Sara Ghazanfari

Homepage LinkedIn GitHub Publications sg7457@nyu.edu +1 (917) 386 7139

Experiences

• **GenAI Research Intern**, **Adobe Inc.**, San Jose, US May. 2025 - Aug. 2025

Designed and released **SpotEdit**, a benchmark for visually-guided image editing. Introduced the first hallucination evaluation subset, revealing vulnerabilities in GPT-4o and other SOTA models. Published results (arXiv 2025) and open-sourced code.

- Conceptualized and led SpotEdit, a publicly released benchmark evaluating visually-guided image editing across diffusion, autoregressive, and hybrid models.
- Designed a novel hallucination evaluation subset that assesses model failures when visual cues are missing—revealing GPT-4o's vulnerability and substantial performance gaps among state-of-the-art models
- Benchmarked leading models (OmniGen2, BAGEL, GPT-4o), showcasing their varying weaknesses; e.g., OmniGen2 excelled at visual following, while BAGEL showed stronger background fidelity.

• **Research Assistant**, **New York University (NYU)**, NY, US Jan. 2023 - Jan. 2027

My research focuses on advancing the visual understanding capabilities of multimodal large language models, with a particular emphasis on enhancing their spatial reasoning and perceptual abilities.

Research Interests: Large Multimodal Models, Multimodal LLMs, Video LLMs, Multimodal Perception

- Proposed a novel framework for video LLMs that grounds reasoning in explicit frame references, improving interpretability and performance on complex video question-answering tasks.
- Introduced a unified benchmark and models for multimodal perceptual similarity tasks, uncovering key insights into the generalization challenge
- Presented an efficient modality adaptation module that aligns visual and textual representations, boosting cross-modal performance and robustness in Multi-Modal Large Language Models.

• **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, 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 NLP pipeline for Persian news (tokenizer, lemmatizer, CNN-based classifiers).
- Improved topic classification accuracy by 15% with optimized CNN architectures.

Education

Ph.D. Candidate Jan. 2023 - Jan. 2027

Electrical and Computer Engineering Department, New York University (NYU) GPA: 3.9/4.0
Advisor: Siddharth Garg, Farshad Khorrami
Research Area: Deep Learning, Computer Vision, Multimodal Large Models, Multimodal LLMs
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 Department, Sharif University of Technology (SUT) GPA: 4.0/4.0

Lead-author Publications

SpotEdit: Evaluating Visually-Guided Image Editing Methods ArXiv 2025
Chain-of-Frames: Advancing Video Understanding in Multimodal LLMs ArXiv 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

Skills

Programming Languages: Python, C, Java, R.

Machine Learning & Deep Learning: Diffusion Models, Autoregressive Models, Vision Foundation Models, Multimodal Large Language Models (MLLM), Large Multimodal Models (LMM), Vision-Language Models (VLMs), Multimodal Perception Systems, Unified Models for Multimodal Understanding & Generation, Visual Reasoning.

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

Frameworks & Tools: PyTorch, scikit-learn, spaCy, Django, Flask, Jupyter Notebook, PyCharm, Git/GitHub, Unix Shell, MATLAB, Microsoft Office.

Research & Development Skills: Industry-integrated research, Debugging, Collaborative Problem Solving, Technical Engineering, Team Management, High-Performance Computing (HPC), Architecture Design.

Projects

SpotEdit: Evaluating Visually-Guided Image Editing Methods

We present SpotEdit, a comprehensive benchmark designed to systematically assess visually-guided image editing methods. More specifically, our benchmark is the first to include a dedicated component on hallucination, highlighting how leading models, such as GPT-4o, fail in performing the editing task.

Advancing Video Understanding in Multimodal LLMs via Frame-Aware Reasoning

We propose a method to enhance the performance of video LLMs on a variety of video understanding benchmarks that involve complex temporal and spatial reasoning tasks. Our approach, Chain-of-Frames (CoF), introduces a novel frame-aware chain-of-thought reasoning strategy that explicitly incorporates temporal information into the reasoning process by directly referencing video frames within the CoT traces.

Towards Unified Benchmark and Models for Multi-Modal Perceptual Metrics

We propose UniSim-Bench, the first benchmark to track the progress of perceptual similarity metrics across uni- and multimodal tasks. Furthermore, we propose UniSim, a set of multi-task perceptual models which are a first step towards general-purpose perceptual metrics.

EMMA: Efficient Visual Alignment in Multimodal LLMs

Our lightweight approach, EMMA (Enhanced Multi-Modal Adaptation), leverages CLIP's text encoder to generate instruction encodings and, by exploiting this initial alignment, demonstrates that the modality adaptation module can be simple while still enhancing the alignment between visual and textual modalities.

LipSim: A Provably Robust Perceptual Similarity Metric

We demonstrate the vulnerability of the SOTA perceptual similarity metric based on an ensemble of ViT-based feature extractors to adversarial attacks. We then propose a framework to train a robust perceptual similarity metric called LipSim (Lipschitz Similarity Metric) with provable robustness guarantees.

R-LPIPS: An Adversarially Robust Perceptual Similarity Metric

The very first work to show the vulnerabilities of the extensively-used perceptual metric, LPIPS, to adversarial attacks and further propose the use of Adversarial Training to build a new Robust Learned Perceptual metric, R-LPIPS , that leverages adversarially trained deep features.

Teaching Experiences

Head TA, Machine Learning	Feb. 2021 - Aug. 2021
TA, Modern Information Retrieval	Feb. 2020 - Aug. 2020
TA, Machine Learning	Feb. 2019 - Aug. 2019

Volunteer Experiences

- Reviewer for CVPR 2025 and NeurIPS 2025, contributing to the peer-review process in machine learning.
- Volunteered as a mathematics teacher at underprivileged schools, over the course of one semester.