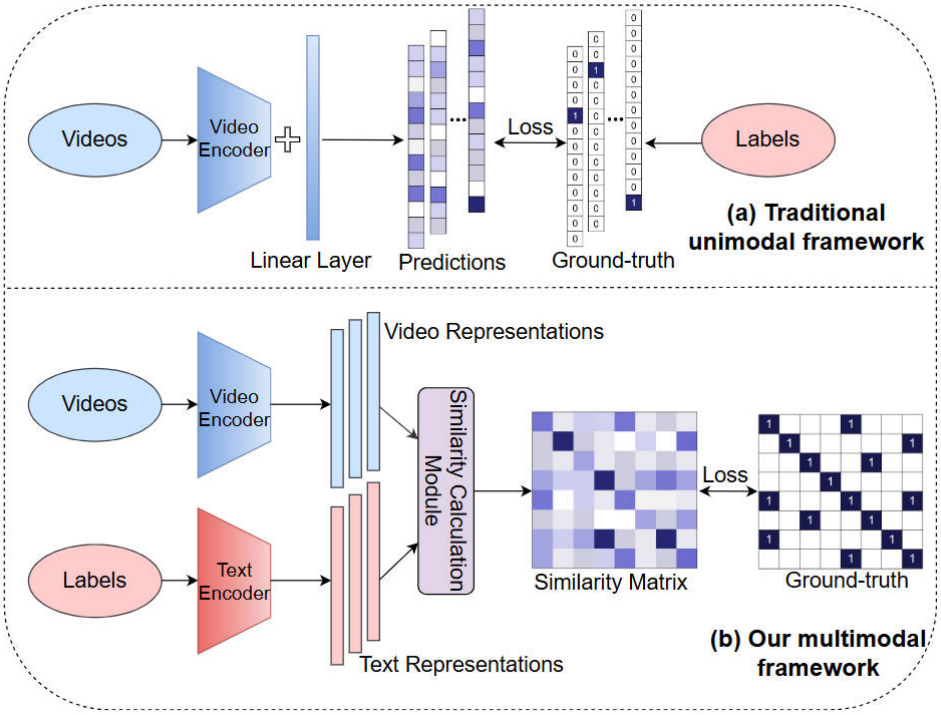
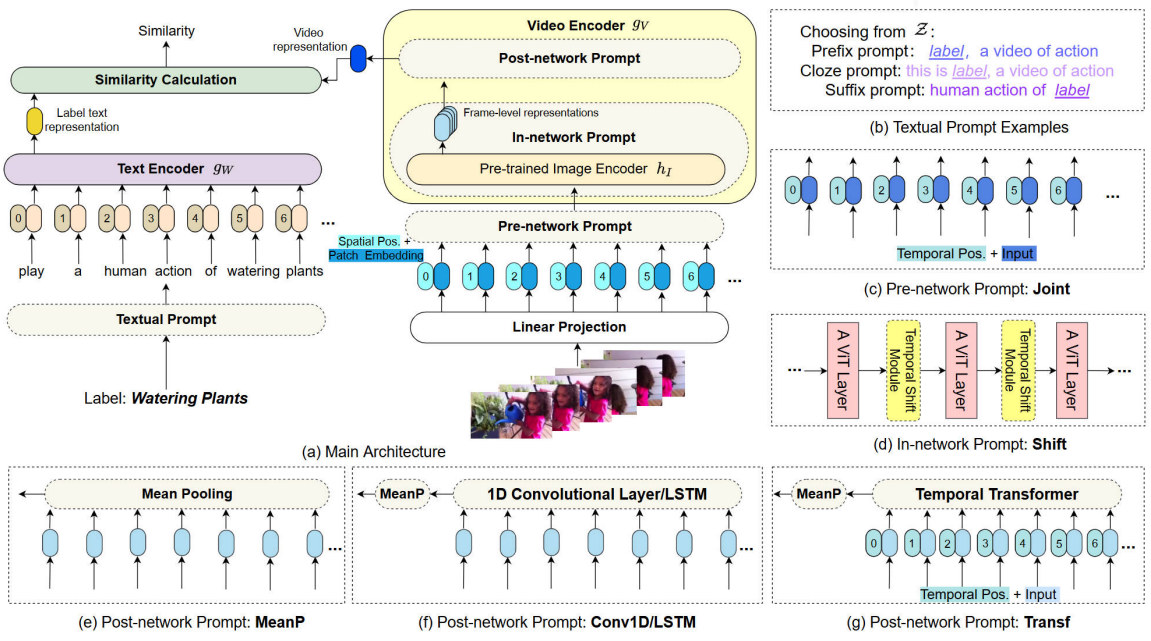
ActionCLIP: A New Paradigm for Video Action Recognition

动机：



动作识别本质是分类，虽然是在视频领域当中，也不妨将CLIP迁移过来，这样一来便不需要再使用带Labels的数据来做监督训练了（图中的上半部分），而可以使用海量的web-data（类似CLIP，又有不同在于，这篇paper训练是supervision-fashion的，同时不是简单地使用cross entropy loss，使用的是KL divergence）

Architecture：



注意：只有文本处理这块中的prompt是“传统”意义上的prompt，而视频处理这块的prompt就不再是是传统“意义”上的prompt了

解释：

Textual Prompt：(b)

分为三类，prefix（前缀） prompt、cloze（完形填空） prompt和suffix（后缀） prompt，与视频处理过程中的三类“prompt”对应

Pre-network Prompt：(c)——Joint

将spatial pos token和temporal pos token合并到一起

In-network Prompt：(d)——Shift

zero-cost、zero-memory，parameter-free，而且效果显著

Post-network Prompt：(e)~(g)

这个和视频文本检索中的CLIP4clip十分相似

实验：

写法：

提一个问题，做一次Ablation实验