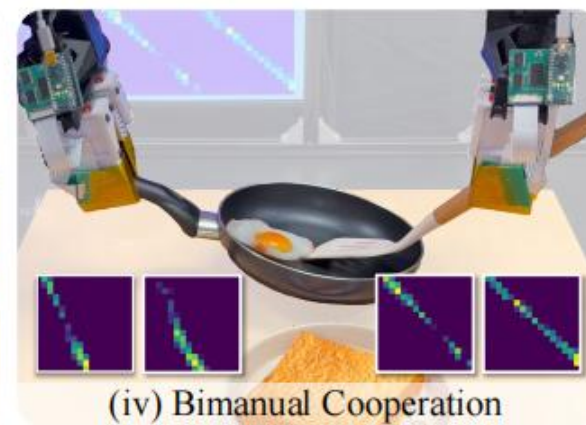
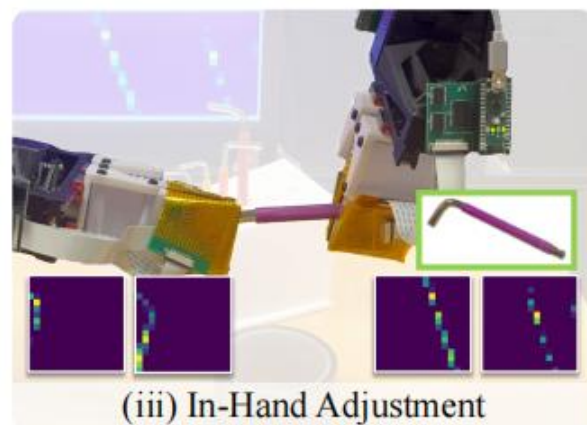
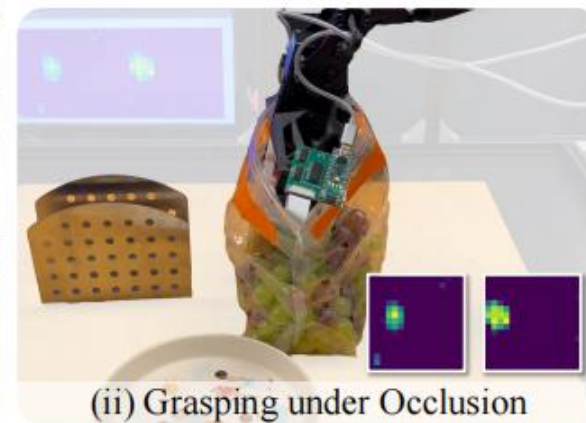
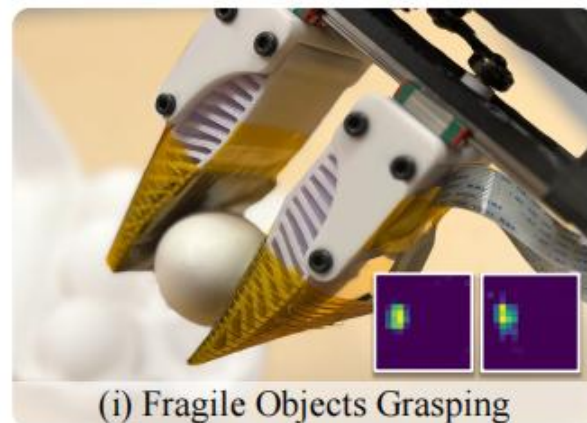
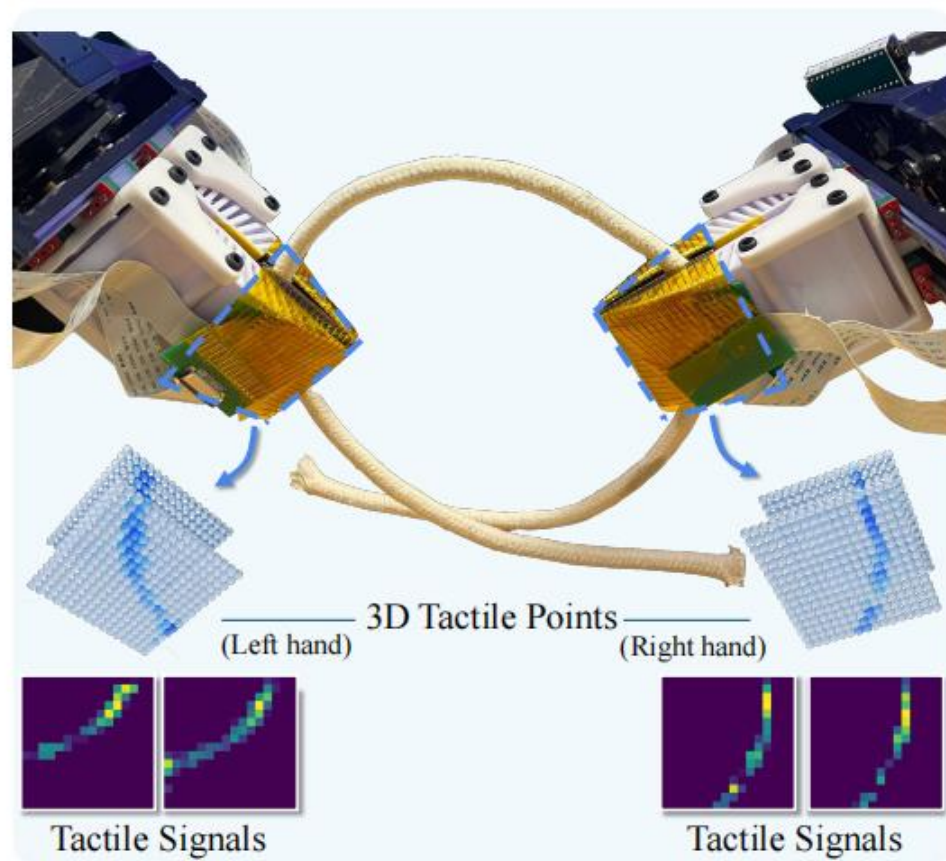


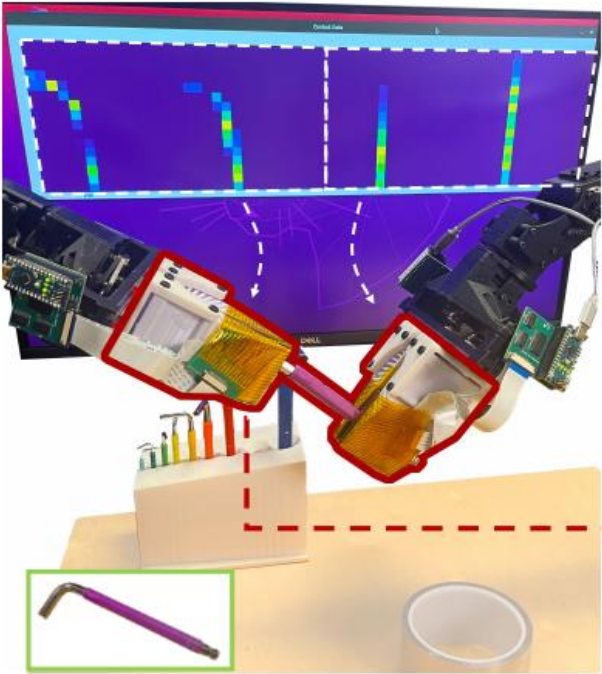
3D-ViTac: Learning Fine-Grained Manipulation with Visuo-Tactile Sensing



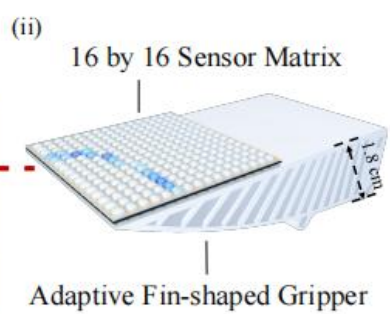
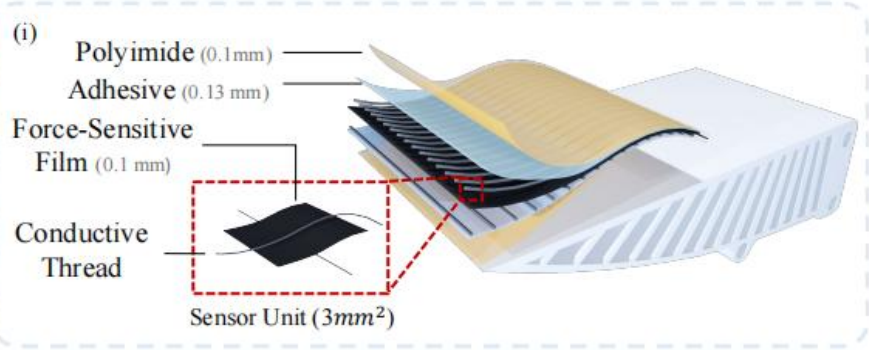
Contact-rich 典型场景

(i)精确施力 (ii)遮挡 (iii)调姿定位 + 插入 (iv)多工具配合

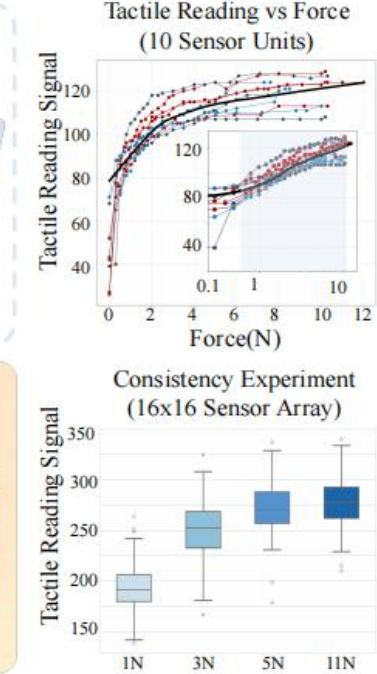
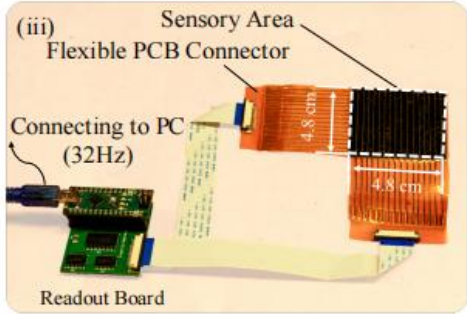
主要贡献点1：触觉硬件结构



(a) Tactile Sensing Platform



(b) Tactile Integrated Soft Gripper Design



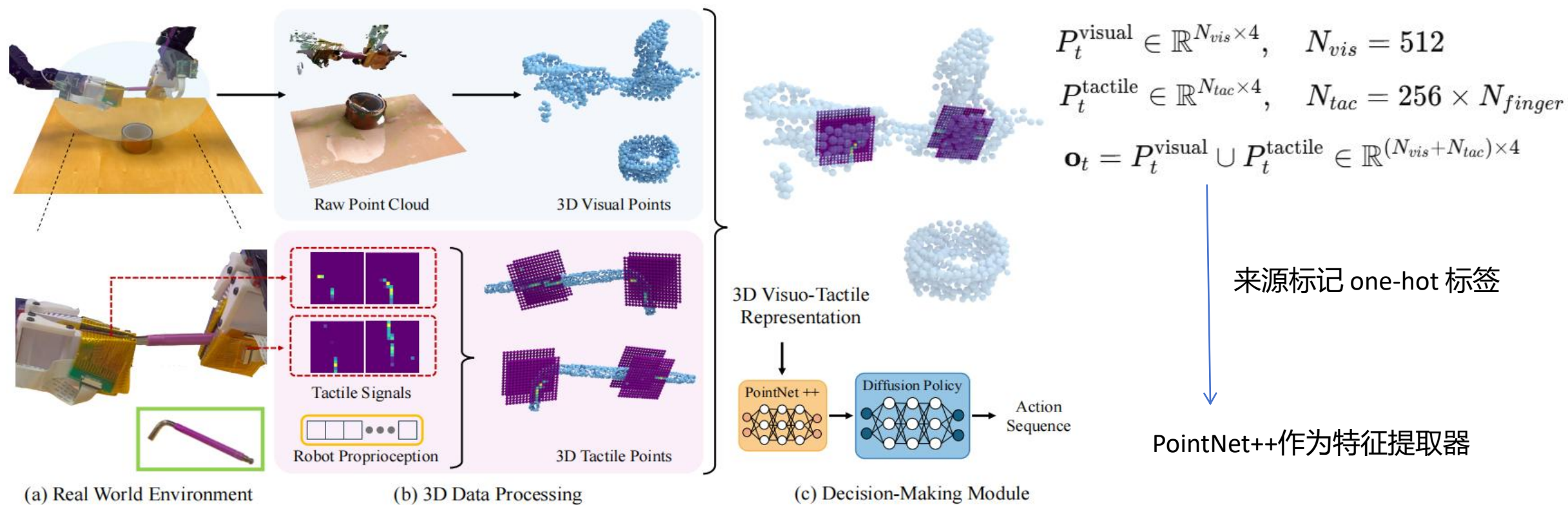
(c) Physical Characteristic

- 阻式传感矩阵 (16×16分辨率)
- 3D 打印软夹持器
- 读取速率达到 32.2 Hz
- 响应与加载力呈对数线性关系

类别	代表传感器	优势	主要问题
光学触觉传感器	GelSight、DIGIT、OmniTact、GelSlim 3.0	高分辨率 ;能输出触觉图像	体积大：厚重且难以安装于灵活手指上， 刚性高：不适合柔性抓取或大面积接触， 视野受限：光学系统需固定方向观察
电容/电阻/压阻型二维阵列传感器	uSkin、ReSkin	可大面积覆盖 ;成本较低 ;柔性材料可选	分辨率低：如4×4、8×8阵列，难感知细节 信号离散或二值化：缺连续压力分布信息
商业集成高性能触觉模块	BioTac、SynTouch、TakkTile	高灵敏度 ;多模态信号（力、温度、湿度）	价格昂贵，封闭系统，难以二次开发或集成较大体积，柔性差

主要贡献点2：视触融合训练策略

- 多模态融合：视觉点云 + 触觉点云 → 统一三维表示
- 模仿学习框架：PointNet++ & Diffusion Policy



三视角
Realsense D435 ×2 + D455 ×1

通过正向运动学 (FK)
将触觉数据转换为 3维坐标

PointNet++ 网络架构

步骤	操作	功能
① Sampling	从原始点云中选取关键点 (如 Farthest Point Sampling)	缩小规模, 建立局部邻域
② Grouping	为每个采样点找到其 邻近点 (基于半径或 k近邻)	构建局部点集
③ PointNet	对每个局部点集单独应用一个小型 PointNet	提取局部特征并汇聚 (max pool)

非结构化的点云

$$\mathbf{o}_t = P_t^{\text{visual}} \cup P_t^{\text{tactile}} \in \mathbb{R}^{(N_{vis}+N_{tac}) \times 4}$$

来源标记 one-hot 标签

由多个 Set Abstraction (SA) 层组成, 每个 SA 层得到分层聚合局部特征

$$\text{MLP}(\mathbf{p}_i) = \phi(\mathbf{W}_n \cdot \phi(\cdots \phi(\mathbf{W}_1 \cdot \mathbf{p}_i + \mathbf{b}_1) \cdots) + \mathbf{b}_n)$$

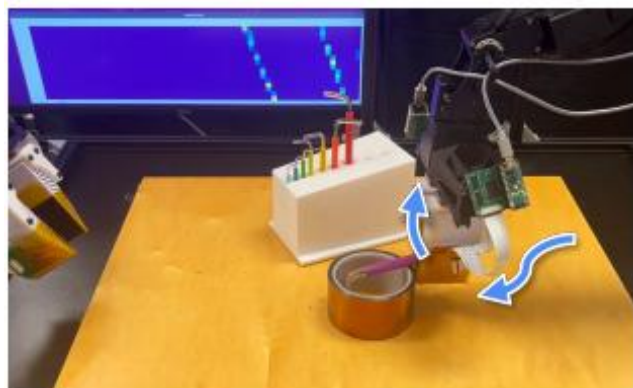
三层 SA 层和两层全连接层组成的 PointNet++ 特征提取器配置

层次	参数配置	说明
SA-1	Sample 64点, 半径0.04, 邻域内采样16个点 MLP: [64, 64, 128]	提取局部小范围特征
SA-2	Sample 16点, 半径0.08, 邻域采样32个点 MLP: [128, 128, 256]	进一步汇聚更大区域特征
SA-3	Global abstraction 层 (所有点) MLP: [256, 512, 1024]	汇聚全局特征
FC-1	1024 → 512	全连接层, 降维
FC-2	512 → 256	再降维, 为后续Diffusion Policy做条件输入

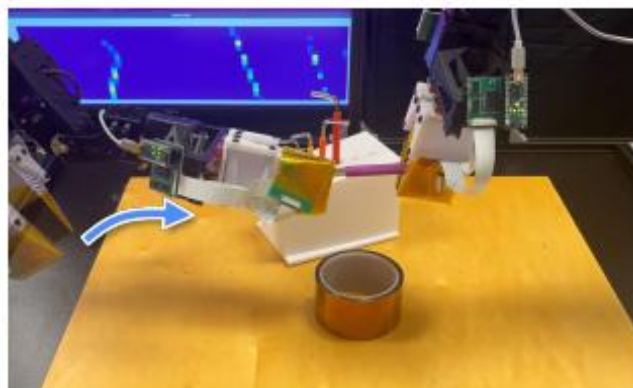
对不同范围的无序集合提取局部语义, 得到全局特征—256 维特征向量 z

状态向量 z +随机高斯噪声+专家动作序列 (30~50 条演示, 双臂远程操控系统采集)

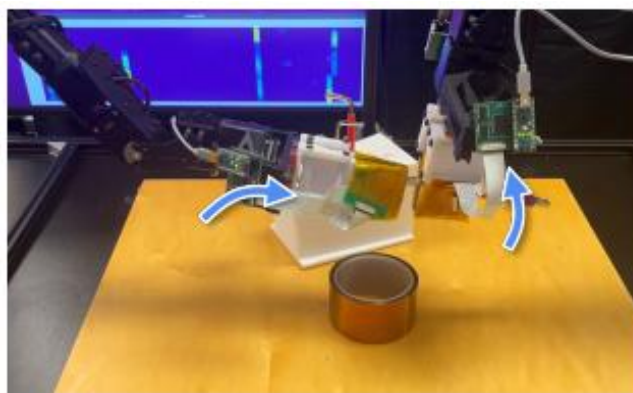
PointNet++ 中“局部区域聚合”如何捕捉触觉模式？



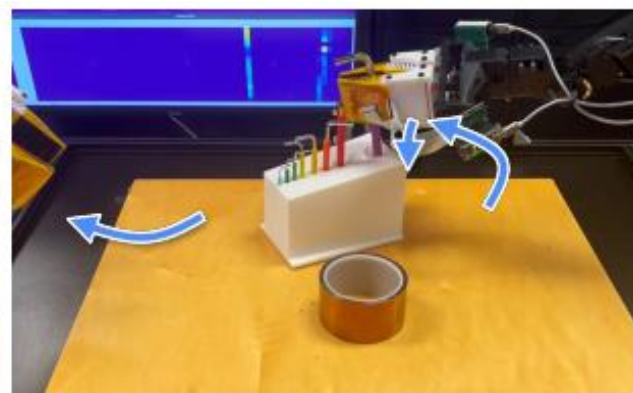
(a) Right Hand Grasp



(b) Left Hand Grasp



(c) Adjust and Rotate



(d) Insertion



Init State (Hex Key Collection)

一个SA层的某采样点 p_i 落在夹爪触觉阵列，球形邻域采样后得到的都是触觉点云信息，



$$\text{MLP}(\mathbf{p}_i) = \phi(\mathbf{W}_n \cdot \phi(\cdots \phi(\mathbf{W}_1 \cdot \mathbf{p}_i + \mathbf{b}_1) \cdots) + \mathbf{b}_n)$$

空间排布+压强分布

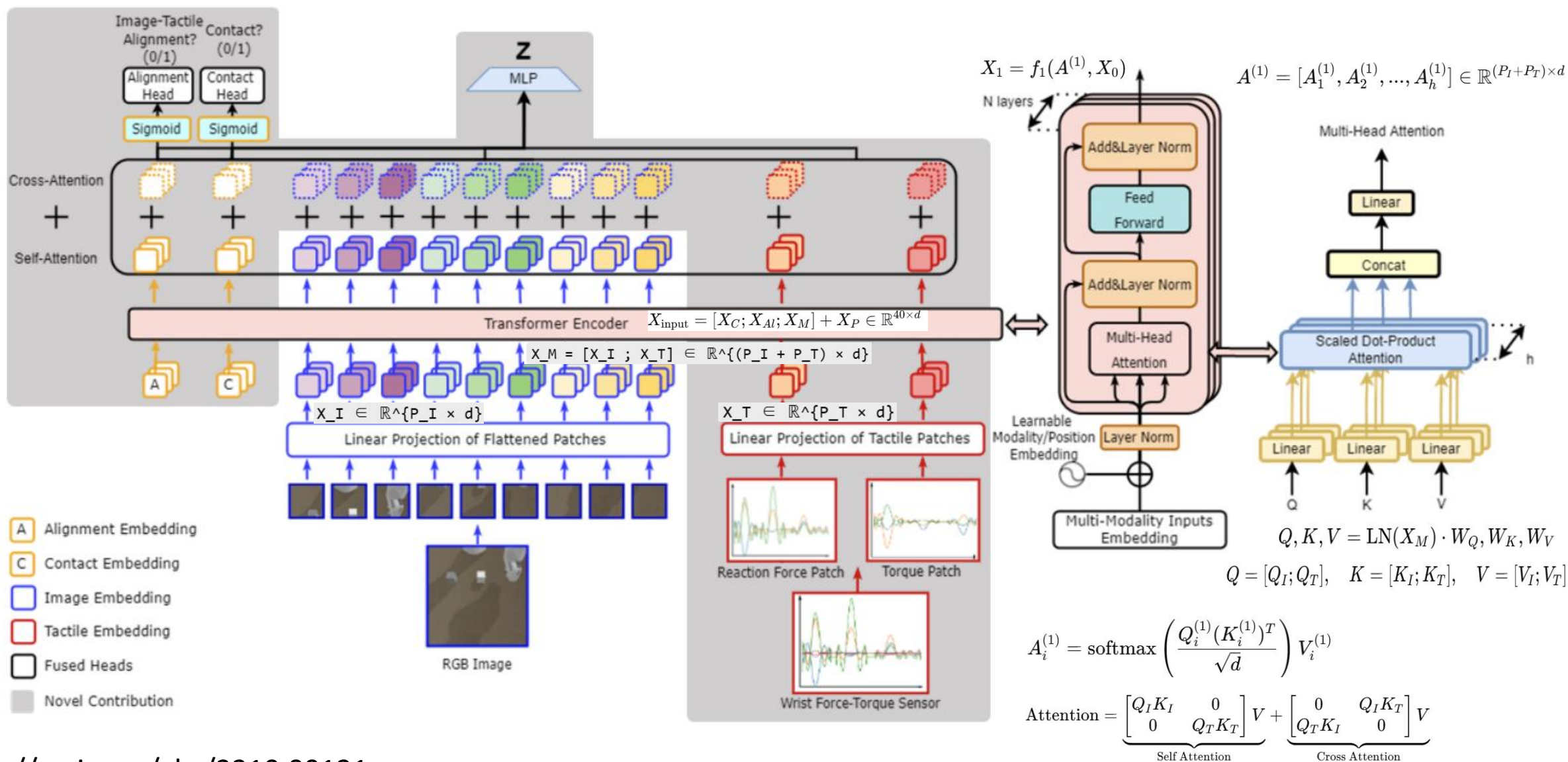
语义特征：“不均匀夹持”、“大力抓取”、“物体偏心”

视触的非结构化3D点云— PointNet++

VS

2D图像和六维力传感器的信息融合—VTT

Visuo-Tactile Transformers for Manipulation



视触融合机制

特征空间融合 VS 三维空间融合

力觉/光学触觉传感器 VS 高精度触觉传感阵列

对比项	VTT	3D-ViTac
融合方式	Self-attention + Cross-modal Attention (显式)	输入点云已融合 (视觉点+触觉点合并)
融合层次	多层 Transformer encoder 逐层交互	PointNet++ 一开始即处理融合后的点云
时序对齐机制	引入 Alignment Embedding 学习时序对齐	利用 FK 映射触觉点云到空间坐标, 与视觉点云合并
接触语义融合	引入 Contact Embedding 与 Contact Loss	显式利用触觉接触形成接触点点云

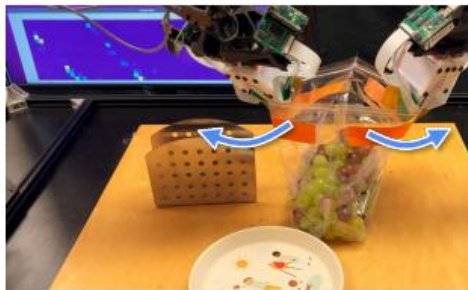
横向对比实验



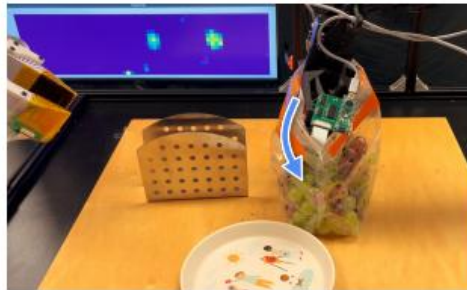
Init State (Fruit Preparation)



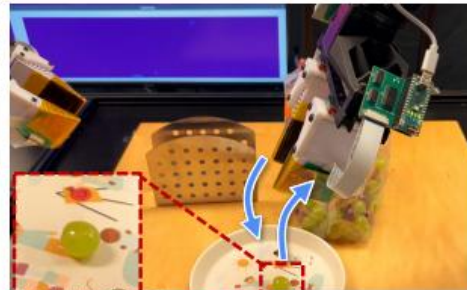
(a) Grasp and Place Plate



(b) Open Plastic Bag



(c) Grasp Grape



(d) Place Grape

Tasks Requiring Fine-Grained Force Information

Modalities	Egg Steaming (30 demos)				Fruit Preparation (30 demos)			
	Open Egg Tray	Grasp Egg	Place Egg	Whole Task	Grasp and Place Plate	Open Plastic Bag	Grasp Grapes	Whole Task
RGB Only	0.95	0.75	0.60	0.50	0.95	0.75	0.45	0.45
RGB w/ Tactile Image	0.95	0.90	0.80	0.70	0.90	0.80	0.70	0.70
PC. Only	1.00	0.75	0.65	0.55	1.00	0.85	0.50	0.45
PC. w/ Tactile Image	1.00	0.85	0.70	0.70	0.95	0.90	0.75	0.65
PC. w/ Tactile Points (Ours)	1.00	0.90	0.90	0.85	0.95	0.85	0.80	0.80

Tasks Requiring In-Hand State Information

Modalities	Hex Key Collection (30 demos)				Sandwich Serving (50 demos)			
	Right Hand Grasp	Left Hand Grasp	In-Hand Adjustment	Whole Task	Grasp Serving Spoon	Tilt Pot	Get Fried Egg	Whole Task
RGB Only	0.90	0.85	0.55	0.45	1.00	0.95	0.70	0.60
RGB w/ Tactile Image	0.95	0.90	0.70	0.60	1.00	1.00	0.85	0.75
PC. Only	1.00	0.90	0.65	0.65	1.00	1.00	0.75	0.65
PC. w/ Tactile Image	0.95	0.85	0.60	0.50	1.00	0.90	0.75	0.65
PC. w/ Tactile Points (Ours)	1.00	0.95	0.95	0.90	1.00	1.00	0.90	0.85

对所有策略2,000 个 epoch

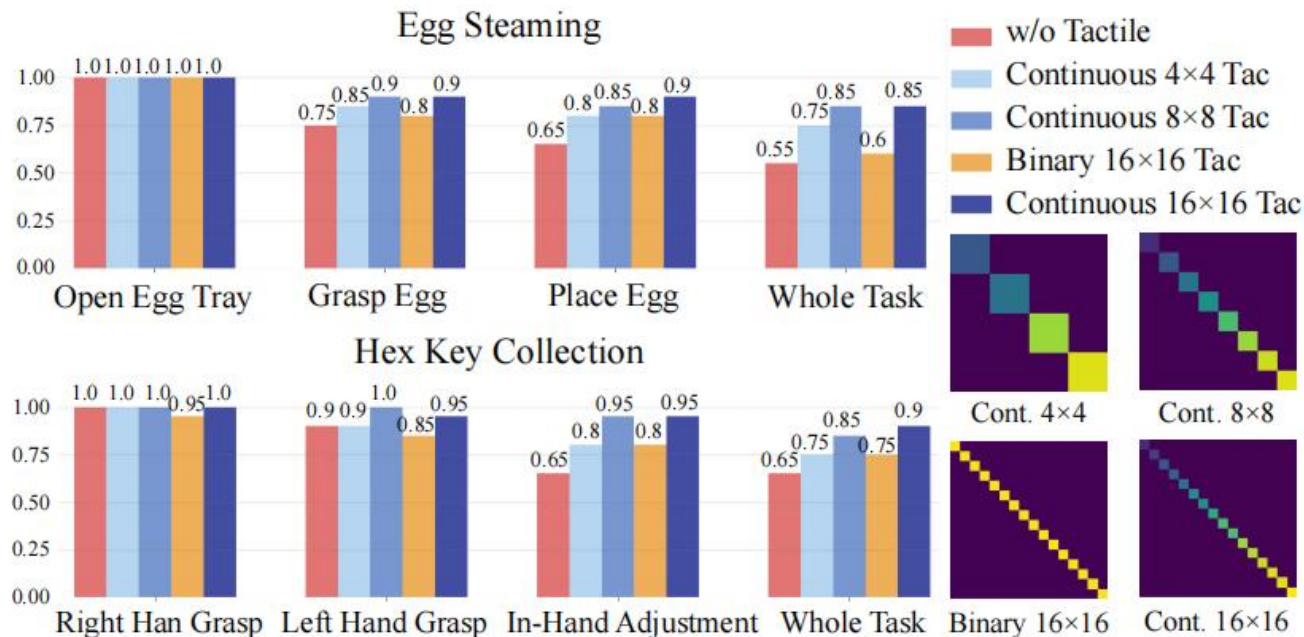
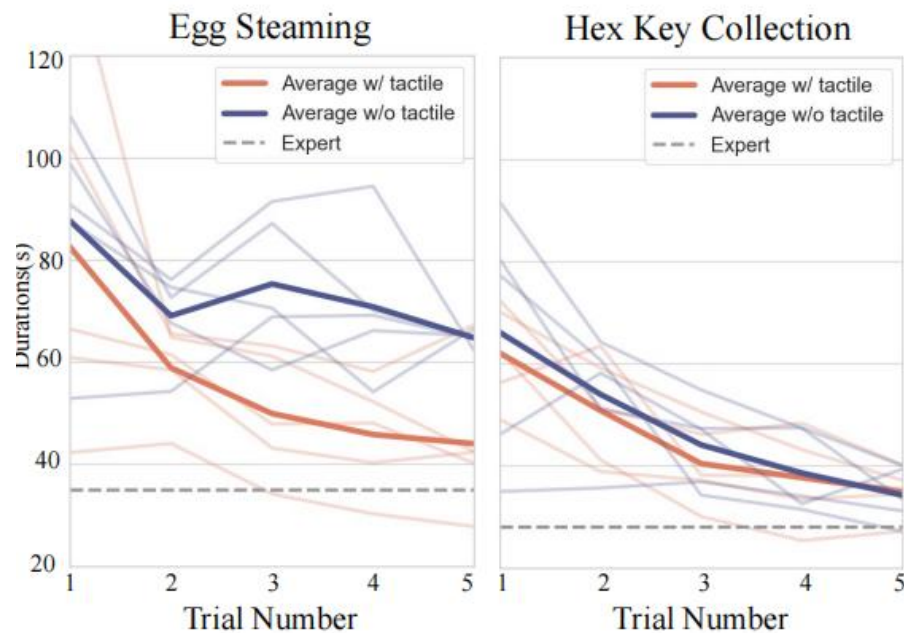
(1) 仅 RGB, 基于图像的扩散策略
(2) CNN 提触觉图像的特征, 通过扩散策略的不同分支处理多视图 RGB 图像和触觉图像。

(3) 仅 PC, PointNet++ 提多视图视觉点云特征用于扩散策略。

(4) CNN 提触觉图像的特征, PointNet++ 提多视图视觉点云, 通过不同的分支融合, 以实现扩散策略。

消融实验

Settings	Egg Cooking (30 demos)				Hex Key Collection (30 demos)			
	Open Egg Tray	Grasp Egg	Place Egg	Whole Task	Right Hand Grasp	Left Hand Grasp	In-Hand Adjustment	Whole Task
Tactile Points Only	0.70	0.55	0.10	0.10	0.50	0.30	0.30	0.15
Single Cam PC. Only	0.90	0.65	0.60	0.50	0.80	0.60	0.45	0.30
Multi Cam PC. Only	1.00	0.75	0.65	0.55	1.00	0.90	0.65	0.65
Single Cam PC. w/ Tactile Points	0.95	0.90	0.85	0.80	0.85	0.80	0.75	0.70
Multi Cam PC. w/ Tactile Points	1.00	0.90	0.90	0.85	1.00	0.95	0.95	0.90



批判性&建设性问题