Member of Technical Staff
Synthetic Data Generation

Location: Bay Area, CA · Type: Full-time · Team: Nvex R&D (Synthetic Data)

About Alchedata

Alchedata is building the self-improving data layer for Physical AI. Robots fail in the real world in ways no lab benchmark predicts. Our platform, Nvex, turns those failures into evidence: it diagnoses why a model failed, decides what data or environment would fix it, and proves whether the next checkpoint is actually better before it is promoted.

Frontier models can now label video, build simulation twins and generate synthetic data. We use them as labor inside the loop. Our job is the part they cannot do for a customer on their own: governed, auditable evidence that a fix worked on that customer's robots, tasks and real-world data. In short, frontier models can generate the fix; Nvex proves it worked.

About The Role

We are looking for a Member of Technical Staff to lead synthetic data generation at Alchedata. You will build the systems that generate agentic data, images and video at scale, and post-train the generative models behind them, with a focus on data for world models, VLMs and robot policies.

Generating data is getting cheaper. Knowing which generated data actually makes a model better is not. In this role you will do both: build state-of-the-art generation pipelines, and measure rigorously which synthetic data improves a downstream model and which does not. Your work ships into Nvex and into data we deliver to Physical AI teams and leading AI labs.

What You Will Work On

Agentic data generation

- Build pipelines that generate agentic trajectories: multi-turn interactions, tool and function calls, and computer-use tasks
- Design executable environments and verifiable-reward tasks for reinforcement learning post-training
- Generate reward-model, judge and preference data, with automated checks on correctness and diversity
- Extend agentic data from digital tasks toward embodied tasks, where an agent plans and acts in a physical or simulated world

Image and video generation, with a focus on post-training

- Post-train image and video generation models (SFT, preference optimization, RL, distillation) to produce controllable, physically plausible video of robots and scenes
- Build action-, camera- and trajectory-conditioned video generation for robot manipulation and navigation
- Generate targeted video for specific failure cases, such as rare objects, lighting, occlusion and contact-rich interactions
- Improve temporal consistency, physical realism and controllability, and define metrics that capture them

Synthetic data for world models and VLMs

- Generate training and post-training data for world models: rollouts, counterfactuals and imagined experience
- Build VLM post-training data, including dense captions, spatial and temporal QA, failure reasoning and hard negatives
- Combine generated, simulated and real data into recipes that cover the gaps in a model's behavior

Quality, verification and data value

- Build automated quality filters and verifiers: physical plausibility, consistency with conditioning, deduplication and diversity
- Run controlled experiments that show which synthetic data improves a downstream model, and by how much
- Feed what you learn back into generation, so the pipeline produces more of what helps and less of what does not

Pipelines at scale

- Build and operate generation pipelines on GPU clusters, optimizing throughput, latency and cost
- Use and orchestrate frontier models and open models in the same pipeline, with clean, configurable interfaces

Minimum Qualifications

- MS or PhD in Computer Science, Machine Learning or a related field, or equivalent industry experience
- 3+ years of hands-on work in generative modeling, synthetic data generation or multimodal machine learning
- Hands-on post-training of image or video generation models (for example diffusion or flow-matching models), or of LLMs / VLMs
- Experience building large-scale data pipelines with distributed inference (for example vLLM, SGLang or similar)
- Strong Python and PyTorch, and solid software engineering habits
- Ability to design experiments that tie data changes to model outcomes

Preferred Qualifications

- Publications at NeurIPS, ICML, ICLR, CVPR, ICCV, ECCV, CoRL or similar
- Experience with world models or video world simulators, including action-conditioned generation
- Experience generating data for agentic, tool-use or RL post-training, including environment and reward design
- Experience with VLM post-training data: captioning, grounding, video QA or preference data
- Background in robotics or embodied AI, or with simulation-rendered data
- Open-source contributions to generative modeling or data tooling with real adoption

What Success Looks Like In Your First Six Months

- An agentic data pipeline is producing quality-verified data that a customer or internal model team is using
- A post-trained video generation model is producing data that measurably improves a world model or VLM on our internal benchmark
- Data-quality and data-value metrics you defined are running as standard checks in our generation pipelines

Why Join

- Work on a central question in AI right now: which generated data actually makes models better
- Real robots, real deployment failures and real model decisions, not only offline benchmarks
- Publish: we encourage papers and open releases under our PEARL research program
- Early-team ownership and meaningful equity