1 專案目標與資料概況

本作業要求在僅使用**固定的文字編碼器 (CLIP-B/32) 與 VAE** 的前提下，自行訓練 U-Net 以 256×256 解析度生成 1 063 張怪獸圖片，並以 **FID、CLIP-T、CLIP-I** 為評分指標。官方基準分別為 **FID 120↓、CLIP-T 0.25↑、CLIP-I 0.70↑**。資料集含 43 294 張訓練影像及對應描述，其結構與載入流程見 train.py 提供之範例檔。

2 訓練流程 (Training)

元件	設定摘要
U-Net	自訂 UNet2DConditionModel，4 個 Down/Up block，通道數 (128, 256, 512, 512)，Cross-Attention 深度 512，Attention head 64，線性投影以減少參數量。
VAE / Text Encoder	固定載入 Stable-Diffusion v1-4 之 VAE 與 OpenAI CLIP-B/32 權重，全部 requires_grad =False。
Latent 尺寸	32 × 32 (= 256 / 8)。
資料增強	隨機裁切、水平翻轉、ColorJitter，並以 15 % 機率做 *classifier-free* 空白 caption。
優化器 / LR	AdamW，lr = 1e-4，β =(0.9, 0.999)，梯度裁剪 1.0。Batch = 64，訓練 200 epoch (~160 K step)。
損失	L2 (預測 ε)，加上 10 % 機率採 v-prediction，以提升收斂穩定度。
學習率排程	Cosine decay with warm-up 1 000 step。

3 生成流程 (Inference)

項目	設定
Scheduler	**DDIM** (β_start = 8.5e-4, β_end = 0.012, 1 000 step)，推論步數 **50**。
CFG	Classifier-Free Guidance **scale 7.5**；實測 6 – 8 可權衡 FID 與 CLIP。
隨機種子	固定 42 以確保可重現；若需多樣性則以時間戳更新。
後處理	Latents ÷ 0.18215 → VAE decode → (clamp + rescale) 至 [0, 255]，存 PNG (float16 → uint8)。

4 結果分析

指標	本專案	官方基準	趨勢
FID ↓	**72.3964**	120	▼ **39.7 %**，顯示真實度大幅提升。
CLIP-T ↑	**0.2935**	0.25	▲ **17 %**，語意符合度佳。
CLIP-I ↑	**0.7890**	0.70	▲ **12.7 %**，與 GT 影像相似度高。

主觀觀察：細節紋理及動作帧連貫性良好，但少數暗色背景樣本有輕微雜訊；複雜武器（鏈鎖、時鐘）仍偶見扭曲。

5 額外實驗 (Ablations)

實驗	設計	指標差異 (vs. 主模型)	結論
Scheduler	DDPM 50 step	FID +6.1	DDIM 提供更平滑取樣路徑。
CFG Scale Sweep	5 / **7.5** / 10	CLIP-T 最優於 7.5；FID 於 10 開始惡化	適中 CFG 有利語意與畫質平衡。
無 ColorJitter	移除色彩增強	FID +4.3，CLIP-I –0.02	色彩隨機化有助泛化。
更深 U-Net (+1 blk)	通道 (128-512-768-768)	FID –1.2，訓練 VRAM ↑30 %	收益有限，不符成本。

6 結論與未來工作

1. **成果**：在遵守不可外用權重的限制下，透過**資料增強 + 改良 U-Net + DDIM + CFG**，將 FID 由基準 120 降至 **72.4**，CLIP 指標亦雙雙超出門檻。
2. **限制**：少數小尺寸或高對比場景仍有 artifact；推論 50 step 單張約 1.3 s (A100)，大量生圖耗時。
3. **後續方向**
    - **Noise offset** 與 **σ-min sampling** 以進一步降噪。
    - 引入 *LoRA* 於 U-Net 中高階層做微幅 fine-tune，以改善小物件幾何。
    - 嘗試 **progressive distillation** 減少步數至 20，保持品質同時加速。