**accepted**

|  |
| --- |
| **工作概述**  测一些闭源模型的rebuttal case，对比看看生成的rebuttal有什么问题 |

**已有经验**

一个有效的反驳应具备以下特点：

相关性： 直接回应原始论点的核心观点 。

连贯性： 逻辑结构清晰，易于理解 。

说服力： 能够改变听众或读者的观点，尽管当人们意识到是 AI 参与时，批判性可能会增加 。

事实准确性/有根据性： 基于正确信息，避免幻觉 。这一点至关重要，因为 LLM 即使从事实正确的信息源获取数据，如果对上下文理解错误，也可能生成错误信息 。

逻辑严谨性： 避免逻辑谬误 。

立场明确： 清晰地表达对立观点 。

简洁性： 以信息丰富且简洁的方式阐述观点 ，尽管 LLM 可能难以满足特定的长度限制 。

**💎模型review与rebuttal回答**

|  |
| --- |
| 🌟rebuttal所用的prompt  你是一名经验丰富的学术论文作者，正在准备对已有的审稿人评论进行rebuttal。你的目标是专业、建设性地回应每个问题，以争取论文被接收。请严格遵循以下规则生成rebuttal：  输入信息（用户将提供）  - 审稿人评论：[在此粘贴审稿人的完整评论，包括所有编号点]  - 论文摘要：[在此粘贴论文摘要]  - 关键证据：[在此粘贴与评论直接相关的论文段落、图表编号或实验数据（例如："Section 3.2, Figure 4"）]  - 修改计划：[可选：说明你已/将如何修改论文，例如"将在Camera-ready版本中补充实验" ]    rebuttal生成规则  1. 整体要求：  - 语气：礼貌、尊重、积极（例："感谢审稿人指出..."而非"审稿人错误地认为..."）。  - 长度：每个评论点的回应≤150字；总rebuttal≤500字。  - 禁止：编造未提供的证据、攻击审稿人、回避问题或过度承诺修改。    2. 结构化回应（对每个审稿人评论点）：  - 步骤1: 确认理解（20字内）：简要重述评论，表明你认真阅读了（例："我们理解审稿人关注[具体问题]..."）。  - 步骤2: 核心回应（80字内）：  - 若评论正确 → 承认并说明修改计划（例："您是对的，我们将在Section 4添加讨论..."）。  - 若评论是误解 → 用\*\*具体证据\*\*澄清（例："根据Section 3.1的Eq. (5)，[解释]；这与评论中假设不同，因为..."）。  - 若评论是建议 → 评估可行性（例："我们将在未来工作探索此方向，但当前版本受限于[原因]..."）。  - 步骤3: 行动承诺（20字内）：明确论文修改动作（例："已在修订稿中更新Figure 3"）。    3. 全局结构：  - 开头：1句感谢（例："感谢审稿人的细致评论，帮助我们提升论文质量。"）。  - 主体：按审稿人评论编号逐点回应（格式：Comment X: [回应]）。  - 结尾：1句开放性承诺（例："我们乐意根据反馈进一步完善论文。"）。    现在生成rebuttal  基于以下用户输入，输出纯文本rebuttal（无需额外解释）：  审稿人评论：  [用户粘贴评论]  论文摘要：  [用户粘贴摘要]  关键证据：  [用户粘贴证据]  修改计划：  [用户粘贴修改计划] 中文输出回答 |

* 论文https://openreview.net/forum?id=EMZBuEEk7jJ

**Gpt 太快了**

**💎模型review部分**

**论文“利用自监督方法识别粗粒度独立因果机制”审稿意见**

**一、工作概述**

本文研究了在多个生成机制及环境特定变换混合的数据中，如何恢复独立因果机制（Independent Causal Mechanisms，ICM）。作者提出了一种新颖的生成模型——ICM，利用**退化混合先验**（degenerate mixture prior）使得在每次采样时仅有一个机制子空间被激活。理论上，在可逆性假设下，作者证明了模型在“粗粒度”层面是可识别的——任意与数据分布一致的模型，仅能在子空间层面做置换或内部变换。为实现该模型，作者设计了一个自监督的 GAN 风格网络，结合辅助编码器、Wasserstein 损失、循环一致性与潜变量重构项。在 MNIST、FashionMNIST 和 Sprites 数据集上的实验表明，ICM 不仅更好地恢复了真实潜变量结构，而且在受控“环境转移”下，作为下游分类器也表现出更强的鲁棒性。

**二、优点**

1. **创新的先验设计**  
    引入退化混合先验，巧妙地在模型先验中内置了“同一时刻只激活一个机制”的 ICM 原则。
2. **可识别性解析**  
    在光滑可逆假设下，给出了定理证明：任何重建数据分布的模型，都只能在子空间层面做置换或内部可逆变换，从而保证了“粗粒度”可辨识性，填补了无监督因果机制辨析的理论空白。
3. **实用的自监督训练框架**  
    将 WGAN、辅助编码器与循环一致性有效地结合，避免了过于复杂的对抗训练，且有利于潜变量的可逆映射学习。
4. **充分的实证验证**

* 合成数据实验：成功恢复已知的退化混合与高斯潜变量。
* 潜变量遍历：清晰展示了不同机制（如数字的风格因子）间的分离效果。
* 鲁棒性测试：在多种环境转移设置下，ICM 相较 VAE、β-VAE、Ada-GVAE、VaDE 均有显著（最高约 9.8 个百分点）性能提升。

**三、弱点与建议**

1. **粗粒度 vs. 细粒度解耦**  
    “粗粒度”解耦仅保证机制间分离，但允许机制内部混合，可能影响对细粒度因子的可解释性（如笔画粗细与倾斜角度的区分）。  
   **建议**：讨论在需要细粒度解耦的场景下，能否通过层次化先验或多级混合模型扩展当前框架。
2. **理论假设与实际差距**  
    可识别性依赖于生成器与编码器的光滑可逆性，而 GAN 网络在实践中往往不严格满足。  
   **建议**：在实验中提供相应的可逆性诊断指标（如雅可比矩阵条件数分布、循环一致性误差统计）以佐证近似可逆性。
3. **基线与度量选择**

* 未与可逆流模型（如 i-VAE）等近期具有可识别性的模型进行对比。
* 仅通过下游鲁棒性间接衡量解耦效果，建议补充常用解耦度量（MIG、DCI、FactorVAE 分数等）以强化定量分析。

1. **超参数敏感性与机制数选择**  
    机制数 KGK\_G 与每个子空间维度需手动设定，缺乏敏感性分析。  
   **建议**：增加消融实验，考察 KGK\_G 取值过少或过多，以及子空间维度变化对性能的影响，为实践提供指导。
2. **论文表述与结构**

* 核心证明与关键推理较多地放在附录，建议在正文中增加直观说明或示例帮助理解。
* 符号略显冗余（如同时使用 zCz\_C 与 one-hot 指示）；可考虑精简。
* 建议在第 4 式处添加损失项的“可视化”总结框，配以要点说明以提升可读性。

**四、详细意见**

1. **第 2.2 节 退化先验**  
    动机清晰，建议对比一下非退化混合先验在 KL 散度或样本分布重叠上的量化差异。
2. **定理 4**  
    对“几乎处处互补子空间”假设的严格性可再补充说明，实践中子空间或许存在少量重叠。
3. **训练稳定性**  
    附录提到在非退化混合时 σ 很小时训练不稳定，建议补充对 IC M 中 GAN 训练稳定性的更多定量说明。
4. **Sprites 实验**  
    当前仅在下游分类器层面测试颜色偏移鲁棒性，无法直接反映 ICM 捕捉颜色变换的能力，可在讨论中予以说明。

**五、总体评价与推荐**

**推荐等级：强烈接受（Strong Accept）**  
 本工作在理论与实践两方面均有显著创新：退化混合先验、可识别性定理、自监督 GAN 框架及丰富的实证。若能补充上述澄清与诊断分析，将进一步提升论文说服力和实用价值。

我期待作者的回复与讨论。

**📝rebuttal部分**

感谢审稿人的细致评论，帮助我们提升论文质量。

Comment 1:  
**确认理解**：我们理解您关注“粗粒度与细粒度解耦”的可解释性问题。  
**核心回应**：正如 Section 3.2 所述，当前模型侧重于机制间分离，细粒度因子亦可通过在子机制内添加层次化先验扩展（参见 Figure 4）。由于篇幅限制，我们将在附录中提供该扩展思路。  
**行动承诺**：已在附录中补充相关讨论。

Comment 2:  
**确认理解**：您指出可逆性假设在实践中可能不严格成立。  
**核心回应**：实验中，我们计算了循环一致性误差（Table 2）与 Jacobian 条件数分布（Section 4.1），结果表明网络近似满足可逆性假设。  
**行动承诺**：已在 Section 4.1 中添加误差统计。

Comment 3:  
**确认理解**：您建议增加流模型基线对比。  
**核心回应**：由于 i-VAE 依赖部分监督标签，当前框架无标签设置下难以直接对比。未来工作我们将探索无监督流模型的集成。  
**行动承诺**：在讨论中已补充此局限性说明。

Comment 4:  
**确认理解**：您关心超参数敏感性及机制数量选择。  
**核心回应**：在 Supplementary Table 3 中，我们展示了 KGK\_G 在 {3,5,7} 和子空间维度 {8,16,32} 时的性能曲线，结果稳定。  
**行动承诺**：已在附录表格中补充该消融结果。

我们乐意根据反馈进一步完善论文。

**gemini 2.5pro**

**论文评审：使用自监督方法识别粗粒度独立因果机制**

**审稿人评分：** 5/5 (强烈推荐接收)

1. **总结**

本文提出了一种仅从观测数据中学习独立因果机制 (ICM) 的新颖方法。其核心贡献在于一种新的生成模型，作者称之为 ICM 模型，该模型利用了**简并混合先验 (degenerate mixture prior)**。这种先验旨在施加一个结构性约束，即数据样本在生成时仅激活一个数据生成机制。

作者提供了一个理论证明，表明在模型可逆的条件下，他们的方法可以导出一个**可识别模型 (identifiable model)**。这种可识别性在一定变换范围内成立，而这些变换能够保留作者所定义的“**粗粒度解耦 (coarse-grained disentanglement)**”——意味着不同数据生成机制之间、以及生成机制与变换机制之间的潜在变量保持非纠缠状态。

为了满足其理论的可逆性要求，作者提出了一种实用的**自监督训练方法**。该方法使用一个辅助编码器和循环一致性损失来学习生成器逆过程的近似。论文的实证评估非常广泛，包含了在 MNIST、FashionMNIST 和 Sprites 数据集上的实验。结果表明，与几个强大的基线模型相比，所提出的模型实现了更优的解耦，并在环境变化的情况下带来了更好的下游分类性能。

1. **优势**

这是一篇杰出的、执行得非常出色的论文，具有以下几个关键优势：

* **高度的重要性与新颖性：** 论文解决了因果表示学习中一个至关重要且极具挑战性的问题——可识别性。所提出的解决方案——利用简并混合先验来施加结构性约束以实现可识别性——既优雅又有效。在无需监督的情况下实现这一点，是该领域向前迈出的重要一步。
* **坚实的理论基础：** 论文的观点不仅仅是经验性的。**定理 4 (Theorem 4)** 为该方法为何有效提供了坚实的理论基础。我尤其欣赏作者在讨论可识别性的确切性质（“粗粒度”）及其局限性时的坦诚。这种严谨的治学态度大大增强了其贡献的可信度。
* **实用且动机明确的方法：** 自监督训练框架是一种将理论上的可逆性要求付诸实践的巧妙方式。损失函数（公式 4）的设计经过深思熟虑，它将用于保证样本质量的 WGAN 目标，与直接强制实现所需模型属性的自监督和循环一致性损失结合在一起。
* **全面且令人信服的实验：** 实验验证详尽且有说服力。
* 合成数据实验（**图 5**）作为一个完美的“概念验证”，清晰地展示了该模型恢复真实潜在结构的能力，而标准 VAE 则失败了。
* 定性可视化结果（**图 6**）清楚地展示了所期望的解耦效果，这与 β-VAE 的纠缠结果（**图 7**）形成鲜明对比。
* 基于环境变化下下游任务性能的定量评估（**表 1**），是衡量所学表示效用的一种有原则且实用的方法。其在多个强大基线上取得的一致且显著的提升令人印象深刻。
* 附录中的消融研究堪称典范，为损失函数的每个组成部分以及简并混合先验本身的必要性提供了强有力的证据。
* **无与伦比的清晰度：** 论文的写作水平极高，逻辑流程清晰易懂。图中插图的运用，特别是**图 2**（模型对比）和**图 3**（简并与非简并先验对比），非常有效地直观传达了核心概念。

1. **不足与作者问答**

这篇论文非常扎实，我在此提出的几点主要是为了澄清或讨论一些次要的局限性，而非重大的缺陷。

1. **关于机制数量（KG）的选择：** 作者在附录 B.5 中提出了一种手动、迭代的程序来选择数据生成机制的数量 KG。虽然这是一个实用的解决方案，但它确实引入了一个可能需要一些精力来调整的手动超参数。

* 您能否评论一下模型对这一选择的敏感性？例如，如果将 KG 设置得远大于“真实”的机制数量，会发生什么？性能是否会平滑地下降？
* 您是否考虑过任何自动化这一选择过程的方法？例如，在训练过程中观察到的某个指标（如与编码器预测或生成器输出相关的指标）能否为选择合适的 KG 值提供线索？

1. **对更复杂因果结构的可扩展性：** 论文假设的数据生成过程是一个清晰的两步复合过程：x=MT(MG(zG))。论文提到将此扩展到顺序复合是未来的工作方向。

* 您能否详细说明这种扩展可能面临的挑战？例如，在一个拥有 MT2(MT1(MG(zG))) 的系统中，当前简并先验的公式及其相应的可识别性证明是否仍然适用，还是需要进行重大修改？

1. **变换机制（MT）的性质：** 实验主要集中在很大程度上是全局性和几何性的变换（旋转、宽度、厚度）或与颜色相关的变换（zT）。但框架本身似乎更具通用性。

* 您是否考虑过模型如何处理更抽象或更局部的变换？例如，如果一个变换只影响对象的特定部分（比如改变 FashionMNIST 中衬衫的袖子样式），这是否仍能被一个共享的 zT 有效地捕捉？

1. **计算成本：** 所提出的模型架构涉及一个生成器、一个辅助编码器和一个判别器，它们都需要联合训练。您能否简要评论一下这种方法的计算开销（例如，训练时间、内存占用）与实验中使用的基于 VAE 的基线模型相比如何？
2. **结论**

我投票**强烈推荐接收**。这篇论文为一个重要的机器学习问题提供了一个新颖、有理论依据且经过经验验证的解决方案。这是一项高质量的贡献，写作精良，对其局限性保持透明，并由一系列强有力的实验所支持。我提出的几点次要问题旨在引发讨论，并不会削弱这项工作的整体卓越性。我相信这篇论文会引起社区的极大兴趣。

**📝rebuttal部分**

好的，这是我作为论文作者，针对前述审稿意见生成的Rebuttal。

感谢审稿人的细致评论，这些宝贵的意见帮助我们进一步提升了论文质量。我们对您的疑虑和建议逐一回应如下：

评论 1: 关于机制数量 (KG) 的选择

我们理解审稿人对机制数量

KG 选择方法的关注。您是对的，目前的方法依赖手动调整，如附录B.5 (1)(1)(1)(1)所述，我们提供了一个实用的验证流程。我们的实验（例如在MNIST中使用15个机制处理10个类别 (2)）表明，当

KG 大于真实机制数时，模型具有鲁棒性。自动化选择是一个很好的方向，但这超出了本文的当前范围。我们将在附录中补充关于 KG 选择鲁棒性的讨论。

评论 2: 对更复杂因果结构的可扩展性

我们理解审稿人对模型向更复杂因果结构扩展性的疑问。这是一个极具价值的建议。诚如审稿人所言，扩展到顺序复合的机制是一个重要的研究方向，正如我们在结论部分的未来工作 3 中所讨论的。这可能需要对先验和可识别性证明进行非平凡的修改，是我们未来研究的重点。我们将在结论部分更明确地强调这一未来研究方向。

评论 3: 变换机制 (MT) 的性质

我们理解审稿人关于变换机制

MT 普适性的问题。我们的理论框架（例如，模型 x=g(zG,zT) (4)）本身是通用的，并未对变换的局部性或全局性做出限制。尽管我们的实验主要关注全局变换以便于和先前工作对比，但该框架有潜力捕捉更抽象或局部的变换。我们将在实验部分简要说明我们框架的通用性。

评论 4: 计算成本

我们理解审稿人对计算成本的关心。审稿人观察正确，我们的模型包含生成器、编码器和判别器 (G, E, D) ，相比标准VAE (通常是编码器和生成器) 增加了一个判别器，因此训练开销略高。我们认为，通过引入判别器获得的更优分布匹配和最终实现的显著性能提升 6，证明了这一额外成本是合理的。我们将在附录中增加关于计算成本的简短讨论。

再次感谢您的宝贵意见，我们乐意根据反馈进一步完善论文。

**💎模型review部分**

**Qwen3-235B-A22B-2507**

论文审阅意见

论文概述

本论文《Identifying Coarse-grained Independent Causal Mechanisms with Self-supervision》提出了一种基于退化混合先验(degenerate mixture prior)的生成模型，用于学习粗粒度解耦的独立因果机制。作者证明了该模型在特定条件下是可识别的，并提出了一种自监督训练方法来实现这一目标。论文在MNIST、FashionMNIST和Sprites数据集上进行了实验验证，表明该方法在环境转移下能提高下游预测模型的准确性。

## 主要贡献

1. \*\*创新的模型设计\*\*：提出使用退化混合先验来学习解耦机制，解决了传统非退化混合先验的两个关键问题——单个潜在变量必须编码多种变化类型，以及干预潜在变量可能导致数据样本从一个混合成分转移到另一个。

2. \*\*可识别性理论证明\*\*：在退化混合先验条件下，证明了生成模型在三种保持粗粒度解耦的变换下是可识别的，这是一个重要的理论贡献。

3. \*\*实用的自监督训练方法\*\*：提出了一种比竞争训练方法更易实现的自监督训练方法，不需要额外的标记数据或辅助数据集。

4. \*\*全面的实验验证\*\*：通过潜变量空间遍历和下游分类任务的准确性评估，证明了所提方法的有效性。

## 优点

1. 问题定位准确，针对解耦表示学习中的可识别性问题提出了有价值的解决方案。

2. 理论分析严谨，提供了清晰的可识别性证明，弥补了现有工作中可识别性分析的不足。

3. 退化混合先验的设计巧妙，有效解决了机制解耦的关键挑战。

4. 实验设计合理，包括合成数据实验、定性可视化和定量评估，多角度验证了方法的有效性。

5. 与现有工作相比，该方法不需要额外的监督信号或辅助数据集，更具实用性。

## 需要改进的地方

1. \*\*粗粒度解耦的实际影响\*\*：论文承认其解耦是"粗粒度"的（同一机制内的潜在变量可能仍然纠缠），但没有充分讨论这种粗粒度解耦在实际应用中的局限性。例如，在需要精细控制的生成任务中，这种局限性可能会影响性能。建议作者更详细地分析粗粒度解耦在不同应用场景中的影响。

2. \*\*机制数量的选择方法\*\*：论文附录B.4中提到选择机制数量的方法是通过潜变量空间遍历来观察，这在实际应用中可能不够系统化，尤其对于复杂数据集。建议提供更自动化的机制数量选择方法或理论依据。

3. \*\*可逆性假设的实践可行性\*\*：可识别性证明依赖于生成模型的可逆性，但论文没有充分讨论在实践中神经网络如何保证这种可逆性，以及当可逆性不完全满足时对结果的影响。建议提供更多的实证分析来验证模型的可逆性程度。

4. \*\*与最新解耦方法的比较\*\*：虽然与VAE、β-VAE等进行了比较，但建议增加与2021-2022年最新解耦方法的对比，如Dittadi等人关于解耦表示在现实环境中迁移的工作。

5. \*\*计算效率分析\*\*：论文缺乏对所提方法计算复杂度和训练时间的分析，这对于评估方法的实用性很重要。建议添加与其他基线方法在计算资源消耗方面的比较。

6. \*\*应用场景的局限性\*\*：论文假设数据遵循混合分布，每个混合成分由特定机制生成，但没有充分讨论这种方法在不符合此假设的数据集上的表现。建议讨论方法的适用边界和可能的扩展方向。

## 具体建议

1. 在实验部分，建议增加对粗粒度解耦局限性的定量分析，例如比较在需要精细控制的任务中，该方法与完全解耦方法的性能差异。

2. 建议在附录中提供更详细的消融实验，分析损失函数中各组件对最终性能的贡献。

3. 论文提到在MNIST(R)数据集上，当环境转移过大时，所提方法的优势会减弱（第22页）。建议深入分析这一现象，并讨论可能的解决方案。

4. 在方法部分，建议更清晰地解释退化混合先验如何避免传统方法中的问题，可以考虑添加更直观的图示说明。

5. 建议讨论所提方法在更复杂、高维数据集（如ImageNet）上的可扩展性。

## 总体评价

这是一篇具有重要理论贡献和实用价值的研究论文。作者成功地解决了独立因果机制学习中的可识别性问题，并提出了一个不需要额外监督信号的实用解决方案。尽管存在一些需要改进的地方，但论文的创新性和实证结果使其成为因果表示学习领域的一个有价值的贡献。我建议在解决上述问题后接受该论文。

## 推荐意见

小修后接受（Minor Revision）。作者应针对上述问题进行澄清和改进，但不需要进行重大修改。

**📝rebuttal部分**

感谢审稿人的细致评论，帮助我们提升论文质量。

Comment 1: 我们理解审稿人关注粗粒度解耦的实际影响。根据Section 6.3和Appendix C.5，粗粒度解耦足以分离关键机制（如MNIST中数字类型与样式），环境转移下下游任务准确率仍提高9.75%。实验表明粗粒度解耦已保留环境鲁棒性益处。将在Section 7补充适用边界讨论。

Comment 2: 我们理解审稿人对机制数量选择方法的关注。Appendix B.4已提供实用流程：通过潜变量遍历观察变化类型，机制数≥真实数时模型能容忍差异（如MNIST用15机制处理10类）。将在Section 4.1详细说明选择方法。

Comment 3: 我们理解审稿人对可逆性假设的关切。Section 4的自监督训练通过损失函数L中的λG和λT项确保z≈E(G(z))，Appendix C.1显示移除这些项会导致解耦失败。将在Section 3.2补充神经网络实现可逆性的讨论。

Comment 4: 我们理解审稿人希望与最新方法比较。我们已与Ada-GVAE等比较，Dittadi等人(2021)设置不同（关注域转移解耦而非机制识别）。将在Related Work添加针对性对比分析。

Comment 5: 我们理解审稿人对计算效率的关注。Appendix B.2显示训练时间仅比VAE多15%，自监督方法比竞争训练更高效。将在实验部分添加计算效率比较表格。

Comment 6: 我们理解审稿人对应用场景局限性的关注。Appendix C.5分析了环境转移过大时性能下降情况。将在Section 7扩展适用边界和可能扩展方向的讨论。

我们乐意根据反馈进一步完善论文。

**grok3 等待时间比较长**

**💎模型review部分**

以下是对会议论文《通过自监督识别粗粒度独立因果机制》的审稿意见，基于提供的文档内容。审稿包括总结、优点、缺点、改进建议和总体评价。我已将文档摘录中的信息融入审稿，确保内容清晰且自洽，供作者参考。审稿后，我预计会有一个反驳阶段，作者可以回应我的意见，我们将进一步讨论以完善论文。

**对《通过自监督识别粗粒度独立因果机制》的审稿意见**

**总结**

该论文提出了一种新颖的生成模型，称为ICM（独立因果机制）模型，旨在使用自监督方法识别数据中因果机制的解纠缠表示。作者引入了“粗粒度解纠缠”，即同一机制内的潜在变量可能纠缠，但不同机制间的变量独立。该方法使用退化混合先验分布学习这些机制，无需标记或辅助数据。论文提供了可识别性的理论保证，并通过特定条件下的证明支持。模型在合成数据和真实数据集（MNIST、FashionMNIST和Sprites）上评估，与VAE、β-VAE、Ada-GVAE和VaDE等基线相比，在环境变化下的下游分类精度提高了高达75%。

**优点**

* **重要问题领域：** 解决因果表示学习中的关键问题，对构建鲁棒且可解释的AI系统至关重要。
* **创新方法：** ICM模型通过退化混合先验和粗粒度解纠缠提供新视角，避免依赖标记数据。
* **理论严谨性：** 可识别性证明（如定理4）为模型主张提供坚实基础。
* **实用性：** 自监督方法使其可扩展，适用于标签稀缺的现实数据集。
* **令人信服的结果：** 实验显示对基线的显著改进，定量指标（表1）和定性可视化（图6和图7）支持有效性。

**缺点与改进建议**

1. **粗粒度解纠缠的定义模糊：**

* **问题：** “粗粒度解纠缠”是核心贡献，但定义和与标准解纠缠的区别不清晰（第3页）。
* **建议：** 提供明确定义和简单示例（如MNIST中分离数字身份和风格），说明其必要性。

1. **退化混合先验的解释不足：**

* **问题：** 退化混合先验对可识别性至关重要，但解释不足，图3对比未充分说明其作用（第5-6页）。
* **建议：** 详细阐述“退化”属性并与第3节理论结果关联。

1. **下游模型选择缺乏依据：**

* **问题：** 实验使用GBT（第13页），但未解释选择理由或验证普适性。
* **建议：** 解释GBT选择并测试其他模型（如神经网络）以展示普适性。

1. **数据生成过程假设的局限性：**

* **问题：** 模型假设两阶段确定性过程和机制独立（第4页），未探讨假设违反的影响。
* **建议：** 讨论假设现实性并加入敏感性分析或局限讨论。

1. **环境变化细节不足：**

* **问题：** 环境变化是关键指标，但正文中描述不足（第13页），附录图14提供部分信息。
* **建议：** 在正文中描述变化模拟方式并解释其测试解纠缠的有效性。

1. **与相关工作的比较不足：**

* **问题：** 相关工作部分（第10页）缺乏详细比较以突出ICM进展。
* **建议：** 扩展此部分，添加表格或段落对比ICM与先前方法，强调独特贡献。

**总体评价**

该论文在因果表示学习领域提出有前景的贡献，方法新颖、理论严谨、结果令人信服。解决缺点——澄清概念和增强实验背景——将提升影响力。建议\*\*修订后接受\*\*，等待反驳阶段回应。

我期待作者的反驳，以澄清这些问题并讨论潜在修订。这种对话将有助于完善论文，确保其贡献被充分理解并提升说服力。

**📝rebuttal部分**

你好！根据你的要求，我基于审稿人评论生成了以下rebuttal，回应了所有关键问题。以下是完整内容：

**开头：**

感谢审稿人的细致评论，帮助我们提升论文质量。

**Comment 1: 粗粒度解纠缠的定义模糊**

我们理解审稿人关注粗粒度解纠缠概念的清晰度。粗粒度解纠缠允许同一机制内的潜在变量纠缠，但确保不同机制间的变量独立。例如，在MNIST中，数字身份（机制）与风格（变量）分离，但风格内变量可纠缠。Section 2.1已简要说明，但我们将进一步澄清。我们将在修订稿Section 2.1添加具体示例。

**Comment 2: 退化混合先验的解释不足**

我们理解审稿人希望更详细解释退化混合先验。退化混合先验强制每个样本由单一机制生成，确保机制分离。Section 3.1的Eq. (5)定义了先验，Figure 3展示了其与非退化先验的对比。我们将在正文中更明确地阐述其作用。我们将在修订稿Section 3.1补充解释。

**Comment 3: 下游模型选择缺乏依据**

我们理解审稿人关注GBT选择的合理性。GBT因其对表格化潜在数据的适用性而选，符合Locatello et al. (2019, 2020)标准。我们将在Camera-ready版本中补充神经网络实验，以展示普适性。我们将在Camera-ready版本中补充实验。

**Comment 4: 数据生成过程假设的局限性**

我们理解审稿人关注假设的现实性。假设基于Peters et al. (2017)，适用于多种场景。我们将在Section 4讨论假设违反时的影响，并计划在未来工作中探索相关性机制。我们将在修订稿Section 4添加讨论。

**Comment 5: 环境变化细节不足**

我们理解审稿人希望更详细的环境变化描述。环境变化模拟了笔画粗细和旋转变化，测试模型对机制变化的鲁棒性。Figure 14展示了变化示例。我们将在正文中简要描述。我们将在修订稿Section 5.1补充描述。

**Comment 6: 与相关工作的比较不足**

我们理解审稿人希望更detailed的比较。我们将在Section 1.1添加表格，对比ICM与非线性ICA和解纠缠方法，突出ICM的独特贡献，如自监督和粗粒度解纠缠。我们将在修订稿Section 1.1添加表格。

**结尾：**

我们乐意根据反馈进一步完善论文。

**📝rebuttal**

|  |
| --- |
| 🌟rebuttal所用的prompt  你是一名经验丰富的学术论文作者，正在准备对已有的审稿人评论进行rebuttal。你的目标是专业、建设性地回应每个问题，以争取论文被接收。请严格遵循以下规则生成rebuttal：  输入信息（用户将提供）  - 审稿人评论：[在此粘贴审稿人的完整评论，包括所有编号点]  - 论文摘要：[在此粘贴论文摘要]  - 关键证据：[在此粘贴与评论直接相关的论文段落、图表编号或实验数据（例如："Section 3.2, Figure 4"）]  - 修改计划：[可选：说明你已/将如何修改论文，例如"将在Camera-ready版本中补充实验" ]    rebuttal生成规则  1. 整体要求：  - 语气：礼貌、尊重、积极（例："感谢审稿人指出..."而非"审稿人错误地认为..."）。  - 长度：每个评论点的回应≤150字；总rebuttal≤500字。  - 禁止：编造未提供的证据、攻击审稿人、回避问题或过度承诺修改。    2. 结构化回应（对每个审稿人评论点）：  - 步骤1: 确认理解（20字内）：简要重述评论，表明你认真阅读了（例："我们理解审稿人关注[具体问题]..."）。  - 步骤2: 核心回应（80字内）：  - 若评论正确 → 承认并说明修改计划（例："您是对的，我们将在Section 4添加讨论..."）。  - 若评论是误解 → 用\*\*具体证据\*\*澄清（例："根据Section 3.1的Eq. (5)，[解释]；这与评论中假设不同，因为..."）。  - 若评论是建议 → 评估可行性（例："我们将在未来工作探索此方向，但当前版本受限于[原因]..."）。  - 步骤3: 行动承诺（20字内）：明确论文修改动作（例："已在修订稿中更新Figure 3"）。    3. 全局结构：  - 开头：1句感谢（例："感谢审稿人的细致评论，帮助我们提升论文质量。"）。  - 主体：按审稿人评论编号逐点回应（格式：Comment X: [回应]）。  - 结尾：1句开放性承诺（例："我们乐意根据反馈进一步完善论文。"）。    现在生成rebuttal  基于以下用户输入，输出纯文本rebuttal（无需额外解释）：  审稿人评论：  [用户粘贴评论]  论文摘要：  [用户粘贴摘要]  关键证据：  [用户粘贴证据]  修改计划：  [用户粘贴修改计划] 中文输出回答  同步自文档: <https://uniquestudio.feishu.cn/docx/UlivdH2Bmoj1Cbxij2jcyDYnnkd#FD4idQNG4sx5uKbNEfxc0KVQntf> |

**🔍问题分析**

**问题1 openreview上的置信度不高，大模型基本没有考虑这点，反而从文章内容入手**

**问题2 大模型review的时候会对文章内容评价过高，基本不存在偏激的评价，可能与事实不符**

**问题3 大模型review的时候过于机械化，比较生硬，并且没有打分或者打分与事实不一样**

**🚀未来计划**

|  |
| --- |
| **目标与计划** |

错误分类

逻辑谬误： LLM 经常误解任务的逻辑要求 。例如：

转移话题（Red Herrings）： 误导性推论，以弥补真实论点的缺失 。

稻草人谬误（Strawman Fallacy）： 歪曲或夸大对手的论点，使其更容易被攻击 。这在反驳中尤其具有欺骗性。

草率概括（Hasty Generalization）： 基于小样本得出宽泛结论 。

虚假两难（False Dilemma）： 在存在更多选项的情况下只提出两个选项 。

诉诸权威/情感/传统/大众（Appeal to Authority/Emotion/Tradition/Popular Belief）： 依赖非逻辑的说服策略 。

循环论证（Begging the Question）： 使用仅被假定为真的前提来论证某事为真 。

事实不准确/幻觉：

生成无根据或不正确的主张 。幻觉可能源于基础模型中数据的缺失、领域专业知识的不足或模型在处理更长上下文时的性能下降 。

即使从事实正确的信息来源获取数据，如果对上下文理解错误，也可能误解上下文 。

如果所需信息不在训练集中，LLM 可能会凭空捏造答案 。

连贯性和一致性问题：

缺乏逻辑流畅性，出现矛盾陈述 。

在联想或语义关系方面存在困难 。

尽管被认为有效，但在维护连贯辩论方面仍存在困难 。

重复和冗余：

严重的重复和质量下降，尤其是在生成较长文本时 。

过于频繁地重复短语 。

误解上下文/提示：

未能完全理解提示的上下文，导致偏离主题或不相关的反驳 。

在代码生成中生成“垃圾代码”或“无意义片段”，这类似于论证性文本中的不相关内容 。

在理解高级指令方面存在困难，特别是较小的模型 。

偏见和毒性：

从网络来源的训练数据中习得固有偏见 。

生成带有偏见或毒性的内容 。

风格限制：

难以匹配人类的风格变化，偏爱特定的语法结构（例如，现在分词短语、名词化） 。这可能使反驳听起来不自然或机械化。