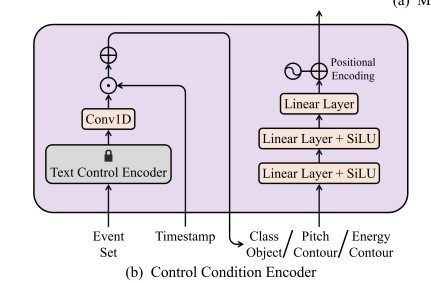
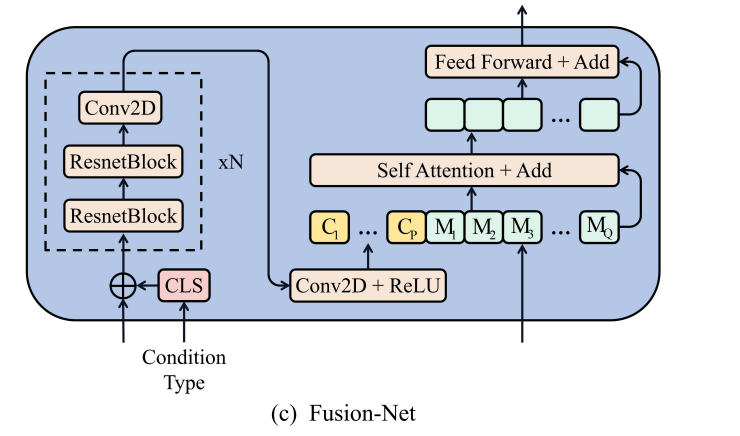
## Audio Generation with Multiple Conditional Diffusion Model



时间顺序控制：利用源自FLAN-T5-LARGE的语言知识来生成事件类的语义表示，并将其与时间戳一起转换为类对象（语义层面编码），强调了图像生成的bbox标签映射令牌中得到类标签和box坐标，对应到音频生成就是时间戳，时间戳编码也是构建一个事件类型X帧数的矩阵，用二进制编码（标签层面编码）

具体而言：FLAN-T5-LARGE将声音事件类别转换为语义表示，经过1\*1卷积层后转化为label embedding，将label embeddeing和 timestamp相乘



为了支持多个控制条件，需要将控制条件标准化到相同的维度。然后利用一个共享的多层感知器对它们进行编码，使用条件类型作为提示来区分融合网络中的不同控制条件

## Soundify: Matching Sound Effects to Video

