The Role

My client is looking for someone who can rigorously evaluate the quality of generated outputs, compare model performance against alternative approaches, and validate whether simulation assets are effective for robotics training. This includes running benchmarks, building evaluation infrastructure, performing inference across multiple models, and teleoperating robotic hands to validate real-world manipulation performance.

This role will play a critical part in maintaining quality standards across the organization by identifying issues, quantifying performance gaps, and ensuring systems meet a high technical bar.

What You'll Do

Benchmarking & Evaluation

- Design and execute benchmarks to measure the quality of 3D generation, reconstruction, and articulation outputs
- Compare internal pipelines against state-of-the-art and open-source alternatives using both standard and custom metrics (geometric accuracy, physical plausibility, texture quality, articulation correctness)
- Build automated evaluation pipelines that run continuously across model and pipeline updates
- Create dashboards and reports that surface quality trends across the organization

Model Inference & Comparison

- Run inference across multiple models and frameworks to benchmark performance
- Profile inference speed, memory usage, and output quality across different hardware configurations
- Identify failure modes and edge cases, then develop targeted test suites around them
- Maintain a living comparison matrix of approaches for different tasks within the pipeline

Teleoperation & Real-World Validation

- Teleoperate robotic hands and manipulators to evaluate object behavior in manipulation tasks
- Validate sim-to-real transfer performance by testing policies trained on simulated assets against physical robots
- Develop test protocols for grasping, manipulation, and articulated object interaction
- Document and quantify differences between simulation behavior and real-world performance

Dataset & Ground Truth

- Help curate evaluation datasets with high-quality ground truth annotations
- Build tools for efficient annotation review and quality control
- Design evaluation splits focused on generalization, including novel categories, out-of-distribution inputs, and edge cases
- Set up and maintain experimental rigs in both simulation and real-world environments

---

What We're Looking For

Must Have

- Strong Python skills
- Experience with ML model evaluation, including metrics design, statistical analysis, and benchmark construction
- Experience running inference with frameworks such as PyTorch, ONNX, and TensorRT
- A methodical, detail-oriented approach to testing and measurement
- Ability to work with 3D data including meshes, point clouds, and images
- Strong written communication skills for documentation and reporting

Strongly Preferred

- Hands-on experience with robotic teleoperation
- Familiarity with robotics simulation tools such as Isaac Sim, MuJoCo, PyBullet, or similar platforms
- Experience with 3D quality metrics including Chamfer Distance, F-score, IoU, LPIPS, or FID-style metrics for 3D outputs
- Background in sim-to-real transfer evaluation
- Experience with GPU profiling and inference optimization

- Experience setting up robotic systems for evaluation and building functional robotics stacks

Nice to Have

- Experience with dexterous manipulation or robotic hands
- Familiarity with ROS/ROS2
- Background in QA or test engineering for ML systems
- Experience building evaluation dashboards using tools such as Streamlit or Grafana

---

Compensation & Benefits

- Competitive salary adjusted for local market conditions
- Meaningful equity for full-time early employees
- Remote-first environment with flexible working hours
- Access to GPU compute infrastructure and robotic hardware
- Direct collaboration with company leadership and technical founders