PROBLEM STATEMENT On AIGC

Problem Statement:

- Create a data filtering pipeline that could remove low-quality text-video pairs from the dataset and thus improve video generation quality with less cost of model training.

Background :

- AI Generated Content (AIGC) has reshaped the way of creation and innovation in the past year of 2023 and is going to have significant impacts in the coming 2024. However, the heavy model training process has generated significant amounts of CO2 emission and thus raised the conciseness of governments. The performance of the generative model is heavily dependent on the quality of the dataset: datasets with imperfections could slow down the model convergence speed and result in sub-optimal solutions. Thus, high-quality compact datasets could reduce the training loop for getting high-quality generative models.

Requirements (where applicable) :

- Students participating in this project must fulfill the following requirements
- Familiar with Python and programming basics: data structure, deep neural network and Pytorch

Goals and Objectives:

- A data mining pipeline that could select 10k text-video pairs that maximize the video generation performance of a pre-trained video generation model. To achieve this, instead of actually fine-tuning the model, the following metrics are used:
    - CLIP Score: the quality of the selected dataset should be accessed based on its CLIP score between the re-generated text descriptions.
    - Video based performance benchmarking: https://github.com/mbzuai-oryx/Video-ChatGPT/tree/main/quantitative_evaluation

Dataset:

We will provide a noisy dataset with 1M text-video pairs.

Judging Criteria -

A panel of peer judges spanning across our product, operations, engineering, security, and design teams was charged with evaluating projects based on three key criteria:

1. **Completeness** - Successfully executing the project concept, implementing the main features/functionality, and creating a usable and well-designed product with a smooth experience
2. **Creativity / Innovation** - unique, different, or novel compared to other similar solutions in terms of data mining algorithm, evaluation metrics, and other relevant aspects.
3. **Technical Accomplishment** - The final score based on the provided benchmark. This is important for the accessment of technical accomplishment and the project value.
4. **Product value/Functionality** - a solution that effectively addresses the recognized problem and has potential for impact and usefulness.
5. **Quality of presentation (For Final)**
    1. successful demo of the hack. Good coverage of the problem, solution idea, objective, main functionalities of the hack, and other key content.
    2. handling the Q&A session well and answering questions to the judges satisfaction.

Completeness	20%
Creativity / Innovation	20%
Technical accomplishment	30%
Product value / Functionality	30%
Polishness of the solution	20%