Sara Ghazanfari

Homepage LinkedIn GitHub Publications sg7457@nyu.edu

About me

I am a PhD candidate in ECE department at NYU, researching multimodal large language models with a focus on video understanding, cross-modal alignment, and perceptual similarity. My work spans synthetic data construction, post-training of vision-language models for complex visual reasoning, parameter-efficient architecture design, and systematic evaluation of foundation models across temporal, spatial, and physical reasoning tasks. This summer, I am joining Apple as an Applied Research Scientist Intern to apply this research to continue the same line of work.

Experiences

- **Applied Research Scientist Intern**, **Apple Inc.**, Cupertino, US — May. 2026 – Aug. 2026
  - Working on video understanding in generative models.
- **GenAI Research Intern**, **Adobe Inc.**, San Jose, US — May. 2025 – Aug. 2025
  - Released SpotEdit (NeurIPS W 2025), a comprehensive benchmark evaluating visually-guided image editing across diffusion, autoregressive, and hybrid model architectures.
  - Designed the first hallucination evaluation subset for image editing, exposing systematic failure modes in GPT-4o and other state-of-the-art models when visual cues are absent.
  - Benchmarked leading models (OmniGen2, BAGEL, GPT-4o), identifying distinct capability profiles: OmniGen2 excelled at visual instruction following while BAGEL showed stronger background fidelity.
- **Research Assistant**, **New York University (NYU)**, NY, US — Jan. 2023 – Jan. 2027
  - Independently drove the full research cycle — problem identification, hypothesis formulation, implementation, and publication — delivering several first-author papers at CVPR, ICLR, TMLR, and NeurIPS within 3 years.
  - Conducted research across Multimodal LLMs, Vision-Language Models, and video understanding — with hands-on experience in model training, instruction tuning, and parameter-efficient fine-tuning (LoRA, adapter-based cross-modal alignment modules).
  - Built end-to-end research pipelines spanning synthetic data generation (Habitat, Kubric, CLEVRER), benchmark construction, and large-scale inference across 15+ state-of-the-art MLLMs including GPT-5.4, Gemini-3, Qwen3-VL, and LLaVA.
  - Designed rigorous evaluation frameworks incorporating human baselines, bootstrap confidence intervals, and sim-to-real correlation analyses — establishing reproducible methodology for diagnosing MLLM performance gaps across temporal, spatial, and physical reasoning tasks.
- **Technical Team Lead**, Narvan Startup Studio, Iran — Oct. 2021 – Jan. 2023
  - Led a team of 6 engineers to deliver a cryptocurrency exchange platform (web + mobile).
  - Architected backend using PostgreSQL, Redis, and Kafka; mentored junior engineers and coordinated cross-team technical decisions.
- **Data Engineer**, Narvan Startup Studio, Iran — Aug. 2020 – Oct. 2021
  - Built streaming data pipelines using Kafka with SQL (PostgreSQL) and NoSQL (Elasticsearch, InfluxDB, Redis).
  - Developed an NLP pipeline for Persian news (tokenizer, lemmatizer, CNN-based classifiers), improving topic classification accuracy by 15%.

Education

Ph.D. Candidate — Jan. 2023 – Dec. 2026

Electrical and Computer Engineering, New York University (NYU) — GPA: 3.9/4.0
Advisors: Siddharth Garg, Farshad Khorrami
Research Area: Multimodal Large Language Models (MLLMs), Vision-Language Models (VLMs), Foundation Models, Generative AI, Computer Vision, Deep Learning
Courses: Probability & Stochastic (A), Deep Learning (A), Algorithmic Machine Learning & Data Science (A^-), Linear Systems (A), Machine Learning for Cybersecurity (A).

Master of Science — Sep. 2018 – Jul. 2021

Computer Engineering, Sharif University of Technology (SUT) — GPA: 4.0/4.0

Lead-author Publications

SYNCR: A Cross-Video Reasoning Benchmark with Synthetic Grounding	ArXiv 2026
Chain-of-Frames: Advancing Video Understanding in Multimodal LLMs	CVPR 2026
SpotEdit: Evaluating Visually-Guided Image Editing Methods	NeurIPS W 2025
Towards Unified Benchmark and Models for Multi-Modal Perceptual Metrics	CVPR W 2025
EMMA: Efficient Visual Alignment in Multimodal LLMs	TMLR 2025
LipSim: A Provably Robust Perceptual Similarity Metric	ICLR 2024
R-LPIPS: An Adversarially Robust Perceptual Similarity Metric	ICML W 2023

Awards & Honors

- School of Engineering (SOE) Fellowship, Dept. of Electrical and Computer Engineering, NYU 2023
- Computer Science and Engineering Fellowship, University of California San Diego (UCSD) 2023

Skills

Research Expertise: Large Language Models (LLMs), Vision-Language Models (VLMs), Foundation Models, Generative AI, Diffusion Models, Autoregressive Models, Transformer Architectures, Fine-tuning, Instruction Tuning, Zero-shot Evaluation, Benchmark Design, Model Evaluation, Natural Language Processing (NLP).

Programming: Python, C, Java, R.

ML Frameworks & Tools: PyTorch, HuggingFace Transformers, HuggingFace Accelerate, DeepSpeed, FSDP, vLLM, scikit-learn, spaCy, Weights & Biases (wandb).

Infrastructure: CUDA, SLURM/HPC, Docker, Git/GitHub, Unix Shell, Jupyter Notebook, Blender, PyBullet, MATLAB.

APIs: OpenAI API, Gemini API.

Databases & Data Engineering: PostgreSQL, Redis, Elasticsearch, InfluxDB, Kafka, Django, Flask.

Selected Projects

SYNCR: A Cross-Video Reasoning Benchmark with Synthetic Grounding
A controlled synthetic benchmark for cross-video reasoning built using Habitat, Kubric, and CLEVRER simulator engines, providing programmatically verified ground truth for temporal, spatial, physical, and topological variables. SYNCR contains 8,163 QA pairs across 9,650 unique videos and evaluates MLLMs across eight tasks — revealing a 37-point gap between the best model (52.5%) and the human baseline (89.5%).

Chain-of-Frames (CoF): Advancing Video Understanding in Multimodal LLMs
A frame-aware chain-of-thought reasoning strategy for video LLMs that grounds model outputs in explicit frame references throughout the reasoning process. CoF improves performance on complex temporal and spatial QA benchmarks by anchoring each reasoning step to specific video frames, enabling more interpretable and accurate video understanding.

UniSim: Unified Benchmark and Models for Multi-Modal Perceptual Metrics
UniSim-Bench is the first benchmark to unify progress tracking for perceptual similarity metrics across both uni- and multimodal tasks, uncovering key insights into the generalization challenge. Paired with UniSim, a family of multi-task perceptual models serving as a first step toward general-purpose perceptual metrics.

EMMA: Efficient Visual Alignment in Multimodal LLMs
A lightweight modality adaptation module (<0.2% parameter overhead) that leverages CLIP's text encoder to generate instruction-aware visual representations, efficiently fusing visual and textual encodings with minimal added parameters. EMMA yields notable improvements on both specialized and general MLLM benchmarks.

LipSim: A Provably Robust Perceptual Similarity Metric
A provably robust perceptual similarity metric built on 1-Lipschitz neural networks with knowledge distillation from state-of-the-art ViT-based models, offering certified robustness guarantees against adversarial attacks. LipSim addresses demonstrated vulnerabilities of existing perceptual metrics while maintaining competitive natural performance.