Recent Advances on Unified Model for Voice and Audio Generation (Part 2)

2023-11-20

Outline

- **Part 1:** Purely LLM based Audio Generation
  - Speech-X, **Make-A-Voice / MVoice**
  - UniAudio
- **Part 2:** Incorporating Diffusion into Audio Generation
  - ☆ **AudioLDM 2**
- **Discussion :** Beyond Unified Model for Audio Generation
  - Towards Final Fantasy: Any-to-Any Generation
- Appendix: Understanding Diffusion Models

回顾：LLM Based Audio Generation Foundation Model

特点一：音频的离散化

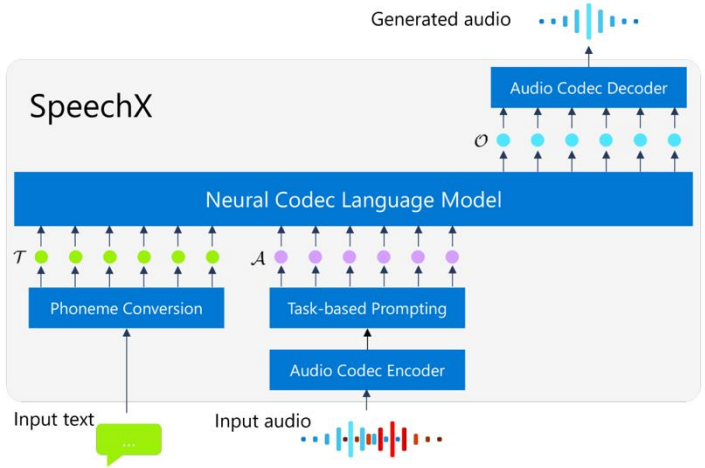
- 音频离散化 + LLM 建模，成为一种主流选择
  - 抛弃连续表征，效果能否达到足够好？
  - HiFi-Codec / FunCodec 等优化 codec 的工作

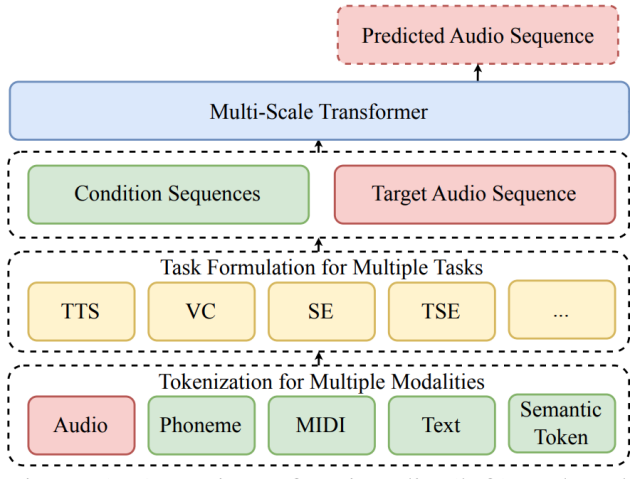
特点二：Unified LLM（多任务）

- 单个模型解决所有任务 or 各模型
  - 假设：不同类型的音频之间是有一些共享的信息的

特点三：解决长序列建模问题

- 方案一：自回归+非自回归
  - VALL-E / Speech-X
- 方案二：减少建模的序列长度
  - Make-A-Voice / UniAudio （RVQ 层数为 3）
- 方案三：使用更适合长序列建模的 Transformer
  - MVoice / UniAudio





UniAudio

Incorporating Diffusion into Audio Generation

AudioLDM 2: Learning Holistic Audio Generation with Self-supervised Pretraining

- 论文将 LLM 与 Diffusion 相结合，提出一种音频生成的通用解决方案
- 思路：采用通用的语音表征（LOA，Language Of Audio） 作为中间特征

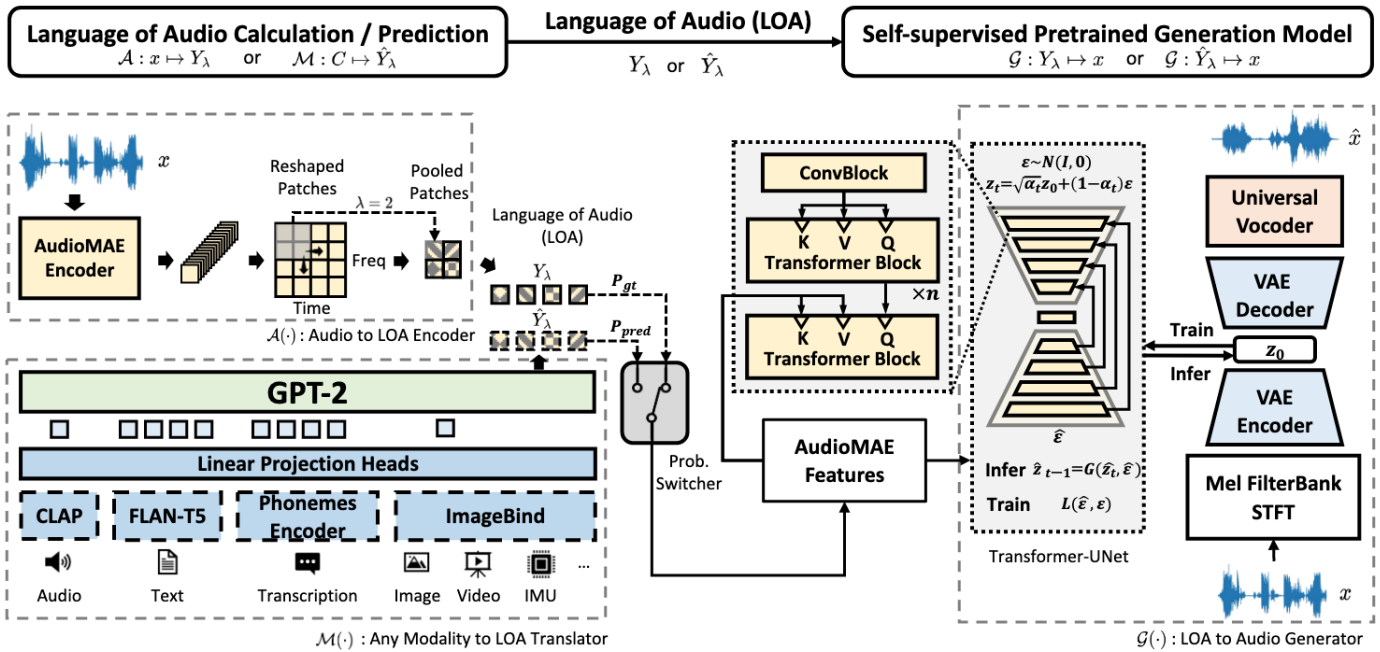


Fig. 1. The overview of the *AudioLDM 2* architecture. The AudioMAE feature is a proxy that bridges the *conditioning information to LOA translation* stage (modelled by GPT-2) and the *LOA to audio generation* stage (modelled by the latent diffusion model). The probabilistic switcher controls the probability of the latent diffusion model using the ground truth AudioMAE (P_{gt}) and the GPT-2 generated AudioMAE feature (P_{pred}) as the condition. Both the AudioMAE and latent diffusion models are self-supervised pre-trained with audio data.

Incorporating Diffusion into Audio Generation

AudioLDM 2: Learning Holistic Audio Generation with Self-supervised Pretraining

核心思想：Regeneration Learning

问题定义　$H: C \rightarrow x$

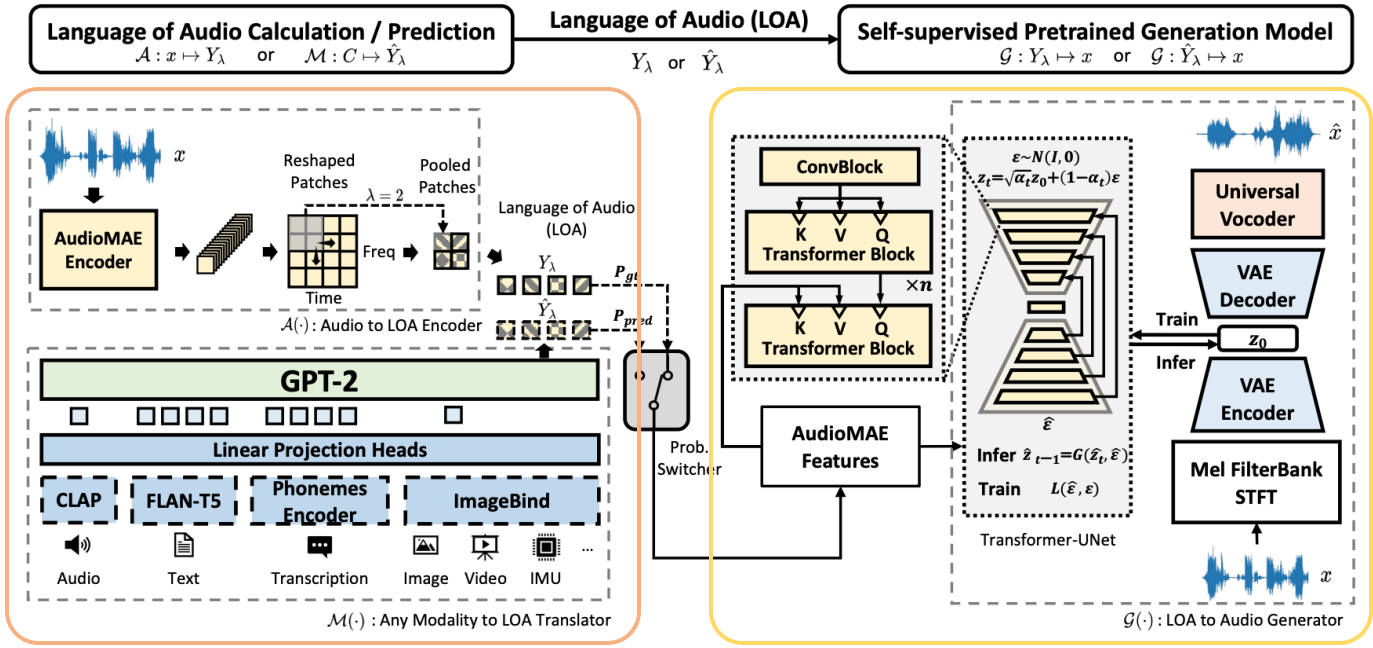
- C：各种模态的 condition 输入
- x：目标音频

问题转换　$H = G \cdot M: C \rightarrow Y' \rightarrow x$

- $M: C \rightarrow Y'$　基于 condition 生成中间表征 （LLM）
- $G: Y' \rightarrow x'$　将中间特征还原为音频波形（Diffusion）

注意：$Y \rightarrow x'$ 可以使用大量无标注数据自监督训练

$$H' = A \cdot G: x \rightarrow Y \rightarrow x'$$

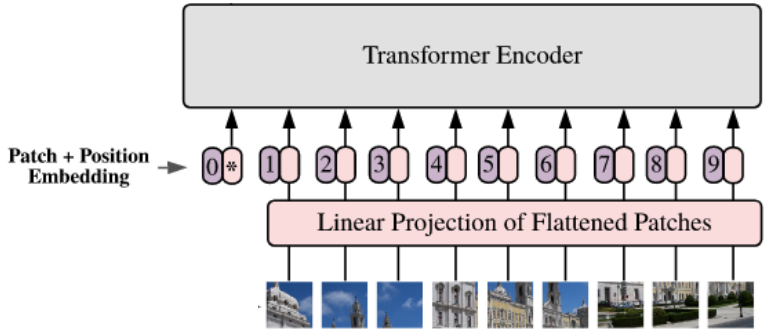


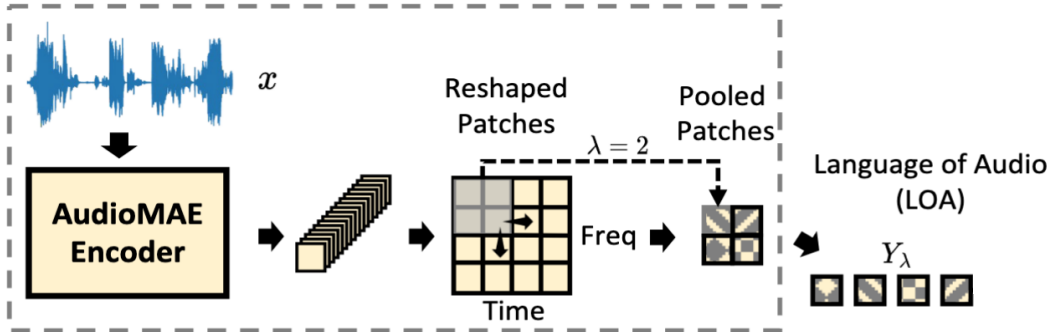
Incorporating Diffusion into Audio Generation

AudioLDM 2: Learning Holistic Audio Generation with Self-supervised Pretraining

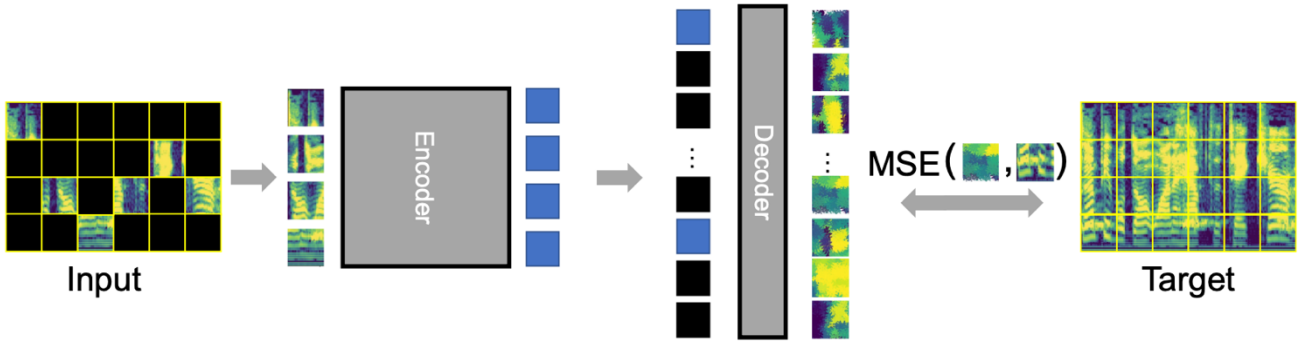
模块一： AudioMAE → Language of Audio（LOA）

- 将 Encoder 的输出作为音频的表征
- **连续空间的表征**，不需要进行聚类或 VQ 的离散化





λ 表示二维池化的 kernel_size 和 stride



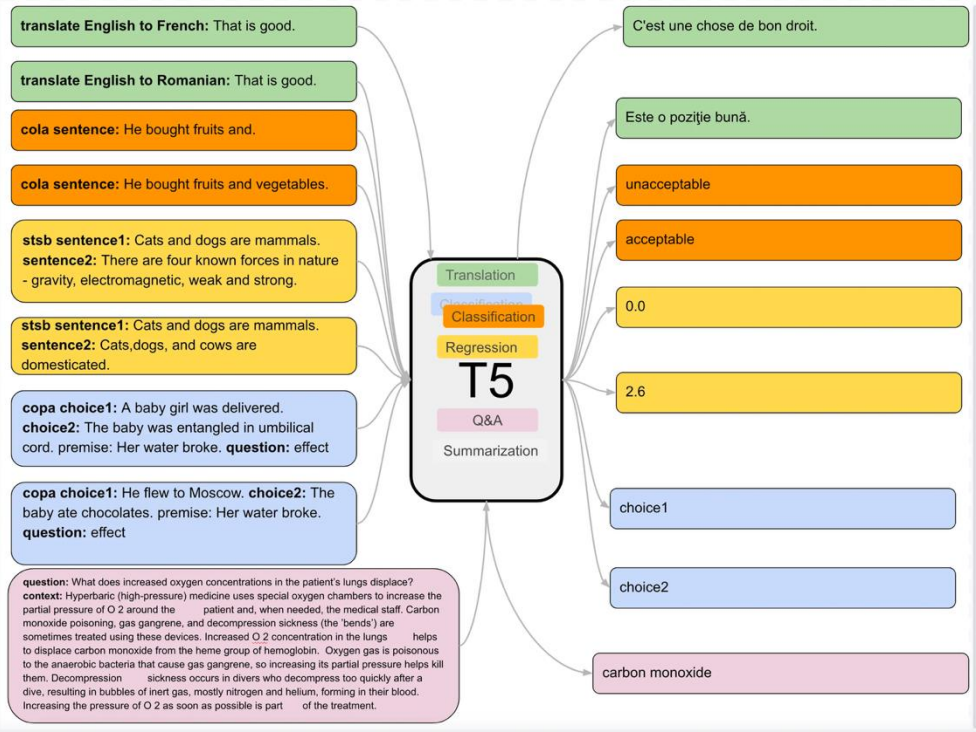
- 80% 的 patch 被 mask
- 只用被 mask 的 patch 计算 MSE

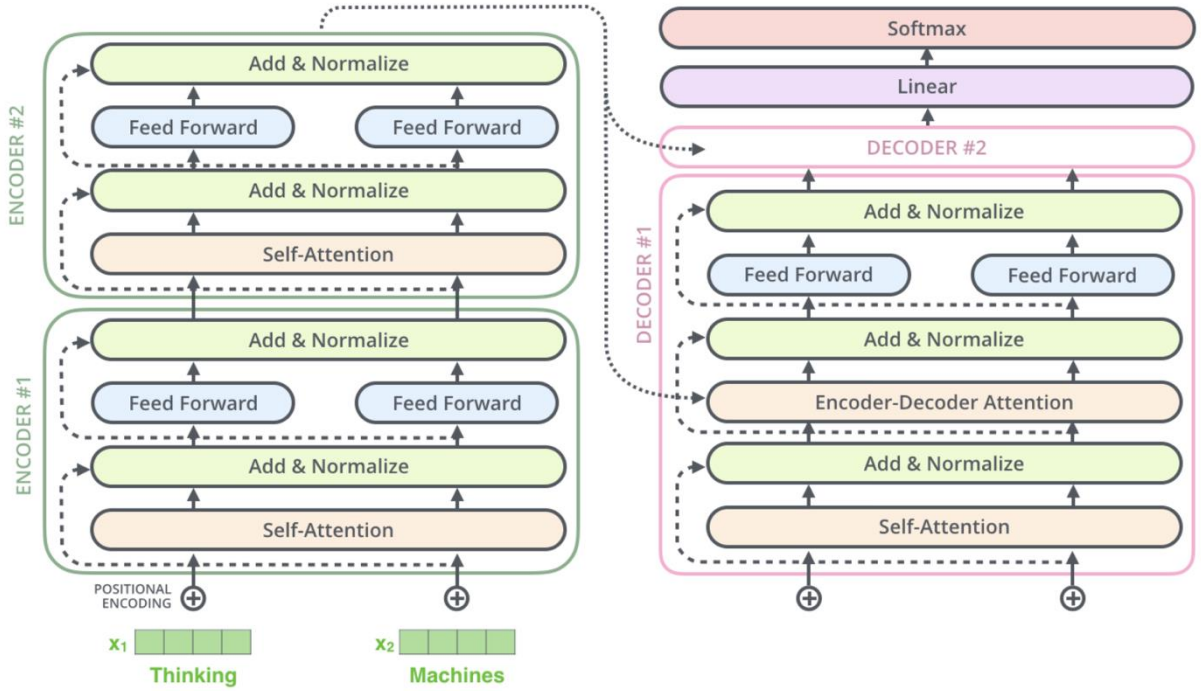
Incorporating Diffusion into Audio Generation

AudioLDM 2: Learning Holistic Audio Generation with Self-supervised Pretraining

模块二：各输入模态的 Encoders

1. FLAN-T5 Encoder（Text 输入）



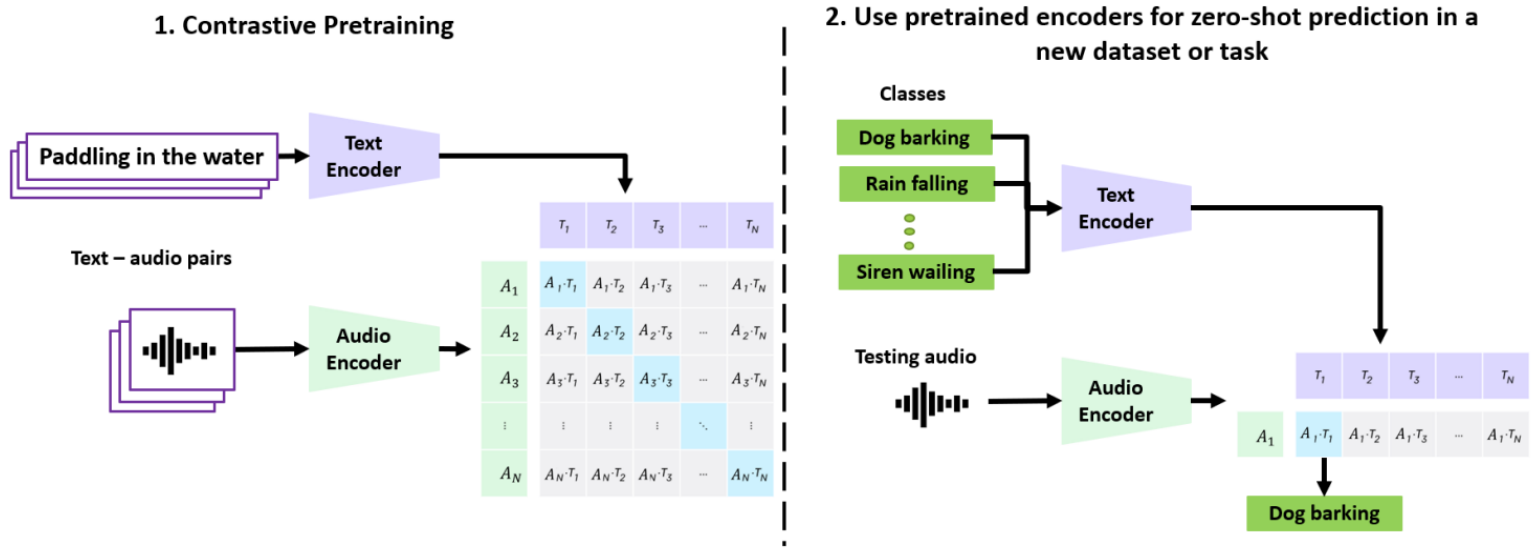


Incorporating Diffusion into Audio Generation

AudioLDM 2: Learning Holistic Audio Generation with Self-supervised Pretraining

模块二：各输入模态的 Encoders

2. CLAP（Audio 输入）



- 目的：Dual Encoder 方案，将文本和音频两个模态映射到相同的语义空间
- 应用：便于计算两种模态数据的相似度，用于分类、排序、打分等任务

(1) **BERT**

- [CLS] 的embedding 作为全局表征

(2) **PANNS**

- 基于 Mel 特征序列的多层二维卷积
- Global pooling 的输出作为音频表征

(3) 各 Encoder 增加全连接层映射到需的维度 (1024 维)

(4) 两个 Encoder 解冻参数，基于两个方向上的**对称交叉熵损失函数**进行对比预训练

$$L_s = \frac{1}{2D}\sum_{i=1}^{D}(l_1 + l_2)$$

$$l_1 = \log \frac{\exp(E_i^x \cdot E_i^y/\tau)}{\sum_{j=1}^{D}\exp(E_i^x \cdot E_j^y/\tau)}$$

$$l_2 = \log \frac{\exp(E_i^y \cdot E_i^x/\tau)}{\sum_{j=1}^{D}\exp(E_i^y \cdot E_j^x/\tau)}$$

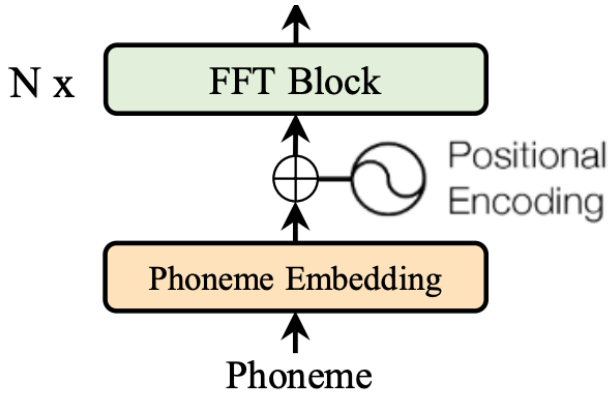
Incorporating Diffusion into Audio Generation

AudioLDM 2: Learning Holistic Audio Generation with Self-supervised Pretraining

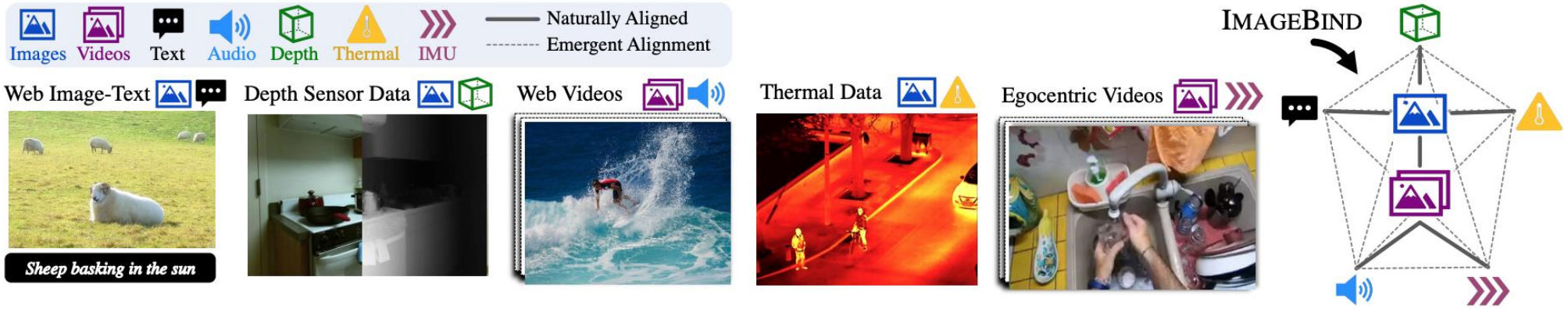
模块二：各输入模态的 Encoders

3. Phoneme Encoder（Phoneme 输入）

- g2p 方案: espeak
- 多层 Transformer Encoder 的堆叠



4. ImageBind（Image 模态输入）



Incorporating Diffusion into Audio Generation

AudioLDM 2: Learning Holistic Audio Generation with Self-supervised Pretraining

模块二：各输入模态的 Encoders

4. ImageBind

https://audioldm.github.io/audioldm2

目标：将更多的模态（7个）映射到相同的表征空间

作用：

- 接入到下游任务时，训练阶段可以只用其中一种模态
- 推理时，使用任意其他模态通过相应的 Encoder 之后作为模型输入，达到相似的效果

ImageBind 训练

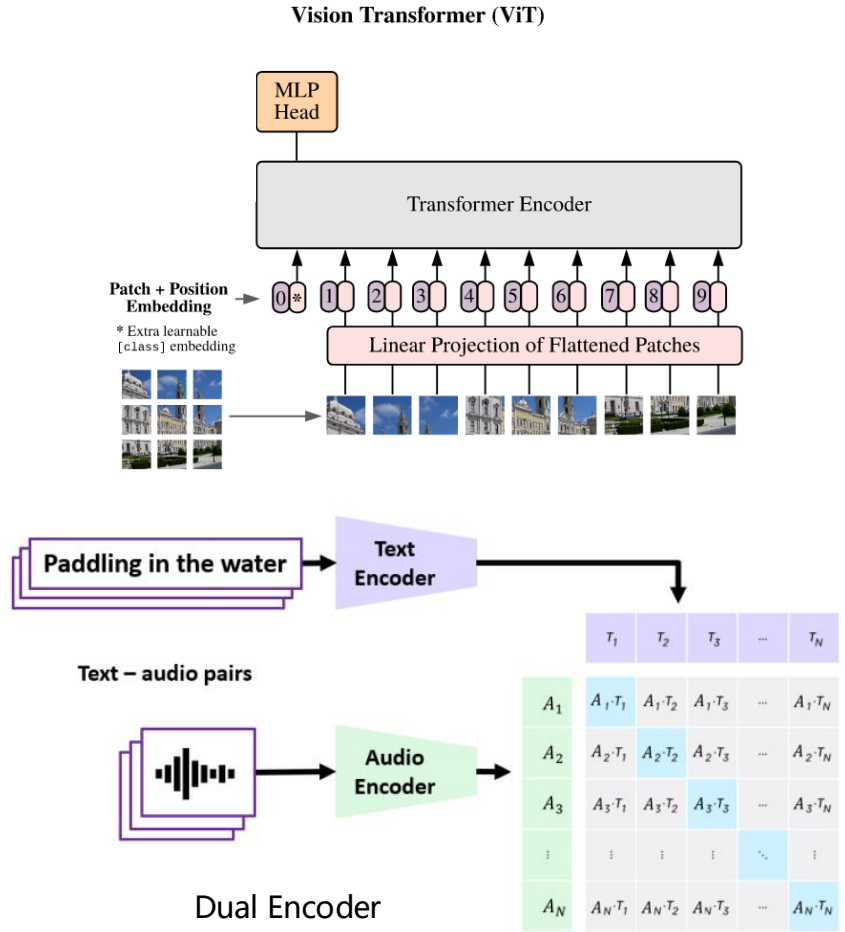
- image 核心模态，使用 <image, audio> / <image, video> … 等 pair 数据训练 CLIP
  - Text Encoder : BERT（CLIP 初始化）
  - Image/Video Encoder: ViT（CLIP 初始化）
  - Audio Encoder: AST (Audio Spectrogram Transformer)

泛化效果：<audio, text> 的 zero-shot 分类任务同样达到了 SOTA 结果

AudioLDM2 应用于 image-to-audio 的任务

训练阶段不需要 <image, audio> 的 pair 数据

- 训练使用 Audio Encoder 接受音频输入
- 推理使用 Image Encoder 接收图片输入



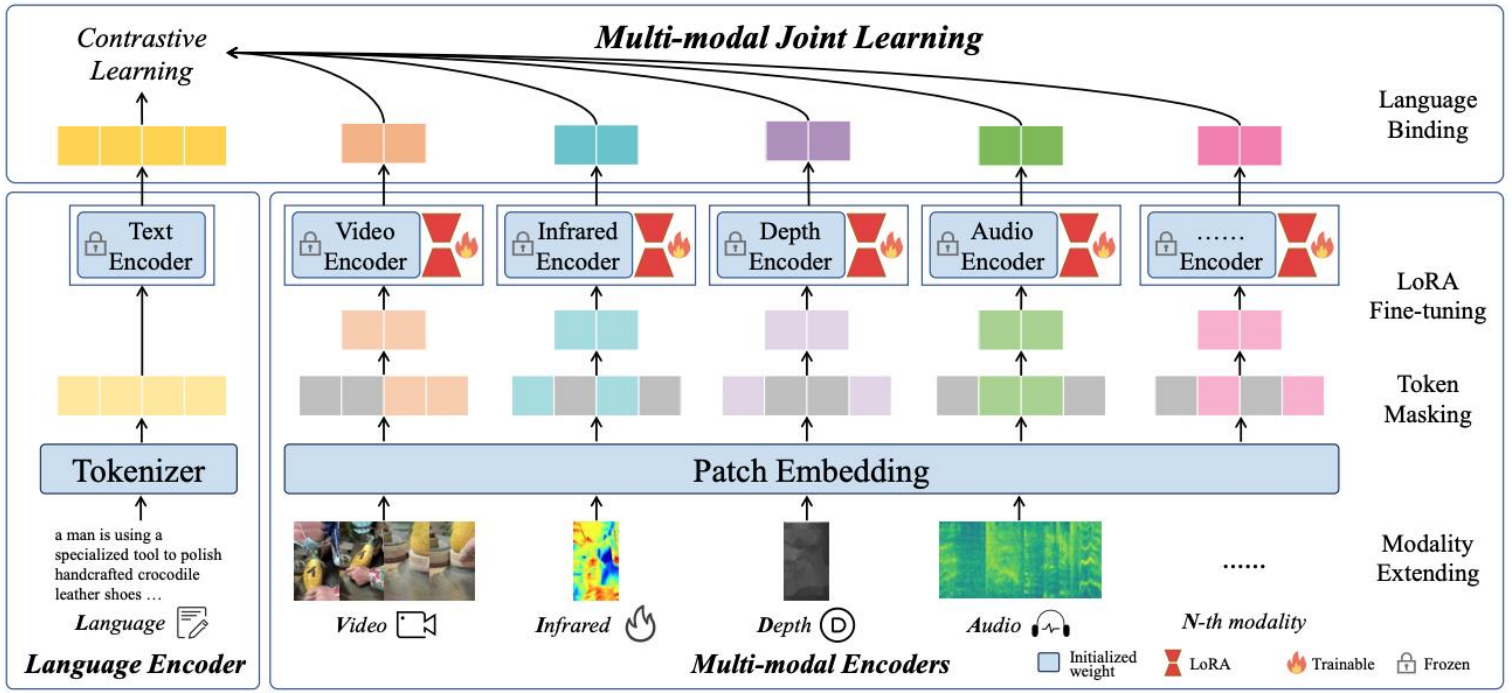
Incorporating Diffusion into Audio Generation

AudioLDM 2: Learning Holistic Audio Generation with Self-supervised Pretraining

模块二：各输入模态的 Encoders

4. ImageBind

拓展 → LanguageBind



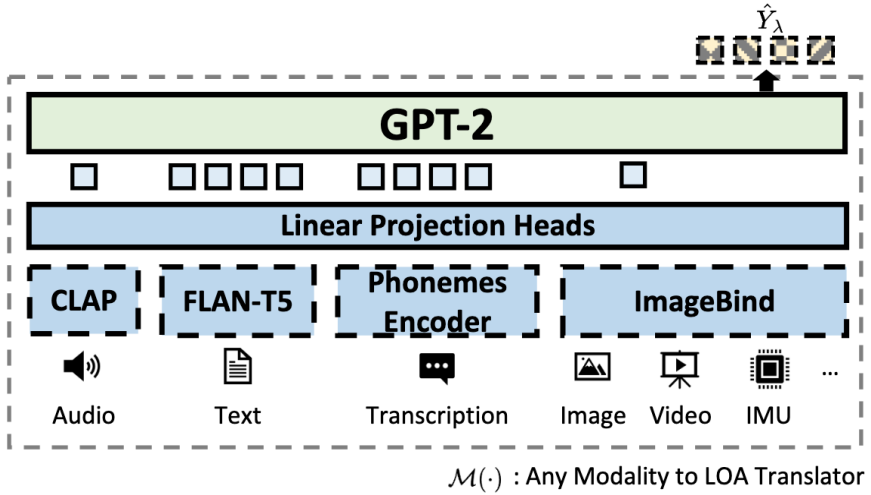
Incorporating Diffusion into Audio Generation

AudioLDM 2: Learning Holistic Audio Generation with Self-supervised Pretraining

模块三：GPT-2

第一阶段　$H = M \cdot G:\ C \to Y' \to x$

- $M: C \to Y'$　基于 condition 生成中间表征（AudioMAE）



输入：各模态 Encoder 的输出，映射到相同维度

- CLIP/ImageBind 输出的是 global embedding
- FLAN-T5/Phoneme Encoder 输出的是 embedding 序列

不同任务的数据注入方式

- text-to-audio/music
  - CLAP 的 text encoder （文本全局表征）
  - FLAN-T5 Encoder（细节语义表征）
- text-to-speech
  - CLAP 的 text encoder （文本全局表征）
  - Phoneme Encoder（具体合成文本）
- Prompt based TTS
  - 训练：CLAP 的 audio encoder
  - 推理：CLAP 的 text encoder
- image-to-audio
  - 训练：ImageBind 的 audio encoder
  - 推理：ImageBind 的 image encoder

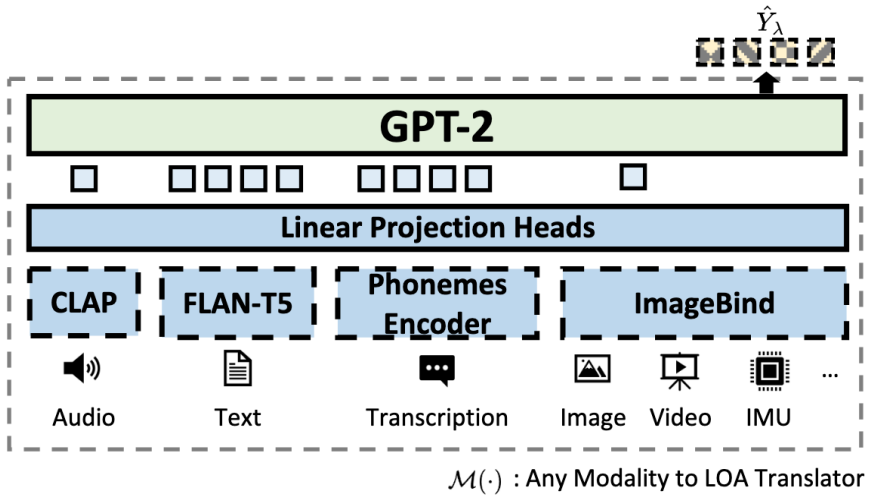
Incorporating Diffusion into Audio Generation

AudioLDM 2: Learning Holistic Audio Generation with Self-supervised Pretraining

模块三：GPT-2

第一阶段　$H = M \cdot G:\ C \rightarrow Y' \rightarrow x$

- $M: C \rightarrow Y'$　基于 condition 生成中间表征（AudioMAE）



模型训练

- 输出从原本的分类任务修改为回归任务
  - Loss：从交叉熵 CE 修改为 MSE
- 仍然采用 teacher-forcing 的训练策略

$$\text{argmax}_\theta\ \mathbb{E}_C\ [Pr(y_1|C;\theta)\prod_{l=2}^{L} Pr(y_l|y_1,...,y_{l-1},C;\theta)]$$

模型推理

- 从输入的 condition 序列自回归预测 AudioMAE 连续特征

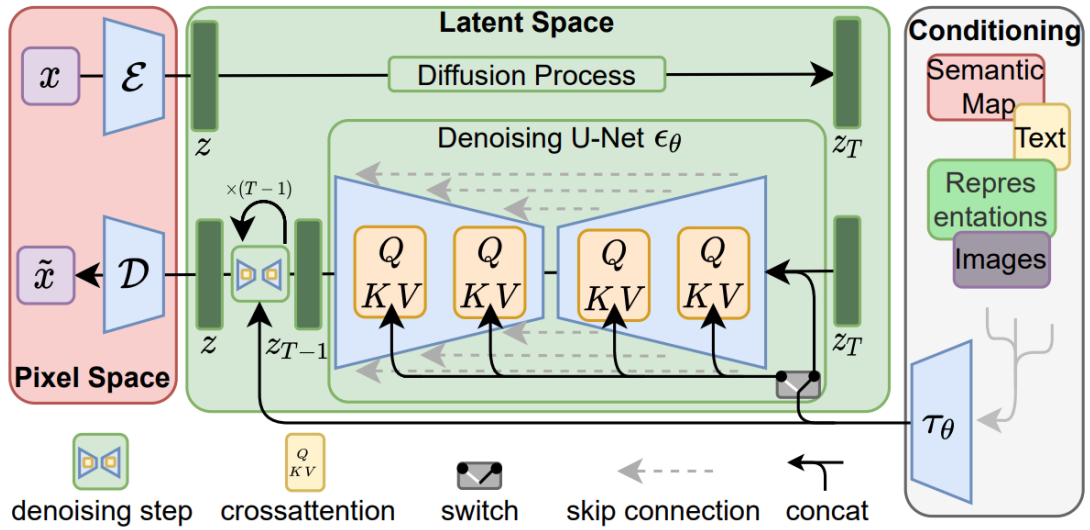
Incorporating Diffusion into Audio Generation

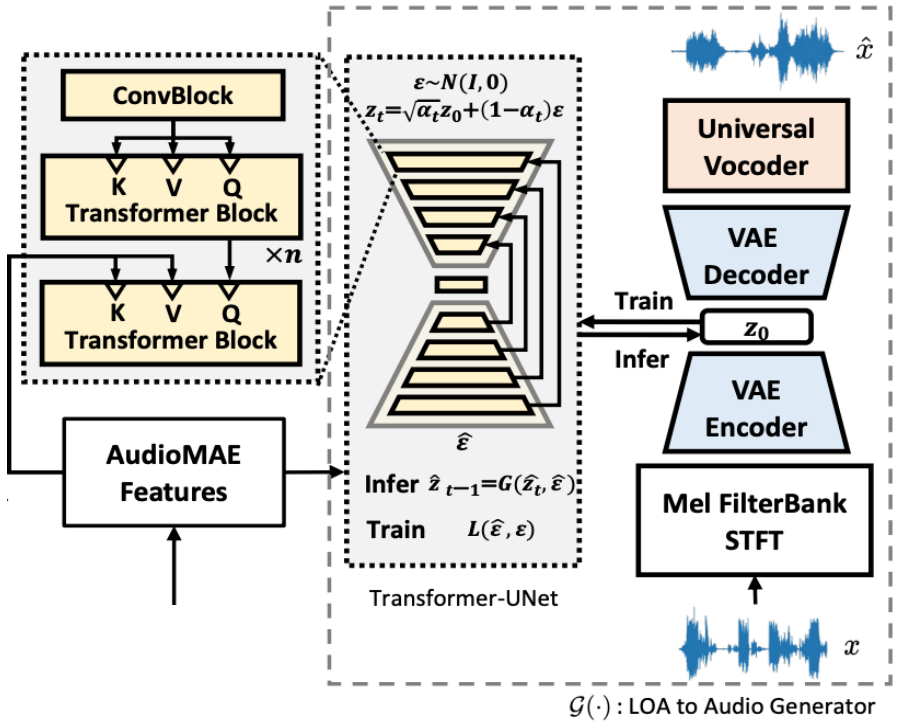
AudioLDM 2: Learning Holistic Audio Generation with Self-supervised Pretraining

模块四：LDM（Latent Diffusion Model）

第二阶段： $G: Y' \rightarrow x'$　将中间特征还原为音频波形

关键词：VAE / LDM / UNet / Cross-Attention / Classifier-Free Guidance





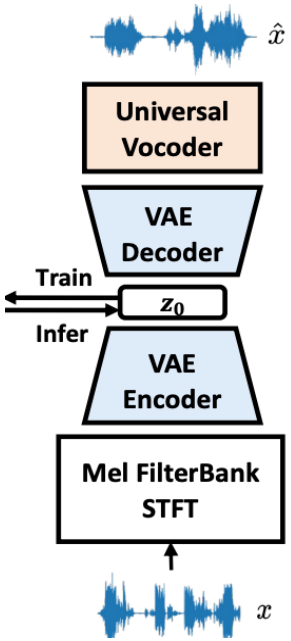
Incorporating Diffusion into Audio Generation

AudioLDM 2: Learning Holistic Audio Generation with Self-supervised Pretraining

模块四：LDM（Latent Diffusion Model）

第二阶段：$G: Y' \rightarrow x'$　将中间特征还原为音频波形

（1）感知压缩（VAE 建模 Latent 表征）



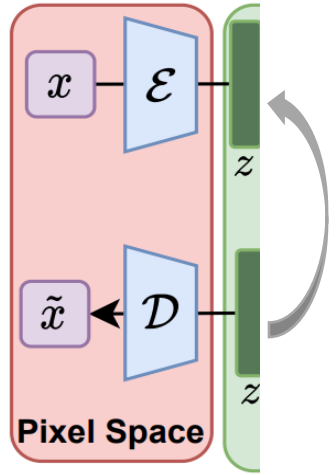
核心思路：在梅尔特征上使用 VAE 得到 Latent 表征

- Encoder 编码为维度更低的 Latent 表征
- Decoder 将 Latent 表征还原为梅尔，并通过通用声码器还原为波形
- 增加 KL 散度正则化，使得 z 的分布趋向标准正态分布
  - 为了减小对重建效果的影响，实际 KL loss 系数很小

VAE 的作用：

- 数据压缩，在维度更低的隐空间内建模更加高效

Stable Diffusion 中，除了 KL 正则化，还尝试了 VQ 正则化，使用 VQ 前一层的 embedding 作为 latent 表征（VQ 被吸收为 Decoder 的一部分）



Incorporating Diffusion into Audio Generation

AudioLDM 2: Learning Holistic Audio Generation with Self-supervised Pretraining

模块四：LDM（Latent Diffusion Model）

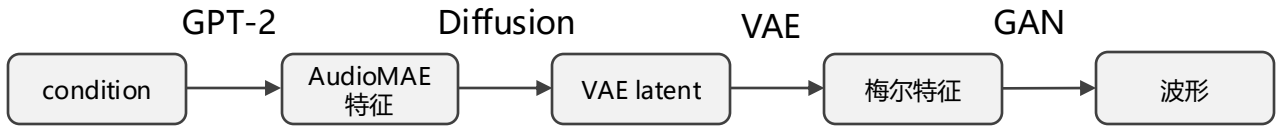
第二阶段： $G: Y' \rightarrow x'$　将中间特征还原为音频波形

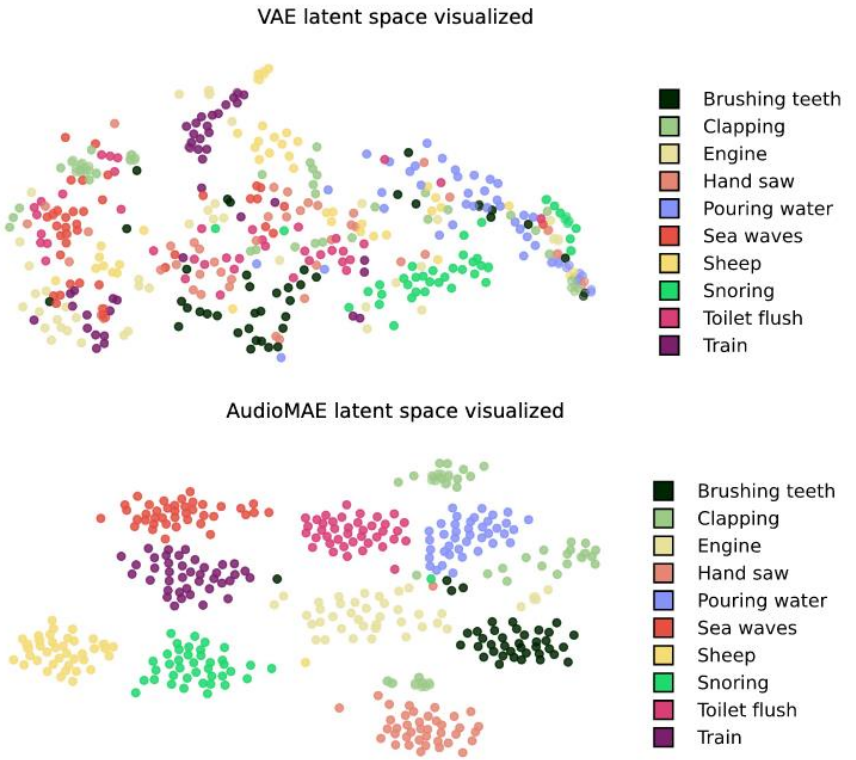
（1）感知压缩（VAE 建模 Latent 表征）

疑问：为什么要用 audioMAE 和 VAE 两种表征、不直接用 GPT-2 预测连续的 VAE latent?

- VAE 以梅尔特征恢复为主要训练目标，用于建模 acoustic 表征
- audioMAE 基于类似 MLM 的损失函数，用于建模 semantic 表征

总结：相当于更长的 Regeneration Learning 流程





Incorporating Diffusion into Audio Generation

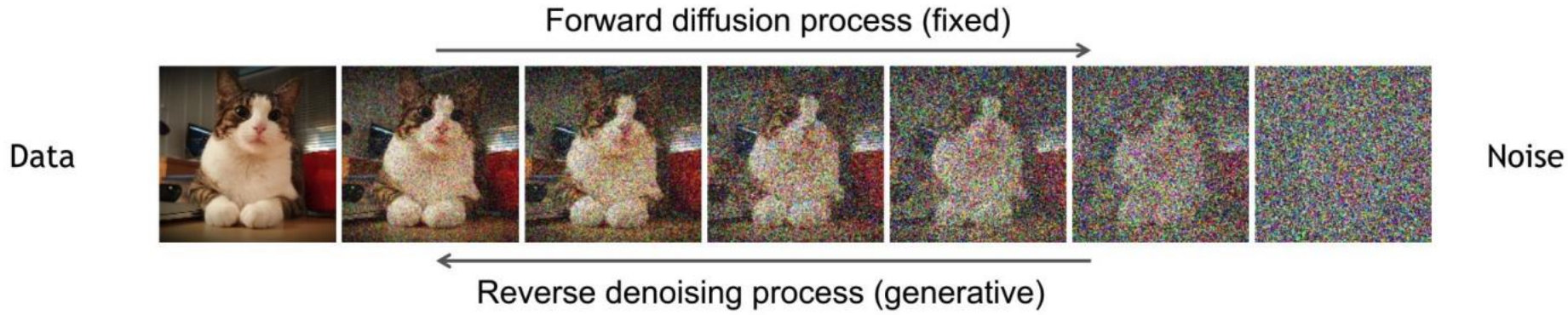
AudioLDM 2: Learning Holistic Audio Generation with Self-supervised Pretraining

模块四：LDM（Latent Diffusion Model）

第二阶段： $G: Y' \rightarrow x'$　将中间特征还原为音频波形

(2) Diffusion Model（DDPM）

- Diffusion 的前向：逐步向输入图像添加随机高斯噪声，最终希望得到高斯噪声
- Diffusion 的反向：从高斯噪声中生成图片



Incorporating Diffusion into Audio Generation

AudioLDM 2: Learning Holistic Audio Generation with Self-supervised Pretraining

模块四：LDM（Latent Diffusion Model）

DDPM 反向过程：去噪

Noise Predictor

$$\mathbb{E}_{\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\frac{\beta_t^2}{2\sigma_t^2\alpha_t(1-\bar{\alpha}_t)}\left\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0+\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon},t)\right\|^2\right]$$

$$L_{\text{simple}}(\theta):=\mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\left\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0+\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon},t)\right\|^2\right]$$

最大似然目标转换为最小化噪声的 MSE

详细推导参考《Understanding Diffusion Models》

$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ 的均值 $\boldsymbol{\mu}_\theta(\boldsymbol{x}_t,t)=\frac{1}{\sqrt{\alpha_t}}\boldsymbol{x}_t-\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)$

方差 $\sigma_q^2(t)=\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}$ $t>1$

DDPM 训练和推理（采样）

Algorithm 1 Training

1: **repeat**
2: $\mathbf{x}_0\sim q(\mathbf{x}_0)$
3: $t\sim\text{Uniform}(\{1,\dots,T\})$
4: $\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})$
5: Take gradient descent step on
$\nabla_\theta\left\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0+\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon},t)\right\|^2$
6: **until** converged

Algorithm 2 Sampling

1: $\mathbf{x}_T\sim\mathcal{N}(\mathbf{0},\mathbf{I})$
2: **for** $t=T,\dots,1$ **do**
3: $\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})$ if $t>1$, else $\mathbf{z}=\mathbf{0}$
4: $\mathbf{x}_{t-1}=\frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t-\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t,t)\right)+\sigma_t\mathbf{z}$
5: **end for**
6: **return** $\mathbf{x}_0$

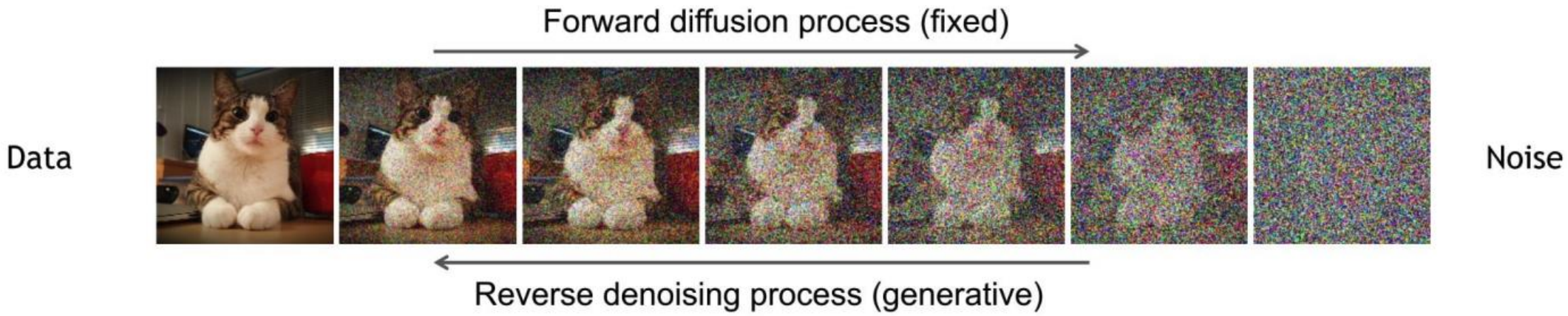
Understanding Diffusion Models

https://arxiv.org/abs/2208.11970

Understanding Diffusion Models

DDPM（Denoising Diffusion Probabilistic Models）概述

- DDPM 的前向：逐步向输入图像添加随机高斯噪声，最终希望得到高斯噪声
- DDPM 的反向：从高斯噪声中生成图片



Understanding Diffusion Models

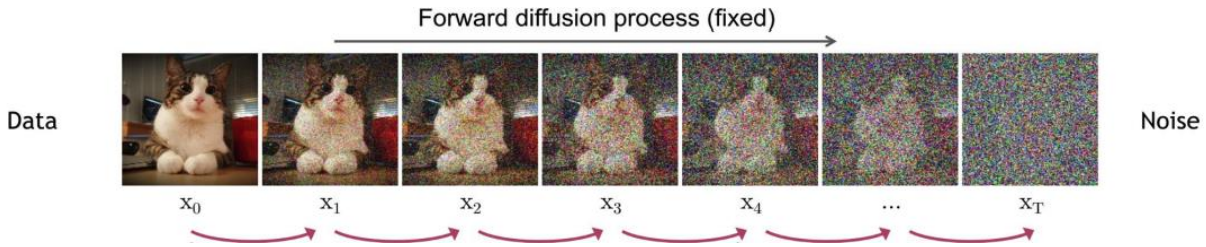
DDPM 前向过程：加噪

$$\boldsymbol{x}_t = \sqrt{\alpha_t}\boldsymbol{x}_{t-1} + \sqrt{1-\alpha_t}\boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

$$\boldsymbol{x}_t \sim q\left(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}\right) = \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{\alpha_t}\boldsymbol{x}_{t-1}, (1-\alpha_t)\mathbf{I}\right)$$

$$\boldsymbol{x}_t \sim q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\mathbf{I})), \quad \bar{\alpha}_t = \prod_{s=1}^{t}\alpha_s$$

为什么是这种形式?



$$q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) = \prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$$

出发点 → $\boldsymbol{x}_t = a_t\boldsymbol{x}_{t-1} + b_t\boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 线性加噪，高斯分布；$\boldsymbol{a_t}$ 和 $\boldsymbol{b_t}$ 是预先设定好的参数，与 t 有关

$$\boldsymbol{x}_{t-1} = a_{t-1}\boldsymbol{x}_{t-2} + b_{t-1}\boldsymbol{\varepsilon}_{t-1}$$

$$\begin{aligned}
\boldsymbol{x}_t &= a_t\boldsymbol{x}_{t-1} + b_t\boldsymbol{\varepsilon}_t \\
&= a_t(a_{t-1}\boldsymbol{x}_{t-2} + b_{t-1}\boldsymbol{\varepsilon}_{t-1}) + b_t\boldsymbol{\varepsilon}_t \\
&= a_t a_{t-1}\boldsymbol{x}_{t-2} + a_t b_{t-1}\boldsymbol{\varepsilon}_{t-1} + b_t\boldsymbol{\varepsilon}_t \\
&= \dots \\
&= (a_t \dots a_1)\boldsymbol{x}_0 + (a_t \dots a_2)b_1\boldsymbol{\varepsilon}_1 + (a_t \dots a_3)b_2\boldsymbol{\varepsilon}_2 + \cdots + a_t b_{t-1}\boldsymbol{\varepsilon}_{t-1} + b_t\boldsymbol{\varepsilon}_t
\end{aligned}$$

$(a_t \dots a_2)^2 b_1^2 + (a_t \dots a_3)^2 b_2^2 + \cdots + a_t^2 b_{t-1}^2 + b_t^2$ 噪声分布合并后的方差

$\boldsymbol{x}_t = (a_t \dots a_1)\boldsymbol{x}_0 + \sqrt{(a_t \dots a_2)^2 b_1^2 + (a_t \dots a_3)^2 b_2^2 + \cdots + a_t^2 b_{t-1}^2 + b_t^2}\,\bar{\boldsymbol{\varepsilon}}_t \quad \bar{\boldsymbol{\varepsilon}}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 转换后的递推式

Understanding Diffusion Models

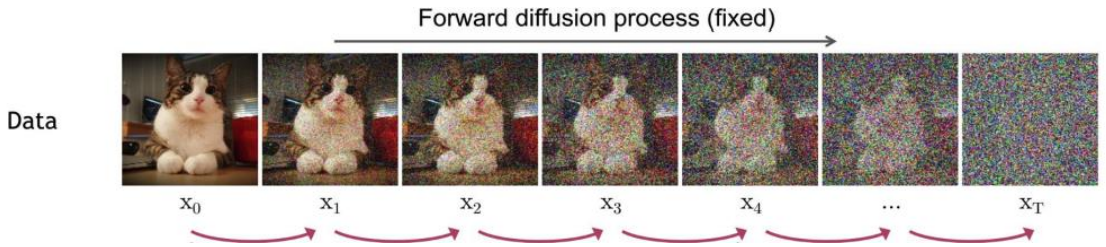
DDPM 前向过程：加噪

$$\boldsymbol{x}_t = \sqrt{\alpha_t}\boldsymbol{x}_{t-1} + \sqrt{1-\alpha_t}\boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

$$\boldsymbol{x}_t \sim q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{\alpha_t}\boldsymbol{x}_{t-1}, (1-\alpha_t)\mathbf{I}\right)$$

$$\boldsymbol{x}_t \sim q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\mathbf{I})), \quad \bar{\alpha}_t = \prod_{s=1}^{t}\alpha_s$$

为什么是这种形式?



$$q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) = \prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$$

$$\boldsymbol{x}_t = (a_t \dots a_1)\boldsymbol{x}_0 + \sqrt{(a_t \dots a_2)^2 b_1^2 + (a_t \dots a_3)^2 b_2^2 + \dots + a_t^2 b_{t-1}^2 + b_t^2}\,\bar{\boldsymbol{\varepsilon}}_t$$

$$(a_t \dots a_2)^2 b_1^2 + (a_t \dots a_3)^2 b_2^2 + \dots + a_t^2 b_{t-1}^2 + b_t^2$$

噪声分布合并后的方差

$$\begin{aligned}
&(a_t \dots a_1)^2 + (a_t \dots a_2)^2 b_1^2 + (a_t \dots a_3)^2 b_2^2 + \dots + a_t^2 b_{t-1}^2 + b_t^2 \\
=&(a_t \dots a_2)^2 a_1^2 + (a_t \dots a_2)^2 b_1^2 + (a_t \dots a_3)^2 b_2^2 + \dots + a_t^2 b_{t-1}^2 + b_t^2 \\
=&(a_t \dots a_2)^2 (a_1^2 + b_1^2) + (a_t \dots a_3)^2 b_2^2 + \dots + a_t^2 b_{t-1}^2 + b_t^2 \\
=&(a_t \dots a_3)^2 \left(a_2^2(a_1^2 + b_1^2) + b_2^2\right) + \dots + a_t^2 b_{t-1}^2 + b_t^2 \\
=&a_t^2 \left(a_{t-1}^2 \left(\dots \left(a_2^2(a_1^2 + b_1^2) + b_2^2\right) + \dots\right) + b_{t-1}^2\right) + b_t^2
\end{aligned} = 1$$

$$a_t^2 + b_t^2 = 1$$

$$(a_t \dots a_2)^2 b_1^2 + (a_t \dots a_3)^2 b_2^2 + \dots + a_t^2 b_{t-1}^2 + b_t^2 = \mathbf{1} - (a_t \dots a_1)^2$$

$$\boldsymbol{x}_t = (a_t \dots a_1)\boldsymbol{x}_0 + \sqrt{(a_t \dots a_2)^2 b_1^2 + (a_t \dots a_3)^2 b_2^2 + \dots + a_t^2 b_{t-1}^2 + b_t^2}\,\bar{\boldsymbol{\varepsilon}}_t \qquad \bar{a}_t = (a_t \dots a_1)^2$$

$$\boldsymbol{x}_t = \sqrt{\bar{a}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{a}_t}\bar{\boldsymbol{\varepsilon}}_t, \quad \bar{\boldsymbol{\varepsilon}}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

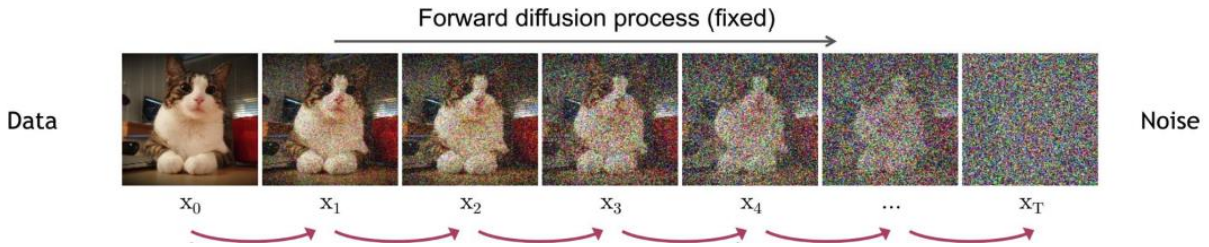
Understanding Diffusion Models

DDPM 前向过程：加噪

$$\boldsymbol{x}_t = \sqrt{\alpha_t}\boldsymbol{x}_{t-1} + \sqrt{1-\alpha_t}\boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

$$\boldsymbol{x}_t \sim q\left(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}\right) = \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{\alpha_t}\boldsymbol{x}_{t-1}, (1-\alpha_t)\mathbf{I}\right)$$

$$\boldsymbol{x}_t \sim q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\mathbf{I})), \quad \bar{\alpha}_t = \prod_{s=1}^{t}\alpha_s$$



$$q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) = \prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$$

说明一

- $0 < \alpha_t < 1$
- $t \to \infty$ 时，$\overline{\alpha_t} \to 0$，$1 - \overline{\alpha_t} \to 1$，$x_t \to N(x_t; 0, \mathbf{I})$ 符合预期

说明二

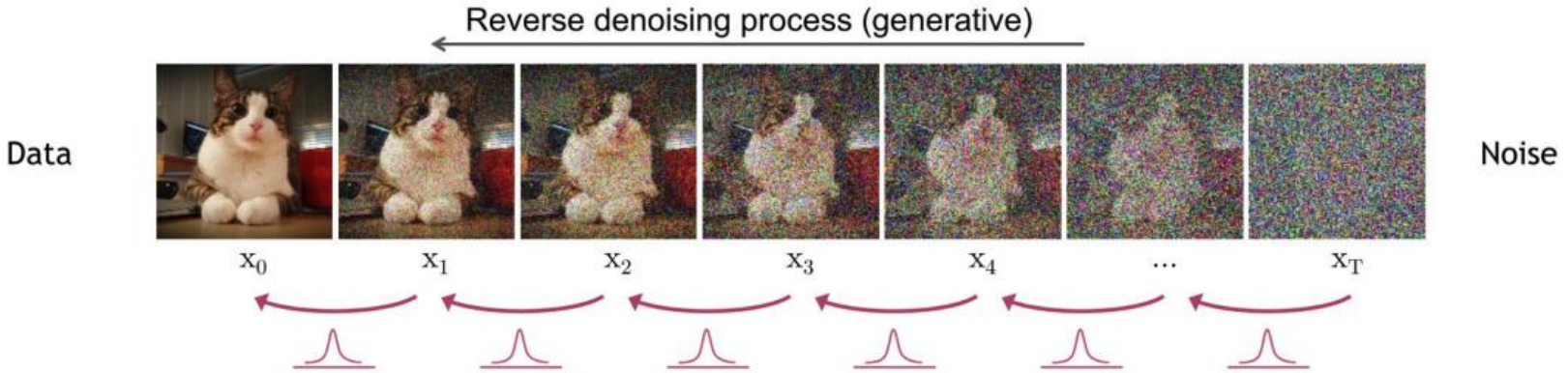
- 前向扩散过程不需要学习，只预设好各步的 α_t 参数，并在每一步采样噪声
- 前向扩散过程中任意时刻分布$q(x_t)$ 可以直接由原始 x_0和噪声的超参数 α_t~α_1 直接计算，不需要逐步迭代

说明三

- 前向过程中，通常加噪幅度随着时间 t 逐步变大（初期加噪小，后期加噪大）：$\alpha_1 > \alpha_2 > \ldots > \alpha_T$
- noise schedule: 线性 / 二次 / …（比如：$\alpha_t = 1 - 0.02t/T$）

Understanding Diffusion Models

DDPM 反向过程：去噪



$$\boldsymbol{x}_t = \sqrt{\alpha_t}\boldsymbol{x}_{t-1} + \sqrt{1-\alpha_t}\boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

$$\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\bar{\boldsymbol{\varepsilon}}_t, \quad \bar{\boldsymbol{\varepsilon}}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

对应的 ε_t 未知，不能直接求解

- 目标：通过高斯噪声 x_T, 生成 x_0
- 数学形式：已知 $p(\boldsymbol{x}_T) = \mathcal{N}(\boldsymbol{x}_T; \mathbf{0}, \mathbf{I})$ ，还原出 x_0
- 拆解问题：已知 x_t , 希望还原 x_{t-1}
  - 实际可以理解为： x_0 表示真实样本， $x_{1:T}$ 表示 latent 变量
  - 学习目标：建模 $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$

Understanding Diffusion Models

DDPM 优化目标

- 目标：通过高斯噪声 x_T，生成 x_0
- 数学形式：已知 $p(\boldsymbol{x}_T) = \mathcal{N}(\boldsymbol{x}_T; \mathbf{0}, \mathbf{I})$ ，还原出 x_0
- 拆解问题：已知 x_t，希望还原 x_{t-1}
  - 未知条件： $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$

生成模型优化目标：最大化似然（针对观测样本数据）

- 右侧推导出 ELBO 的形式（Latent Variable Model）
- 其中的第三项为

$$-\sum_{t=1}^{T-1} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_{t-1},\boldsymbol{x}_{t+1}|\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}(q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) \parallel p_\theta(\boldsymbol{x}_t|\boldsymbol{x}_{t+1}))\right]}_{\text{consistency term}}$$

问题：两个分布对应的时刻不一致，同时出现两个额外随机变量： x_{t-1}和 x_{t+1}，这种优化方案一般是 sub-optimal 的

$$\begin{aligned}
\log p(\boldsymbol{x}) &= \log \int p(\boldsymbol{x}_{0:T}) d\boldsymbol{x}_{1:T} && \text{边缘概率分布} \\
&= \log \int \frac{p(\boldsymbol{x}_{0:T}) q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} d\boldsymbol{x}_{1:T} \\
&= \log \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\right] && \text{期望的定义} \\
&\geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\right] && \text{log 上凸函数的 Jensen 不等式} \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)\prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{\prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})}\right] && p(\boldsymbol{x}_{0:T}) = p(\boldsymbol{x}_T)\prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T) p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\prod_{t=2}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_T|\boldsymbol{x}_{T-1})\prod_{t=1}^{T-1} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})}\right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T) p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\prod_{t=1}^{T-1} p_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_T|\boldsymbol{x}_{T-1})\prod_{t=1}^{T-1} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})}\right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T) p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)}{q(\boldsymbol{x}_T|\boldsymbol{x}_{T-1})}\right] + \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \prod_{t=1}^{T-1} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})}\right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\right] + \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T|\boldsymbol{x}_{T-1})}\right] + \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\sum_{t=1}^{T-1} \log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})}\right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\right] + \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T|\boldsymbol{x}_{T-1})}\right] + \sum_{t=1}^{T-1} \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})}\right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\right] + \mathbb{E}_{q(\boldsymbol{x}_{T-1},\boldsymbol{x}_T|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T|\boldsymbol{x}_{T-1})}\right] + \sum_{t=1}^{T-1} \mathbb{E}_{q(\boldsymbol{x}_{t-1},\boldsymbol{x}_t,\boldsymbol{x}_{t+1}|\boldsymbol{x}_0)}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})}\right] \\
&= \underbrace{\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\right]}_{\text{reconstruction term}} - \underbrace{\mathbb{E}_{q(\boldsymbol{x}_{T-1}|\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}(q(\boldsymbol{x}_T|\boldsymbol{x}_{T-1}) \parallel p(\boldsymbol{x}_T))\right]}_{\text{prior matching term}} \\
&\quad - \sum_{t=1}^{T-1} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_{t-1},\boldsymbol{x}_{t+1}|\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}(q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) \parallel p_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{x}_{t+1}))\right]}_{\text{consistency term}}
\end{aligned}$$

Understanding Diffusion Models

DDPM 优化目标

- 目标：通过高斯噪声 x_T，生成 x_0
- 数学形式：已知 $p(\boldsymbol{x}_T) = \mathcal{N}(\boldsymbol{x}_T; \mathbf{0}, \mathbf{I})$ ，还原出 x_0
- 拆解问题：已知 x_t，希望还原 x_{t-1}
  - 未知条件： $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$

生成模型优化目标：最大化似然（针对观测样本数据）

- 根据马尔可夫性，引入额外条件 x_0

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{x}_0)$$

- 根据贝叶斯公式

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{x}_0) = \frac{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)q(\boldsymbol{x}_t|\boldsymbol{x}_0)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}$$

- 其中第三项变为最小化 KL 散度

$$-\sum_{t=2}^{T} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) \| p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))\right]}_{\text{denoising matching term}}$$

$$\begin{aligned}
\log p(\boldsymbol{x}) &= \log \int p(\boldsymbol{x}_{0:T}) d\boldsymbol{x}_{1:T} && \text{边缘概率分布} \\
&= \log \int \frac{p(\boldsymbol{x}_{0:T}) q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} d\boldsymbol{x}_{1:T} \\
&= \log \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\right] && \text{期望的定义} \\
&\geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\right] && \text{log 上凸函数的 Jensen 不等式} \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)\prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{\prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})}\right] && p(\boldsymbol{x}_{0:T}) = p(\boldsymbol{x}_T)\prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\prod_{t=2}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_1|\boldsymbol{x}_0)\prod_{t=2}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})}\right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\prod_{t=2}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_1|\boldsymbol{x}_0)\prod_{t=2}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{x}_0)}\right] && \text{引入马尔可夫性 } q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{x}_0) \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_T)p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)}{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} + \log \prod_{t=2}^{T} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{x}_0)}\right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)}{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} + \log \prod_{t=2}^{T} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{\frac{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)q(\boldsymbol{x}_t|\boldsymbol{x}_0)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}}\right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)}{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} + \log \prod_{t=2}^{T} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{\frac{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)\cancel{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}}{\cancel{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}}}\right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)}{\cancel{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}} + \log \frac{\cancel{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}}{q(\boldsymbol{x}_T|\boldsymbol{x}_0)} + \log \prod_{t=2}^{T} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)}\right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)}{q(\boldsymbol{x}_T|\boldsymbol{x}_0)} + \sum_{t=2}^{T} \log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)}\right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\right] + \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T|\boldsymbol{x}_0)}\right] + \sum_{t=2}^{T} \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)}\right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\right] + \mathbb{E}_{q(\boldsymbol{x}_T|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T|\boldsymbol{x}_0)}\right] + \sum_{t=2}^{T} \mathbb{E}_{q(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)}\right] \\
&= \underbrace{\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\right]}_{\text{reconstruction term}} - \underbrace{D_{\mathrm{KL}}(q(\boldsymbol{x}_T|\boldsymbol{x}_0) \| p(\boldsymbol{x}_T))}_{\text{prior matching term}} - \sum_{t=2}^{T} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) \| p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))\right]}_{\text{denoising matching term}}
\end{aligned}$$

特别说明：在各种噪声强度下，可以省略不计？

Understanding Diffusion Models

DDPM 优化目标

引理

$$\boldsymbol{x}_{t-1} \sim q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) \quad \text{是高斯分布}$$

已知条件

$$\boldsymbol{x}_t \sim q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\mathbf{I}))$$

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{x}_0)$$

结论

$$\boldsymbol{x}_{t-1} \sim q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$$

高斯分布

- 均值: $\underbrace{\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar{\alpha}_t}}_{\mu_q(\boldsymbol{x}_t, \boldsymbol{x}_0)}$
- 方差: $\underbrace{\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{I}}_{\boldsymbol{\Sigma}_q(t)}$

$$\begin{aligned}
q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) &= \frac{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{x}_0)q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}{q(\boldsymbol{x}_t|\boldsymbol{x}_0)} \\
&= \frac{\mathcal{N}(\boldsymbol{x}_t; \sqrt{\alpha_t}\boldsymbol{x}_{t-1}, (1-\alpha_t)\mathbf{I})\mathcal{N}(\boldsymbol{x}_{t-1}; \sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0, (1-\bar{\alpha}_{t-1})\mathbf{I})}{\mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\mathbf{I})} \\
&\propto \exp\left\{-\left[\frac{(\boldsymbol{x}_t - \sqrt{\alpha_t}\boldsymbol{x}_{t-1})^2}{2(1-\alpha_t)} + \frac{(\boldsymbol{x}_{t-1} - \sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0)^2}{2(1-\bar{\alpha}_{t-1})} - \frac{(\boldsymbol{x}_t - \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0)^2}{2(1-\bar{\alpha}_t)}\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\frac{(\boldsymbol{x}_t - \sqrt{\alpha_t}\boldsymbol{x}_{t-1})^2}{1-\alpha_t} + \frac{(\boldsymbol{x}_{t-1} - \sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0)^2}{1-\bar{\alpha}_{t-1}} - \frac{(\boldsymbol{x}_t - \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0)^2}{1-\bar{\alpha}_t}\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\frac{(-2\sqrt{\alpha_t}\boldsymbol{x}_t\boldsymbol{x}_{t-1} + \alpha_t\boldsymbol{x}_{t-1}^2)}{1-\alpha_t} + \frac{(\boldsymbol{x}_{t-1}^2 - 2\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_{t-1}\boldsymbol{x}_0)}{1-\bar{\alpha}_{t-1}} + C(\boldsymbol{x}_t, \boldsymbol{x}_0)\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[-\frac{2\sqrt{\alpha_t}\boldsymbol{x}_t\boldsymbol{x}_{t-1}}{1-\alpha_t} + \frac{\alpha_t\boldsymbol{x}_{t-1}^2}{1-\alpha_t} + \frac{\boldsymbol{x}_{t-1}^2}{1-\bar{\alpha}_{t-1}} - \frac{2\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_{t-1}\boldsymbol{x}_0}{1-\bar{\alpha}_{t-1}}\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left[(\frac{\alpha_t}{1-\alpha_t} + \frac{1}{1-\bar{\alpha}_{t-1}})\boldsymbol{x}_{t-1}^2 - 2\left(\frac{\sqrt{\alpha_t}\boldsymbol{x}_t}{1-\alpha_t} + \frac{\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0}{1-\bar{\alpha}_{t-1}}\right)\boldsymbol{x}_{t-1}\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\frac{\alpha_t(1-\bar{\alpha}_{t-1}) + 1 - \alpha_t}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\boldsymbol{x}_{t-1}^2 - 2\left(\frac{\sqrt{\alpha_t}\boldsymbol{x}_t}{1-\alpha_t} + \frac{\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0}{1-\bar{\alpha}_{t-1}}\right)\boldsymbol{x}_{t-1}\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\frac{\alpha_t - \bar{\alpha}_t + 1 - \alpha_t}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\boldsymbol{x}_{t-1}^2 - 2\left(\frac{\sqrt{\alpha_t}\boldsymbol{x}_t}{1-\alpha_t} + \frac{\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0}{1-\bar{\alpha}_{t-1}}\right)\boldsymbol{x}_{t-1}\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\frac{1-\bar{\alpha}_t}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\boldsymbol{x}_{t-1}^2 - 2\left(\frac{\sqrt{\alpha_t}\boldsymbol{x}_t}{1-\alpha_t} + \frac{\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0}{1-\bar{\alpha}_{t-1}}\right)\boldsymbol{x}_{t-1}\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left(\frac{1-\bar{\alpha}_t}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\right)\left[\boldsymbol{x}_{t-1}^2 - 2\frac{\left(\frac{\sqrt{\alpha_t}\boldsymbol{x}_t}{1-\alpha_t} + \frac{\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0}{1-\bar{\alpha}_{t-1}}\right)}{\frac{1-\bar{\alpha}_t}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}}\boldsymbol{x}_{t-1}\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left(\frac{1-\bar{\alpha}_t}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\right)\left[\boldsymbol{x}_{t-1}^2 - 2\frac{\left(\frac{\sqrt{\alpha_t}\boldsymbol{x}_t}{1-\alpha_t} + \frac{\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0}{1-\bar{\alpha}_{t-1}}\right)(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{x}_{t-1}\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left(\frac{1}{\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}}\right)\left[\boldsymbol{x}_{t-1}^2 - 2\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar{\alpha}_t}\boldsymbol{x}_{t-1}\right]\right\} \\
&\propto \mathcal{N}(\boldsymbol{x}_{t-1}; \underbrace{\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar{\alpha}_t}}_{\mu_q(\boldsymbol{x}_t, \boldsymbol{x}_0)}, \underbrace{\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{I}}_{\boldsymbol{\Sigma}_q(t)})
\end{aligned}$$

Understanding Diffusion Models

DDPM 优化目标

- 已有结论

$$\boldsymbol{x}_{t-1} \sim q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) \propto \mathcal{N}(\boldsymbol{x}_{t-1}; \underbrace{\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar{\alpha}_t}}_{\mu_q(\boldsymbol{x}_t, \boldsymbol{x}_0)}, \underbrace{\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{I}}_{\boldsymbol{\Sigma}_q(t)})$$

方差只和 α_t 有关，每个时间步方差是固定的

$$-\sum_{t=2}^{T} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) \parallel p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))\right]}_{\text{denoising matching term}}$$

- 设计 $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$
  - **假设一：** $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ 也是高斯分布
  - **假设二：** 方差与 $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$ 相同，都是 $\sigma_q^2(t) = \frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}$
    - 将均值表示为 $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$ 不能带 x_0，因为 x_0 未知；与 t 相关

Understanding Diffusion Models

DDPM 优化目标

高斯分布性质 $$D_{\mathrm{KL}}(\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu}_x,\boldsymbol{\Sigma}_x)\parallel\mathcal{N}(\boldsymbol{y};\boldsymbol{\mu}_y,\boldsymbol{\Sigma}_y))=\frac{1}{2}\left[\log\frac{|\boldsymbol{\Sigma}_y|}{|\boldsymbol{\Sigma}_x|}-d+\mathrm{tr}(\boldsymbol{\Sigma}_y^{-1}\boldsymbol{\Sigma}_x)+(\boldsymbol{\mu}_y-\boldsymbol{\mu}_x)^T\boldsymbol{\Sigma}_y^{-1}(\boldsymbol{\mu}_y-\boldsymbol{\mu}_x)\right]$$

$$-\sum_{t=2}^{T}\underbrace{\mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)\parallel p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))\right]}_{\text{denoising matching term}}$$

$$\begin{aligned}
&\arg\min_{\boldsymbol{\theta}} D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)\parallel p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))\\
&=\arg\min_{\boldsymbol{\theta}} D_{\mathrm{KL}}(\mathcal{N}(\boldsymbol{x}_{t-1};\boldsymbol{\mu}_q,\boldsymbol{\Sigma}_q(t))\parallel\mathcal{N}(\boldsymbol{x}_{t-1};\boldsymbol{\mu}_{\boldsymbol{\theta}},\boldsymbol{\Sigma}_q(t)))\\
&=\arg\min_{\boldsymbol{\theta}}\frac{1}{2}\left[\log\frac{|\boldsymbol{\Sigma}_q(t)|}{|\boldsymbol{\Sigma}_q(t)|}-d+\mathrm{tr}(\boldsymbol{\Sigma}_q(t)^{-1}\boldsymbol{\Sigma}_q(t))+(\boldsymbol{\mu}_{\boldsymbol{\theta}}-\boldsymbol{\mu}_q)^T\boldsymbol{\Sigma}_q(t)^{-1}(\boldsymbol{\mu}_{\boldsymbol{\theta}}-\boldsymbol{\mu}_q)\right]\\
&=\arg\min_{\boldsymbol{\theta}}\frac{1}{2}\left[\log 1-d+d+(\boldsymbol{\mu}_{\boldsymbol{\theta}}-\boldsymbol{\mu}_q)^T\boldsymbol{\Sigma}_q(t)^{-1}(\boldsymbol{\mu}_{\boldsymbol{\theta}}-\boldsymbol{\mu}_q)\right]\\
&=\arg\min_{\boldsymbol{\theta}}\frac{1}{2}\left[(\boldsymbol{\mu}_{\boldsymbol{\theta}}-\boldsymbol{\mu}_q)^T\boldsymbol{\Sigma}_q(t)^{-1}(\boldsymbol{\mu}_{\boldsymbol{\theta}}-\boldsymbol{\mu}_q)\right]\\
&=\arg\min_{\boldsymbol{\theta}}\frac{1}{2}\left[(\boldsymbol{\mu}_{\boldsymbol{\theta}}-\boldsymbol{\mu}_q)^T\left(\sigma_q^2(t)\mathbf{I}\right)^{-1}(\boldsymbol{\mu}_{\boldsymbol{\theta}}-\boldsymbol{\mu}_q)\right]\\
&=\arg\min_{\boldsymbol{\theta}}\frac{1}{2\sigma_q^2(t)}\left[\|\boldsymbol{\mu}_{\boldsymbol{\theta}}-\boldsymbol{\mu}_q\|_2^2\right]
\end{aligned}$$

建模方式一：最小化 KL 散度的优化目标变为最小化均值的 MSE loss

Understanding Diffusion Models

DDPM 优化目标

已推导过的结论 $$\boldsymbol{\mu}_q(\boldsymbol{x}_t, \boldsymbol{x}_0) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar{\alpha}_t}$$

假设三： $$\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)}{1-\bar{\alpha}_t}$$ 用模型根据上一步 x_t 和 t 估计 x_0

$$
\begin{aligned}
&\arg\min_{\boldsymbol{\theta}} D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) \parallel p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)) \\
&= \arg\min_{\boldsymbol{\theta}} D_{\mathrm{KL}}(\mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q(t)) \parallel \mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_q(t))) \\
&= \arg\min_{\boldsymbol{\theta}} \frac{1}{2\sigma_q^2(t)} \left[\left\| \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)}{1-\bar{\alpha}_t} - \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar{\alpha}_t} \right\|_2^2 \right] \\
&= \arg\min_{\boldsymbol{\theta}} \frac{1}{2\sigma_q^2(t)} \left[\left\| \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)}{1-\bar{\alpha}_t} - \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar{\alpha}_t} \right\|_2^2 \right] \\
&= \arg\min_{\boldsymbol{\theta}} \frac{1}{2\sigma_q^2(t)} \left[\left\| \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t} (\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) - \boldsymbol{x}_0) \right\|_2^2 \right] \\
&= \arg\min_{\boldsymbol{\theta}} \frac{1}{2\sigma_q^2(t)} \frac{\bar{\alpha}_{t-1}(1-\alpha_t)^2}{(1-\bar{\alpha}_t)^2} \left[\|\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) - \boldsymbol{x}_0\|_2^2 \right]
\end{aligned}
$$

建模方式二：最小化 KL 散度的优化目标变为 最小化 x_0 预测值与真实值之间的 MSE loss

Understanding Diffusion Models

DDPM 优化目标

已知条件并代入 $\boldsymbol{x}_0 = \frac{\boldsymbol{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_0}{\sqrt{\bar{\alpha}_t}}$

$$
\begin{aligned}
\boldsymbol{\mu}_q(\boldsymbol{x}_t, \boldsymbol{x}_0) &= \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar{\alpha}_t} \\
&= \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\frac{\boldsymbol{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_0}{\sqrt{\bar{\alpha}_t}}}{1-\bar{\alpha}_t} \\
&= \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + (1-\alpha_t)\frac{\boldsymbol{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_0}{\sqrt{\alpha_t}}}{1-\bar{\alpha}_t} \\
&= \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t}{1-\bar{\alpha}_t} + \frac{(1-\alpha_t)\boldsymbol{x}_t}{(1-\bar{\alpha}_t)\sqrt{\alpha_t}} - \frac{(1-\alpha_t)\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_0}{(1-\bar{\alpha}_t)\sqrt{\alpha_t}} \\
&= \left(\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} + \frac{1-\alpha_t}{(1-\bar{\alpha}_t)\sqrt{\alpha_t}}\right)\boldsymbol{x}_t - \frac{(1-\alpha_t)\sqrt{1-\bar{\alpha}_t}}{(1-\bar{\alpha}_t)\sqrt{\alpha_t}}\boldsymbol{\epsilon}_0 \\
&= \left(\frac{\alpha_t(1-\bar{\alpha}_{t-1})}{(1-\bar{\alpha}_t)\sqrt{\alpha_t}} + \frac{1-\alpha_t}{(1-\bar{\alpha}_t)\sqrt{\alpha_t}}\right)\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\boldsymbol{\epsilon}_0 \\
&= \frac{\alpha_t - \bar{\alpha}_t + 1 - \alpha_t}{(1-\bar{\alpha}_t)\sqrt{\alpha_t}}\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\boldsymbol{\epsilon}_0 \\
&= \frac{1-\bar{\alpha}_t}{(1-\bar{\alpha}_t)\sqrt{\alpha_t}}\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\boldsymbol{\epsilon}_0 \\
&= \frac{1}{\sqrt{\alpha_t}}\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\boldsymbol{\epsilon}_0
\end{aligned}
$$

假设四： $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) = \frac{1}{\sqrt{\alpha_t}}\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$

$$
\begin{aligned}
&\arg\min_{\boldsymbol{\theta}} D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) \parallel p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)) \\
&= \arg\min_{\boldsymbol{\theta}} D_{\mathrm{KL}}(\mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q(t)) \parallel \mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_q(t))) \\
&= \arg\min_{\boldsymbol{\theta}} \frac{1}{2\sigma_q^2(t)}\left[\left\|\frac{1}{\sqrt{\alpha_t}}\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) - \frac{1}{\sqrt{\alpha_t}}\boldsymbol{x}_t + \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\boldsymbol{\epsilon}_0\right\|_2^2\right] \\
&= \arg\min_{\boldsymbol{\theta}} \frac{1}{2\sigma_q^2(t)}\left[\left\|\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\boldsymbol{\epsilon}_0 - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)\right\|_2^2\right] \\
&= \arg\min_{\boldsymbol{\theta}} \frac{1}{2\sigma_q^2(t)}\left[\left\|\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}(\boldsymbol{\epsilon}_0 - \hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t))\right\|_2^2\right] \\
&= \arg\min_{\boldsymbol{\theta}} \frac{1}{2\sigma_q^2(t)}\frac{(1-\alpha_t)^2}{(1-\bar{\alpha}_t)\alpha_t}\left[\|\boldsymbol{\epsilon}_0 - \hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)\|_2^2\right]
\end{aligned}
$$

建模方式三：最小化 KL 散度的优化目标变为 最小化噪声预测的 MSE loss

↓

模型作为 Noise Predictor

Understanding Diffusion Models

DDPM 优化目标

最小化 KL 散度的优化目标变为 最小化噪声预测的 MSE loss

$$\mathbb{E}_{\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\frac{\beta_t^2}{2\sigma_t^2\alpha_t(1-\bar{\alpha}_t)}\left\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0+\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon},t)\right\|^2\right]$$

$$L_{\text{simple}}(\theta):=\mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\left\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0+\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon},t)\right\|^2\right]$$

$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ 的均值 $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)=\frac{1}{\sqrt{\alpha_t}}\boldsymbol{x}_t-\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)$

方差 $\sigma_q^2(t)=\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}$ t > 1

Algorithm 1 Training

1: **repeat**
2: $\mathbf{x}_0\sim q(\mathbf{x}_0)$
3: $t\sim\text{Uniform}(\{1,\dots,T\})$
4: $\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})$
5: Take gradient descent step on
$\nabla_\theta\left\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0+\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon},t)\right\|^2$
6: **until** converged

Algorithm 2 Sampling

1: $\mathbf{x}_T\sim\mathcal{N}(\mathbf{0},\mathbf{I})$
2: **for** $t=T,\dots,1$ **do**
3: $\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})$ if $t>1$, else $\mathbf{z}=\mathbf{0}$
4: $\mathbf{x}_{t-1}=\frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t-\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t,t)\right)+\sigma_t\mathbf{z}$
5: **end for**
6: **return** $\mathbf{x}_0$

Incorporating Diffusion into Audio Generation

AudioLDM 2: Learning Holistic Audio Generation with Self-supervised Pretraining

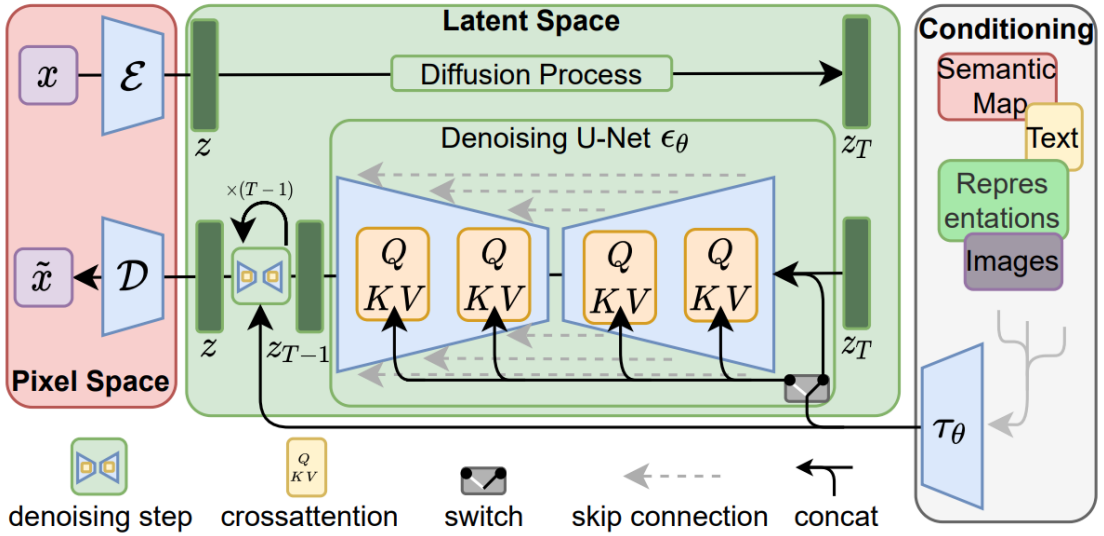
模块四：LDM（Latent Diffusion Model）

- Diffusion Model
- Latent Diffusion Model
- Conditional Latent Diffusion Model

$$L_{DM} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\right]$$

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x),\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t)\|_2^2\right]$$

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|_2^2\right]$$



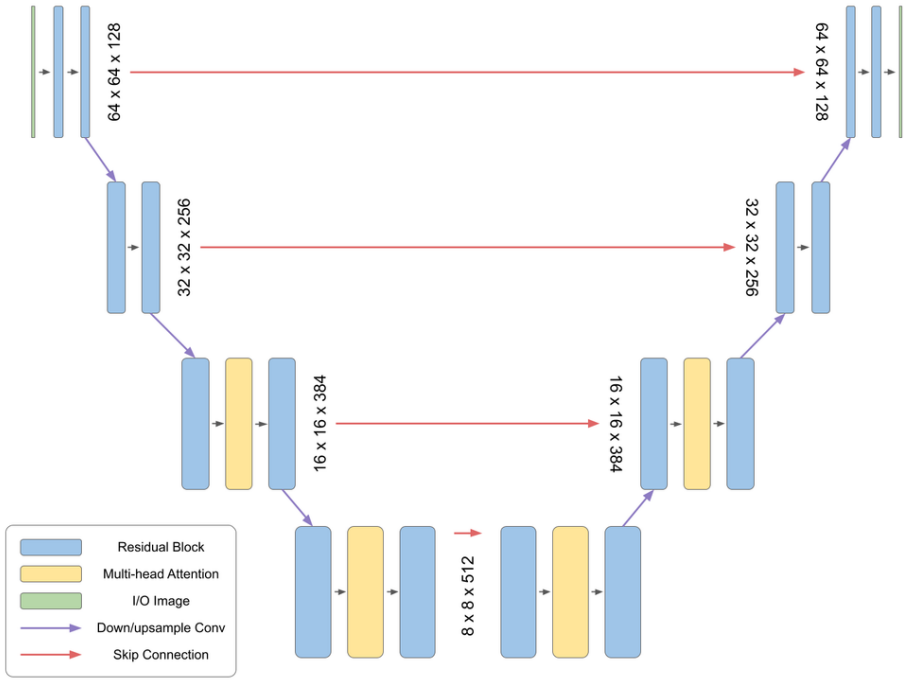
如何加入 Condition?

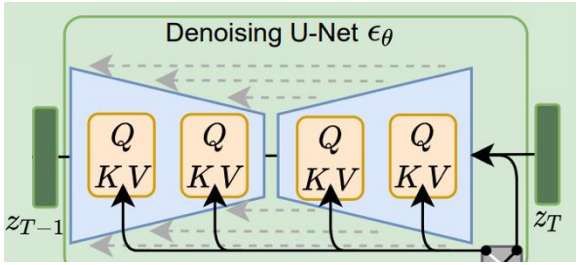
Incorporating Diffusion into Audio Generation

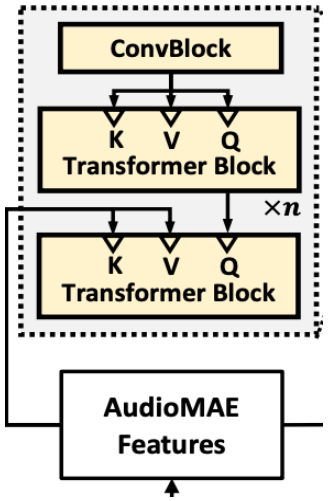
AudioLDM 2: Learning Holistic Audio Generation with Self-supervised Pretraining

模块四：LDM（Latent Diffusion Model）

Noise Predictor: U-Net + Cross-Attention + Time Condition







$$Q = W_Q^{(i)} \cdot \varphi_i(z_t),\ K = W_K^{(i)} \cdot \tau_\theta(y),\ V = W_V^{(i)} \cdot \tau_\theta(y)$$

Time-Condition: t 通过绝对位置编码加入到模型中

Incorporating Diffusion into Audio Generation

AudioLDM 2: Learning Holistic Audio Generation with Self-supervised Pretraining

模块四：LDM（Latent Diffusion Model）

CFG: Classifier Free Guidance

$$p(x \mid y) = \frac{p(y \mid x) \cdot p(x)}{p(y)}$$ y 是 condition

$$\Longrightarrow \log p(x \mid y) = \log p(y \mid x) + \log p(x) - \log p(y)$$

$$\Longrightarrow \nabla_x \log p(x \mid y) = \nabla_x \log p(y \mid x) + \nabla_x \log p(x),$$

Classifier

简单理解：增强 condition 的作用，降低生成结果的丰富度

- 训练阶段随机 10% 将 condition 置为空
- 推理阶段控制采样的引导强度：3.5

CFG $$p(y \mid x) = \frac{p(x \mid y) \cdot p(y)}{p(x)}$$

$$\Longrightarrow \log p(y \mid x) = \log p(x \mid y) + \log p(y) - \log p(x)$$

$$\Longrightarrow \nabla_x \log p(y \mid x) = \nabla_x \log p(x \mid y) - \nabla_x \log p(x).$$

$$\nabla_x \log p_\gamma(x \mid y) = \nabla_x \log p(x) + \gamma \nabla_x \log p(y \mid x)$$

$$\nabla_x \log p_\gamma(x \mid y) = (1-\gamma) \nabla_x \log p(x) + \gamma \nabla_x \log p(x \mid y)$$

$$\bar{\epsilon}_\theta\left(\mathbf{x}_t, t, y\right) = (w+1)\epsilon_\theta\left(\mathbf{x}_t, t, y\right) - w\epsilon_\theta\left(\mathbf{x}_t, t\right)$$

```python
clip_model = ...  # 加载一个官方的 clip 模型

text = "一只狗"  # 输入文本
text_embeddings = clip_model.text_encode(text)  # 编码条件文本
empty_embeddings = clip_model.text_encode("")  # 编码空文本
text_embeddings = torch.cat(empty_embeddings, text_embeddings)  # 把它俩 concate 到一起作为条件

input = get_noise(...)  # 从高斯分布随机取一个跟输出图像一样 shape 的噪声图

for t in tqdm(scheduler.timesteps):

    # 用 unet 推理，预测噪声
    with torch.no_grad():
        # 这里同时预测出了有文本的和空文本的图像噪声
        noise_pred = unet(input, t, encoder_hidden_states=text_embeddings).sample

    # Classifier-Free Guidance 引导
    noise_pred_uncond, noise_pred_text = noise_pred.chunk(2)  # 拆成无条件和有条件的噪声
    # 把【"无条件噪声"指向"有条件噪声"】看做一个向量，根据 guidance_scale 的值放大这个向量
    #  (当 guidance_scale = 1 时，下面这个式子退化成 noise_pred_text)
    noise_pred = noise_pred_text + guidance_scale * (noise_pred_text - noise_pred_uncond)

    # 用预测出的 noise_pred 和 x_t 计算得到 x_t-1
    latents = scheduler.step(noise_pred, t, latents).prev_sample
```

Incorporating Diffusion into Audio Generation

AudioLDM 2: Learning Holistic Audio Generation with Self-supervised Pretraining

模块四：LDM（Latent Diffusion Model）

CFG: Classifier Free Guidance

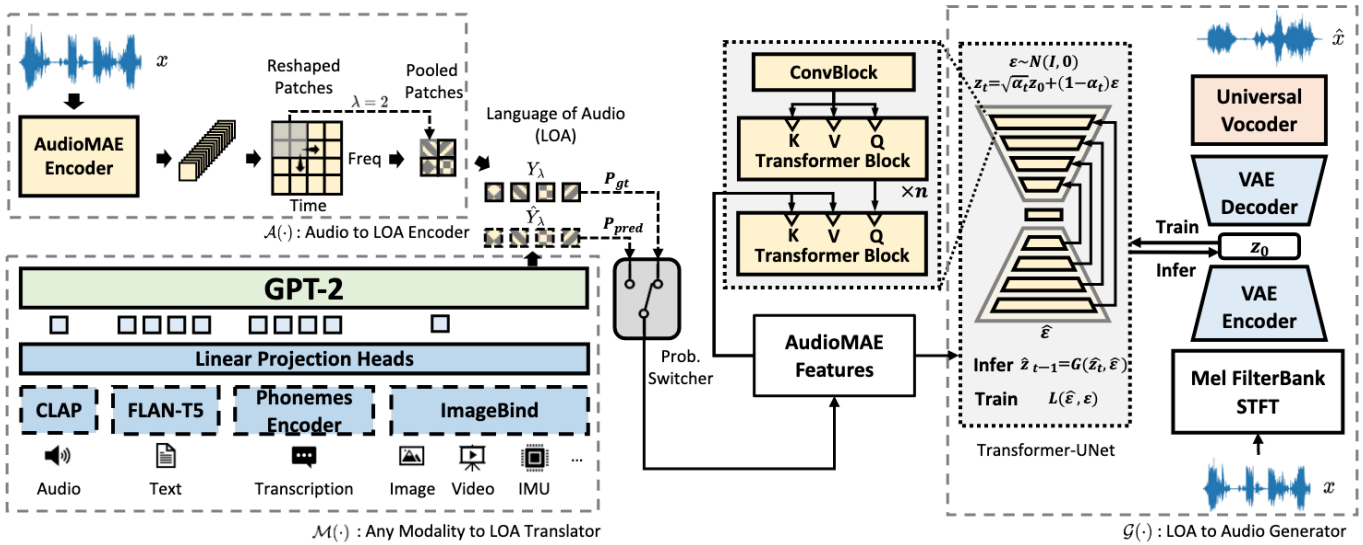
- full face epic portrait, male wizard with glowing eyes, elden ring, matte painting concept art, beautifully backlit, swirly vibrant color lines, majestic, cinematic aesthetic, smooth, intricate
- 全脸史诗肖像，发光眼睛的男巫师，埃尔登戒指，哑光绘画概念艺术，美丽的背光，漩涡般充满活力的色彩线条，雄伟，电影美学，光滑，复杂

Incorporating Diffusion into Audio Generation

AudioLDM 2: Learning Holistic Audio Generation with Self-supervised Pretraining

完整的训练策略

- 模块一：AudioMAE → 预训练好，冻结参数保持不变
- 模块二：各模态 Encoder → 预训练好，冻结参数保持不变（Phoneme Encoder 除外）
- 模块三：GPT-2 → 有标注数据训练，teacher-forcing 预测 AudioMAE 特征，使用 MSE loss
- 模块四：
  - VAE → 自监督训练，用于提取 latent 表征，冻结参数保持不变
  - LDM → 先自监督训练，使用 GT AudioMAE 训练 LDM
- 增加联合训练步骤：GPT-2 + LDM 采用 prob-switcher，GT= 0.25，pred = 0.75



Algorithm 1 Training

1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $t \sim \text{Uniform}(\{1, \dots, T\})$
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: Take gradient descent step on
$$\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) \right\|^2$$
← 加入 AudioMAE 特征
6: **until** converged

Algorithm 2 Sampling

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \dots, 1$ **do**
3: $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
4: $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z}$ ← 加入 AudioMAE 特征
5: **end for**
6: **return** $\mathbf{x}_0$

推理流程

输入 文本/图像 Condition → AudioMAE → LDM 多次采样得到 z0 → VAE 恢复梅尔特征 → Vocoder 恢复波形

Incorporating Diffusion into Audio Generation

AudioLDM 2: Learning Holistic Audio Generation with Self-supervised Pretraining

核心思想：Regeneration Learning

问题定义　$H: C \rightarrow x$

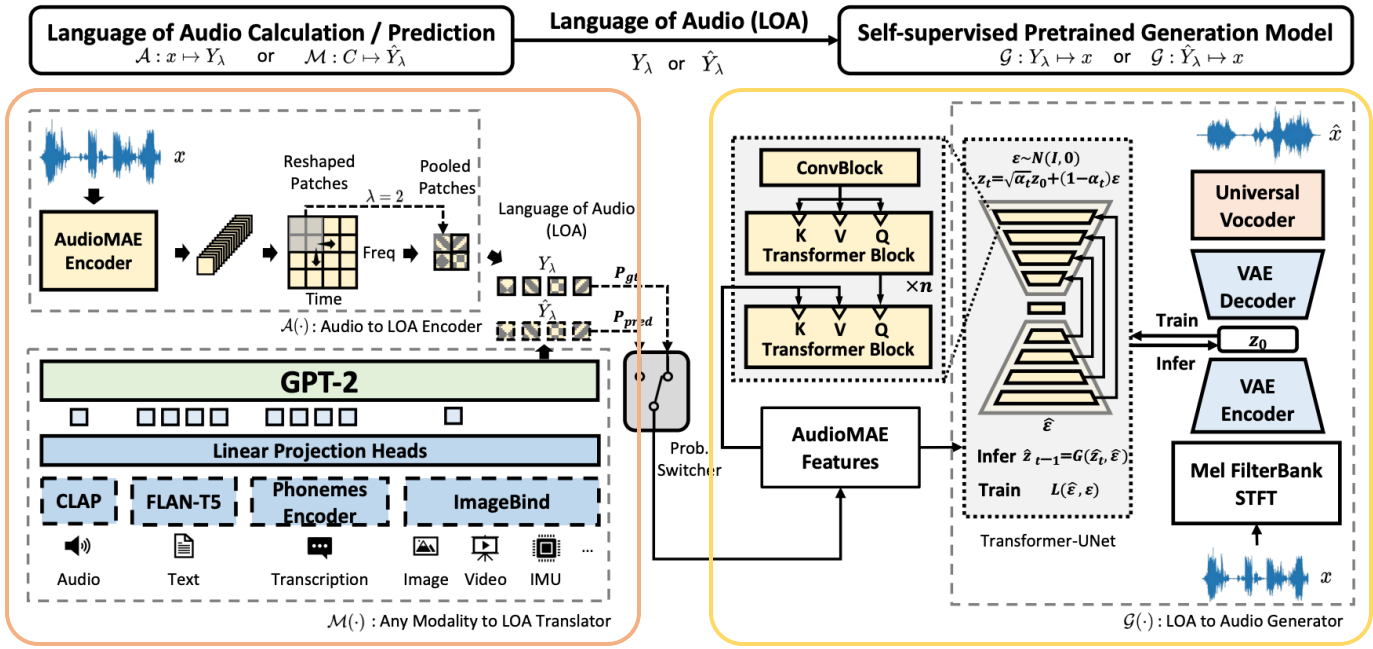
- C：各种模态的 condition 输入
- x：目标音频

问题转换　$H = G \cdot M: C \rightarrow Y' \rightarrow x$

- $M: C \rightarrow Y'$　基于 condition 生成中间表征 （LLM）
- $G: Y' \rightarrow x'$　将中间特征还原为音频波形（Diffusion）

注意：$Y \rightarrow x'$ 可以使用大量无标注数据自监督训练

$$H' = A \cdot G: x \rightarrow Y \rightarrow x'$$



Incorporating Diffusion into Audio Generation

AudioLDM 2: Learning Holistic Audio Generation with Self-supervised Pretraining

TABLE I

PERFORMANCE COMPARISON ON THE AUDIOCAPS EVALUATION SET. *AudioLDM 2* OUTPERFORMS PREVIOUS APPROACHES BY A LARGE MARGIN ON BOTH SUBJECTIVE AND OBJECTIVE EVALUATION.

Model	Duration (h)	Param	FAD↓	KL↓	CLAP (%)↑	OVL ↑	REL ↑
GroundTruth	-	-	-	-	25.1	4.04	4.08
AudioGen-Large	6824	1 B	1.82	1.69	-	-	-
Make-an-Audio	3000	453 M	2.66	1.61	-	-	-
AudioLDM-Large-FT	9031	739 M	1.96	1.59	-	-	-
AudioLDM-M	9031	416 M	4.53	1.99	14.1	3.61	3.55
Make-an-Audio 2	3700	937 M	2.05	1.27	17.3	3.68	3.62
TANGO	145	866 M	1.73	1.27	17.6	3.75	3.72
AudioLDM 2-AC	145	346 M	1.67	1.01	**24.9**	3.88	**3.90**
AudioLDM 2-AC-Large	145	712 M	**1.42**	**0.98**	24.3	**3.89**	3.87

Incorporating Diffusion into Audio Generation

AudioLDM 2: Learning Holistic Audio Generation with Self-supervised Pretraining

TABLE III
PERFORMANCE COMPARISON ON THE MUSICCAPS EVALUATION SET. THE SUPERSCRIPT † INDICATES RESULTS REPRODUCED USING PUBLICLY AVAILABLE IMPLEMENTATIONS. THE OPEN-SOURCE VERSION OF MUSICGEN-MEDIUM EXCLUDES VOCAL SOUNDS, RESULTING IN SLIGHTLY INFERIOR PERFORMANCE COMPARED TO THE ORIGINAL REPORT [33]. ALL GENERATED AUDIO CLIPS WERE RESAMPLED TO 16KHZ PRIOR TO EVALUATION.

Model	FAD↓	KL↓	CLAP (%)↑	OVL↑	REL↑
GroundTruth	-	-	25.3	3.82	4.26
Riffusion	14.80	2.06	19.0	-	-
Mousai	7.50	1.59	-	-	-
MeLoDy	5.41	-	-	-	-
MusicLM	4.00	-	-	-	-
MusicGen-Medium	3.4	1.23	32.0	-	-
MusicGen-Medium†	4.89	1.35	29.1	3.37	3.38
AudioLDM-M†	3.20	1.29	**36.0**	3.03	3.25
AudioLDM 2-MSD	4.47	1.32	29.4	**3.41**	3.30
AudioLDM 2-Full	**3.13**	**1.20**	30.1	3.34	**3.54**

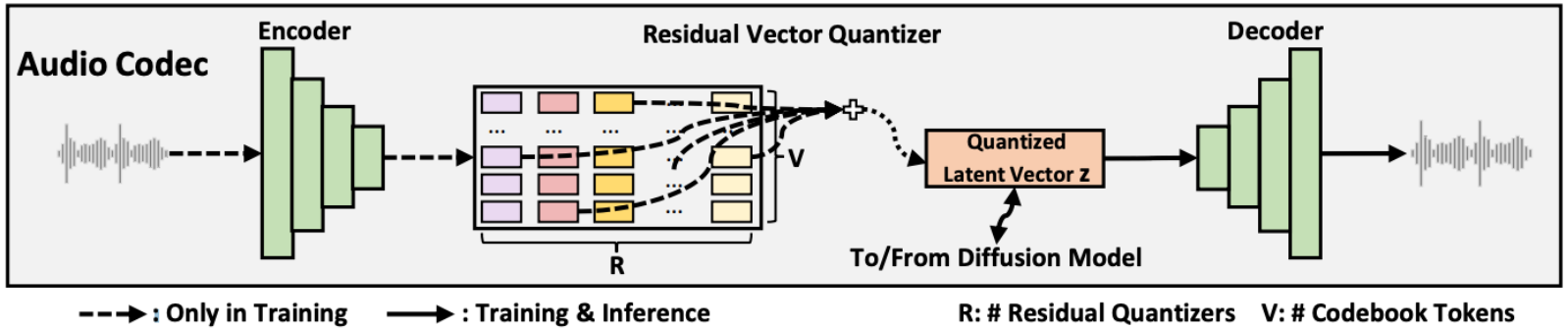
TABLE IV
TEXT-TO-SPEECH PERFORMANCE EVALUATED ON THE LJSPEECH TEST SET.

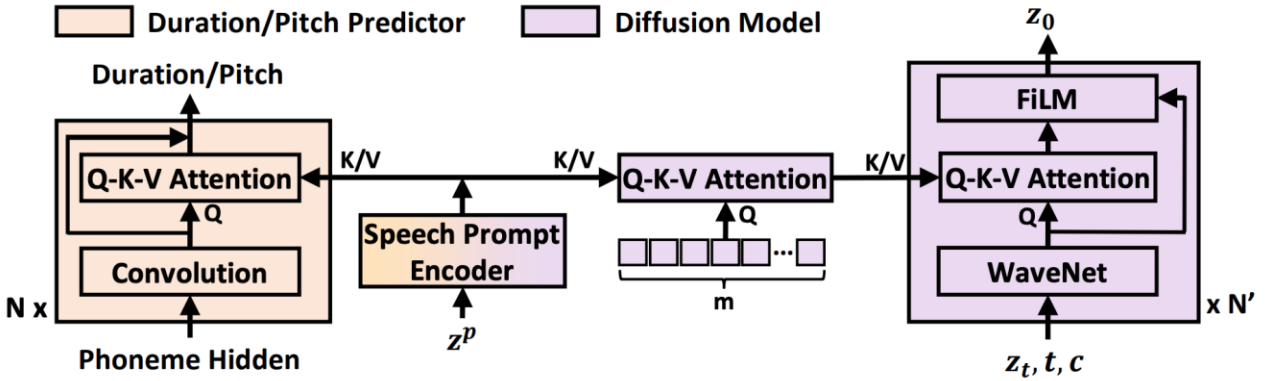
Model	Mean Opinion Score↑
GroundTruth	4.63 ± 0.08
GT-AudioMAE	4.14 ± 0.13
FastSpeech2	3.78 ± 0.15
AudioLDM 2-LJS	3.65 ± 0.21
AudioLDM 2-LJS-Pretrained	4.00 ± 0.13

Incorporating Diffusion into Audio Generation

Comparative Study

- Text-to-Speech: NaturalSpeech2

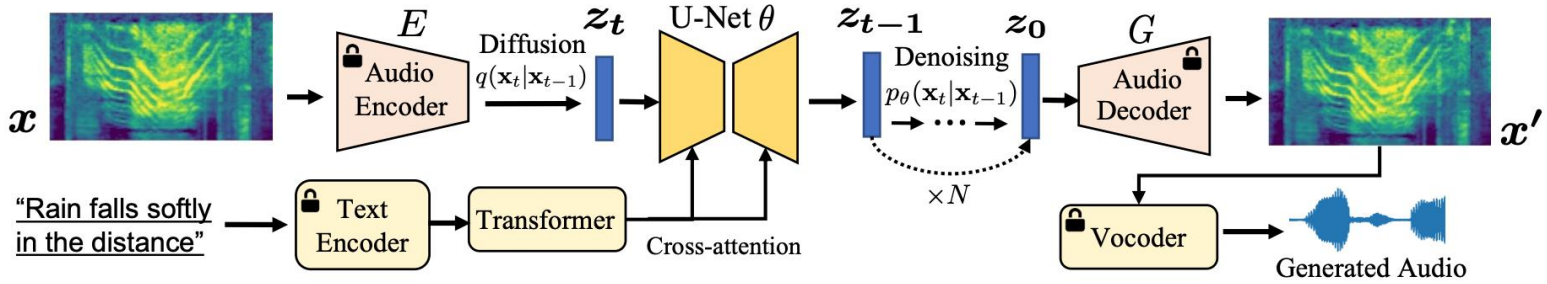




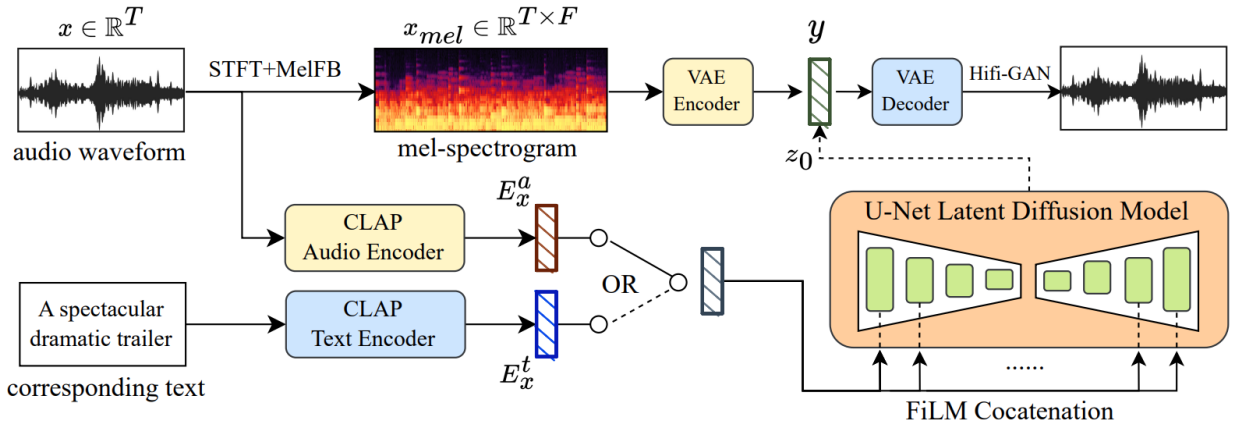
Incorporating Diffusion into Audio Generation

Comparative Study

- Text-to-Audio: Make-an-Audio 1/2



- Text-to-Music: MusicLDM



Incorporating Diffusion into Audio Generation

Comparative Study

- Text-to-Audio: TANGO

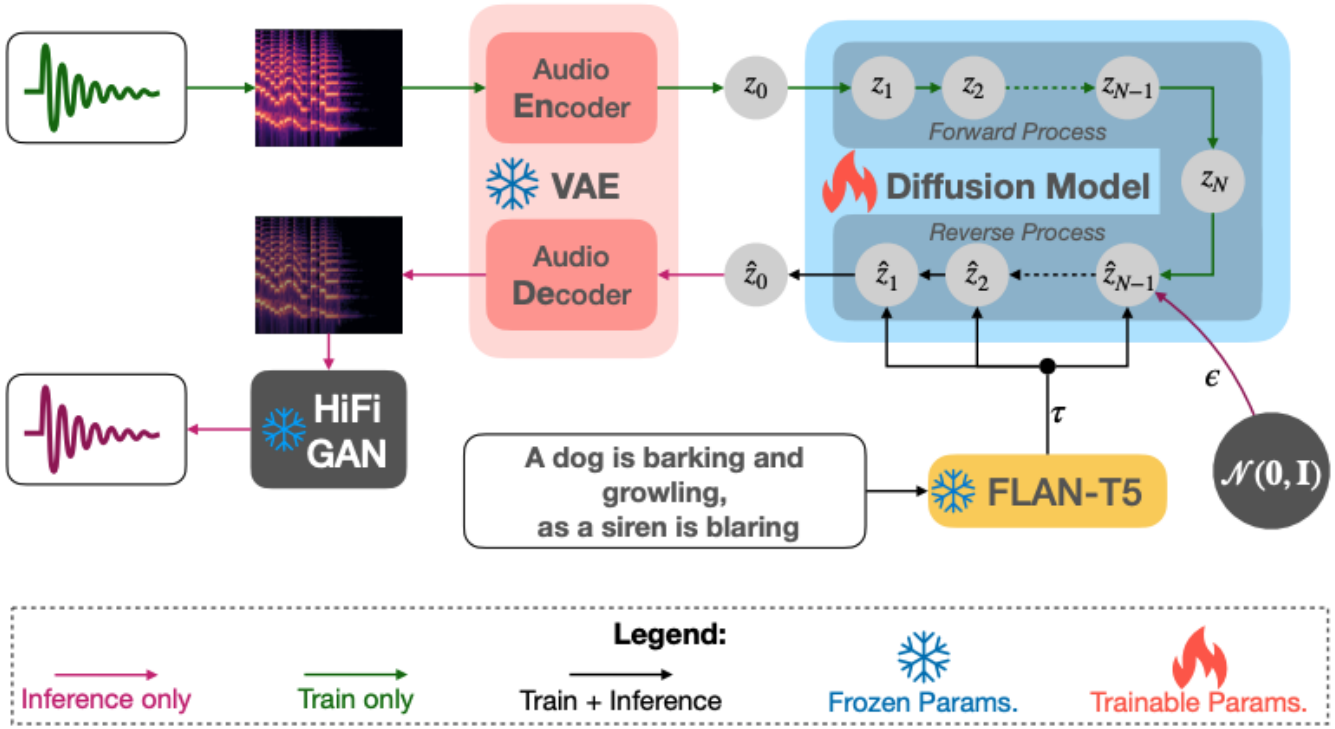
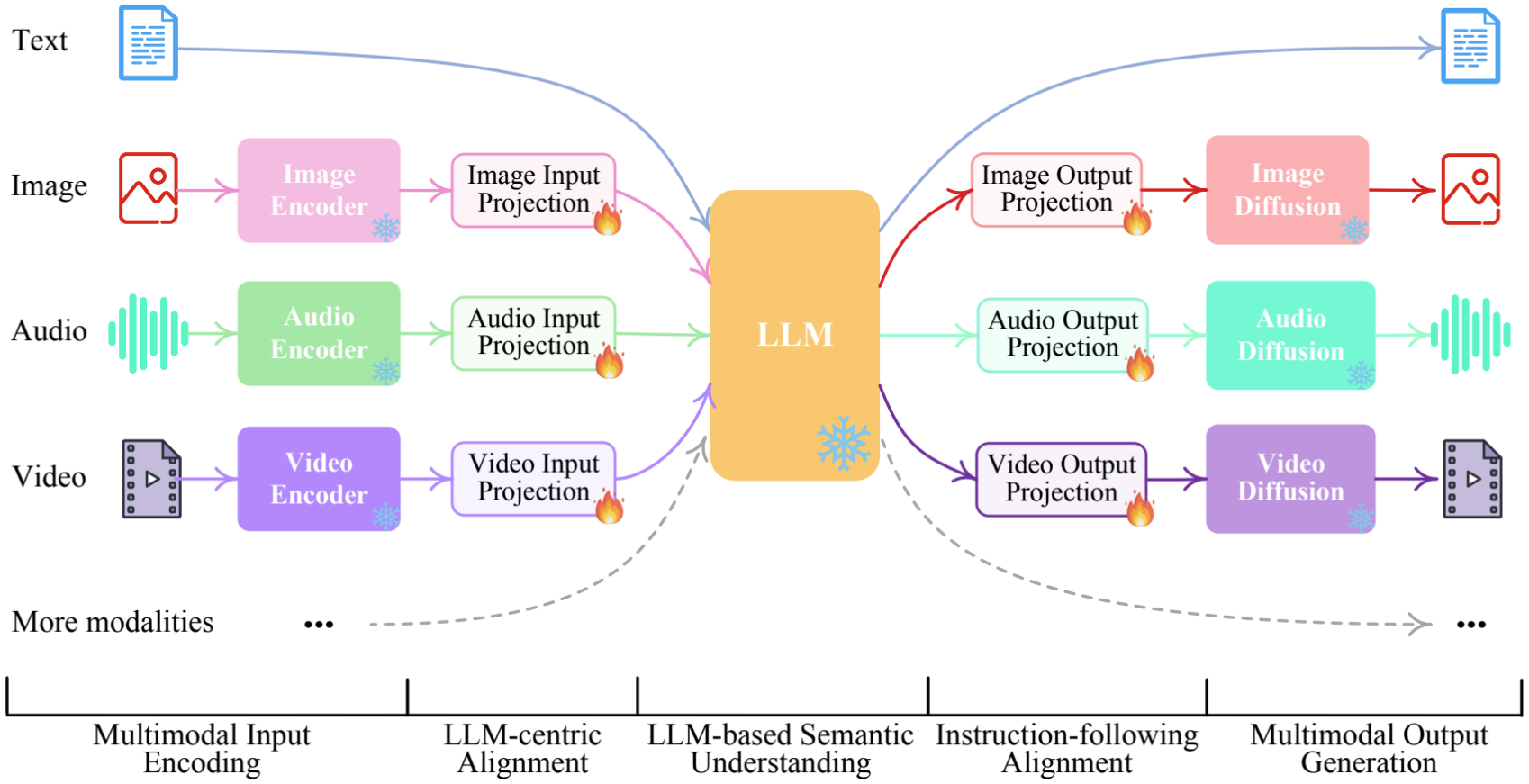


Figure 1: Overall architecture of TANGO.

Final Fantasy: Any-to-Any Generation

NExT-GPT: Any-to-Any Multimodal LLM



Thanks!