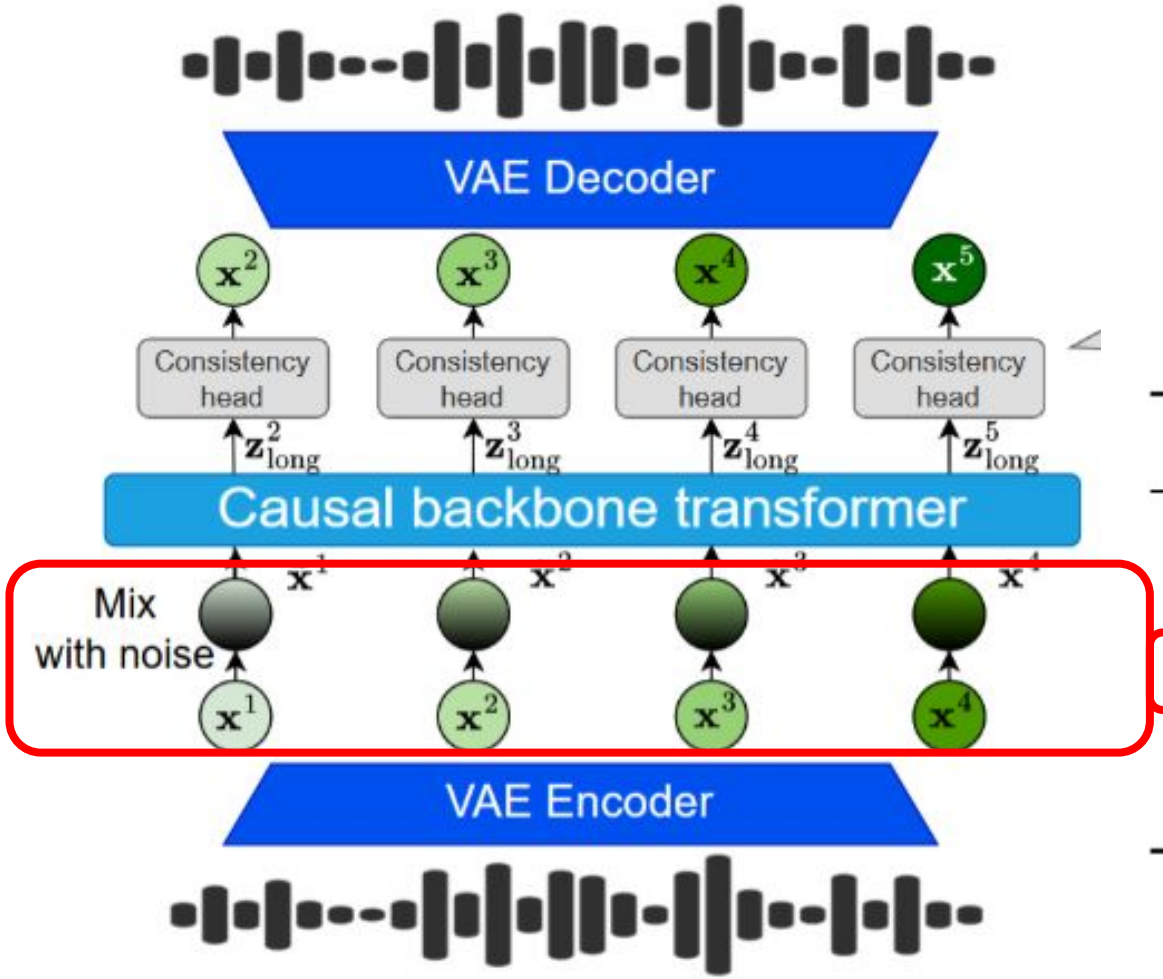
text-to-speech,
speech continuation

text-to-music,
music continuation

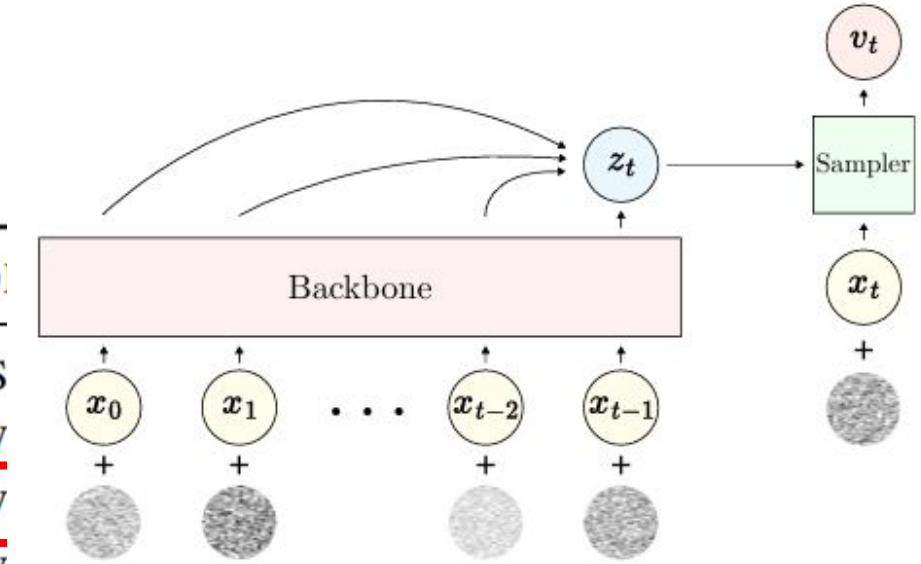




Noise Injection → solving error accumulation at inference



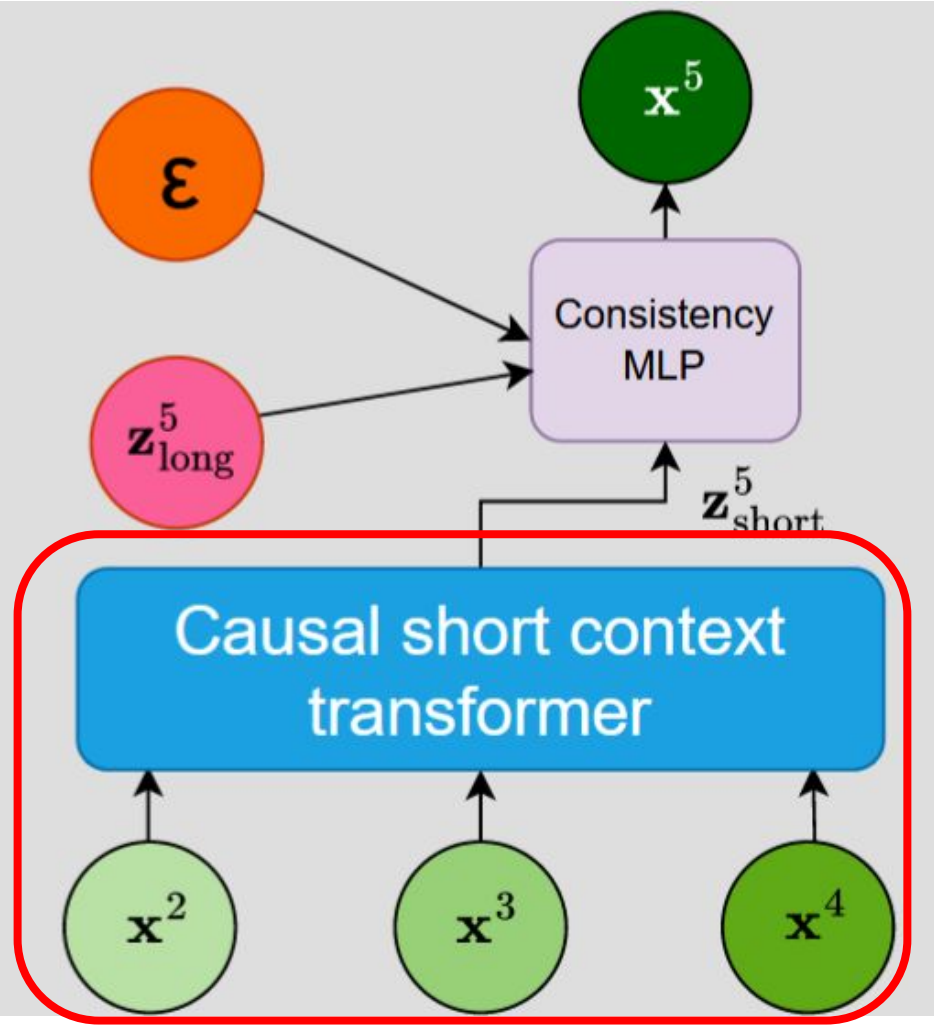
***FAD**: distribution-level distance between ...g space.*



Mo...	FAD (↓)
Bas...	**0.93 ± 0.06**
W...	1.32 ± 0.09
W...	1.63 ± 0.11
W/O SHORT CONTEXT TRANSFORMER	4.03 ± 0.16
W/O ANY OF THE ABOVE	8.38 ± 0.17

Pasini, Marco, et al. "Continuous autoregressive models with noise augmentation avoid error accumulation." *arXiv preprint arXiv:2411.18447* (2024).

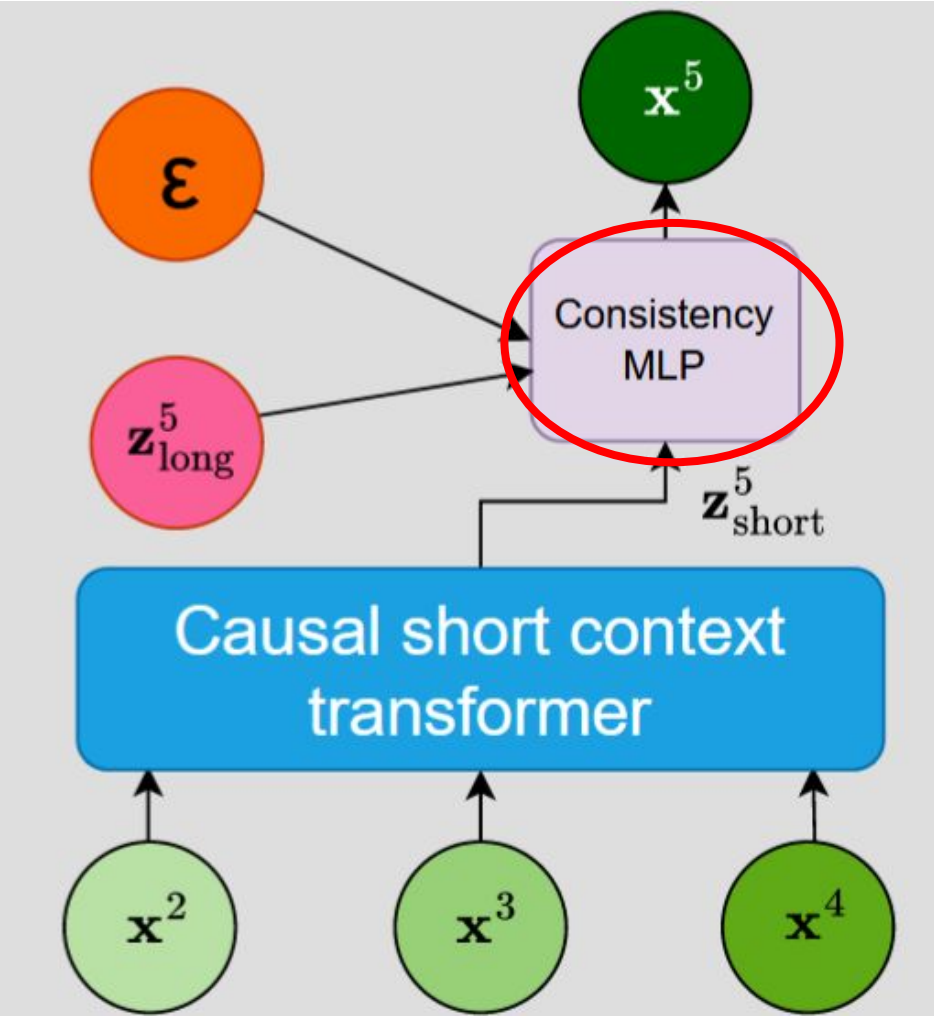
Short-Context Transformer → supply local, high-resolution context



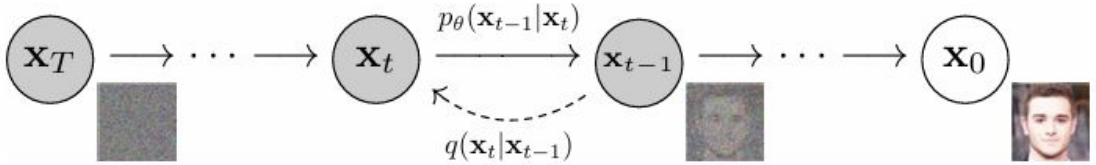
K in short-context Transformer	FAD
5	0.85 ± 0.05
10	0.76 ± 0.04
20	0.73 ± 0.04
40	0.78 ± 0.04

MODEL VARIANT	FAD (↓)
BASE (CALM - CONSISTENCY - 4 STEPS)	$\mathbf{0.93 \pm 0.06}$
W/O HEAD BATCH MULTIPLIER	1.32 ± 0.09
W/O NOISE AUGMENTATION	1.63 ± 0.11
W/O SHORT CONTEXT TRANSFORMER	4.03 ± 0.16
W/O ANY OF THE ABOVE	8.38 ± 0.17

Consistency Model ⟶ accelerating inference

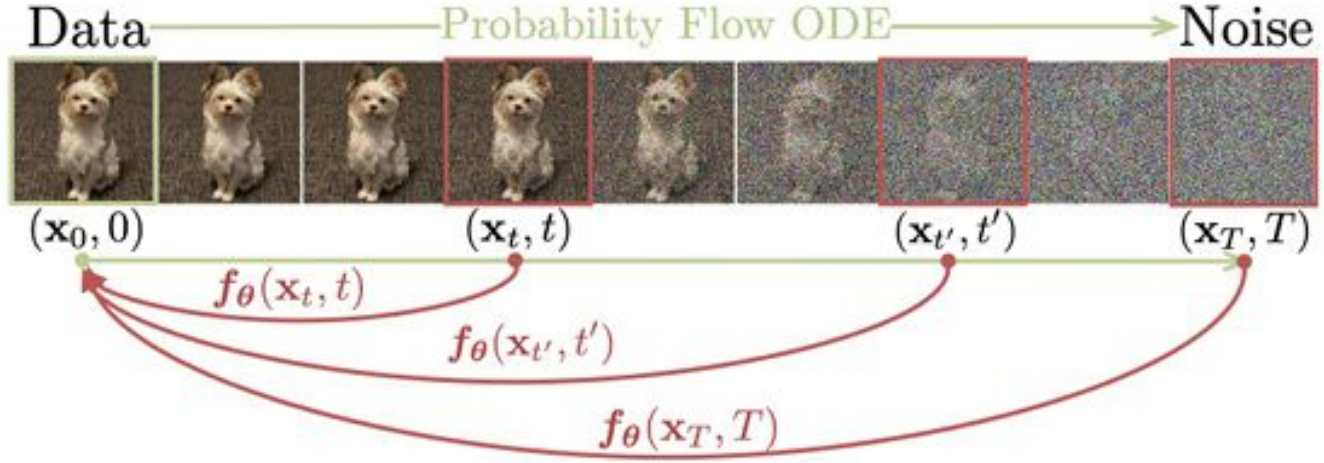


Diffusion Models



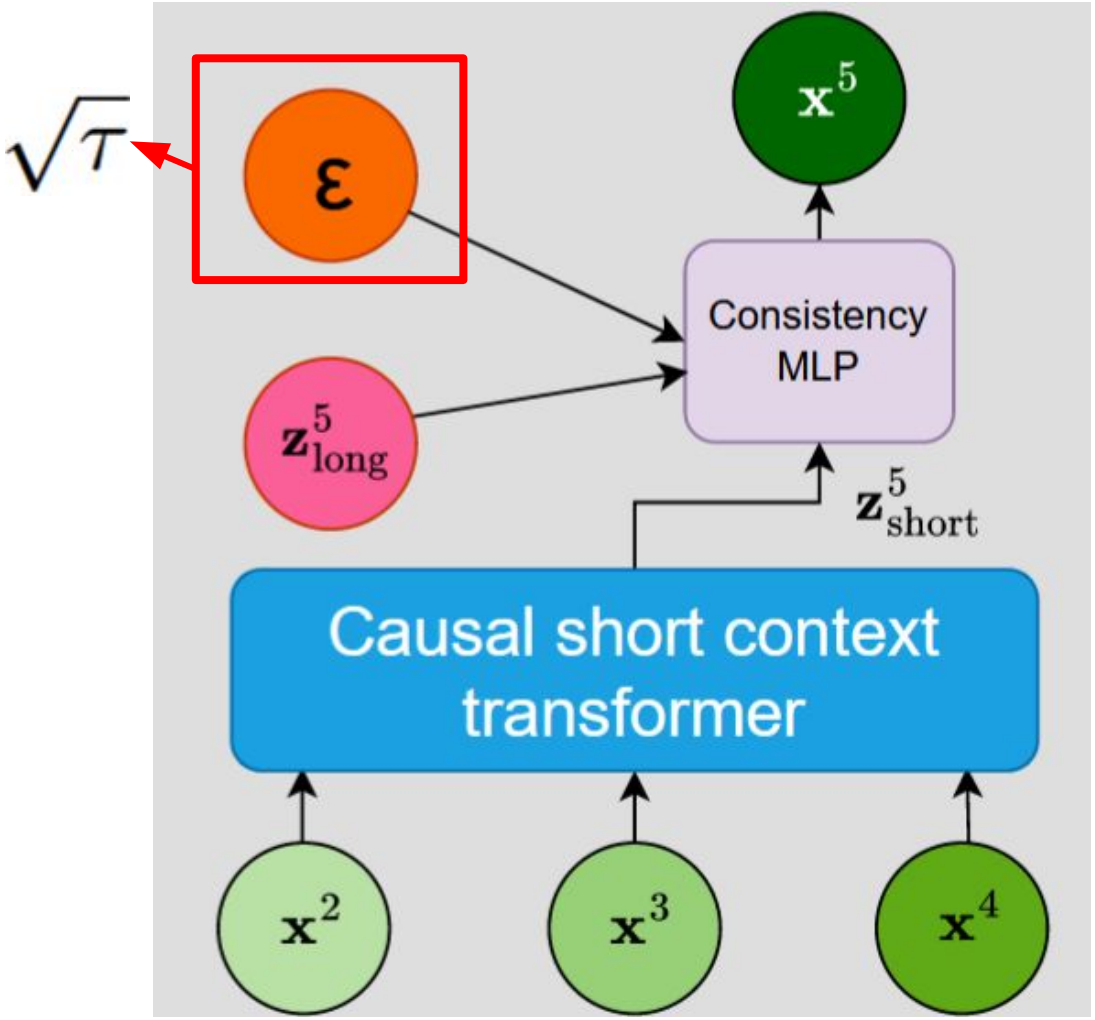
Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." *Advances in neural information processing systems* 33 (2020): 6840-6851.

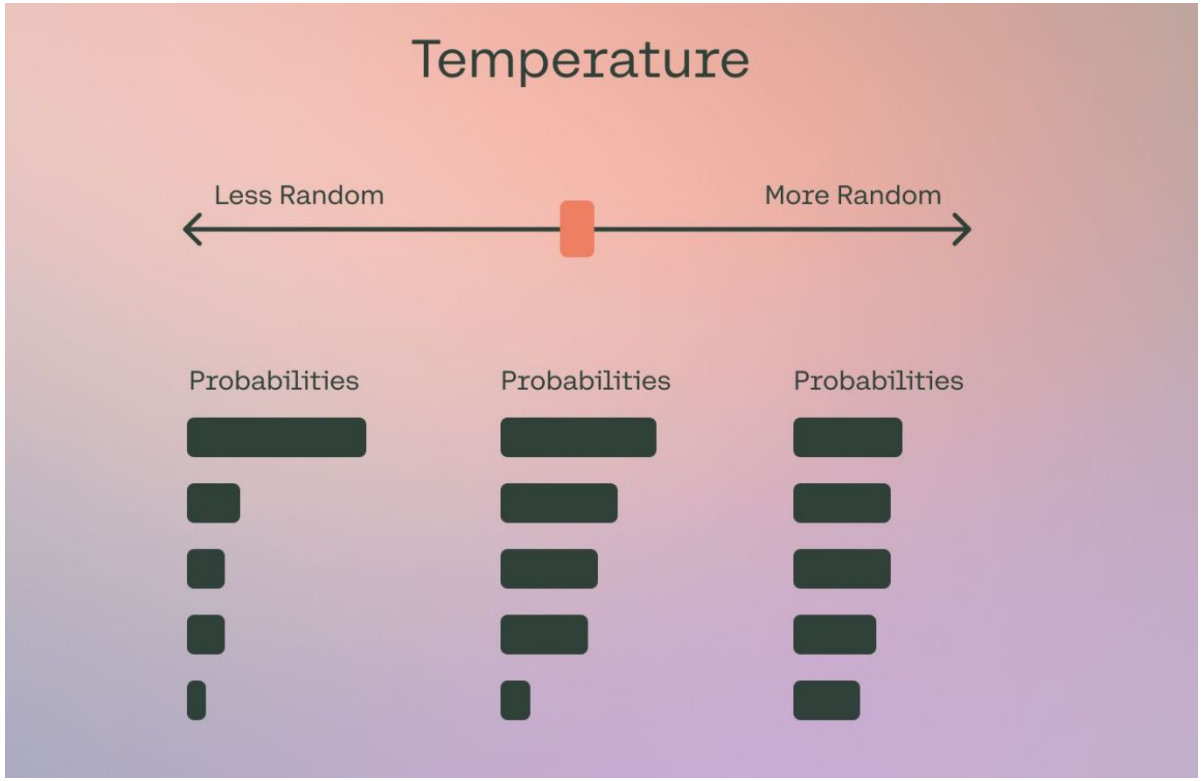
Consistency Models



Song, Yang, et al. "Consistency models." (2023).

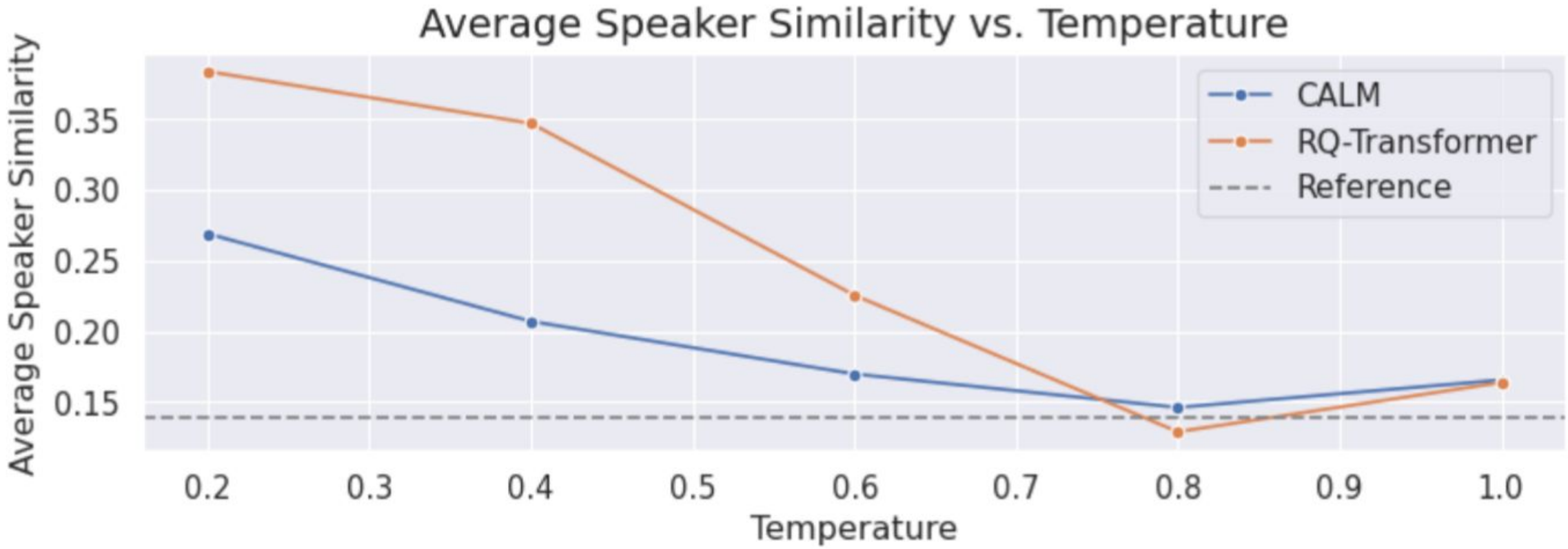
Gaussian Temperature sampling → approximates temperature sampling



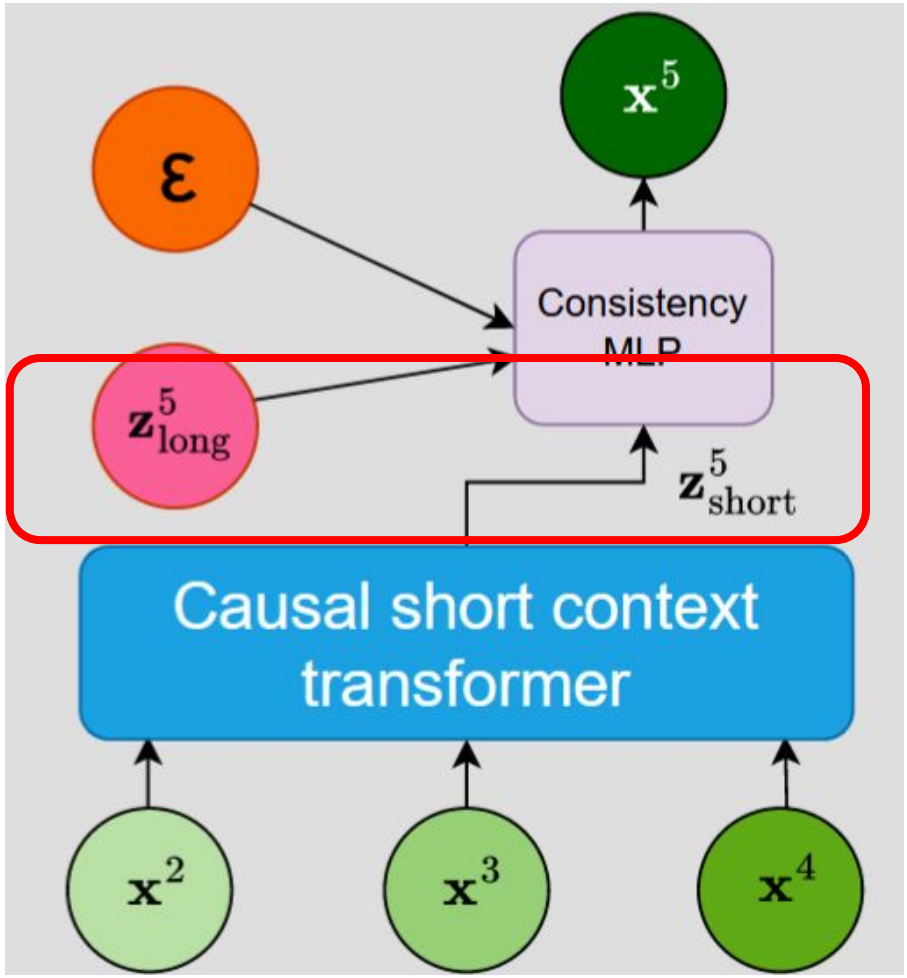


https://cohere.com/blog/llm-parameters-best-outputs-language-ai

Gaussian Temperature sampling ⟶ approximates temperature sampling

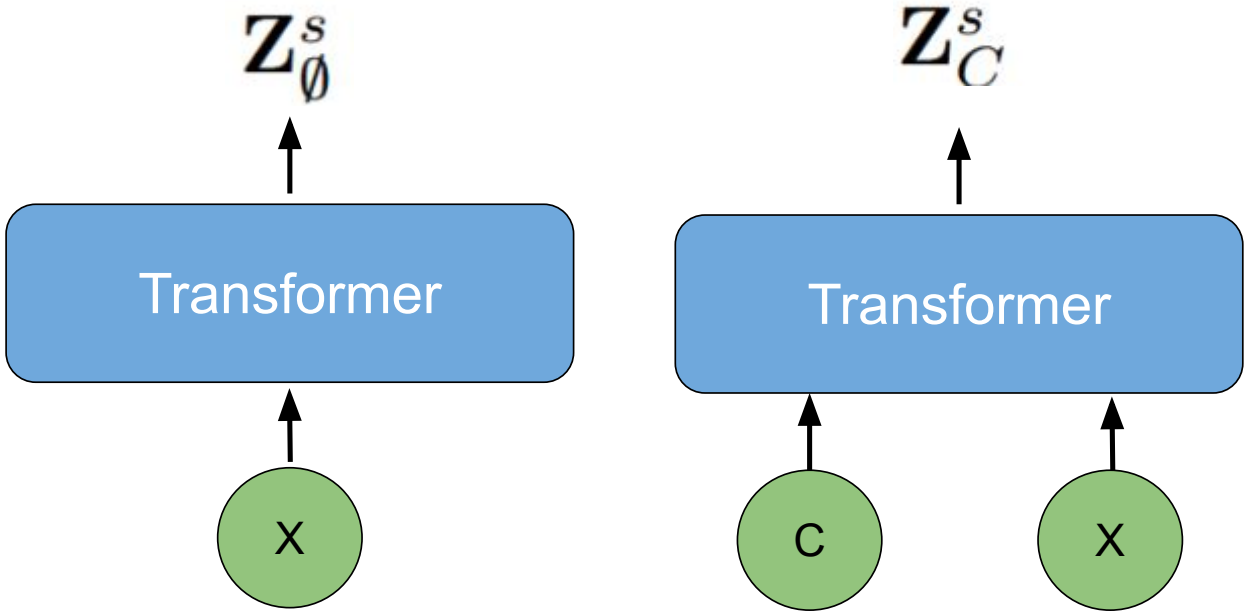


Latent Classifier Free Guidance

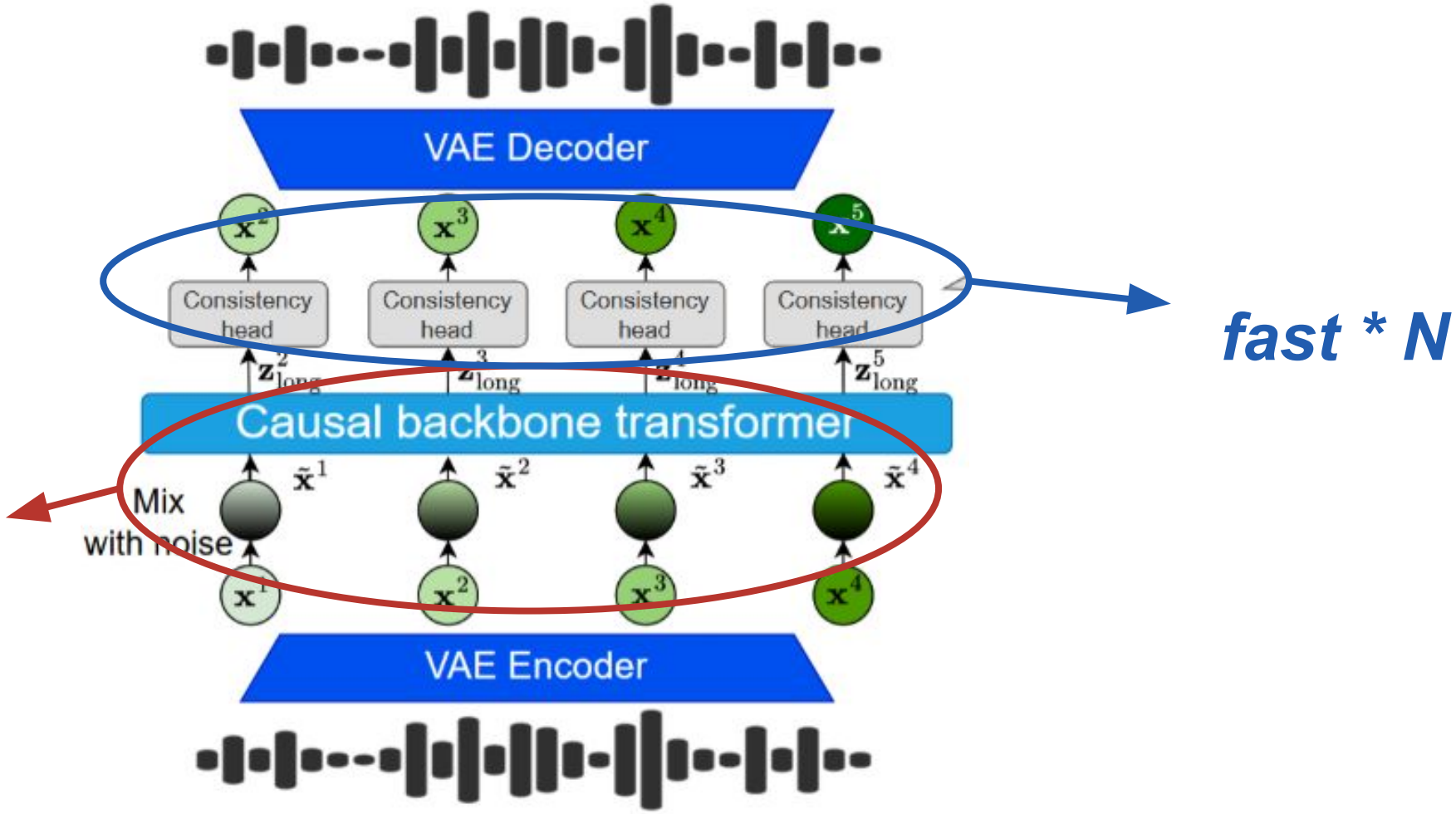


improve conditioned generation quality

$$\bar{\mathbf{Z}}^s_{\text{CFG}} = \mathbf{Z}^s_{\emptyset} + \alpha(\mathbf{Z}^s_C - \mathbf{Z}^s_{\emptyset})$$



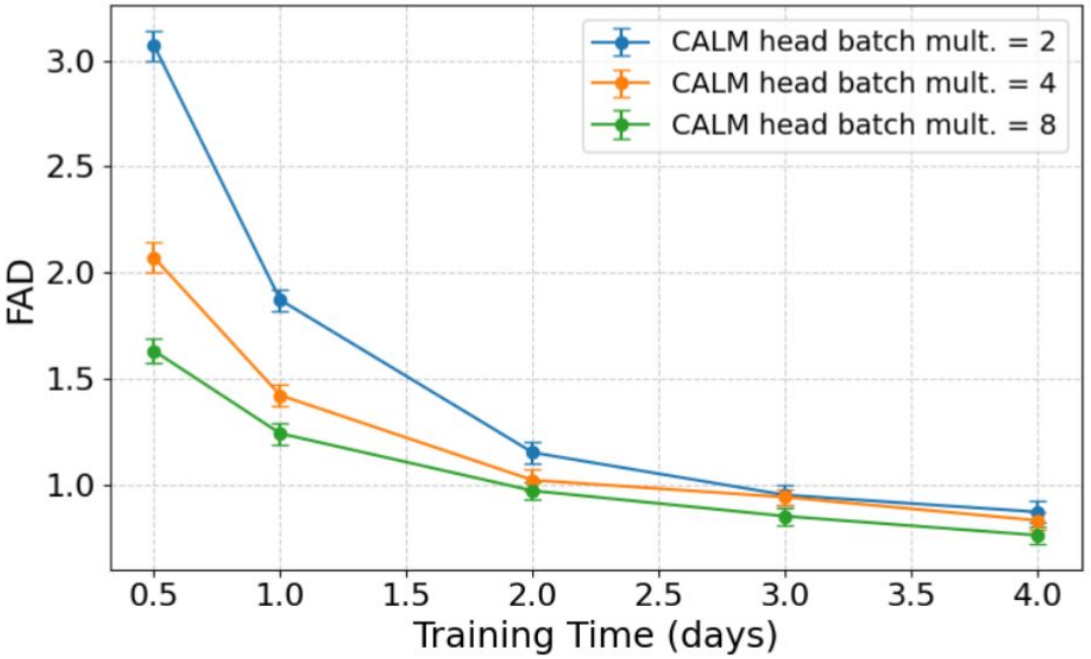
Head Batch Multiplier ⟶ accelerating training



*fast * N*

*slow * 1*

Head Batch Multiplier ⟶ accelerating training



Model Variant	FAD (↓)
Base (CALM - Consistency - 4 steps)	**0.93 ± 0.06**
w/o Head Batch Multiplier	1.32 ± 0.09
w/o Noise Augmentation	1.63 ± 0.11
w/o Short Context Transformer	4.03 ± 0.16
w/o Any of the above	8.38 ± 0.17

EXPERIMENTS AND RESULTS

Speech

- Speech Continuation
- Text-to-Speech Generation

Music

- Music Continuation
- Text-to-Music Generation

EXPERIMENTS AND RESULTS ---- Speech Continuation

prompt

Model Type	Sampling temperature	Overall Speedup (↑)	Sampler Speedup (↑)	% Time in Sampler (↓)	PPX (↓)	VERT (↓)	Acoustic Quality(↑)	Meaningfulness Elo (↑)	Meaningfulness Rank (↓)
Reference	–	–	–	–	20.2	25.2	4.02 ± 0.11	2180 ± 30	–
RQ-transformer 8 RVQ	1.0	×1.0	×1.0	26.7%	52.4	36.3	2.42 ± 0.12	1841 ± 25	4
RQ-transformer 8 RVQ	0.8	×1.0	×1.0	26.7%	26.8	33.1	2.75 ± 0.14	1870 ± 30	3
CALM - Consistency - 1 step	1.0	×1.3	×12.3	2.9%	42.9	34.3	2.82 ± 0.13	1947 ± 28	2
CALM - Consistency - 1 step	0.8	×1.3	×12.3	2.9%	**23.8**	**31.2**	$\mathbf{3.45 \pm 0.14}$	$\mathbf{2023 \pm 27}$	**1**

Model Size: 2.2B

PPX: semantic meaningfulness
VERT: diversity
Acoustic Quality: Acoustic Quality by human evaluation
Elo: human preference strength in pairwise comparisons

EXPERIMENTS AND RESULTS ---- Text-to-Speech

Model	Num. parameters	WER (↓)	CER (↓)	SIM (↑)	Acoustic Quality (↑)	Speaker SIM Human ELO (↑)
Reference	–	2.23	–	0.69	61.8 ± 2.4	1953 ± 24
Reference (with VAE)	–	–	–	0.57	–	–
F5 TTS (NFE=32) (Chen et al., 2025)	336M	2.42	–	0.66	54.7 ± 2.8	2032 ± 18
DSM (16 RVQ CFG=3 w.r.t text and audio prompt) (Zeghidour et al., 2025)	750M	1.95	–	**0.67**	60.2 ± 2.4	**2112 ± 20**
DiTAR (NFE=10) (Jia et al., 2025)	600M	2.39	–	**0.67**	–	–
SALAD (NFE=20) (Turetzky et al., 2024)	350M	–	0.74	0.54	–	–
CALM w/ LSD (NFE=1, CFG=1.5 w.r.t text)	**313M**	**1.81**	**0.57**	0.52	**61.1 ± 2.3**	1966 ± 23

Model Size: 302M

WER: Word Error Rate
CER: Character Error Rate
SIM: Speaker Similarity

EXPERIMENTS AND RESULTS ------- Music Continuation

prompt

Model	Overall Speedup (↑)	Sampler Speedup (↑)	% time in Sampler (↓)	FAD (↓)	Acoustic Quality (↑)	Enjoyment Elo (↑)	Enjoyment Rank (↓)
Reference	–	–	–	–	3.84 ± 0.08	2166 ± 33	-
RQ-transformer 32 RVQ (baseline)	$\times$ 1.0	$\times$ 1.0	57.7%	1.06 ± 0.06	2.85 ± 0.07	1824 ± 29	4
RQ-transformer 16 RVQ	$\times$ 1.5	$\times$ 2.2	38.0%	1.43 ± 0.07	2.76 ± 0.07	1781 ± 29	5
CALM - Consistency - 1 Step	$\times$ **2.2**	$\times$ **19.3**	**6.6%**	0.83 ± 0.04	2.90 ± 0.07	1857 ± 28	2
CALM - Consistency - 4 Steps	$\times$ 1.9	$\times$ 5.4	20.1%	$\mathbf{0.71 \pm 0.05}$	$\mathbf{3.07 \pm 0.07}$	1847 ± 24	3
CALM - TrigFlow - 100 Steps	$\times$ 0.3	$\times$ 0.2	86.6%	$\mathbf{0.64 \pm 0.04}$	$\mathbf{3.12 \pm 0.07}$	$\mathbf{1921 \pm 29}$	**1**
MusicGen Medium	$\times$ 1.3	–	0.0%	1.72 ± 0.12	2.62 ± 0.07	1761 ± 33	6

Model Size: 1.35B

FAD: distance between real and generated audio in feature space

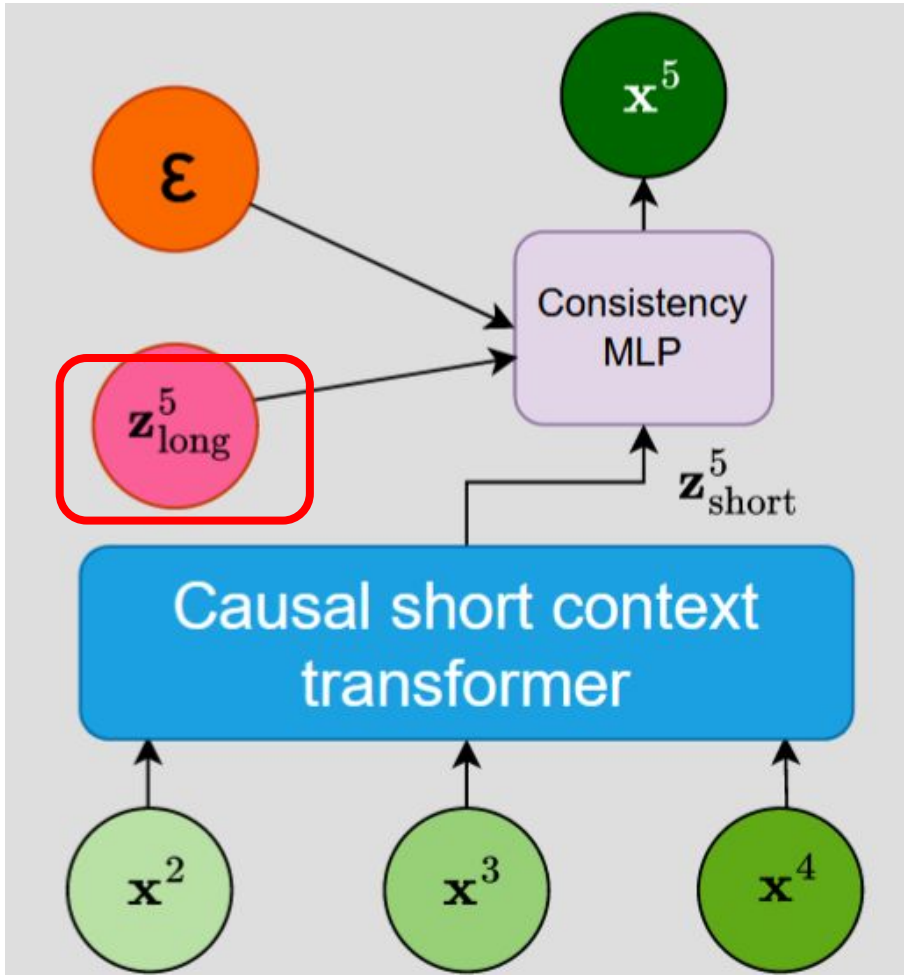
EXPERIMENTS AND RESULTS ---- Text-to-Music

Model	Overall Speedup (↑)	Sampler Speedup (↑)	% time in Sampler (↓)	FAD (↓)	Acoustic Quality (↑)	Enjoyment Elo (↑)	Enjoyment Rank (↓)
Reference	–	–	–	–	3.75 ± 0.15	2104 ± 24	–
RQ-transformer 32 RVQ (CFG=3)	× 1.0	× 1.0	57.7%	$\mathbf{0.93 \pm 0.07}$	$\mathbf{3.16 \pm 0.15}$	$\mathbf{1998 \pm 24}$	1
CALM - Consistency - 4 Steps (CFG=2)	× **1.9**	× 5.4	20.1%	$\mathbf{0.91 \pm 0.08}$	$\mathbf{3.11 \pm 0.14}$	$\mathbf{1998 \pm 24}$	1
MusicGen Medium	× 1.3	–	0.0%	1.93 ± 0.14	2.54 ± 0.14	1946 ± 25	3

Model Size: 1.35B

FAD: distance between real and generated audio in feature space

Latent Distillation ⟶ Pocket TTS



$$\bar{\mathbf{Z}}^s_{\text{CFG}} = \bar{\mathbf{Z}}^s_{\emptyset} + \alpha(\bar{\mathbf{Z}}^s_C - \bar{\mathbf{Z}}^s_{\emptyset})$$

$\bar{\mathbf{Z}}^s_{\text{distill}}$

24 layers ↓ *6 layers*

303M ↓ *100M*

Latent Distillation ⟶ Pocket TTS

Table 12: Comparison of TTS models on Librispeech test-clean.

Model	Param size (gen. only)	WER (↓)	Audio Quality (ELO) (↑)	Speaker Sim (ELO) (↑)	Faster than real-time CPU
F5-TTS (Chen et al., 2025)	336M	2.21	1949 ± 27	1946 ± 26	×
DSM	750M	1.84	1959 ± 25	2037 ± 21	×
Chatterbox Turbo	350M	3.24	2055 ± 23	2012 ± 22	×
Kokoro	82M	1.93	No voice cloning		✓
Pocket TTS (ours)	**100M**	**1.84**	2016 ± 25	1898 ± 26	✓

Reflection

Strengths

- Trying new paradigm
- Very technical

Limitations

- Using too many tricks
- Unclear experimental methodology and result details

Thank you for your attention!

Q&A

The Story from Discrete to Continuous

Discrete Domain

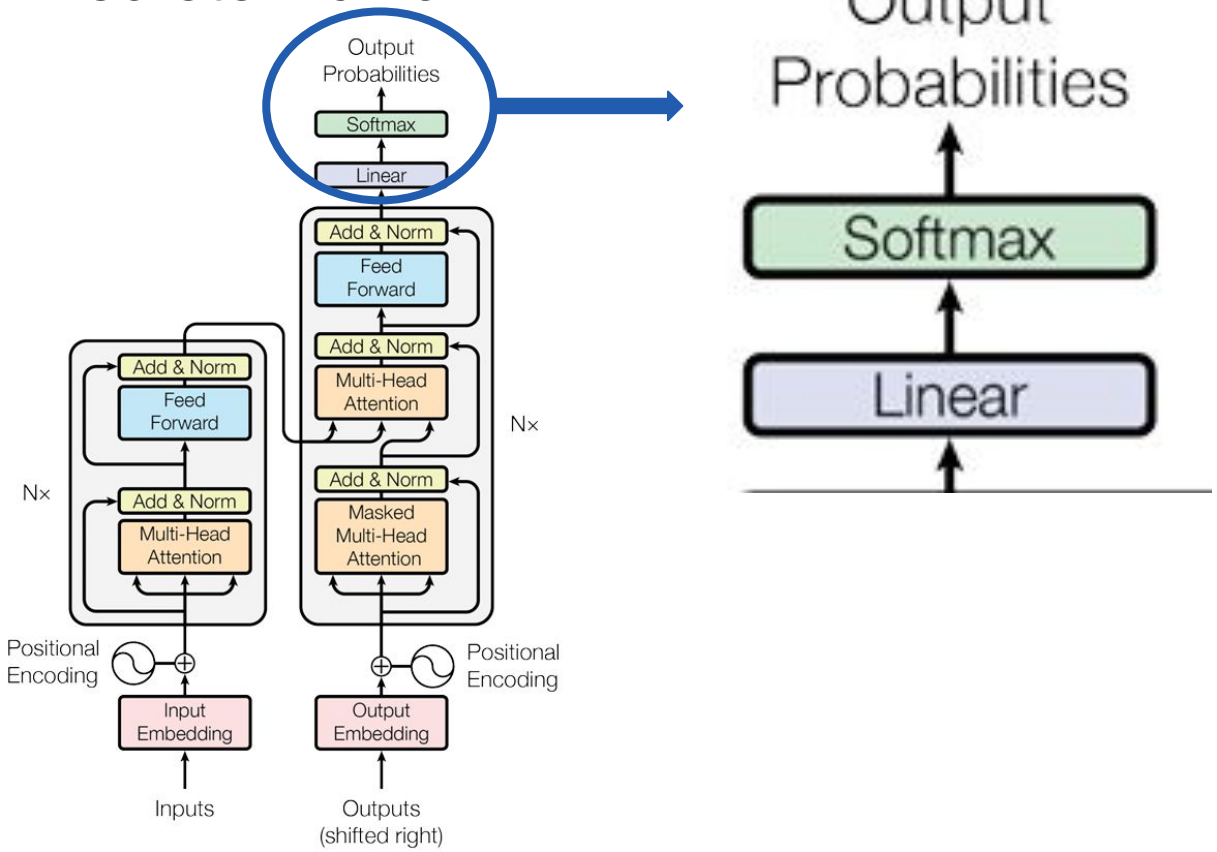


Figure 1: The Transformer - model architecture.

Must have finite vocabulary!!!

Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

Extra Slides – codebook interleaving pattern

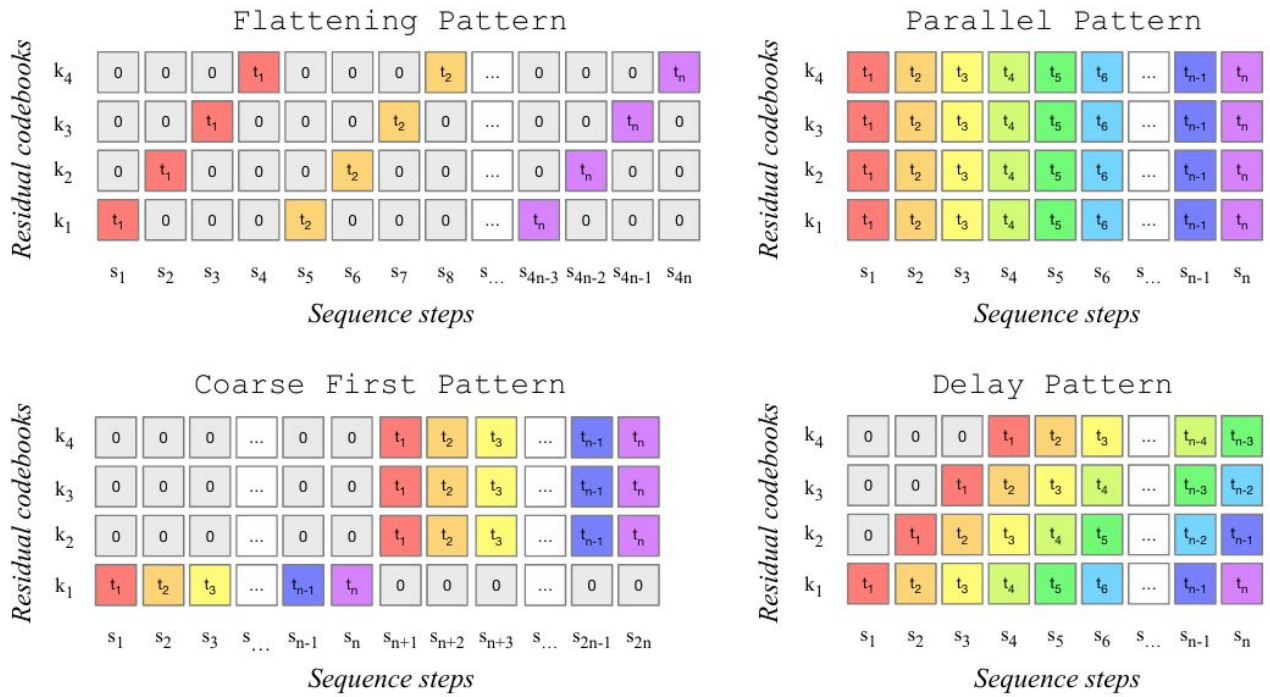
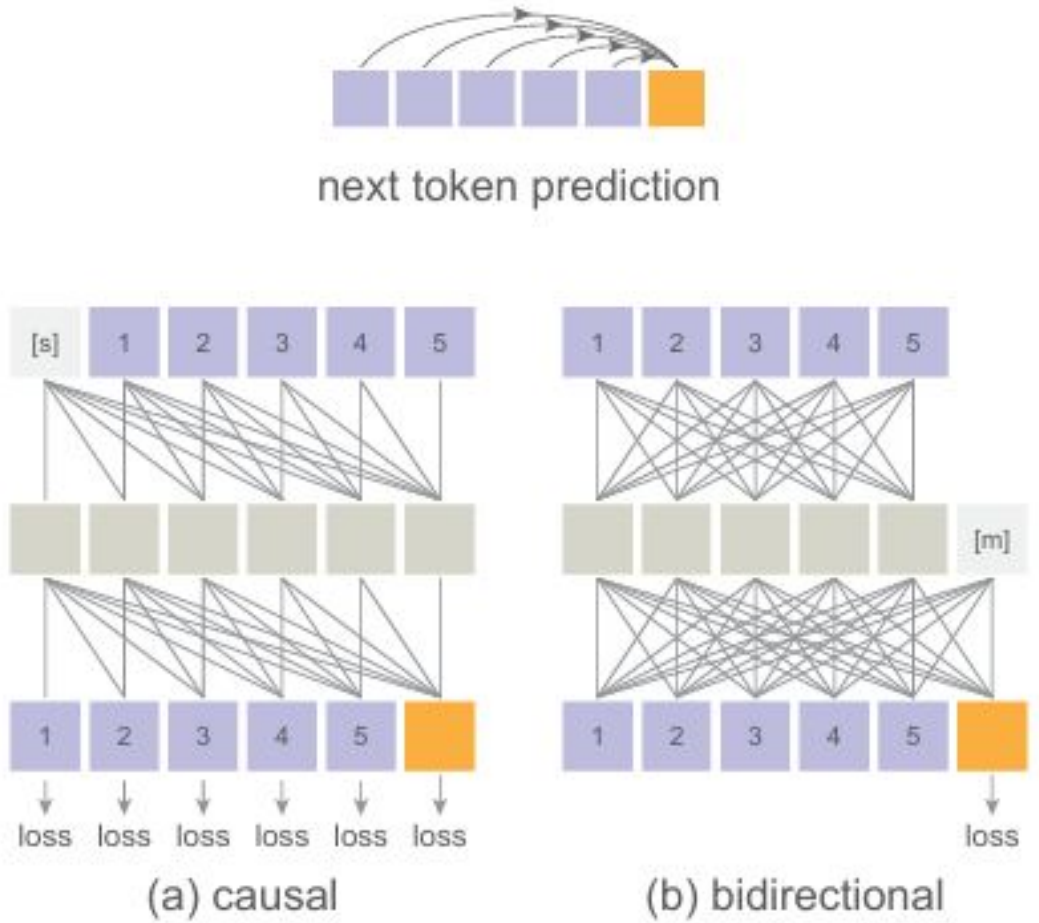


Figure 1: Codebook interleaving patterns presented in Section 2.2. Each time step $t_1, t_2, \ldots, t_n$ is composed of 4 quantized values (corresponding to $k_1, \ldots, k_4$). When doing autoregressive modelling, we can flatten or interleave them in various ways, resulting in a new sequence with 4 parallel streams and steps $s_1, s_2, \ldots, s_m$. The total number of sequence steps S depends on the pattern and original number of steps T. 0 is a special token indicating empty positions in the pattern.

Copet, Jade, et al. "Simple and controllable music generation." *Advances in neural information processing systems* 36 (2023): 47704-47720.

Extra Slides – MAR



Li, Tianhong, et al. "Autoregressive image generation without vector quantization." *Advances in Neural Information Processing Systems* 37 (2024): 56424-56445.

Extra Slides – short-context Transformer

Table 8: **FAD after 500K training steps for different short-context Transformer contexts K.**

K in short-context Transformer	FAD
5	0.85 ± 0.05
10	0.76 ± 0.04
20	0.73 ± 0.04
40	0.78 ± 0.04

Extra Slides – Consistency Models

Consistency Models

Algorithm 2 Consistency Distillation (CD)

Input: dataset $\mathcal{D}$, initial model parameter $\boldsymbol{\theta}$, learning rate η, ODE solver $\Phi(\cdot,\cdot;\boldsymbol{\phi})$, $d(\cdot,\cdot)$, $\lambda(\cdot)$, and μ

$\boldsymbol{\theta}^- \leftarrow \boldsymbol{\theta}$

repeat

Sample $\mathbf{x} \sim \mathcal{D}$ and $n \sim \mathcal{U}[\![1, N-1]\!]$

Sample $\mathbf{x}_{t_{n+1}} \sim \mathcal{N}(\mathbf{x}; t_{n+1}^2 \boldsymbol{I})$

$\hat{\mathbf{x}}^{\boldsymbol{\phi}}_{t_n} \leftarrow \mathbf{x}_{t_{n+1}} + (t_n - t_{n+1})\Phi(\mathbf{x}_{t_{n+1}}, t_{n+1}; \boldsymbol{\phi})$

$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^-; \boldsymbol{\phi}) \leftarrow$

$\quad \lambda(t_n) d(\boldsymbol{f}_{\boldsymbol{\theta}}(\mathbf{x}_{t_{n+1}}, t_{n+1}), \boldsymbol{f}_{\boldsymbol{\theta}^-}(\hat{\mathbf{x}}^{\boldsymbol{\phi}}_{t_n}, t_n))$

$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^-; \boldsymbol{\phi})$

$\boldsymbol{\theta}^- \leftarrow \operatorname{stopgrad}(\mu \boldsymbol{\theta}^- + (1-\mu)\boldsymbol{\theta})$

until convergence

Algorithm 3 Consistency Training (CT)

Input: dataset $\mathcal{D}$, initial model parameter $\boldsymbol{\theta}$, learning rate η, step schedule $N(\cdot)$, EMA decay rate schedule $\mu(\cdot)$, $d(\cdot,\cdot)$, and $\lambda(\cdot)$

$\boldsymbol{\theta}^- \leftarrow \boldsymbol{\theta}$ and $k \leftarrow 0$

repeat

Sample $\mathbf{x} \sim \mathcal{D}$, and $n \sim \mathcal{U}[\![1, N(k)-1]\!]$

Sample $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$

$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^-) \leftarrow$

$\quad \lambda(t_n) d(\boldsymbol{f}_{\boldsymbol{\theta}}(\mathbf{x} + t_{n+1}\mathbf{z}, t_{n+1}), \boldsymbol{f}_{\boldsymbol{\theta}^-}(\mathbf{x} + t_n \mathbf{z}, t_n))$

$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^-)$

$\boldsymbol{\theta}^- \leftarrow \operatorname{stopgrad}(\mu(k)\boldsymbol{\theta}^- + (1-\mu(k))\boldsymbol{\theta})$

$k \leftarrow k + 1$

until convergence

Song, Yang, et al. "Consistency models." (2023).

Extra Slides – Consistency Models

Consistency Models

Algorithm 2 Consistency Distillation (CD)

Input: dataset $\mathcal{D}$, initial model parameter $\boldsymbol{\theta}$, learning rate η, ODE solver $\Phi(\cdot,\cdot;\boldsymbol{\phi})$, $d(\cdot,\cdot)$, $\lambda(\cdot)$, and μ

$\boldsymbol{\theta}^- \leftarrow \boldsymbol{\theta}$

repeat

Sample $\mathbf{x} \sim \mathcal{D}$ and $n \sim \mathcal{U}[\![1, N-1]\!]$

Sample $\mathbf{x}_{t_{n+1}} \sim \mathcal{N}(\mathbf{x}; t_{n+1}^2 \boldsymbol{I})$

$\hat{\mathbf{x}}^{\boldsymbol{\phi}}_{t_n} \leftarrow \mathbf{x}_{t_{n+1}} + (t_n - t_{n+1})\Phi(\mathbf{x}_{t_{n+1}}, t_{n+1}; \boldsymbol{\phi})$

$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^-; \boldsymbol{\phi}) \leftarrow$

$\lambda(t_n) d(\boldsymbol{f}_{\boldsymbol{\theta}}(\mathbf{x}_{t_{n+1}}, t_{n+1}), \boldsymbol{f}_{\boldsymbol{\theta}^-}(\hat{\mathbf{x}}^{\boldsymbol{\phi}}_{t_n}, t_n))$

$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^-; \boldsymbol{\phi})$

$\boldsymbol{\theta}^- \leftarrow \operatorname{stopgrad}(\mu \boldsymbol{\theta}^- + (1-\mu)\boldsymbol{\theta})$

until convergence

Algorithm 3 Consistency Training (CT)

Input: dataset $\mathcal{D}$, initial model parameter $\boldsymbol{\theta}$, learning rate η, step schedule $N(\cdot)$, EMA decay rate schedule $\mu(\cdot)$, $d(\cdot,\cdot)$, and $\lambda(\cdot)$

$\boldsymbol{\theta}^- \leftarrow \boldsymbol{\theta}$ and $k \leftarrow 0$

repeat

Sample $\mathbf{x} \sim \mathcal{D}$, and $n \sim \mathcal{U}[\![1, N(k)-1]\!]$

Sample $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$

$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^-) \leftarrow$

$\lambda(t_n) d(\boldsymbol{f}_{\boldsymbol{\theta}}(\mathbf{x} + t_{n+1}\mathbf{z}, t_{n+1}), \boldsymbol{f}_{\boldsymbol{\theta}^-}(\mathbf{x} + t_n \mathbf{z}, t_n))$

$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^-)$

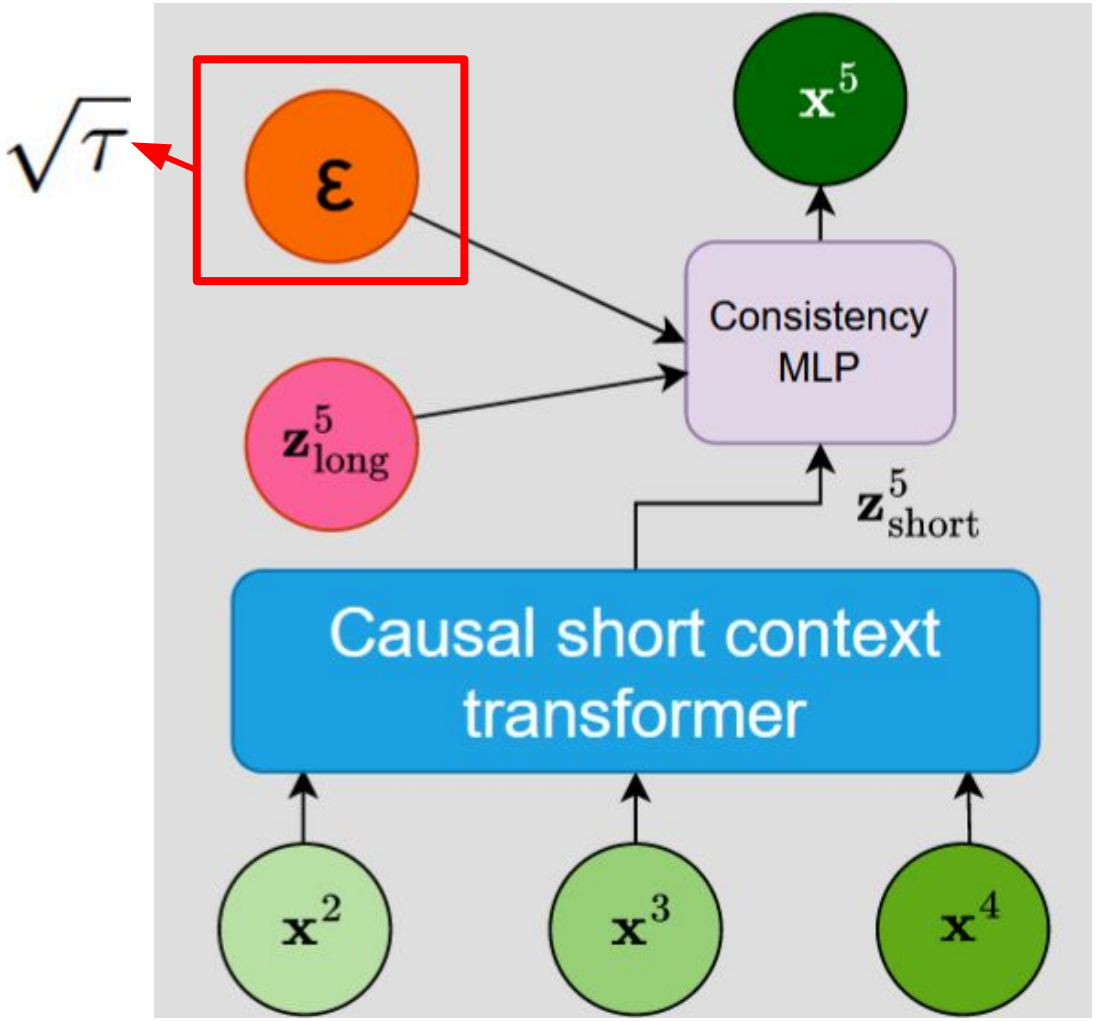
$\boldsymbol{\theta}^- \leftarrow \operatorname{stopgrad}(\mu(k)\boldsymbol{\theta}^- + (1-\mu(k))\boldsymbol{\theta})$

$k \leftarrow k + 1$

until convergence

Song, Yang, et al. "Consistency models." (2023).

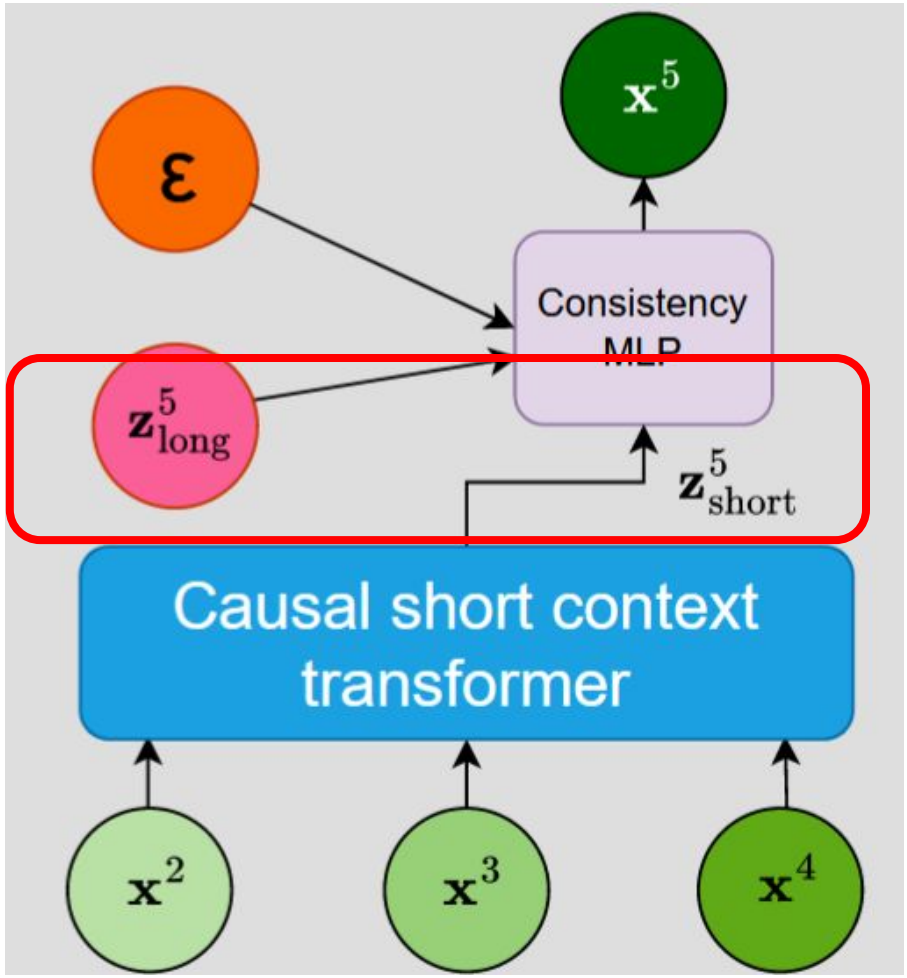
Extra Slides – Gaussian Temperature sampling



Discrete Model:

$$P_T(i) = \frac{\exp(z_i/T)}{\sum_{j=1}^{V} \exp(z_j/T)}.$$

Extra Slides – Latent Classifier Free Guidance



Diffusion Model:

$$\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t|y) = \underbrace{\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)}_{\text{unconditional score}} + \underbrace{s(\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t|y) - \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t))}_{\text{unconditional to conditional}}$$

Consistency Model:

$$\bar{\mathbf{Z}}^s_{\text{CFG}} = \mathbf{Z}^s_{\emptyset} + \alpha(\mathbf{Z}^s_C - \bar{\mathbf{Z}}^s_{\emptyset})$$

Extra Slides – Hyperparameters

Table 13: VAE hyperparameters

	Music VAE	Speech continuation VAE
General		
Sample rate	32kHz	24kHz
Frame rate	25Hz	12.5Hz
Latent dimension	128	32
Architecture		
Convolutions ratios	8, 8, 5, 4	6, 5, 4, 4, 4
Num transformer encoder layers	4	8
Num transformer decoder layers	4	8
Transformer context	30s	10s
Training parameters		
Batch size	64	64
Audio sample length	12s	12s
KL loss weight	0.01	0.01
Reconstruction loss	✓	✗
Distillation loss weight	✗	25
LR Schedule	cosine	cosine
Learning rate	$8 \cdot 10^{-4}$	$8 \cdot 10^{-4}$

Extra Slides – Hyperparameters

Table 14: Model and training hyperparameters

	Music	Speech continuation	Text-to-Speech
Backbone Transformer			
Model dimension	1536	2560	1024
MLP dimension	6336	10560	4096
Number of heads	24	20	16
Number of layers	48	24	24
Learning rate	$1 \cdot 10^{-4}$	$5 \cdot 10^{-5}$	1e-4
Number of parameters	1.35B	2.2B	302M
RQ-Transformer head (RVQ model)			
Model dimension	1024	1024	✗
MLP dimension	4096	4096	✗
Number of heads	16	16	✗
Number of layers	6	6	✗
Number of parameters	701M	701M	✗
Consistency sampler head (ours)			
Number of layers	12	6	6
MLP dimension	3072	512	512
Gating	SiLU	SiLU	SiLU
Number of parameters	601M	10M	10M
Short context transformer (ours)			
Model dimension	1536	✗	✗
MLP dimension	6336	✗	✗
Number of heads	24	✗	✗
Number of layers	4	✗	✗
Context	10	✗	✗
Number of parameters	113M	✗	✗
Audio embedding manipulations			
Center and normalize	✓	✓	✓
Noise before entering backbone	✓	✗	✗
Training parameters			
Head batch multiplier	8	8	8
Optimizer	AdamW $\beta_1 = 0.9, \beta_2 = 0.95$	AdamW $\beta_1 = 0.9, \beta_2 = 0.95$	AdamW $\beta_1 = 0.9, \beta_2 = 0.95$
Batch size	48	144	128
Audio sample length	30s	300s	60s
LR Schedule	cosine	cosine	cosine
Learning rate	$1 \cdot 10^{-4}$	$2 \cdot 10^{-4}$	1e-4
GPU used	16 H100	48 H100	8 H100
Number of training steps	500k	150k	400k
Initial checkpoint	✗	Helium-1 (Kyutai, 2025)	✗
Inner monologue	✗	✓	✗
Acoustic delay	-	2 frames (160 ms)	-

Extra Slides – Data

Speech Continuation
Train: Large-scale English & French speech
Test: 1,000 30s clips (Objective) & 50 internal English samples (Human)

Text-to-Speech (TTS) & Pocket TTS
Train: 88,000 hours of public datasets (GIGASpeech, VoxPopuli, etc.)
Test: Librispeech test-clean

Music Continuation
Train: 400K songs (20K hours) from LAION-Disco-12M
Test: 4,000 clips from LAION-Disco-12M test set

Text-to-Music Generation
Train: LAION-Disco-12M (using CLAP features as pseudo-text)
Test: MusicCaps benchmark & LAION-Disco-12M test set

Extra Slides ---- Text-to-Speech

Model	Num. parameters	WER (↓)	CER (↓)	SIM (↑)	Acoustic Quality (↑)	Speaker SIM Human ELO (↑)
Reference	–	2.23	–	0.69	61.8 ± 2.4	1953 ± 24
Reference (with VAE)	–	–	–	0.57	–	–
F5 TTS (NFE=32) (Chen et al., 2025)	336M	2.42	–	0.66	54.7 ± 2.8	2032 ± 18
DSM (16 RVQ CFG=3 w.r.t text and audio prompt) (Zeghidour et al., 2025)	750M	1.95	–	**0.67**	60.2 ± 2.4	**2112 ± 20**
DiTAR (NFE=10) (Jia et al., 2025)	600M	2.39	–	**0.67**	–	–
SALAD (NFE=20) (Turetzky et al., 2024)	350M	–	0.74	0.54	–	–
CALM w/ LSD (NFE=1, CFG=1.5 w.r.t text)	**313M**	**1.81**	**0.57**	0.52	**61.1 ± 2.3**	1966 ± 23

LSD: Lagrangian Self-Distillation

Extra Slides – Text-to-Music

Table 11: Text-to-music results on the MusicCaps dataset.

Model	FAD (↓)	KLD (↓)	CLAP (↑)
Reference	–	–	0.30
MusicLM (Agostinelli et al., 2023)	4.00	–	–
MusicGen Medium (Copet et al., 2023)	3.40	1.23	0.37
AudioLDM2 (Liu et al., 2024)	3.13	1.20	0.43
Noise2Music (Huang et al., 2023)	2.10	–	–
Jen-1 (Li et al., 2025)	2.00	1.29	–
MusicFlow (Unidirectional LM + FM) (Prajwal et al., 2024)	2.69	1.23	0.52
MusicGen Medium (retrained with CLAP)	2.70	1.37	0.39
RQ-transformer 32 RVQ	2.56	1.35	0.43
CALM - Consistency - 4 Steps	2.14	1.30	0.44