SARVESH BASKAR

+91 8754583044 ◇ baskarsarvesh@gmail.com
LinkedIn ◇ Website ◇ Google Scholar ◇ GitHub

SUMMARY

Dual-degree CS and Physics graduate from BITS Pilani with extensive research experience in Large Language Models (LLMs), Vision-Language Models (VLMs), and agentic reasoning at UMD and UMBC. Co-author of 4 peer-reviewed publications (ACL Findings, NAACL) and 3 preprints under review (EMNLP, ECCV). My research interests span Multimodal AI, Vision-Language Models, LLM Agents, Reasoning & Planning, Visual Grounding, Embodied AI, Robotics, Computer Vision, and Vision-Language-Action Models.

EDUCATION

Birla Institute of Technology and Science (BITS) Pilani Sep 2020 to July 2025
Dual Degree:
Bachelor of Engineering in Computer Science Cumulative GPA: 8.46/10.00
Master of Science in Physics
Thesis Title: An Approach to Resolving the Persona Knowledge Gap in LLMs during Multi-Turn Conversations

SKILLS

Programming: Python, C/C++, Bash, SQL, LaTeX
ML/AI: PyTorch, TensorFlow, OpenCV, Transformers, vLLM, LangChain, LlamaIndex, DSPy, RAG, VLMs
Systems: Git, Linux, Docker, Slurm, Redis, ChromaDB, FAISS, Ollama, MCP, HuggingFace

PUBLICATIONS

Peer-Reviewed Publications

- Agrawal, Aakriti, Gouthaman KV, Rohith Aralikatti, Gauri Jagatap, Jiaxin Yuan, **Sarvesh Baskar**, Vijay Kamarshi, Andrea Fanelli, and Furong Huang. "Towards Mitigating Hallucinations in Large Vision-Language Models by Refining Textual Embeddings."
  ACL Findings, 2026.
- **Baskar, Sarvesh**, Manas Gaur, Srinivasan Parthasarathy, and Tanmay Tulsidas Verlekar. "(CPER) From Guessing to Asking: An Approach to Resolving Persona Knowledge Gap in LLMs during Multi-Turn Conversations."
  NAACL SRW, 2025.
- **Baskar, Sarvesh**, Muhammad R. Islam, Zikui Cai, Ankit Nakhawa, Anirudh Satheesh, Tom Goldstein, and Furong Huang. "The Low-Frequency Trap: Why Scaling Doesn't Solve Simple Temporal Counting."
  ICLR 2026 Workshop on I Can't Believe It's Not Better, 2026.
- Mohseni, Seyedreza, **Sarvesh Baskar**, Arman Hossain, Aritran Piplai, Edward Raff, and Manas Gaur. "Analyzing Chain of Thought (CoT) Approaches in Control Flow Code Deobfuscation Tasks."
  WoRMA, 2026.

Preprints and Under Review

- **Baskar, Sarvesh**, Muhammad R. Islam, Zikui Cai, Anirudh Satheesh, Tom Goldstein, and Furong Huang. "Parametric Capability Profiling for Multimodal Video Reasoning."
  Under Review ECCV, 2026.
- Islam, Muhammad R., **Sarvesh Baskar**, Zikui Cai, Anirudh Satheesh, Keenan Powell, Tom Goldstein, and Furong Huang. "Executable Reasoning Traces for Diagnosing Multimodal Video Reasoning."
  Under Review ECCV, 2026.
- Suri, Manan, **Sarvesh Baskar**, and Dinesh Manocha. "Video2LoRA: Parametric Video Internalization for Vision-Language Models."
  Under Review EMNLP, 2026.

ACADEMIC EXPERIENCE

University of Maryland, College Park Remote
Research Assistant (Furong Huang's Lab) Jul 2025 - Present

- Collaborate with PhD students and a postdoctoral researcher to scope tasks and plan experiments for large vision-language models and video reasoning.
- Developed hypernetwork architectures to generate SmolVLM2 LoRA weights directly from intermediate video representations, bypassing context-heavy visual token concatenation at query time.
- Implemented textual embedding-refinement experiments in OpenQwen2VL to mitigate object hallucination and enhance visual grounding.
- Run controlled experiments and ablations where I tracked configurations, metrics, and results for reproducibility.
- Read literature, synthesize relevant papers to inform methods, and deliver concise weekly updates with findings, figures, and next steps.

University of Maryland, Baltimore County (UMBC) MD, USA
Research Assistant (KAI2 Lab) — Advisor: Dr. Manas Gaur Aug 2024 - Jul 2025

- Led experiments, analysis, and writing for a first-author NAACL 2025 SRW paper on conversation preference elicitation for LLMs.
- Collaborate with PhD students to scope tasks and plan experiments on persona knowledge gap resolution and code reasoning.
- Lead developer for PlanForge, which extends AutoTOS to harder ACPBench classical planning domains in collaboration with IBM where I defined experiments, added benchmarks, and standardized analysis.
- Developed LLM-driven planning systems, code deobfuscation modules, and evaluation pipelines using Python.
- Co-wrote sections of the Samsung START 2025 proposal, including scope, methodology, milestones, and expected outcomes.

TCS Research - Datalab, APPCAIR, BITS Pilani Goa, India
Student Researcher (APPCAIR) Jan 2024 - May 2024

- Conducted comprehensive literature reviews to scope tasks, identify baselines, datasets, and evaluation protocols for automated planning and code synthesis.
- Translated high level research goals into concrete experimental plans and weekly milestones with professors and industry collaborators.
- Implemented discussed methodology pipelines in Python with version control, unit tests, and clear documentation, achieving +30% code yield through iterative prompt repair.
- Prepared weekly written reports and presentations that summarized results, insights, and next steps for faculty and industry stakeholders.
- Coordinated code handoff by organizing repositories, packaging notebooks for maintainability, and contributed to manuscript sections on related work and methods.

Birla Institute of Technology and Science (BITS) Pilani Goa, India
Graduate Teaching Assistant Aug 2023 - May 2024
BITS Pilani, Goa Campus (Department of Physics) *Goa, India*

- Assisted instruction for the graduate-level course *Mathematical Methods of Physics* (Fall 2023-24).
- Graded assignments and problem sets
- Conducted demonstrations for mathematical methods, including integration and differential equations.
- Supported exam evaluation; provided feedback to students and instructors to improve learning outcomes.

SELECTED RESEARCH PROJECTS

Video2LoRA: Zero-Token Video Question Answering (UMD) *Mar 2026 - Present*

- **Problem:** Video reasoning in VLMs is extremely token-heavy, causing high latency, large context window usage, and context degeneration during long-video question answering.
- **What I built:** A perceiver hypernetwork running alongside a VLM to predict adapter weights in a single forward pass, eliminating visual tokens from query context and enabling composition of independent video chunk adapters in rank space.
- **Impact and Results:** Reduces query-time visual tokens by 1500x and TTFT by up to 80x while maintaining accuracy; under review at EMNLP 2026.

PlanForge: Executable Planner Generation from Natural Language (UMBC) *May 2025 - Present*

- **Problem:** Traditional direct LLM plan generation from natural language is highly brittle, unverified, and generalizes poorly to multi-constraint tasks with multiple valid solutions.
- **What I built:** An Architect–Builder–Runner framework that converts natural-language planning problems into executable solver programs rather than direct textual plans. Designed an Architect stage for state/logic inference, a Builder pipeline for solver code generation/repair, and a deterministic Runner to verify outputs.
- **Impact and Results:** Evaluated PlanForge on ACPBench, Tower of Hanoi, Frogs Jumping, and NaturalPlan, achieving a 100% success rate on NaturalPlan trip planning, meeting planning, and calendar scheduling, demonstrating flawless transfer to multi-constraint planning tasks with multiple valid solutions.

VisAlign: Mitigating Hallucinations via Refining Textual Embeddings (UMD) *Oct 2025 - Jan 2026*

- **Problem:** Large vision language models tend to over rely on language tokens because visual embeddings are simply appended to the text stream, which leads to weak visual grounding and hallucination.
- **What I did:** Implemented the VisAlign refinement in the OpenQwen2VL stack and ran the full pretraining and evaluations. Integrated average pooled visual features into textual embeddings at inference, executed experiments, and prepared the result tables and plots for the paper.
- **Impact and Results:** Integrated VisAlign into OpenQwen2VL and ran full pretraining plus eval scripts. Preliminary benchmarks suggest lower hallucination rates with improved grounding; published at ACL Findings 2026.

Multi-Step Reasoning using Process Reward Models (UMD) *Jul 2025 - Aug 2025*

- **Problem:** Outcome-only supervision in math reasoning rewards the final answer but ignores the quality of intermediate steps, leading to brittle reasoning chains.
- **What I built:** An end-to-end PRM pipeline that generates multi-step trajectories, assigns weak step-level labels using verifier/self-consistency signals, and uses that score at inference for PRM-guided answers.
- **Impact and Results:** Majority voting was consistently strongest with a 92% accuracy compared to PRM with a 68% accuracy on GSM8K.

Automated Planning and Code Synthesis for Continual ML Ideation (APPCAIR) *Jan 2024 - May 2024*

- **Problem:** Turning new research ideas into runnable changes in an existing ML pipeline is slow and inconsistent. Teams need automatic extraction of ideas from papers and automatic plan to code execution.
- **What was built:** An LLM driven system that mines techniques and goals from a set of papers with a RAG workflow, then produces a modification plan, and generates, debugs, and finalizes code.
- **Impact and Results:** Internal runs showed about a 30% increase in executable code yield from optimized prompting and iterative repair.

INDUSTRY EXPERIENCES

Avyott Goa, India / Remote
AI Consultant Aug 2025 - Present

- Designed and shipped a multimodal RAG chatbot embedded inside client products, able to process text, images, and tables for grounded and cited answers.
- Wrote and deployed Model Context Protocol (MCP) servers to connect with multiple ticketing systems, enabling chat driven ticket creation, updates, and status sync.
- Implemented a clarification step that asks targeted questions before answering, based on my thesis research, which improved personalization and intent capture.
- Deployed the product for enterprise use at IPCA and Taj, integrating with internal knowledge bases and existing workflows.

Techisy Goa, India / Remote
Summer Intern May 2024 - July 2024

- Built an AI-powered duplicate bill identification system (95% accuracy) combining OCR, OpenAI embeddings, and CLIP, reducing manual review time.
- Created a RAG-based resume matching solution using LangChain, Ollama (Llama 3), and FAISS to align resumes with job descriptions.
- Deployed prototypes and iterated with stakeholders to improve screening efficiency.

ENGLISH TESTS

IELTS (Academic): Overall: 7.5 | Listening: 7.5 | Reading: 7.5 | Speaking: 8.0 | Writing: 7.0