吕九峦-答辩记录

## 答辩中提出的主要问题及回答的简要情况记录

问：研究现状存在的挑战要分析是什么？

答：现有纵向联邦学习方法过度依赖大量标记和对齐数据，而实际应用中往往面临高昂的标注成本和领域专家稀缺的情况。尤其当只有一方拥有标签信息时，联合模型训练的复杂度显著提高。同时，不同参与方的数据分布存在显著差异，导致模型难以有效学习。尽管RSCFed采用加权距离聚合、FedMatch通过客户端间一致性来应对异质性，但在复杂应用场景中的效果仍有待提升。

问：在VFPU框架中，如何保证数据隐私不被泄露？

答：在VFPU框架中，采取了多层次的隐私保护机制：首先，参与方仅共享中间计算结果（如梯度、预测值）而非原始特征数据；其次，对于共享的中间结果，我们采用了安全聚合协议，确保任何参与方都无法从接收到的信息中推断出其他方的私有信息；此外，对于特别敏感的场景，还设计了差分隐私机制，通过向共享值添加校准噪声来进一步增强隐私保护。实验表明，这些机制能在保障模型性能的同时，有效防止隐私泄露风险。

问:FedPSG-PUM的核心创新点是什么？与现有方法相比有何优势？

答：FedPSG-PUM的核心创新在于将联邦半监督学习与生成模型技术结合，提出分阶段处理策略：跨方特征相关性分析：通过隐私保护的Spearman秩相关计算，筛选高相关性特征，指导后续生成策略。联邦生成模型填补：对低相关性特征，采用VF-GAIN等联邦生成模型，学习多方数据分布以填补缺失值。

问：FedPSG-PUM的局限性及未来方向？

答：联邦半监督学习需多轮迭代，VF-GAIN的对抗训练增加通信成本；τ和α需通过验证集调优，可能影响泛化性；引入元学习或贝叶斯优化自动选择τ和α。 适配图像、文本等多模态数据，提升方法普适性。

问：算法中正例比例估计环节对最终性能有多大影响？

答：正例比例估计是VFPU算法的关键环节，对最终性能有显著影响。我们的实验表明，当估计误差在±10%范围内时，模型性能下降相对可控；但若误差超过15%，模型性能将明显下降。

问：如何评估模型在联邦环境下的通信效率？

答：从三个维度评估联邦环境下的通信效率：首先，通信轮次数量，即达到模型收敛需要的总通信次数；其次，单轮通信量大小，即每次交互传输的数据量；第三，整体时间消耗，包括通信延迟和计算时间。实验结果表明，与基准方法相比，VFPU在保持相似精度的情况下，通信轮次减少了约20%，整体训练时间缩短了25-30%。

## 答辩委员所提出的论文修改意见

1. 研究背景及意义这部分介绍了联邦学习的背景，特别是纵向联邦学习（VFL）及其面临的挑战，如标注成本高、对齐样本不足等。然后引出了提出的方法VFPU和FedPSG-PUM。整体逻辑连贯，但可能存在文献引用不足的问题。例如，在提到联邦学习和纵向联邦学习时，引用了部分文献，但对于PU学习的引用可能不够全面，特别是PU学习在联邦场景中的应用是否有更多相关研究需要确认。此外，问题提出的部分可能需要更具体地指出当前研究的空白，例如现有方法在处理未标记数据缺失时的局限性。
2. 文献引用更新：建议检查是否有更新的相关研究文献（2022-2023年）可以补充，以体现研究的前沿性。表格优化：表1中的"性能"一列内容不够清晰，建议重新组织或者增加解释。
3. 第三章问题分析与定义部分详细描述了数据分布和问题设定，但可能存在术语使用不一致的情况，比如“样本ID空间”是否前后一致，是否需要进一步解释。此外，例子中的A、B、C三方描述是否清晰，是否有重复或冗余的信息。
4. 术语一致性“自助法”建议统一为“Bootstrap方法”或“自助采样法”，并在首次出现时附加英文（如“自助法（Bootstrap）”）。“未标记数据缺失的PU学习问题”缩写“UDD-PU”需在首次出现时全称标注。
5. 图表呈现图4-1(FedPSG-PUM算法总体流程)中的文字较小，建议调整字体大小或重新设计图表布局，确保可读性。实验结果图表中(如图4-2和图4-3)建议添加误差线或置信区间，以展示结果的统计稳定性。