HW6 - Conditional Diffusion Model Report

1. 訓練方式

本作業實作一套文字條件式圖像生成模型，依據助教給 sample code（含 CLIP、VAE），加上自定義的 UNet 模型。

受限於設備限制，batch size 僅設 32，並使用 bfloat16 混和精度和 gradient accumulation steps 進行訓練。

模型設定：

- **文字編碼器**：CLIPTextModel (openai/clip-vit-base-patch32)
- **VAE**：AutoencoderKL (CompVis/stable-diffusion-v1-4/vae)
- **UNet**：
  - 架構：4-stage down/up blocks，部分 stage 加入 Cross Attention
  - Cross Attention 維度：512
  - 每層 block 的 channel 為 [256, 384, 512, 768]
  - Attention head dim 設為 16
  - 使用 UNet2DConditionModel 作為 backbone，支援文字條件輸入

訓練設定：

若要接續某個 checkpoint 訓練，可以設定 ckpt_path 路徑，會接續訓練。

- **Img size**：256 × 256（VAE latent 為 32 × 32）
- **Optimizer**：Adam
- **Learning rate**：1e-4
- **Scheduler**：ReduceLROnPlateau（因後期震盪）
- **損失函數**：MSE loss
- **Noise Scheduler**：DDPMScheduler
- **AMP**：使用 torch.amp.autocast 加速訓練（bfloat16）
- **Gradient Accumulation**：16 steps

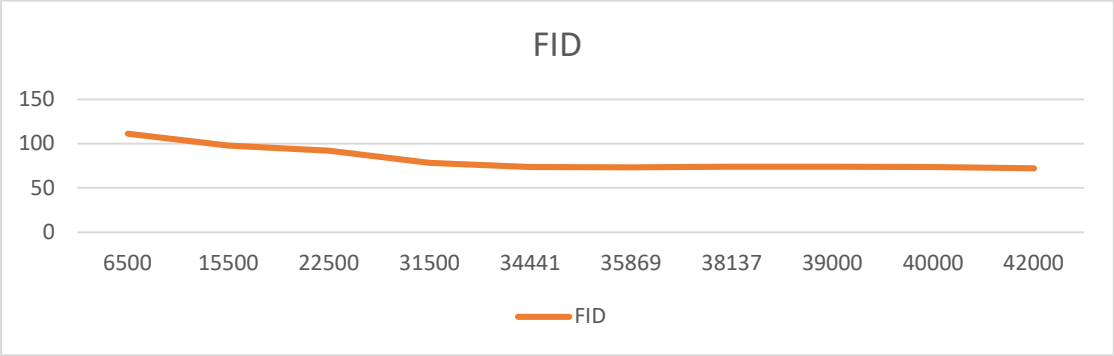
2. 生成方式

測試階段流程：

- **Scheduler**：使用 DDIMScheduler = 50 進行反向推理
- **Latent 初始化**：使用標準 gaussian distribution（1, 4, 32, 32）
- **生成圖片輸出**：將 latent 透過 VAE 還原並正規化至 [0, 1]，儲存為 256*256 的 PNG 檔案

3. 結果分析與觀察

- **表現**：
  - 模型能有效學習文字描述與圖像對應，具有基本語意對齊與風格穩定性，對於簡單描述（如顏色、形狀）學習效果不錯，圖像細節如「武器」、「盔甲」的完成率意外穩定。
  - FID 會隨著訓練時間下降到，大約 step 到 30,000 的時候開始進入震盪（FID 在 8.90 左右），大約 35,000 後進入震盪，FID 不穩定。
- **限制**：
  - 顏色與動作（如：attack）結合時，會看不清楚內容顯得模糊
  - 每個角色收斂的時間不一致，如：單一雪毛怪人收斂很久都沒有穩定表現，但頭上有小企鵝王的雪毛怪人卻比較早成功收斂穩。
  - 較複雜的角色（如：人物有較多裝備、小丑）會需要更多 epoch 才能收斂完成。



4. 額外實驗

- **不使用 scheduler 測試**：
  - 進入震盪後會較快崩壞、overfitting。
- **測試 set_timesteps 差異**：
  - 圖片會變細緻但 FID 差異不大。

- 由於最後一天才有空改助教新增的 noise scheduler ，雖有比沒改前一開始的表現較好，但因時間有限跑不到後面因此效果沒有未改的好。