DLCV-HW2

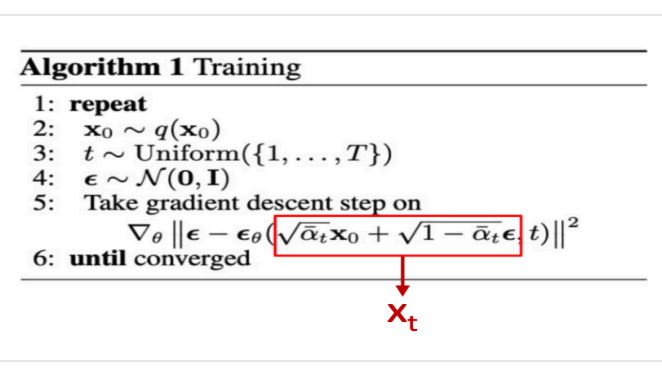
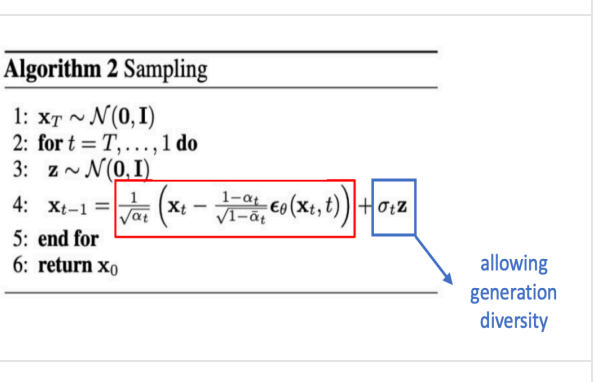
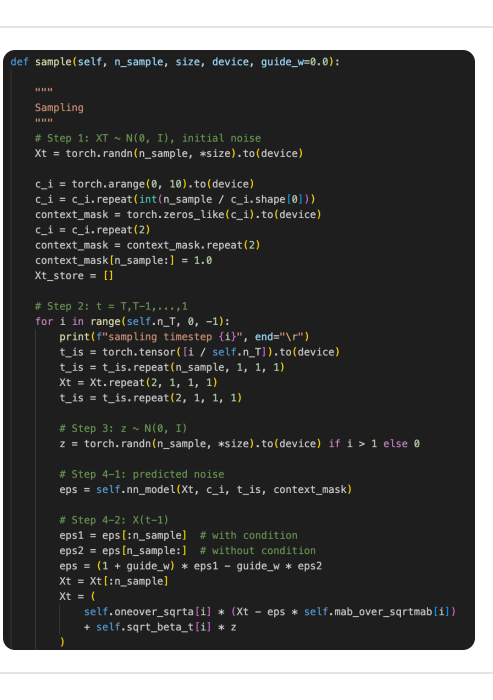
- Student ID: B09611007

Problem 1: Diffusion Models

Sample random noise from normal distribution to generate 50 conditional images for each digit (0-9) on MNIST-M & SVHN datasets. Your script should save total 1000 outputs in the assigned folder for further evaluation.

1. Describe your implementation details and the difficulties you encountered.

- DDPM (https://arxiv.org/abs/2006.11239) (Reference: Classifier-Free Diffusion Guidance (https://arxiv.org/abs/2207.12598), Conditional_Diffusion_MNIST (https://github.com/TeaPearce/Conditional_Diffusion_MNIST))

Training	Sampling
	
	

- A modified UNet that supports conditional learning by adding **time embedding** (time steps in the diffusion process) and **context embedding** (labels)

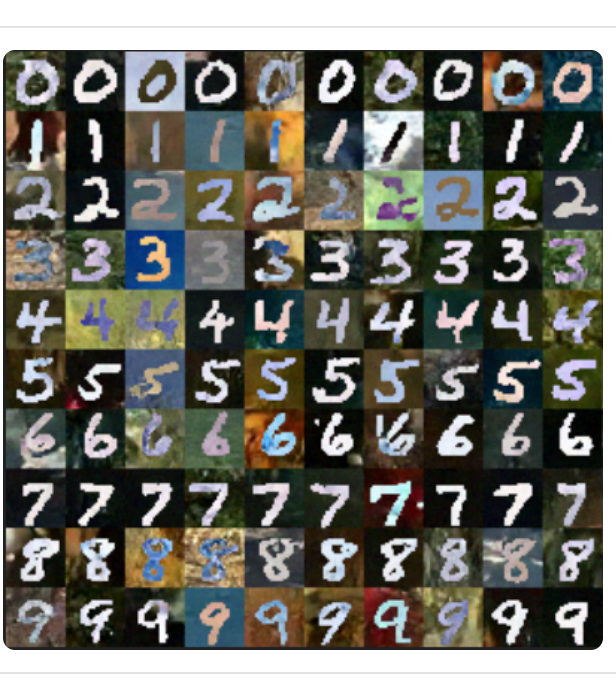
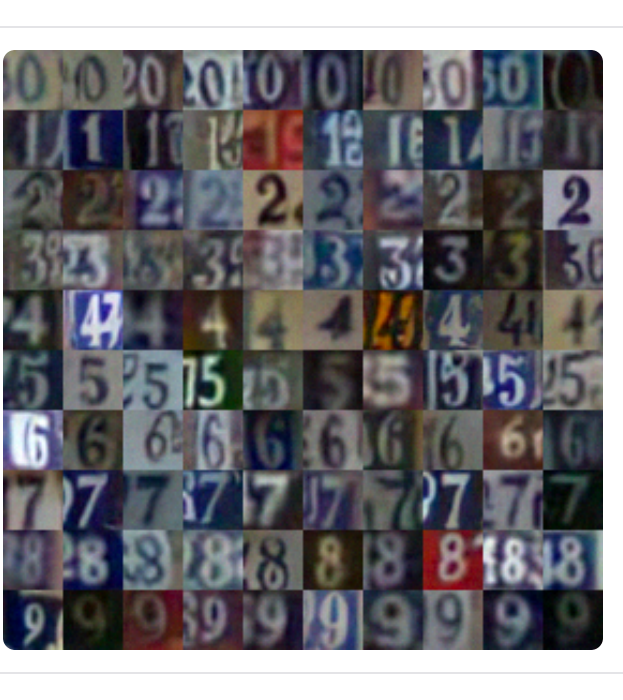
- Training Parameters
  - Epochs: 100, batch size: 256, learning rate: 2e-4
  - Total steps for diffusion model (T): 500
  - Context dropout rate: 0.3
  - Loss: MSE (between actual eps & predicted eps)

- Evaluation (by digit classifiers)

	MNIST-M	SVHN	Average
Acc.	99.80%	99.20%	99.50%

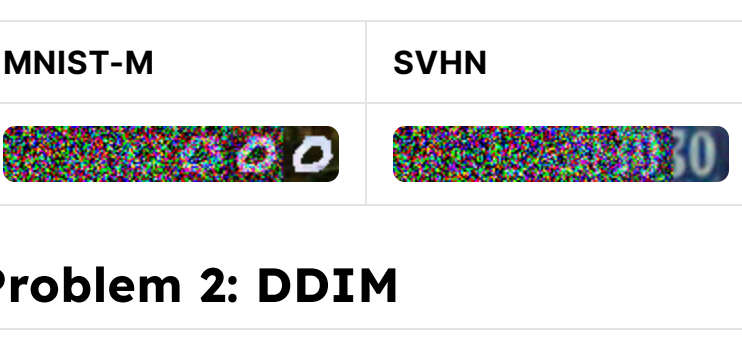
- Challenges
  - Hard to debug
  - Context dropout sometimes affects image quality

2. Please show 10 generated images for each digit (0-9) from both MNIST-M & SVHN dataset in your report.

MNIST-M	SVHN
	

3. Visualize a total of six images from both MNIST-M & SVHN datasets in the reverse process of the first "0" with different time steps.

MNIST-M	SVHN



Problem 2: DDIM

In this problem, you need to implement the DDIM algorithm to generate face images. In addition, you will need to further analyze the properties of DDIM.

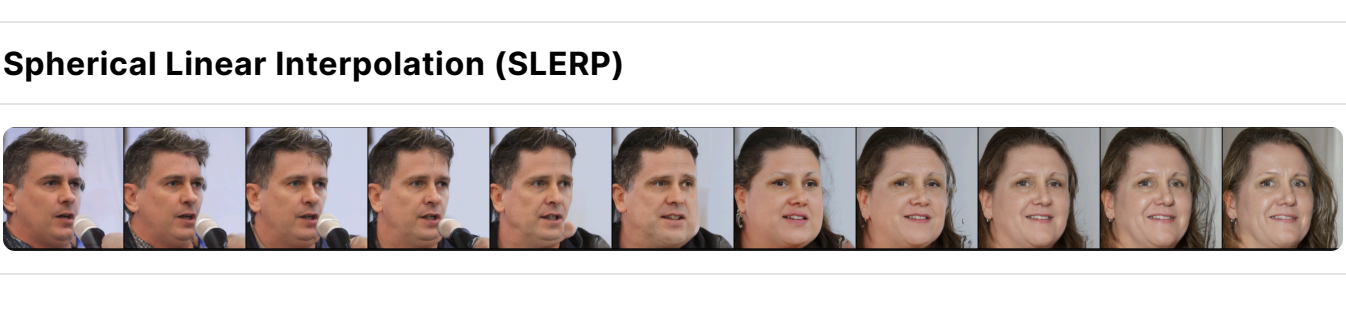
1. Please generate face images of noise 00.pt (http://00.pt) ~ 03.pt (http://03.pt) with different eta in one grid. Report and explain your observation in this experiment. (This following image is just for illustration.)

eta = 0.0	eta = 0.25	eta = 0.5	eta = 0.75	eta = 1.0

- Avg. MSE: 1.56 (between generated images and ground truths)

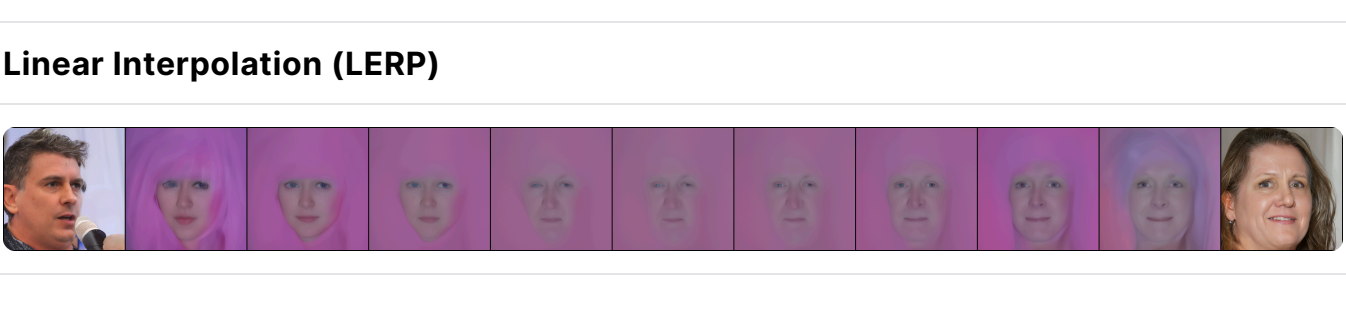
2. Please generate the face images of the interpolation of noise 00.pt (http://00.pt) ~ 01.pt (http://01.pt).
The interpolation formula is spherical linear interpolation, which is also known as slerp. What will happen if we simply use linear interpolation? Explain and report your observation. (There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

Spherical Linear Interpolation (SLERP)

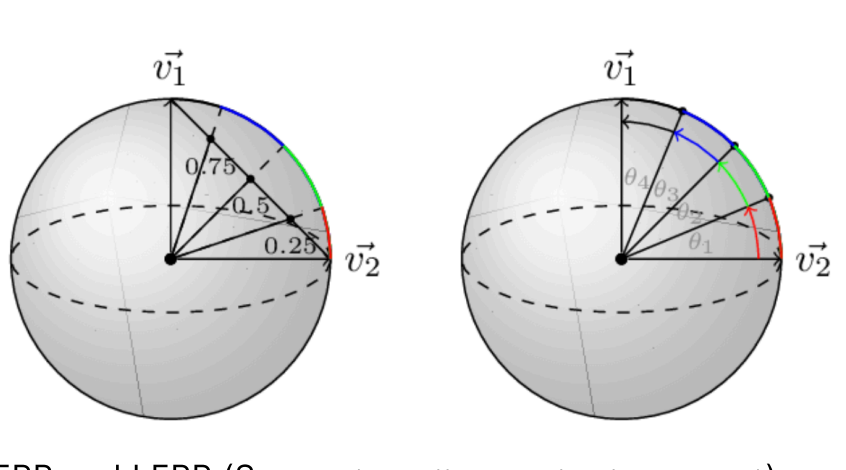


- SLERP ensures that the interpolation occurs **along the surface of a hypersphere** (the space in which the noise vectors exist).
- This smooth interpolation happens because the **angle** between the noise vectors is maintained, preserving the spatial relationships in the latent space of the diffusion model.

Linear Interpolation (LERP)



- Linear interpolation (LERP) simply linearly mixes the two noise vectors without considering their position on a hypersphere.
- This can result in images that look less coherent because the interpolated points might not lie on the surface of the latent space's natural geometry.



SLERP and LERP (Source (https://arxiv.org/abs/2003.14297))

Problem 3: Personalization

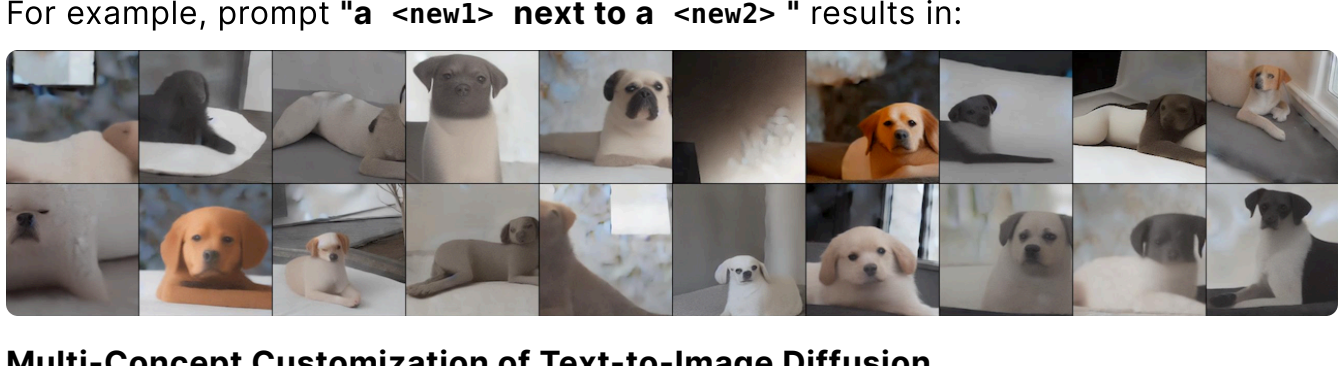
1. Conduct the CLIP-based zero shot classification on the hw2_data/clip_zeroshot/val, explain how CLIP do this, report the accuracy and 5 successful/failed cases.

Success #1	Success #2	Success #3	Success #4	Success #5
"chair"	"camel"	"wardrobe"	"bee"	"cup"

Failed #1	Failed #2	Failed #3	Failed #4	Failed #5
"skunk"	"beetle"	"beetle"	"girl"	"clock"

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.

- The model might selectively present only one of them.
- A fusion of features may lead to unexpected shapes or distorted details.
- For example, prompt **"a `<new1>` next to a `<new2>` "** results in:



- **Multi-Concept Customization of Text-to-Image Diffusion** (https://openaccess.thecvf.com/content/CVPR2023/papers/Kumari_Multi-Concept_Customization_of_Text-to-Image_Diffusion_CVPR_2023_paper.pdf)
  - **Joint training on multiple concepts:** combine the training datasets for each individual concept and train them jointly.
  - Optimize them along with **cross-attention key and value** matrices for each layer.
  - Restricting the weight update to cross-attention key and value parameters leads to better results for composing two concepts compared to methods like DreamBooth, which fine-tunes all the weights.