

HOMWORK 4

MULTI-MODAL FOUNDATION MODELS *

10-423/623/723 GENERATIVE AI
<http://423.mlcourse.org>

OUT: October 23, 2025
DUE: November 3, 2025
TAs: Maxwell, Cassandra

Instructions

- **Collaboration Policy:** Please read the collaboration policy in the syllabus.
- **Late Submission Policy:** See the late submission policy in the syllabus.
- **Submitting your work:** You will use Gradescope to submit answers to all questions and code.
 - **Written:** You will submit your completed homework as a PDF to Gradescope. Please use the provided template. Submissions can be handwritten, but must be clearly legible; otherwise, you will not be awarded marks. Alternatively, submissions can be written in $\text{\LaTeX}$. Each answer should be within the box provided. If you do not follow the template, your assignment may not be graded correctly by our AI assisted grader and there will be a **2% penalty** (e.g., if the homework is out of 100 points, 2 points will be deducted from your final score).
 - **Programming:** You will submit your code for programming questions to Gradescope. We will examine your code by hand and may award marks for its submission.
- **Materials:** The data that you will need in order to complete this assignment is posted along with the writeup and template on the course website.

Question	Points
$\text{\LaTeX}$ Template Alignment	0
Latent Diffusion Model (LDM)	7
VQ-VAEs	8
CLIP	4
VLMs	8
Programming: Text-to-Image Generation	26
Code Upload	0
Collaboration Questions	2
Total:	55

*Compiled on Friday 24th October, 2025 at 22:39

1 $\text{\LaTeX}$ Template Alignment (0 points)

1.1. (0 points) **Select one:** Did you use $\text{\LaTeX}$ for the entire written portion of this homework?

☐ Yes

☐ No

1.2. (0 points) **Select one:** I have ensured that my final submission is aligned with the original template given to me in the handout file and that I haven't deleted or resized any items or made any other modifications which will result in a misaligned template. I understand that incorrectly responding yes to this question will result in a penalty equivalent to 2% of the points on this assignment.

Note: Failing to answer this question will not exempt you from the 2% misalignment penalty.

☐ Yes

2 Latent Diffusion Model (LDM) (7 points)

- 2.1. (1 point) **Math:** A latent diffusion model uses a compression factor of $f = 8$ to downsample. If an input image is 512×512 pixels, what are the dimensions of the latent representation of this image?

- 2.2. (2 points) **Short answer:** Why does a latent diffusion model run diffusion in a latent space instead of pixel space?

- 2.3. **Short answer:** Standard cross-attention for a diffusion-based text-to-image model defines the queries $\mathbf{Q}$ as a function of the pixels (or latent space) $\mathbf{Y} \in \mathbb{R}^{m \times d_y}$, and the keys $\mathbf{K}$ and values $\mathbf{V}$ as a function of the text encoder output $\mathbf{X} \in \mathbb{R}^{n \times d_x}$.

$$\mathbf{Q} = \mathbf{Y}\mathbf{W}_q, \quad \mathbf{K} = \mathbf{X}\mathbf{W}_k, \quad \mathbf{V} = \mathbf{X}\mathbf{W}_v$$

(where $\mathbf{W}_q \in \mathbb{R}^{d_y \times d}$ and $\mathbf{W}_k, \mathbf{W}_v \in \mathbb{R}^{d_x \times d}$) and then applies standard attention:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\mathbf{Q}\mathbf{K}^T / \sqrt{d})\mathbf{V}$$

Now, suppose you instead defined a new formulation where the values are a function of the pixels (or latent space): $\mathbf{V} = \mathbf{Y}\mathbf{W}_v$ where $\mathbf{W}_v \in \mathbb{R}^{d_y \times d}$.

- 2.3.a. (2 points) What goes wrong mathematically in the new formulation?

- 2.3.b. (2 points) Intuitively, why doesn't the new formulation make sense? Briefly begin with an explanation of what the original formulation of cross-attention is trying to accomplish for a single query vector, and why this new formulation fails to accomplish that.

3 VQ-VAEs (8 points)

3.1. The objective function for a VQ-VAE contains two terms in addition to the reconstruction loss term:

- The vector-quantization loss, $\|\text{sg}[z_e(x)] - e\|_2^2$ and
- The “commitment” loss, $\beta\|z_e(x) - \text{sg}[e]\|_2^2$

where sg is the stopgradient operator and e is the latent embedding vector closest to the output of the encoder $z_e(x)$.¹

In this question, you will examine the impact of these terms and the stopgradient operator. Let

$$\mathcal{L} = \|\text{sg}[z_e(x)] - e\|_2^2 + \beta\|z_e(x) - \text{sg}[e]\|_2^2 \quad (1)$$

$$\tilde{\mathcal{L}} = \|z_e(x) - e\|_2^2 + \beta\|z_e(x) - e\|_2^2 \quad (2)$$

3.1.a. (1 point) **Math:** What is the gradient of $\mathcal{L}$ with respect to $z_e(x)$?

3.1.b. (1 point) **Math:** What is the gradient of $\mathcal{L}$ with respect to e ?

3.1.c. (1 point) **Math:** What is the gradient of $\tilde{\mathcal{L}}$ with respect to $z_e(x)$?

3.1.d. (1 point) **Math:** What is the gradient of $\tilde{\mathcal{L}}$ with respect to e ?

3.1.e. (2 points) **Short answer:** Given your findings from the previous parts, how would you describe the impact of the stopgradient operator on the optimization of these two terms? How do the gradients with and without the stopgradient operator differ?

¹For the purposes of this question, we are still assuming that the encoder outputs a single vector; in practice, a true VQ-VAE encoder would output multiple vectors and these terms in the objective function would be sums over these vectors.

- 3.2. (2 points) In a VQ-VAE, the codebook embeddings e need to be learned during training. Recall that the VQ-VAE objective includes the vector-quantization loss $\|\text{sg}[z_e(x)] - e\|_2^2$, where $z_e(x)$ is the encoder output, e is the nearest codebook embedding, and $\text{sg}[\cdot]$ is the stop gradient operator. Using your answer from 3.1, describe in words what the codebook embeddings are being optimized to do.

4 CLIP (4 points)

- 4.1. (4 points) **Short answer:** In [Radford et al. \(2021\)](#), the original CLIP paper, they pre-trained the image and text encoders using a *symmetric cross-entropy loss*. Formally, given a mini-batch of N (image, caption) pairs, let $e_i^{(n)}$ be the image embedding of the n^{th} image normalized to have unit norm and let $e_t^{(n)}$ be the text embedding of the n^{th} caption, also normalized to have unit norm.

The symmetric cross-entropy loss consists of two terms for each (image, caption) pair, a term that compares the image embedding to all N caption embeddings and a term that compares the caption embedding to all N image embeddings. Formally, let

$$\ell_{i \rightarrow t}^{(n)} = -\log \left(\frac{\exp(e_i^{(n)} \cdot e_t^{(n)})}{\sum_{m=1}^N \exp(e_i^{(n)} \cdot e_t^{(m)})} \right) \quad (3)$$

$$\ell_{t \rightarrow i}^{(n)} = -\log \left(\frac{\exp(e_i^{(n)} \cdot e_t^{(n)})}{\sum_{m=1}^N \exp(e_i^{(m)} \cdot e_t^{(n)})} \right) \quad (4)$$

where $\cdot$ is the dot-product operation between two (same-length) vectors. The CLIP objective can then be expressed as

$$\ell = \frac{1}{N} \sum_{n=1}^N \frac{\ell_{i \rightarrow t}^{(n)} + \ell_{t \rightarrow i}^{(n)}}{2} \quad (5)$$

Show that minimizing this objective function is equivalent to jointly maximizing the cosine similarity of an image embedding and its corresponding caption embedding while minimizing the cosine similarity of all other pairs of image and caption embeddings.

Hint: It may be helpful to start with $\min(\ell)$, and follow through with the mathematical logic until you show that it evaluates to $\max_n \{e_i^{(n)} \cdot e_t^{(n)}\}$ and $\min_{n \neq m} \{e_i^{(n)} \cdot e_t^{(m)}, e_i^{(m)} \cdot e_t^{(n)}\}$.

5 VLMs (8 points)

This section will familiarize you with the working of VLMs. For simplicity, we will be focusing on the PaliGemma2 Model. We encourage you to read this [Research Paper](#) and this [Article](#) to better understand the details of the model.

- 5.1. (1 point) **Select one:** Which component converts images into visual tokens in PaliGemma 2?
- ☐ Gemma 2 language model
 - ☐ SigLIP vision encoder
 - ☐ Linear projection layer
 - ☐ Autoregressive decoder
- 5.2. (1 point) **Select one:** What is the primary purpose of the projection layer in multimodal models like PaliGemma2?
- ☐ To downsample the image size for faster processing
 - ☐ To align vision encoder outputs with language model's embedding space
 - ☐ To generate more high-fidelity images from text descriptions
 - ☐ To filter out irrelevant visual information before training attention layers
- 5.3. (1 point) **Select one:** Why does PaliGemma2 *primarily* use a pre-trained vision encoder like SigLIP instead of training a vision component from scratch?
- ☐ To save on computational resources only
 - ☐ Because training vision models is impossible within this framework
 - ☐ To leverage the visual understanding capabilities present in pre-trained vision models
 - ☐ Because processing raw images directly with neural networks leads to worse output
- 5.4. (1 point) **Select one:** What advantage does using a pre-trained language model like Gemma provide in the PaliGemma2 architecture?
- ☐ It eliminates the need for any image processing components
 - ☐ It brings language understanding capabilities to visual reasoning tasks
 - ☐ It makes the system completely language-independent
 - ☐ It automatically generates training data for the vision encoder

- 5.5. (2 points) **Short Answer:** With an image resolution of 224×224 pixels and a patch size of 14×14 pixels, Pali-Gemma 2 generates 256 visual tokens. How would the number of tokens change if we increased the resolution to 448×448 pixels while maintaining the same patch size of 14×14 pixels? Why is this change important for the model's performance?

- 5.6. (2 points) **Short Answer:** VLMs are great for tasks that they have been trained on, but still can be tricked by adversarial examples. For this question, first take a normal image and ask a proprietary VLM like GPT or Gemini to densely caption it (i.e., describe what's happening in the image in detail), and paste the dense caption. Next, find/create an adversarial image and question pair that a VLM gets completely wrong. Attach a screenshot of this exchange, like the example below:



how many legs does the animal in this image have



The animal in the image is a deer, and like all deer, it has **four legs**. In the image, all four legs are visible—two in the foreground and two partially obscured in the background.

Dense Caption

Screenshot

6 Programming: Text-to-Image Generation (26 points)

Introduction

Text-to-image generation has revolutionized how we create visual content, with models like DALL-E, Stable Diffusion, and Midjourney demonstrating remarkable capabilities. These large-scale models require extensive computational resources and massive datasets. Such models have moved us closer to the objective of being able to combine the knowledge of an LLM with the visual quality of a image generation model for a chat-like generation experience. In this assignment, you'll explore the fundamental principles of LLM-conditioned image generation by combining a pretrained LLM and class-conditional diffusion model into a single text-to-image generation model.

The core challenge in text-to-image generation is bridging the semantic gap between language and vision. While diffusion models excel at generating high-quality images and language models capture rich semantic information, connecting these modalities efficiently remains non-trivial. In this assignment, you will train a **Q-Former** (Query Former) architecture proposed by [Perciever IO](#), and used to connect text and image features in [BLIP-2 \(Li et al., 2023\)](#) and [MetaQueries](#). The Q-Former uses a fixed number of learnable query tokens to extract relevant visual features from arbitrary length language model representations.

Your will:

- Train a Q-Former to use language model text embeddings from GPT-2 as conditioning for a pre-trained diffusion model with a Diffusion Transformer (DiT) backbone, while keeping both GPT-2 and the DiT model weights frozen.
- Generate 32×32 images from simple text prompts like "fluffy cat" or "red airplane".
- Examine cross-attention maps to determine what information the Q-Former is extracting from the pretrained LLM.

Note: CIFAR-10 images are only 32×32 pixels, so generated images may appear blurry or lack fine details. This is expected due to the low resolution, not a sign of poor model performance. We highly recommend that you take a look at CIFAR-10 for reference.

Background

The Text-to-Image Pipeline:

Our goal is to generate images from text descriptions. We have three components:

1. **GPT-2 (Frozen):** A pretrained language model that converts text into semantic feature vectors
2. **DiT Diffusion Model (Frozen):** A pretrained class-conditional diffusion transformer trained on the [CIFAR dataset](#)
3. **Q-Former (Trained):** A lightweight adapter that translates language features into visual conditioning, allowing us to provide precise captions for cifar-like images, and extrapolate beyond class-based generation

Image Denoising with Our Qformer Implementation

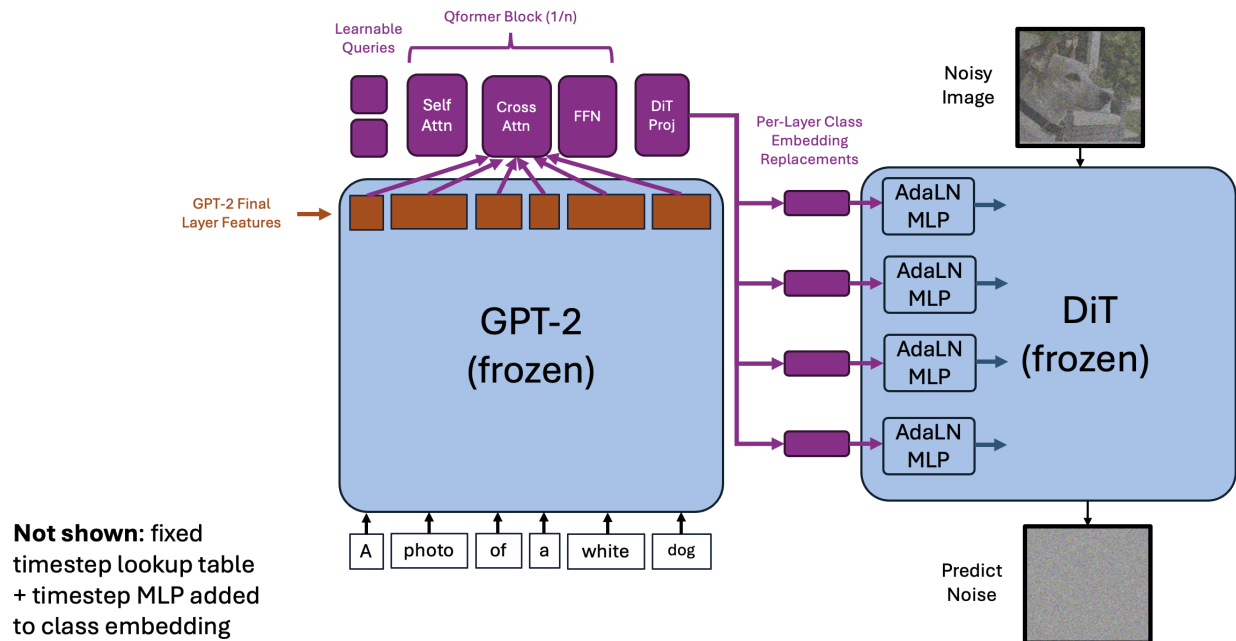


Figure 1: Our Q-Former architecture for text-to-image generation.

First, descriptive text is tokenized and passed through GPT-2 to produce **per-token text features** (bottom-left, white boxes up to orange layer). Next, we initialize **learnable query embeddings**, and transform them with transformer blocks that perform self-attention, cross-attention with the **per-token text features**, and a per-token feed-forward network (FFN) (top-left, orange layer to purple blocks). Then, a DiT projector outputs a set of **class embedding features** (middle, purple block). During the forward pass of the Diffusion Transformer (DiT), the class embedding vector for the current image class is replaced with the **per-layer class embedding features** at every layer (from purple middle and top-right noisy image to bottom-right predicted noise).

During training, only the purple layers are trained, while all other weights are frozen.

Why Q-Former?

Why don't we use GPT-2 features directly as conditioning for the diffusion model? Here are some challenges:

- GPT-2 outputs a variable length sequence of 768-dimensional vectors, but DiT expects a single 768-dimensional vector for conditioning.
- GPT-2 features live in “language space” while DiT needs “visual conditioning space,” and we would not expect these representations to be aligned in any sense.

The Q-Former solves this by learning to extract relevant information from language features and projecting it into the format DiT expects.

Diffusion Models for Image Generation:

Denoising Diffusion Probabilistic Models (DDPMs) (Ho et al., 2020) generate images by iteratively denoising random Gaussian noise. The model learns to predict and remove noise step-by-step:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(x_t, t, c) \right) + \sigma_t z$$

where ϵ_θ is the noise prediction network (DiT), c is the conditioning information, and α_t are noise schedule parameters.

For this assignment, we use a pretrained [Diffusion Transformer \(DiT\)](#) trained on CIFAR-10 with class label conditioning. The sampling process is already implemented for you. To perform the conditioning, the pretrained model has a learned class embedding for each of the 10 classes, which is used to modulate each transformer layer via an Adaptive LayerNorm. This takes the class embedding as input and outputs LayerNorm scale and shift parameters:

$$x = (\text{AdALN}(\mathbf{F}_{\text{class embedding}}))(x)$$

For more details, see the Adaptive LayerNorm section of the original [Diffusion Transformer paper](#), or the `adaLN_modulation` function in `dit.py`.

Classifier-Free Guidance:

To improve prompt adherence, the provided code uses classifier-free guidance (Ho & Salimans, 2022), which amplifies the effect of conditioning:

$$\tilde{\epsilon}_\theta(x_t, t, c) = \epsilon_\theta(x_t, t, \emptyset) + w \cdot (\epsilon_\theta(x_t, t, c) - \epsilon_\theta(x_t, t, \emptyset))$$

where w is the guidance scale. During training, we randomly drop text conditioning with probability 0.1 to enable this technique.

Q-Former Architecture:

The Q-Former uses learnable query tokens that extract information from text features through attention mechanisms:

- **Learnable Queries:** A small set (2-8) of learned embedding vectors that act as "questions" to ask the language model
- **Self-Attention:** Queries interact with each other to build coherent representations
- **Cross-Attention:** Queries attend to GPT-2 hidden states to extract semantic information relevant for image generation
- **Feed-Forward Networks:** Process the attended features
- **Output Projection:** Project to the conditioning dimension expected by DiT

The Q-Former processes text features $\mathbf{F}_{\text{text}} \in \mathbb{R}^{B \times S \times 768}$ through multiple layers:

$$\begin{aligned} \mathbf{Q}^{(0)} &= \text{LearnedEmbedding}(\text{query_ids}) \\ \mathbf{Q}^{(l)} &= \mathbf{Q}^{(l-1)} + \text{SelfAttn}(\text{LN}(\mathbf{Q}^{(l-1)})) \end{aligned}$$

$$\mathbf{Q}^{(l)} = \mathbf{Q}^{(l)} + \text{CrossAttn}(\text{LN}(\mathbf{Q}^{(l)}), \mathbf{F}_{\text{text}})$$

$$\mathbf{Q}^{(l)} = \mathbf{Q}^{(l)} + \text{FFN}(\text{LN}(\mathbf{Q}^{(l)}))$$

where LN denotes LayerNorm. The final queries are projected to produce per-layer conditioning vectors for the DiT model. See Figure 1 for details.

Starter Code

The files are organized as follows:

```
handout/
  requirements.txt
  image_caption_data.py
  train_qformer.py
  eval_qformer.py
  run_in_cloud.ipynb
  dit.py
  ddpm.py
  attention_visualization.py
```

The codebase is partially built off of [this](#) open source diffusion implementation. We will use DDPM and the [CIFAR Dataset](#) for our implementation.

Below is a description of what you will find in each file. Functions are already implemented and provided for you, unless specified otherwise with TO DO.

1. `image_caption_data.py`: Functions for the image/caption data and building/caching LLM embeddings
 - `get_cifar_names`: Gets CIFAR class names
 - `get_synonyms`: Gets synonyms for CIFAR class names, for data augmentation
 - `_maybe_apply_synonym`: Possibly swaps out class names for synonyms
 - `get_text_captions`: Retrieves text captions from cached text caption path
 - `class CIFARTextDataset`: Dataset for training
 - `prompts_to_padded_hidden_states`: **TODO**: Function to turn text into GPT-2 embeddings
 - `build_text_states`: Uses `get_text_captions` to retrieve and optionally augment text captions with synonyms or dense labels
 - `cache_all_text_states`: Used to cache all GPT-2 embeddings before training for fast and more compute efficient training
2. `train_qformer.py`: Training script
 - `setup_optimizer_and_scheduler`: **TODO**: Given a model and a set of arguments, set up the optimizer, scheduler, and loss criterion
 - `maybe_load_resume_states`: Allows for restarting from a given optimizer and weight state in the case of disconnection

- `run_qformer_inference`: General inference function
 - `train_cifar10_qformer_dense_captions`: Main training loop
3. `eval_qformer.py`: Evaluation script
 - `eval_cifar10_qformer`: Main inference code
 4. `run_in_cloud.ipynb`: Cloud based ipynb task that has scripts to run training/inference
 5. `dit.py`: Model architecture
 - `DiT_Llama`: Pretrained diffusion transformer
 - `QueryEmbedder`: **TODO**: Q-Former implementation (init is provided for you)
 - `Attention`: Multi-head attention with cross-attention support
 - `FeedForward`: FFN with SwiGLU activation
 6. `ddpm.py`: DDPM pipeline
 - `DDPM`: forward diffusion and sampling of DDPM
 - `ddpm_schedules`: schedules for DDPM sampling
 7. `attention_visualization.py`: Visualization of cross-attention maps
 8. Download links to weights/datasets:
TODO: You should download the three items below. The URLs are in `run_in_cloud.ipynb`.
 - **Pretrained DiT weights**: A DiT with our architecture, trained on the class-conditional CIFAR dataset task
 - **Dense CIFAR text captions**: A dataset of dense CIFAR captions for enhancing text to image training
 - **Cached GPT-2 text embeddings**: Cached GPT-2 embeddings for the dense CIFAR caption dataset

Implementation

You will implement three key functions that enable Q-Former training and help you understand what it learns:

1. `prompts_to_padded_hidden_states`: Extract features from a pretrained LLM
2. `setup_optimizer_and_scheduler`: Configure which parameters train and how
3. `QueryEmbedder.forward`: Implement Q-Former to process text features

All other components (cross attention, training loop, inference) are provided. Your focus is on understanding how efficient fine-tuning works and analyzing the learned attention patterns.

Part 1: Extract LLM Features from any layer of GPT-2

Implement `prompts_to_padded_hidden_states` in `shared.py`.

```
def prompts_to_padded_hidden_states(
    prompts: List[str],
    gpt2,
    tokenizer,
    gpt2_layer_index: int,
    device: torch.device,
):
    """
    Convert a list of text prompts to a single padded tensor
    of GPT-2 hidden states.

    Returns the hidden states from that specific layer along
    with masks as a tuple (hidden_states, masks)
    """
```

Your task:

- For all prompts, do the following:
 - Call the tokenizer with `return_tensors="pt"` to get valid inputs for GPT-2
 - Call GPT-2 on these inputs, with `output_hidden_states` set to `True`. You can call `outputs.hidden_states` after this to get the hidden states
 - Given these hidden states, use the `gpt2_layer_index` to determine exactly which layer of the hidden states to keep for this prompt
- After you have a list of the kept hidden states for all prompts, perform max-length padding of the hidden states, so that the final hidden state size is

(Batch_size, max_token_length_in_batch, gpt2_hidden_dimension).
- **NOTE:** Also return a boolean padding tensor of size (Batch_size, max_seq) where `True` indicates that a token is not a padded token

Part 2: Setup Optimizer and Scheduler

Implement `setup_optimizer_and_scheduler(model, args)` in `train_qformer.py`.

```
def setup_optimizer_and_scheduler(model, args):
    """
    Configure training: freeze/unfreeze parameters, initialize,
    create optimizer/scheduler.

    Returns: optimizer, scheduler, criterion
    """
```

Step 1: Freeze and Unfreeze Parameters (Parameter-Efficient Training)

The key insight: We want to train only the Q-Former while keeping DiT (~100M) and GPT-2 (~117M) frozen. This makes training fast and prevents catastrophic forgetting of pretrained capabilities.

Your task:

- Set `requires_grad = False` for all model parameters, except for those in the query embedder

Step 2: Create Optimizer

Your task:

Create `torch.optim.AdamW` with:

- Only parameters where `requires_grad=True`
- `lr=args.lr, weight_decay=args.weight_decay, eps=args.adam_epsilon`
- `betas=(0.9, 0.95)`

Step 3: Create Learning Rate Scheduler

We might expect the first few steps of training to be very unstable, as we're replacing the class embedding from a pretrained model with random noise from a randomly initialized Q-Former. We can add warmup steps to smoothly begin training and determine good momentum and RMS parameters for AdamW.

Your task:

If `args.warmup_steps > 0`, create a linear warmup scheduler:

- Use `torch.optim.lr_scheduler.LambdaLR`
- The lambda function should return: `min(1.0, (step + 1) / warmup_steps)`
- This gradually increases learning rate (LR) from 0 to `args.lr` over the first `warmup_steps`
- If `warmup_steps <= 0`, set `scheduler = None`

Why warmup? Warmup gradually ramps up the learning rate to prevent unstable updates before model weights and optimizer statistics are well-initialized.

Step 4: Create Loss and Return*Your task:*

- Set criterion to MSE loss
- Return designated parameters.

Part 3: Q-Former ForwardImplement forward in `./mindiffusion/dit.py`.

```
def forward(self, cross_attention_states, batch_size=1,
            cross_attention_mask=None):
    """
    Forward pass of the Q-Former query embedder.

    Returns: Final per-layer features
             (batch_size, num_dit_layers, conditioning_hidden_size)
    """
```

Your task:

- Get the set of learnable query embeddings from the query table using indices 0 to `num_queries-1`.
- Expand the queries to shape `(batch_size, num_queries, transformer_hidden_size)` and put the queries on the same device as the `cross_attention_states`.
- For each block in `self.blocks`, do the following:
 - Apply the self-attention block (i.e. `x -> x + self_attn(attn_norm(x))`).
 - Apply the cross-attention block (i.e. `x -> x + cross_attn(cross_norm(x, cross_attn_states, cross_attention_mask))`).
 - Apply the forward-feed block (i.e. `x -> x + ffn(ffn_norm(x))`).
 - You are implementing the equations under Q-Former Architecture.
- Use `output_proj` to project the resulting shape to the `conditioning_hidden_size`.
- Map the queries to the per-layer features using the `query_to_layer` linear layer to get a tensor with a shape of `(batch_size, num_dit_layers, conditioning_hidden_size)`.
- Returns per-layer features.

Hint: If you are stuck, look in `__init__()` to see what you have to work with.

Part 4: Training

We train using cached embeddings from GPT-2² with the following text prompts:

- simple text prompts with class name: "a photo of a dog"
- simple text prompts with a synonym for class name: "a photo of a four legged animal that barks"
- dense text prompts with class name: "a photo of a black dog in a grassy field"
- dense text prompts with a synonym for class name: "a photo of a four legged animal that barks in a grassy field"

These different captions allow the Qformer to learn how to transform semantic text information into dense conditioning information to the pretrained and frozen DiT. The training loop is provided but you should understand how it works:

Algorithm 1 Q-Former Training (Provided)

```

1: Load pretrained DiT: model.load_state_dict(torch.load(pretrained_path))
2: Set up the optimizer, scheduler, and criterion
3: for each epoch do
4:   for each batch  $(x_{\text{img}}, y_{\text{caption}})$  from densely labeled, randomly captioned CIFAR-10 do
5:     Sample timestep:  $t \sim \text{Uniform}(1, T)$ 
6:     Sample noise:  $\epsilon \sim \mathcal{N}(0, I)$ 
7:     Add noise:  $x_t = \sqrt{\bar{\alpha}_t}x_{\text{img}} + \sqrt{1 - \bar{\alpha}_t}\epsilon$ 
8:     Load pre-computed text embedding  $y_{\text{emb}}$  for text  $y_{\text{caption}}$ 
9:     or calculate  $y_{\text{emb}} = \text{extract\_features}(\text{gpt2}, y_{\text{caption}})$ 
10:    Predict noise:  $\epsilon_{\text{pred}} = \text{DiT}(x_t, t, y_{\text{emb}}, \text{text})$ 
11:    Compute loss:  $\mathcal{L} = \|\epsilon - \epsilon_{\text{pred}}\|^2$ 
12:    Backpropagate and update only Q-Former parameters
13:  Generate sample images to monitor training progress
14:  Save Q-Former weights

```

²we could have had you generate these, but it would add unnecessary extra time to the already long assignment

Running Training:

See the `run_in_cloud.ipynb` file for the training script. Keep all hyperparameters from these files, but feel free to inspect `train_qformer.py` to see all the arguments that you can use and how they work.

NOTE: If your training gets killed after 5 or more epochs, the script will automatically save both your weights and optimizer states every 5 epochs. Use `--resume_optimizer_path`, `--resume_model_path`, and `--start_epoch` to restart from an epoch that is a multiple of 5.

Recommended: Sign in to Google Colab to claim Colab Pro, which is available when you verify with your CMU email. We strongly advise using the A100 GPU when it's available, but if not, using a L4 or T4 should work as well.

Running Inference:

Using the `run_in_cloud.ipynb` file, run `eval_qformer.py` in Colab to run inference.

NOTE: The `inference.yaml` function has a few examples of how to run inference, but you might have to add more to it to generate all the results for empirical questions.

Understanding the Big Picture

Your Implementation	What It Enables
Freeze/unfreeze parameters	Preservation of pretrained model weights and fast training
Optimizer & scheduler	Stable gradient updates
QFormer forward	Understanding how the QFormer model works

After implementing these functions, the training loop (provided) will:

1. Load cached text embeddings from the provided link
2. Train only the Q-Former using your optimizer
3. Save samples each epoch for monitoring

Key Insight: You're implementing the "glue code" that makes efficient multi-modal fine-tuning work. The architecture is provided; your job is to configure *how* it learns and to instrument it for analysis.

Empirical Questions

6.1. Q-Former Architecture Understanding + Dataset Dive

- 6.1.a. (2 points) Looking at the code, we observe that each of the Q-Former transformer block has 11.8M parameters, with 2 such blocks. Combining that with the number of parameters of the **query embeddings** and **output projections** how many trainable parameters does the Q-Former have given the config in the training script? (round to nearest million).

Suppose the GPT-2 has 125M parameters, and DiT has 18M parameters, what percentage of the weights in the entire process are trained (not frozen) in our method?

- 6.1.b. (2 points) Paste 3 lines from the densely captioned CIFAR dataset. The indices should match with the CIFAR dataset.

6.2. Q-Former Training Understanding

Consider our Q-Former setup with the following specifications:

- the DiT is pretrained on a large corpus of high quality images I
- the LLM is pretrained on a large corpus of text T
- the Q-Former training set is on a small set of medium quality images I' with corresponding text T'

We will consider two scenarios:

- Scenario A : Same as in the homework; we only train the Q-Former and keep the LLM and DiT fixed.
- Scenario B : We will unfreeze parameters of the DiT and LLM along with the Q-Former while training.

6.2.a. (2 points) Compare and contrast scenario A and B in terms of final asymptotic loss value when trained continuously. Would you expect one scenario to have a lower asymptotic loss, and if so, which one?

6.2.b. (2 points) Consider running a general text generation validation set (i.e. for a token prediction task) on the LLM used at the end of training in scenarios A and B . Which would you expect to have higher mean token accuracy on this general validation set?

6.3. Q-Former Training

As indicated in `run_in_cloud.ipynb`, train your Q-Former for 50 epochs with `num_queries=4`. [Expected runtime on T4: 2-3 hours] [Expected runtime on A100: 1 hour]

- 6.3.a. (2 points) Paste your training Exponential Moving Average (EMA) loss curve for the first 1000 steps.



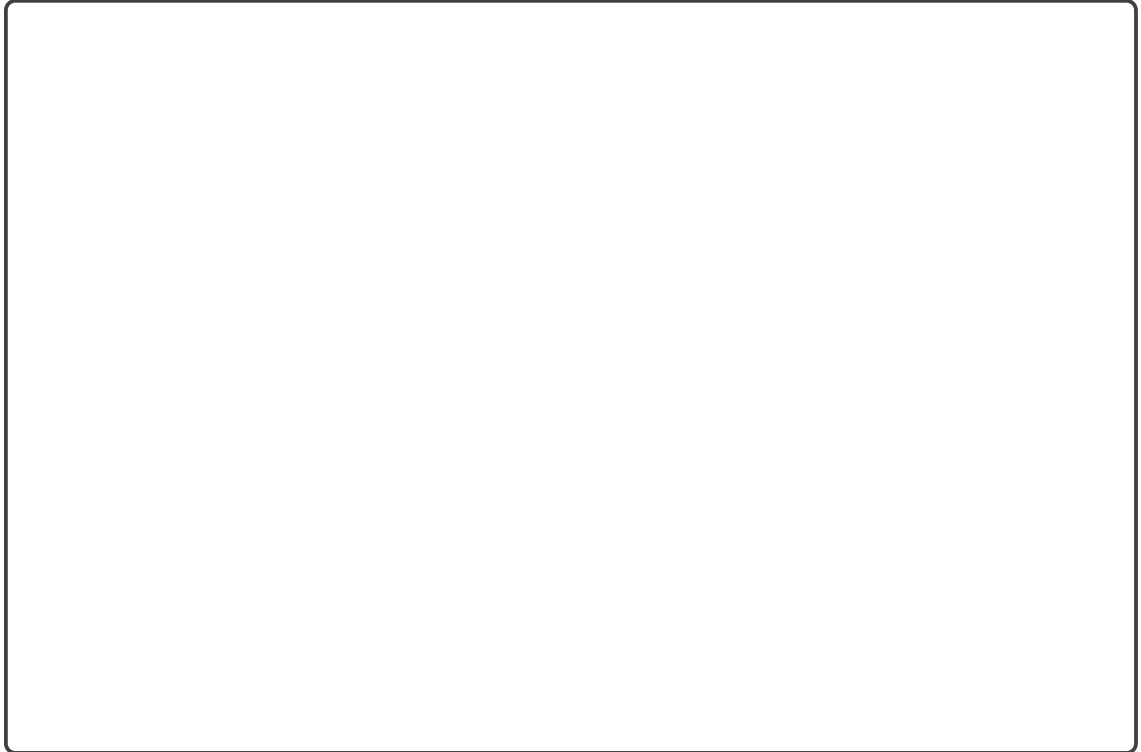
- 6.3.b. (3 points) Paste the sample grid from epochs 5, 25, and 49 (these should be saved in wandb). How does image quality/prompt adherence improve during training?



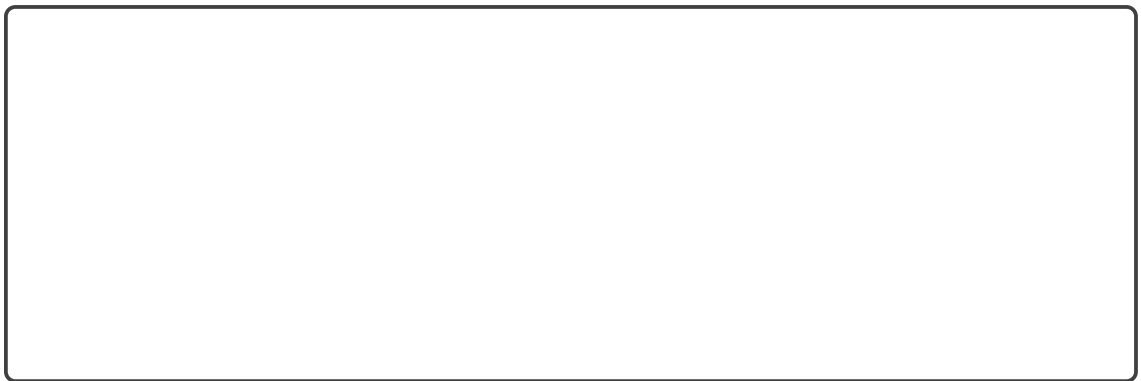
6.4. Classifier-Free Guidance

Experiment with different CFG scales with the original pretrained model: $w \in \{1.0, 2.0, 3.0, 4.0\}$ for the class “deer” (class 4).

- 6.4.a. (2 points) Paste the 4 images generated with different guidance scales. How does the appearance change as you increase w ? Which of these actually use the conditional information from the text embeddings?



- 6.4.b. (1 point) Based on your results, what seems to be a good guidance scale for this model?



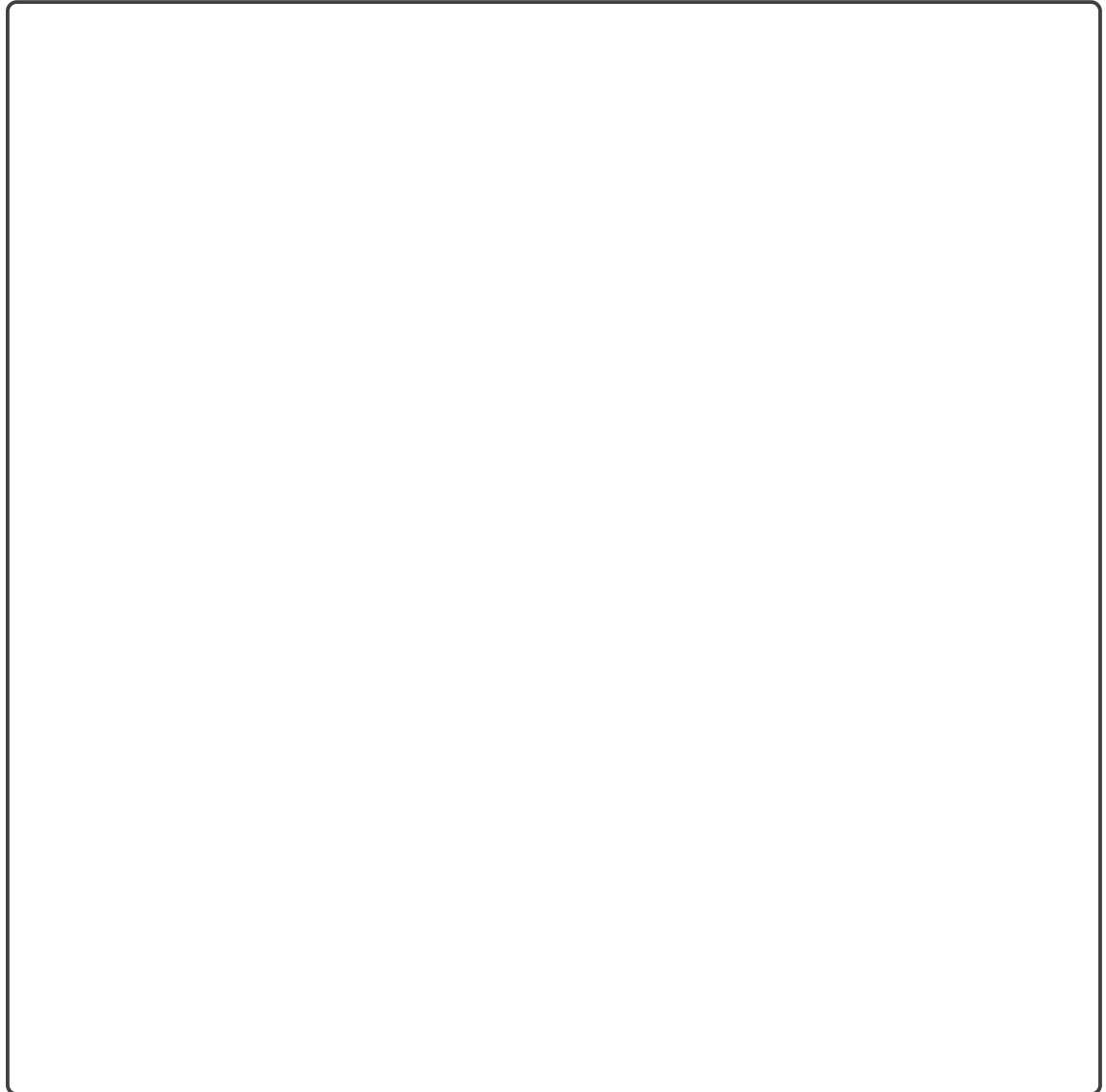
6.5. Text-Conditioned Generation

After training, generate images from custom text prompts.

6.5.a. (2 points) Generate 8 images of cars under two conditions with cfg 3.0:

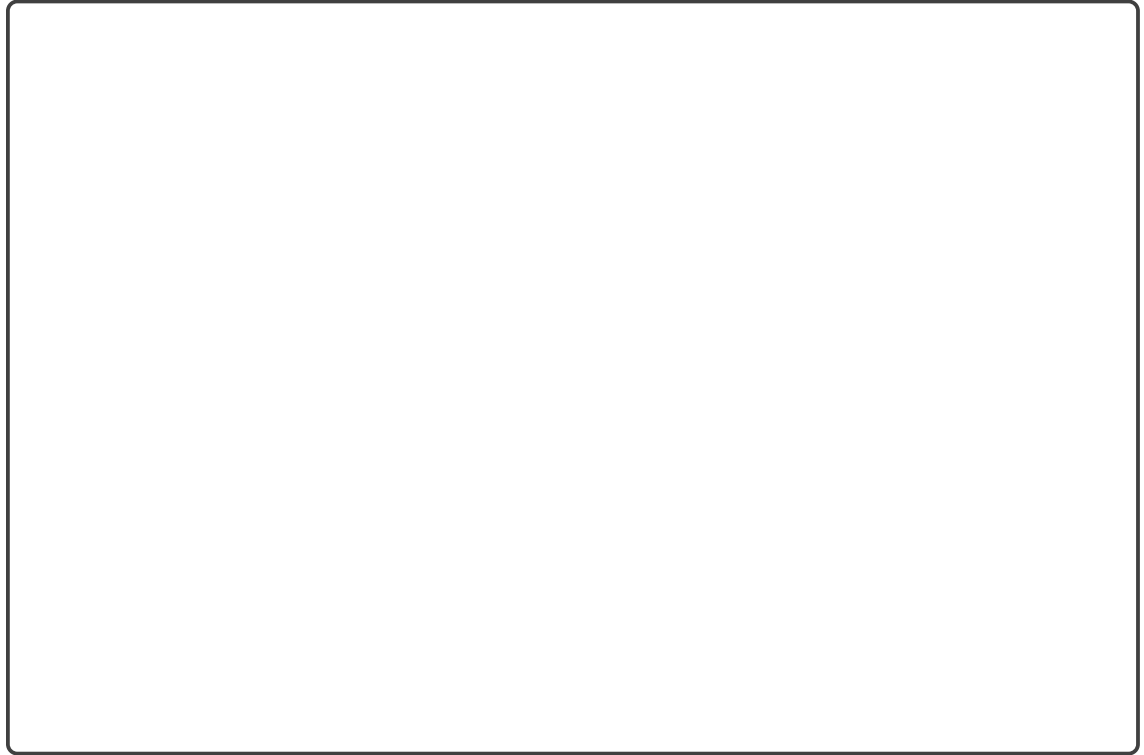
- 8 images with the pretrained class-conditional DiT with class label for automobile
- 8 images with the Qformer DiT with text “a photo of a blue automobile with a bright blue sky”

Paste all 8 images. Is there any noticeable difference between the two sets of images?



- 6.5.b. (2 points) Compare generation of OOD underspecified prompts. Generate 4 photos with caption: “a photo of a vehicle that goes on grass” and compare with 4 photos with caption: “a photo of a vehicle that goes in water” with cfg 3.0.

Paste the 8 images: How do these images differ?



6.6. Cross-Attention Maps

Finally, we will examine the cross-attention maps between the LLM text features and the Q-Former. The Q-Former uses cross-attention to query the text embeddings. By saving these attention patterns, we can visualize which words the model focuses on when generating specific image features.

- 6.6.a. (6 points) Using the text prompt “a photo of a red airplane with a green field in the background”, generate the attention maps of both layers of the Q-Former at diffusion timesteps $t = 500$ and $t = 100$. Paste these 4 attention maps as well as the final image.

Looking at the attention maps, what do you notice about what tokens are getting attended to and what are getting completely ignored? Why might this be the case?

Hint: Take a look at [Darcet et al. \(2023\)](#) and [Xiao et al. \(2023\)](#).

7 Code Upload (0 points)

7.1. (0 points) Did you upload your code to the appropriate programming slot on Gradescope?

Hint: The correct answer is ‘yes’.

☐ Yes

☐ No

For this homework, you should upload `train_qformer.py`, `dit.py`, and `image_caption_data.py`.

8 Collaboration Questions (2 points)

After you have completed all other components of this assignment, report your answers to these questions regarding the collaboration policy. Details of the policy can be found in the syllabus.

- 8.1. (1 point) Did you collaborate with anyone on this assignment? If so, list their name or Andrew ID and which problems you worked together on.

- 8.2. (1 point) Did you find or come across code that implements any part of this assignment? If so, include full details.