秦正

电话 & 微信: 17737732966 | Homepage | 谷歌学术 | qinzheng@stu.xjtu.edu.cn |
出生日期: 2000-07-13

教育背景

特伦托大学, 人工智能专业, 访问博士 | 导师:Nicu Sebe 教授	2025.09—2026.07
西安交通大学, 控制科学与工程专业, 博士 | 导师:王乐教授	2021.09—2026.09
哈尔滨工业大学, 机器人工程专业, 学士	2017.09—2021.06

实习经历

蚂蚁集团 | 多模态交互组 | 多模态大模型科研实习生 |*2025.03—2025.12*

1、 实习内容 1-多模态大模型：HumanSense：从多模态感知到共情情境回应 **AAAI26 一作**

主导设计了“以人为中心的感知-交互策略”多维度评测基准，并完成了从数据采集、清洗、标注到模型训练与评测的全流程闭环。**研究目的**: 面向真实人际交互场景，系统性提升多模态大模型对人类情感、隐含意图与社会语境的理解能力，并评估其是否能够生成合理、共情且情境匹配的响应。

- **构建 HumanSense 评测体系**：形成 3882 个样本、15 个任务; 提出四层金字塔评测框架（基础感知 → 社会关系理解 → 情感/意图推理 → 同理心响应生成），补足现有 MLLM Benchmark 对人际交互与情感因素覆盖不足。
- **系统评估主流多模态模型的“情感理解与推理瓶颈”**：对比不同模型在 L1–L4 任务表现，结论为基础感知接近饱和，而 **多模态整合与高层推理（L3/L4）**与人类差距明显，定位问题主要在推理/决策能力不足。
- **提出 Multi-stage、Modality-Progressive GRPO 训练范式**：选择 Qwen25-Omni 为 Baseline，将人机交互任务拆解为多阶段决策过程，引导模型按**视觉 → 听觉 → 全模态**逐步融合信息进行推理学习，提升复杂语境下推理深度与响应一致性。
- **提出 Prompt Enhancement 策略**：从 GRPO 训练中提炼稳定的推理结构并固化为提示模板，使未额外训练的模型在 HumanSense 上的情感理解与交互推理能力获得提升。

2、 实习内容 2-多模态大模型: MLLM 中的 ID-consistency 问题 **Ongoing**

聚焦视频大模型时序身份一致性 (ID-consistency) 缺陷，践行了“问题拆解—机理归因—评测验证—算法优化”的研究链路。**研究目的**: 提升 MLLM 在视频中对目标的 ID-consistency 能力。

- **ID-consistency 失效机理剖析与量化建模**。系统拆解视频人数统计任务，通过单帧对照实验将 MLLM 的核心瓶颈定位于跨时间“身份维持失败”。基于思维链（CoT）揭示模型错误归因机制，定位简单场景的“多数偏差”和复杂场景的“少数偏差”。并创新引入“人员动态变化熵”指标，从统计学层面验证了场景混乱度对模型准确率的负相关影响。
- **评测体系构建**：构建覆盖外观主导（电影场景）与位置主导（舞蹈 / 运动视频）的评测数据，设计了三个层面的任务，用于系统评估 MLLM 的 ID-consistency 能力
- **设计动态身份档案策略**: 引入结构化外部记忆, 引导 MLLM 实时记录并更新目标的唯一编号与细粒度外观特征，有效降低了复杂场景下的身份混淆率。
- **设计自适应分辨率感知**：通过 Context-learning 赋予模型自主评估感知难度的能力：若自主评估发现目标像素过小等因素，则动态决定调用大分辨率进行重编码。该自适应策略实现了计算资源的最优动态分配。

蚂蚁集团 | 数字人算法组 | 数字人算法科研实习生 |*2024.05—2025.02*

1、 实习内容 1-多模态视频生成/数字人: VersaAnimator: 多模态控制的说话人视频生成框架 **ACMMM25 一作**

研究目的: 实现“音频驱动 + 文本可编辑语义动作”的高保真全身说话人视频生成，并支持多尺度与任意画幅。

- **设计双分支多模态动作生成器**：引入 VQ-VAE 将异构的 3D 动作统一离散化为 Motion Tokens，并设计双分支 Transformer 架构。创新性地解耦动作控制逻辑：由音频信号控制基础肢体节律，由 Text Prompt 精准引导高语

义动作（如特定手势），生成的全维 3D Token 可适配任意画幅尺寸（含全身腿部动作）。

- **提出 Token2Pose 平滑映射策略**：针对 3D 到 2D 转换过程中常见的动作僵硬与失真问题，设计 Token-to-pose Translator，将 3D Motion Tokens 平滑映射为 2D 姿态序列。该设计有效缓解了跨维度转换的机械感，并显著增强了肢体运动的细粒度表现力。
- **设计多模态控制的 video diffusion 模型**：设计高保真渲染的两阶段训练范式。Stage 1 微调 SVD 骨干网络以强化面部特写细节；Stage 2 引入音姿双重控制，利用多尺度（头部至全身）同步数据联合训练。最终融合音频、图像与身份特征，实现语音精准驱动面部、2D 姿态引导肢体的高保真影视级视频生成。

2、 实习内容 2-动作生成: Diverse-T2M: 引入不确定性的多样 3D 动作生成 **TCSVT 一作** under review

研究目的: 解决现有生成模型难以兼顾“文本一致性”与“生成多样性”的痛点。基于 RVQ-VAE 与 Mask Transformer 重构底层链路，摒弃传统文本到动作的“一对一”强绑定，在保证语义约束的前提下，拓宽生成多样性的上限。

- **提出噪声驱动的不确定性建模范式**：首次将纯噪声信号作为多样性载体融入 Transformer 架构。赋予模型在相同文本指令下（如“假装擦拭物体”），智能推理并生成站立、弯腰等多种合理运动轨迹的能力。
- **设计变分隐空间与随机采样机制**：构建变分隐空间（Variational Latent Space），将文本平滑投影为高维概率分布。结合自研隐空间采样器（Latent Space Sampler），在推理阶段引入随机采样，实现动作序列的高质量拓展。

刷新 SOTA 与商业落地验证：在 HumanML3D 与 KIT-ML 数据集上维持顶尖文本对齐精度的同时，核心多样性指标（Multimodality）超越此前 SOTA 方法 30%。业界影响力: 算法在内部被部署为服务，并供 Galacean 引擎调用，可用于驱动蚂蚁庄园小鸡和数字人 Luna，部署同学反馈效果很好。

伊利诺伊大学芝加哥分校 | *Wei Tang* 教授团队 | 计算机视觉科研实习生 |*2022.05—2024.03*

1、 实习内容 1-视觉导航：RSRNav: 基于空间关系推理的图像目标导航方法 **TCSVT 一作** under review

研究目的: 解决视觉导航 agent 用语义特征**无法提供准确方位信息**以及**训练/应用视角不一致导致**的性能下降问题;

- **设计“感知-关系-动作”导航新体系**：创新性地提出一种显式推理范式。通过构建目标与观测值之间的多尺度互相关矩阵，将黑盒特征提取升级为空间几何关系推理，为智能体提供鲁棒的导航指引。

  基于 PPO 的导航策略优化：采用 Actor-Critic 架构结合 PPO 强化学习算法进行端到端训练。融合空间互相关线索与历史状态，定制多层次复合奖励函数（结合步进中的距离/视角缩减奖励、控制路径冗余的时间惩罚，及位置 + 位姿双标准的终局奖励），精准引导模型高效收敛。

实验结果：在 Gibson、MP3D 和 HM3D 三大基准数据集上均超过 SOTA 性能。

2、 实习内容 2-多目标跟踪：MotionTrack: 基于长短期运动的多目标跟踪框架。 **CVPR23 一作**

研究目的: 面向人群密集与严重遮挡的复杂真实场景，解决多目标跟踪中频繁发生的 ID Switch（身份切换）问题,

- **构建数据驱动的社会力运动学模型**: 摒弃传统卡尔曼滤波的简单线性运动假设, 利用图神经网络实现社会力建模, 通过密集场景下目标间的避让与跟随行为，显著提升了目标短期轨迹预测的精度与物理合理性。
- **设计历史轨迹回溯机制**：针对严重遮挡导致的目标丢失重识别难题，提出长距离历史轨迹回溯策略。基于提取的位置特征，校验历史轨迹过渡至新目标的时空一致性，大幅降低了复杂场景下的 ID Switch 发生率。

该方法摆脱复杂 Re-ID 依赖，**在极低计算开销**与硬件友好约束下显著提升了目标身份保持能力。业界影响力: 该工作已获得 Google Scholar 210+ 次引用，其核心技术因低算力、易部署的特性受到大疆（DJI）自动驾驶团队邀请，并被认为具备在低功耗视觉芯片上的落地潜力。

3、 实习内容 3-多目标跟踪：GeneralTrack: 多应用场景统一跟踪框架 **CVPR24 一作**

研究目的: 旨在打破传统跟踪算法跨场景泛化性差、高度依赖人工先验融合特征的技术瓶颈。

- **探究跟踪失效底层机理**: 开展跨场景数据统计分析，量化评估了目标密度、运动复杂度和形变幅度对跟踪稳定性的影响，精准解析了限制现有算法泛化性的核心元素。
- **设计由点及面的像素级关联范式**: 创新性地引入了由点及面的关联机制，通过构建 4D 相关体捕捉像素级微观运动特征，彻底摆脱了对目标框尺度敏感的固有缺陷。

在 MOT17/20、DanceTrack 及 BDD100K 等**多个异构数据集上刷新 SOTA** 纪录，证明了极强的泛化能力。

创业项目

AR 景区互动小程序｜合作对象：洛阳市洛邑古城、丽景门 |*2023.11—2024.01*

具体内容: 开发了一个基于 AR 的景区互动小程序。用户打开后，摄像头识别景区大门或预设场景，就会触发 3D 凤凰特效，围着城墙飞舞；夜晚还可以在手机屏幕中看到我实现的烟花特效。更有趣的是，当扫描到特定贴图，图中的形象就会变成 3D 建模出现在屏幕里，游客可以和它合影。**结果**: 整个项目从与景区谈判、组建二人小型团队、打通技术链路，产品迭代测试、到最终版权售卖 1.8 万元，虽然金额不大，但让我积累了宝贵的商业需求洞察、谈判技巧、团队协作以及产品落地经验。

个人荣誉

- 国家奖学金（博），2025
- 潍柴动力奖学金（博），2025
- 优秀研究生（博），2023，2024
- 一等新生奖学金，2021
- 哈尔滨工业大学优秀毕业生，2021
- 一等学业奖学金，2018，2019，2020

科研成果

- MotionTrack: Learning Robust Short-term and Long-term Motions for Multi-Object Tracking. Qin Z, Zhou S, Wang L, Duan J, Hua G, Tang W. **CVPR2023**.
- Towards Generalizable Multi-Object Tracking. Qin Z, Wang L, Zhou S, Fu P, Hua G, Tang W. **CVPR2024**.
- Referencing Where to Focus: Improving Visual Grounding with Referential Query. Wang Y, Tian Z, Guo Q, Qin Z, Zhou S, Yang M, Wang L. **NIPS2024**
- RefDetector: A Simple yet Effective Matching-based Method for Referring Expression Comprehension. Wang Y, Tian Z, Qin Z, Zhou S, Wang L. **AAAI2025**
- Towards Precise Embodied Dialogue Localization via Causality Guided Diffusion. Wang H, Wang L, Qin Z, Wang Y, Hua G, Tang W. **CVPR2025**
- Versatile Multimodal Controls for Whole-Body Talking Human Animation. Qin Z, Zheng R, Wang Y, Li T, Zhu Z, Yang M, Yang M, Wang L. **ACM MM2025**
- HumanSense: From Multimodal Perception to Empathetic Context-Aware Responses through Reasoning MLLMs. Qin Z, Zheng R, Wang Y, Li T, Yuan Y, Chen J, Wang L. **AAAI2026**
- Spatial Matters: Position-Guided 3D Referring Expression Segmentation Wang Y, Tian Z, Wang L, Qin Z, Zhou S. **CVPR2026**
- Single-Shot and Multi-Shot Feature Learning for Multi-Object Tracking. Li Y, Zhou S, Qin Z, Wang L, Wang J, Zheng N. **TMM2024**
- Robust Noisy Label Learning via Two-Stream Sample Distillation. Bai S, Zhou S, Qin Z, Wang L, Zheng N. **TMM2025**
- Semantic and Kinematics Guidance for RMOT. Li Y, Zhou S, Qin Z, Wang L. **TMM2025**
- Injecting Position and Relation Prior for Dense Video Captioning. Li Y, Zhou S, Qin Z, Lin J, Sun X, Wu K, Wang L. (Submitted for **TIP**)
- From Mapping to Composing: A Two-Stage Framework for Zero-shot Composed Image Retrieval. Wang Y, Tian Z, Guo Q, Qin Z, Zhou S, Yang M, Wang L. (Submitted for **TCSVT**)
- Embracing Aleatoric Uncertainty: Generating Diverse 3D Human Motion. Qin Z, Wang L, Wang Y, Yang M, Rong C, Yang M, Zheng N. (Submitted for **TCSVT**)
- RSRNav: Reasoning Spatial Relationship for Image-Goal Navigation. Qin Z, Wang L, Wang Y, Zhou S, Hua G, Tang W. (Submitted for **TCSVT**)
- RAMP: Iterative Refinement and Adaptive Multi-granularity Perception for Embodied Dialog Localization. Wang H, Wang L, Qin Z, Zhou S, Hua G. (Submitted for **PR**)