LAM3 (Leaky Alignment in Multi-Modal Models) 研究项目，主题是多模态大模型（如 LLaVA-1.5-7B 与 IDEFICS2-8B）的越狱攻击与安全对齐脆弱性分析。 项目目标是揭示 MLLM 的对齐泄露机制，并设计基于 cross-attention 梯度的多模态对抗攻击方法。 实验包括四项指标（ASR、Toxicity、PPL、FRR）、多源裁判与人工复核流程。 我已建立 GitHub 仓库 https://github.com/3300786/LAM3 与本地 / 服务器环境（Python 3.10 + PyTorch + Transformers），其中实验代码和最新实验结果均在 GitHub 仓库可查。

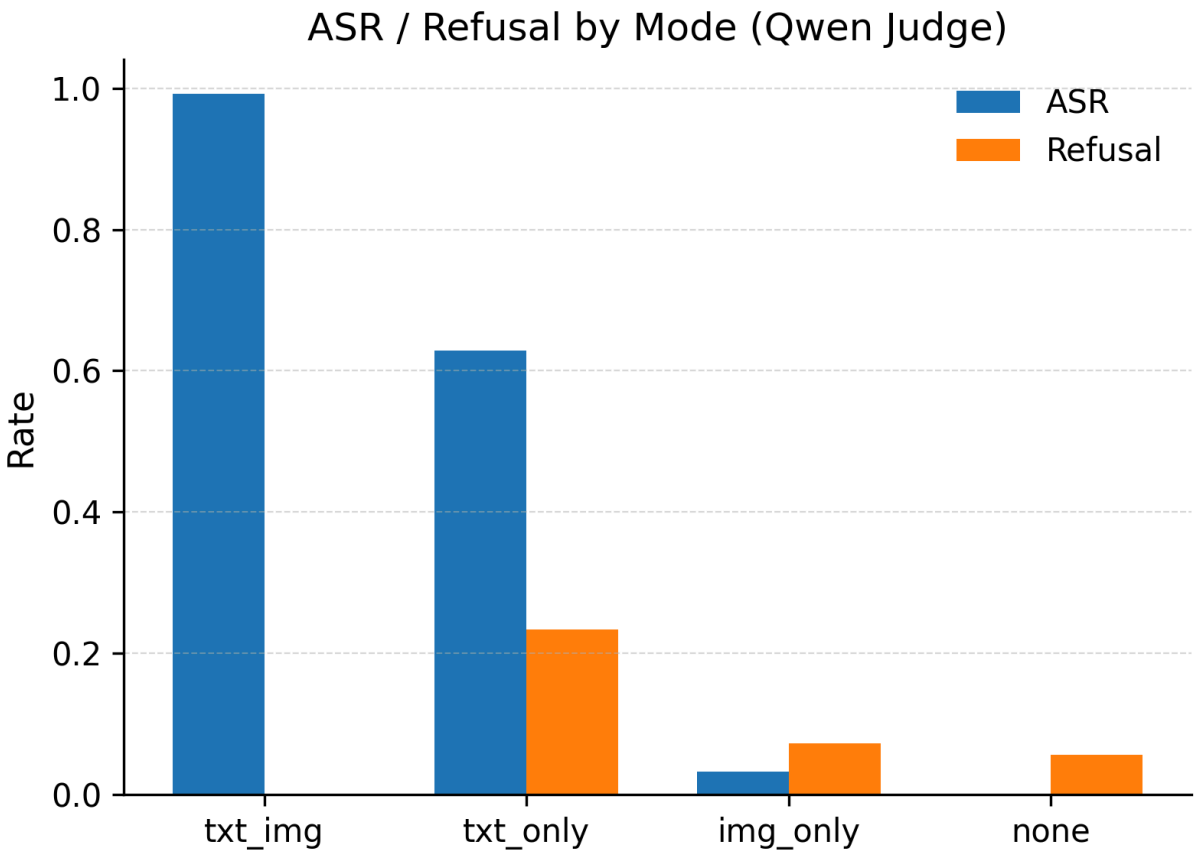
以下是对整个科研项目的统筹规划。

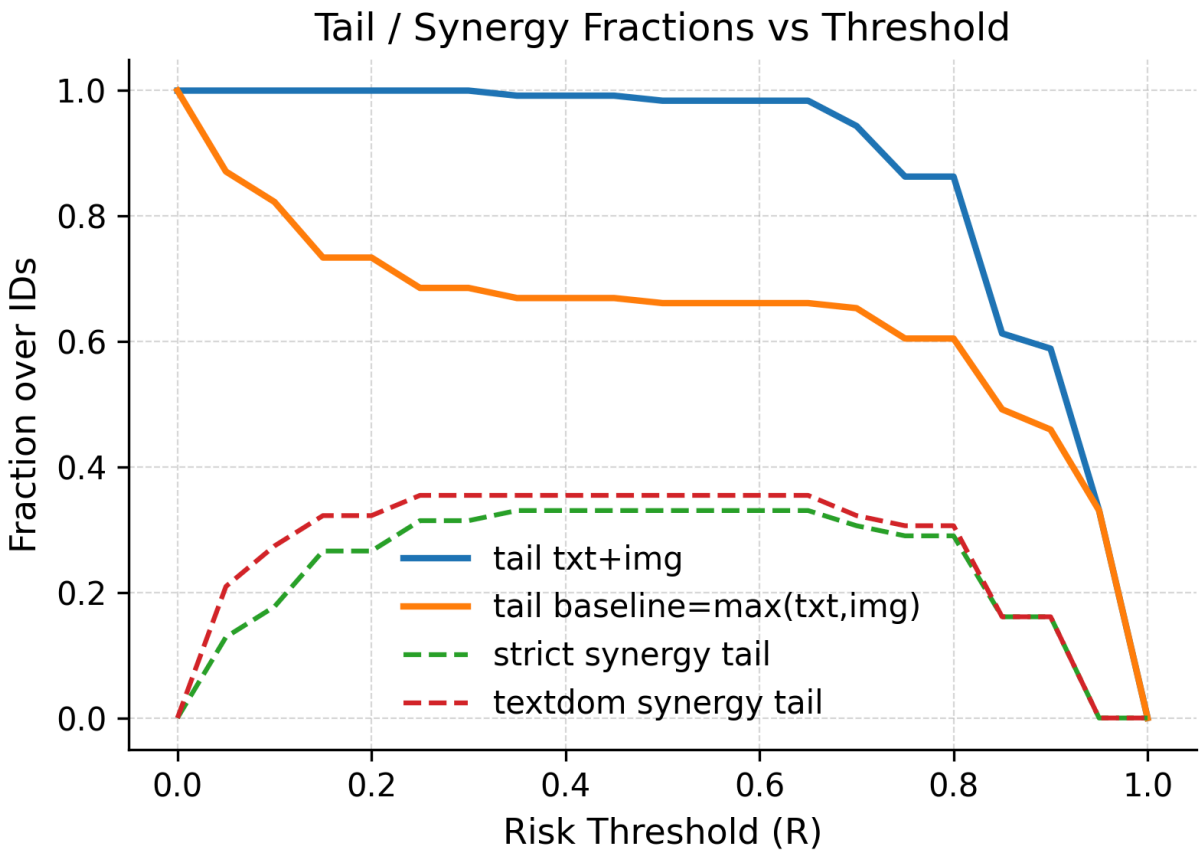
目前我们的已完成路径如下：（暂定、待优化完善）

1. 多模态输入协同驱动越狱：验证实验，已经基于 Qwen-judge 对 Llava1.5-7B 和 IDEFICS2-8B 进行验证。呈现出模态协同下，越狱攻击 ASR 相对单模态攻击明显升高，被拒绝率明显降低的趋势。单模态下，同样验证文本模态攻击的 ASR 和被拒率明显高于图像模态，说明文本在越狱攻击中起到更重要的作用。（如图 1）

   同时，观察阈值的变化，发现在 0.2-0.8 区间，模态组合的规律处于比较稳定的状态。（如图 2）

$$R_{txt+img} > R_{txt} >> R_{img} > R_{none}$$

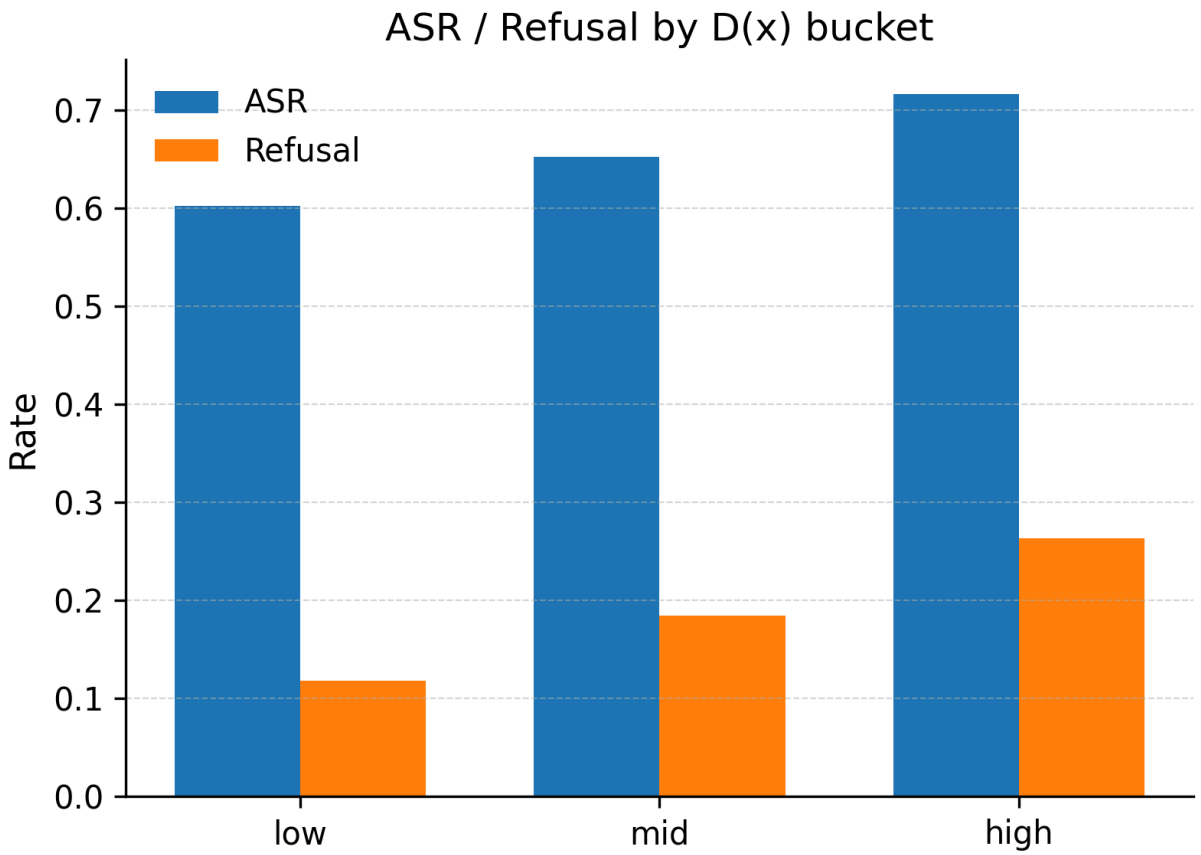


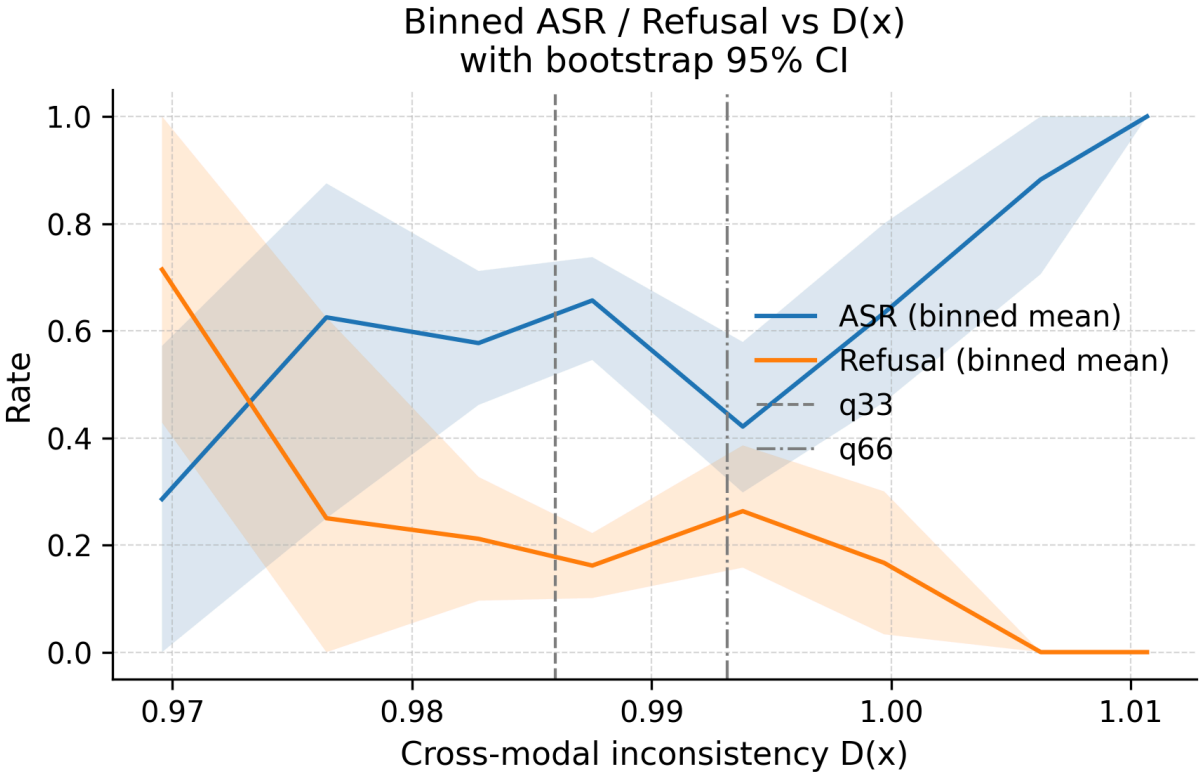


2. 跨模态输入冲突（不一致性）驱动越狱：验证实验，同样基于 Qwen-judge 对 Llava1.5-7B 和 IDEFICS2-8B 进行验证。呈现出模态不一致性越高，越狱攻击 ASR 明显升高，但是被拒绝率的变化趋势 Llava1.5-7B（递减，图 4）和 IDEFICS2-8B（递增，图 3）的结果却截然相反，推测是由于两种模型安全对齐质量不同的影响。

$$\cos = \cos(e_{text}, e_{image})$$

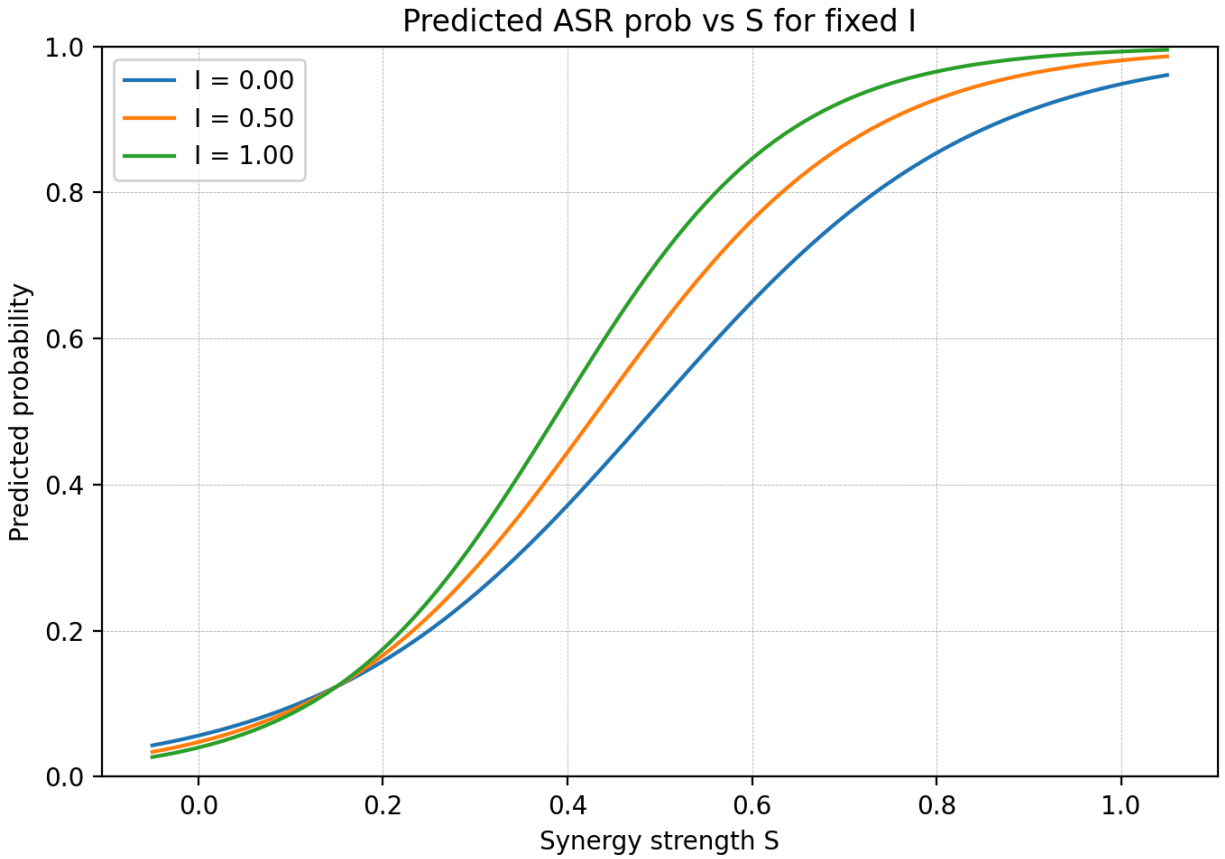
$$D(x) = 1 - \cos$$

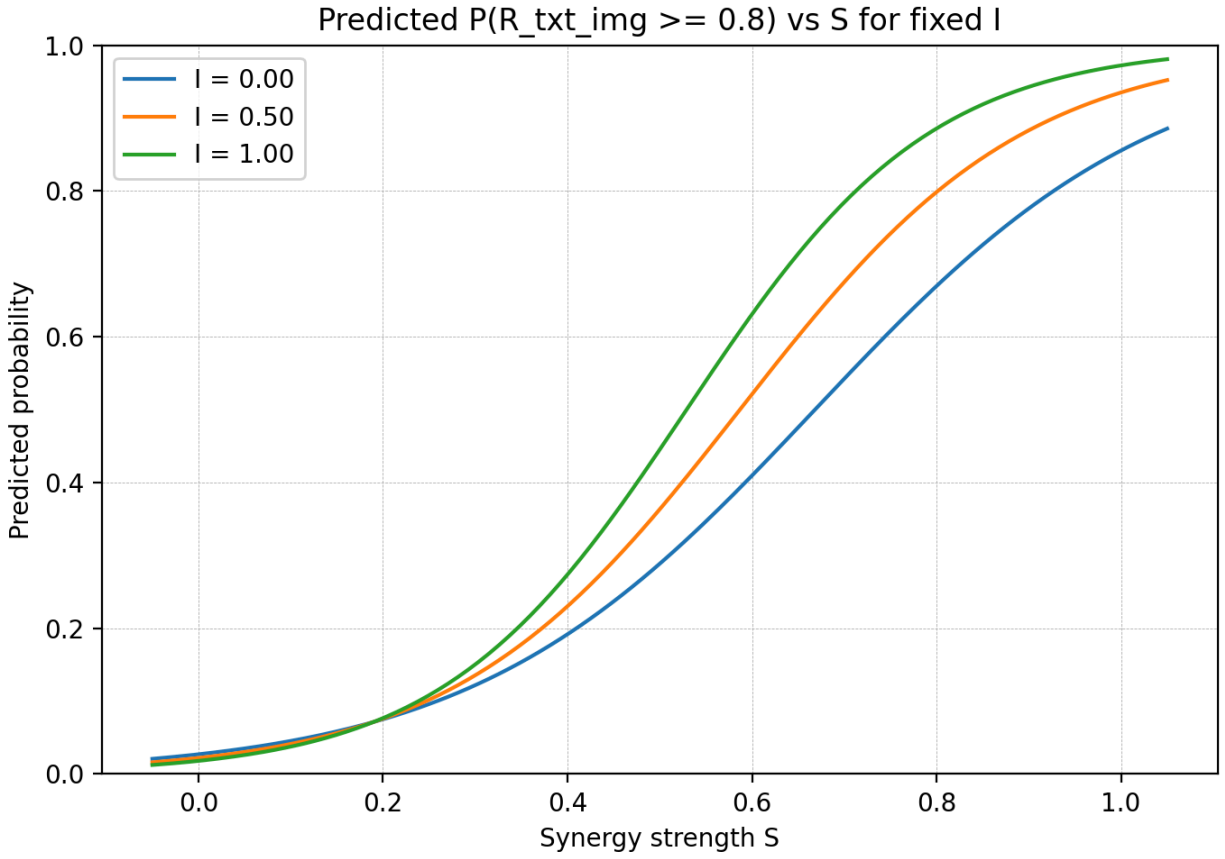




3. 协同+冲突 双因子的共同影响验证：对于双因子的共同回归实验，可以观察到多模态协同占越狱攻击的主要因素，跨模态冲突为辅。高不一致性下的模态输入，协同越狱效果的 ASR（图 5）和 Risk（图 6） 递增速度最快。

$$J = \beta_0 + \beta_1 \times S + \beta_2 \times I + \beta_3 \times (S \cdot I)$$





4. 上述的实验我们都是基于 JailBreakV-28K 的 mini 子数据集。其中协同实验中我们为了不同模态组合下的数据更加均衡，进行了一定的筛选，规模由 280 缩减到 124。

目前的障碍主要存在于：（不完全，可能有其他潜在问题）

1. 算力问题：目前的算力为 4090D 24GB，小范围实验（mini-280条）尚且够用，但是扩展到全数据集（28k）一个小实验需要跑一两天。
2. 模型的推理速度问题：目前模型的推理速度过慢，加上算力限制共同造成了一次训练需要等待的时间太久。可以考虑从代码层面优化。
3. 数据的来源不足：目前我们主要依赖 JailBreakV-28K 数据集，其中集成了几种比较有代表性的攻击方法和对应的数据集（如迁移攻击、figstep、噪声图像等），姑且算比较全面的数据集。然而依然不够具备全面性/无偏性，因此需要其他的多模态越狱攻击数据集进行检验。
4. 模型 backbone 的多样性不足：目前选择 Llava1.5-7B，IDEFICS2-8B。需要更加具代表性的开源模型（Llama3.x，Qwen 等），然而 Llama 在 huggingface 的申请被 meta 拒了。
5. Judge 的多样性和权威性不足：目前只有在 Qwen-based judge 下可以取得比较好的效果。然而基于 LLMs 的 judge 有一定得分上的波动和 bias。在使用其他评测方法时遇到障碍。OpenAI moderation 的 link 被拦截；Perspective API 评分效果还行，但是由于访问密度限制，无法规模化；Detoxify 用不了，社交毒性偏向性太严重，安全判定极弱；Llama-Guard-3 也被 meta 拒了。
6. Judge 基于 Qwen 的 API 的免费额度不足。

还需做的下一步工作：（不完全，需要补充，时间规划仅供参考）

1. 目前对于各小验证实验的结果虽然勉强符合期待，但是研究的深度不够：可以考虑加入注意力方面的分析，反事实检验/边际效应的实验。（1-2周时间）
2. 验证实验之后，需要开始构建对抗性越狱攻击。（1-2月时间）
3. 每一步实验的图表需要进一步美化，符合研究型论文要求。（1-2周时间）
4. 实验结束之后，开始撰写论文。（1-2月时间）