Arnav Bendre

Roorkee, Uttarakhand, India
9897076017 bendrearnav6@gmail.com Linkedin Github

Research Interests

- **Multimodal generative modeling** and the reliability of vision-language models, efficient diffusion sampling and distillation, and the measurement, mechanisms, and **mitigation of hallucination in VLMs**. Drawn to rigorous, falsifiable research questions, with an eye toward results that also hold up in practice.

Education

Indian Institute of Technology, Roorkee **Aug 2024 – May 2028**
Bachelor of Technology in Energy Engineering *Roorkee, Uttarakhand*

Publications

B-DENSE: Branching For Dense Ensemble Network Supervision Efficiency Paper
Accepted, ICLR 2026 Workshop on Deep Learning for Theory (DeLTa) *2026*

Rethinking CD: A Reproducibility Study and Extension on the Ineffectiveness of Contrastive Decoding at Mitigating Object Hallucinations in MLLMs Paper
arXiv preprint *2026*

Experience

Data Science Group, IITR **Website**
Core Member - student-led ML/AI research group *Jan 2025 - Present*

- Research on multimodal generative models and vision–language reliability: efficiency of diffusion sampling, and the measurement and mitigation of hallucination in VLMs.
- Placed in multiple Kaggle and ML competitions, implementing transformers and CNNs from scratch, with rigorous EDA driving architecture and feature-engineering choices; also ran a Kaggle-style onboarding challenge.

Projects

BranchDistillation | *Python, PyTorch, HuggingFace, diffusers, Optuna* **Dec 2025**

- Built a diffusion-distillation framework supervising the student on the teacher's full denoising trajectory via K-fold branched output channels (K·C instead of C), **recovering intermediate denoising steps** discarded by standard distillation.
- Cut FID against Progressive Distillation on CIFAR-10 (39.66 →20.81 at 128 steps) and against SFD on ImageNet-64 at NFE 2 (10.25 →9.57) for a $\sim 0.01\%$ FLOP increase and identical inference-time complexity.
- Used PyTorch and Python to create scripts to run the training, distillation, and inference on Lightning.ai(A100/L4).

Rethinking Contrastive Decoding | *Python, API, PyTorch, HuggingFace* **July 2026**

- Led a reproducibility study and extension stress-testing contrastive decoding (VCD, ICD, SID) across LLaVA-1.5 (7B/13B) and Qwen2.5-VL on POPE, MME, and CHAIR.
- Showed CD is **statistically indistinguishable** from spurious, **non-visual control** methods on precision and hallucination rate, isolating its effect to a single recall shift (+12.21pp, $p<0.001$) rather than genuine visual grounding.
- Traced a directional anomaly in **ICD's Yes-rate shift** and conducted a **logit-lens analysis** to localize the internal layers where the intervention breaks down, establishing that CD's apparent gains are a distributional artifact, not corrected perception.

VLM-Reliability: Modality Gaps and Vision Refusal | *Python, PyTorch, HuggingFace* **April 2026**

- Probed the modality gap in LLaVA-1.5-7B by extracting text and image refusal directions from residual-stream activations and comparing them geometrically..
- Found their cosine similarity never exceeds 0.3 while the text-side norm consistently dominates the image-side norm, identifying **the modality gap as a primary driver of weak vision-side refusal.**

Visual-CoT | *Agentic workflow, Python, API, PyTorch* **May 2026**

- Built an end-to-end, resumable pipeline turning raw ChartQA samples into a supervised fine-tuning corpus of structured visual chain-of-thought traces (chart summary →extraction →computation →answer).
- Sampled 3 traces per item from a frontier VLM teacher, then filtered through five staged gates — schema, answer correctness, numeric grounding against the source data table, arithmetic re-verification, and 2-of-3 self-consistency — keeping the highest-grounded survivor (95%yield).

Technical Skills

Languages and Tools: Python, LaTeX, Shell
Libraries and Frameworks: PyTorch, TensorFlow, OpenCV, pandas, NumPy, Matplotlib, scikit-learn, HuggingFace Transformers/diffusers/datasets