DLCV_HW2_report

B10901098 電機四 蔡承恩

Problem 1

1. (5%) Describe your implementation details and the difficulties you encountered.

Implementation Summar

1. **Model Architecture (`p1_model.py`):**
   - `ConvBlock`, `U_encoder`, **and** `U_decoder`**:** Basic building blocks for the U-Net with residuals, pooling, and transposed convolution for encoding and decoding.
   - `EmbedFC`**:** Embeds conditional labels and time steps for diffusion conditioning.
   - `Unet`**:** The main model, embedding class and dataset information, and processing it through a U-Net architecture with skip connections.
   - `DDPM_framework`**:** Implements the forward and reverse diffusion process, managing noise addition/removal with conditional embeddings, dropout, and scheduling.
2. **Training Script (`p1_train.py`):**
   - **Data Loading:** Combines MNIST-M and SVHN datasets, each with their digit labels and dataset indicators.
   - **Training Loop:** Trains with Adam and Cosine Annealing LR scheduler, with mixed-precision training for efficiency and checkpointing.
3. **Inference Script (`p1_inference.py`):**
   - **Image Generation:** Loads a model checkpoint and generates images conditioned on each digit and dataset combination.

Difficulties Encountered

1. **Conditional Label Integration:** Embedding and combining digit classes and dataset labels accurately, handling mask expansions to control label influence.
2. **Data Alignment:** Ensuring consistent preprocessing and handling two distinct datasets with different label formats.

3. **Inference Precision:** Ensuring generated images reflect specified digit classes and dataset labels accurately.

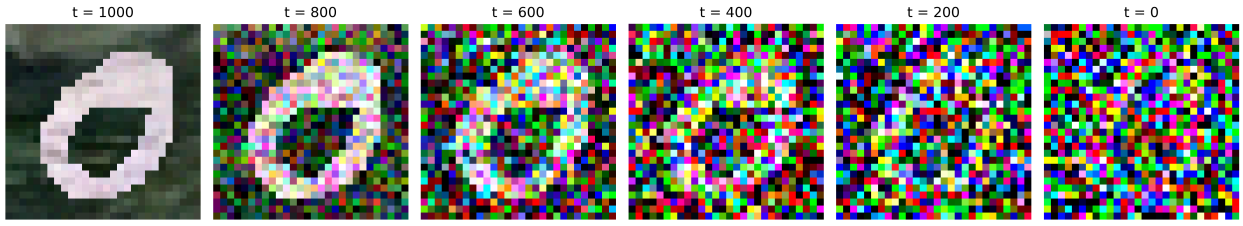
2. (5%) Please show 10 generated images for each digit (0-9) from both MNIST-M & SVHN dataset in your report. You can put all 100 outputs in one image with columns indicating different noise inputs and rows indicating different digits. [see the below MNIST-M example, you should visualize BOTH MNIST-M & SVHN]

MNIST-M

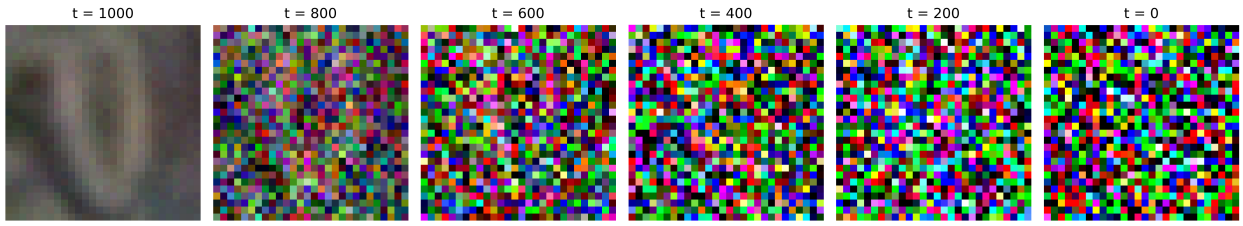
SVHN

3. (5%) 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. [see the MNIST-M example below, but you need to visualize BOTH MNIST-M & SVHN]

MNIST-M



SVHN



Problem 2

1. (7.5%) Please generate face images of noise 00.pt ~ 03.pt with different eta in one grid. Report and explain your observation in this experiment. (This following image is just for illustration.

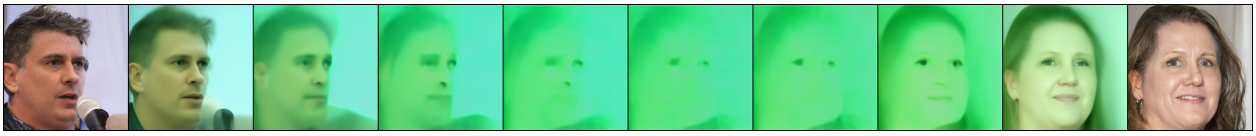
2. (7.5%) Please generate the face images of the interpolation of noise 00.pt ~ 01.pt. The interpolation formula is spherical linear interpolation , which is also known as slerp. What will happen if we simply use linear interpolation? Explain and report your observation. (There should be two images in your report, one for spherical linear and the other for linear)

Spherical Linear Interpolation

The transition is significantly smoother and more natural-looking. The intermediate faces maintain better quality and sharpness throughout the sequence. There are no unnatural color artifacts, and the facial features transform more coherently from one identity to another.

Linear Interpolation

The transition appears less natural with a noticeable green tint in the middle frames. The intermediate faces show blurry and distorted features, particularly in the middle of the sequence where the transformation is most pronounced.



Comparison

Spherical linear interpolation (SLERP) produces superior results because it:

- Maintains consistent quality across all interpolation steps
- Preserves facial features better during transition
- Creates a more natural and smooth progression between source and target images
- Avoids the unnatural artifacts and color distortions seen in linear interpolation

Problem 3

1. (7.5%) 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 (Contrastive Language-Image Pre-training) is a powerful model developed by OpenAI that bridges the gap between visual and textual understanding. It learns to associate images with their corresponding textual descriptions by training on a vast dataset of image-caption pairs. CLIP consists of two main components:

1. **Image Encoder:** Processes images and maps them into a high-dimensional embedding space.
2. **Text Encoder:** Processes text prompts and maps them into the same embedding space.

Zero-shot classification refers to the ability of a model to correctly classify instances from classes it hasn't explicitly been trained on. Here's how CLIP achieves this:

1. **Embedding Generation:**
   - **Text Prompts:** For each class label, a textual description (prompt) is created, such as "A photo of a dog."
   - **Images:** Each image in the dataset is processed by the image encoder to obtain its embedding.
2. **Similarity Computation:**
   - The text prompts are also processed by the text encoder to obtain their embeddings.
   - For each image, the cosine similarity between its embedding and each text prompt embedding is calculated.
3. **Prediction:**
   - The class label corresponding to the text prompt with the highest similarity score is assigned to the image.

Zero-Shot Classification Accuracy: 71.36% (1784/2500)

5 Successful Cases:





1. Image: 27_469.png | Ground Truth: bee | Prediction: bee
2. Image: 34_477.png | Ground Truth: castle | Prediction: castle
3. Image: 14_475.png | Ground Truth: apple | Prediction: apple
4. Image: 5_459.png | Ground Truth: clock | Prediction: clock
5. Image: 38_462.png | Ground Truth: mountain | Prediction: mountain

---

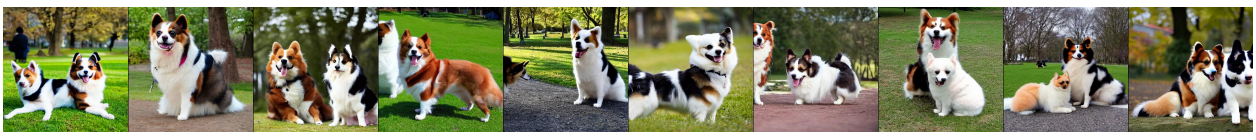
5 Failed Cases:

1. Image: 23_470.png | Ground Truth: pine_tree | Prediction: palm_tree
2. Image: 0_494.png | Ground Truth: bicycle | Prediction: oak_tree
3. Image: 24_453.png | Ground Truth: house | Prediction: pine_tree
4. Image: 47_466.png | Ground Truth: otter | Prediction: elephant
5. Image: 47_456.png | Ground Truth: otter | Prediction: willow_tree

---

2. (7.5%) 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.

"A <newdog> next to a <newcat> in a park."



"A portrait of a <newdog> and a <newcat> in the style of <newartist>."

"A <newcat> sitting on top of a <newdog>."

Observations from Multi-Concept Generation

Composition Challenges

- The generated images show inconsistent quality when combining multiple personalized concepts (<newdog> and <newcat>)
- While individual features are preserved, the spatial relationships and interactions between concepts appear less natural
- The style consistency (<newartist>) seems to affect both objects simultaneously but may dilute the distinct characteristics of each concept
- The model struggles with precise positioning and relative scaling between the personalized concept

Related Research

A significant paper addressing these multi-concept challenges is "Custom Diffusion: Multi-Concept Customization of Text-to-Image Diffusion".

Key Methodology

- Introduces concept-specific fine-tuning of the cross-attention layers
- Implements a novel "attention pooling" mechanism to better handle multiple concepts
- Uses concept-specific learning rates based on concept similarity to prevent interference

Technical Innovations

- Employs gradient-based concept separation to maintain distinct features
- Introduces a cross-concept attention mask to control concept bleeding
- Implements adaptive prompt weighting to balance multiple concept representations

Advantages Over Basic Textual Inversion

- Better preservation of individual concept characteristics
- More natural composition of multiple concepts

- Improved spatial relationship handling
- Reduced training time per concept
- Better scalability with multiple concepts

The Custom Diffusion approach specifically addresses the limitations we observe in basic Textual Inversion when combining multiple concepts, offering a more robust solution for multi-concept personalization.