RL-Pruner: RL 负责各层剪枝比例, 选择影响最小的剪，最后用原始模型KD 应用于CNN，适合于ResNet、GoogLeNet、MobileNet等

FedSpaLLM: 把联邦学习放在LLMs上，利用l0 norm来聚合重要权重，并且利用考虑资源需求的层采样策略。 使用了非结构剪枝， 数据local，每个client有一个mask，聚合的时候选取都被剪的部分

CipherPrune: 主要是隐私相关，加密token，动态剪枝，确定需要剪枝部分；给重要的部分用高阶多项式，不重要的低阶，主要偏向加密计算方向

Mutual Effort for Efficiency:  利用跨分支的similarity来指导剪枝。 SSL目前需要跨分支信息，传统剪枝会破坏这一部分，所以针对similarity进行剪枝 （token在各个分支之间找到相似token，然后按照跨分支排序，直接按最相似剪掉）

Fast and Effective Weight Update for Pruned Large Language Models： 剪枝逐层处理，通过calibration数据对每层独立剪枝 使用二阶信息 更快收敛

启发：可能能替换wanda / sparseGPT对每层剪枝

Turbo: Informativity-Driven Acceleration Plug-In for Vision-Language Models：定义Information Degree来计算数据的重要性，删除相似token，并且按照语义价值，保留Information Deg最高的 主要是图像上处理 图像理解；图像生成

Pareto Prompt Optimization： Prompt 学习，RL，学习支配/或者非支配损失，找到最好的pareto front, （对应多任务相对最优）提升prompt性能。

启发：如果做压缩，可能可以对压缩之后的提示词找到pareto front

Adaptive Prompt-Tailored Pruning: 关注文生图模型，提出prompt based pruning :将提示词编码之后转换为architecture code，决定模型剪枝方式（Unet/channel/attention head）使用对比学习，相似提示词映射到差不多的剪枝部分

启发 可能可以动态剪枝，对不同prompt用不同层？

LLAMAFLEX：用MLP训练router，实现对剪枝数量以及比例的估算对深度（layer）宽度（attn）剪枝 不用设定剪枝比例，并且使用策略感知补偿剪枝损失，学习ek剪枝比例，通过对scale以及shift调整提升表达能力

启发：策略感知可能可以用