DLCV Hw-2

Name: 李維釗

Student ID: R12942089

Problem 1: Conditional Diffusion Models

1. Describe your implementation details and the difficulties you encountered.
    - I refer to the github repo [link], and adjust the ContextUnet to fit the different condition method.
    - Because the task contain 2 conditions, the method to tackle the two conditions will be important. I represent the condition using two one-hot encoded vectors: one for Dataset_type (with 2 possible classes) and one for Number (with 10 possible classes). These two vectors are concatenated to form a final conditional embedding of length 12. In this embedding, exactly 2 positions will be set to 1 (one from each vector), while the remaining 10 positions are set to 0.

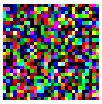
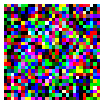
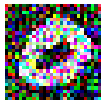
```
MNIST-M acc = 0.9980 (correct/total = 499/500)
SVHN acc = 0.9920 (correct/total = 496/500)
acc = 0.9950
```

2. Please show 10 generated images **for each digit (0-9) from both MNIST-M & SVHN dataset** in your report. You can put all 100 outputs in one image with columns indicating different noise inputs and rows indicating different digits.
    - Mnist-M
    - SVHN

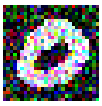
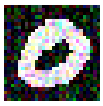
3. Visualize a total of six images from **both MNIST-M & SVHN datasets** in the reverse process of the **first "0"** in your outputs in (2) and with **different time steps**.

- Mnist-M

Step = 0 Step = 200 Step = 400

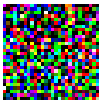
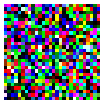
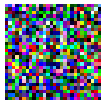




Step = 440 Step = 480 Step = 500

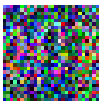
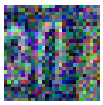




- SVHN

Step = 0 Step = 200 Step = 400

Step = 440 Step = 480 Step = 500

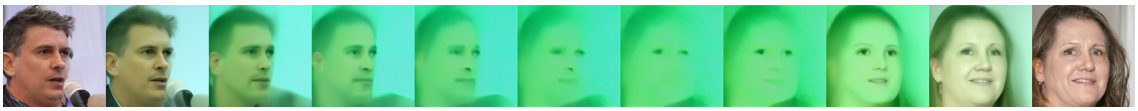




Problem 2: DDIM

1. Please generate face images of noise **00.pt ~ 03.pt with different eta** in one grid. Report and explain your observation in this experiment.

- 當eta值=0時，denoise 的過程為 DDIM，是完全確定的過程。在支持精確的逆過程、適合圖像編輯和插值、生成結果更穩定這幾項優點外，他也代表生成結果會缺乏多樣性、趨於生成遵循某些規則的圖像。相反的當eta值=1時，denoise 過程為 DDPM，是完全隨機的。使得生成結果更加多樣、細節更加自然隨機。相對的，DDPM並不支持精確的逆過程。因此在這張圖裡面可以發現當eta值較小的情況下圖像的輪廓、性別其實差異不大。當eta大於0.5後生成的結果就比較隨機，在eta=0.75, 1時的生成結果都相對的較自然一些。

1. Please generate the face images of the interpolation of noise **00.pt ~ 01.pt.**
    a. Linear



b. Slerp

- 由於 Diffusion model 的 noise 空間是一個高維空間，有效的 noise 向量實際上分布在一個高維球面上（因為通常會進行 normalization）這個空間中的每個點都對應到一個可能的圖像。SLERP 在兩點之間進行插值時，會沿著球面的大圓路徑進行。這保證了所有插值點都位於單位球面上，都可以有不錯的生成結果。當在兩點之間進行線性插值時，插值路徑是一條直線。在高維空間中，直線路徑會穿過球體內部，導致以線性差值的 noise 進行 denoise 後的生成結果會有模糊、扭曲或不自然的特徵。

Problem 3: Personalization

1. Conduct the CLIP-based zero shot classification on the hw2_data/clip_zeroshot/val, explain how CLIP do this, report the accuracy and 5 successful/failed cases.

   - CLIP model利用contrastive learning的方式來訓練兩個encoder，分別處理文字與圖像。在訓練過程中會不斷更新 encoders，逐漸使對應 text-image pair 的 embeddings 越接近。因此在 inference 時，將文字與圖像同時輸入後，model 會算出兩個對應的 embeddings，而後再根據cosine similarity來計算兩個 embeddings 之間的相似程度，距離最近的分類機率越高。
   - Classification report (accuracy = 56%)

```
              precision    recall  f1-score   support

           0       0.98      0.90      0.94        50
           1       0.69      0.98      0.81        50
           2       0.81      0.76      0.78        50
           3       0.87      0.52      0.65        50
           4       0.72      0.68      0.70        50
```

5	0.91	0.80	0.85	50
6	1.00	0.10	0.18	50
7	1.00	0.48	0.65	50
8	0.61	0.34	0.44	50
9	0.95	0.76	0.84	50
10	0.89	0.34	0.49	50
11	0.75	0.54	0.63	50
12	0.97	0.76	0.85	50
13	0.05	0.62	0.09	50
14	0.82	0.66	0.73	50
15	0.49	0.48	0.48	50
16	0.88	0.42	0.57	50
17	0.59	0.60	0.59	50
18	0.34	0.76	0.47	50
19	0.54	0.40	0.46	50
20	0.81	0.60	0.69	50
21	0.87	0.68	0.76	50
22	0.84	0.32	0.46	50
23	0.35	0.76	0.48	50
24	0.95	0.38	0.54	50
25	0.92	0.46	0.61	50
26	0.96	0.52	0.68	50
27	0.84	0.62	0.71	50
28	0.81	1.00	0.89	50
29	1.00	0.10	0.18	50
30	0.76	0.62	0.68	50
31	0.91	0.40	0.56	50
32	0.68	0.88	0.77	50
33	0.57	0.16	0.25	50
34	0.96	0.44	0.60	50
35	0.60	0.18	0.28	50
36	0.64	0.32	0.43	50
37	0.57	0.54	0.56	50
38	0.90	0.72	0.80	50
39	0.64	0.96	0.77	50
40	0.92	0.66	0.77	50

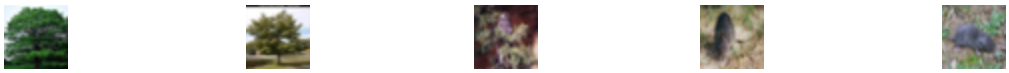
```
              41       1.00      0.06      0.11        50
              42       0.95      0.70      0.80        50
              43       0.97      0.60      0.74        50
              44       0.84      0.76      0.80        50
              45       0.98      0.80      0.88        50
              46       0.95      0.42      0.58        50
              47       0.75      0.06      0.11        50
              48       0.39      0.64      0.48        50
              49       0.84      0.98      0.91        50

        accuracy                           0.56      2500
       macro avg       0.78      0.56      0.60      2500
    weighted avg       0.78      0.56      0.60      2500
```

- 以下呈現幾個正確與錯誤分類的結果。
  - Successful cases

  - Failed cases (true ⇒ pred)

13 ⇒ 18 18 ⇒ 23 23 ⇒ 13 33 ⇒ 8 48 ⇒ 30

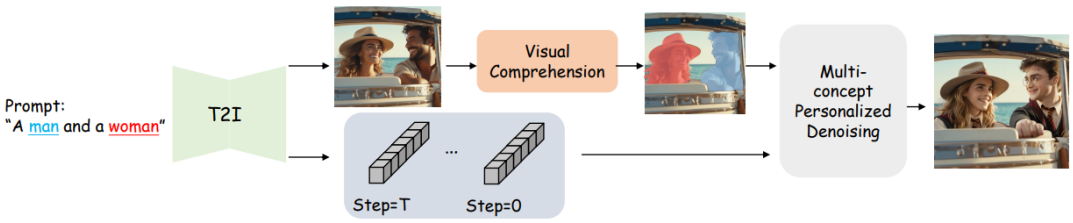


2. What will happen if you simply generate an image containing multiple concepts (e.g., a <new1> next to a <new2>)? You can use your own objects or the provided cat images in the dataset. Share your findings and survey a related paper that works on multiple concepts personalization, and share their method.
   - multiple concepts with prompt: a `<new1> in a city in a style of <new2>`

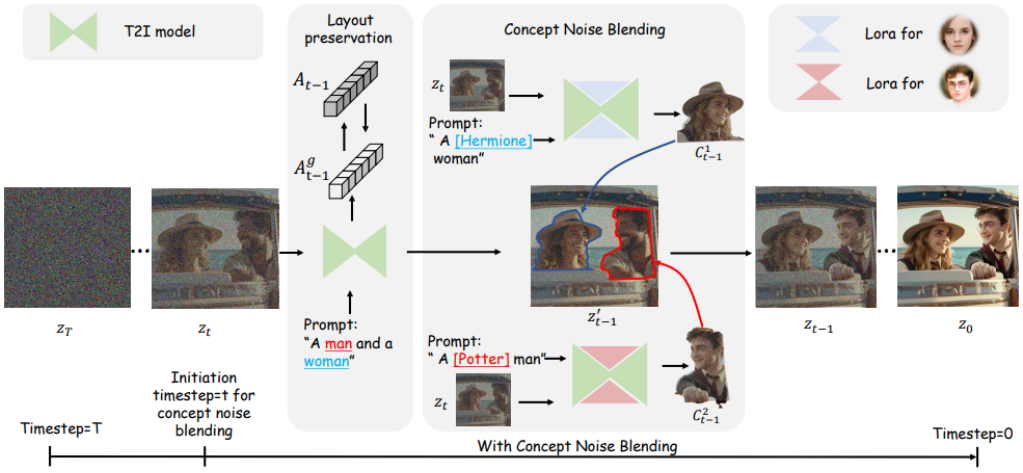
     Good sample 前後景畫面不和諧 <new1> 保留不完全



- OMG: Occlusion-friendly Personalized Multi-concept Generation
  - 這篇論文點出目前 multiple concepts personalization 的問題在於目標物的保留、遮蔽以及前景與背景之間不和諧問題。並提出了一個兩階段的生成方式來解決。



OMG model framework



Multi-concept Personalized Denoising

1. 根據上述提到的問題，第一階段利用 T2I model來生成圖像的同時，也學習圖像的 layout 以及視覺理解訊息（target identities 的圖像位置資訊）來使後續降噪過程可以使用這些圖像位置資訊。

2. 透過前階段學到的資訊，來進行降噪。在降噪方法的部分，本篇提出了 latent level 和 attention-level noise 的混和，並且在不同的 identities 使用不同的 lora layer 來進行訓練，以保留最大程度的 identities 資訊，並且可以在對不同目標進行轉換的同時，學習到遮擋的部分，也保留了前、後景之間的不協調。