**5 讨论（还没开始写，以下是写作思路）**

**5.1 理论解释**

1. Mode 与 Raha-Baran 的差异机理
   * 在前文（第4节）对比了这两种清洗策略的性能差异，这里可进一步从理论层面探讨：
     + Mode 策略：简单、高效，主要假设大多数特征维度具有明显主频值（或常用值）。在数值型或缺失程度较低的数据集下，能够快速修复部分噪声/空值；但面对复杂语义冲突或知识库约束时，缺乏足够的上下文推理能力，容易引入偏差。
     + Raha-Baran 策略：依赖更丰富的上下文与规则，实现对各种语义、规则冲突的精准修复。但其算法开销和对先验知识的依赖度更高。
   * 将这些结果与已有文献进行对比，说明本研究如何证实了“适度复杂的清洗策略在错误率不高时效果最佳，而在错误率极端时也可能受限”等结论；同时可引用其他学者的类似或相反发现，与之形成呼应或对照。
2. 层次聚类 (HC) 的稳健性与适用性
   * 结合聚类原理：HC 逐步合并或拆分簇，不依赖随机初始化；因此在轻度噪声和缺失值场景下通常更稳健。
   * 与 KMeans、DBSCAN 等对比：
     + KMeans 假设球形分布，对噪声更敏感；
     + DBSCAN/OPTICS 在噪声识别和ϵ\epsilonϵ参数选择不当时容易出现极端聚类形态。
   * 理论视角可再次佐证：HC 在适当的规模和噪声水平下能更好地逼近基准结果，尤其当清洗策略能减少无效或混乱的数据点时，HC 的逐步分层更能体现出稳定优势。
3. 特殊现象的机理解释
   * 高于100%的得分：可从“内部指标的放大效应”切入，说明在距离度量或簇间离散度被极端拉大的场合，DB\_score、Silhouette\_score 等会呈现过高评价。
   * 不收敛或超时：可将失败场景归因于参数初始值过于极端、数据分布被错误修复扭曲，以及部分算法复杂度（如层次聚类在大规模数据下的O(n2)甚至O(n3)开销过高等。

5.2 方法与评估指标的深入分析

1. 自动化调优（Optuna）对性能与稳定性的提升
   * 在第4节已经提到Optuna搜参的成果，这里可更深入讨论：
     + Optuna 的 TPE / 贝叶斯优化在多大程度上缩小了盲目搜索空间？
     + 是否有场景（如超高维度或数据规模非常大）需要新的采样策略或并行化搜索？
   * 可以结合已有文献或 AutoML 相关研究，说明本研究的选择如何符合实际聚类任务在无监督场景中的需求。
2. 内部聚类指标（DB\_score 与 Silhouette\_score）的局限
   * 在第4节多次提到这些指标导致部分组合“爆分”。在本节可更系统地指出：
     + 高分不一定代表与真实业务意义吻合；
     + 当簇数过多或初始分区极端时，DB\_score与Silhouette\_score可能被高估；
     + 同时，这些指标对噪声敏感度较高，当清洗策略改变数据分布后，很容易产生过于理想化的聚类效果。
   * 建议在实际应用中，若需要更可解释或稳健的聚类结构，可在内部指标之外结合外部指标或专家评估。

**5.3 清洗与聚类的协同作用**

1. 协同提升聚类性能的机制
   * 在结果分析（第4节）中已指出“清洗能提升聚类效果”，此处可结合更多参考文献与理论模型，阐述“噪声/缺失值的减少如何降低簇内离散度、增强簇间分离度”。
   * 如果有可能，可插入一个示意性流程图，展示“清洗 -> 数据分布优化 -> 超参数调优 -> 聚类模型收敛更好”的框架。
2. 清洗策略对协同作用的影响
   * 对比 Mode 与 Raha-Baran 对于协同提升的不同贡献路径：
     + Mode更简单，但若数据本身多语义冲突，协同效果有限；
     + Raha-Baran能深度修复语义与逻辑冲突，在高维、具有复杂错误类型的数据上对聚类帮助更大。

5.4 局限性与展望

1. 数据集与清洗策略的限制
   * 第4节聚焦在40个数据集，这里可从通用性与外部有效性等学术角度指出：
     + 数据规模是否足够广泛？
     + 数据类型以数值或可转化为数值居多，是否对文本/图像/多模态数据也能适用？
     + Mode 与 Raha-Baran都属于已知的典型清洗策略，但对更复杂的知识库约束或跨表数据一致性，是否需要新的清洗范式？
2. 超参数优化范围的局限性
   * Optuna 在聚类中虽有助于自动找优，但依赖于用户定义的搜索空间。若搜索范围不合理，仍可能错过全局最优。
   * 在大规模或高维数据中，优化时长与内存开销均需进一步平衡。
3. 聚类算法种类的局限
   * 本研究聚焦于6种经典算法，暂未涉及深度学习聚类（如 Deep Embedded Clustering 等）。
   * 未来可研究对高维度（>100特征）、大量非线性特征或图结构数据的聚类策略适用性；或与自监督学习、表示学习相结合。
4. 对实际应用的挑战
   * 可以最后强调：真实场景中并非只有数值与少量的字符串字段，可能存在更多异构数据或实时流数据等，需要考虑在线清洗与在线聚类的联合。
   * 对于强噪声或错误率极高（>50%）的情况，需要更深入的领域知识或交互式修复过程。