# 23.11.6~23.11.12周报

## 一：本周工作

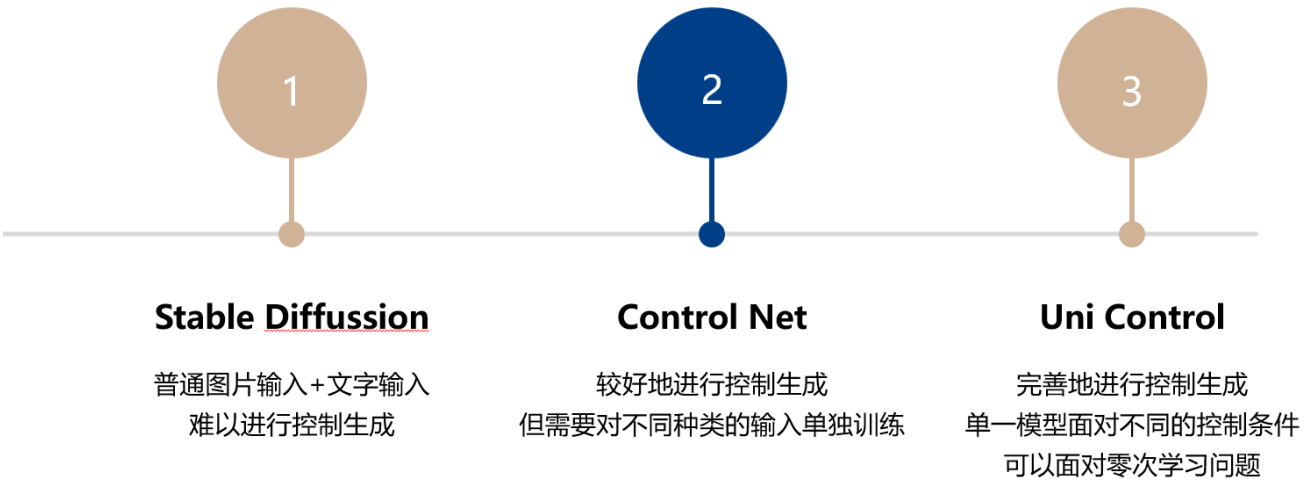
- Uni-control 论文阅读
- 通用大模型微调方式大致调研

## 二：下周计划

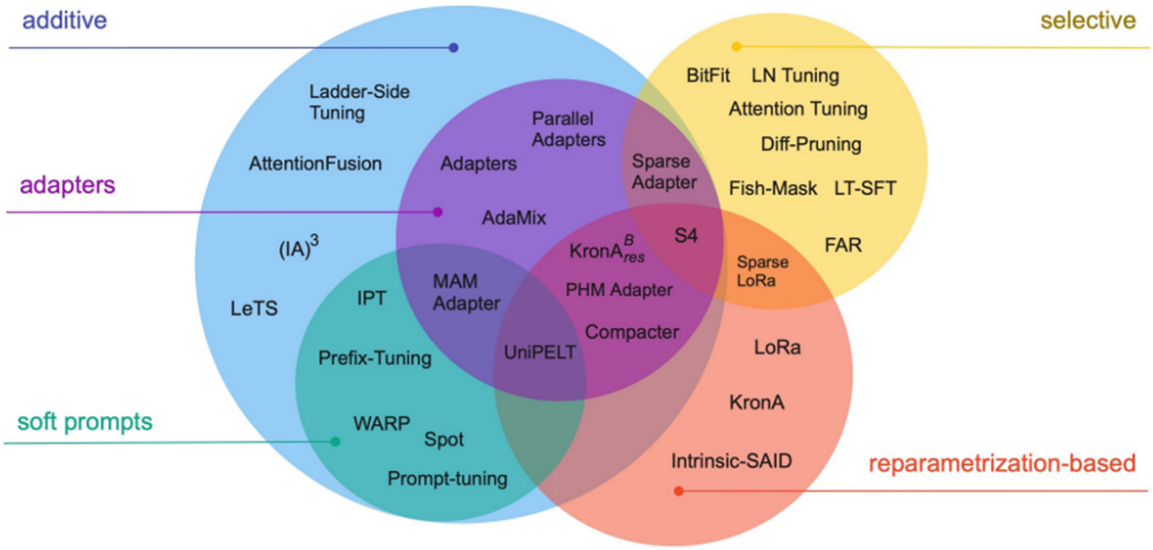
- 浅读sig graph2022年的五篇最佳论文（了解这些工作的意义，突破，但不深究其具体实现方式）
- 对sig graph会议大体方向有一个具体的了解
- 对目前图形学的一些热点方向有一定了解

## 三：结论

- Control Net是基于Stable Difussion模型，使用HyperNet方式进行微调，使模型具有了控制生成的能力。
- Uni-control是基于Control Net的工作，使用MOE将多个模型合并为一个，并提高了生成效果。




- 通用大模型的微调方式主要可以分为四个部分，而且互相之间各有交叉。

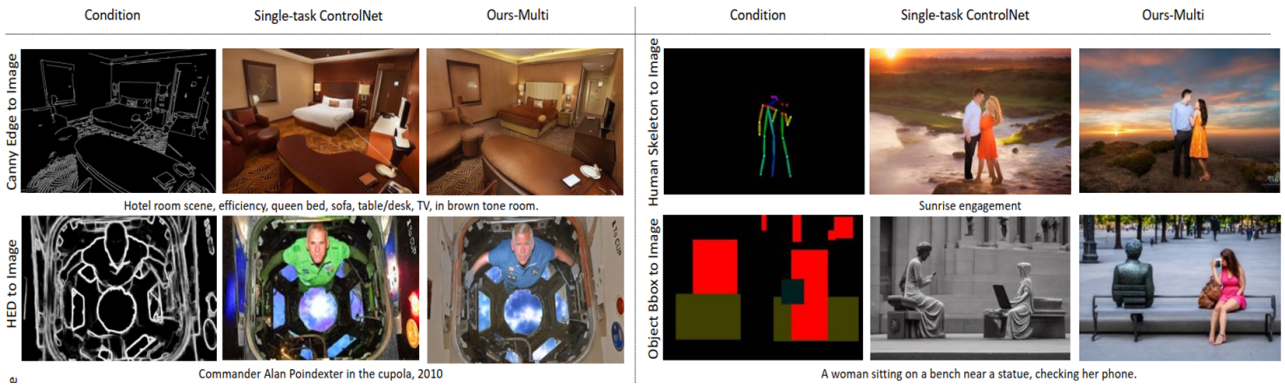



# 四：详细工作

## 4.1 Unicontrol论文介绍

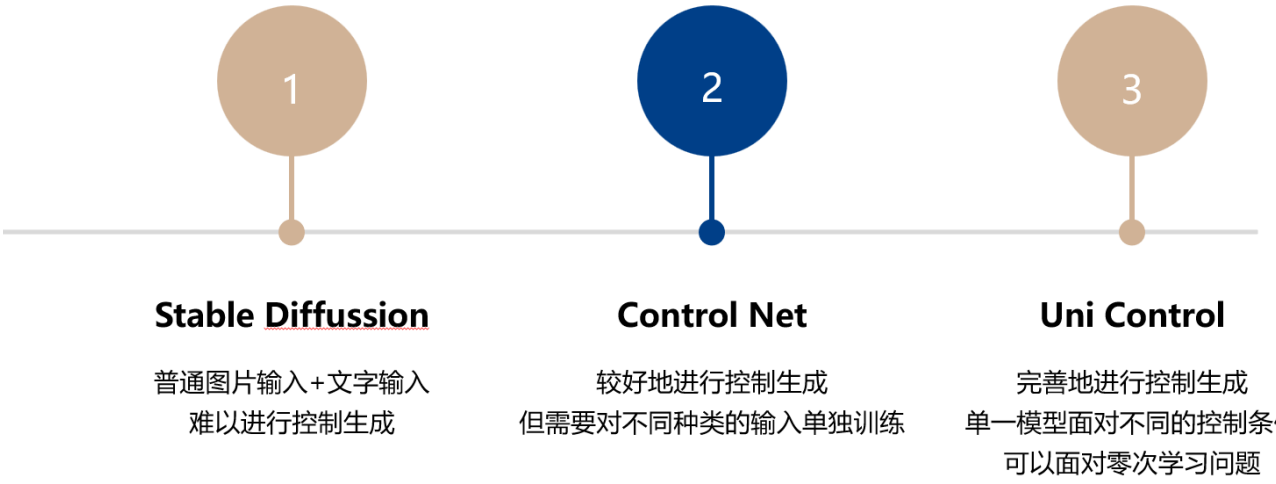
Unicontrol是一篇Arxiv上的预发表论文，主体研究人员来自Salesforce AI Research，这是一个与Open AI类似的人工智能服务提供商。

这项工作的主题是基于Stable Difussion的控制生成。这里的控制生成指的是根据Visual Condition进行图片生成，包括深度图，边界图，切分图，人物姿态等。



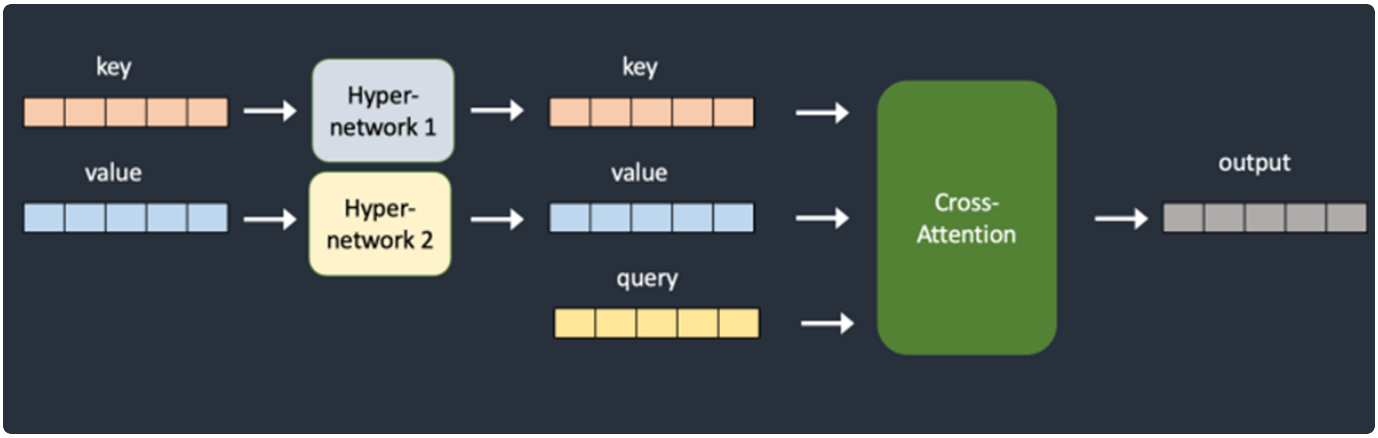

这篇论文的工作很大一部分都是基于另一篇论文ControlNet中的工作。所以我也进一步阅读了这篇论文，其在ICCV会议上发表，主要就是基于Stable Diffusion的模型微调，使其可以基于特定的图像进行控制生成。由于不同种类的输入所蕴含的信息是大不相同的，所以在这篇文章中，不同种类的控制生成分别需要训练独有的模型。

ControlNet还启发了许多其他的论文，例如Uni-ControlNet就与本次所研究的UniControl有着几乎一致的标题和十分类似的工作——都是把ControlNet中的多个模型进行了整合。

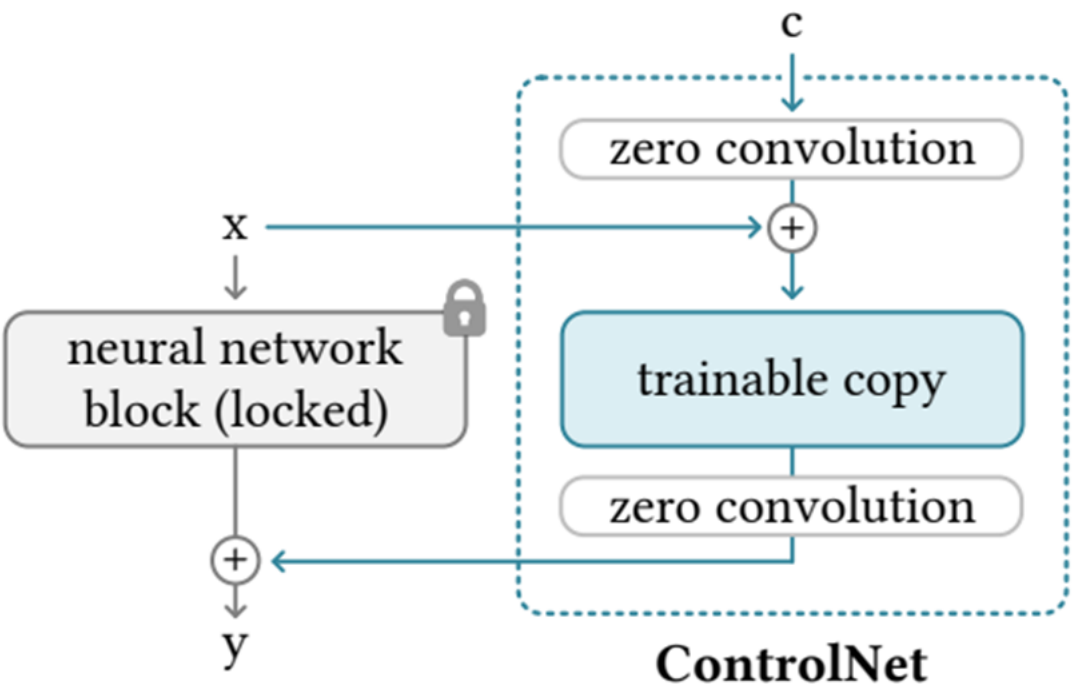



### 4.2 实现方式

HyperNet是由Novel AI开发的SD微调技术（不同于16年的HypereNet模型），主要作用于SD模型的Cross-Attention环节，在不改变原有参数的情况下，通过两个额外的小型神经网络来改变key和value的值。然后再针对这两个小型神经网络进行训练，由于规模显著减小，所以可以降低训练的成本。

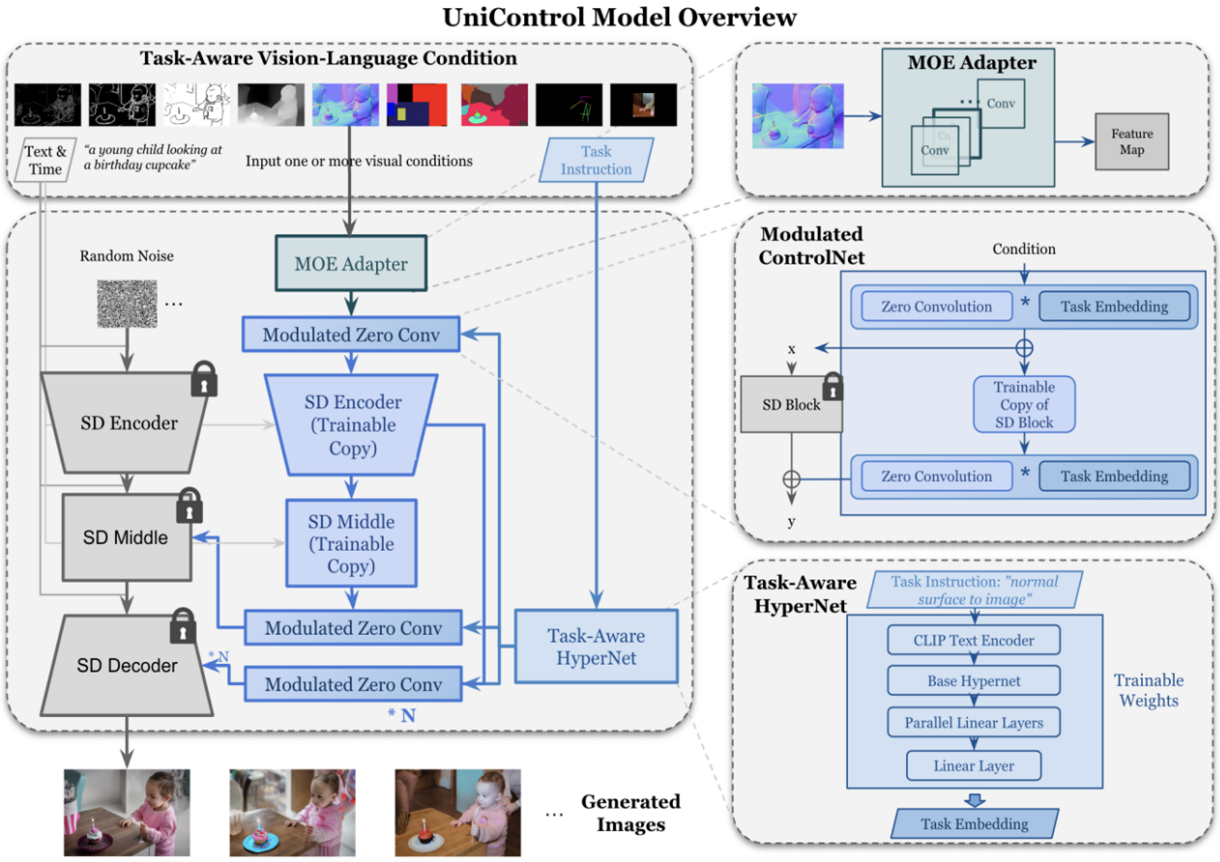



ControlNet中使用的HyperNet则是将原本的网络进行了固定（图中左侧锁住的模块），再复制了一个相同的网络作为外接的可训练网络。同时各个网络之间以零卷积层（1*1卷积）连接，保证了在未训练时模型输出与原模型一致。然后将额外的输入c从新加入的一侧进行输入，经过运算后加回到输出的y之中，以改变整个模型的输出



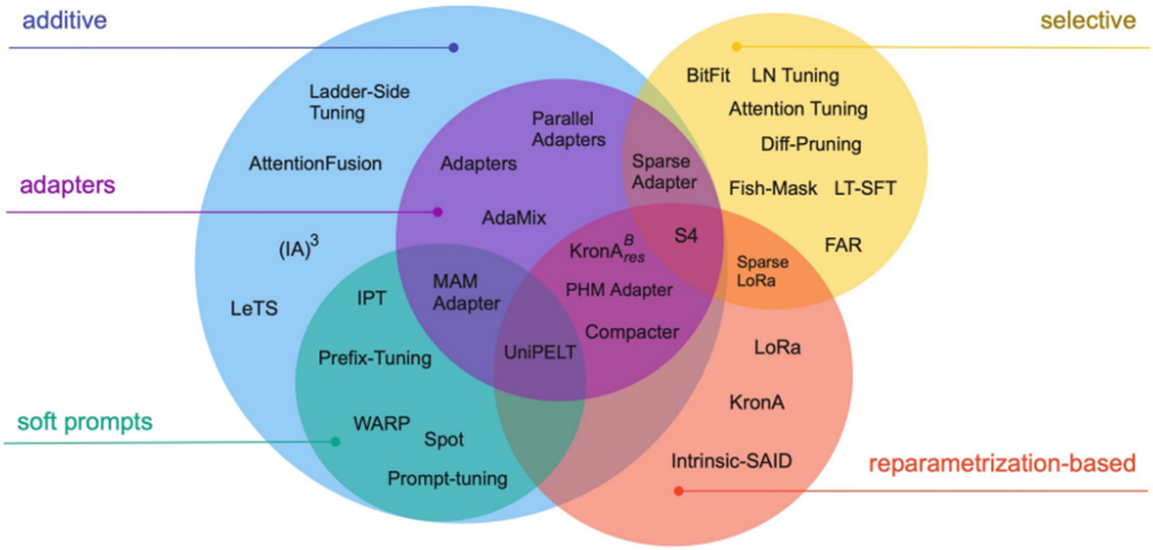

而Unicontrol则在Control的基础上采用了MOE的方式进行训练。MOE（Mixture of Experts），最早于1991年就已经提出了。其主要思想时通过训练数个结构一致的专家模型，对结果共同做出判断。每个专家模型意见的权重由其与输入的契合程度所决定。现行常用的MOE进一步添加了Gating Network机制，极大提高专家模型数量，但每次预测只交由数个模型，大多数模型不起作用。

本文中MOE用来将ControlNet的多个模型集合在同一模型中，且赋予了其整体更好的表现和处理零次训练问题的能力




## 4.3 通用大模型微调调研

微调这一概念可以追溯到18年推出的BERT模型（NLP），也就是根据一个预训练模型，然后针对任务进行特定的微调，可以取得优秀的效果。此时的微调还是采用的全量参数微调（FFT），但在预训练模型的参数规模不断膨胀后，固定大部分参数，而去调整小部分参数的PEFT，成为了实践的主流。PEFT又可以根据下图主要分为三大类别，其中的additive类别可以进一步细分，形成四大类别，但彼此之间仍有交叉。




Additive方式的主要思路是给预训练模型额外加入一些参数，然后通过对这些新参数进行训练来实现模型的微调。其中又最普遍使用的又分为adapters方式和soft prompts方式。Selective方式试图去忽略掉模型的结构，而去稀疏的调整极少数参数来实现微调。Re-based方式利用低秩表征来最小化可训练的参数量。

**Adapter方式涉及在Transformer子层后引入新的小型网络**

- Adapter Tuning，也是这一类别中最基础的方法，该方法设计了Adapter结构，并将其嵌入Transformer的结构里面，训练时只对这些新增的Adapter和Layer Norm进行训练
- Adapter Fusion 基于Adapter的优化，将整个学习部分划分成了两个阶段，从而提高下游的任务表现
- Adapter Drop，对Adapter进行了剪枝，抛弃了低层Transformer的Adapter结构，在不影响效果的前提下，大幅提高了推理速度，并小幅提高训练速度。
- Adapter Mix，采用了MOE策略，引入了多个Adapter结构，从而提升面对复合任务时的准确率。

**Soft prompt的原始思想是通过给输入加入一些特定的修改来提高对特定任务的处理能力。为了将这一在离散空间中寻找修改的问题变为一个优化问题，引入了“Soft”这一理念，也就是将嵌入的内容通过梯度下降进行优化**

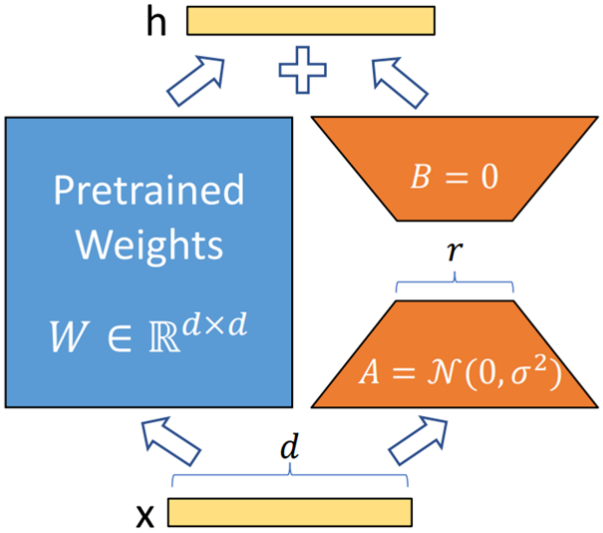
- Prefix prompt，在每一层上都带上一个Virtual Token作为前缀，以适应不同的任务
- Prompt tuning，针对不同的任务，仅在输入层引入Virtual Token，是对Prefix prompt的简化
- P-tuning，将prompt转化为了可以学习的Embedding层，并使用MLP+LSTM的方式来对Prompte Embedding进行处理，并且token的插入位置也不固定在前缀上。

**Slective方式都会以某种标准选取部分参数进行训练**

- BitFit，仅对bias参数进行修改，数据量极小

- DiffPurning，会对部分参数进行mask处理
- FAR（Freeze and Reconfigure），挑选参数并做冻结或激活处理
- FishMask，根据Fisher矩阵来决定哪些参数需要更新（Fisher矩阵可以表示参数的重要性，常用于剪枝）

**Reparametrization-based方式最典型的代表为LoRA**




- LoRA算法的主要思想就是，将训练d*d的矩阵，转化为训练一个d*r的降维矩阵和一个r*d的升维矩阵，来减小参数。在性能上可以匹敌全参量微调
- AdaLoRA是对LoRA的改进，通过重要性评分来区分对待不同的参数矩阵。更重要的保留更高秩，不重要的就采用低秩
- KronA,采用了克罗内积来降低了LoRA的计算开支
- QLoRA,通过低精度的数据储存类型（4bit），和计算数据类型（BFloat16),以及双量化技术，进一步减小了微调所需要的内存。

## 五：Reference

《Scaling Down to Scale Up: A Guide to Parameter-Efficient Fine-Tuning》

《Adding Conditional Control to Text-to-Image Diffusion Models》

《UniControl: A Unified Diffusion Model forControllable Visual Generation In the Wild》