Additional Material and Clarifications:

- Restriction on Training Datasets
    a. You are allowed to use any external dataset or augmentation technique for training. Consider the CIFAKE dataset as a template as to how the test set will look like (the class labels, image size (by downsampling) of the test set would follow CIFAKE)

- Test Dataset composition
    a. The test dataset will consist of real images and fake images generated from Stable Diffusion, PixArt, GigaGAN and Adobe Firefly with adversarial perturbations introduced to prevent detection (example reading - [1], [2])

- Evaluation for the task 2
    a. Task 2 explanations are only required for images you have classified as AI generated images in Task 1. You will receive scores only for the subset of AI generated images you have classified correctly.
    b. The first part of the evaluation would be the intersection over union on the subset of the artefacts selected from the list of possible artefacts.
    c. The second part of the evaluation would be based on content similarity with human generated explanations for each selected artefact. This will be done using GPT4.
    d. The explanation score would only be considered for the subset of selected artefacts that match the human annotated ground truth.

- Guidelines for explanation generation for AI generated images
    a. A file with the list of artefacts is attached. (reading on artefacts in AI generated images - [3])
    b. The output must be generated as a json with a list of dictionaries. Each dictionary has the artefact names are keys and an explanation as to why the image has this artefact as the values.
    c. The explanations should be limited to 50 words.
    d. A general location in the image (top left, top right, etc.) can be provided for artefacts that can be localized.

- Model-specific information
    a. Task I: smaller model (millions of parameters) + low latency
    Example:
    Standard architectures like: **ResNet-50**.
    (Parameters: ~23M, **Inference Time**: ~200ms on CPUs (excluding model loading time), **Size**: ~98MB**.)**
    Note: anything lower would attract more points, and the latency/efficiency scoring will be relative across teams
    b. Task II: Teams are allowed to open-source large language models (upto a model size of ~13B parameters e.g. Llama, Mistral family of models), but third-party APIs are not allowed

- Overall evaluation (100%)
    a. Solution Accuracy (50%): Performance metrics as specified for each task.
        i. Task I Classification Accuracy: 25%
        ii. Task II Classification Accuracy: 40%
        iii. Task II Explanation correctness: 35%

    b. Approach (35%): Innovative techniques, efficiency, and generalizability of the solution.
        i. Task I Innovation: 10%
        ii. Task I Efficiency: 40%
        iii. Task II Innovation: 50%

    c. Presentation (15%): Final presentation, Quality of documentation, clarity of code

Note: Innovation involves aspects like - Novelty, Technical depth, Impact Potential e.g. dealing with adversarial perturbations, pipelines for artefact selection, scalability of the proposed approach, etc.