Text-Based Speech Editing

From a Perspective of Mask-Based Generative Models

2023-01-12

Contents

Background

基于文本的语音编辑

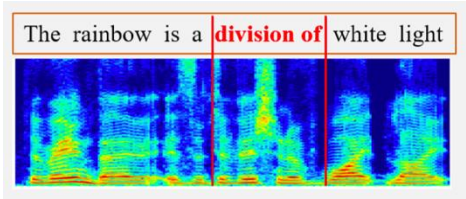
问题定义：通过对文本的编辑，实现对语音相应位置的修改

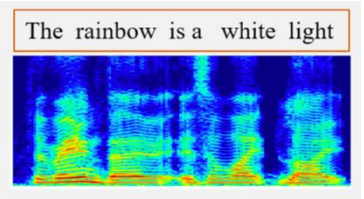
研究背景：录制的语音后来发现存在问题，需要修改语音内容

- 解决方案一：重新录制
  - 成本高：专业的设备、录音场地、人力/时间成本等
  - 时效性低：修复周期长，录制流程冗长不便于即时修改
  - 难以实现：录音的说话人联系不到或音色发生变化
- 解决方案二：人工编辑
  - 人工要求较高，修改处边界可能需要逐帧修改
  - 只能 cut/copy/paste，难以**插入**录音中原本不存在的内容

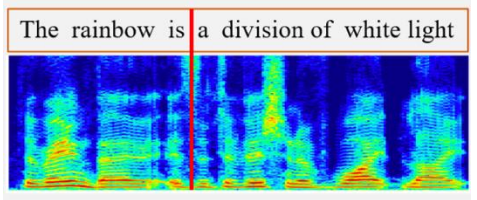
语音编辑的类型

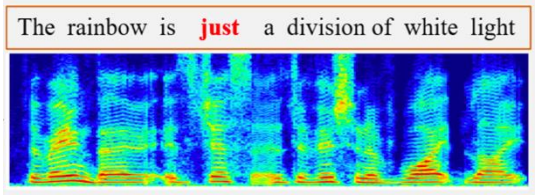
Deletion: 删除操作



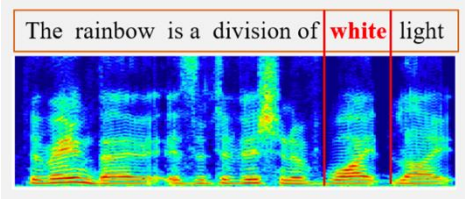


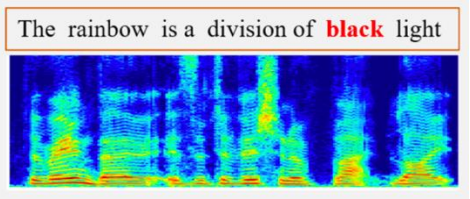
Insertion: **插入操作（难点）**





Replace: 替换操作 ≈ 删除 + **插入**





Background

语音编辑的四大要求/难点

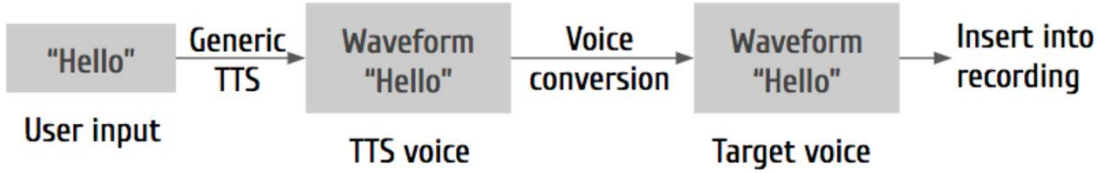
- 位置的准确性 (Location Accuracy)
- 内容的正确性 (Content Clarity)
- **音色的一致性 (Speaker Identity)**
- **韵律的连续性 (Prosody Continuity)**

Alignment → 对齐模型

Text, Timbre, Prosody → TTS

语音编辑（插入）的简单思路

- 方案一：TTS 模型合成待插入文本，再拼接到语音的目标位置
  - 音色统一性较差 + 韵律连续性不好
- 方法二：先合成出待插入片段，再语音转换成目标音色
  - 计算存在冗余；韵律连续性同样无法保证
  - 语音转换效果不一定更好，多系统**级联**，错误累积



Introduction: Text-Based Speech Inpainting

https://arxiv.org/abs/2202.07273

SPEECHPAINTER: TEXT-CONDITIONED SPEECH INPAINTING

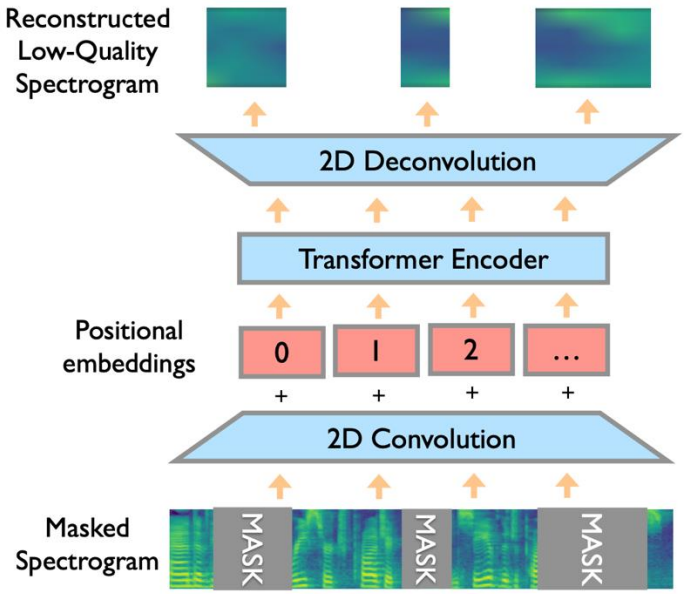
Zalán Borsos, Matt Sharifi, Marco Tagliasacchi*

Google Research

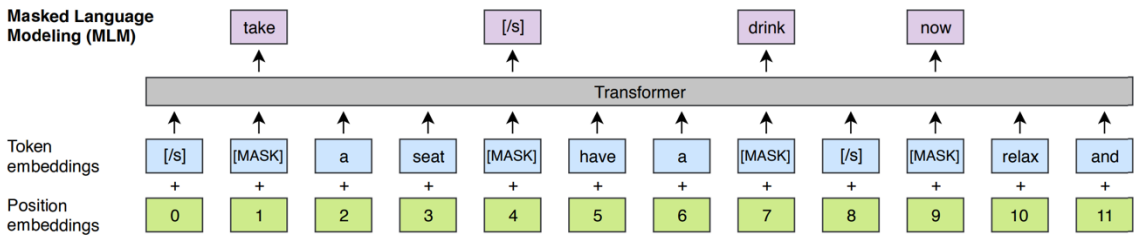
Speech Inpainting

语音修复

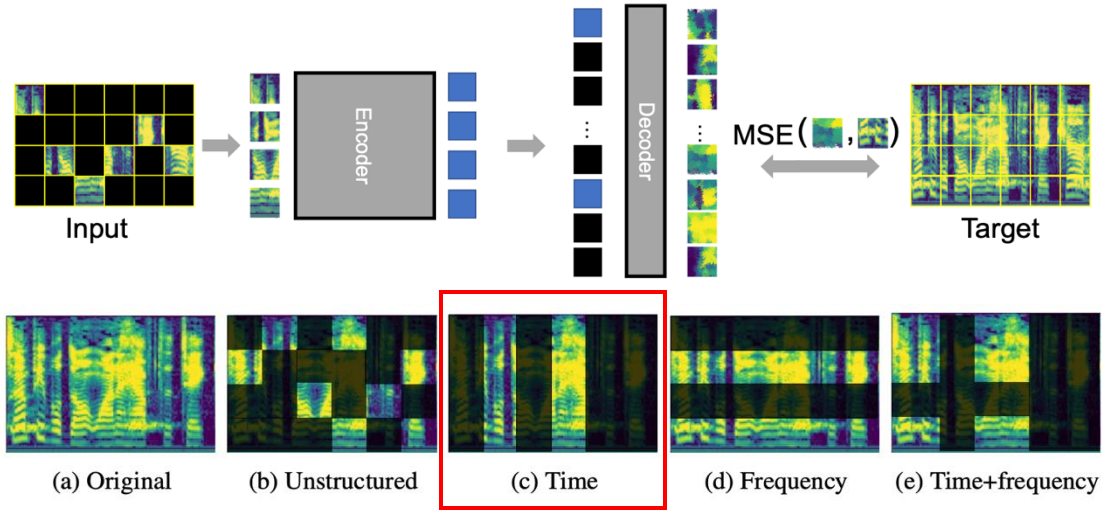
问题定义：恢复/修复出语音波形中缺失的部分



Masked Acoustic Model (MAM), Baidu



Masked Language Model, Google



Audio Masked Auto-Encoder (Audio MAE), Meta AI

- Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).
- Chen, J., Ma, M., Zheng, R. and Huang, L., 2020. Mam: Masked acoustic modeling for end-to-end speech-to-text translation. arXiv preprint arXiv:2010.11445.
- Xu, H., Li, J., Baevski, A., Auli, M., Galuba, W., Metze, F. and Feichtenhofer, C., 2022. Masked autoencoders that listen. arXiv preprint arXiv:2207.06405.

SpeechPainter (Google)

基于文本的语音修复

输入：

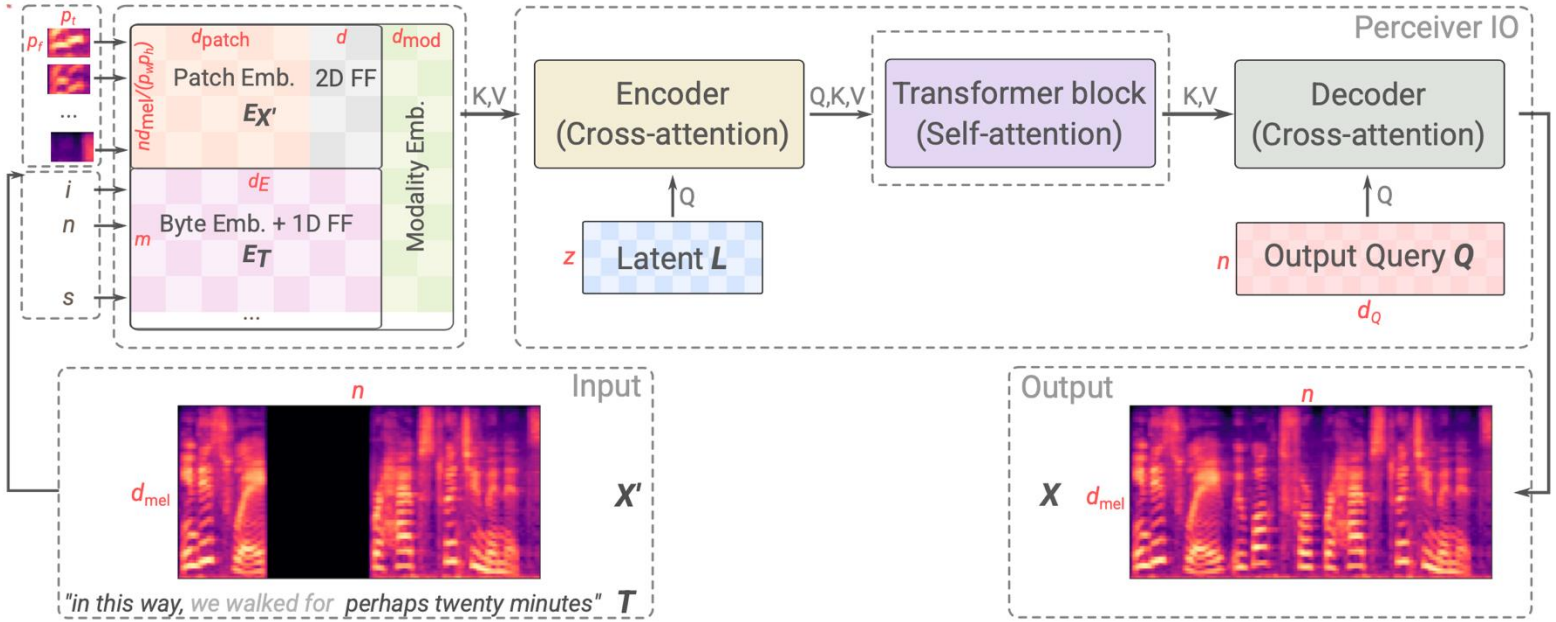
- 完整的文本(音素)序列
- 时域上不完整的语音

输出：

- 修复后的完整语音

应用：

- 语音语法错误修正
- 错误发音修正
- 智能丢包补偿
- 移除背景突发噪声



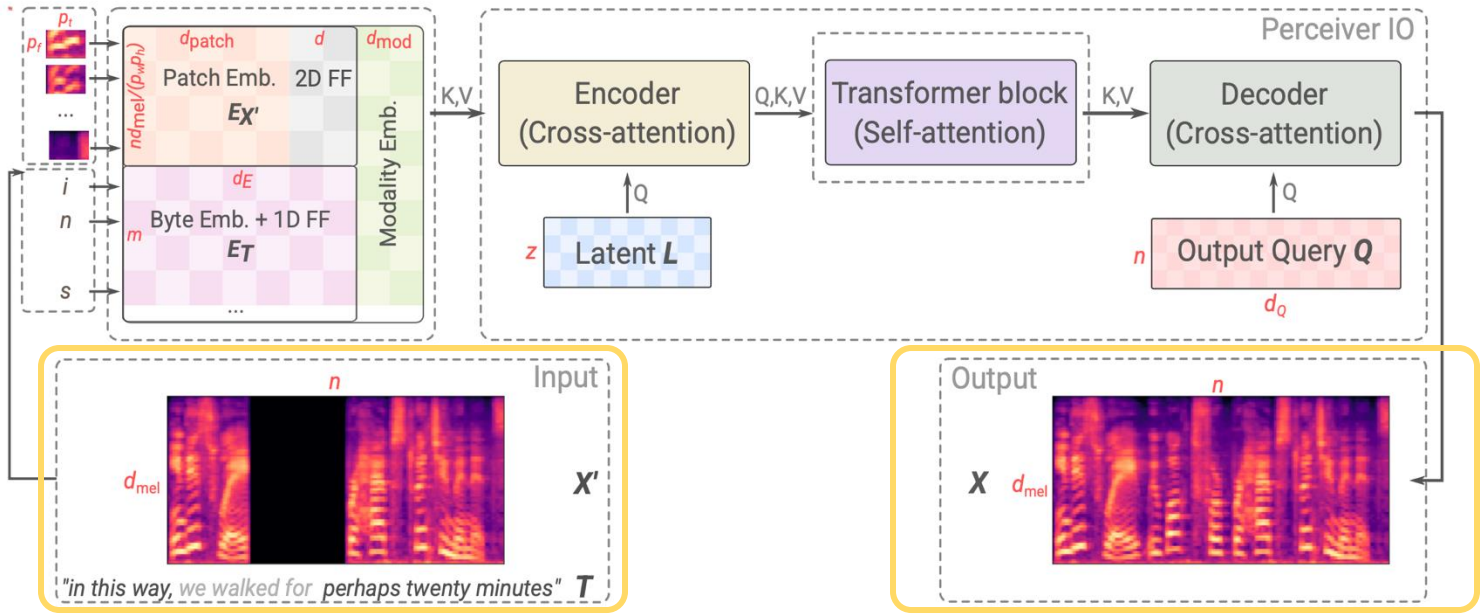
- 不需要将语音和文本对齐，文本只需要包含目标的文本子串
- 论文效果：文本作为补充条件，得到比纯语音修复更好的效果
  - 纯语音修复：缺失部分在 250 ms 以内
  - 增加了文本之后： 缺失部分的长度允许在 1s 以内

SpeechPainter (Google)

基于文本的语音修复

1. 模型的输入输出

- 输入X ‘: 含有待恢复区的的语音特征
  - n: 帧数；d_mel: 声学特征维度
- 输入T: 包含完整目标子串的文本
- 输出X：目标修复出完整的语音特征
  - n: 帧数；d_mel: 声学特征维度



2. 优化目标

$$\mathcal{L}_{\mathcal{G}}^{\text{rec}} = \frac{1}{nd_{\text{mel}}}\mathbb{E}_{(X,T)}\left[\|X - \mathcal{G}(X', T)\|_1\right]$$

SpeechPainter (Google)

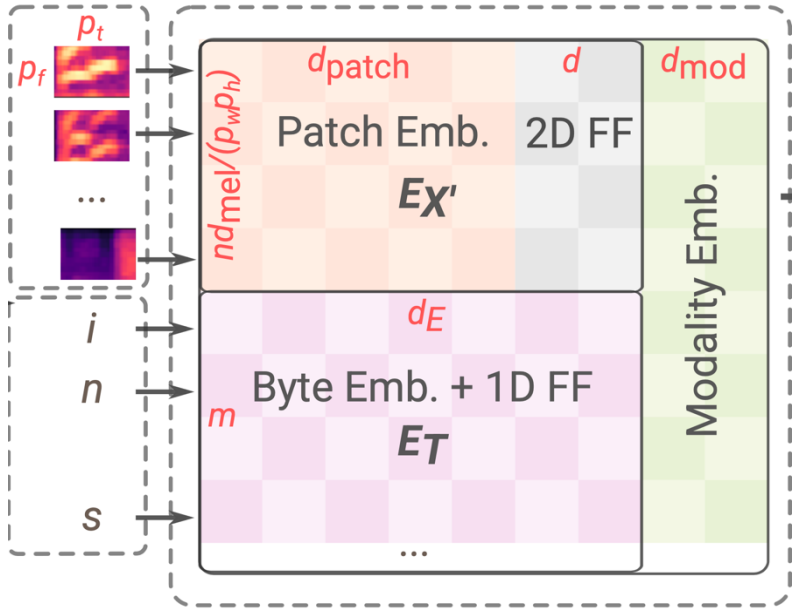
基于文本的语音修复

3. 语音文本的 embedding 策略

- **Patch Embedding**
  - p_t × p_f 互不交叠的 patch, flatten 经过全连接映射到 d_patch 维度
  - patch 个数为 nd_mel/(p_t * p_f)
  - embedding 维度 d_E = d_patch + d_FF (2-D 的维度)
- **Fourier Feature Encoding (1-D, 2-D)**

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

- **Byte Embedding**
  - 将每个 char 的 UTF-8 编码映射为 d_E 维的 embedding
  - 1-D 的 FF encoding 维度同样为 d_E，与 Byte Embedding 相加
- **Modality Embedding**
  - 随机初始化的可学的 embedding，维度为 d_mod

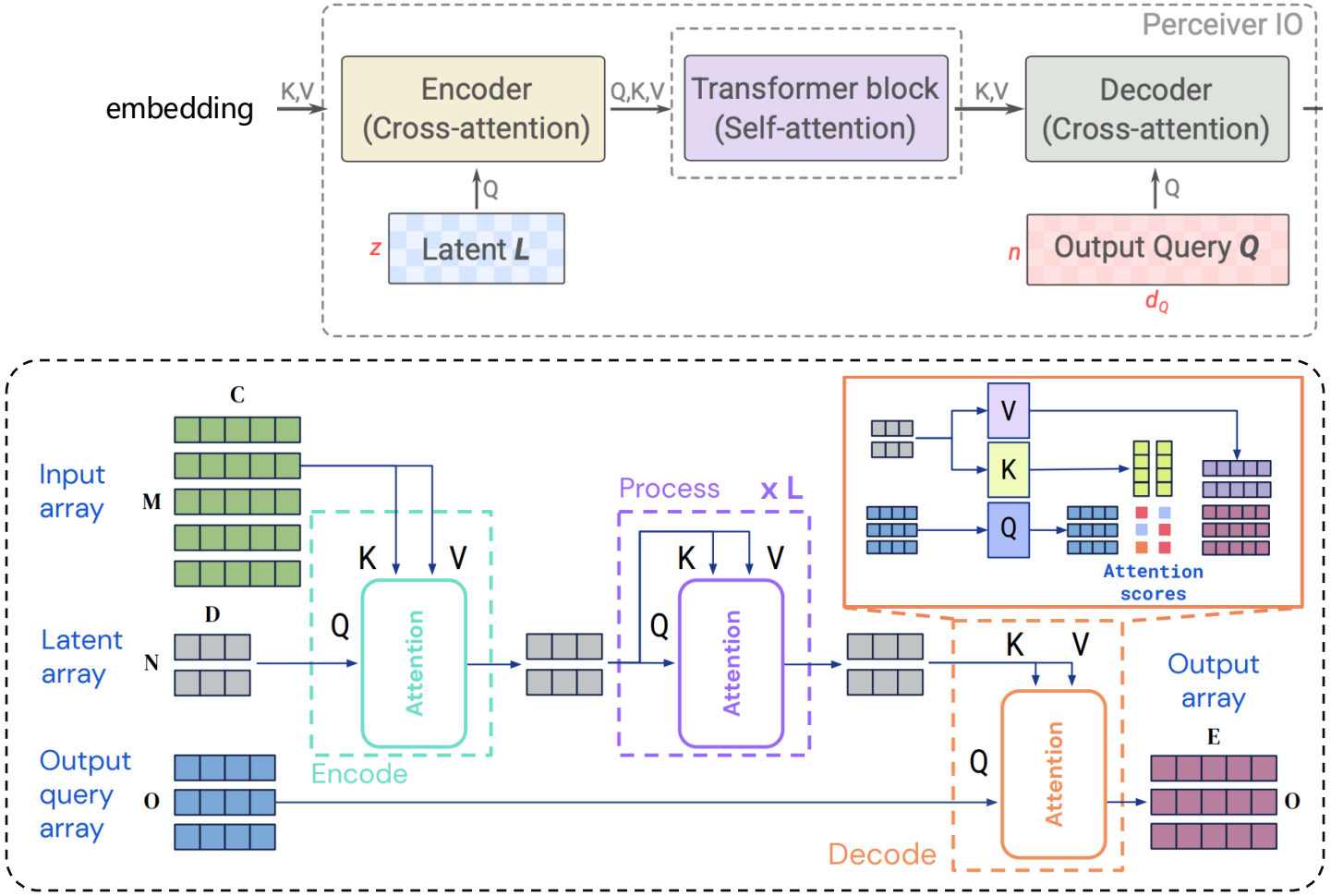


SpeechPainter (Google)

基于文本的语音修复

4. 基于 Perceiver IO 的编解码模块

Encode → Process → Decode



之前工作：

- 多模态数据用独立的模块对不同信息进行特征抽取，再用额外的模型进行融合。

Perceiver IO：

- 单一的神经网络架构，适应不同模态的输入，作为处理任意信息/任意任务的通用模型。

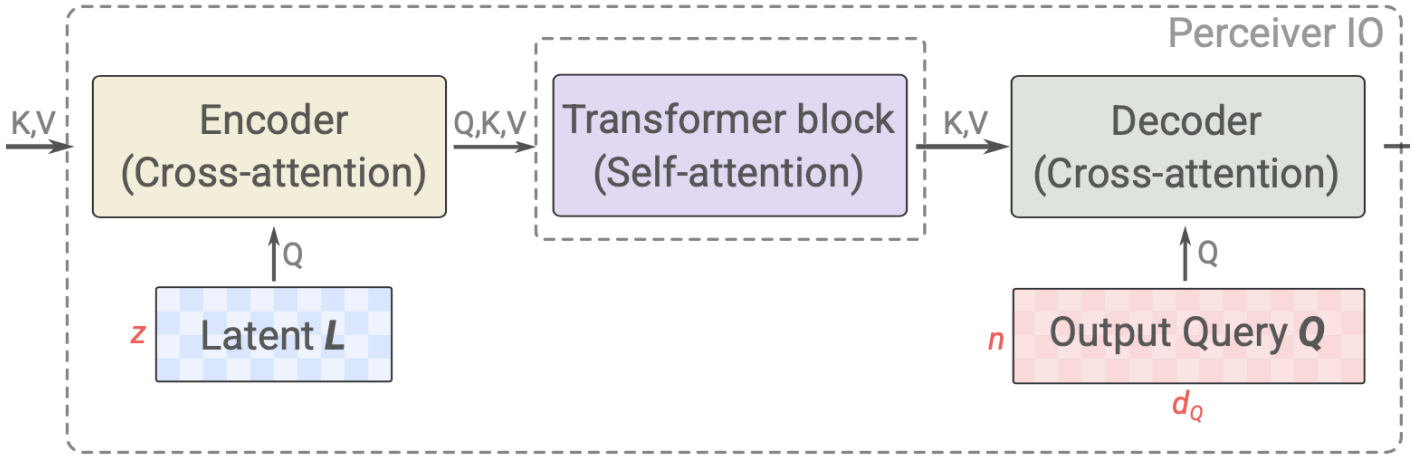
- Jaegle, Andrew, et al. "Perceiver IO: A general architecture for structured inputs & outputs." arXiv preprint arXiv:2107.14795 (2021).

SpeechPainter (Google)

基于文本的语音修复

4. 基于 Perceiver IO 的编解码模块

Encode → Process → Decode



- 核心：利用与任务相关的特定查询向量，与隐变量一起生成输出。
- 优势：模型可以生成任意的目标输出，同时隐变量特征保持与特定任务无关的通用特性，灵活地适应不同任务的输出需求。

- Jaegle, Andrew, et al. "Perceiver IO: A general architecture for structured inputs & outputs." arXiv preprint arXiv:2107.14795 (2021).

demo: https://google-research.github.io/seanet/speechpainter/examples

SpeechPainter (Google)

实验细节

参数设计

$$n = 240, d_{mel} = 64$$

$$p_t = 8, p_f = 4$$

$$d_{patch} = 768, d_{FF} = 256, d_{mod} = 16$$

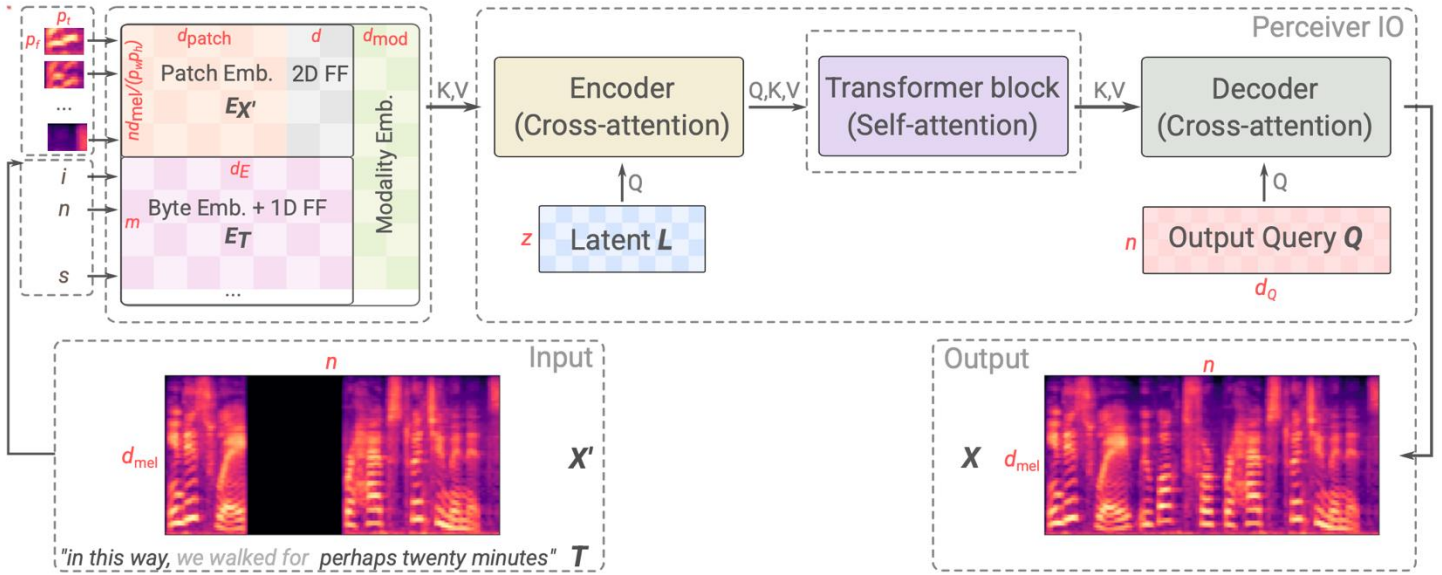
$$d_E = d_{patch} + d_{FF} = 1024$$

$$m = 500, z = 512$$

$$d = 1024, d_Q = 512$$

$$k = 16, \text{头数} h = 16, \text{隐层} d = 1024$$

声码器：**MelGAN**



	MOS		
	TTS	SpeechPainter	Original (Target)
LibriTTS clean	3.06 ± 0.10	3.60 ± 0.09	3.68 ± 0.09
LibriTTS other	2.77 ± 0.12	3.12 ± 0.10	3.30 ± 0.12
VCTK	3.38 ± 0.10	3.70 ± 0.09	3.98 ± 0.08
Aggregated	3.07 ± 0.06	3.48 ± 0.06	3.66 ± 0.06

A New Perspective: Speech-Text Joint Pretraining

https://arxiv.org/abs/2203.09690

A^{3}T: Alignment-Aware Acoustic and Text Pretraining for Speech Synthesis and Editing

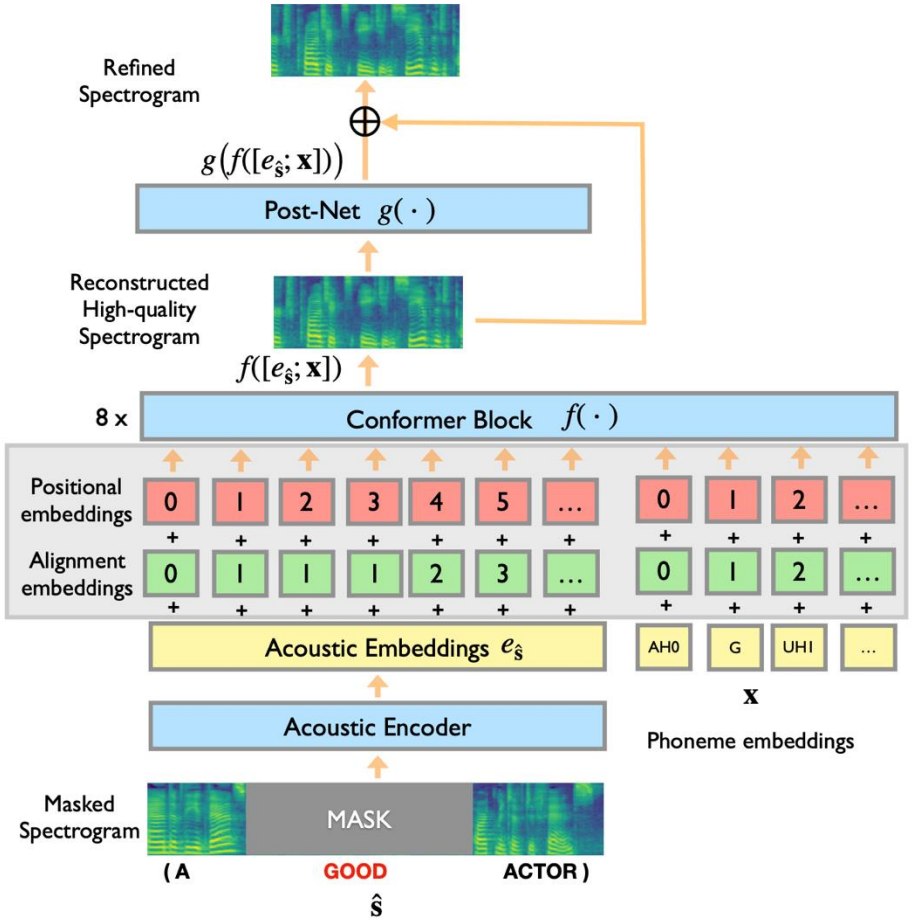
He Bai [*1] Renjie Zheng [*2] Junkun Chen [3] Xintong Li [2] Mingbo Ma [2] Liang Huang [3]

A3T: Alignment-Aware Acoustic and Text Pretraining

文本-语音 联合预训练

- 语音/文本模态联合预训练，常用于分类/识别/理解性任务
- **语音合成/语音编辑** 属于生成类的任务

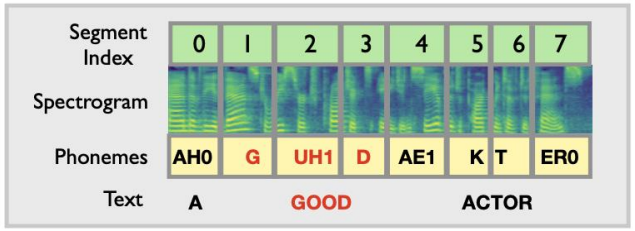
- A3T：以生成语音为训练目标，两大类应用
  - **语音编辑**
  - **Prompt-Based TTS**



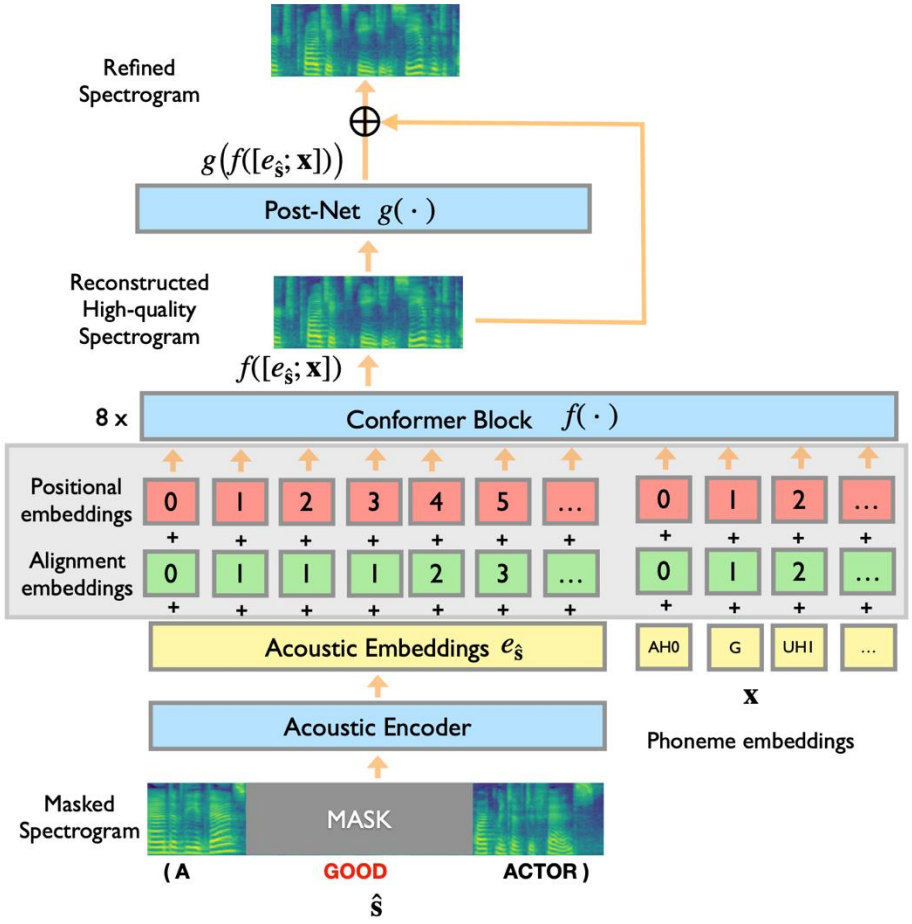
A3T: Alignment-Aware Acoustic and Text Pretraining

A3T 的结构

- 输入：音素序列 + 部分 mask 的语音特征
- 输出：重建完整的语音
- Separate Encoders
  - Acoustic Encoder：全连接层 + 非线性层
  - Phoneme Embeddings
- Shared Encoder/Decoder：Conformer Block
- Cross-Modal Alignment Embedding

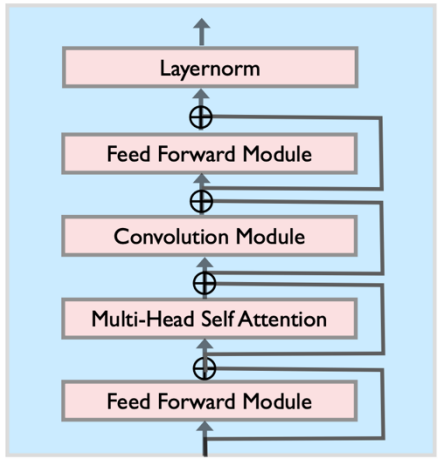


  - 按照音素的出现先后进行 one-shot 编码
- Training Objective: Speech Reconstruction
  - 只针对 mask 的部分进行重建预测

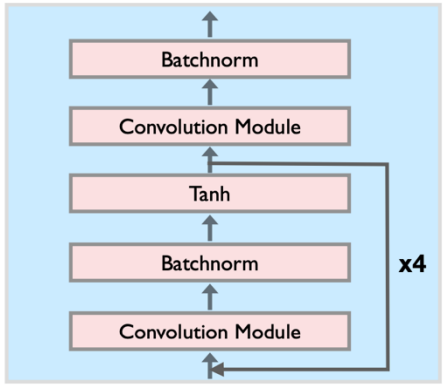


A3T: Alignment-Aware Acoustic and Text Pretraining

A3T 的细节

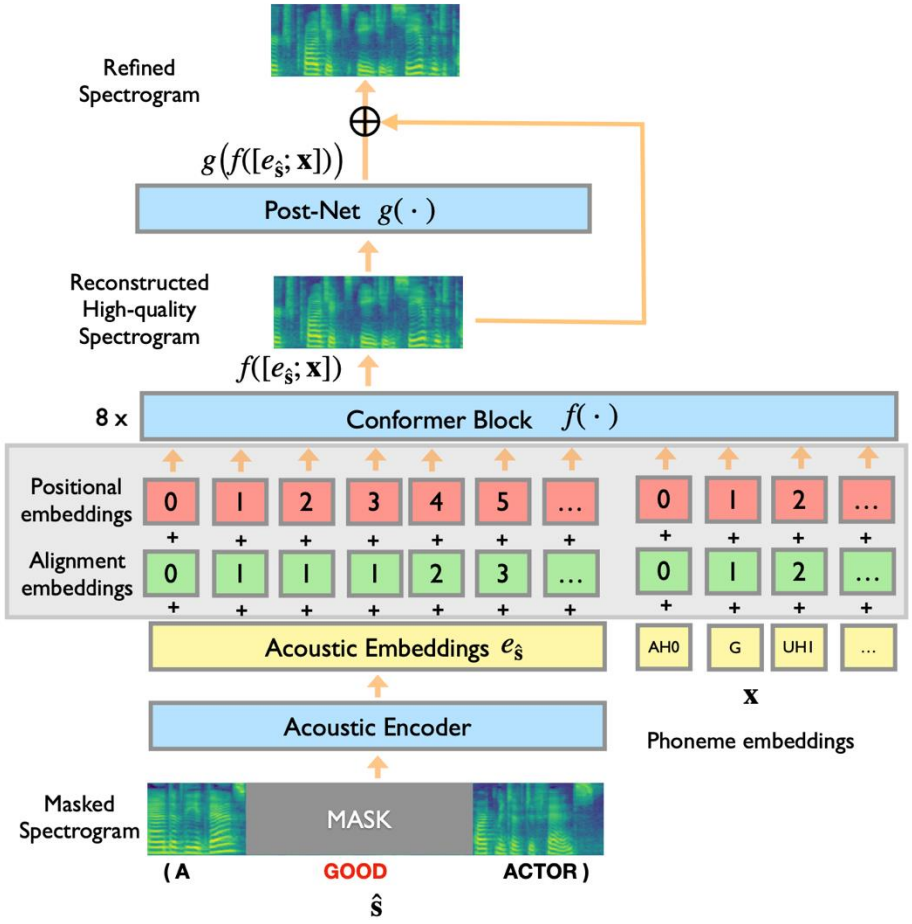


(c) Conformer Block.



(d) Post-Net.

$$\ell_{\mathbf{s}}(D_{\mathbf{s},\mathbf{x}}) = \sum_{\langle \mathbf{s},\mathbf{x} \rangle \in D_{\mathbf{s},\mathbf{x}}} \| \underbrace{f([e_{\hat{\mathbf{s}}};\mathbf{x}]) + g\big(f([e_{\hat{\mathbf{s}}};\mathbf{x}])\big)}_{\text{refined spectrogram}} - \mathbf{s} \|_1 + \| \underbrace{f([e_{\hat{\mathbf{s}}};\mathbf{x}])}_{\text{reconstructed spectrogram}} - \mathbf{s} \|_1$$



A3T: Alignment-Aware Acoustic and Text Pretraining

应用一：基于文本的语音编辑

流程细节

1. 原始文本+原始音频对齐，获取各音素边界
2. 修改后的完整文本，预测持续时长 duration
3. duration 调整

d'_i **duration predictor** 预测的时长值
$\tilde{x}$ 原始文本的全部音素
$\tilde{d}_j$ 强制对齐获取的真实时长值
d'_j duration predictor 预测的时长值

$$\hat{d}_i = d'_i \frac{\sum_{j=1}^{|\tilde{x}|} \tilde{d}_j}{\sum_{j=1}^{|\tilde{x}|} d'_j}$$

4. 插入待预测的 mask 帧

$\hat{x}$ 修改后的音素(插入的音素)

插入 mask 的帧数：$\sum_{i=1}^{\hat{x}} \hat{d}_i \times \frac{sr}{h}$

5. 使用 A3T 模型预测 mask 部分的声学特征
6. 使用 PWG 声码器生成目标片段的波形
7. 插入到原始语音波形中，得到编辑后的语音

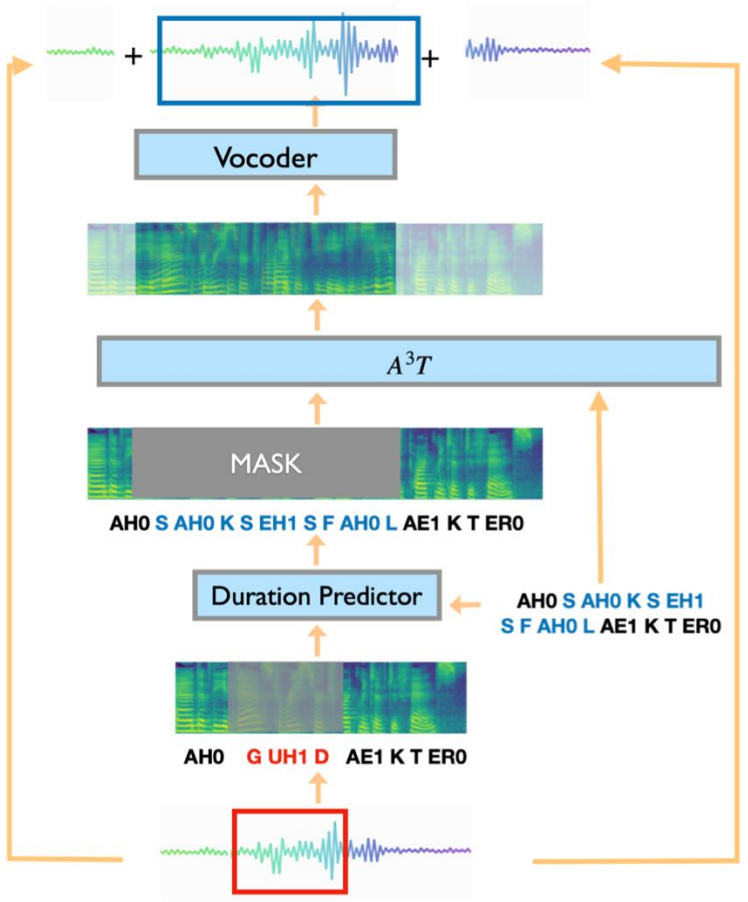


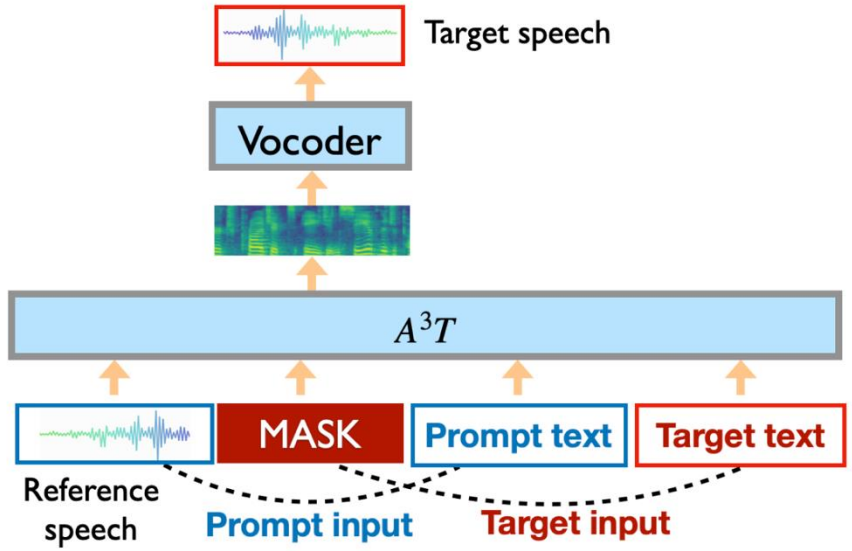
Figure 3. Speech editing pipeline.

A3T: Alignment-Aware Acoustic and Text Pretraining

应用二：One-Shot 语音合成

流程细节

- Prompt：**目标说话人的一条参考语音 + 对应音素/文本 → One-Shot**
- 输入：目标文本(音素)
- 根据目标文本预测时长 duration，计算目标语音的帧数
- 将 Prompt 的语音和相应帧数的 Mask 拼接
- 将 Prompt 的文本与目标文本拼接
- 拼接后的语音与文本作为 A3T 模型输入，预测被 Mask 的语音特征
- 语音特征经过声码器转换为目标的波形



(b) Prompt-based decoding.

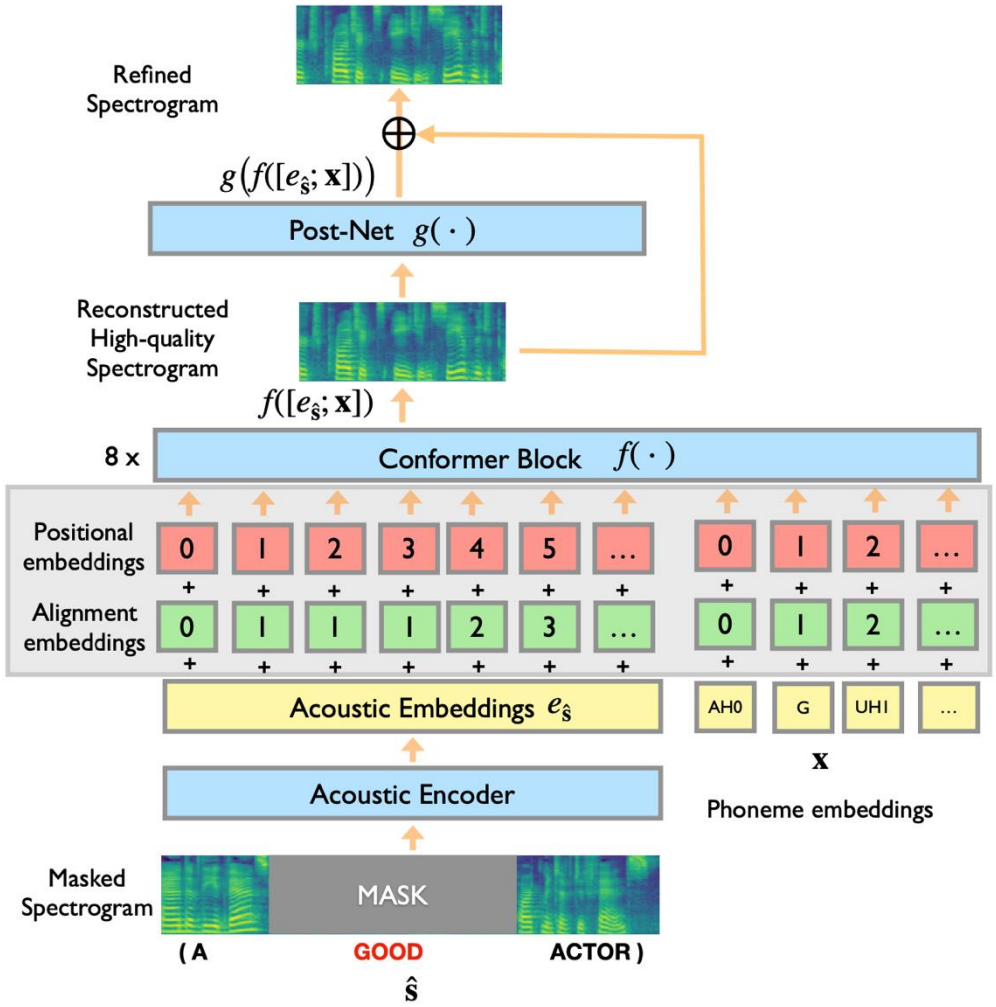
A3T: Alignment-Aware Acoustic and Text Pretraining

实验细节

- 对齐工具：HTK
- **外部 Duration Predictor:** 基于 FastSpeech2
- 声码器：Parallel WaveGAN
- Conformer
  - Encoder: 4 层，kernel_size=7
  - Decoder: 4 层，kernel_size = 31
- Alignment Embedding / Phone Embedding: 384 维
- Spectrogram: 帧级别的 80 维的 fbank 特征
- 模型总参数量：67.7 M
- 训练策略：每条选择 80% 的音素对应的帧进行 mask

MaskRate	Seen MCD ↓	Unseen MCD ↓
20%	10.35	12.22
50%	7.75	9.99
80%	8.72	9.68

Table 3. MCD scores of A^{3}T pretrained in different masking rates with VCTK.



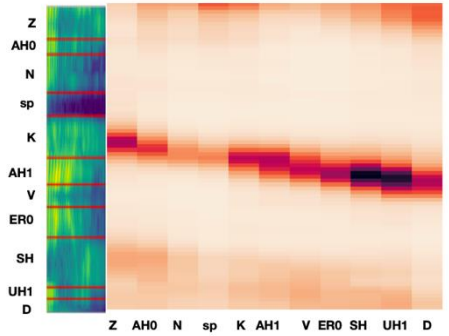
A3T: Alignment-Aware Acoustic and Text Pretraining

实验结果一：语音重建

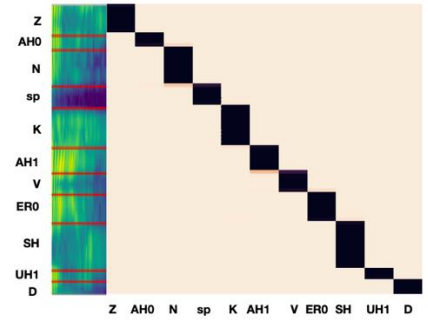
- 测试数据：30 条音频
- 测试前准备：Mask 1/3 音素对应的语音特征
- 评价指标：MCD (Mel Cepstral Distortion)
- 消融实验：

Example	Model	MCD ↓
Fig. 5(b)	A^3T	8.09
Fig. 5(c)	- Alignment Embeddings	10.73
Fig. 5(d)	- Conformer	12.43
Fig. 5(e)	- Post-Net	12.94
Fig. 5(f))	- L1 loss	11.55

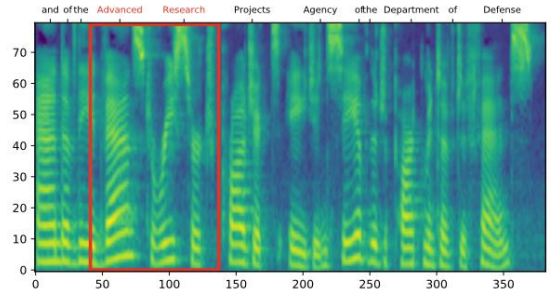
- 是否使用 alignment embedding 带来的影响



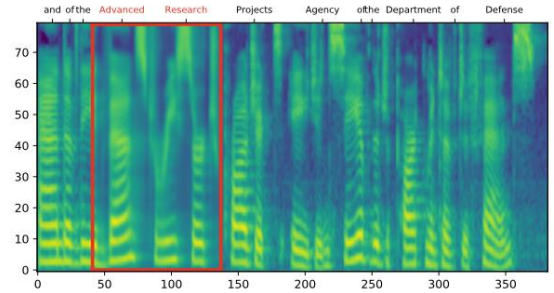
(a) Attention map of A^3T w/o alignment embeddings



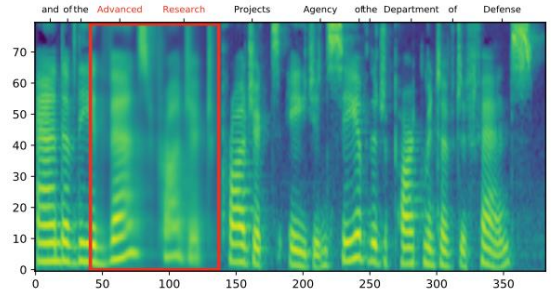
(b) Attention map of A^3T with alignment embeddings



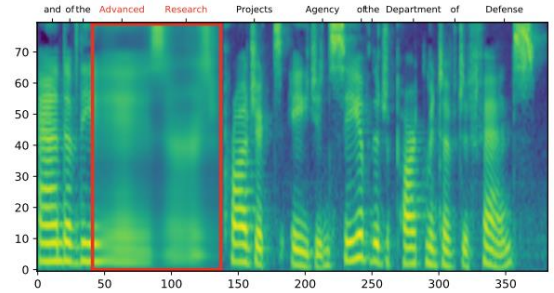
(a) Groundtruth spectrogram from LJSpeech.



(b) Reconstructed spectrogram by A^3T.



(c) Reconstructed spectrogram based on A^3T in (b) without segment embedding.



(d) Reconstructed spectrogram based on A^3T in (c) with Transformer instead of Conformer.

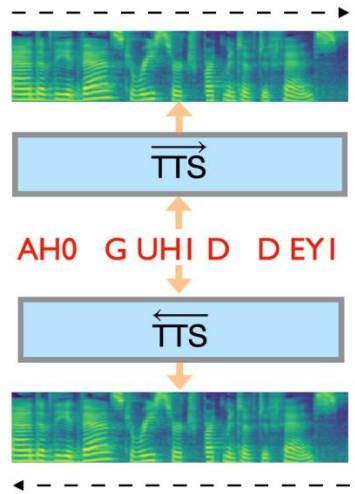
A3T: Alignment-Aware Acoustic and Text Pretraining

实验结果二：语音编辑

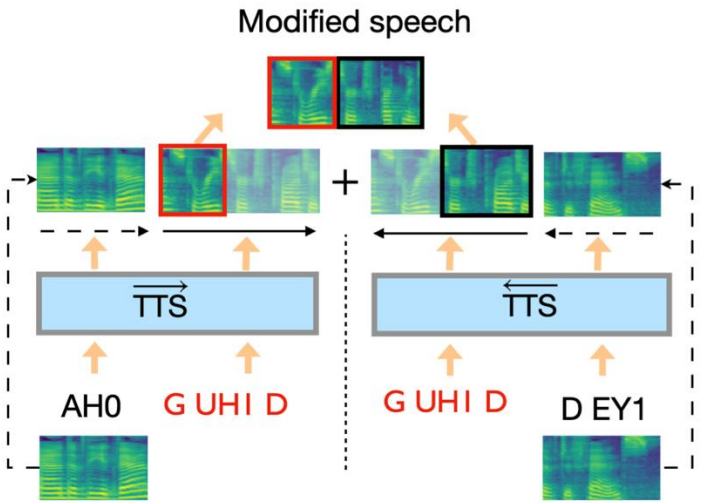
A3T 的亮点一：

完美契合语音编辑任务的预训练模式

- Baseline 1: 根据修改后的文本重新合成语音
- Baseline 2: 只合成插入文本对应的语音，插入到原始语音
- Baseline 3: 合成全部语音，根据对齐结果，将插入的部分提取出来，插入到原始语音
- Tan et al.: EditSpeech



(b) EditSpeech train



(c) EditSpeech decode

Model	VCTK MCD ↓	LJSpeech MCD ↓
Baseline 1/3	10.66	10.32
Baseline 2	12.06	10.91
A^3T	**7.76**	**9.26**
w/o Alignment Emb.	11.37	10.30

Model	Insert	Replace
Baseline 1	3.02 ± 0.20	2.64 ± 0.16
Baseline 2	2.89 ± 0.17	2.70 ± 0.16
Baseline 3	2.89 ± 0.17	2.44 ± 0.16
Tan et al. (2021)	3.50 ± 0.16	3.58 ± 0.16
A^3T	$\mathbf{3.53} \pm 0.17$	$\mathbf{3.65} \pm 0.15$
w/o Alignment Emb.	2.48 ± 0.21	1.98 ± 0.17

Tan, Daxin, et al. "Editspeech: A text based speech editing system using partial inference and bidirectional fusion." 2021 IEEE ASRU 2021.

A3T: Alignment-Aware Acoustic and Text Pretraining

实验结果三：One-Shot 语音合成

- Baseline 1: FastSpeech 2 + X-vectors
- Baseline 2: FastSpeech 2 + GST

Seen Speaker

- 30 条测试文本
- 15 个人进行评分

Unseen Speaker

- 20 条测试文本
- 15 个人进行评分

A3T 的亮点二：

将目标说话人参考音频当作 prompt → TTS 新思路

Model	Seen	Unseen
FastSpeech 2	3.33 ± 0.10	3.78 ± 0.10
+GST (Wang et al., 2018)	3.42 ± 0.10	3.81 ± 0.11
A^3T	3.61 ± 0.09	3.90 ± 0.10
Groundtruth	3.94 ± 0.08	4.09 ± 0.10

Table 6. The MOS evaluation (↑) for **speaker similarity** on multi-speaker TTS on VCTK with 95% confidence intervals. The FastSpeech2 model is equipped with X-vectors (Snyder et al., 2018).

Model	Seen	Unseen
FastSpeech 2	3.34 ± 0.11	3.85 ± 0.11
+GST (Wang et al., 2018)	3.27 ± 0.11	3.72 ± 0.11
A^3T	3.63 ± 0.10	3.94 ± 0.11
Groundtruth	4.04 ± 0.08	4.05 ± 0.10

Table 7. The MOS evaluation (↑) for **speech quality** on multi-speaker TTS on VCTK with 95% confidence intervals. The FastSpeech2 model is equipped with X-vectors (Snyder et al., 2018).

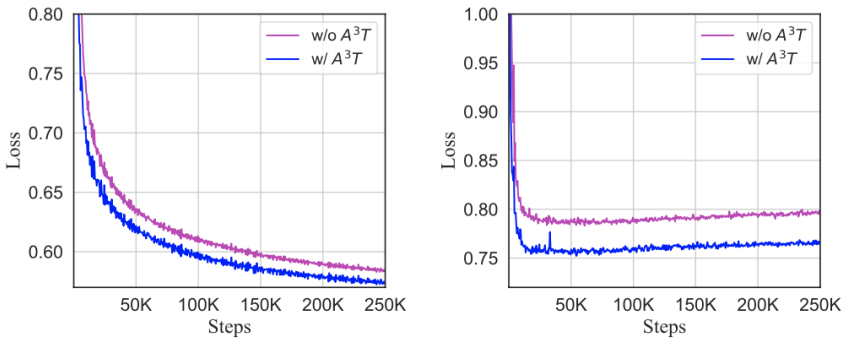
demo: https://educated-toothpaste-462.notion.site/Demo-b0edd300e6004c508744c6259369a468

A3T: Alignment-Aware Acoustic and Text Pretraining

实验结果四：针对 TTS 的 finetune

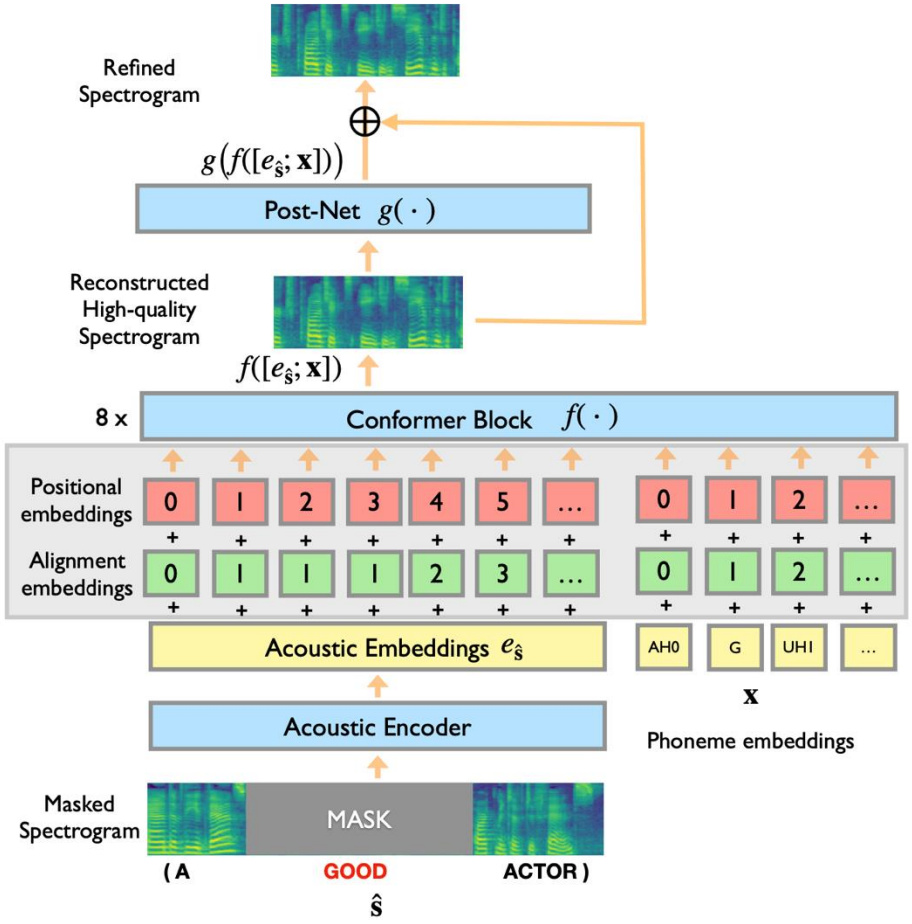
Method	Pretrain Data	MOS ↑
Groundtruth		4.07 ± 0.07
FastSpeech 2	none	3.63 ± 0.07
A^3T	LibriTTS	3.72 ± 0.07
A^3T	ASR + Speech	$\mathbf{3.77} \pm 0.07$

Table 8. MOS evaluation for multi-speaker speech synthesis.



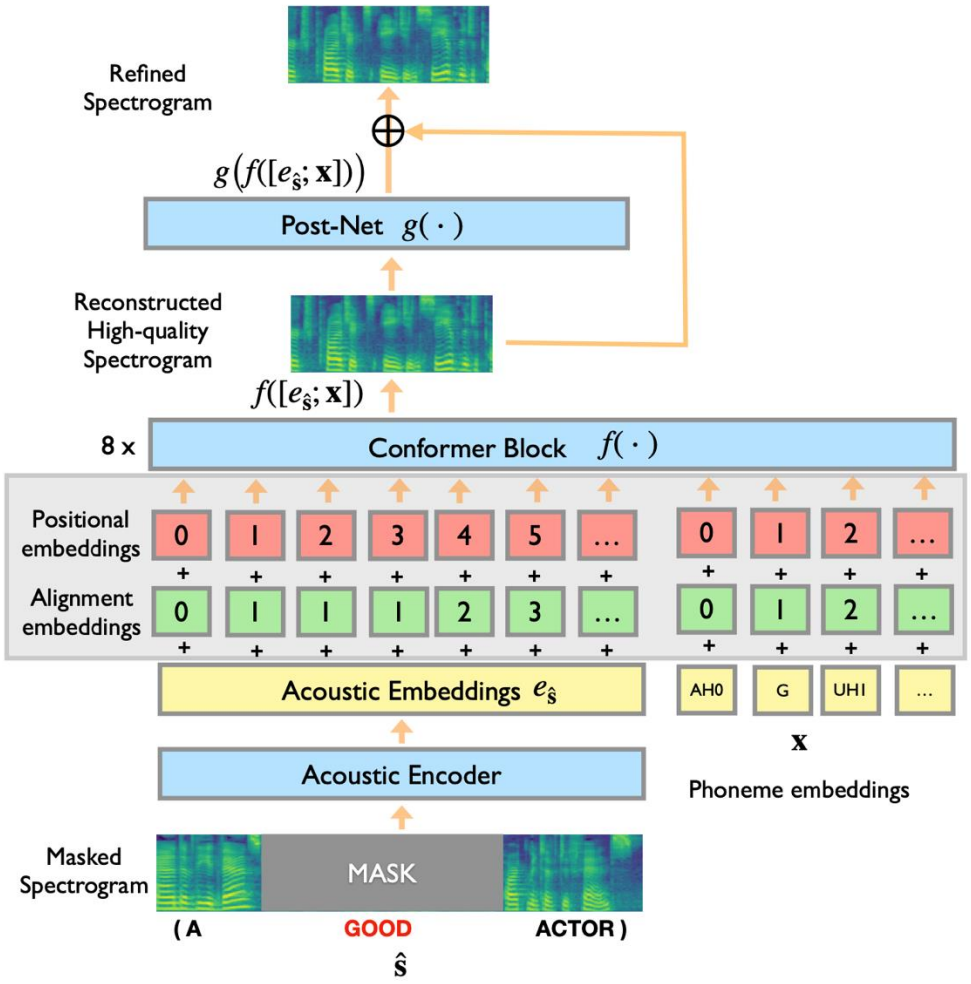
(a) Training Loss (b) Validation Loss

Figure 11. Training and validation loss using LibriTTS between TTS models initialized with and without A^3T.



ERNIE-SAT (Baidu)

https://arxiv.org/abs/2211.03545



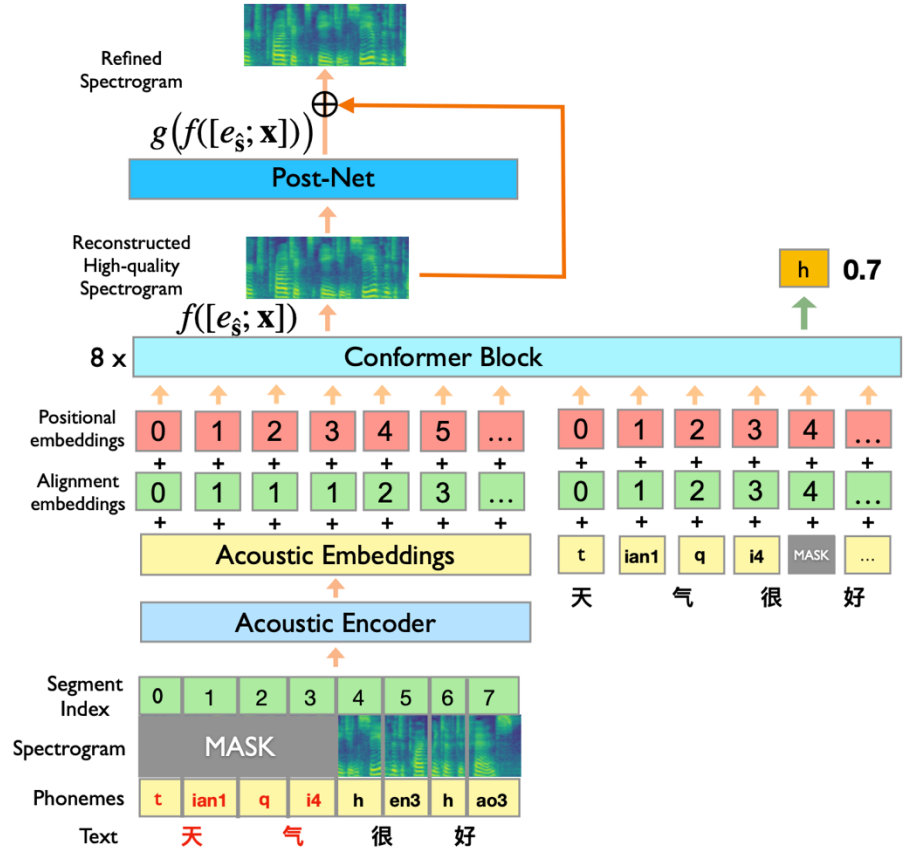


Fig. 1. ERNIE-SAT model architecture.

ERNIE-SAT (Baidu)

相对于 A3T 的修改

1. 文本增加 mask 策略

- 语音特征上被 mask 的相应音素，文本上不再做 mask

2. 中英文音素集混合，采用纯中文/纯英文语料训练

- 使得模型具备 cross-lingual 的能力

目标：将 A3T 的单语种扩充至 Cross-Lingual

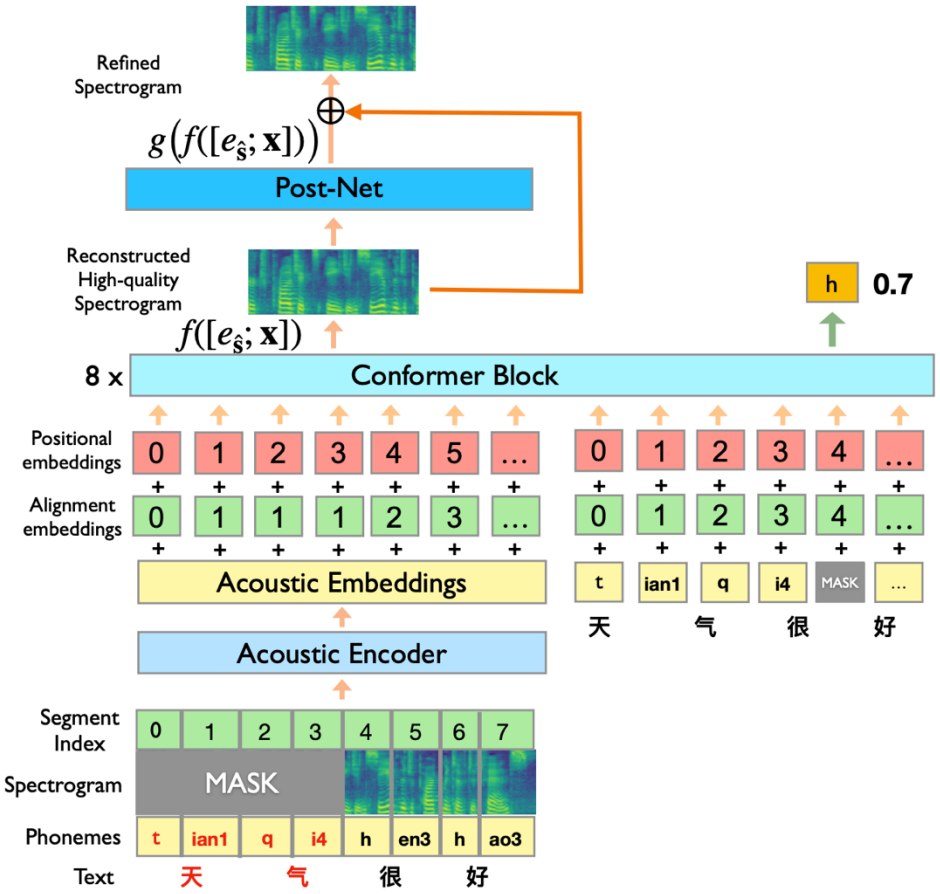


Fig. 1. ERNIE-SAT model architecture.

demo: https://brainy-candy-0a2.notion.site/ERNIE-SAT-DEMO-6f3ef7fbea944d9db46ba2770ab6693d

ERNIE-SAT (Baidu)

实验一：跨语种音色克隆

Model	Unseen
Tacotron 2 + X-vectors + GST	3.33 ± 0.16
FastSpeech 2 + X-vectors + GST	3.49 ± 0.14
ERNIE-SAT	3.58 ± 0.14

Table 1. The MOS for **speech quality** on cross-lingual multi-speaker TTS with 95% confidence intervals.

Model	Unseen
Tacotron 2 + X-vectors + GST	3.30 ± 0.17
FastSpeech 2 + X-vectors + GST	3.45 ± 0.16
ERNIE-SAT	3.53 ± 0.11

Table 2. The MOS for **speaker similarity** on cross-lingual multi-speaker TTS with 95% confidence intervals.

实验二：跨语种语音编辑

Model	Unseen
Tacotron 2 + X-vectors + GST	3.07 ± 0.18
FastSpeech 2 + X-vectors + GST	3.37 ± 0.15
ERNIE-SAT	3.45 ± 0.17

Table 3. The MOS for **speech quality** on cross-lingual multi-speaker speech editing with 95% confidence intervals.

Model	Unseen
Tacotron 2 + X-vectors + GST	3.10 ± 0.19
FastSpeech 2 + X-vectors + GST	3.34 ± 0.15
ERNIE-SAT	3.51 ± 0.11

Table 4. The MOS for **speaker similarity** on cross-lingual multi-speaker speech editing with 95% confidence intervals.

Rethinking on a Unified Paradigm for Generation and Editing

A3T: From Text to Speech

应用场景：语音编辑/语音合成

输入输出及目标

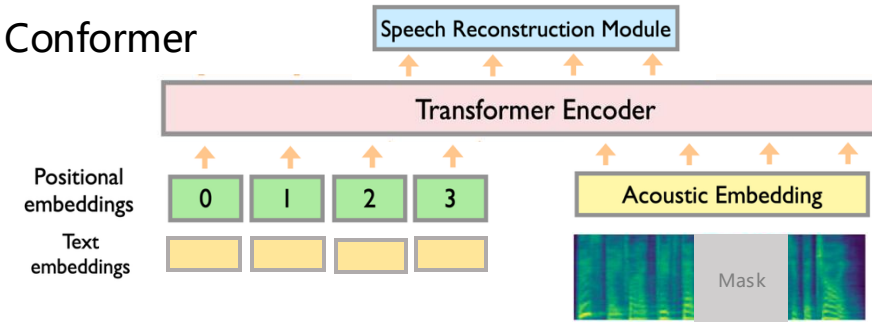
- 完整输入的模态：文本
- 加入 Mask 的模态：语音
- 预测的目标模态：语音
- 预测目标：预测被 mask 的语音特征

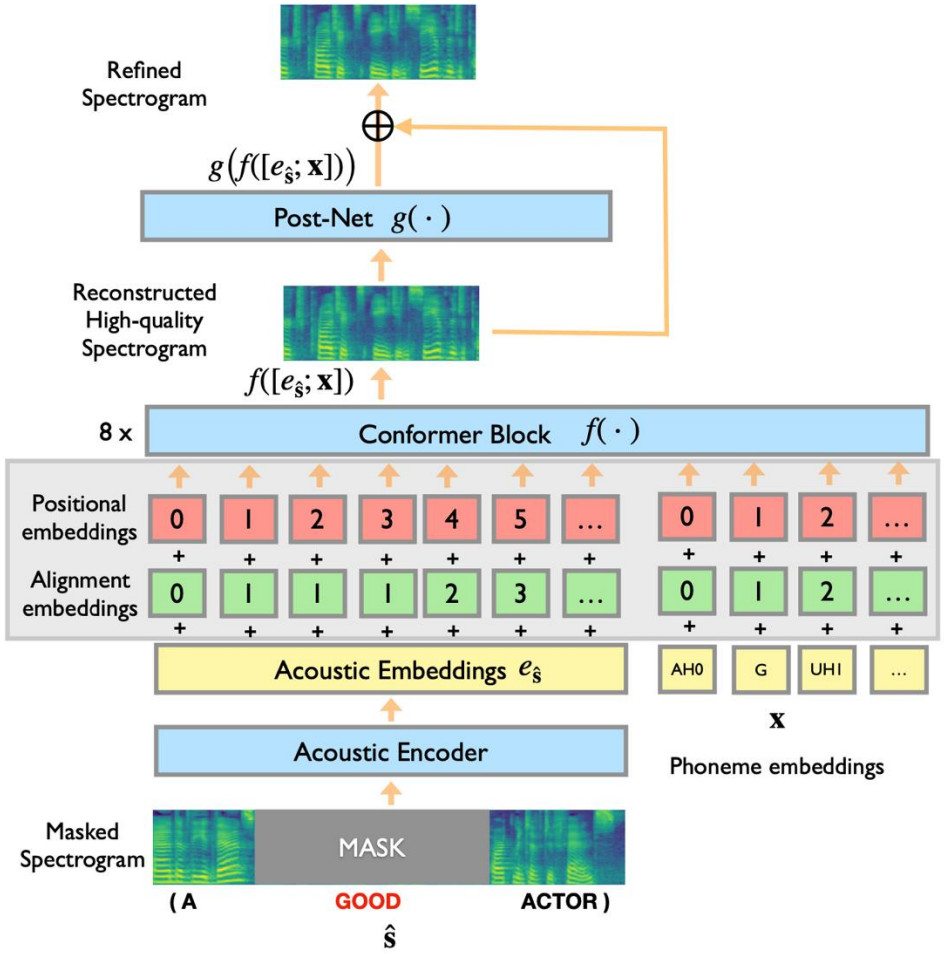
各自模态的处理

- 输入模态 Encoder：Text embedding
- 目标模态 Encoder: Acoustic Encoder

多模态融合的方式

- 多模态融合方式：对齐编码 & 拼接输入
- 多模态 Encoder: Conformer





FAT-MLM：From Speech to Text

应用场景：语音识别/语音翻译

输入输出及目标

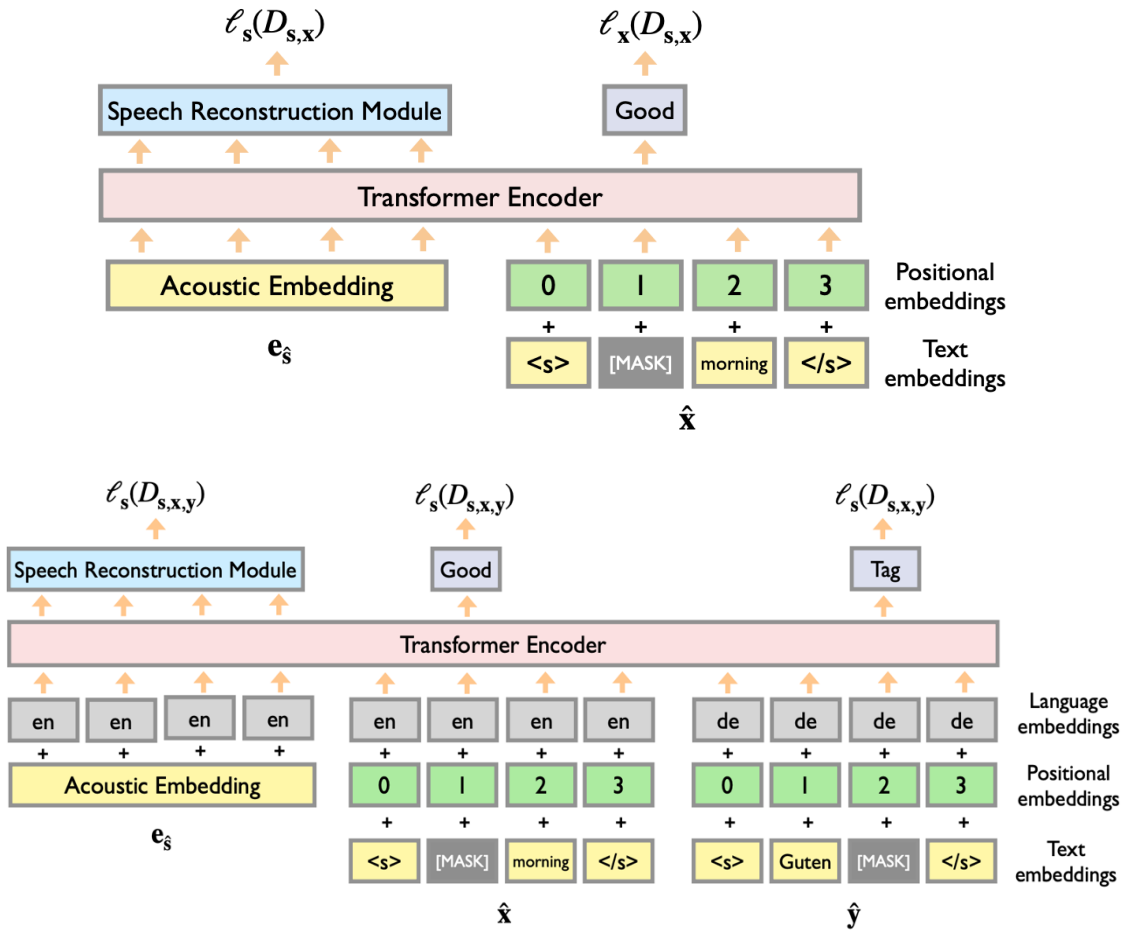
- 输入的两个模态：语音和文本 (均采取 mask 策略)
- 预测的目标模态：语音和文本
- 预测目标：预测被 mask 的语音特征和文本

各自模态的处理

- 输入模态 Encoder：Text embedding
- 目标模态 Encoder：Transformer

多模态融合的方式

- 多模态融合方式：位置编码 & 拼接输入
- 多模态 Encoder：Transformer



- Zheng, Renjie, et al. "Fused acoustic and text encoding for multimodal bilingual pretraining and speech translation." PMLR, 2021.

BEIT-3: From Image to Text

应用场景：图像描述 (Image Captioning)

输入输出及目标

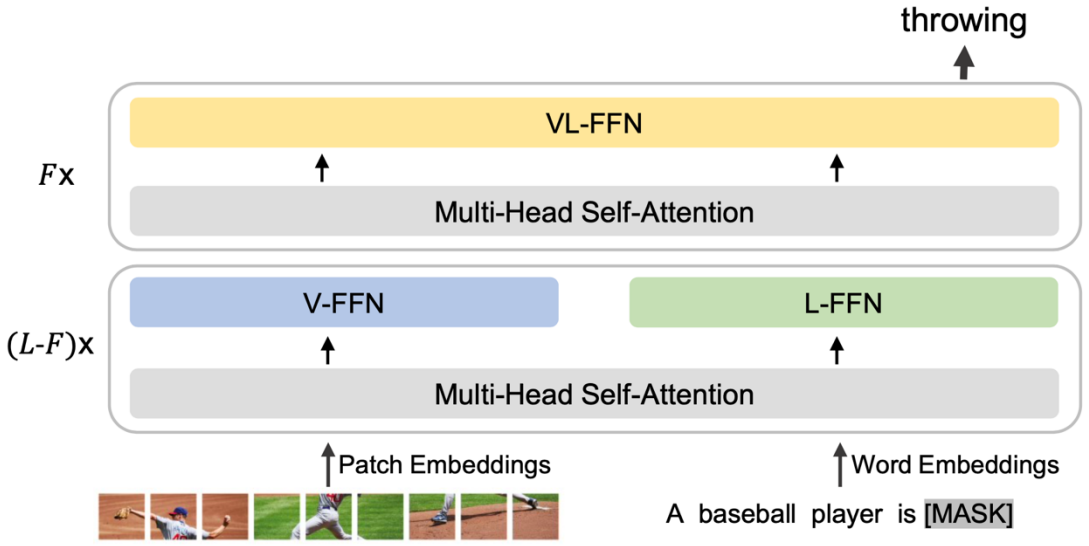
- 完整输入的模态：图像
- 加入 Mask 的模态：文本
- 预测的目标模态：文本
- 预测目标：预测被 mask 的文本

各自模态的处理

- 输入模态 Encoder：Text embedding
- 目标模态 Encoder：Patch Embedding

多模态融合的方式

- 多模态融合方式：拼接输入
- 多模态 Encoder：Transformer



(e) Image-to-Text Generation

Image Captioning (COCO)

- Wang, Wenhui, et al. "Image as a foreign language: Beit pretraining for all vision and vision-language tasks." arXiv preprint arXiv:2208.10442 (2022).

MUSE: From Text to Image

Muse: Text-To-Image Generation via Masked Generative Transformers

Huiwen Chang * Han Zhang * Jarred Barber † AJ Maschinot † José Lezama Lu Jiang Ming-Hsuan Yang
Kevin Murphy William T. Freeman Michael Rubinstein † Yuanzhen Li † Dilip Krishnan †

Google Research

https://arxiv.org/abs/2301.00704

推理速度比Stable Diffusion快2倍，生成、修复图像谷歌一个模型搞定，实现新SOTA

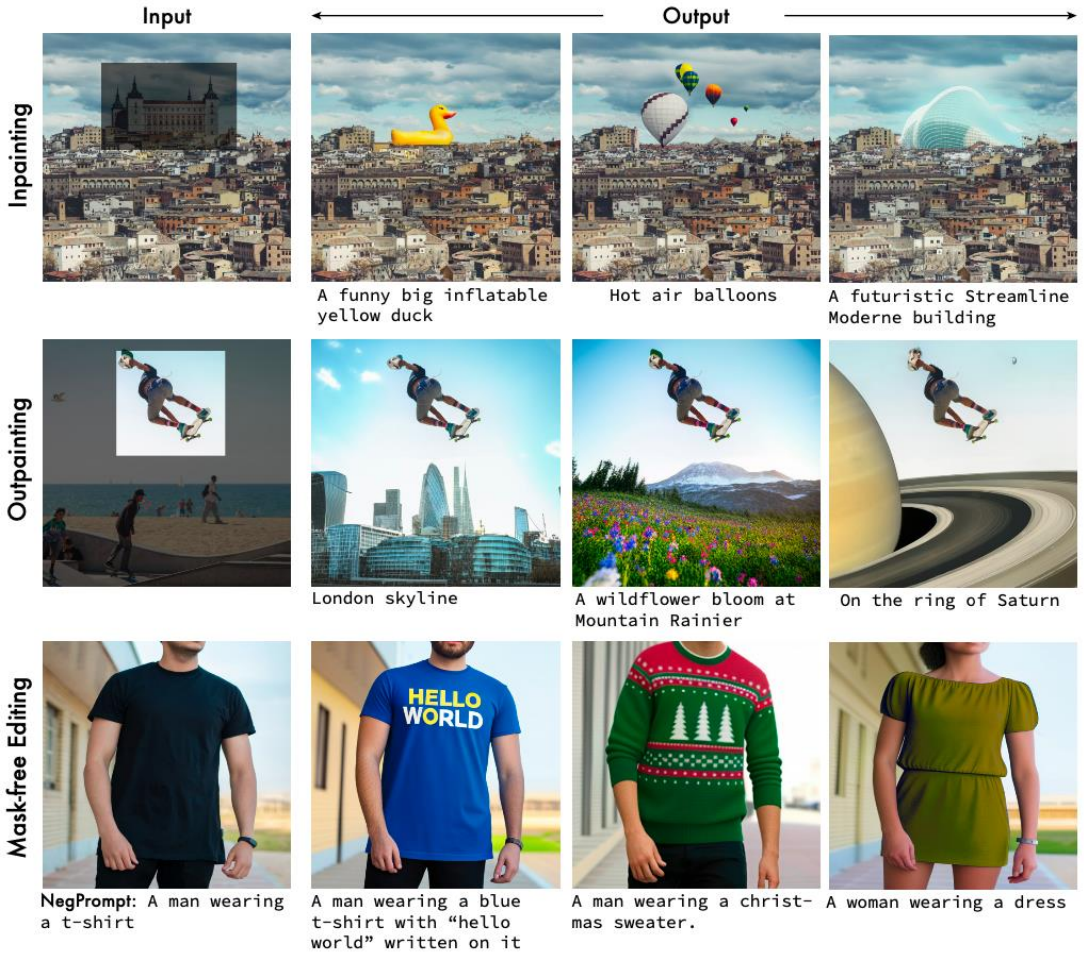
机器之心 2023-01-05 13:10 发表于北京

机器之心报道

机器之心编辑部

> 图像生成领域越来越卷了！

文本到图像生成是 2022 年最火的 AIGC 方向之一，被《science》评选为 2022 年度十大科学突破。最近，谷歌的一篇文本到图像生成新论文《Muse: Text-To-Image Generation via Masked Generative Transformers》又引起高度关注。



MUSE: From Text to Image

应用场景：图像编辑/图像生成

输入输出及目标

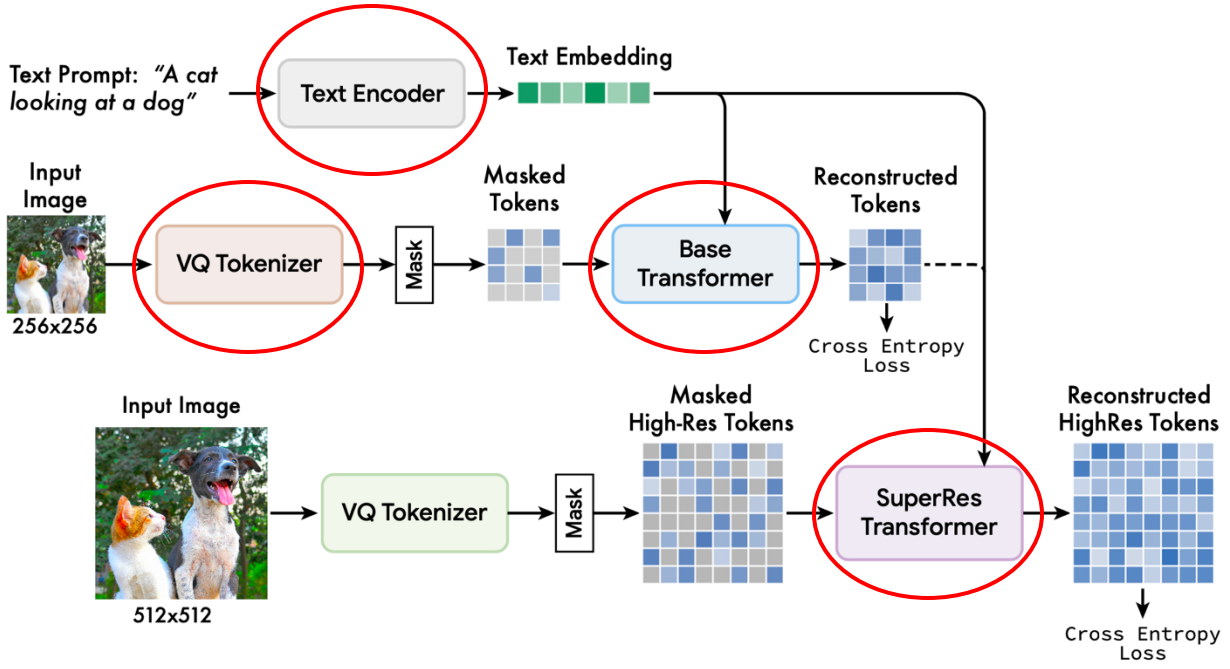
- 完整输入的模态：文本
- 加入 Mask 的模态：图像
- 预测的目标模态：图像
- 预测目标：预测被 mask 的图像

各自模态的处理

- 文本模态 Encoder：Text Encoder (LLM)
- 图像模态 Encoder：VQ Tokenizer

多模态融合的方式

- 多模态融合方式：拼接 & Cross-Attention
- 多模态 Encoder：Transformer



MUSE: From Text to Image

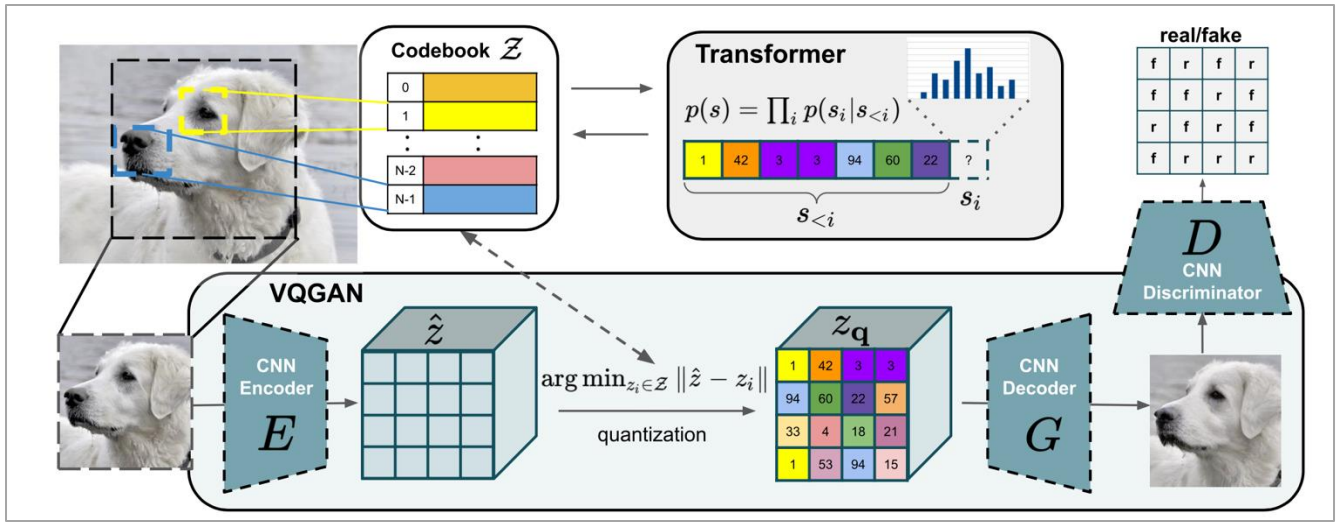
1. Text Encoder: T5-XXL 大型语言模型 (LLM) 编码器

- 物体（名词）
- 动作（动词）
- 特点（形容词）
- 空间关系（介词）
- 个数及组合关系...

to high-quality image generation. The embeddings extracted from an LLM such as T5-XXL (Raffel et al., 2020) carry rich information about objects (nouns), actions (verbs), visual properties (adjectives), spatial relationships (prepositions), and other properties such as cardinality and composition. Our hypothesis is that the Muse model learns to map these rich

2. VQ Tokenizer: 将图片转为 token

- VQ-GAN：完全卷积的 Encoder / Decoder
- Encoder 进行降采样，Decoder 进行升采样
- 低分辨率：256 × 256，f = 16, 16 × 16 个 token
- 高分辨率：512 × 512，f = 8， 64 × 64 个 token
- 双重作用
  - 提取图像中高级别的语义信息
  - 降低后续 Transformer 的计算量，提高模型推理效率



MUSE: From Text to Image

3. Base Transformer:

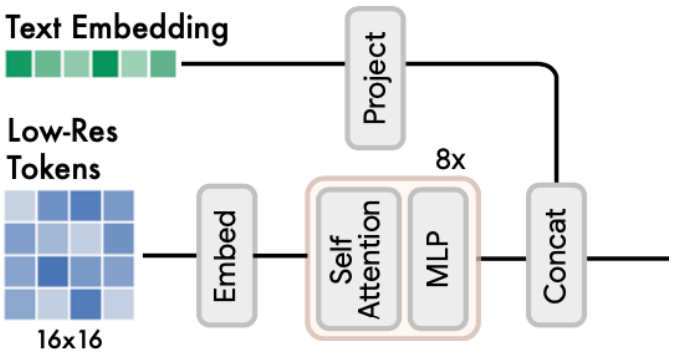
- 文本不做 mask，图像 token 进行 mask
- 2-D positional encoding

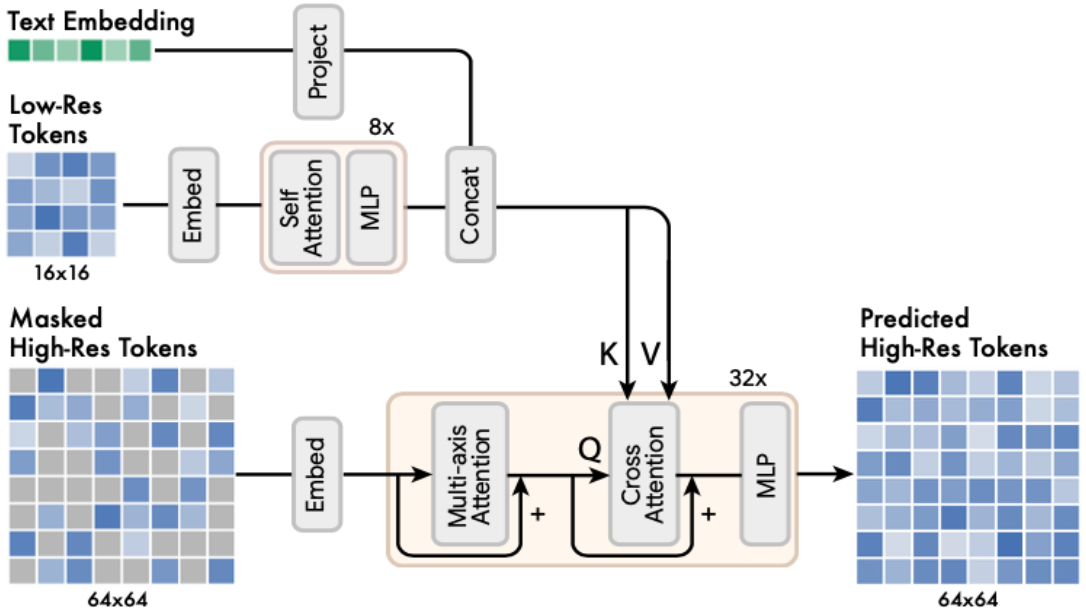
- 输出经过全连接，预测图像 token 对应的 id
- 预测 mask 部分的 token 类别，交叉熵损失函数

4. SuperRes Transformer:

- 使用另一个 VQ Tokenizer

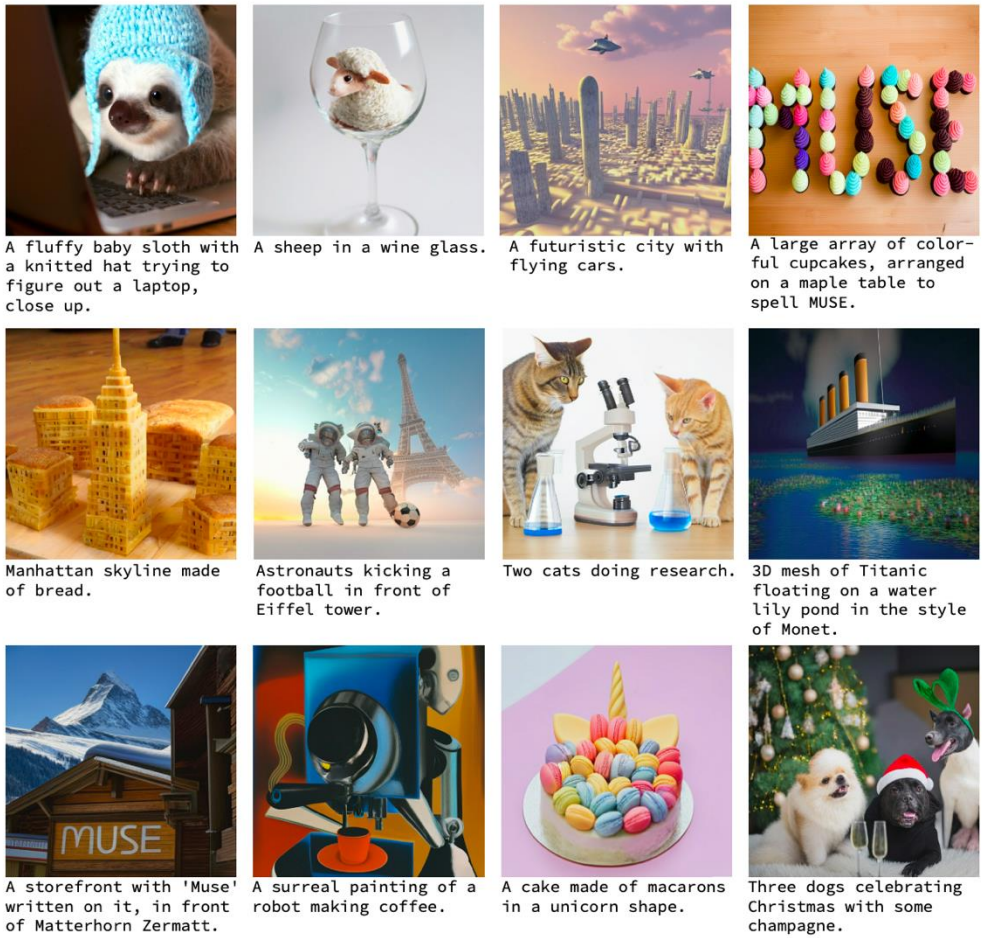
生成的图片是 VQ-GAN Decoder 恢复得到的，
固定其他参数，使用数据微调 VQ-GAN Decoder



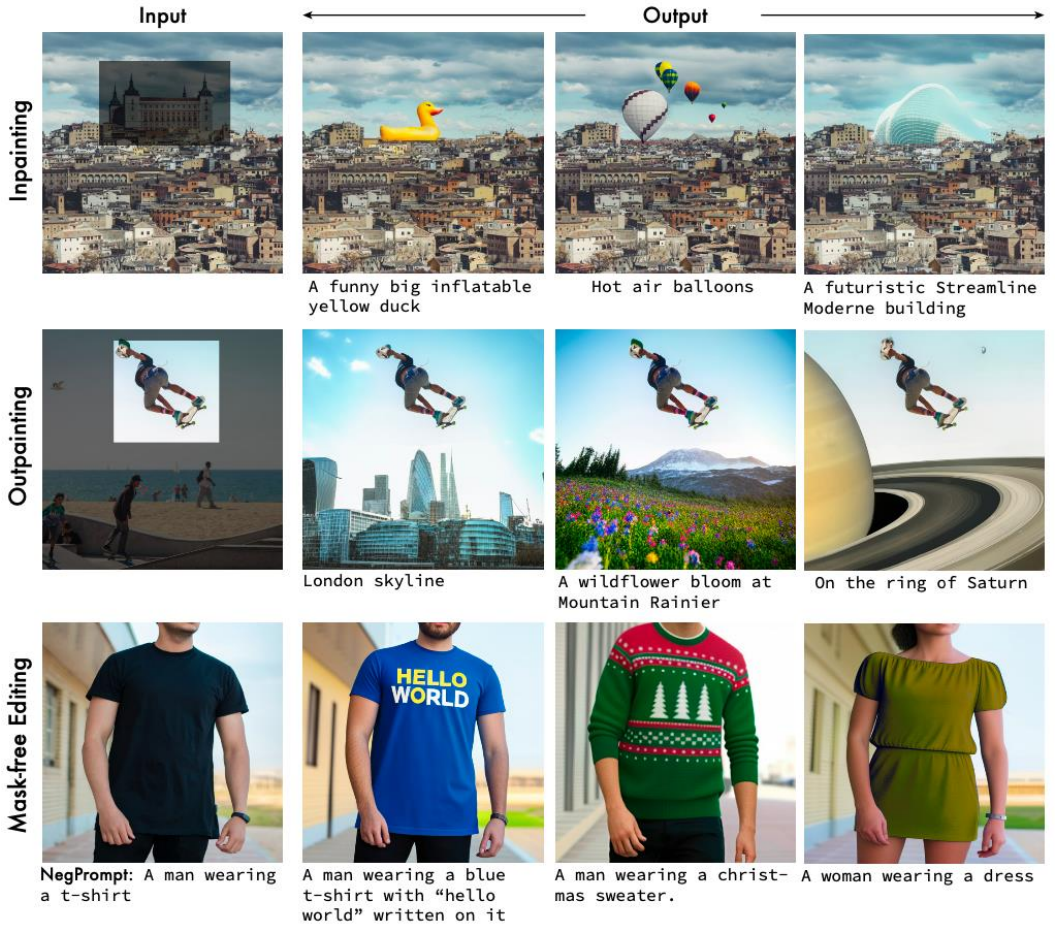


MUSE: From Text to Image

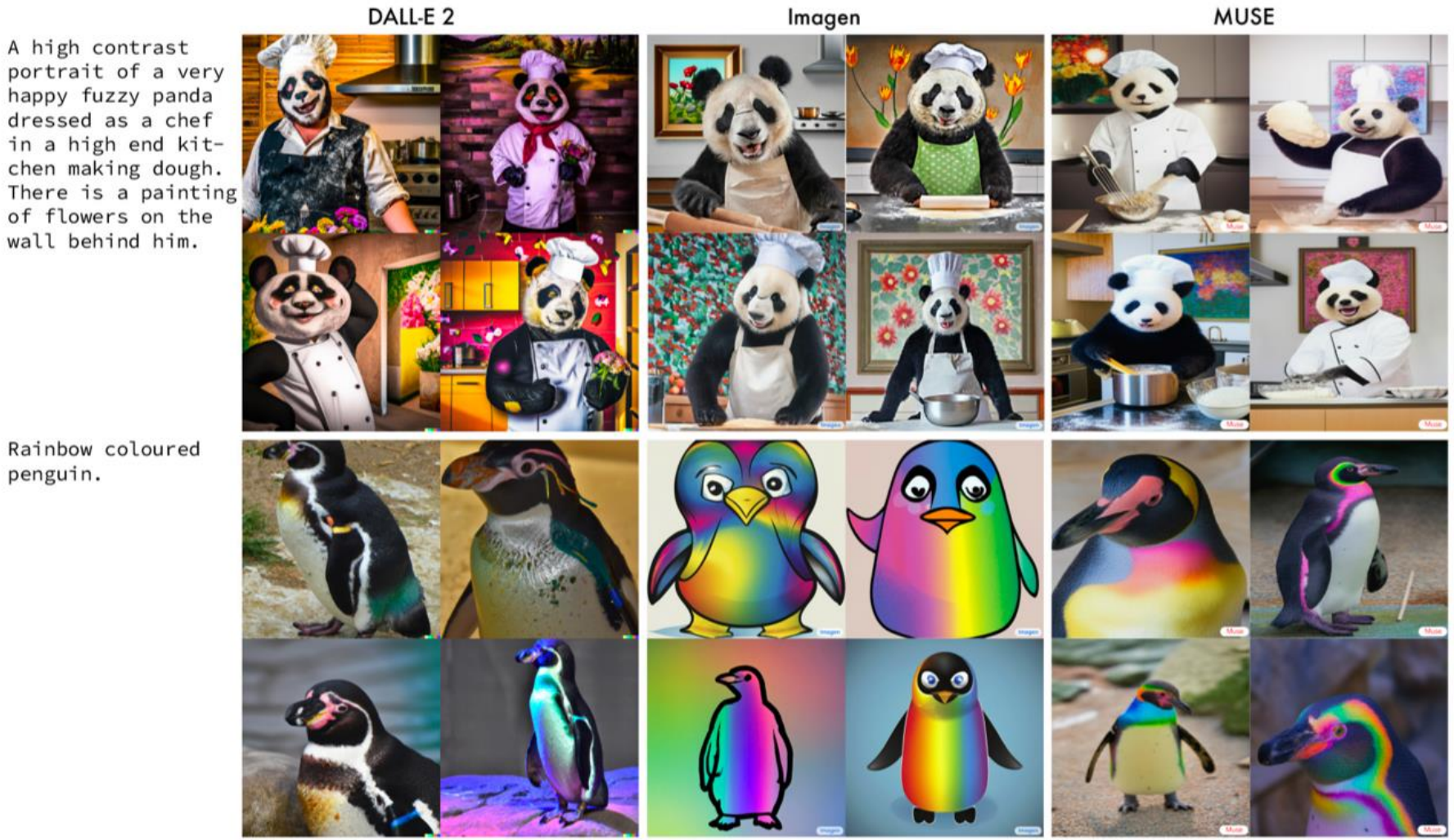
应用一：图像生成



应用二：图像编辑



MUSE: From Text to Image



MUSE: From Text to Image

A stack of 3 books. A green book is on the top, sitting on a red book. The red book is in the middle, sitting on a blue book. The blue book is on the bottom.

A real flamingo reading a large open book. a big stack of books is piled up next to it. dslr photograph

MUSE: From Text to Image

效果又快又好，担心为坏人所用，
所以先不开源，也不让测试 demo

Fr´echet Inception Distance (FID)

直观感受，**FID是反应生成图片和真实图片的距离，数据越小越好。**专业来说，FID是衡量两个多元正态分布的距离，其公式如下

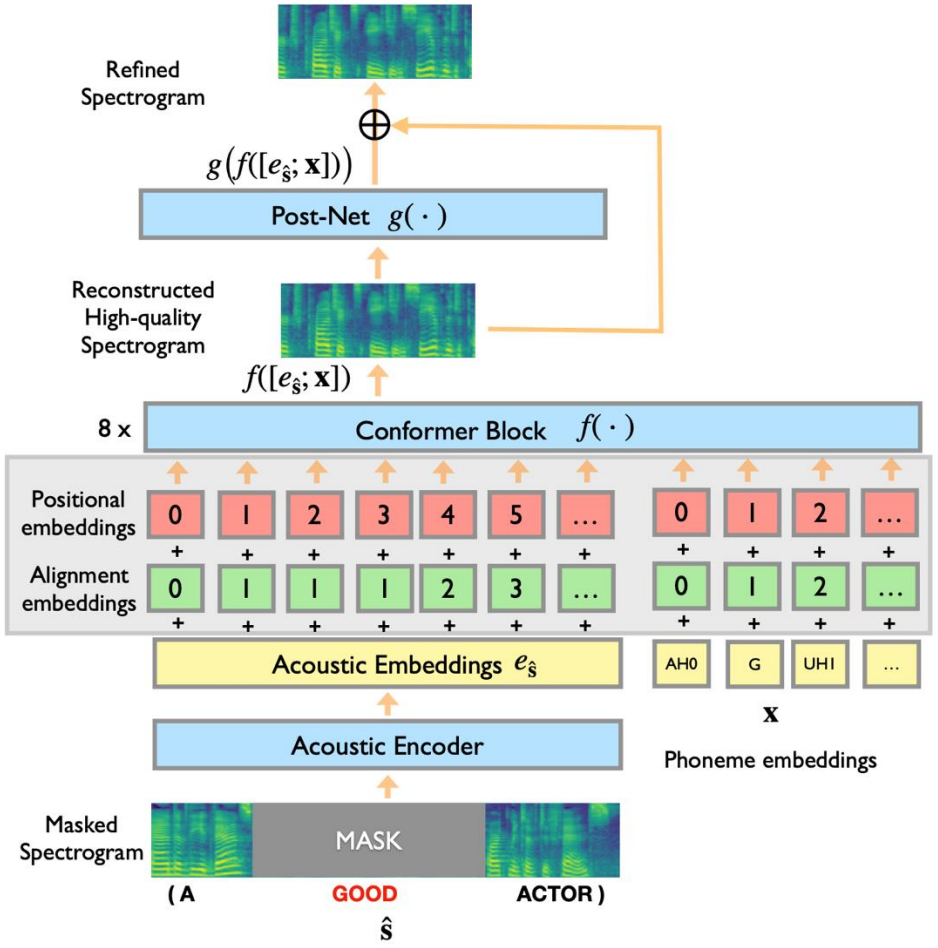
$$FID = ||\mu_r - \mu_g||^2 + Tr(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2})$$

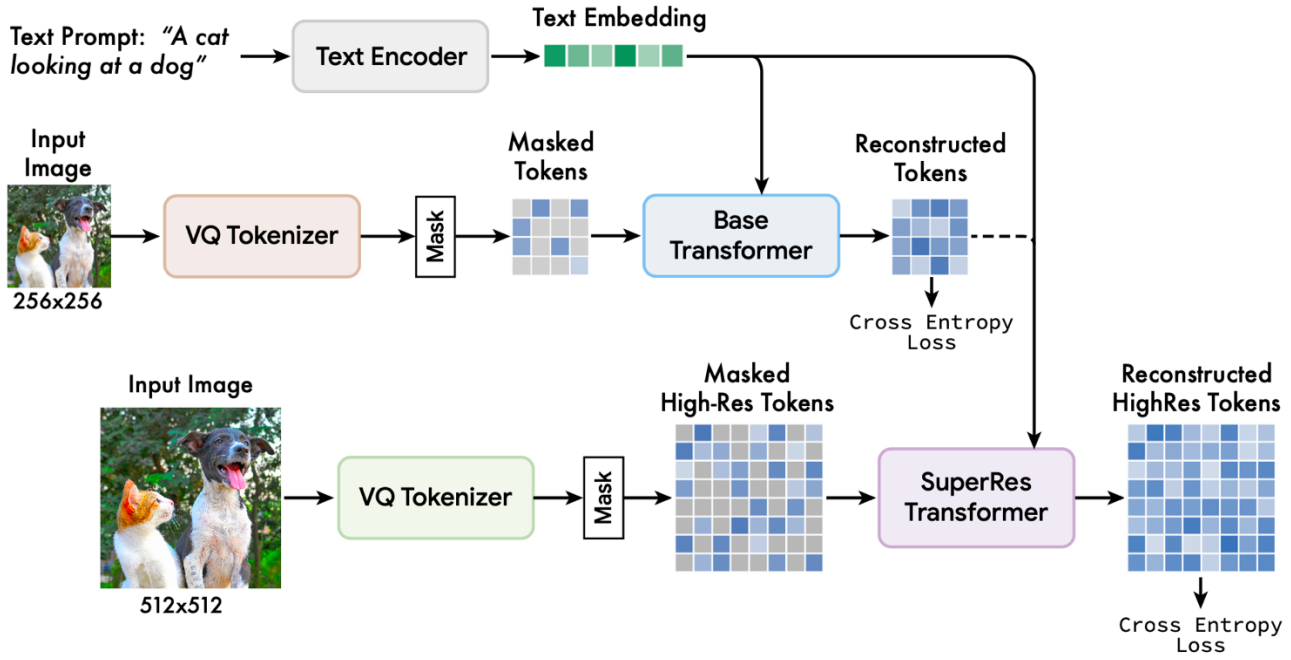
- μ_r ： 真实图片的特征均值
- μ_g ： 生成图片的特征均值
- Σ_r ： 真实图片的协方差矩阵
- Σ_g ： 生成图片的协方差矩阵
- Tr ： 迹

Approach	Model Type	Params	FID-30K	Zero-shot FID-30K
AttnGAN (Xu et al., 2017)	GAN		35.49	-
DM-GAN (Zhu et al., 2019)	GAN		32.64	-
DF-GAN (Tao et al., 2020)	GAN		21.42	-
DM-GAN + CL (Ye et al., 2021)	GAN		20.79	-
XMC-GAN (Zhang et al., 2021)	GAN		9.33	-
LAFITE (Zhou et al., 2021)	GAN		8.12	-
Make-A-Scene (Gafni et al., 2022)	Autoregressive		7.55	-
DALL-E (Ramesh et al., 2021)	Autoregressive		-	17.89
LAFITE (Zhou et al., 2021)	GAN		-	26.94
LDM (Rombach et al., 2022)	Diffusion		-	12.63
GLIDE (Nichol et al., 2021)	Diffusion		-	12.24
DALL-E 2 (Ramesh et al., 2022)	Diffusion		-	10.39
Imagen-3.4B (Saharia et al., 2022)	Diffusion		-	7.27
Parti-3B (Yu et al., 2022)	Autoregressive		-	8.10
Parti-20B (Yu et al., 2022)	Autoregressive		3.22	7.23
Muse-3B	Non-Autoregressive		-	7.88

Model	Resolution	Time
Imagen	256 × 256	9.1s
Imagen	1024 × 1024	13.3s
LDM (50 steps)	512 × 512	3.7s
LDM (250 steps)	512 × 512	18.5s
Parti-3B	256 × 256	6.4s
Muse-3B	256 × 256	0.5s
Muse-3B	512 × 512	1.3s

MUSE: From Text to Image





A Unified Paradigm

- **「修复」/「编辑」**通常针对语音和图像模态

其他应用

- 语音和图像间直接的转换，文本通常充当媒介：图像 → 语音？语音 → 图像？
- 各个模态的内部转换：文本 → 文本 (翻译)，语音 → 语音 (speech-to-speech 翻译)，图像 → 图像 (image2image)

论文工作	A3T	FAT-MLM	BEIT_V3	MUSE
应用场景	语音合成/语音编辑	语音识别/语音翻译	图像描述	图像生成/图像编辑
A 模态	文本	语音	图像	文本
B 模态	语音	文本	文本	图像
被 mask 的模态	语音	语音和文本	文本	图像
A 模态 Encoder	embedding	embedding	embedding	PLM
B 模态 Encoder	FNN	Transformer	embedding	VQ-Tokenizer
模态融合方式	对齐编码+拼接输入	位置编码+拼接输入	拼接输入	拼接输入+Cross-Attention
多模态 Encoder	Conformer	Transformer	Transformer	Transformer
预测目标	被 mask 的语音特征	被 mask 的语音和文本	被 mask 的文本	被 mask 的图像

...

Recent Advances on Text-Based Speech Editing (Optional)

Speech Inpainting → Speech Editing

基于文本的语音编辑 (MSRA)

对于插入的文本，联合预测 duration 时长，此时相当于 Speech Inpainting 的建模问题

语音恢复的缺点分析

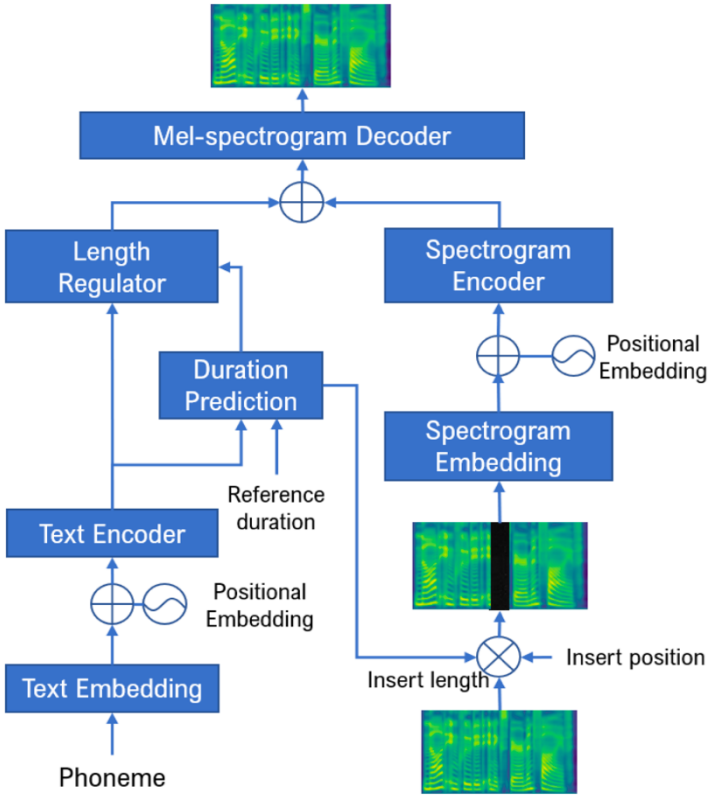
- 能够预测的时长短
- 需要预留/插入相应的 gap，使用不灵活

Table 2: *Speech naturalness Mean Opinion Score (MOS) test*

Method	MOS
Ground truth	4.57
Baseline[8]	2.66
Ours	**3.86**

Table 4: *Speaker similarity Mean Opinion Score (MOS)*

Method	MOS
Baseline[8]	2.69
Ours	**3.41**



- Tang, Chuanxin, et al. "Zero-Shot Text-to-Speech for Text-Based Insertion in Audio Narration." *arXiv preprint arXiv:2109.05426* (2021).

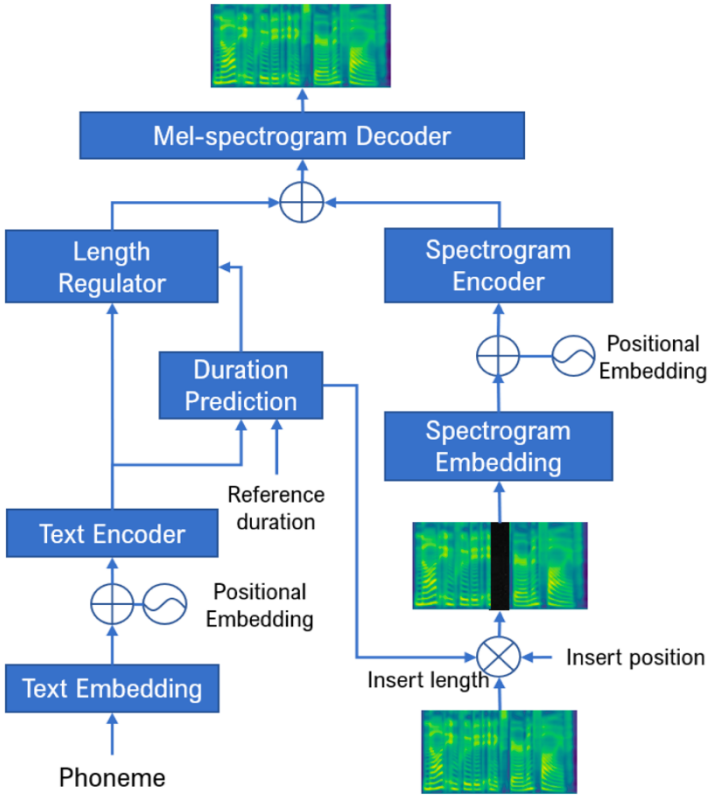
Speech Inpainting → Speech Editing

基于文本的语音编辑 (MSRA)

对于插入的文本，联合预测 duration 时长，此时相当于为 Speech Inpainting 的建模问题

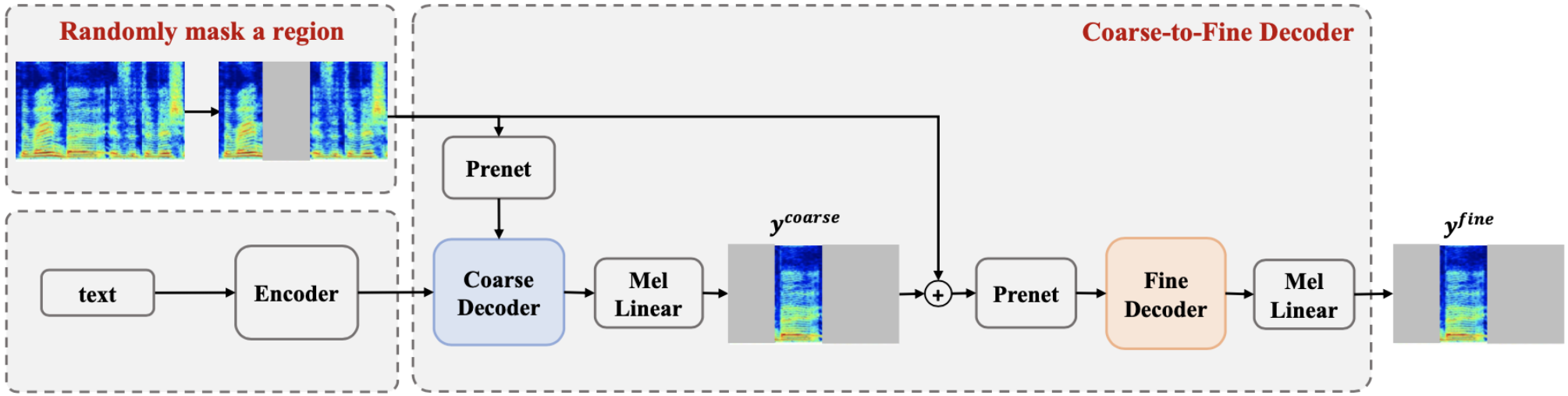
语音编辑的流程梳理

- 将原始文本与现有语音进行强制对齐，找到编辑区域
- 基于修改后文本，预测编辑区域的发音时长
  - 插入对应时长的 mask
  - 将编辑区域替换为 mask
- 预测编辑区域被 mask 的梅尔谱
  - 合成语音后，再与非编辑区域的原始波形拼接
  - 与非编辑区域的原始梅尔谱拼接，再合成波形



- Tang, Chuanxin, et al. "Zero-Shot Text-to-Speech for Text-Based Insertion in Audio Narration." *arXiv preprint arXiv:2109.05426* (2021).

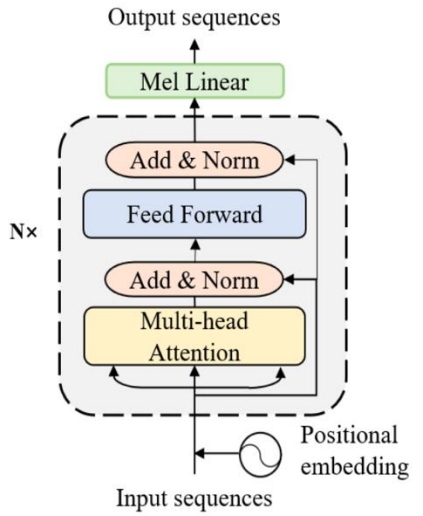
CampNet (中科院自动化所)



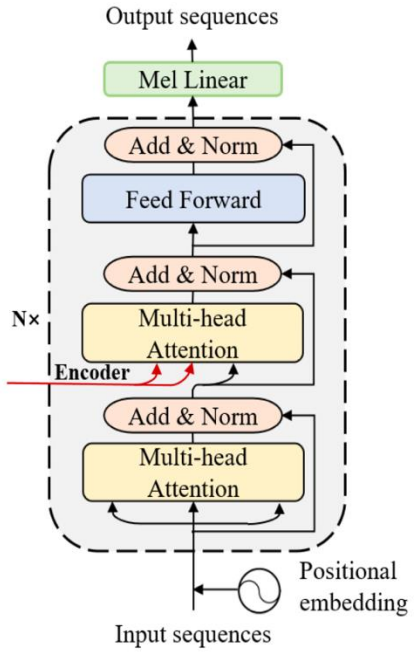
System A	Scores A(%)	N/P Neural(%)	Scores B(%)	System B
CoareDecoder	22.25	47.00	30.75	OneDecoder
FineDecoder	43.25	32.75	24.00	OneDecoder
FineDecoder	44.75	35.00	20.25	CoareDecoder

- Wang, Tao, et al. "CampNet: Context-Aware Mask Prediction for End-to-End Text-Based Speech Editing." arXiv preprint arXiv:2202.09950 (2022).

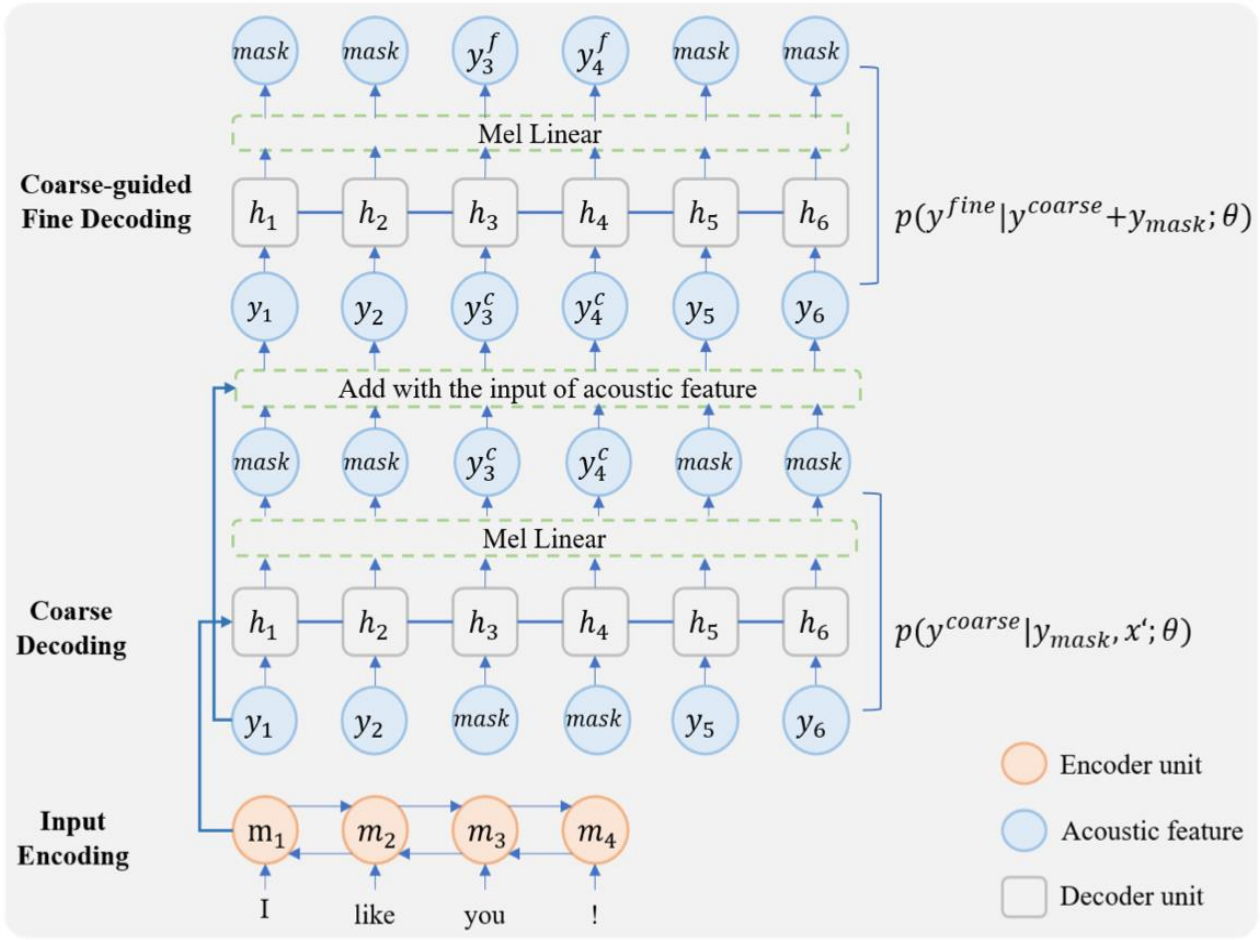
CampNet (中科院自动化所)



(a) The structure of Encoder and Fine Decoder

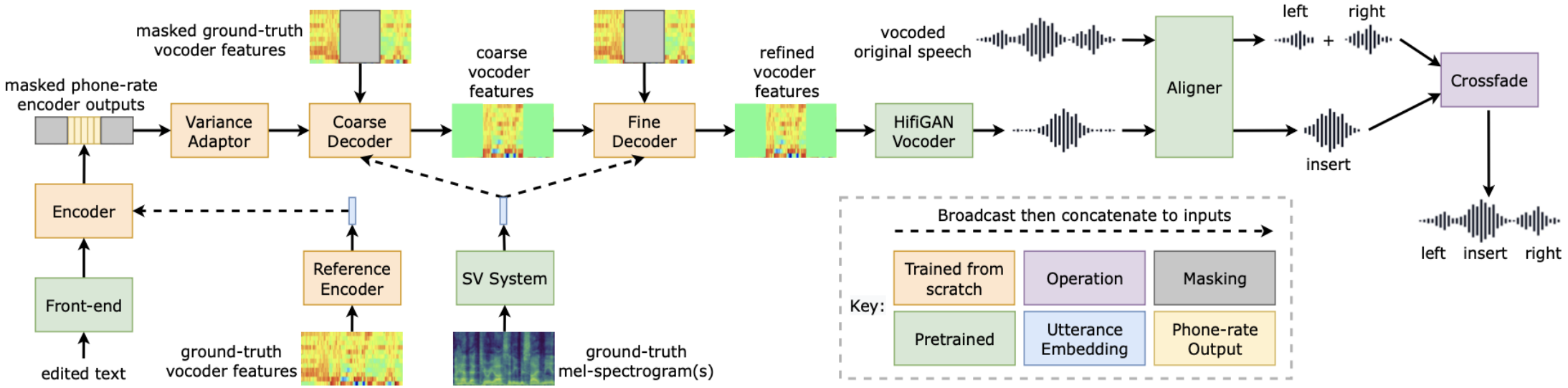


(b) The structure of Coarse Decoder



- Wang, Tao, et al. "CampNet: Context-Aware Mask Prediction for End-to-End Text-Based Speech Editing." arXiv preprint arXiv:2202.09950 (2022).

Text-Based Voice Editing, TBVE (Meta)



	No Embedding			Spk Embedding			Utt Embedding		
	Clar	Spk	Pros	Clar	Spk	Pros	Clar	Spk	Pros
No Ref Enc	3.94	4.03	3.92	3.98	4.02	3.93	3.86	4	3.96
Ref Enc	4.04	4.02	4.02	3.97	4.02	3.91	3.96	3.98	4.02

- Fong, Jason, et al. "Towards zero-shot Text-based voice editing using acoustic context conditioning, utterance embeddings, and reference encoders." arXiv preprint arXiv:2210.16045 (2022).

Thanks!