CS671_New: Deep Learning and Applications
PROJECT PROPOSALS

Project-ID: P1
Project Mentor: Richa

Project Title: Self-Supervised Learning in DINO-style Teacher-Student learning framework for Videos.

Problem Statement:

Video object segmentation/detection/classification is a fundamental problem in computer vision. Traditional supervised approaches require large amounts of pixel-level annotated video data, which is expensive, time-consuming, and often impractical to obtain at scale. Moreover, such models tend to generalize poorly when applied to unseen object categories or real-world scenarios.

This project addresses the challenge of learning robust object segmentation/detection/classification in videos. By leveraging self-supervised learning, the model learns meaningful spatio-temporal representations from unlabeled videos using intrinsic cues such as motion consistency, appearance similarity, and temporal coherence. The objective is to develop a segmentation framework that can effectively identify and track objects across frames while reducing dependency on labeled data and improving generalization across diverse video domains.

Models to be used: Video Vision Transformer or CNN-Transformer hybrid to be used in DINO-style Teacher-Student learning framework.

Resources:

1. https://www.v7labs.com/blog/contrastive-learning-guide
2. https://openaccess.thecvf.com/content/ICCV2021/html/Caron_Emerging_Properties_in_Self-Supervised_Vision_Transformers_ICCV_2021_paper
3. https://arxiv.org/abs/1805.01978

Project-ID: P2
Project Mentor: Inam-ul-Haq Gulzar, Peeyush Kumar Singh

Project Title: Generative Models for Cross-Modal Medical Image Synthesis (CT → MRI)

Problem Statement:

Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) are two widely used and complementary medical imaging modalities. CT provides fast, high-resolution structural information, particularly for bone and dense tissues, while MRI excels at soft-tissue contrast and functional imaging without exposing patients to ionizing radiation. Despite their complementary nature, acquiring both scans for a single patient is often impractical due to cost, time constraints, scanner availability, and contraindications associated with MRI.

In many clinical workflows—such as radiotherapy planning, image-guided surgery, and multimodal analysis—access to both CT and MRI is highly desirable. However, requiring multiple imaging sessions increases patient burden and healthcare costs. Recent advances in generative modeling have opened the possibility of synthesizing MRI-like images directly from CT scans, enabling virtual modality conversion without additional acquisitions.

This project aims to explore and develop generative models that learn mappings from CT to MRI, with an emphasis on image fidelity, perceptual quality, and **a**n indispensable focus on downstream clinical utility. You will get to work at the intersection of medical imaging, deep learning, and generative modeling, gaining hands-on experience with real clinical data and modern architectures.

You will mainly work on implementation of variants of cyclic models, on multi context (2D, 2.5D, 3D) data. The aim is to train the networks with minimal paired data availability.

Methodology:

1) Data Curation & Simulation

Paired data:

1) SynthRAD data;
2) Proprietary Data (Pre-processed to a ready to develop format)

2) Metrics:

1. Standard Objective IQA: PSNR, SSIM, MAE

2. Perceptual: LPIPS
3. Task utility: Downstream performance evaluation such as segmentation, classification.

3) Ablation Studies

1. Analyze what networks perform best on each task, and why.
2. Architecture: 2D/2.5D/3D;
3. Losses: +Edge, +SSIM, etc.
4. Possible additional Conditioning

Project ID: P3

Mentor: Soma Chakraborty (D21006@students.iitmandi.ac.in)

Project Title:

Make It Till You Gait It: Generating body position-aware Wearable IMU Signals

Problem Statement:

To build a generative model that can create **plausible IMU signals** just by knowing where the sensor is worn on the body (e.g., "head", "right foot").

Methodology:

1. Data Understanding (A gait dataset with **IMU - accelerometer+gyroscope signals from 16 body locations)**

- Explore IMU signals from different body locations
- Observe differences in amplitude, frequency, and smoothness

2. Signal Generation

Model: A diffusion model (e.g. text-conditioned DDPM)

Input:

- Random noise schedule
- Body location label (as an embedding from CLIP)

Output:

- A synthetic IMU time series for that body location (the goal is **not perfect reconstruction**, but capturing typical motion patterns).

3. Signal-Level Evaluation

Compare generated and real IMU signals using:

- RMSE / correlation
- Power spectral density (frequency content)
- Visual inspection of waveforms

4. Task-Level Evaluation (Most Important Part!)

Train and test ML models for downstream tasks:

- Gait anomaly detection (classification)

- Action recognition (classification)
- Person identification (one-to-many matching)

Compare performance when using:

1. Real IMU signals
2. Generated IMU signals
3. A mix of both

Project ID: P4

Mentor: Soma Chakraborty (D21006@students.iitmandi.ac.in)

Project Title:

MotionSense: ***Understanding First-Person Vision-based Human Motion from Head-Mounted IMU and Back***

Problem Statement:

To connect (align) motion information (representation) coming from:

- ego-centric action videos, and
- head-mounted IMU signals,

even when the two data sources are not paired (i.e., they do not come from the same person or time).

Methodology:

1: Use Pre-trained Feature Extractors

- a pre-trained model that extracts motion features from ego-centric video clips (any action recognition model), and
- a pre-trained model that extracts motion features from head-mounted IMU data (an MAE pretrained on gait signals).

These models are treated as **black boxes**—no need to train large networks from scratch.

2: Learn to Align the Features (using the concept of Cycle-GAN)

Students will implement small neural networks that:

- map video features into the IMU feature space, and
- map IMU features into the video feature space.

Training is done using:

- adversarial learning (to make mapped features look realistic),
- cycle consistency (mapping there and back should preserve information),
- distance preservation (walking styles should remain distinguishable).

All work happens in feature space, making the work computationally light.

3: Evaluate the Learned Representations

Instead of visualizing signals, you will evaluate embeddings using:

- cross-modal retrieval or identification (video → IMU and IMU → video),
- simple classification or clustering tasks,
- qualitative analysis of embedding similarity.

Project-ID: P5

Project Mentor: Sushovan

Project Title: Agentic Civic Engagement framework

Problem Statement:

This proposal outlines the development of an intelligent agentic framework that automates the process of converting activist video content from social media platforms (Instagram, YouTube) into structured government complaints. The system will identify relevant government authorities, extract contact information and portal details, intelligently populate complaint forms, execute email communications with evidence attachments, and maintain escalation trails through state and central government authorities. This framework leverages agentic AI, computer vision, natural language processing, and workflow automation to democratize civic participation and accelerate environmental protection mechanisms.

Target Scope: Environmental violations, civic infrastructure issues, pollution incidents, waste management violations, and sustainability concerns captured by citizen activists.

Methodology:

The framework will operate as a modular, multi-stage agentic system that transforms activist video content into government-coordinated complaint actions:

Input Processing Pipeline

- Extract video/image from Instagram, YouTube via public APIs or direct upload
- Apply computer vision to detect issue type (environmental, infrastructure, sanitation, pollution)
- Extract geolocation metadata (GPS, landmark recognition, or manual specification)
- Apply NLP to extract context, severity indicators, and environmental violations

Authority Mapping Engine

- Query knowledge graph of government authorities indexed by jurisdiction, issue type, and capability
- Retrieve contact details, portal URLs, email addresses, and complaint form specifications
- Identify hierarchical escalation path (local → district → state → central)
- Flag authorities with multi-channel capabilities (online portal, email, WhatsApp, CPGRAMS)

Intelligent Form Population

- Parse government complaint forms and identify required fields
- Autonomously populate fields using extracted video context and user metadata
- Validate data against form requirements (character limits, field formats, mandatory fields)
- Generate natural language summaries suitable for different portal styles

Multi-Channel Execution

- Submit complaints via CPGRAMS API integration
- Fill and submit forms on ministry and state pollution board portals
- Compose and send formal emails to identified authorities with video/image attachments
- Generate CC chains to ensure state and central oversight

Escalation and Tracking

- Maintain complaint registry with unique tracking IDs
- Implement automated escalation logic based on response SLAs
- Route to higher authorities if local response is insufficient
- Provide citizen dashboard for complaint status tracking

Project-ID: P6

Project Mentor: Sushovan

Project Title: Efficient VideoQA using query based frame selection and Test-Time adaptation on mobile devices

Problem Statement:

Video question-answering on live videos is still challenging in resource-constrained devices due to its limited memory and compute. But it can be made computationally efficient by selecting only query-relevant frames before feeding into vision-language models (VLMs). It builds on recent peer-reviewed advances in query-adaptive sampling to reduce token costs and latency while maintaining or improving accuracy on benchmarks like MVBench. The problem focuses on enabling live Video QA and Assistance for Blind and Low Vision community.

Methodology:

Standard VideoQA pipelines uniformly sample frames from long videos, leading to high compute demands as VLMs process redundant tokens. Query-based methods address this by dynamically selecting semantically relevant frames via submodular mutual information (SMI) or cross-modal retrieval, boosting accuracy.

Technical Approach

- Query-Aware Frame Selection: Extract candidate frames via uniform downsampling, then rank using SMI or CLIP-based similarity to the query embedding for top-k selection (e.g., k=8-16).

- Multi-Resolution Adaptation: Assign higher resolutions to high-relevance frames and lower to others, minimizing tokens while preserving details; integrate with small VLMs like SmolVLM and others.

Evaluation:

On Videoa11y dataset and MVBench.

Background : We have already baselined SmolVLM variants for Blind and Low-Vision accessibility and deployed on mobile devices, assessing their performance and latency.

Published the work - https://aclanthology.org/2025.mmloso-1.8/

This work would build on top of that for more novelties.

Project-ID: P7
Project Mentor: Jyoti Nigam

Project Title: Who is talking to me ?

Task:
An egocentric video provides a unique lens for studying social interactions because it captures utterances and nonverbal cues from each participant's unique view and enables embodied approaches to social understanding. Progress in egocentric social understanding could lead to more capable virtual assistants and social robots. Computational models of social interactions can also provide new tools for diagnosing and treating disorders of socialization and communication such as autism, and could support novel prosthetic technologies for the hearing-impaired.

While the Ego4D dataset can support such a long-term research agenda, our initial Social benchmark focuses on multimodal understanding of conversational interactions via attention and speech. Specifically, we focus on identifying communicative acts that are directed towards the camera-wearer, as distinguished from those directed to other social partners: Talking to me (TTM): given a video and audio segment with the same tracked faces and an additional label that identifies speaker status, classify whether each visible face is talking to the camera wearer. The TTM task is defined as a frame-level prediction y, which stands in contrast to audio analysis tasks where labels are often assigned at the level of audio frames or segments. A desired model must be able to make a consolidated decision based on the video and audio cues over the time course of an utterance. For example, if the speaker turns their head to the side momentarily while speaking to the camera-wearer, then a frame where the speaker is looking away would have y = 1.

Data:

- Input: Video and audio segments from the same tracked talking person.
- Output: A binary label indicating if the person is talking to the camera wearer and a score (probabilty).

Starter code:
Please find the code to reproduce existing work baselines and to get started on this challenge here.

Ego4D Episodic Memory Benchmark

EGO4D is the world's largest egocentric (first person) video ML dataset and benchmark suite.
For more information on Ego4D or to download the dataset, read: Start Here.
The Episodic Memory Benchmark aims to make past video queryable and requires localizing where the answer can be seen within the user's past video. The repository contains the code needed to reproduce the results in the Ego4D: Around the World in 3,000 Hours of Egocentric Video.

Project-ID: P8
Project Mentor: Jyoti Nigam

Task: **VQ2D: *Visual Queries with 2D Localization***

This task asks: “When did I last see [this]?” Given an egocentric video clip and an image crop depicting the query object, the goal is to return the last occurrence of the object in the input video, in terms of the tracked bounding box (2D + temporal localization). The novelty of this task is to upgrade traditional object instance recognition to deal with video, and particularly ego-video with challenging view transformations.

Data:

- Input: Long, untrimmed video clip, query frame (time at which the query is made), and a static visual crop and textual name of the object
- Output: Temporally contiguous set of per-frame bounding boxes around the most recent occurrence of the object

Note:

- The task input additionally includes the raw textual name of the object.
- We have cleaned the VQ2D annotations to fix a small subset of cases where the bounding boxes were incorrectly rotated in the frame. This primarily affects the train and val splits. The test split remains unaffected.
    - The updated annotations (version v1.0.5) can be downloaded using the Ego4D CLI as follows: `python -m ego4d.cli.cli --aws_profile_name ego4d --datasets annotations -y --version v2 --output_directory <PATH TO OUTPUT DIR>`
    - The annotations will be downloaded to `<PATH TO OUTPUT DIR>/v2/annotations/vq_*.json`.

Starter code:

Please find the code to reproduce our paper baselines and to get started on this challenge here.

Project-ID: P9
Project Mentor: Jyoti Nigam

Title: NLQ: *Natural Language Queries*

Task:

Natural Language Queries (NLQ) is a task in the Ego4D Episodic Memory Benchmark. The motivation behind the NLQ task is to enable searching through an egocentric video using a natural language query. The system responds to a query by providing a temporal window localized in the video, from which the answer to the query can be deduced. More concretely, given an egocentric video V and a natural language query Q, the goal is to identify a response track r, such that the answer to Q can be deduced from r. The response track should be a set of temporally contiguous frames within V. Given the episodic nature of this task, r should be sufficient to answer Q, without the additional need for V or any external knowledge bases.

This task asks, "What/when/where....?" -- general natural language questions about the video past. Given a video clip and a query expressed in natural language, the goal is to localize the temporal window within all the video history where the answer to the question is evident. The task is novel because it requires searching through video to answer flexible linguistic queries. For brevity, these example clips illustrate the video surrounding the ground truth (whereas the original input videos are each ~8 min).

Data:

- Input: Long, untrimmed video clip V and a natural language query Q
- Output: Temporally contiguous set of frames r within V, such that the answer to Q can be deduced from r

Starter code: Please find the code to reproduce existing work paper baselines and to get started on this challenge here.

Project-ID: P10
Project Mentor: Jyoti Nigam

Title: EEG based visual brain decoding: EEG to Video Reconstruction

Task:

Our visual experiences are composed of continuously evolving scenes caused by the movement of objects and viewing perspective. To investigate the mechanism of our visual system, various neuroimaging techniques have been used to analyze brain activities, especially non-invasive methods like functional Magnetic Resonance Imaging (fMRI), magnetoencephalography(MEG), and electroencephalography(EEG). Compared to fMRI and MEG which need to be recorded by large and expensive medical devices, EEG is relatively low-cost and portable and thus has been applied across many human visual studies. Besides the fundamental classification tasks, reconstructing visual perceptions from corresponding brain signals helps to advance the understanding of our visual neural system.

Recently, some works reconstruct high-quality two-second videos from a single fMRI data frame. However, limited by the low temporal resolution of fMRI, these video generation frameworks lack the ability of capturing high dynamic changes. To this end, we propose to address the problem of video reconstruction from given EEG Signals.

EEG Dynamic Vision Dataset and Benchmarks:

The SJTU EEG Dataset for Dynamic Vision (SEED- DV) is constructed by a group of researchers. Then two benchmarks were built on the SEED-DV dataset: EEG visual perception classification benchmark and video reconstruction benchmark. The purpose of building this new dataset is to answer the following research questions:

RQ1 : Whether can we decode dynamic visual information from EEG signals?

RQ2 : If yes, which visual information can be decoded?

RQ3 : To what extent can we reconstruct video from EEG signals?

Hence, they carefully selected video clips suitable for studying dynamic vision and annotated their meta information.

Input: EEG signal Segments

Output: Reconstructed Video

Project-ID: P11
Project Mentor: Samsung PRISM

Project Title: Evaluating Style by Analogy: Few-Shot Visual In-Context Metrics for Style Transfer

Problem Statement:
The evaluation of style transfer methods is currently hindered by the "subjectivity gap". Traditional metrics (e.g., Gram matrices, LPIPS) measure low-level texture statistics or perceptual distance but fail to capture the semantic nuances of artistic style. Similarly, modern semantic metrics (e.g., CLIP-Score) are static, they evaluate image content but struggle to quantify the fidelity of a specific stylistic transformation. A "cubist" generation might have high CLIP alignment but fail to respect the specific geometric rules of a reference artist. Furthermore, relying on zero-shot VLM evaluation leads to high variance and hallucination, as the model lacks a ground-truth definition of the target style within its immediate context.

We propose that style transfer should be evaluated as a Visual Analogy task, rather than a static quality assessment. Current systems lack the ability to observe input-target pairs (golden exemplars) to understand the intended transformation logic before scoring. Without these few-shot examples, automated evaluators cannot distinguish between a stylistic artifact (e.g., intended brush strokes) and a generation error (e.g., noise). This proposal outlines the development of a VLM-based evaluator that utilizes In-Context Learning (ICL) to define style dynamically. By conditioning the model on high-quality reference pairs, we aim to create a metric that correlates highly with human aesthetic judgment, capable of serving as a robust reward signal for training and RLHF.

Expectations:
- **Human Alignment:** The proposed metric must achieve a higher Spearman/Pearson correlation with human preference ratings than current state-of-the-art metrics (CLIP-Score, PickScore, LPIPS, and DINO).
- **Generalization:** The model must demonstrate the ability to evaluate unseen or custom styles effectively when provided with just 1-3 exemplar pairs, without requiring fine-tuning.
- **Explainability:** The model should provide textual rationales for its scores, correctly identifying specific stylistic failures (e.g., "color palette matches, but stroke texture is missing") in at least 80% of negative cases.
- **Reward Signal Viability:** Demonstrating that optimizing against this new metric results in qualitatively better image generation compared to optimizing against standard style losses.

References:

1. *Chen, S. et al.* ADIEE: Automatic Dataset Creation and Scorer for Instruction-Guided Image Editing Evaluation. In ICCV 2025
2. *Baraldi, L. et al.* What Changed? Detecting and Evaluating Instruction-Guided Image Edits with Multimodal Large Language Models. In ICCV 2025
3. *Jiao, Q. et al.* Img-Diff: Contrastive Data Synthesis for Multimodal Large Language Models. In CVPR 2025

Project-ID: P12

Project Mentor: Geetanjali Sharma (d19062@students.iitmandi.ac.in)

Project title : Forehead Creases Based Presentation Attack Detection (PAD) (New Work)

1. **Problem Statement :** Face recognition systems are vulnerable to **presentation attacks**, such as printed photos, images displayed on screens (e-display), and synthetically generated faces. Most existing PAD systems focus on the **entire face**, while **forehead creases**—a stable and texture-rich facial region—remain largely unexplored. This project aims to **design and evaluate a Presentation Attack Detection (PAD) system using only forehead creases**, making it robust to partial facial visibility and occlusion (e.g., masks).

2. Project Objectives

The main objectives of this project are:

1. To **create a custom PAD dataset** consisting of:
    - Bona fide (real) forehead images
    - Printed photo attacks
    - E-display (screen replay) attacks
    - Synthetic (GAN-generated) attacks
2. To **extract and analyze forehead crease patterns** for distinguishing real and fake presentations.
3. To **train and evaluate a PAD model** using forehead-only information.
4. To study the **effectiveness of forehead creases** as a soft biometric cue for PAD.

3. Dataset Preparation

3.1 Bona Fide (Real) Dataset (I will provide you)

- Capture real face images using a mobile phone or webcam.
- Extract **forehead region only** using:
    - Manual cropping or

    - Face landmark–based cropping.
- Ensure variations in:
    - Illumination
    - Expression
    - Distance

Label: real

3.2 Printed Attack Dataset (Label: print_attack)

- Print real face images on paper.
- Capture photos of the printed images using a camera.
- Crop only the **forehead region**.

3.3 E-Display (Replay) Attack Dataset (Label: edisplay_attack)

- Display forehead images on:
    - Mobile phone, and Laptop screen
- Re-capture the displayed images using a camera.
- Crop extra region to get complete forehead area.

3.4 Synthetic Attack Dataset (Label: synthetic_attack)

- Generate fake forehead with creases images using:
    - GAN-generated faces (StyleGAN and Diffusion based model)
    - Online synthetic face generators
- Crop forehead region.

4. Model Design (PAD Classification)

Task Definition (Classification)

- **Binary classification**: Bona fide (real), Attack (print + e-display + synthetic)
- **Multi-class classification**: Real, Print attack, E-display attack, Synthetic attack

5. Recommended Models

- CNN-based model (ResNet50 / MobileNet-V2, Efficient-Net)
- Vision Transformer based models(ViT, Swim Transformer, and new models like vision mamba, mamba vision etc)
- Foundational model (DINO, DINO-V2, CLIP)
- Build new architecture model and compare the performance of existing models(CNN, Vision transformer and foundational models)

6. Evaluation Metrics

- Accuracy, Precision, Recall, F1-score
- APCER (Attack Presentation Classification Error Rate)
- BPCER (Bona Fide Presentation Classification Error Rate)
- ACER

Project-ID: P13

Project Mentor: Geetanjali Sharma (d19062@students.iitmandi.ac.in)

Project title : Brownian Bridge Diffusion–Based Instance Segmentation of Forehead Creases (New Work)

Problem Statement : Forehead creases are an important soft biometric feature, especially when facial regions are partially occluded. However, instance-level segmentation of forehead creases is challenging due to their fine structure, low contrast, overlapping patterns, and variations caused by illumination, expressions, and presentation attacks. Existing segmentation methods often fail to preserve precise boundaries and structural consistency. Standard diffusion models also introduce unnecessary randomness, leading to blurred or unstable predictions. Therefore, there is a need for a structure-aware segmentation approach that can accurately capture forehead crease instances.

Project Objectives

1. To develop a **Brownian Bridge Diffusion–based model** for instance segmentation of forehead creases.
2. To improve **boundary accuracy and structural preservation** of fine crease patterns.

Methodology

- Extract forehead region features using a deep convolutional backbone.
- Formulate instance segmentation as a **Brownian Bridge diffusion process** conditioned on ground-truth crease masks.
- Perform guided denoising from noise to structured crease instances.
- Generate final instance masks using the denoised representations.

Performance Evaluation Metrics

- Dice Coefficient, Intersection over Union (IoU), Precision, Recall, and F1-score
- Boundary-based accuracy (for fine crease evaluation)

Dataset : Provided to you on request.

Project-ID: P14,
Project Mentor: Geetanjali Sharma (d19062@students.iitmandi.ac.in)

Project title : LLM-Based Quality Assessment of Segmented Forehead Creases for Reliable Biometric Recognition (New Work)

Problem Statement : Although accurate instance segmentation of forehead creases is essential for biometric recognition, not all segmented forehead regions are suitable for reliable matching. Poor-quality segmented images may lack visible crease patterns due to low resolution, blur, illumination variations, facial expressions, occlusion, or presentation attacks. Using such low-quality samples for recognition can significantly degrade system performance. Existing quality assessment methods are mostly handcrafted or CNN-based and fail to capture semantic and structural cues related to crease visibility. Therefore, there is a need for an intelligent, semantic-aware quality assessment framework that can automatically evaluate whether segmented forehead images contain sufficient and reliable crease information for further biometric processing.

Project Objectives

1. To design a **quality assessment framework** for segmented forehead crease images.
2. To fine-tune a **Qwen-3 Large Language Model (LLM)** for predicting the quality of segmented forehead creases.
3. To classify segmented samples into **high-quality and low-quality** categories based on crease presence and clarity.
4. To select only **high-quality samples** for downstream forehead-based biometric recognition.

Methodology

- Perform **forehead crease instance segmentation** using the Brownian Bridge Diffusion model.
- Construct a **quality-labeled dataset** of segmented forehead images (e.g., clear creases, faint creases, no creases, noisy segmentation).
- Convert visual features and segmentation statistics into **LLM-compatible prompts** describing crease visibility, continuity, and density.
- Fine-tune the **Qwen-3 model** to predict the quality level of each segmented forehead image.
- Filter and retain only **high-quality segmented samples** for biometric recognition.

Evaluation Metrics

- Quality classification accuracy.
- Precision, Recall, and F1-score (High vs Low quality).
- Quality metrics: SSIM, PSNR

Project-ID: P15

Project Mentor: Geetanjali Sharma (d19062@students.iitmandi.ac.in)

Project title: Brownian Bridge Diffusion–Based Anomaly Detection for Iris Presentation Attack Detection

Problem Statement - Iris presentation attack detection (PAD) is a critical component of biometric security systems, aimed at distinguishing bona fide iris samples from spoofing attacks such as printed images, e-display attacks, and synthetically generated irises. Although recent approaches based on convolutional neural networks, transformer models, and foundation models have achieved promising performance, they primarily rely on supervised classification and often fail to generalize to unseen attack types and sensor variations. These methods tend to learn decision boundaries rather than the intrinsic structure of genuine iris textures, leading to performance degradation under cross-dataset and cross-attack scenarios.

Diffusion models offer a generative framework capable of learning complex data distributions. However, their potential for iris PAD has not been fully explored. Standard diffusion processes introduce unnecessary stochasticity and lack structural constraints, which limits their effectiveness in modeling fine iris texture patterns. Therefore, there is a need for a structure-aware and distribution-driven PAD framework that can robustly detect presentation attacks without relying solely on attack-specific supervision.

Project Objectives

1. To develop a **Brownian Bridge Diffusion–based framework** for iris presentation attack detection.
2. To model the **intrinsic distribution of bona fide iris textures** using a structure-preserving diffusion process.
3. To detect presentation attacks as **anomalies** based on diffusion reconstruction and denoising behavior.
4. To improve robustness against **unseen attack types** and cross-sensor variations.

Methodology

- Preprocess and normalize iris images using a standard iris segmentation and normalization pipeline.
- Train a **Brownian Bridge Diffusion Model** using only bona fide iris samples, where the diffusion process is conditioned to bridge from noise to clean iris textures.

- During inference, apply the trained diffusion model to both genuine and attack samples and analyze denoising trajectories.
- Compute a **PAD score** based on reconstruction error, denoising stability, or diffusion consistency.
- Classify samples as bona fide or attack using a threshold on the PAD score.

Evaluation Metrics

- Attack Presentation Classification Error Rate (APCER)
- Bona fide Presentation Classification Error Rate (BPCER)
- Average Classification Error Rate (ACER)
- Equal Error Rate (EER)
- Detection Error Tradeoff (DET) curves

Ablation Study 1 (Diffusion type comparison)
This ablation study justifies *why Brownian Bridge diffusion* is needed.

- Standard DDPM
- Conditional Diffusion
- Brownian Bridge (Proposed)

Ablation Study 2 (Diffusion type comparison)
This ablation study shows efficiency–accuracy tradeoff.

- Number of denoising steps: 10 / 25 / 50 / 100
- Compute all Evaluation metrics and inference time.

Project-ID: P16
Project Mentor: Sushovan

Project Title: RL based finetuning of lightweight Video LLMs to enhance alignment to Blind and Low Vision focused descriptions

Problem Statement:

Video question-answering on live videos is being done mostly by Video LLMs, but the general pretraining of LLMs makes them respond to queries which are readable and understandable by normal human beings and not much by the Blind and Low Vision (BLV) community. The BLV focused descriptions should follow the guidelines that is followed by major media platforms like Netflix, Amazon, etc. and that is because the BLV user needs more detailed, specific information of the objects in the scene, with spatial awareness which the general VLMs may lack. On top of that as our end application is to be deployed in a mobile device, lightweight VLMs may be lower in performance than the larger ones. The solution is to finetune the small VLMs using alignment techniques like RLHF or DPO on the query and response pairs generated by the larger VLMs prompted with accessibility guidelines.

Methodology:

Adopt the paper's adaptive LUV-space differencing for 3-4 query-relevant keyframes, enabling efficient video processing. Curate BLV pairs from AVCaps/Charades + human annotations: preferred (detailed BLV: "Person turning left 2m ahead, uneven pavement") vs. rejected (generic: "Person walking"). Augment with RAG for ambience/social contexts. Use TRL/DPO on SmolVLM2-500M: reward model scores BLV alignment (e.g., via custom LLM evals from paper); LoRA/QLORA for consumer GPU efficiency.

Evaluation:

Use paper's frameworks on AVCaps/Charades held-out sets + MVBench subset for temporal reasoning; human BLV rater prefs and CIDEr/BLEU. Ablate RL vs. SFT; deploy on-device for latency/FPS.

Background : We have already baselined SmolVLM variants for Blind and Low-Vision accessibility and deployed on mobile devices, assessing their performance and latency.

Published the work - https://aclanthology.org/2025.mmloso-1.8/

This work would build on top of that for more novelties.

Project-ID: P17
Project Mentor: Sushovan

Project Title: RL based finetuning of lightweight Video LLMs to enhance alignment to Blind and Low Vision focused descriptions

Problem Statement:

Video question-answering on live videos is being done mostly by Video LLMs, but the general pretraining of LLMs makes them respond to queries which are readable and understandable by normal human beings and not much by the Blind and Low Vision (BLV) community. The BLV focused descriptions should follow the guidelines that is followed by major media platforms like Netflix, Amazon, etc. and that is because the BLV user needs more detailed, specific information of the objects in the scene, with spatial awareness which the general VLMs may lack. On top of that as our end application is to be deployed in a mobile device, lightweight VLMs may be lower in performance than the larger ones. The solution is to finetune the small VLMs using alignment techniques like RLHF or DPO on the query and response pairs generated by the larger VLMs prompted with accessibility guidelines.

Methodology:

Adopt the paper's adaptive LUV-space differencing for 3-4 query-relevant keyframes, enabling efficient video processing. Curate BLV pairs from AVCaps/Charades + human annotations: preferred (detailed BLV: "Person turning left 2m ahead, uneven pavement") vs. rejected (generic: "Person walking"). Augment with RAG for ambience/social contexts. Use TRL/DPO on SmolVLM2-500M: reward model scores BLV alignment (e.g., via custom LLM evals from paper); LoRA/QLORA for consumer GPU efficiency.

Evaluation:

Use paper's frameworks on AVCaps/Charades held-out sets + MVBench subset for temporal reasoning; human BLV rater prefs and CIDEr/BLEU. Ablate RL vs. SFT; deploy on-device for latency/FPS.

Background : We have already baselined SmolVLM variants for Blind and Low-Vision accessibility and deployed on mobile devices, assessing their performance and latency.

Published the work - https://aclanthology.org/2025.mmloso-1.8/

This work would build on top of that for more novelties.

Project-ID: P18
Project Mentor: Sushovan

Project Title: Enhancing Explainability in MRI and CT Imaging using Foundation models

Problem Statement:

This project proposes finetuning a vision-language foundation model to improve explainability in MRI and CT scans by generating natural language rationales alongside visual attention maps for diagnostic decisions. Building on recent advances in multimodal models, it adapts 2D/3D vision-language models (VLMs) for medical imaging to provide interpretable outputs trusted by clinicians. The approach leverages pre-trained foundation models like LLaVA-Med or MedBLIP, fine-tuned for explainable abnormality detection and report generation. Recent CVPR, ECCV, and ICCV papers demonstrate VLMs' potential in medical imaging. For instance, "Adapting Vision-Language Models for 3D CT/MRI Understanding" (ICCV 2025 workshop) introduces slice selection and instruction tuning to boost F1 scores from 0.07 to 0.53 in diagnostics, enhancing interpretability via task-aligned reasoning. SilVar-Med (CVPR 2025 workshop) pioneers speech-driven VLMs with reasoning datasets for abnormality interpretation, addressing black-box limitations in clinical settings. Merlin (related foundation model) processes full 3D CT volumes with EHR integration for report generation, emphasizing multimodal pre-training.

Methodology:

Adapt open-source VLMs (e.g., 3D-CT-GPT or MedBLIP) using LoRA fine-tuning on paired image-report data, incorporating attention mechanisms like Grad-CAM for visual highlights and text prompts for rationales (e.g., "Explain tumor localization in this MRI"). Input: 3D MRI/CT volumes with radiology reports; output: classification, heatmap, and textual justification. Training involves contrastive alignment and instruction tuning on synthetic diagnostic dialogues for faithfulness.

Dataset:

Primary datasets include TCIA (The Cancer Imaging Archive) for diverse MRI/CT cancer scans with annotations (e.g., brain tumors, lung nodules) and RSNA Abdominal Trauma CT (4,274 studies) for injury detection/segmentation. Supplementary: BraTS for MRI brain tumors and CT-RATE for chest CT with grounded VQA pairs. These provide de-identified, annotated volumes suitable for VLM pre-training and evaluation.

Evaluation:

Model performance uses AUROC, F1-score, and clinical utility via Cox PH for outcome prediction. Explainability metrics include Grad-CAM localization accuracy (IoU with ground-truth), perturbation faithfulness (prediction drop on masked regions), and XAlign for explanation fidelity. Human evaluation by radiologists assesses rationale coherence via BLEU/ROUGE for reports and Likert scores for trust.

Project-ID: P19

Project Mentor: Bhavesh Kapil (d24023@students.iitmandi.ac.in)

Project Title

Uncertainty-Guided Self-Supervised Learning for Missing-Modality Brain MRI (Reconstruction + Downstream Segmentation)

Problem Statement

Multi-modal brain MRI (e.g., T1, T1ce, T2, FLAIR) is central to reliable tumor analysis and clinical decision-making. However, in real-world settings, one or more modalities are frequently missing due to differences in acquisition protocols, scanner availability, patient motion, time constraints, cost, or corrupted sequences. This missing-modality issue causes a major performance drop in downstream tasks such as tumor segmentation, grading, and response assessment, since most modern pipelines assume a complete set of modalities during both training and inference.

Traditional supervised approaches for modality synthesis or imputation typically require paired, fully-available multi-modal ground truth, which is expensive and often unrealistic at scale. Moreover, these methods usually produce a single deterministic reconstruction and do not quantify confidence, making them risky in safety-critical medical settings especially under domain shift scanners or rare pathology patterns.

This project addresses the challenge of learning robust, generalizable representations and reconstructions under missing modalities without relying on exhaustive manual annotations. The key idea is to use self-supervised learning to learn cross-modal structure from unlabeled MRI volumes, while explicitly incorporating epistemic uncertainty to (i) identify unreliable pseudo-targets/regions, (ii) weight learning signals, and (iii) enable uncertainty-aware downstream inference.

Objective

Develop an uncertainty-guided self-supervised framework that can:
1. Infer / reconstruct missing MRI modalities from available ones (cross-modal completion).
2. Learn modality-robust 3D representations that remain effective when modalities are absent at test time.
3. Improve downstream tasks (especially 3D tumor segmentation) under missing-modality conditions.
4. Produce voxel-level and/or volume-level uncertainty maps that correlate with reconstruction/segmentation errors for reliability assessment.

Resources

1. https://www.sciencedirect.com/science/article/pii/S0925231226000561?via%3Dihub
2. https://ieeexplore.ieee.org/abstract/document/10984423
3. https://ieeexplore.ieee.org/abstract/document/10444695

Project-ID: P20

Project Mentor: Bhavesh Kapil (d24023@students.iitmandi.ac.in) and Parul Chaudhary(s23109@students.iitmandi.ac.in)

Project Title

Self-Supervised Learning for Robust Deepfake Detection Under Domain Shift

Problem Statement

Deepfakes, synthetically generated or manipulated face videos and images, pose serious risks to digital trust, enabling misinformation, identity fraud, and reputational harm. Modern deepfake generation methods (GANs, diffusion-based face swapping, reenactment, and neural rendering) are improving rapidly, making artifacts harder to detect. While supervised deepfake detectors achieve strong performance on known datasets, they often fail to generalize to: (i) unseen manipulation methods, (ii) new camera/compression pipelines (social media), (iii) different lighting/pose demographics, and (iv) low-quality, heavily compressed videos. Additionally, collecting and labeling large-scale deepfake datasets for every new manipulation type is costly and quickly becomes outdated.

This project addresses the challenge of building a generalizable deepfake detection system with reduced dependency on labeled deepfake data. The key idea is to leverage self-supervised learning (SSL) on large-scale unlabeled real videos/images to learn manipulation-invariant but forensic-sensitive representations. By learning robust spatio-temporal and frequency-domain features through SSL pretext tasks (e.g., masked modeling, contrastive learning, temporal consistency), the detector can better identify subtle inconsistencies in face dynamics, texture, and compression—improving performance on unseen deepfake types and distribution shifts.

Objectives

1. Learn strong forensic representations using SSL on unlabeled real face videos/images.
2. Improve cross-dataset generalization for deepfake detection (train on one dataset, test on another).
3. Detects deepfakes across diverse conditions: compression, low resolution, occlusion, motion blur, and varied demographics.
4. Provide localization cues (frame-level or region-level heatmaps) indicating manipulated regions.

Resources:

1. https://www.mdpi.com/3241072
2. https://www.sciencedirect.com/science/article/pii/S026288562500006X?via%3Dihub
3. https://arxiv.org/abs/2511.17181
4. https://arxiv.org/abs/2511.17181

Project-ID: P21

Project Mentor: Bhavesh Kapil (d24023@students.iitmandi.ac.in) and Pushap

Project Title

Uncertainty Estimation for EEG Emotion Recognition Using the Foundation Model

Problem Statement

EEG-based emotion recognition is highly sensitive to subject variability, noise, and dataset shift, causing deep models (including foundation-model backbones) to produce overconfident but incorrect predictions. This is unsafe for affective BCIs and clinical/behavioral monitoring, where incorrect emotional state inference should be flagged or abstained rather than presented with high confidence.

This project proposes a reliability-focused pipeline that uses the EEG foundation model as a transferable representation backbone and augments it with principled uncertainty quantification (UQ) and calibration, enabling the system to: predict emotion labels (valence/arousal), and 2) report how confident it is (epistemic/aleatoric proxies) under cross-subject and corrupted settings.

Objectives:

1. Model fine-tuning code (linear probe + adapters/LoRA).
2. UQ module comparison: MC-dropout vs ensembles vs TTA (+ optional conformal).
3. Reliability report: calibration tables, risk–coverage curves, per-subject performance.

Resources:

1. https://arxiv.org/abs/2405.18765
2. https://ieeexplore.ieee.org/document/11271584
3. https://www.biorxiv.org/content/10.1101/2025.07.09.663220v2
4. https://doi.org/10.1016/j.neunet.2025.107363

Project-ID: P22

Project Mentor: Arya Pulkit (S23084@students.iitmandi.ac.in) and Asif Hoda

Project Title: Advancing Reliable Reasoning in SLMs Through Hallucination Reduction.

Problem Statement:

Small language models are more prone to hallucinations, while also exhibiting weaker reasoning capabilities. To improve reasoning, we often rely on distillation from larger LLMs. However, due to the parametric constraints of small models, this distillation process can further amplify hallucinations. Our goal is to address and resolve this trade-off by developing methods that improve reasoning without increasing hallucination.

Objective:

To design and evaluate methods that improve reasoning capabilities in small language models while minimizing hallucinations, by developing constraint-aware reasoning distillation techniques that transfer useful reasoning behavior from large language models without overloading the limited parametric capacity of smaller models. The objective includes balancing reasoning depth, factual consistency, and generalization

Project-ID: P23

Project Mentor: Arya Pulkit

Project Title: Generalised DeepFake Video Detection.

Problem Statement:

Current deep learning-based video deepfake detection methods face challenges in detecting diverse fake content, as they often generalize well within the same model family but struggle across different families. Many methods are fitted to specific datasets or patterns, limiting their adaptability to new generative models.

Objective:

Our aim is to address the growing challenge of distinguishing real videos from highly realistic AI-generated counterparts and develop a generalised deepfake detection approach.

Project-ID: P24

Project Mentor: Arya Pulkit

Project Title: Explanablity in Generalised DeepFake Video Detection using Foundation Models

Problem Statement:

DeepFake generation techniques are rapidly evolving, making existing video detection systems brittle and poorly generalizable to unseen manipulation methods. While foundation models improve detection performance across diverse DeepFake types, they often operate as black boxes, providing limited insight into why a video is classified as fake. This lack of explainability reduces trust, hinders error analysis, and limits adoption in high-stakes domains such as media forensics and digital security.

Objective:

Develop an explainable DeepFake video detection framework using foundation models that generalizes to unseen manipulation techniques while providing interpretable, human-understandable explanations for its predictions.

Project-ID: P25

Project Mentor: Parul Chaudhary (s23109@students.iitmandi.ac.in) and Prof. Arnav Bhavsar Vinayak (arnav@iitmandi.ac.in)

Project Title: Forensic Explainability for AI-Generated Image Detection

Problem Statement: AI-generated images from GANs and diffusion models are increasingly indistinguishable from authentic photographs, threatening digital media integrity across journalism, legal systems, and social platforms. Current deep learning detectors can flag images as synthetic with high accuracy but fail to explain *what specific characteristics* triggered the classification, whether suspicious facial asymmetries, unnatural textures, abnormal lighting patterns, or hidden algorithmic fingerprints. This opacity is unacceptable in forensic contexts where investigators must understand *why* an image was flagged: courts require evidence beyond binary verdicts, journalists need verifiable manipulation indicators, and content moderators must justify decisions to users. Existing explainable approaches either highlight suspicious spatial regions without revealing underlying generation artifacts, or expose frequency-domain anomalies without connecting them to visible manipulations, leaving critical gaps in forensic understanding of what actually reveals an image as AI-generated.

Objective: Develop an explainable detection framework using CLIP's vision transformer that not only classifies images as real or AI-generated, but explicitly identifies and visualizes the spatial artifacts (facial inconsistencies, texture anomalies, boundary distortions) and frequency-domain fingerprints (spectral periodicities, upsampling signatures) that flag the image as synthetic, providing comprehensive forensic evidence of what characteristics expose AI generation across diverse GAN and diffusion-based models.

Resources:

1. https://arxiv.org/abs/2404.18649
2. https://proceedings.mlr.press/v119/frank20a/frank20a.pdf
3. https://arxiv.org/abs/2409.07913
4. https://arxiv.org/abs/2503.20188
5. Efficient Explainable Face Verification Based on Similarity Score Argument Backpropagation
6. SIDA: Social Media Image Deepfake Detection, Localization and Explanation with Large Multimodal Model
7. Towards Universal Fake Image Detectors That Generalize Across Generative Models

Project-ID: P26

Project Mentor: Parul Chaudhary (s23109@students.iitmandi.ac.in) and Prof. Arnav Bhavsar Vinayak (arnav@iitmandi.ac.in)

Project Title: LocalForge: A Fine-Grained Dataset for Image Manipulation Detection and Localization

Problem Statement:

Existing deepfake detection datasets predominantly focus on holistic manipulations (complete face synthesis, full face-swaps) or object-level edits (adding/removing entire objects), overlooking the fine-grained attribute manipulations increasingly prevalent in targeted misinformation: changing clothing colors to misrepresent individuals at specific events, altering accessories to fabricate associations, modifying facial attributes to manipulate perceived identities, or adjusting textures to create false narratives. These subtle, attribute-level forgeries are operationally challenging to detect yet forensically critical, as adversaries exploit them to create plausible deniability while spreading disinformation. Current benchmarks lack systematic evaluation of detection models' ability to identify and localize such fine-grained manipulations across diverse semantic categories (clothing attributes, facial accessories, texture modifications, background context changes), and no large-scale dataset exists with comprehensive ground-truth masks showing exactly what attributes were manipulated, hindering development of forensic detection systems where identifying *what specific attribute was changed* is as important as detecting *that manipulation occurred*.

Objective:

Create a large-scale benchmark of fine-grained attribute-level forgeries with automatic ground-truth localization masks and evaluate state-of-the-art detection methods to establish baseline performance on attribute-level forgery localization; producing a comprehensive forensic dataset with structured taxonomy of manipulation types (clothing color changes, accessory additions, texture alterations, facial attribute modifications) that challenges existing detectors and enables systematic advancement in localized manipulation detection.

Project-ID: P27

Project Mentor: Kajal (s23083@students.iitmandi.ac.in)

Project Title: Image Generation for Medical Images Using Diffusion Models

Problem Statement:

Deep learning methods in medical imaging require large, diverse, and well-annotated datasets. However, for many diseases and imaging modalities, medical image datasets are scarce or not publicly available due to privacy concerns, ethical restrictions, and high annotation costs. This data limitation hinders the development, generalization, and evaluation of reliable AI models for medical diagnosis and analysis. Therefore, there is a need for effective methods to generate realistic, and clinically meaningful medical images that can supplement limited datasets.

Objective:

The objective of this work is to generate realistic medical images using diffusion-based generative models and to systematically evaluate the quality and clinical usability of the generated samples. Furthermore, the study aims to analyze whether synthetic medical images can effectively support downstream tasks such as classification and segmentation. To validate their practical relevance, segmentation models will be trained and evaluated on the generated dataset, assessing whether the synthetic data preserves meaningful anatomical and disease-related features.

Project-ID: P28

Project Mentor: Kajal (s23083@students.iitmandi.ac.in)

Project Title: Efficient Medical Image Segmentation Using Self-Distillation

Problem Statement:

Medical image segmentation models are often computationally heavy, making them unsuitable for real-time and resource-constrained clinical environments. Despite good accuracy, their large model size and inference latency limit practical deployment for disease-specific medical segmentation tasks.

Objective:

The objective of this project is to apply self-distillation using a teacher–student framework, where the student model learns from the teacher's soft predictions to improve performance. The aim is to compress the model size while preserving segmentation accuracy, enabling efficient and real-time medical image segmentation for disease diagnosis and clinical applications.

Project-ID: P29

Project Mentor: Pushap Singh (erpd2201@students.iitmandi.ac.in) and Prof. Arnav Bhavsar Vinayak (arnav@iitmandi.ac.in)

Project Title: Does Cleaner Pretraining Always Win? A Study of Cross-Modal Neural Compression for Brain BioSignal (iEEG & scalp EEG)

Problem Statement:

Neural compression models such as BrainCodec[1] have shown that training on cleaner biosignals (iEEG) can yield representations that generalize better to noisier modalities (scalp EEG), while preserving downstream task performance at high compression ratios. This suggests neural compression can act not only as a storage tool but as a representation bottleneck useful for pretraining or bootstrapping large-scale EEG foundational models.

Objectives:

This project will:

(A) Reproduce the iEEG to EEG transferability claims under varied, controlled settings.
(B) Evaluate two candidate pipelines for leveraging compressed signals for foundation model style pretraining:

> (a) Pretrain a transfer pipeline and integrate a foundation model in latent space.
> (b) Train a shallow combined pipeline and progressively add capacity as noisier data is included.

Deliverables:

1. Reproduce training and testing pipeline of [1].
2. Extension of the compression pipeline to a foundational model similar to Latent Diffusion modelling [2].

References:

1. The Case for Cleaner Biosignals: High-fidelity Neural Compressor Enables Transfer from Cleaner iEEG to Noisier EEG | OpenReview
2. High-Resolution Image Synthesis With Latent Diffusion Models

Project-ID: P30
Project Mentor: Munish Daroch

Project Title: Multi-Modal Medical Image Fusion Using Generative Models.

Problem Statement:
Multi-modal medical imaging modalities such as CT and MRI provide complementary anatomical and functional information critical for clinical diagnosis. However, single-modality analysis often fails to capture complete structural and pathological characteristics. Existing medical image fusion approaches typically rely on handcrafted features or supervised learning, requiring paired data or manual annotations, which limits scalability and generalization.

This project aims to develop a generative framework for multi-modal medical image fusion using Deep generative models. By learning shared latent representations from unannotated CT, MRI, and other imaging modalities, the proposed method aims to generate fused images that preserve structural consistency and diagnostically relevant information across modalities, enabling further use in disease diagnosis.

Resources:

- **Dataset:** SynthRAD2023 Grand Challenge dataset: generating synthetic CT for radiotherapy [**LINK**], The Harvard public medical dataset [**LINK**]
- **Baseline:** Mask-difuser: A masked diffusion model for unified unsupervised image fusion[**LINK**], Fusionmamba: Dynamic feature enhancement for multimodal image fusion with mamba[**LINK**]
- **Evaluation of Metrics:** Visual Information Fidelity (VIF), Structural Similarity Index (SSIM), Multi-Scale SSIM (MS-SSIM), Standard Deviation (SD), and Entropy (EN).

Project-ID: P31
Project Mentor: Munish Daroch

Project Title: Anatomy-Aware Denoising Framework for Low-Dose CT at Variable Dosage.

Problem Statement:
Low-Dose CT (LDCT) imaging is essential for reducing patient radiation exposure, but lowering the X-ray dose significantly increases noise, reducing diagnostic quality. Clinical practice requires balancing radiation safety with image fidelity, especially in oncology and lung screening, where fine anatomical details are critical.

Conventional denoising approaches trained with L1/L2 pixel-wise losses tend to oversmooth images, erasing subtle but clinically relevant structures. While perceptual losses such as SSIM improve structural similarity, they remain insensitive to the fine anatomical details that radiologists rely upon.

This project proposes an anatomy-aware denoising framework for LDCT reconstruction. We will first simulate LDCT images at multiple dose levels (10%, 20%, 25%, 50%, 70%) using high-dose CT scans from the reference dataset. Then, a deep learning model will be designed to restore high-dose CT quality from these LDCT inputs. To ensure clinical safety, the framework will incorporate a novel anatomy-aware loss function that prioritizes preserving structures such as lung parenchyma, vessels, and tumor regions, and will be compared against traditional L1, L2, and SSIM-based objectives.

The goal is to develop a denoising pipeline that balances noise suppression with anatomical fidelity, validated through both quantitative image metrics and anatomical/clinical relevance assessments.

Resources:

- **Dataset:** High-Dose CT scans (public datasets such as AAPM Low-Dose CT Grand Challenge **[LINK]**, LoDoPab CT**[Link]**, or institutional NDCT data).
- **Simulation Tool:** LDCT simulation using noise insertion at 10%, 20%, 25%, 50%, 70% dose levels: ASTRA toolbox.
- **Baseline Models:** FoundDiff: Foundational Diffusion Model for Generalizable Low-Dose CT Denoising [**LINK**], ASCON **[LINK],** Anatomy-Aware Low-Dose CT Denoising via Pretrained Vision Models and Semantic-Guided Contrastive Learning **[LINK],**
- **Evaluation Tools:** Image similarity metrics (PSNR, SSIM, RMSE)

Project-ID: P32
Project Mentor: Munish Daroch

Project Title: 3D MRI-to-Synthetic CT Translation Using Diffusion-Based Generative Models

Problem Statement:
Computed Tomography (CT) plays a critical role in clinical applications such as radiation therapy planning and surgical guidance due to its accurate tissue density representation. However, CT imaging involves ionizing radiation, which limits its repeated use. Magnetic Resonance Imaging (MRI) is radiation-free and provides superior soft-tissue contrast but lacks electron density information required for CT-dependent clinical workflows.

MRI-to-synthetic CT (sCT) translation seeks to generate CT-equivalent images from MRI data, enabling MRI-only clinical pipelines. Existing supervised approaches based on convolutional neural networks or generative adversarial networks often suffer from limited structural consistency, poor bone representation, and training instability, particularly in 3D volumetric settings. Moreover, capturing fine-grained anatomical details across entire volumes remains challenging.

This project aims to develop a supervised 3D diffusion-based generative framework for MRI-to-CT translation using paired MRI–CT volumes. By modeling the conditional distribution of CT images given MRI inputs, the proposed method seeks to generate high-fidelity 3D synthetic CT volumes with improved anatomical accuracy and tissue density estimation.

Resources:

- **Dataset:** SynthRAD2023 Grand Challenge dataset: generating synthetic CT for radiotherapy [LINK]
- **Baseline Models:** Synthetic CT generation from MRI using 3D transformer-based denoising diffusion model [LINK]
- **Evaluation Tools:** Image similarity metrics (PSNR, SSIM, RMSE)

Project-ID: P33
Project Mentor: Sayan Shaw, Dr. Sneha Singh

Project Title: Robust Preprocessing of Multimodal Medical Images for Outcome Enhancement Using a Hybrid Architectures

Problem Statement:

Medical image analysis involves several important tasks such as classification, segmentation, synthesis, fusion etc etc. The performance of all these tasks strongly depends on one fundamental step → **accurate image alignment.** In multimodal imaging scenarios involving CT, MRI, and PET, each modality offers complementary information, and effective integration of this data depends on accurate spatial alignment. Without accurate registration, downstream tasks may suffer from poor generalizability, reduced accuracy, and unstable performance across different datasets and modalities.

In this project, we aim to develop a **novel and robust preprocessing framework** for unimodal as well as multimodal medical images based on deformable registration and B-spline interpolation techniques. The objective is to design a novel registration approach that can handle nonlinear deformations, intensity variations across modalities, and preserve important anatomical structures during alignment. The proposed framework will focus on improving **alignment accuracy**, **computational complexity**, and **structural consistency**, thereby enhancing the performance of above mentioned downstream tasks.

Evaluation Metrics:

The proposed framework will be evaluated using quantitative metrics such as Dice Similarity Coefficient (DSC), Target Registration Error, Structural Similarity Index (SSIM), Average Symmetric Surface Distance, Jacobian determinant regularity etc to assess alignment accuracy. In addition, qualitative evaluation will be performed through visual inspection by project mentors and/or domain experts to validate the results.

Datasets: will be attached shortly

Papers to Start With:
non-rigid registration, accelerating B-spline Interpolation on GPUs, deep learning based registration, a novel loss function for registration, deformable registration, deep feature matching based B-spline

Project-ID: P34
Project Mentor: Sumit Maan, Dr. Sneha Singh

Project Title: Tissue Segmentation, lesion characterization and Outcome Prediction in Rectal Cancer using Histopathology Images

Problem Statement:

1. Accurate prognostic prediction in rectal cancer patients undergoing direct surgical resection remains limited due to the inadequate predictive power of conventional TNM staging and isolated clinicopathologic variables. Existing models fail to effectively integrate multiparametric MRI and structured clinical data, and often rely on manual tumor segmentation and single-task frameworks.
2. Conventional histopathology AI models are restricted by task-specificity and limited generalizability across heterogeneous digitization protocols and diverse patient populations due to significant domain shifts. These frameworks often fail to leverage complex contextual interactions within the tissue microenvironment, necessitating the development of robust, general-purpose foundation models such as the Clinical Histopathology Imaging Evaluation Foundation (CHIEF) model for systematic pan-cancer diagnostic and prognostic evaluation.
3. Existing pathology frameworks are often limited by a reliance on task-specific fine-tuning and a failure to integrate multimodal supervisory signals from clinical reports, which restricts their generalizability across gigapixel whole-slide images (WSIs) particularly in low-data regimes like rare cancer diagnosis. This study will addresses these constraints by developing a multimodal foundation model that bridges the semantic gap between local patch embeddings and global clinical endpoints through large-scale vision-language alignment, enabling robust zero-shot classification and cross-modal retrieval without extensive architectural specialization.

   Evaluation Metrics: This study will addresses this deficiency by utilizing Analysis of Variance (ANOVA) to systematically evaluate the impact of critical parameters including layer depth, dropout rates, and attention-based patch selection on diagnostic metrics like F1-score and AUC-ROC, facilitating a transition from heuristic-driven design toward statistically validated and interpretable clinical AI.

Reference Paper and resources:

- A pathology foundation model for cancer diagnosis and prognosis prediction | Nature
- Multitask deep learning model based on multimodal data for predicting prognosis of rectal cancer: a multicenter retrospective study | BMC Medical Informatics and Decision Making | Springer Nature Link
- Understanding the Impact of Deep Learning Model Parameters on Breast Cancer Histopathological Classification Using ANOVA
- A multimodal whole-slide foundation model for pathology | Nature Medicine

Project-ID: P35
Project Mentor: Peeyush Kumar Singh, Dr. Sneha Singh

Project Title: Natural Image reconstruction using fMRI and Diffusion Models

Problem Statement:

Core problem: How can we accurately decode and reconstruct the visual images that a person is viewing directly from their brain activity (fMRI signals), bridging the gap between neural representations and visual perception?
This fundamental challenge in computational neuroscience and brain-computer interfaces involves:

Primary Challenges:

1. Neural-to-Visual Translation Gap: Converting sparse, noisy fMRI voxel activations (~4600 dimensions) into rich, detailed natural images (128×128×3 pixels). Mapping fundamentally different data modalities with no direct correspondence.
2. Information Reconstruction Complexity: Recovering both precise pixel-level details (textures, edges, colors, shapes) AND high-level semantic content (object categories, scenes, spatial relationships) from limited neural signals. Current methods achieve either clarity OR semantic accuracy, but not both simultaneously.
3. Hierarchical Visual Processing Integration: Human visual cortex processes information hierarchically (V1→V2→V3→V4→HVC). Existing models fail to effectively leverage this multi-level brain organization to guide reconstruction.
4. Data Scarcity and Signal Variability: Extremely limited paired fMRI-image datasets (typically 1000-1500 samples). High inter-subject variability and attention fluctuations during scanning sessions.

Quantitative metrics: aHash, Histogram Similarity, Structural Similarity (SSIM)

Datasets: Horikawa17 dataset: 1200 training images across 150 categories with fMRI from 7 brain regions (V1, V2, V3, V4, LVC, HVC, VC) for 5 subjects. Link

Base Paper: Link