Risheek Garrepalli

risheek1993@gmail.com | 541-908-5873

SUMMARY

Staff ML Researcher with 9+ years across generative modeling, RL, and foundation models.
Research grounded in RL and probabilistic modeling with representations and uncertainty as the lens across diffusion for vision and language, inference efficiency, and sequential decision making.
Published 20+ papers at ICML, NeurIPS, CVPR; shipped world-first on-device GenAI.

LINKS

LinkedIn: risheekg
Scholar: risheekg

CORE STRENGTHS

Foundation Models
- LLMs / MLLMs / VLAs
- Diffusion: Vision & Text
- World Models / Sequential Learning

Training / Inference
- Speculative / dLLM Sampling
- Distillation
- FSDP / ZeRo Sharding

Learning & Uncertainty
- RL / Imitation Learning / DAgger
- Structured Prediction
- Uncertainty / Robustness

SERVICE

Qualcomm Innovation Fellowship 2022–2025

Program Committee:

CVPR 2025 Workshop on Efficient Large Vision Models

Reviewer (~50 papers):

ICML, NeurIPS, ICLR, AAAI, CVPR, ECCV, ICCV, WACV

WORK EXPERIENCE

QUALCOMM AI RESEARCH | Staff Deep Learning Researcher

June 2021 -- Present | San Diego, CA

- **Diffusion Distillation (DDIL):** Identified trajectory diversity collapse as a core failure mode in distilled diffusion models. Proposed a **DAgger-based stable training framework** connecting covariate shift across sampling steps to imitation learning, unifying DMD2, LADD, and Clockwork. Shipped **world-first sub-0.6s Stable Diffusion on mobile**. *[The Verge]* | **MWC'23, Snapdragon Summit**.
- **LLM & dLLM Inference:** Led research on **speculative decoding** for autoregressive models and on-policy distillation for efficient decoding emphasis on deployment-grounded algorithmic wins over benchmark-only results. *currently working on speculation with distillation framework for dLLM Inference.*
- **World Models & Sequential Perception:** Developed E2E autonomous perception systems inspired by structured prediction and RL, including **Future Depth** (trained without teacher forcing) for temporally consistent sequential visual modeling and uncertainty-aware dense prediction.
- **Expressivity in Hard Regimes:** Worked on long-tail and difficult-to-amortize settings (**GeRO**, **RoCA**), motivated by the need for greater native expressivity in deep learning rather than relying on heuristics or superficial scaling.
- **Foundation Model Training & Inference:** Architected multi-node training pipelines (**FSDP/ZeRO**, model sharding) for large diffusion and foundation models.

OREGON STATE UNIVERSITY | MS / PhD Candidate, CS/AI

2016 -- 2021 | Advisors: Tom Dietterich & Alan Fern

- **Robustness & Uncertainty:** Developed open-set recognition, novelty detection, and calibrated uncertainty methods using VAEs, normalizing flows, contrastive learning, and deep ensembles.
- **RL & Planning Foundations:** Risk-sensitive RL, constrained MDPs, Bayesian inference, and structured prediction under advisors with deep roots in RL and sequential decision making.

PLAYSTATION (SIE) R&D | Reinforcement Learning Research Intern

Summer 2018 & Winter 2019 | San Mateo, CA

- Built **DeepMimic/GAIL-style** imitation learning pipelines for humanoid locomotion in Bullet simulation: MoCap retargeting, reference tracking, and adversarial training for stable bipedal gaits.
- Re-architected the simulation stack (ODE solver, first-order hold, PID interfaces) to decouple policy from motor and contact dynamics, achieving ~10× **reproducibility improvement** and more stable RL training.

SIRENA TECHNOLOGIES | Robotics Systems Engineer -- Controls

2014 -- 2016 | Bangalore, India

- Founding engineer for Nino, a consumer humanoid robot. Led end-to-end system design from concept to functional prototype.
- Designed kinematic structure (DoF allocation, joint placement via static load analysis), **EKF-based sensor fusion** (IMU, F/T sensors, encoders), and **3D Linear Inverted Pendulum** gait generation.

EDUCATION

MANIPAL UNIVERSITY | BE Electrical & Electronics

2014 | Focus: Control Theory

SELECTED PUBLICATIONS

1. Diffusion Models & Generative AI

How score matching and masked objectives shape representations — and what that buys in trajectory diversity, compositionality, distillation stability, and inverse problems

[1] **Risheek Garrepalli**, S. Mahajan, M. Hayat, and F. Porikli, "Improved diffusion distillation with **Imitation Learning**," *arXiv preprint arXiv:2410.11971*, 2024.

[2] **Risheek Garrepalli***, S. Kadambi*, S. Borse, M. Hyatt, and F. Porikli, "Madi: Masking-augmented diffusion with inference-time scaling for visual editing," *arXiv preprint arXiv:2507.13401*, 2025.

[3] A. Habibian, A. Ghodrati, N. Fathima, G. Sautiere, **Risheek Garrepalli**, F. Porikli, and J. Petersen, "Clockwork diffusion: Efficient generation with model-step distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition **(CVPR)***, 2024.

[4] S. Borse, S. Kadambi, N. P. Pandey, K. Bhardwaj, V. Ganapathy, S. Priyadarshi, **Risheek Garrepalli**, R. Esteves, M. Hayat, and F. Porikli, "Foura: Fourier low-rank adaptation," in *Thirty-Eighth Annual Conference on Neural Information Processing Systems **(NeurIPS)***, 2024.

2. Discrete Diffusion for Language

Contrasting token-level NTP with sequence-level diffusion objectives, how objective and denoising process reshape internal geometry and reasoning degrees of freedom — unlocking multi-token atomicity, redundancy exploitation, and richer sampling like Sequential Monte Carlo over AR

[1] S. Agrawal, **Risheek Garrepalli**, R. Goel, M. Lee, C. Lott, and F. Porikli, "Spiffy: Diffusion llm speculative decoding," *arXiv preprint arXiv:2509.18085*, 2025.

[2] R. Goel, **Garrepalli, Risheek**, S. Agrawal, C. Lott, M. Lee, and F. Porikli, "Skip to the good part: Representation structure & inference-time layer skipping in diffusion vs. autoregressive llms," *arXiv preprint arXiv:2603.07475*, 2026.

[3] J. Piskorz, C. Pinneri, A. Correia, M. Alfarra, **Garrepalli, Risheek**, and C. Louizos, "Masks can be distracting: On context comprehension in diffusion language models," *arXiv preprint arXiv:2511.21338*, 2025.

[4] Z. Qin, R. Goel, M. Gagrani, **Garrepalli, Risheek**, M. Lee, and Y. Sun, "Confu: Contemplate the future for better speculative sampling," *arXiv preprint arXiv:2603.08899*, 2026.

3. Vision-Language Agents & Reinforcement Learning

Iterative Proposal Refinement with dense finite horizon feedback and distribution matching in navigating the explore-exploit goldilocks toward richer functional priors over structured representations

[1] R. Yasarla, D. Hegde, S. Han, H.-P. Cheng, Y. Shi, M. Sadeghigooghari, S. Mahajan, A. Bhattacharyya, L. Liu, R. **Garrepalli**, *et al.*, "Generative scenario rollouts for end-to-end autonomous driving," *arXiv preprint arXiv:2601.11475*, 2026.

[2] R. Yasarla, S. Han, H.-P. Cheng, L. Liu, S. Mahajan, A. Bhattacharyya, Y. Shi, R. **Garrepalli**, H. Cai, and F. Porikli, "Roca: Robust cross-domain end-to-end autonomous driving," *arXiv preprint arXiv:2506.10145*, 2025.

[3] R. Yasarla, M. K. Singh, H. Cai, Y. Shi, J. Jeong, Y. Zhu, S. Han, R. **Garrepalli**, and F. Porikli, "Futuredepth: Learning to predict the future improves video depth estimation," in ***ECCV 2025***, Springer, 2025, pp. 440–458.

[4] R. **Garrepalli**, "Temporal difference learning for recurrent neural networks.," in *Conference on Cognitive Computational Neuroscience (CCN)*, 2019.

4. AI Safety & Uncertainty Quantification

Decoupling representational competence from information loss, and the downstream safety tradeoffs that follow

[1] S. Liu*, R. **Garrepalli***, T. Dietterich, A. Fern, and D. Hendrycks, "Open category detection with PAC guarantees," in *Proceedings of the 35th International Conference on Machine Learning **(ICML)***, 2018, pp. 3169–3178.

[2] R. **Garrepalli**, "Oracle analysis of representations for deep open set detection," *arXiv preprint arXiv:2209.11350*, 2022.

5. Computer Vision

Structured prediction and generative priors as a lens for efficient, uncertainty-aware CV: representations as the shared substrate

[1] R. **Garrepalli**, J. Jeong, R. C. Ravindran, J. M. Lin, and F. Porikli, "Dift: Dynamic iterative field transforms for memory efficient optical flow," in ***CVPRW 2023***, 2023, pp. 2219–2228.

[2] R. Yasarla, H. Cai, J. Jeong, Y. Shi, R. **Garrepalli**, and F. Porikli, *Mamo: Leveraging memory and attention for monocular video depth estimation*, 2023. arXiv: `2307.14336 [cs.CV]`. [Online]. Available: `https://arxiv.org/abs/2307.14336`.

[3] S. Borse, D. Das, H. Park, H. Cai, **Garrepalli, Risheek**, and F. Porikli, "Dejavu: Conditional regenerative learning to enhance dense prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition **CVPR***, 2023, pp. 19 466–19 477.