Milestone 2: Modeling and Retrieval

Deadline: 31st October [midnight]

Problem Statement

Build a lightweight (preferably CPU-only**) image understanding system** that can:

1. **Classify** a new test images into your 10 chosen object classes,
2. **Predict** relevant visual attributes (e.g., color, material, condition) for it, and
3. **Retrieve** matching images for a short text query like *"blue plastic jug"*.

You will use two datasets:

- **Dataset 1:** your own dataset from Milestone 1
- **Dataset 2:** the **pooled dataset** containing images from all teams (will be provided later)

The goal is to understand how dataset **scale and diversity** affect performance and generalization over different models.

Part A: Model Requirement

Each team will:

1. Pick **two lightweight Vision Transformer variants** (examples: **DeiT-Tiny**, **MobileViT**, **CCT**, **TinyViT**, **CvT-Tiny**, or similar).
    a) You are allowed to use Vision Transformer variants with publicly available ImageNet-pretrained weights (you CANNOT use models pretrained on our specific dataset). However, you must perform fine-tuning on it.
    b) **No marks** if you simply pick a pre-trained model with frozen weights and submit that.
    c) The kind of fine-tuning you wish to perform is left up to individual teams.
2. Train **each model on both datasets** — your own and the pooled one.
3. Strictly keep the same training setup and compare the results.

Part B: Analysis

Each team should provide a clear, reasoned **comparison** of:

1. **Model vs. Model:** how the two transformer variants differ in learning speed, stability, and accuracy.

2. **Dataset vs. Dataset:** how model behavior changes between your small, controlled dataset and the larger pooled dataset.
3. **Insights:** why certain patterns appear (e.g., data imbalance, overfitting, better generalization with more diversity).

To make this meaningful, you must include:

- Accuracy and F1 scores (per class and overall)
- Examples of **misclassifications** or **wrong retrievals** with your explanation
- A few **visuals or plots** such as confusion matrix, retrieval gallery, or 2D feature space (t-SNE/PCA)
- Short **written observations** drawn from your own training and testing — not generic or AI-generated text

The analysis should reflect **your team's understanding** of the models' behavior, not just numbers. Part B is purposefully kept open-ended. It is up to you how you perform a thorough analysis of your experiments.

Part C: Deliverables

Submit:

- **Two trained models** (weights + configs) — each trained on both datasets
- **Brief report (6-10 pages)** containing:
    - Your chosen models and rationale
    - Training details (epochs, time, hyperparameters)
    - Comparative results and your key insights
    - 3–5 visual examples (plots or retrieval outputs)
- **Demo Jupyter notebook** that:
    - Predicts class/attributes for an image, and
    - Performs text→image retrieval (showing top-K results)

- ➢ It is highly recommended to create a small, usable prototype (with a minimalistic UI) of your image understanding system which can be demo-ed live to the class.
- ➢ More details about the deliverables may be shared in due course of time.