生成式人工智慧 HW6

學號:313512072 姓名:洪亮

➢ Training

一、訓練環境

- 硬體：單張 NVIDIA A6000（Ada）
- 軟體：DeepSpeed + Hugging Face Accelerate

二、訓練流程

- **編碼階段:**使用 VAE 與 text encoder 將 prompt 與 ground-truth 一起壓縮到 latent space。
- **加噪與去噪:**在潛在張量上套用 DDPM 添加雜訊。將帶噪潛在張量加上文字條件，一起輸入模型進行去噪重建。
- **解碼階段:**將去噪後的潛在特徵通過 VAE Decoder 還原為最終影像。

三、訓練優化工具

- **EMA**：對模型參數進行滑動平均，以提升收斂穩定性。
- **Classifier-Free Guidance**：在推理時透過隨機遮蔽 text condition，增強樣本多樣性與品質。
- **Cosine Learning Rate Scheduler**：採用餘弦退火策略動態調整學習率。

四、訓練策略

- **跳出局部極值:**若訓練停滯或 Loss 長期不降，將學習率調大讓模型脫離局部極值的區域

➢ Inference

首先將多個 prompt 組成 batch，使用 CLIP text encoder 將 prompt 映射至潛空間中，並隨機採樣高斯分布透過 DDPM schedule 進行去噪，最後透過 VAE 還原成圖片，並將像素正規化，輸出最終圖片。

➢ 結果討論

模型:DiT vs unet

一開始先嘗試使用 DiT 模型進行訓練，因為想說其基於 Transformer 的架構能更有效地捕捉圖像中的全局關聯，但產出來的圖片都趨近模糊，一開始以為是訓練時間不夠長導致，但產出來的圖片還是偏模糊，後來在換成 unet 模型，出來的圖片就相對清晰許多

Predict Type:noise vs image

一開始的策略是預測圖片來計算 loss，但效果並不理想，出來的圖片很多都看不清楚，改成計算噪聲的 loss 後，圖片的輪廓就相對比較明顯，所以後來都改成預測噪聲

DDPM vs DDIM

在推理階段有嘗試使用 DDIM，雖然速度變快，但導致圖像細節變得比原本更模糊，所以最終仍選擇使用 DDPM，雖然生成速度較慢，但能產生較為清晰的圖片。

Timestep:

在推理階段嘗試使用不同的 timestep 值，發現值太小，出來的圖片幾乎都是石巨人，根據後面結果，timestep 值趨近 70 到 100 效果比較好。