ISCAS 中国科学院软件研究所学术年会
暨重点实验室科技活动周 2025 第十届

学术论文

见微知著：消除单模态虚假关联的多模态奖励模型泛化方法

The Devil Is in the Details: Tackling Unimodal Spurious Correlations for Generalizable Multimodal Reward Models

李梓超，温学儒，娄杰，季雨秋，陆垚杰，韩先培，张德兵，孙乐

Forty-Second International Conference on Machine Learning (ICML 2025)

李梓超 (lizichao2022@iscas.ac.cn)

研究背景

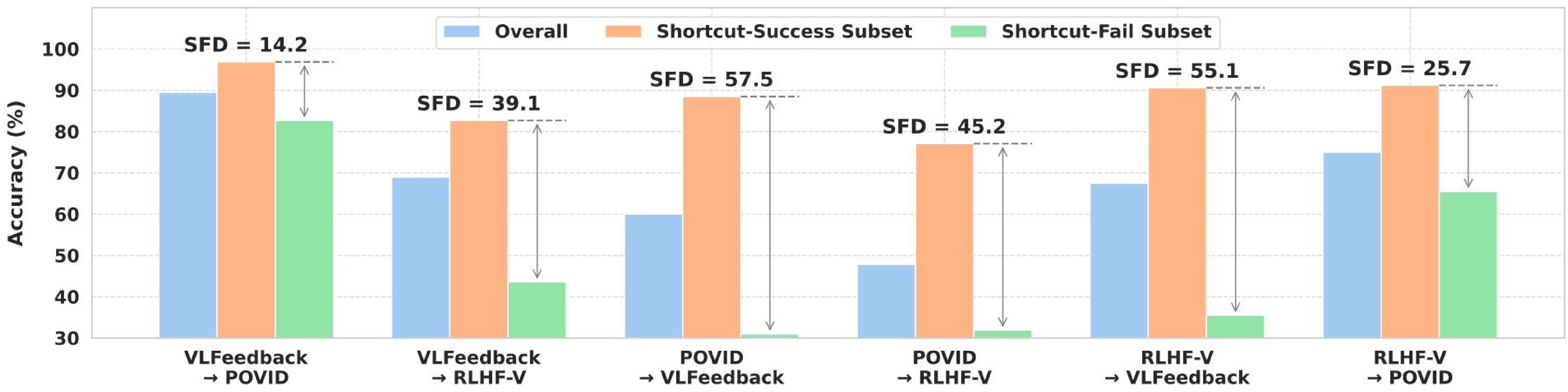
- 多模态奖励模型（MM-RM）在大语言模型偏好对齐中扮演关键角色，而泛化能力至关重要

核心贡献

- 识别单模态虚假关联关键挑战
- 提出捷径感知多模态奖励模型算法
- 增强模型跨分布场景下泛化能力

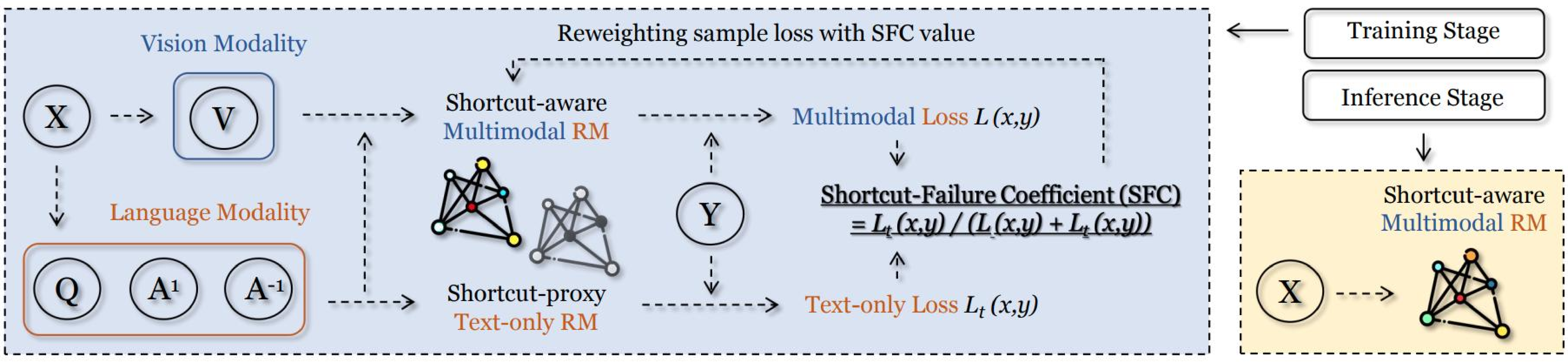
单模态虚假关联

- **纯文本捷径（Text-only Shortcut）**：模型过度依赖与训练标签相关但无法泛化的纯文本特征，阻碍了多模态奖励函数的鲁棒学习



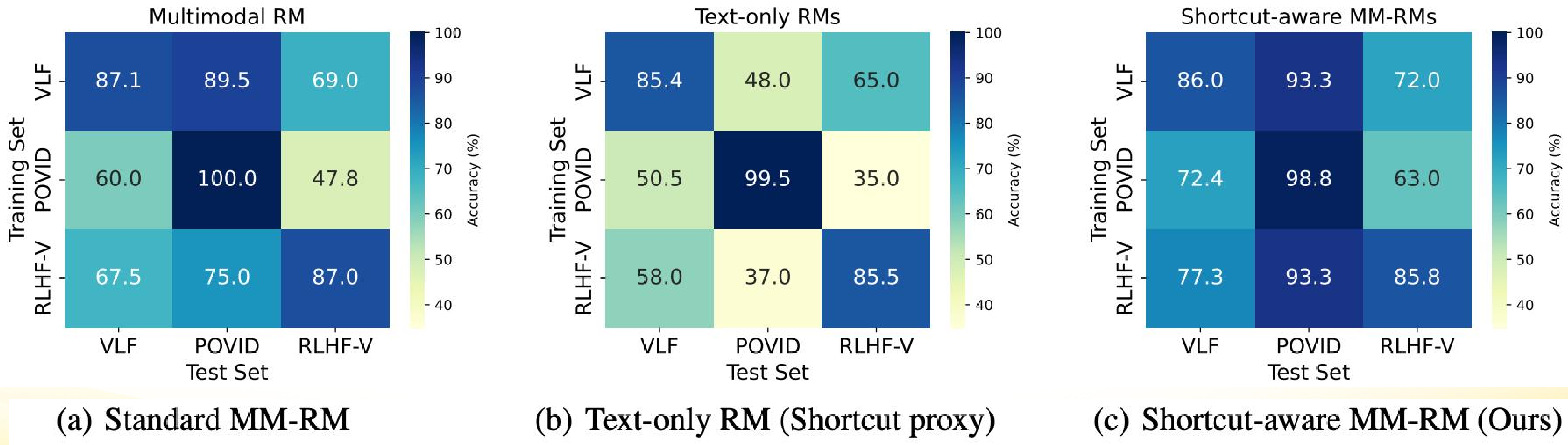
捷径感知学习算法

- 通过识别捷径失效的样本，动态赋予样本权重，**强调多模态理解**



跨分布实验框架

- 捷径感知学习算法有效提升了奖励模型在**分布外（OOD）的泛化能力**



(a) Standard MM-RM

Training Set \ Test Set	VLF	POVID	RLHF-V
VLF	87.1	89.5	69.0
POVID	60.0	100.0	47.8
RLHF-V	67.5	75.0	87.0

(b) Text-only RM (Shortcut proxy)

Training Set \ Test Set	VLF	POVID	RLHF-V
VLF	85.4	48.0	65.0
POVID	50.5	99.5	35.0
RLHF-V	58.0	37.0	85.5

(c) Shortcut-aware MM-RM (Ours)

Training Set \ Test Set	VLF	POVID	RLHF-V
VLF	86.0	93.3	72.0
POVID	72.4	98.8	63.0
RLHF-V	77.3	93.3	85.8