

中文课程名称

课程讲义模板

将封面图命名为 `cover.jpg` 并放入 `figures/` 目录后，会自动显示在这里。

教师： 授课教师
学期： 2026 春季
单位： 学校 / 机构名称
日期： 2026 年 8 月 16 日

目录

第一讲 课程导论	1
1.1 课程说明	1
1.2 表格示例	1
1.3 图片示例	1
1.4 例题与练习	2
第二讲 第二讲标题	3
2.1 知识点一	3
2.1.1 更小的层级	3
2.2 长表格示例	3
第五讲 多模态机器学习方法与模型	4
5.1 多模态表示学习与跨模态对齐	5
5.1.1 单模态表示学习：从原始信号到向量	5
5.1.2 自监督学习与预训练表示	9
5.1.3 跨模态对齐	12
5.2 多模态基础模型结构与训练方法	15
5.2.1 多模态基础模型的通用结构	15
5.2.2 视觉语言模型：图像理解中的多模态建模	18
5.2.3 语音语言模型：语音理解中的序列建模	20
5.2.4 多模态指令微调与对齐	23
5.2.5 图像—语音—文本联合建模	24
5.3 条件生成建模与多模态内容生成	24
5.3.1 条件生成模型基础	25
5.3.2 文本生成图像	26
5.3.3 图像编辑：约束条件下的生成	28
5.3.4 语音生成：连续信号的条件生成	29
5.3.5 音频生成：声音事件与声景的条件生成	30
5.3.6 跨模态交互生成	32
5.4 本章小结	33

附录 A 附录	35
A.1 常用 LaTeX 片段	35
参考资料	36

第一讲 课程导论

学习目标

- 了解本课程的学习目标、评价方式与课堂要求。
- 熟悉讲义中的例题、练习、阅读材料与作业模块。
- 掌握如何在模板中插入章节、表格、图片和重点提示。

1.1 课程说明

这里填写本讲的正文内容。中文可直接输入，不需要额外转码。若要突出关键词，可使用 **关键词强调**，也可以使用普通的 **加粗**、*斜体* 等格式。

重点提示

这里可以写课堂重点、易错点、考试提示，或者需要学生特别记住的概念。

1.2 表格示例

表格建议使用 booktabs 提供的 `\toprule`、`\midrule`、`\bottomrule`，视觉会比普通竖线表格更清爽。

表 1.1: 课程安排示例

周次	主题	任务
第 1 周	课程导论与学习方法	阅读讲义
第 2 周	核心概念一	完成练习
第 3 周	核心概念二	小组讨论

1.3 图片示例

将图片放入 `figures/` 目录，然后用 `\includegraphics` 插入。下面的示例会在 `figures/example.png` 存在时显示图片，否则显示一个占位框。



图片占位：将图片保存为 `figures/example.png`

图 1.1: 图片插入示例

1.4 例题与练习

例题 1

这里放置例题内容。可以包含公式：

$$a^2 + b^2 = c^2$$

也可以写解题步骤或分析。

课堂练习

1. 请用自己的话概括本讲的三个重点。
2. 根据表 1.1，说明第 2 周需要完成什么任务。
3. 观察图 1.1，写出你能得到的信息。

第二讲 第二讲标题

2.1 知识点一

从这里开始写第二讲内容。新增章节时，复制下面结构即可：

```
\chapter{新的一讲}
\section{小节标题}
正文内容……
```

2.1.1 更小的层级

如果一讲内部内容较多，可以继续使用 `\subsection` 或 `\subsubsection` 组织层次。

2.2 长表格示例

如果表格跨页，可以使用 `longtable`。

表 2.1: 长表格结构示例

模块	内容	备注
课前	预习阅读材料	可附链接或二维码图片
课中	讲授、讨论、练习	建议加入例题与即时反馈
课后	作业、复盘、扩展阅读	可列出提交要求

第五讲 多模态机器学习方法与模型

前面几章我们建立了机器学习的数学基础与优化方法、强化学习、大语言模型的预训练与微调（包括指令微调与偏好对齐），以及深度学习中的连续动力学与基于流的生成模型。这些内容大多围绕单一模态、尤其是文本数据展开。然而，真实世界的信息通常以多种模态并存：图像是定义在二维空间上的像素信号，语音与音频是随时间变化的声学信号，文本则是离散的符号序列。它们在数据结构、噪声特性与语义粒度上差异显著，却又常常指向相同的语义——同一只猫，可以是一张照片、一句描述，也可以是一段叫声（图 5.1）。

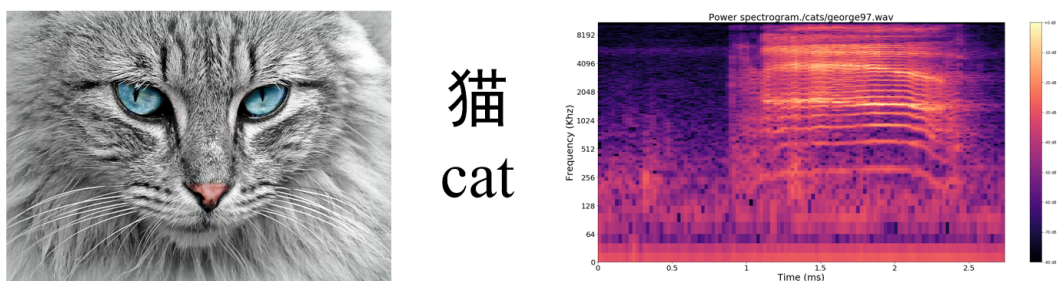


图 5.1: 同一语义“猫”在不同模态中的具体形式：一张猫的图片、文本“猫”（cat）与猫叫声的频谱。表示学习与跨模态对齐所追求的，正是让这三者在向量空间中彼此接近。

本章讨论**多模态机器学习**：研究图像、语音/音频与文本等不同模态如何被统一表示、对齐、建模与生成。全章分为三个部分，依次递进。第一部分是多模态的**表示学习与跨模态对齐**，解决不同模态如何被编码为向量并进入统一语义空间；第二部分是**多模态基础模型的结构与训练**，说明图像、语音如何接入大语言模型，形成理解与交互能力；第三部分是**条件生成建模**，把文本生成图像、图像编辑、语音合成与音频生成，统一表述为在给定条件下学习一个分布并采样的问题。

学习目标

- 掌握图像、语音与文本在计算机中的原始表示形式，以及将其转化为向量表示的动机与方法；
- 理解向量与向量空间的含义，明确“语义相近则表示相近”这一表示学习目标，并据此理解跨模态对齐；
- 掌握以对比学习为代表的跨模态对齐方法及其在图文场景中的实现，了解相关代表工作的发展脉络；
- 理解图像、语音如何接入大语言模型，构成具备多模态理解能力的基础模

型；

- 掌握条件生成建模的统一框架，能够将文本生成图像、图像编辑、语音与音频生成，表述为条件分布 $p(x | c)$ 的学习与采样问题。

5.1 多模态表示学习与跨模态对齐

本节讨论两个基础问题：其一，图像、语音与文本如何从原始数据转化为模型可处理的**向量表示**；其二，不同模态的表示如何在同一空间中建立**语义对应**，即跨模态对齐。这两个问题是后续多模态模型的前提。

5.1.1 单模态表示学习：从原始信号到向量

数据在计算机中的原始形式。 讨论表示学习之前，需要先明确各模态在计算机中以何种形式存在（见图 5.2），因为后续处理都建立在这一原始形式之上。

图像在计算机中通常是一个数值矩阵。屏幕上的图像由许多小方格构成，每个小方格称为一个**像素** (pixel)。一幅灰度图像可表示为一个 $H \times W$ 的矩阵，其中第 (i, j) 个元素是一个整数（常取 0 至 255），表示该像素的明暗，0 为黑、255 为白。彩色图像一般由红、绿、蓝三个通道叠加而成，对应一个 $H \times W \times 3$ 的三维数组。于是一幅 224×224 的彩色图像，大致相当于 $224 \times 224 \times 3 \approx 1.5 \times 10^5$ 个整数。

语音/音频在计算机中通常是一串随时间排列的数值。声音是空气压强的振动，麦克风按固定的时间间隔（采样率，如每秒 16000 次）测量这种振动的强弱，每次测量得到一个数值，称为一个**采样点**，其取值即**振幅**。因此一段语音可视为一个一维的长向量，长度大致正比于时长。

文本在计算机中通常是一串编号。先准备一张**词表** (vocabulary)，把可能出现的词（或子词）逐一编号；一段文本便被转写为这些编号构成的序列。例如词表中“猫”的编号为 352，文本中的“猫”便以整数 352 进入模型。

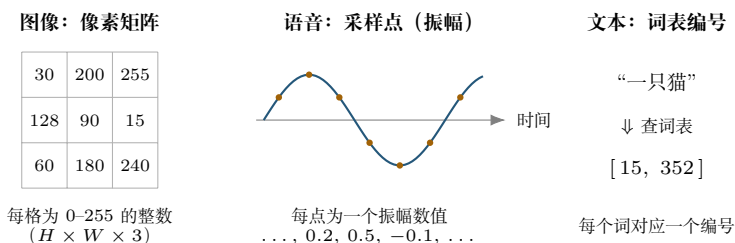


图 5.2: 三种模态在计算机中的原始形式：图像是像素整数矩阵，语音是随时间的振幅采样序列，文本是词表编号序列。

为什么要用向量表示。 上述原始形式虽是数字，却不便于直接学习。一个重要原因是，原始数字之间的数值接近并不等于语义接近：像素值相近的两幅图像内容可能并不相干，内容相近的两幅图像像素值也可能相差很大；词表中编号相邻的两个词往往含义无关。我们希望得到的，是一种语义相近则表示也相近的形式。

为此，机器学习常用**向量**来表示对象。一个向量是一组有序的实数 $\mathbf{x} = (x_1, x_2, \dots, x_d)$ ，其中每个分量 x_k 可理解为对象在某个特征上的取值。也就是说，**向量可以看作对象若干特征的集合**。例如，可设想用不同分量分别刻画一幅图像“是否有毛发”“是否四足”“是否有轮子”等语义特征；若两幅图像在这些特征上取值接近，其向量也接近。真实模型学到的特征并非人为指定，而是从数据中自动学习，但“向量作为特征集合”这一直观通常是成立的。

向量空间与相似度。 全体 d 维向量构成一个 d 维**向量空间**。在这个空间中，每个对象对应一个点，对象间的“相似程度”可用点之间的距离或夹角来量化。常用两种度量：欧氏距离 $\|\mathbf{x} - \mathbf{y}\|_2$ 衡量远近；余弦相似度

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\| \|\mathbf{y}\|} \quad (5.1)$$

衡量方向的一致程度，取值越接近 1 表示方向越一致。表示学习的目标，可以概括为学习一个从原始数据到向量的映射，使得**语义相近的对象在向量空间中距离较小、语义无关的对象距离较大**（图 5.3）。这一点是理解后续内容的关键：跨模态对齐所追求的，正是让表达相同语义的不同模态对象，在同一向量空间中彼此接近。

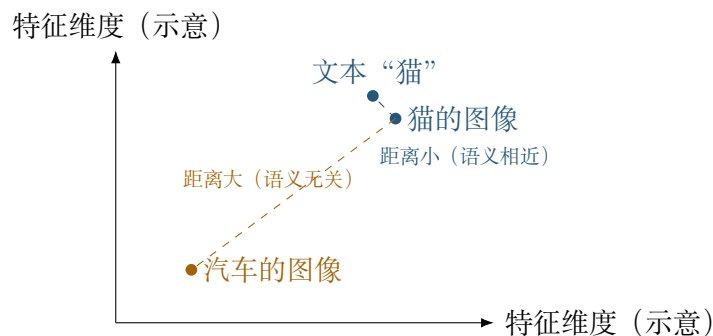


图 5.3: 表示学习的目标：语义相近的对象（无论来自哪个模态）在向量空间中距离较小，语义无关的对象距离较大。

Token：表示的基本单元。 处理图像、语音、文本时，通常不会把整个对象压成单个向量，而是先切分为若干**基本单元**，每个单元各对应一个向量，从而把对象表示为一串向量。这种基本单元称为 **token**。文本的 token 一般是词或子词；图像的 token 常取 16×16 的图块 (**patch**)；语音的 token 可取声学帧或声学单元。把不同模态都整理为“token 序列”的一个好处是：以自注意力为核心的序列模型（Transformer 与大

语言模型) 可以较为统一地处理它们, 这为多模态与大语言模型的结合提供了结构上的便利。

图像表示: 卷积与视觉 token。 如前所述, 原始像素通常不适合直接作为语义表示, 需要一个**编码器**把像素逐层加工为高层特征。图像编码中最经典的运算是**卷积**(convolution)。

卷积的核心是一个小的权重矩阵, 称为**卷积核** K (例如 3×3)。它在输入图像上滑动, 在每个位置与所覆盖的局部区域做逐元素相乘再求和, 得到输出的一个数值; 所有位置的输出组成一张**特征图**。以二维情形为例, 记输入为 I 、卷积核为 K , 则输出特征图 S 在位置 (i, j) 处为

$$S(i, j) = (I * K)(i, j) = \sum_m \sum_n I(i + m, j + n) K(m, n). \quad (5.2)$$

式 (5.2) 表明: 卷积核只关注一小块局部区域 (称为**局部感受野**), 用于检测该处是否出现某种局部模式, 如某方向的边缘或某种纹理; 同一卷积核在整幅图像上共享使用 (图 5.4)。参数共享带来两点好处: 一是显著减少参数量, 二是带来一定的**平移不变性**——同一模式出现在不同位置时, 可由同一卷积核检测到。

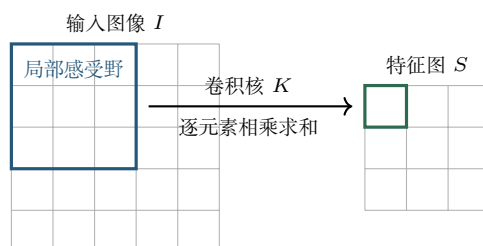


图 5.4: 卷积运算: 卷积核在输入上滑动, 每个局部感受野经“逐元素相乘再求和”得到特征图的一个值。

以边缘检测为例: 取一个用于检测竖直边缘的 3×3 卷积核 (如 Sobel 核), 按式 (5.2) 在图像上滑动, 即可在灰度变化剧烈的竖直边缘处得到较大响应、在颜色平坦处得到接近 0 的响应, 从而“点亮”物体的竖直轮廓 (图 5.5); 换用不同的卷积核, 则可分别检测水平边缘、斜向纹理等不同局部模式。

图像编码器的发展脉络。 把多层卷积叠加起来, 可从低层的边缘、纹理逐层组合出更高层的语义, 这类网络称为**卷积神经网络** (CNN)。CNN 的发展大体经历了从 AlexNet[1] 确立深度卷积网络在图像识别上的有效性, 到 VGG[2]、ResNet[3] 通过残差连接把网络显著加深的过程。另一类做法不依赖卷积, 而是把图像切成固定大小的图块 (patch): 例如把 224×224 的图像切成 14×14 个 16×16 的图块, 每个图块经线性映射成一个向量, 即一个**视觉 token**, 随后用自注意力建模这些 token 间的关系, 这便是 Vision Transformer (其结构见图 5.6) [4]。两类结构的训练目标类似, 都是让所得视觉表示服务于具体任务 (如分类), 并使语义相近的图像表示彼此接近。经过

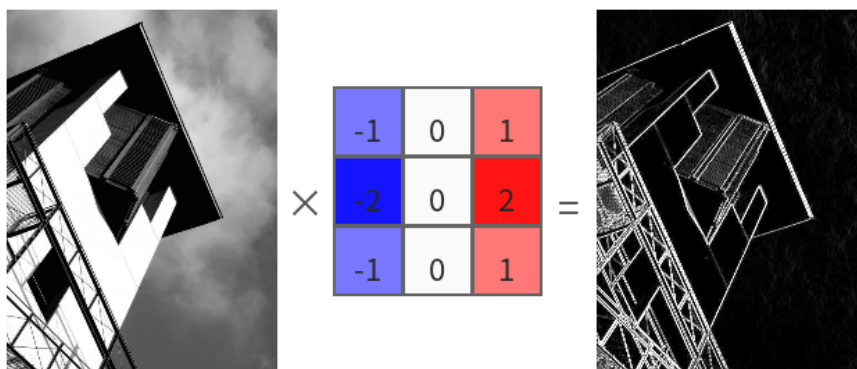


图 5.5: 卷积核用于边缘检测：竖直边缘检测核（中间，Sobel 核）对左侧图像做卷积，得到右侧突出竖直轮廓的特征图。

编码后，图像被表示为一组视觉 token，其序列形式与文本 token 较为接近，这为二者进入同一模型创造了条件；不过二者在语义粒度上仍有差异，视觉 token 更接近局部区域特征，而文本 token 通常已对应较完整的语义单位。

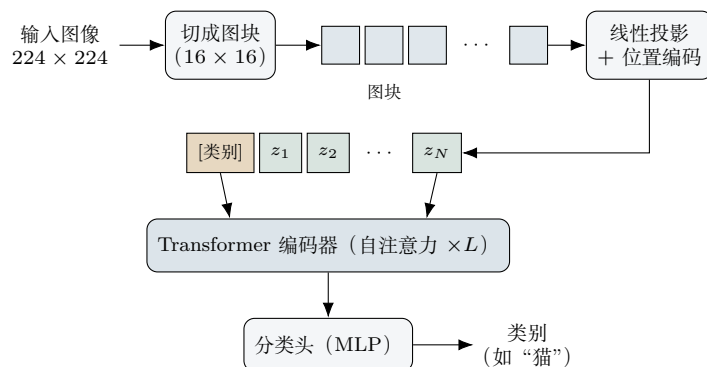


图 5.6: Vision Transformer (ViT)：图像切成固定大小的图块，每块经线性投影加位置编码成为一个视觉 token；连同一个额外的“[类别]” token 一起送入 Transformer 编码器，最后由分类头输出类别。

语音表示：从波形到声学 token。 语音在计算机中的原始形式是一维**波形** (waveform)，即按时间顺序记录的振幅采样点序列。若采样率为 16000 Hz，则一秒语音包含 16000 个数值。波形完整保留了原始声学信号，但其数值随时间快速振荡，且会同时混合说话人音色、发音内容、语速、情绪、环境噪声等多种因素，因此通常不直接把整段波形作为高层语义表示。

一种经典处理方式是先把波形转换为**频谱图** (spectrogram)。具体而言，将连续波形切分为许多短时片段，称为**帧** (frame)；常见帧长约为 20–25 ms，相邻帧之间可有重叠。对每一帧进行短时傅里叶变换 (short-time Fourier transform, STFT)，即可得到该短时间内不同频率成分的强弱。若记波形为 $x[n]$ 、窗函数为 $w[n]$ 、第 t 帧的频

率索引为 k ，则其复频谱可写为

$$X(t, k) = \sum_n x[n + tH] w[n] e^{-j2\pi kn/N}, \quad (5.3)$$

其中 H 是帧移， N 是每帧长度。通常取其模平方 $|X(t, k)|^2$ 作为能量，因此频谱图可以理解为一个“时间 $\times$ 频率”的二维矩阵：横轴表示时间，纵轴表示频率，每个位置的数值表示该时刻附近某个频率成分的能量。

人耳对不同频率的分辨能力并不均匀，低频区域更敏感，因此语音处理中还常将频谱投影到**梅尔尺度**（Mel scale）上，得到**梅尔频谱**（Mel spectrogram）。其基本做法是用一组按梅尔尺度分布的三角滤波器，对频谱能量进行加权汇总，再取对数：

$$M(t, m) = \log \left(\sum_k F_m(k) |X(t, k)|^2 + \epsilon \right), \quad (5.4)$$

其中 $F_m(k)$ 表示第 m 个梅尔滤波器。梅尔频谱保留了语音中与发音内容、音高、共振峰和韵律相关的重要结构，同时压缩了不必要的频率细节，因此长期以来是语音识别、说话人识别和语音生成中的常用输入特征。波形、频谱图与梅尔频谱三种表示的对照见图 5.7。

无论输入是原始波形还是梅尔频谱，语音编码器的目标都是把每一帧或一小段连续信号映射为一个高维向量，从而得到按时间排列的**声学 token** 序列。这里的声学 token 通常是连续向量，并不一定对应一个完整的音节、词或字符；其语义粒度往往比文本 token 更细，还会保留部分说话人、情绪和韵律信息。与图像中“图块 $\rightarrow$ 视觉 token”的过程类似，语音也可视为“时间帧 $\rightarrow$ 声学 token”的过程；区别在于，图像主要在二维空间上展开，而语音沿一维时间轴展开，前后 token 的顺序具有更强的因果性和动态性。同一句话在不同语速下持续时间不同，因此语音 token 序列通常也比对应文本 token 序列长得多。

文本表示：分词、词表与词嵌入。 文本的表示相对直接，其流程如图 5.8 所示：先把一段文字切分为基本单元（词或子词），称为**分词**；每个单元在**词表**中对应一个编号；再通过一张可学习的**词嵌入**（word embedding）表，把编号映射为向量。于是一句话最终成为一串词向量，与图像的视觉 token、语音的声学 token 在形式上一致。

对照图 5.8，以“一只猫”为例：先切分为“一只”“猫”两个 token，查词表得到各自编号，再经词嵌入表取出对应向量。词嵌入的意义与前文一致——把离散编号转化为特征向量，使语义相近的词（如“猫”与“猫咪”）向量也相近。

5.1.2 自监督学习与预训练表示

有了编码器，还需回答它如何学到有用的表示。若依赖“数据+人工标签”的监督训练，则受限于标注的规模与成本；而图像、语音、文本的无标注数据较为丰富。

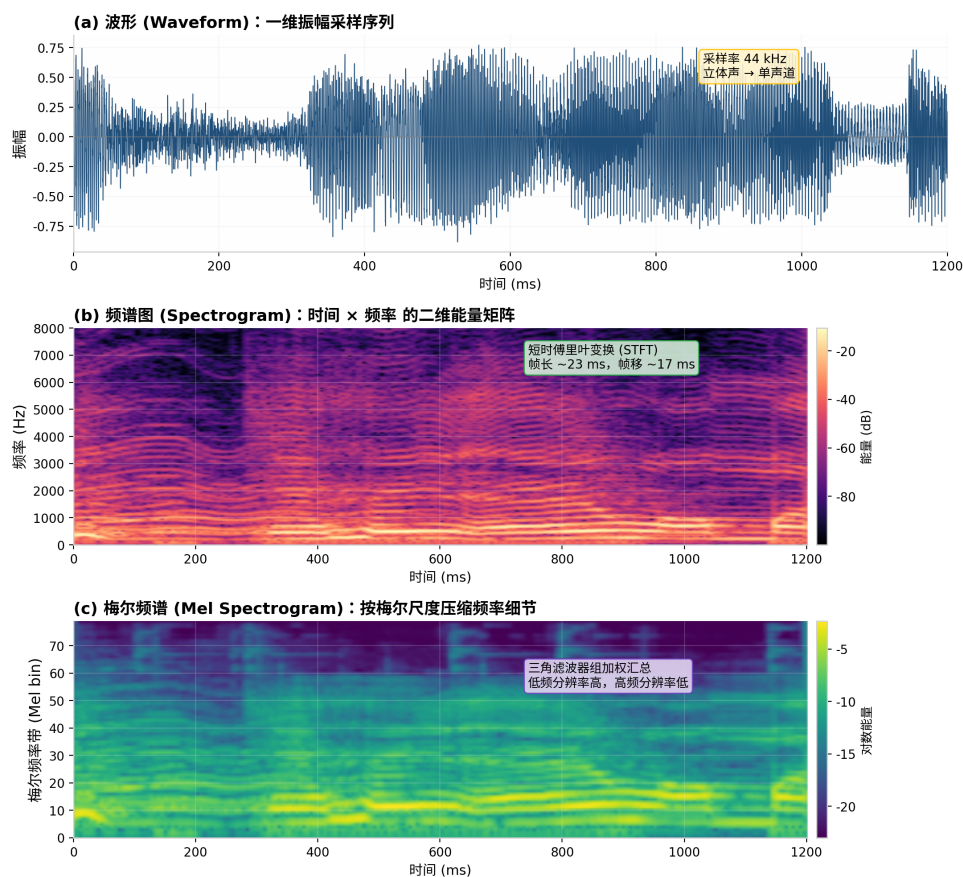


图 5.7: 语音的三种常见表示: (a) 波形, 按时间排列的一维振幅采样序列; (b) 频谱图, 经短时傅里叶变换得到的“时间 × 频率”二维能量矩阵; (c) 梅尔频谱, 按梅尔尺度对频率细节加权压缩后的结果。

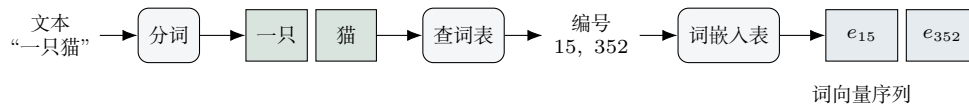


图 5.8: 文本表示流程: 分词得到子词 token, 查词表得到编号, 再由词嵌入表映射为词向量序列。

自监督学习的基本思想，是从数据自身构造监督信号，无需人工标注，从而能够利用大规模数据进行**预训练**，得到可迁移到下游任务的通用表示。下面着重说明其在图像上的两类典型训练目标。

掩码重建。 随机遮挡输入的一部分，让模型根据未遮挡部分重建被遮挡的内容。在图像上，可遮挡若干图块再要求模型恢复其像素或特征；要重建得较好，模型往往需要学到图像的结构与语义规律。其训练目标是**最小化重建误差**，与文本中的掩码语言建模在形式上相近。这类方法的代表包括 MAE[12] 与 BEiT[13]。以 MAE 为例（图 5.9），它随机遮住约 75% 的图块，只把剩余可见图块送入编码器，再由一个轻量解码器重建被遮部分；要补得合理，模型必须理解物体结构与上下文（例如由半只火烈鸟推断其全貌）。训练完成后，这套编码器学到的通用表示可迁移到分类、检测等下游任务。

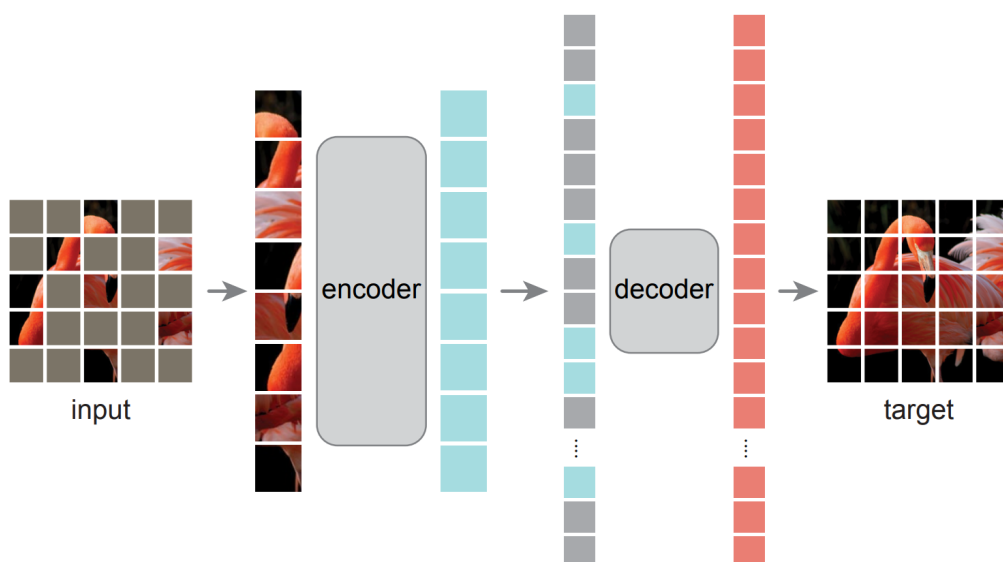


图 5.9: 掩码自编码器 (MAE): 随机遮住大部分图块 (左, input), 仅将可见图块送入编码器, 再由解码器重建被遮部分, 训练目标是还原原图 (右, target)。

对比学习。 对同一对象施加两种不同的随机变换（如对图像做不同的裁剪与颜色扰动），得到的两个视图视为同一语义，称为**正样本对**，要求其表示接近；来自不同对象的视图为**负样本对**，要求其表示远离。其训练目标使正样本对的相似度高于负样本对，从而学到对无关变化（如位置、颜色）较不敏感、对语义较敏感的表达。代表工作包括 SimCLR[5] 与 MoCo[6]；此外 DINO[7] 等自蒸馏方法也取得了较好效果。下一小节的跨模态对齐，可视为对比学习思想从“不同视图”到“不同模态”的推广。

语音的自监督。 语音同样有海量的无标注录音，因此也希望用自监督的方式学到通用表示。它的核心思路与图像的掩码重建一脉相承，可以概括成一句话：**遮住一段语**

音，再让模型根据前后听到的内容把它“猜”回来（图 5.10）。为了猜得准，模型必须理解语音里稳定的发音规律与上下文，这正是我们想要的表示。

不过语音有一个和图像、文本都不同的特点：它是连续变化的声波，相邻的采样点几乎一模一样，信息高度冗余。如果直接要求模型逐点还原原始波形，它只要照抄邻近的点就能蒙混过关，学不到有用的东西。因此语音自监督一般不去还原波形本身，而是预测更“抽象、稳定”的目标。按预测目标的不同，有两种有代表性的做法。

第一种是从若干候选里认出正确答案，代表是 wav2vec 2.0[9]。它先用一个卷积网络把波形转成一串声学向量，随机遮住其中一段，再让 Transformer 根据没被遮住的上下文，为被遮位置给出预测；训练时同时准备好该位置真正的向量和若干来自别处的“干扰向量”，要求模型从中认出正确的那一个。这本质上就是上一段对比学习的思想在时间轴上的应用。

第二种是先给语音编号，再做“完形填空”，代表是 HuBERT[10]。它先用聚类（例如对声学特征做 K-means）把每一帧归入某一类，得到一个类别编号，并把这些编号当作无需人工标注的“伪标签”；预训练时遮住一段语音，让模型根据上下文预测被遮位置属于哪个编号。这样，原本连续的语音就被看成一串离散“单元”，预测任务与文本的掩码语言建模非常相似，模型也因此更容易学到音素一级的结构。

两种做法的预测目标不同（认向量还是猜编号），但共享同一个骨架：**遮挡，再用上下文预测**。这也再次说明，图像、文本与语音的自监督在思想上是相通的。

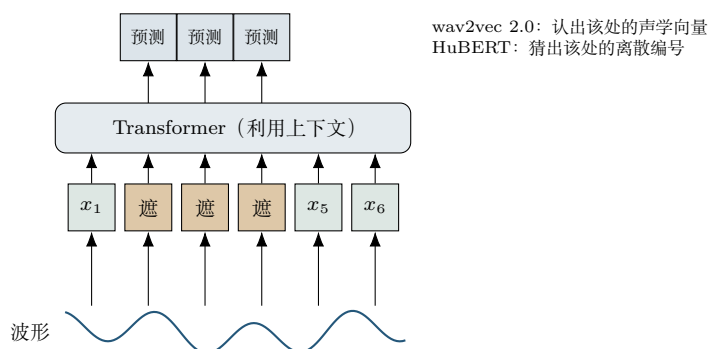


图 5.10: 语音自监督的共同思路：遮住一段语音，让模型根据前后未遮挡的上下文预测被遮部分。wav2vec 2.0 预测该处的声学向量，HuBERT 预测该处的离散编号。

5.1.3 跨模态对齐

前两小节使每个模态各自得到了向量表示，但不同模态的编码器多是各自训练的，其向量分处不同空间，通常难以直接比较。**跨模态对齐**要解决的问题是：把不同模态的表示纳入同一个向量空间，使表达相同语义的对象（如一幅猫的图像与文本“猫”）在该空间中较为接近。

实现对齐的一个常用途径，是利用天然存在的**配对数据**。例如互联网上大量“图像配文字”的样本，本身就标明了哪幅图与哪句话语义对应。据此可训练图像编码器

与文本编码器，使配对样本的表示靠近、非配对样本的表示远离——这正是对比学习从“同一模态的不同视图”推广到“不同模态的配对”的结果。这一思路的代表是 CLIP[14] 与几乎同期的 ALIGN[15]：前者使用较高质量的图文数据，后者表明利用规模更大但噪声更多的网络图文数据同样可行。此后的工作在此基础上做了不少改进，例如 SigLIP[16] 用 sigmoid 形式的损失替代批内 softmax，以改善大批量训练下的稳定性与效率。

以 CLIP 为例，取一批 N 对图文。图像编码器把 N 幅图编码为向量 $\mathbf{I}_1, \dots, \mathbf{I}_N$ ，文本编码器把 N 句话编码为向量 $\mathbf{T}_1, \dots, \mathbf{T}_N$ ，并归一化到同一空间。计算任意一幅图 $\mathbf{I}_i$ 与任意一句话 $\mathbf{T}_j$ 的相似度（可用内积 $\mathbf{I}_i \cdot \mathbf{T}_j$ 表示），得到一个 $N \times N$ 的相似度矩阵（图 5.11）：对角线元素 $\mathbf{I}_i \cdot \mathbf{T}_i$ 为真实配对，应取较高值；非对角线元素为错误配对，应取较低值。训练即调整两个编码器，使该矩阵趋于“对角线高、其余低”。

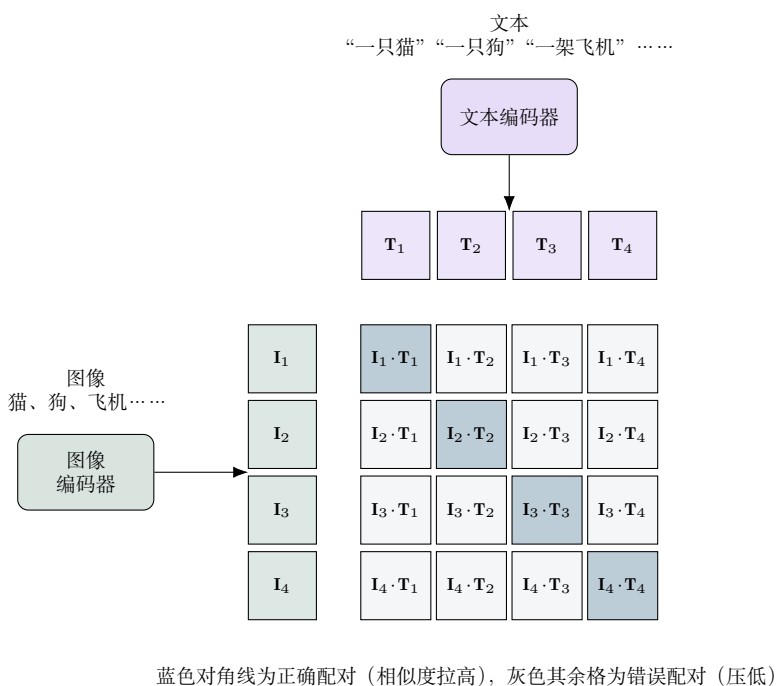


图 5.11: 图文对比对齐 (CLIP): 图像编码器输出 $\mathbf{I}_i$ 、文本编码器输出 $\mathbf{T}_j$ ，两两做内积得到 $N \times N$ 相似度矩阵。训练目标是使对角线元素 $\mathbf{I}_i \cdot \mathbf{T}_i$ 较高、非对角线元素较低。

把这一目标写成损失函数，即对比损失 (InfoNCE)。对第 i 幅图，其“图到文”方向的损失为

$$\ell_i^{\text{图} \rightarrow \text{文}} = -\log \frac{\exp(\cos(\mathbf{I}_i, \mathbf{T}_i)/\tau)}{\sum_{j=1}^N \exp(\cos(\mathbf{I}_i, \mathbf{T}_j)/\tau)}. \quad (5.5)$$

该式可逐项理解：分子是第 i 幅图与其正确配文的相似度，分母是该图与批内全部 N 句话相似度之和。整个分式是一个 softmax，表示“在这 N 句话中，模型判定第 i 句为正确配文的概率”；取负对数后，概率越接近 1、损失越小。 τ 为温度参数， τ 越小，

对“正确配对显著高于错误配对”的要求越严格。将“图到文”与对称的“文到图”两个方向的损失相加，即为完整的对齐目标。

经过对齐，文本与图像被纳入同一向量空间，于是“用文字检索图像”“判断图文是否相符”等任务，大体可归结为在该空间中计算相似度。更重要的是，对齐后的文本表示常被用作图像生成的条件——5.3 的文本生成图像，往往依赖这种图文对应关系，使文本描述能够引导图像的生成方向。

对齐后的图文空间还能直接用于分类，且无需为新类别重新训练：给定一张图片，把若干候选类别分别写成文本“一张关于某类别的照片”（如“猫”“狗”“飞机”），编码后与图像向量计算相似度，相似度最高者即预测类别。只要更换候选文本，就能识别训练时未专门标注过的新类别，这称为**零样本分类**。

语音、音频的对齐。 语音和文本之间天然存在配对关系。例如，语音识别数据提供“语音-文字转写”，语音翻译数据提供“语音-文字翻译”。语音编码器和文本编码器分别将二者映射为向量：

$$\mathbf{s}_i = f_{\text{speech}}(S_i), \quad \mathbf{u}_i = f_{\text{text}}(T_i). \quad (5.6)$$

匹配的语音-文本为正样本，不匹配的组合为负样本。训练仍采用式 (5.5) 的对比损失，使正确配对的表示相互靠近，错误配对的表示相互远离，如图 5.12 所示。

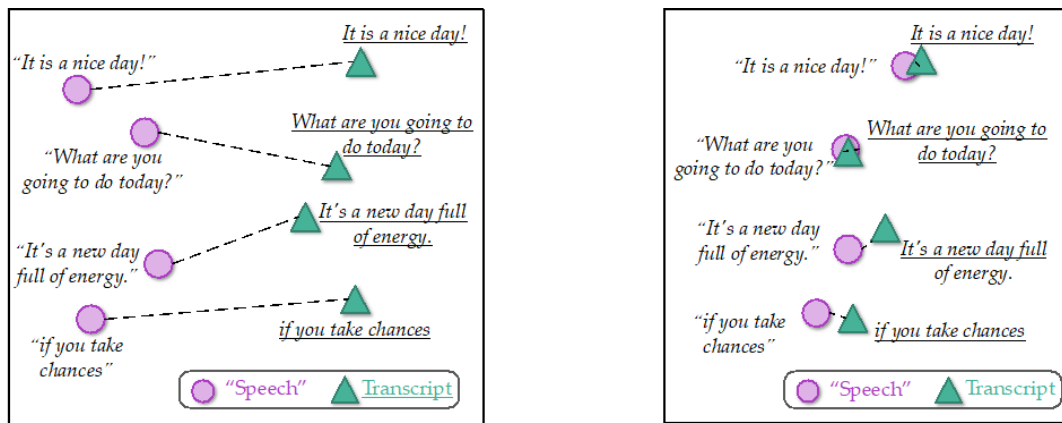


图 5.12: 语音-文本表示的对齐过程。左图为对齐前，右图为对齐后；圆形表示语音，三角形表示对应的文字转写。

句级与词级对齐。 句级对齐将整段语音与整句话分别压缩为一个向量，适合语音检索和语音翻译。词级对齐则进一步确定某段语音帧对应哪个文本 token。例如，WACO[17] 通过拉近语音词片段与对应文本词的表示，提升低资源语音翻译效果。

音频-文本对齐。 更广义的音频还包括音乐、动物叫声、交通声和环境声等。这类音频通常与“狗在吠叫”“雨声伴随雷声”等文字描述配对。CLAP[18] 将 CLIP 中的图像编码器替换为音频编码器，在共享空间中对齐音频和文字描述，如图 5.13 所示。

需要注意。 文字转写主要描述语音中的“说了什么”，通常不会完整记录说话人、情

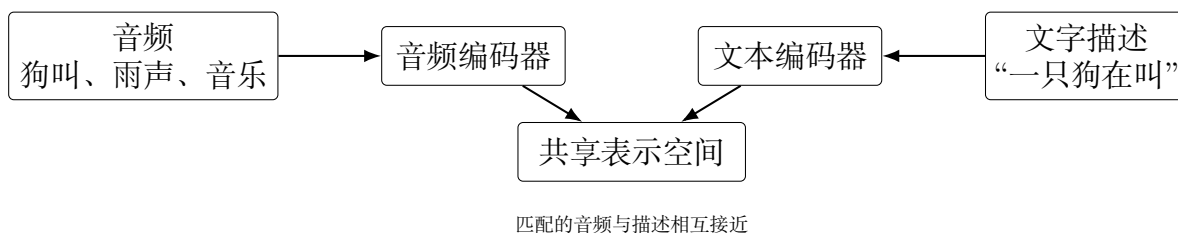


图 5.13: 音频-文本对齐的基本结构。

绪、音色和韵律。因此，只使用转写进行对齐时，模型更容易学习语言内容；若要保留副语言信息，还需加入说话人、情绪或韵律相关的训练目标。

5.2 多模态基础模型结构与训练方法

本节讨论如何把图像、语音等非文本模态接入大语言模型，构成能够理解多模态输入、并按指令完成任务的**多模态基础模型**。需要说明的是，这类模型一般并非把图像模型、语音模型与语言模型简单并置，其关键的机器学习问题，是不同表示空间之间的映射与联合优化。大语言模型自身的预训练、指令微调与对齐方法已在第三部分介绍，本节着重于多模态特有的结构与训练环节。

5.2.1 多模态基础模型的通用结构

现有多模态基础模型大多可归入同一通用结构（图 5.14）：**模态编码器、连接模块与大语言模型**三段相连。模态编码器（即 5.1 所述的图像或语音编码器）把非文本输入编码为高层表示；大语言模型负责语义理解与文本生成；二者之间的连接模块负责把模态表示映射到大语言模型可接受的输入空间。

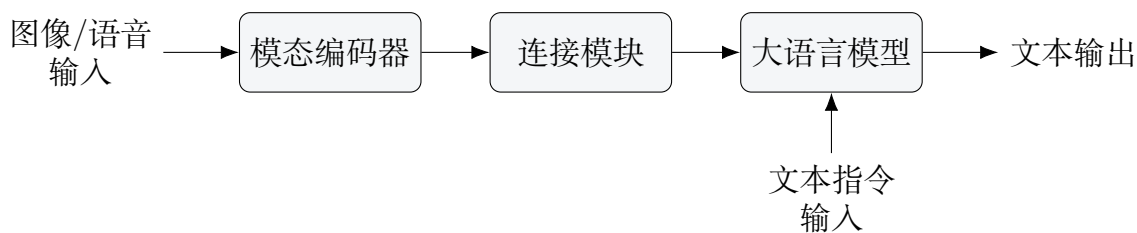


图 5.14: 多模态基础模型的通用结构：编码器负责感知，连接模块负责把模态表示映射到语言模型空间，文本指令与映射后的模态表示一并输入大语言模型，由其完成理解与表达。

按融合发生的早晚看三类结构。 上述通用结构之外，还需回答一个问题：来自不同模态的信息，究竟在模型的哪一步结合到一起？按**融合发生的早晚**，大致可分为三类（图 5.15），它们正对应连接方式由浅到深的不同选择：

- **后期融合**：图像先经一个独立的视觉编码器完整编码，再经投影与文本拼接后接入语言模型，融合发生得较晚、较浅。各模块相对独立，最易复用现成的视觉编码器与语言模型。代表是 LLaVA[21]，其连接模块只是一个轻量投影层。上文给出的通用结构即以这类为典型。
- **中间层融合**：图像与文本先各自编码，再通过**交叉注意力**在编码器与语言模型之间交互。代表是 BLIP-2[20]：它在冻结的图像编码器与冻结的语言模型之间插入一个 Q-Former，用一组可学习查询向量、以交叉注意力从图像特征中提取出少量视觉向量，再送入语言模型；Flamingo[19] 则在语言模型各层插入门控交叉注意力以引入视觉信息。
- **早期融合**：把图像编码成嵌入后，与文本嵌入交错成同一序列，交由一个联合训练的统一模型处理；融合发生在输入端、程度最深，模型甚至可以统一地理解并生成图像与文本。代表是 Emu[28]，它把视觉嵌入与文本 token 放入同一自回归序列端到端训练。

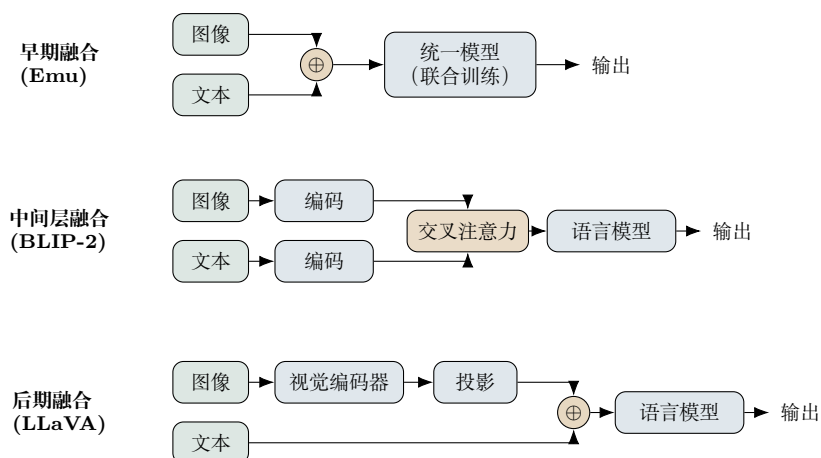


图 5.15: 三类融合按“融合发生的早晚”区分 ($\oplus$ / 交叉注意力为融合点，由上到下依次后移)：早期融合 (Emu) 在输入端就把图文嵌入并入同一序列、统一建模；中间层融合 (BLIP-2) 各自编码后用交叉注意力在编码器与语言模型之间交互；后期融合 (LLaVA) 让视觉先经独立编码器完整编码，末端才与文本拼接接入语言模型。

三者的差别，直观上就是图 5.15 中融合点由上至下逐渐后移：越靠早融合，模态间交互越深、越依赖联合训练；越靠后融合，各模块越独立、越易复用已有的视觉编码器与语言模型。需要说明的是，这一划分并非绝对，实际模型可能兼有多种方式；连接模块的具体形式将在 5.2.2 展开。

连接模块与分阶段训练。 无论属于上述哪一类，连接模块的作用都是把模态编码器输出的向量，转换为若干与文本 token 处于同一空间的向量，相当于把图像或语音“翻译”为语言模型可读取的一段输入。训练上常采用**分阶段策略**：先冻结模态编码器与大语言模型、只训练连接模块，使两个表示空间先行对齐，这样既降低计算开销，又有助于保留语言模型已有的能力；随后再对整体做小幅微调。

语音的接入。 语音和音频同样可以按照“编码器-连接模块-大语言模型”的结构接入。首先，语音编码器将波形或梅尔频谱转换为声学表示：

$$\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_T.$$

连接模块再对这些表示进行**时序压缩**和空间投影，得到较短的语音 token 序列：

$$\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_K, \quad K \ll T.$$

这些语音 token 与文本指令一起输入大语言模型，使其能够完成语音问答、语音翻译和音频理解等任务，如图 5.16 所示。

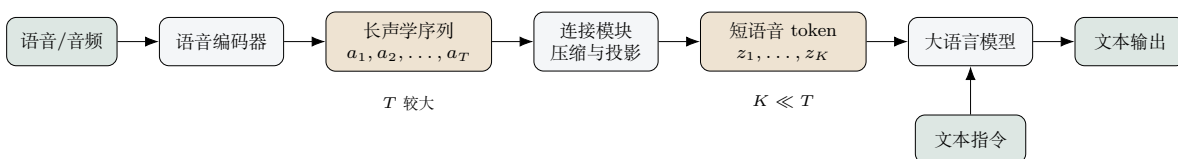


图 5.16: 语音接入大语言模型的基本结构。连接模块先压缩较长的声学序列，再将其投影为大语言模型可读取的语音 token。

语音与图像接入的主要区别在于，语音沿时间轴连续展开，声学 token 通常数量更多。如果直接将全部声学 token 输入语言模型，不仅计算开销较大，还可能挤占文本上下文长度。因此，连接模块常采用卷积下采样、池化、查询向量或注意力压缩等方法，在尽量保留语义和时间顺序的同时缩短序列。

语音的融合。 语音首先经过语音编码器，被转换为按时间排列的声学 token $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_T$ 。这些 token 与文本或图像信息融合时，同样落在前面“融合发生早晚”的框架内：可以把声学 token、图像 token 和文本 token 拼接成统一序列，再交给同一个 Transformer 处理（偏早期融合）；也可以保留独立的语音编码器，让文本 token 通过交叉注意力读取语音特征（偏中间层融合），如图 5.17 所示。

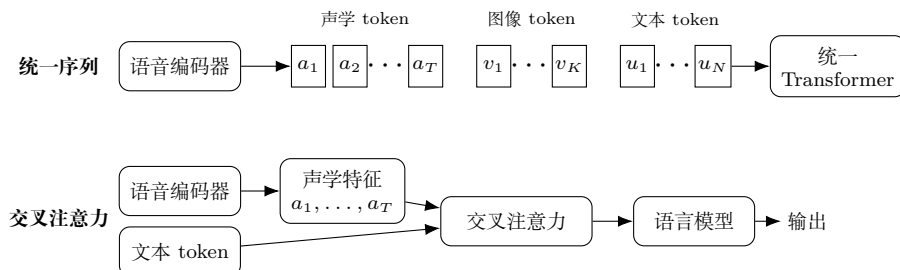


图 5.17: 语音与其他模态的两种常见融合方式。上：将声学、图像和文本 token 拼接为统一序列；下：文本通过交叉注意力读取语音特征。

与文本相比，语音 token 数量更多，并且具有明确的时间顺序。因此，融合前通常需要通过卷积或池化进行下采样，以缩短序列；同时加入位置编码，保留“先说什

么、后说什么”的时序关系。对于语音问答、语音翻译等任务，模型还应避免在处理当前时刻时破坏语音原有的时间结构。

5.2.2 视觉语言模型：图像理解中的多模态建模

把上述通用结构配以图像编码器，即得到**视觉语言模型 (VLM)**：能够接收图像输入，并围绕图像进行描述、问答与推理。从机器学习角度看，视觉语言模型把图像表示作为条件，使大语言模型在该条件下进行文本生成，即建模条件分布 $p(\text{文本} | \text{图像})$ 。

其工作方式如图 5.18 所示：图像经编码器与连接模块转化为若干视觉 token，置于文本提示词之前，与文本 token 共同构成大语言模型的输入；模型随后如同处理普通文本上下文那样，自回归地逐词生成回答。对模型而言，图像相当于上下文的一部分，区别在于这部分来自图像。

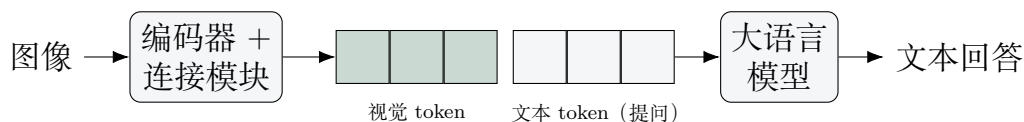


图 5.18: 视觉语言模型：图像被编码为视觉 token，与文本 token 拼接后输入大语言模型，由模型在图像条件下生成回答。

代表工作：连接方式的对比。 不同视觉语言模型的主要差别，往往在于**连接模块**的设计，即如何把视觉编码器的输出接入大语言模型。图 5.19 对比了三种有代表性的做法，三者的视觉编码器（绿色）与大语言模型（蓝色）大体相同，差异集中在中间的连接模块（黄色）。

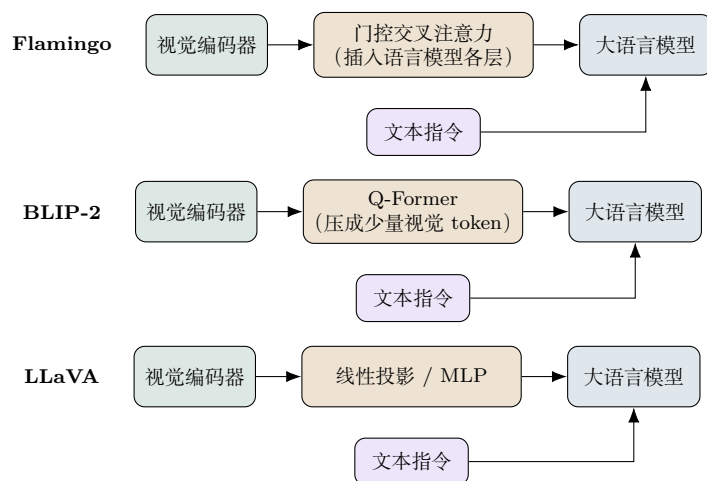
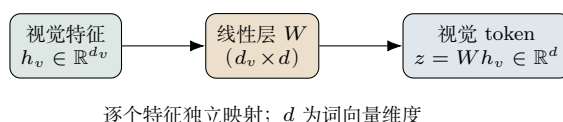


图 5.19: 三种代表性视觉语言模型的连接方式对比：视觉经编码器与连接模块（黄色）接入大语言模型，文本指令（紫色）同样输入大语言模型；三者的差异主要在于中间连接模块的形式。

对照图 5.19：第一行的 **Flamingo**[19] 在冻结的语言模型各层中插入门控交叉注意力，让文本在生成时按需“查看”视觉特征，适合处理图文交错的序列与少样本学

习；第二行的 **BLIP-2**[20] 在冻结的图像编码器与冻结的大语言模型之间接入一个轻量的 **Q-Former**：它带有一组可学习的查询向量（如 32 个），通过交叉注意力从图像特征中提取出固定数量的视觉向量（数量只由查询个数决定，与图像分辨率、图块数量无关），经一个线性层投影后作为“软提示”送入语言模型；训练时只更新 Q-Former 与该投影层，并分两阶段进行——先把 Q-Former 接在图像编码器上学习图文对齐的视觉表示，再把其输出接入语言模型学习看图生成文本，因而参数高效；第三行的 **LLaVA**[21] 采用最简单的**线性投影（或 MLP）**，把视觉特征直接映射到词向量空间，充当视觉 token。可以看出，连接模块从复杂的交叉注意力逐步走向更轻量的投影，而模型效果并未因此下降——这与下面要讲的训练方式密切相关。

三种连接模块的具体结构。 图 5.19 只标出了连接模块的名称，下面进一步说明每一种“黄色方块”内部到底是什么。**线性投影**（图 5.20）最简单：把视觉编码器输出的每个特征向量，用一个矩阵 W （或两层 MLP）映射到与词向量相同的维度，所得向量直接当作视觉 token 插入文本序列。**Q-Former**（图 5.21）引入一组可学习的查询向量，通过交叉注意力从大量图像特征中“问”出所需信息，输出固定数量的少量 token，从而把成百上千个图像特征压缩到几十个，显著缩短序列。**门控交叉注意力**（图 5.22）不改动文本序列，而是在语言模型的层中插入一个交叉注意力子层，让文本隐状态去读取视觉特征；其输出先乘一个门控 $\tanh(\alpha)$ 再残差加回， α 初始为 0，使训练之初几乎不影响原语言模型，因而更稳定。可以看出，从线性投影到 Q-Former 再到门控交叉注意力，连接模块在结构复杂度和对原模型的改动程度上依次增加。



逐个特征独立映射； d 为词向量维度

图 5.20: 线性投影连接模块 (LLaVA)：用一个矩阵 W （或两层 MLP）把视觉特征映射到与词向量同维的空间，直接充当视觉 token。

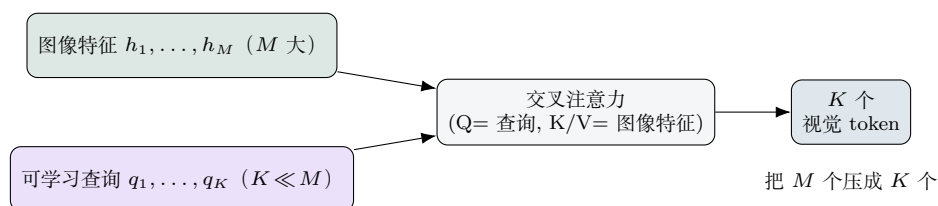


图 5.21: Q-Former 连接模块 (BLIP-2)：一组可学习查询向量通过交叉注意力从大量图像特征中提取信息，输出固定的少量视觉 token，实现压缩。

训练范式的演进。 除连接结构外，训练方式也在演进。早期以图文对齐或图文生成的预训练为主，例如 CLIP[14] 训练出的图文对齐视觉编码器，常被后续模型直接复用为视觉塔。此后**视觉指令微调**成为关键一步：LLaVA 借助较强的纯文本模型，自动构造“图像+指令+回答”数据对模型进行微调，使其能较好地遵循人类指令；

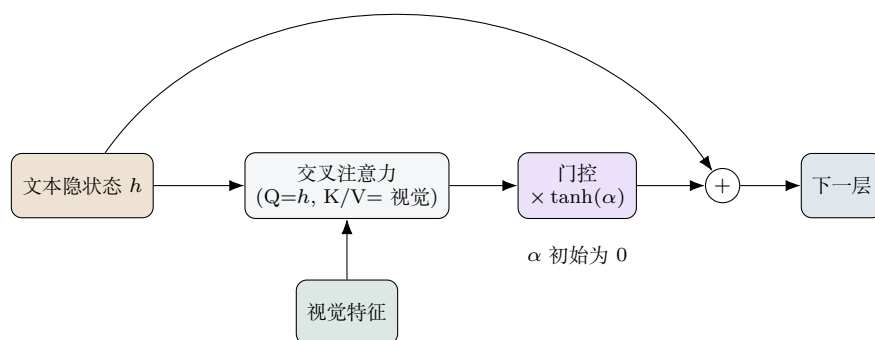


图 5.22: 门控交叉注意力连接模块 (Flamingo): 在冻结的语言模型层中插入交叉注意力读取视觉特征, 输出经门控 $\tanh(\alpha)$ 缩放后残差加回; α 初始为 0, 初始时几乎不干扰原模型。

InstructBLIP[22]、Qwen-VL[23] 等则在指令数据与模型结构上进一步扩展。需要说明的是, 这些模型在不同任务上各有优劣, 且“主流”随时间变化, 此处仅将其视为发展过程中的代表, 而非确定的优劣排序。

重点提示

视觉语言模型可能出现**视觉幻觉**现象: 模型有时会描述出图像中并不存在的内容。其成因之一, 可能是语言先验较强——模型从大量文本中习得了“某场景通常包含某物”的统计倾向, 以致在图像未提供相应证据时仍作出断言。抑制幻觉是多模态对齐训练关注的问题之一。

举一个具体例子: 给定一张“厨房餐桌上放着水果”的图片与问题“图中有几个苹果?”, 模型先由视觉编码器把图像转成视觉 token, 与问题文本一起输入大语言模型, 再据图像内容作答“有两个苹果”; 若图中并无苹果而模型仍答“有一个苹果”, 便是上述视觉幻觉。

5.2.3 语音语言模型: 语音理解中的序列建模

与视觉语言模型相对应, **语音语言模型**把语音表示接入大语言模型, 以完成语音识别、语音理解、语音问答与语音对话等任务, 其结构大体遵循 5.2.1 的通用框架。二者的主要差异在于数据结构: 图像更多体现空间结构, 语音更多体现时间结构, 且语音中常同时包含语言内容、说话人信息、情感与噪声。因此语音语言模型的一个核心, 是把连续声学信号压缩、映射为大语言模型可处理的语义序列。

从语音识别到语音理解。 传统语音识别主要学习

$$p(\text{转写文本} \mid \text{语音}),$$

目标是准确回答“语音中说了什么”。语音语言模型则进一步结合文本指令，学习

$$p(\text{回答} | \text{语音}, \text{指令}),$$

因此输出不必是逐字转写，也可以是问题答案、翻译结果或对话回复。二者的区别如图 5.23 所示。

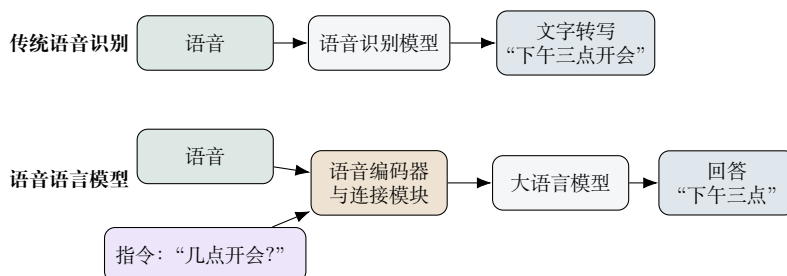


图 5.23: 从语音识别到语音语言模型：传统模型主要输出逐字转写，语音语言模型则结合语音和指令完成问答、翻译与对话等任务。

举例来说，给定一段语音“几点开会？”，传统语音识别只输出其逐字转写“几点开会”，即建模 $p(\text{转写文本} | \text{语音})$ ；而语音语言模型可结合指令直接作答，例如根据日程回答“下午三点”，即建模 $p(\text{回答} | \text{语音}, \text{指令})$ 。同一段语音，前者是“转写”，后者是“理解并回应”。

语音如何进入语言模型。 常见方法有两类，如图 5.24 所示。

1. **连续特征接入**：语音编码器输出连续声学表示，再由连接模块进行压缩和投影，使其维度与文本词向量一致。Qwen-Audio[26] 和 SALMONN[27] 等模型主要采用这种结构。
2. **离散 token 接入**：先将语音量化为有限集合中的离散单元，再把这些单元加入语言模型词表，使同一个模型能够统一处理文本 token 和语音 token。SpeechGPT[25] 采用了这种思路。

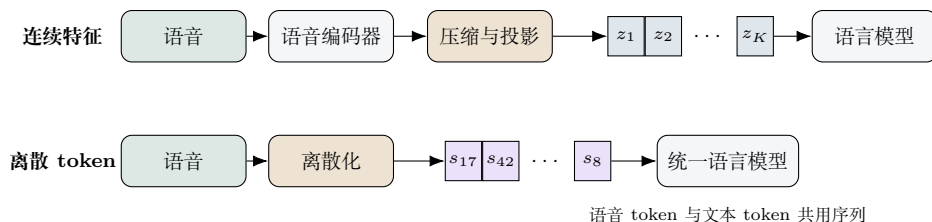


图 5.24: 语音 token 接入语言模型的两种方式。上：将连续声学特征压缩并投影到词向量空间；下：将语音离散化，与文本 token 统一建模。

语音与文本 token 的对齐。 连续接入方法通过连接模块，使语音表示落入语言模型能够读取的词向量空间；离散接入方法则让语音 token 与文本 token 使用同一个自回归模型。训练时，一段语音通常与转写、翻译或回答配对，模型根据语音和指令预测目标文本：

$$\mathcal{L}_{\text{speech}} = - \sum_{l=1}^L \log p(y_l | y_{<l}, \mathbf{z}_{1:K}, \mathbf{q}), \quad (5.7)$$

其中 $\mathbf{z}_{1:K}$ 为语音表示， $\mathbf{q}$ 为文本指令， y_l 为第 l 个目标文本 token。模型必须从语音中提取与答案有关的内容，才能正确预测后续文本，从而逐渐建立语音与文本语义之间的对应关系。

训练能力的逐步扩展。 语音语言模型通常经过以下三个阶段：

1. **语音-文本训练：**利用语音与转写或翻译数据，先学习语音内容与文字之间的对应关系。
2. **语音指令微调：**加入“语音+问题+回答”数据，使模型从简单转写扩展到问答、翻译、摘要和信息提取。
3. **语音对话训练：**使用多轮对话数据，使模型结合前文，持续理解用户语音并生成连贯回复。

Whisper[24] 是大规模多语言语音识别与语音翻译的代表，说明了大规模语音-文本训练可以显著增强跨语言和跨场景的识别能力。在此基础上，SpeechGPT[25] 进一步使用离散语音表示，使语言模型能够接收和生成语音；Qwen-Audio[26] 将训练范围扩展到语音、音乐和环境声等多种音频任务；SALMONN[27] 则把语音编码器、通用音频编码器与语言模型结合，用于语音识别、翻译、音频问答和情感识别等任务。

现实语音中的困难。 真实语音比书面文本更加复杂，主要体现在以下几方面：

- **口语化与方言：**语音中常有重复、停顿、口头语和不完整句子，不同地区的口音与方言也可能在训练数据中分布不均。
- **噪声与混响：**交通声、音乐和远距离录音会遮盖发音内容，使声学表示不稳定。
- **多人重叠：**多人同时说话时，模型还需判断“谁说了什么”，往往需要结合语音分离或说话人分段。
- **长语音与实时性：**语音 token 序列很长，过度压缩可能丢失细节，而保留全部 token 又会增加计算量和回答延迟。

重点提示

语音理解并不完全等于“先转写，再让语言模型回答”。逐字转写主要保留“说了什么”，却可能丢失语气、情绪、停顿、说话人和环境声音。因此，语音语言模型通常仍需直接读取声学表示，而不能只依赖转写文本。

5.2.4 多模态指令微调与对齐

模型具备图像或语音输入能力后，一般仍需通过**指令微调**学会理解用户意图并按规定格式作答，再通过**偏好对齐**使输出更符合人类偏好与安全要求。指令微调、RLHF与DPO等方法本身已在第三部分讨论，本小节只说明其在多模态情形下的特殊之处。

其特殊性主要体现在两方面。其一，**指令数据须包含非文本模态**：多模态指令数据通常由“图像/语音+指令+目标回答”构成，如图像问答、图像描述、图文推理等；其构造成本较高，实践中常借助已有的强文本模型，依据图像的文字标注自动生成大量问答样本（LLaVA即采用了类似做法）。其二，**对齐须兼顾模态一致性**：除一般的有用性与安全性外，多模态对齐往往还需抑制前述视觉幻觉，使模型的回答尽量忠实于实际输入的图像或语音，而非依赖语言先验作出臆测。

语音侧的指令数据。 语音指令数据通常由

语音输入 + 文本指令 + 目标回答

组成。它不再只要求模型逐字转写语音，而是要求模型根据语音内容完成指定任务。常见数据形式如图 5.25 所示。

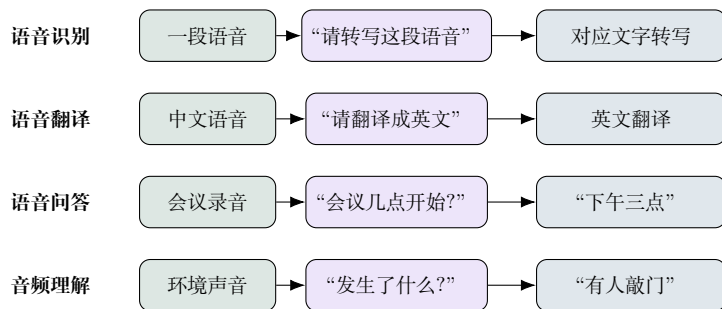


图 5.25: 语音侧指令数据的常见形式：同一模型根据不同指令，对语音完成转写、翻译、问答或音频理解。

已有的语音识别和语音翻译数据可直接改写为指令格式。例如，将“语音-转写”样本包装为“请转写这段语音-目标转写”，将“语音-翻译”样本包装为“请翻译成某种语言-目标译文”。对于语音问答和摘要，还可以先利用语音转写、场景标签或人工标注，再由较强的文本模型生成问题与参考答案。SpeechGPT[25]、Qwen-Audio[26]和SALMONN[27]等工作均通过混合多种语音和音频任务，使模型能够根据指令选择不同的处理方式。

5.2.5 图像—语音—文本联合建模

现实信息往往多模态并存：视频同时包含画面、声音与字幕；课堂同时包含板书、讲解与文字资料。**联合建模**要求模型在同一上下文中同时处理图像、语音与文本，分别对应空间结构、时间结构与符号结构，并在统一表示中完成跨模态的理解与生成。其主要挑战，在于协调这几种异质结构，尽量保持空间、时间与语义上的一致。从双模态系统推广到三模态及以下的统一系统，是近年多模态模型的一个发展方向，一些较新的模型已尝试在统一框架内原生支持多种模态的输入与输出。

以语音/音频为核心的联合场景。 在视频、课堂和会议等场景中，语音或音频通常沿时间连续展开，可作为组织其他模态的**时间轴**。模型需要把同一时刻附近的画面、声音和字幕组合起来，才能完整理解正在发生的事件。

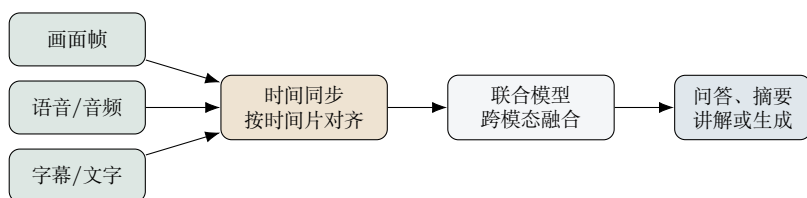


图 5.26: 以语音或音频时间轴为核心的联合建模：先对齐同一时间片内的画面、声音与字幕，再进行统一理解与生成。

视频理解。画面提供人物、物体和动作，语音提供说话内容，环境声音则可提示画面外发生的事件。例如，仅凭画面可能看不出门外有人，而敲门声能够补充这一信息；字幕还可以帮助模型理解口音较重或受噪声干扰的语音。

看图配音讲解。模型也可以同时接收图像和文本指令，再生成相应的语音讲解。例如，根据一幅医学图像生成口头说明，或根据课件内容生成教学讲解。此时不仅要保证语音内容与图像一致，还要控制语速、停顿和语气，使生成结果适合具体场景。

长上下文中的时间同步。对于长视频、课堂录音或多人会议，模型通常先将输入切分为若干时间片，形成

$$(V_1, A_1, T_1), (V_2, A_2, T_2), \dots, (V_L, A_L, T_L),$$

其中 V_l 、 A_l 和 T_l 分别表示第 l 个时间片内的画面、音频和文字。模型需要保留这些时间片的先后顺序，并正确判断一句话、一个动作和一种声音是否发生在同一时刻。若时间对齐错误，就可能把后面的声音解释为前面的画面，从而产生错误的事件描述。

5.3 条件生成建模与多模态内容生成

本节讨论多模态内容的生成。文本生成图像、图像编辑、语音合成、文本生成音频等任务表面各异，但大体可统一表述为同一个数学问题：在给定条件 c 下，学习目

标内容 x 的条件分布并从中采样,

$$\text{学习 } p(x | c) \text{ 并从中采样得到新的 } x. \quad (5.8)$$

其中条件 c 可为文本、图像、语音或指令, 目标 x 可为图像、语音或音频。本节先建立条件生成的基本概念, 再依次讨论文本生成图像、图像编辑、语音生成、音频生成与跨模态交互生成。阅读每一种任务时, 建议对照式 (5.8), 明确其中的 c 与 x 各指什么。

5.3.1 条件生成模型基础

概率分布与采样。 理解生成模型, 先要明确两个基本概念。**概率分布**是对一个随机对象所有可能取值给出概率(离散情形)或概率密度(连续情形)的描述。以图像为例, 可设想在“所有可能图像”构成的高维空间上存在一个分布 $p(x)$: 真实、自然的图像所在区域概率较高, 杂乱无意义的像素组合概率较低。**采样**则是按某分布抽取一个具体取值。需要强调, 机器学习中所说的“生成”, 大体上可以理解为从一个分布中采样。

这一概念对文本与图像是相通的。语言模型在生成时, 每一步给出“下一个词”在词表上的概率分布, 再据此抽取一个词, 逐步连成句子——这就是**采样文字**。图像生成则是在图像分布上抽取一个高维向量(即一幅图像)——这就是**采样向量**。无论目标是离散的词还是连续的图像向量, 生成大体都遵循“先有分布、再采样”的模式。

判别模型与生成模型。 据此可区分两类模型。**判别模型**学习的是条件 $p(y | x)$, 即由输入预测标签, 关注类别之间的分界; 先前多数任务属此类。**生成模型**学习的是数据本身的分布 $p(x)$, 学成之后即可采样出新的、训练集中未必出现过的样本。**条件生成模型**进一步引入条件 c , 学习 $p(x | c)$, 从而能按指定条件采样, 例如要求“生成一只猫”而非任意图像。在多模态生成中, 条件 c 正是连接不同模态的纽带: 文本可作为图像生成的条件, 原图与指令可作为图像编辑的条件。

常见生成建模方法。 如何建模并从复杂分布中采样, 历史上发展出若干方法(表 5.1): 生成对抗网络 [29]、变分自编码器 [30]、扩散模型 [31], 以及第四部分已讨论的基于流的生成模型。本节将以扩散模型为主线, 因其近年在图像、语音、音频生成中应用较广。各方法的数学细节不在此展开; 它们的共同点在于都要回答同一个问题: 如何建模数据分布并从中采样。

表 5.1: 常见生成建模方法（仅列核心思想）

方法	核心思想	代表
自回归	按顺序逐元素建模 $p(x) = \prod_t p(x_t x_{<t})$ ，逐步采样	语言模型、图像自回归
变分自编码器	引入隐变量，优化重建与分布约束的折中	VAE
生成对抗网络	生成器与判别器对抗，提升样本逼真度	GAN
扩散模型	正向逐步加噪、反向逐步去噪并采样	DDPM
基于流的模型	以可逆变换在简单分布与数据分布间建立映射	Normalizing Flow

5.3.2 文本生成图像

文本生成图像建模的是 $p(\text{图像} | \text{文本})$ ：给定一段文字描述，采样出与之语义较为一致的图像。这一方向经历了较明显的方法演变：早期曾用生成对抗网络直接由文本生成图像；DALL·E[33] 将图像离散为 token 后以自回归方式生成；此后扩散模型逐渐成为较主流的一类方法，GLIDE[34]、DALL·E 2[35]、Imagen[36] 在开放域文本生成图像上取得了较好效果，而潜在扩散模型 [37]（即 Stable Diffusion 的基础）通过在低维潜在空间中做扩散，显著降低了计算成本。下面以扩散模型为例，较完整地说明其训练与采样过程，并解释它如何实现式 (5.8) 所要求的“建模分布并采样”。

基本思想：加噪与去噪。 第四章已从连续时间随机微分方程的角度介绍了扩散模型的正向扰动与反向生成过程；本节沿用一致的记号（以 $\mathbf{x}$ 记图像数据、 p_{data} 记数据分布、 $\mathcal{N}(\mathbf{0}, \mathbf{I})$ 记先验分布），并从条件生成的角度，说明文本如何作为条件引导反向过程生成图像。为便于直观理解，这里采用离散时间步的表述，其思想与第四章的连续时间视角一致。扩散模型包含两个方向相反的过程（图 5.27）。**正向过程**对一幅真实图像 $\mathbf{x}_0 \sim p_{\text{data}}$ 逐步加入高斯噪声，经过 T 步后得到接近纯噪声的 $\mathbf{x}_T$ ；这一过程是预先设定的，无需学习。**反向过程**则需要学习：训练一个神经网络，使其能从带噪图像中估计所含噪声，从而每步去除一部分噪声。学会“去噪”之后，生成时便从先验分布 $\mathcal{N}(\mathbf{0}, \mathbf{I})$ 采样出发，反复去噪，逐步还原出一幅图像。

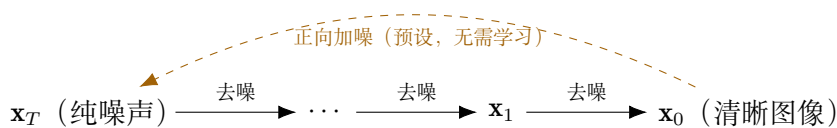


图 5.27: 扩散模型：正向逐步加噪（虚线），反向逐步去噪并采样（实线）。生成从纯噪声开始，沿反向链得到图像。

训练目标。 正向过程加噪到第 t 步，可一次写出

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad (5.9)$$

其中系数 $\bar{\alpha}_t$ 随 t 增大而减小，表示越靠后原图成分越少、噪声成分越多。训练一个网络 $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{c})$ ，令其根据带噪图 $\mathbf{x}_t$ 、步数 t 与条件 $\mathbf{c}$ 估计所加噪声 $\boldsymbol{\epsilon}$ ，目标是估计得尽量准确：

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\mathbf{x}_0, \boldsymbol{\epsilon}, t} \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{c}) \right\|_2^2. \quad (5.10)$$

此处的条件 $\mathbf{c}$ 即文本：文字经文本编码器（常用 5.1.3 中与图像对齐过的文本表示）转为向量后注入网络，使每一步去噪都依赖于文本，从而引导生成朝符合描述的方向进行。

如何采样：从噪声到图像。 训练完成后，可按如下方式采样，对应式 (5.8) 的“从 $p(\mathbf{x} | \mathbf{c})$ 采样”。先从先验分布采样初值 $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ ；随后令 t 从 T 递减到 1，反复执行

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{c}) \right) + \sigma_t \mathbf{z}, \quad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad (5.11)$$

即每一步先用网络估计并去除一部分噪声，再注入少量随机噪声 $\sigma_t \mathbf{z}$ 以维持采样的随机性；迭代至 $t = 1$ 得到的 $\mathbf{x}_0$ 即为生成的图像。可以看出，这条反向链条实际上定义了一个以 $\mathbf{c}$ 为条件、可从中采样的分布 $p_\theta(\mathbf{x}_0 | \mathbf{c})$ ，它对应我们要建模的 $p(\mathbf{x} | \mathbf{c})$ 。由于每次采样的初值不同，同一段文字往往可生成多样的结果。

直观地看，这条采样链条就是让图像“从噪声中逐步显影”：起初画面几乎全是噪点，随着式 (5.11) 的去噪步骤反复推进，轮廓、颜色与细节逐渐浮现，最终得到清晰图像。

增强文本一致性：无分类器引导。 为加强生成对文本的服从程度，实践中常采用无分类器引导 [32]：在每一步同时计算“给定条件”与“不给条件”两种噪声估计，并放大二者之差，

$$\tilde{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t, \mathbf{c}) = \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \emptyset) + w (\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{c}) - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \emptyset)), \quad (5.12)$$

其中 $\emptyset$ 表示不提供条件，引导强度 $w > 1$ 。括号内是“顾及文本”相对“不顾文本”所多出的方向，乘以 w 即放大对文本的服从。一般而言， w 越大，结果越贴合文本，但样本多样性可能下降，需要权衡。

举一个完整的例子：给定文本提示词“一只戴着红色帽子的柯基犬坐在草地上，阳光明媚”，模型从纯噪声出发、在该文本条件下逐步去噪，生成与描述语义一致的图像；改变提示词中的属性（如帽子颜色、背景、画风），生成结果会相应变化；而提高无分类器引导强度 w ，结果会更贴合文字，但多样性随之下降。

5.3.3 图像编辑：约束条件下的生成

文本生成图像是从噪声生成全新图像，**图像编辑**则要在了一幅已有图像上做修改，因而多出一项关键约束：在引入目标改动的同时，尽量保持原图的主体、结构或风格不变。其建模形式可写为

$$p(\text{编辑后图像} \mid \text{原图}, \text{编辑指令}, \text{掩码或结构条件}). \quad (5.13)$$

例如把背景替换为雪山时，模型需改变背景，同时尽量保持人物的面部、姿态与主体结构。图像编辑的一个核心，正是在“目标改动”与“原图保持”之间取得平衡。其一般形式如图 5.28 所示：以原图和某种**编辑条件**为输入，交由生成模型（多为扩散模型）产生结果。

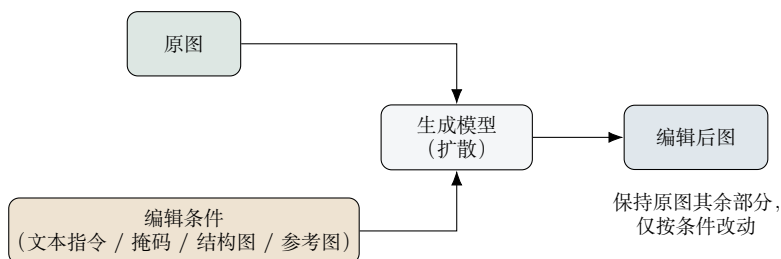


图 5.28: 图像编辑的一般形式：以原图与编辑条件为输入，生成在保持原图其余部分的同时按条件改动的结果。不同代表工作的差异，主要在于“编辑条件”的形式。

对照图 5.28，代表工作的区别主要在于图中“编辑条件”取何种形式，据此可分为以下几类：

- **图像修复与扩展**：在指定的缺失或外延区域生成内容，由掩码约束“仅在此处生成”。
- **文本引导编辑**：以自然语言描述目标，在尽量保留原图的前提下实现该描述，如 SDEdit[38] 与 Prompt-to-Prompt[39]。
- **指令式编辑**：直接给出编辑命令（如“将天空改为黄昏”），模型据此修改，如 InstructPix2Pix[40]。
- **结构控制与身份保持**：以边缘图、深度图、人体姿态等作为额外约束以固定构图，如 ControlNet[41]；或借助少量参考图保持特定对象的身份，如 DreamBooth[47]。

以结构控制为例：当希望生成图像严格遵循某种构图时，可提供一张**结构条件图**——例如人体姿态骨架、线稿或深度图——与文本提示一起输入模型，模型便在保持该结构的前提下按文本生成内容。给定同一姿态骨架，配以不同文本（“穿红裙的女孩”“穿盔甲的战士”），即可得到姿态一致而外观不同的图像。

重点提示

评价图像编辑一般不能只看新内容是否真实，还需考察三点：是否较准确地实现了目标改动；不应改动之处是否得到保持；局部修改是否与整体较为协调。这三者之间的平衡，是图像编辑较之纯生成更为困难之处。

再以指令式编辑为例：给定一张“人物站在城市街道前”的原图与编辑指令“把背景换成雪山”，模型在保持人物面部、姿态与主体结构不变的前提下，仅改动背景区域；若把指令改为“将白天改为黄昏”，则整体光照随之改变，而主体保持一致。

5.3.4 语音生成：连续信号的条件生成

将式 (5.8) 中的目标 x 取为语音、条件 c 取为文本及说话人、情感等，即得到**语音合成**：

$$p(\text{语音} \mid \text{文本}, \text{说话人}, \text{风格}, \text{情感}). \quad (5.14)$$

它与图像生成的一个根本差异在于：图像多为空间结构的一次性生成，语音则是随时间展开的连续序列生成。较好的语音合成不仅要准确传达文本内容，通常还需生成较自然的音色、韵律、语速与停顿。

文本到语音的基本流程。 典型语音合成系统包含三个部分，如图 5.29 所示。首先，**文本前端**将文字规范化，并转换为字符、音素等语言单元；随后，**声学模型**根据文本预测梅尔频谱或声学 token；最后，**声码器**将这些中间表示还原为可播放的语音波形。

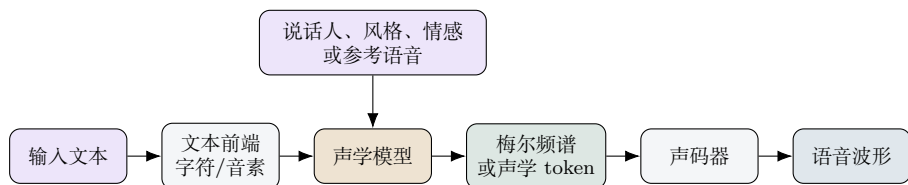


图 5.29: 文本到语音的基本流程。声学模型根据文本和说话人、风格等条件生成声学表示，再由声码器转换为连续语音波形。

举例来说，给定文本“欢迎来到多模态课程”与一段参考语音（指定音色），模型先由文本前端与声学模型生成梅尔频谱，再由声码器还原为语音波形，输出以参考说话人的音色朗读该文本；此时改变情感标签（如“高兴/平静”）或参考语音，便可在内容不变的前提下改变语气与音色，这正体现了语音生成的“一对多”特性。

音色与韵律控制。 同一句文本可以对应多种正确读法，因此语音生成是一种“一对多”任务。说话人条件决定声音像谁，音高、时长和能量共同影响语速、重音、停顿和语气；情感标签则可控制高兴、悲伤或平静等表达方式。模型需要同时满足三项基本要求：内容易懂、声音自然，并符合指定的说话人和表达风格。

零样本合成与声音克隆。 多说话人模型可以从一段参考语音中提取说话人的音色和风格，再使用该声音朗读新的文本。若模型能够为训练中未出现的说话人直接生成语音，

而不需要重新训练，便称为**零样本语音合成**。参考语音提供“用谁的声音说”，输入文本提供“说什么”，二者应尽量相互独立。

连续时间信号的特点。图像生成主要组织二维空间中的像素或图像 token，而语音生成必须沿时间轴保持顺序。一个字或音素通常对应若干连续声学帧，因此模型还需学习文本与语音之间近似单调的时间对应关系。对齐错误可能造成漏词、重复、异常停顿或语速不稳定；生成时间越长，还越难保持音色和韵律的一致。

代表工作。 Tacotron 2[42] 采用“文本到梅尔频谱，再由 WaveNet 声码器生成波形”的两阶段结构；FastSpeech 2[43] 使用非自回归生成，并显式建模时长、音高和能量，提高了速度与可控性。VITS[44] 通过变分推断、流模型和对抗训练，将声学建模与波形生成进行端到端联合训练。VALL-E[45] 则将语音表示为离散神经音频编码 token，像语言模型预测文本 token 一样预测语音 token，并可根据较短的参考语音进行零样本合成。

重点提示

声音克隆的安全与伦理。声音具有明显的个人身份属性，未经许可复制他人声音，可能被用于冒充、虚假信息和诈骗。因此，实际系统应确认声音授权，明确标识合成内容，限制高风险人物或场景的克隆，并结合音频水印、来源记录和伪造语音检测等措施降低滥用风险 [46]。

5.3.5 音频生成：声音事件与声景的条件生成

除语音外，还有更广泛的声音：环境声、音乐片段、声音事件与复杂声景。令模型据文字描述生成此类声音，即一般**音频生成**，建模 $p(\text{音频} | \text{文本描述})$ 。它与语音合成都需生成连续声音信号，但语音有较明确的语言内容与发音规则，一般音频则更强调声音事件、环境层次与时间上的变化，且常有多个声源叠加，建模难度往往更高。

文本到音频的基本流程。 文本到音频生成不是朗读文字，而是根据文字描述还原相应的声音。例如，输入“雨夜的街道上有汽车经过”，模型需要生成雨声、车辆声及相应的空间环境。其基本流程如图 5.30 所示：文字描述先被编码为条件表示，生成模型据此产生频谱、潜在表示或离散音频 token，最后再还原为音频波形。

声音事件与声景生成。 较简单的目标是生成单个**声音事件**，如狗叫、玻璃破碎或汽车鸣笛；更复杂的**声景**则包含多个同时或依次出现的声源。例如，“雨声中，远处有汽车驶过，随后传来雷声”同时规定了声音种类、背景环境和发生顺序。模型不仅要生成每种声音，还要确定它们何时开始、持续多久以及如何相互叠加：先是持续的雨声，中间叠入由远及近的车声，最后才出现雷声。与语音合成不同，这里并没有文字发音，重点在于声音事件的种类、层次与时间结构。

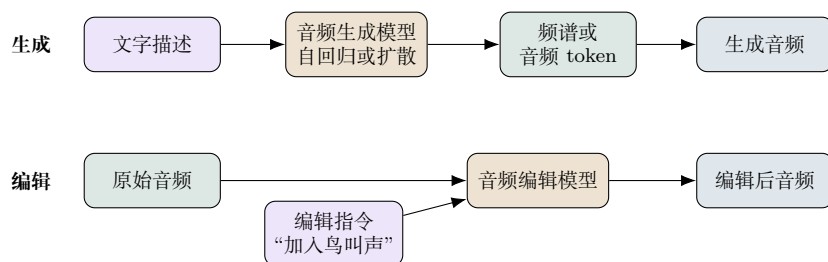


图 5.30: 文本引导的音频生成与编辑。生成模型从文字描述产生新的音频；编辑模型则根据文字指令修改已有音频。

因此，音频生成通常需要满足三个要求：声音内容与文字描述一致，音质具有较强的真实感，并且多个事件在时间上的顺序正确。同一段文字也可能对应许多合理声音，因此模型学习的不是唯一答案，而是符合描述的声音分布。

音频上的扩散建模。 扩散模型训练时逐步向真实音频表示加入噪声，生成时则从随机噪声出发，在文字条件的引导下反复去噪：

$$\mathbf{z}_T \longrightarrow \mathbf{z}_{T-1} \longrightarrow \cdots \longrightarrow \mathbf{z}_0 \longrightarrow \text{音频}. \quad (5.15)$$

为降低计算量，模型通常不直接在高采样率波形上扩散，而是先将音频压缩为频谱或低维潜在表示，在潜在空间完成去噪，再通过解码器和声码器恢复波形。文字表示在每一步去噪中提供条件，使最终声音逐渐接近描述中的事件和环境。

文本引导的音频编辑。 音频生成还可扩展为对已有声音的修改。模型同时接收原始音频和编辑指令，完成“加入雨声”“删除狗叫”“将钢琴替换为吉他”或“修复缺失片段”等操作。编辑时既要改变指定内容，又应尽量保留无关部分，避免整段音频的音色、背景或节奏发生不必要的变化。

代表工作。 AudioGen[48] 将音频量化为离散 token，再使用自回归模型根据文本逐步预测声音；AudioLDM[49] 和 Make-An-Audio[50] 则将扩散模型应用于压缩后的音频表示，提高生成效率与音质。Make-An-Audio 2[51] 进一步强调声音事件的先后顺序和长音频中的时间一致性。AUDIT[52] 则使用“原始音频 + 编辑指令 + 目标音频”数据训练潜在扩散模型，可执行声音添加、删除、替换和片段修复等编辑任务。

总体而言，音频生成是上一小节语音合成从“人类说话声”向一般声音的推广。语音合成重点保证文字发音和说话人特征正确，而一般音频生成还需组织多个声源、背景环境与时间变化，生成具有完整声景结构的连续声音。

5.3.6 跨模态交互生成

前述任务多为单一条件、单向生成。更接近实际应用的，是**多条件、多模态、多轮**的交互生成：

$$p(\text{目标模态} \mid \text{文本, 图像, 语音, 历史上下文, 用户指令}). \quad (5.16)$$

例如：依据图像生成语音讲解；依据语音指令修改图像；依据图像与文本共同生成配套音效；或在多轮对话中不断修订生成结果。其主要难点在于不同模态的结构不一致——图像较重空间、语音/音频较重时间、文本较重逻辑——模型需在生成过程中尽量同时保持语义对应、空间一致、时间同步与用户意图一致。这与本讲前两节相呼应：往往需先具备跨模态对齐与多模态理解能力，才能实现较为可控的多模态生成。近年也出现了一些尝试在统一框架内支持“任意模态到任意模态”生成的工作，如 CoDi[53]。

以语音/音频为目标或条件的交互生成。 语音和音频既可以作为模型需要生成的目标，也可以作为理解用户意图的**条件**，如图 5.31 所示。

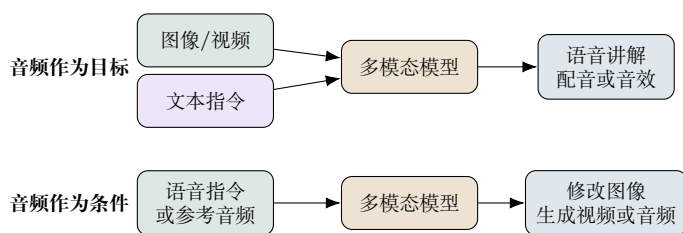


图 5.31: 语音和音频参与交互生成的两种形式：上，依据图像、视频和文字生成语音或音效；下，将语音指令或参考音频作为条件控制其他模态的生成。

视频配音与看图讲解。模型可以根据画面内容和文本要求生成语音讲解。例如，输入一段教学视频，模型需要说明画面中的步骤；输入人物视频时，则可生成与人物动作和场景匹配的配音。此时不仅要保证“说的内容”与画面一致，还要使停顿、语速和音效出现的时间与视频同步。

语音驱动修改。用户也可以直接用语音提出修改要求，如“把背景音乐调小一点”、“在开门时加入脚步声”或“用更平静的语气重新讲一遍”。模型需要结合当前生成结果和历史指令，只修改用户指定的部分，并尽量保留其他内容。

多轮交互可写为

$$Y_t \sim p(Y_t \mid I, A, Q_{\leq t}, Y_{< t}), \quad (5.17)$$

其中 I 表示图像或视频， A 表示参考语音或音频， $Q_{\leq t}$ 表示截至当前轮的用户指令， $Y_{< t}$ 表示此前生成结果。模型每一轮都需根据新指令修订结果，而不是完全重新生成。

一致性与可控性。这类系统通常需要同时考察以下几方面：

- **语义一致性：**生成的语音或音效是否与图像、视频和文字描述相符；

- **时间同步**：声音事件、动作与字幕是否在正确时间出现；
- **指令遵循**：模型是否只修改用户指定的内容；
- **声音质量**：语音是否自然、易懂，音频是否真实且无明显杂音；
- **跨轮保持**：多轮修改后，人物、音色、场景和未编辑内容是否保持一致。

CoDi[53] 等统一多模态生成模型尝试将文本、图像、视频和音频纳入同一框架，使一个模态或多个模态的组合能够控制另一模态的生成。这类模型的关键不只是“能够生成多种模态”，还在于不同输出之间应描述同一事件，并在空间、时间和语义上保持协调。

重点提示

本节讨论的是较为通用的条件生成能力。要在真实场景中落地，通常还需应对噪声、实时性、安全性与可靠性等更严苛的要求。

5.4 本章小结

本章围绕图像、语音/音频与文本的统一处理，按三个层次构建了多模态机器学习的认知框架。在**表示与对齐**层面，我们说明了各模态在计算机中的原始形式，以及将其编码为向量的动机——用向量空间中的距离刻画语义相似度，使语义相近的对象表示相近；在此基础上，通过自监督预训练获得通用表示，并以对比学习为代表，把不同模态纳入同一空间实现跨模态对齐。

在**结构与理解**层面，我们给出了“模态编码器—连接模块—大语言模型”的通用结构，说明图像、语音如何借此接入大语言模型，构成视觉语言模型与语音语言模型，并通过多模态指令微调与对齐获得按指令处理多模态输入的能力。

在**条件生成**层面，我们把文本生成图像、图像编辑、语音合成与音频生成统一表述为条件分布 $p(\mathbf{x} | \mathbf{c})$ 的学习与采样问题，并以扩散模型为例，说明了如何训练去噪网络、如何从先验噪声逐步采样出图像，以及文本条件如何引导生成。

课堂练习

1. 用一句话说明表示学习所追求的目标，并解释为什么希望做到“语义相近的对象，其向量表示也相近”。
2. 一张 224×224 的 RGB 图像，按 16×16 的图块（patch）不重叠地切分，可得到多少个视觉 token？
3. 判断下列说法是否正确并简要说明理由：“视觉语言模型与文本生成图像建模的是同一个条件分布。”

4. 以 CLIP 为例，简述跨模态对齐用什么样的训练目标，把图像与文本纳入同一个向量空间。
5. 某单声道语音的采样率为 16 kHz、时长为 2 秒，问其波形包含多少个采样点？
6. 在基于扩散模型的文本生成图像中，无分类器引导的强度 w 增大，会对生成结果的“文本一致性”和“多样性”分别产生什么影响？
7. 扩散正向过程为 $\mathbf{x}_t = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon$ 。若 $\bar{\alpha}_t = 0.64$ ，求原图 $\mathbf{x}_0$ 与噪声 ϵ 前的系数各为多少。
8. 视觉语言模型与语音语言模型在整体结构上有何共同点？二者所处理数据的主要差异又是什么？

附录 A 附录

A.1 常用 LaTeX 片段

% 插入普通图片

```
\begin{figure}[htbp]
  \centering
  \includegraphics[width=0.8\linewidth]{figures/your-image.png}
  \caption{图片标题}
\end{figure}
```

% 插入三线表

```
\begin{table}[htbp]
  \centering
  \caption{表格标题}
  \begin{tabular}{lll}
    \toprule
    A & B & C \\
    \midrule
    内容 & 内容 & 内容 \\
    \bottomrule
  \end{tabular}
\end{table}
```

参考资料

这里可以列出教材、论文、网站或其他阅读材料。

参考文献

- [1] Krizhevsky A, Sutskever I, Hinton G E. ImageNet Classification with Deep Convolutional Neural Networks. NeurIPS, 2012.
- [2] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition (VGG). ICLR, 2015.
- [3] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. CVPR, 2016.
- [4] Dosovitskiy A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR, 2021.
- [5] Chen T, et al. A Simple Framework for Contrastive Learning of Visual Representations (SimCLR). ICML, 2020.
- [6] He K, et al. Momentum Contrast for Unsupervised Visual Representation Learning (MoCo). CVPR, 2020.
- [7] Caron M, et al. Emerging Properties in Self-Supervised Vision Transformers (DINO). ICCV, 2021.
- [8] van den Oord A, Li Y, Vinyals O. Representation Learning with Contrastive Predictive Coding (CPC). arXiv:1807.03748, 2018.
- [9] Baevski A, Zhou H, Mohamed A, Auli M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. NeurIPS, 2020.
- [10] Hsu W N, et al. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021.
- [11] Baevski A, et al. data2vec: A General Framework for Self-Supervised Learning in Speech, Vision and Language. ICML, 2022.
- [12] He K, et al. Masked Autoencoders Are Scalable Vision Learners (MAE). CVPR, 2022.
- [13] Bao H, Dong L, Piao S, Wei F. BEiT: BERT Pre-Training of Image Transformers. ICLR, 2022.

- [14] Radford A, et al. Learning Transferable Visual Models From Natural Language Supervision (CLIP). ICML, 2021.
- [15] Jia C, et al. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision (ALIGN). ICML, 2021.
- [16] Zhai X, et al. Sigmoid Loss for Language Image Pre-Training (SigLIP). ICCV, 2023.
- [17] Ouyang S, Ye R, Li L. WACO: Word-Aligned Contrastive Learning for Speech Translation. ACL, 2023.
- [18] Elizalde B, Deshmukh S, Al Ismail M, Wang H. CLAP: Learning Audio Concepts From Natural Language Supervision. ICASSP, 2023.
- [19] Alayrac J B, et al. Flamingo: a Visual Language Model for Few-Shot Learning. NeurIPS, 2022.
- [20] Li J, et al. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. ICML, 2023.
- [21] Liu H, et al. Visual Instruction Tuning (LLaVA). NeurIPS, 2023.
- [22] Dai W, et al. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. NeurIPS, 2023.
- [23] Bai J, et al. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. arXiv:2308.12966, 2023.
- [24] Radford A, et al. Robust Speech Recognition via Large-Scale Weak Supervision (Whisper). ICML, 2023.
- [25] Zhang D, et al. SpeechGPT: Empowering Large Language Models with Intrinsic Cross-Modal Conversational Abilities. Findings of EMNLP, 2023.
- [26] Chu Y, et al. Qwen-Audio: Advancing Universal Audio Understanding via Unified Large-Scale Audio-Language Models. arXiv:2311.07919, 2023.
- [27] Tang C, et al. SALMONN: Towards Generic Hearing Abilities for Large Language Models. ICLR, 2024.
- [28] Sun Q, et al. Emu: Generative Pretraining in Multimodality. ICLR, 2024.
- [29] Goodfellow I, et al. Generative Adversarial Nets. NeurIPS, 2014.
- [30] Kingma D P, Welling M. Auto-Encoding Variational Bayes (VAE). ICLR, 2014.

- [31] Ho J, Jain A, Abbeel P. Denoising Diffusion Probabilistic Models (DDPM). NeurIPS, 2020.
- [32] Ho J, Salimans T. Classifier-Free Diffusion Guidance. arXiv:2207.12598, 2022.
- [33] Ramesh A, et al. Zero-Shot Text-to-Image Generation (DALL · E). ICML, 2021.
- [34] Nichol A, et al. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. ICML, 2022.
- [35] Ramesh A, et al. Hierarchical Text-Conditional Image Generation with CLIP Latents (DALL · E 2). arXiv:2204.06125, 2022.
- [36] Saharia C, et al. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding (Imagen). NeurIPS, 2022.
- [37] Rombach R, et al. High-Resolution Image Synthesis with Latent Diffusion Models (Stable Diffusion). CVPR, 2022.
- [38] Meng C, et al. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. ICLR, 2022.
- [39] Hertz A, et al. Prompt-to-Prompt Image Editing with Cross Attention Control. ICLR, 2023.
- [40] Brooks T, Holynski A, Efros A A. InstructPix2Pix: Learning to Follow Image Editing Instructions. CVPR, 2023.
- [41] Zhang L, Rao A, Agrawala M. Adding Conditional Control to Text-to-Image Diffusion Models (ControlNet). ICCV, 2023.
- [42] Shen J, et al. Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions (Tacotron 2). ICASSP, 2018.
- [43] Ren Y, et al. FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. ICLR, 2021.
- [44] Kim J, Kong J, Son J. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech (VITS). ICML, 2021.
- [45] Wang C, et al. Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers (VALL-E). arXiv:2301.02111, 2023.
- [46] National Institute of Standards and Technology. Reducing Risks Posed by Synthetic Content. NIST AI 100-4, 2024.

- [47] Ruiz N, et al. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. CVPR, 2023.
- [48] Kreuk F, et al. AudioGen: Textually Guided Audio Generation. ICLR, 2023.
- [49] Liu H, et al. AudioLDM: Text-to-Audio Generation with Latent Diffusion Models. ICML, 2023.
- [50] Huang R, et al. Make-An-Audio: Text-To-Audio Generation with Prompt-Enhanced Diffusion Models. ICML, 2023.
- [51] Huang J, et al. Make-An-Audio 2: Temporal-Enhanced Text-to-Audio Generation. arXiv:2305.18474, 2023.
- [52] Wang Y, et al. AUDIT: Audio Editing by Following Instructions with Latent Diffusion Models. NeurIPS, 2023.
- [53] Tang Z, et al. Any-to-Any Generation via Composable Diffusion (CoDi). NeurIPS, 2023.