

National University of Computer and Emerging Sciences
Islamabad Campus

Generative AI (AI-4009)

Sessional-II Exam

Course Instructor:

Akhtar Jamil

Total Time (Hrs): 1

Total Marks: 50

Total Questions: 3

Date: April 10, 2025

Roll No

Section

Student Signature

Do not write below this line.

Question No 1. MCQ [25 x 1 = 25]

You must answer the MCQs on the given MCQ Answer sheet. Any question marked somewhere else is not considered. Overwriting an MCQ will result in ZERO marks.

1. What role does the UNet model play in the Stable Diffusion pipeline?
 - A) It predicts the noise added to the latent during diffusion**
 - B) It decodes latent space to images
 - C) It encodes text to token embeddings
 - D) It selects image regions for enhancement
2. A student proposes increasing β_t rapidly toward the final steps ($t = 900 - 1000$). What effect is most likely?
 - A) Gradual generation of finer details
 - B) Loss of semantic content due to excessive noise**
 - C) Improved convergence of the VAE encoder
 - D) The UNet receiving perfect clean signals at early time step t
3. In the context of diffusion models, what does the Diffusion Kernel represent?
 - A) A neural network architecture for inpainting
 - B) A fixed convolution matrix for downsampling
 - C) A probabilistic function modeling step-by-step noise addition**
 - D) A loss function used for image reconstruction

National University of Computer and Emerging Sciences

Islamabad Campus

4. Which of the following best explains the "memoryless" nature of the forward diffusion process modeled as a Markov chain?
- A) The amount of noise added is fixed for all steps regardless of input
 - B) Each step depends only on the original image x_0
 - C) The model forgets all past weights during training
 - D) Each state x_t is computed based solely on the immediate previous state x_{t-1}**

5. The forward diffusion process is defined as:

$$x_t = \sqrt{\alpha_t} x_0 + \sqrt{1 - \alpha_t} \epsilon, \text{ where } \epsilon \sim \mathcal{N}(0, I)$$

As $t \rightarrow T$, what does x_t will look like?

- A) The clean image x_0
 - B) A uniform distribution over pixel space
 - C) A zero matrix
 - D) Pure Gaussian noise**
6. What is the main purpose of incorporating time embeddings (such as sinusoidal or learned positional encodings) into the denoising network used in diffusion models?
- A) To encode spatial location of each pixel
 - B) To control the noise schedule externally
 - C) To enable the model to compute variance adaptively at runtime
 - D) To condition the model on the current timestep t and guide noise prediction accordingly**
7. Why are the weights of the Discriminator and Generator updated in an alternating manner during GAN training?
- A) To prevent premature convergence
 - B) To maintain balanced learning between the Generator and Discriminator**
 - C) To reduce overfitting
 - D) To encourage diverse outputs
8. In StyleGAN, what does "stochastic variation" refer to?
- A) The random changes in the overall structure of the generated image
 - B) The fine details introduced in the image, such as hair, eye color, etc.**
 - C) The modification of the style vector across layers
 - D) The adjustment of resolution at different stages of generation

National University of Computer and Emerging Sciences

Islamabad Campus

9. What does bidirectional attention in BERT mean in practical terms?
- A) Each token attends only to previous tokens
 - B) Separate forward and backward context is computed like in BiRNNs
 - C) Attention flows layer by layer in opposite directions
 - D) Each token attends to all tokens, capturing context from both directions**
10. In the context of a Conditional GAN, where x is the conditional input data, y is the target output, and z is the noise vector, which of the following terms in the loss function ensures that the generated output matches the label y ?
- a). $\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}[\|y - G(x, z)\|_1]$**
 - b). $\mathcal{L}_{cos}(G) = \mathbb{E}_{x,y,z}[1 - \cos(y, G(x, z))]$
 - c). $\mathcal{L}_{BCE}(G) = -\mathbb{E}_{x,y,z}[y \log(G(x, z)) + (1 - y) \log(1 - G(x, z))]$
 - d). None of the above
11. In Conditional GANs, if the term $E_{x \sim p_{data}(x)}[\log D(x|y)]$ is high, what does this indicate?
- a). The discriminator is struggling to distinguish real samples from generated samples.
 - b). The generator is producing samples that match the condition (y) accurately.
 - c). The discriminator is confidently identifying real samples conditioned on (y) as real.**
 - d). The generator has successfully "fooled" the discriminator for most samples.
12. In CycleGAN, let $G: X \rightarrow Y$ represent the generator mapping facial contours x to human faces (y) and ($F: Y \rightarrow X$) represent the inverse mapping. Which of the following is a suitable cycle consistency loss term to ensure that mapping x to y and back to x preserves the original facial contours?
- a). $E_{x \sim p_{data}(x)}[|G(F(x)) - x|_1]$
 - b). $E_{x \sim p_{data}(x)}[|F(G(x)) - x|_1]$**
 - c). $E_{x \sim p_{data}(x)}[|F(G(x)) + x|_1]$
 - d). $E_{x \sim p_{data}(x)}[|G(x) - x|_1]$

National University of Computer and Emerging Sciences

Islamabad Campus

13. Assume that after training the generator of a GAN for 100 epochs, the probability distribution $p_g(x)$ of generated data matches with the real data probability distribution p_{data} . At this stage, if we randomly select a sample from real data and pass to the discriminator, what is the most probable output probability $D(x)$ that the discriminator should assign to x at this point?
- a). 0
 - b). 0.5**
 - c). 1
 - d). Undefined
14. In the Vision Transformer architecture, why is a classification token prepended to the input sequence?
- A) To learn convolution weights for patch attention
 - B) To reduce the number of positional encodings
 - C) To represent the entire image during classification via its final state**
 - D) To signal the end of a patch sequence
15. What is the mathematical reason for dividing the dot product of query and key vectors by $\sqrt{d_k}$ in self-attention?
- A) To prevent the softmax function from producing vanishing gradients**
 - B) To normalize by batch size
 - C) To encourage sharper attention over longer sequences
 - D) To reduce computational complexity
16. The StyleGAN generator starts from a constant learned input tensor instead of directly sampling from the latent space.
- A) True**
 - B) False
17. What **key architectural change** allows **Conditional GANs** to generate **class-specific outputs**?
- A) Use of batch normalization in the generator
 - B) Injecting conditional information y into both the generator and discriminator**
 - C) Using sigmoid activation at the output
 - D) Replacing convolution with max-pooling

National University of Computer and Emerging Sciences

Islamabad Campus

18. How does the Cycle Consistency Loss in CycleGAN help training?
- A) Encourages diversity among outputs
 - B) Aligns latent space distributions of different domains
 - C) Ensures translating from one domain to another and back yields the original image**
 - D) Reduces computational cost
19. Which loss would most directly help a CycleGAN retain color composition when translating paintings to photos?
- A) Adversarial loss
 - B) MSE loss
 - C) Identity loss**
 - D) Gradient penalty
- In StyleGAN, the "style" vector is introduced to:
- a). Define the image resolution for the generator output
 - b). Apply spatial transformations to image patches
 - c). Limit the diversity of generated images
 - d). Control the scale and translation of features in each layer**
20. How does disentangling the latent space benefit StyleGAN?
- a). Reduces training time
 - b). Allows independent control over features for generation**
 - c). Improves resolution
 - d). Lowers computational cost
21. What is the role of noise injection in StyleGAN's generator?
- a). Control global features
 - b). Add fine-grained stochastic details**
 - c). Ensure consistency
 - d). Add redundancy
22. What would happen if [MASK] tokens used in pretraining were also used during fine-tuning?
- A) Model accuracy would improve drastically
 - B) The model would fail to generalize to downstream tasks**

National University of Computer and Emerging Sciences

Islamabad Campus

- C) The model would overfit
 - D) It would behave like a generative decoder
23. Which component in BERT helps it distinguish between tokens from Sentence A and Sentence B during NSP?
- A) Positional embeddings
 - B) Segment embeddings**
 - C) Token embeddings
 - D) Output embeddings
24. BERT can perform text Generation.
- A) True
 - B) False**
25. In multi-head attention with 8 heads, each head uses a separate projection of the same input. What is the main advantage of this design?
- A) Reduces dimensionality of the input sequence
 - B) Encourages parameter sharing for optimization
 - C) Ensures faster gradient propagation through skip connections
 - D) Allows the model to learn attention from multiple subspaces and capture diverse relationships**

National University of Computer and Emerging Sciences

Islamabad Campus

Question No 2. [3 x 5=15]

Write short answers to the following questions.

1. In the forward diffusion process, what does it mean for x_T to converge to an isotropic Gaussian, and how does the variance schedule β_1, \dots, β_T influence this behavior?

An **isotropic Gaussian** refers to a multivariate normal distribution where all directions have the same variance. In the forward diffusion process, as noise is added over many steps, the data distribution x_T gradually loses its structure and converges toward pure Gaussian noise.

This convergence to isotropic Gaussian occurs **only if the variance schedule β_1, \dots, β_T is well-behaved**—i.e., the noise is added incrementally and smoothly.

2. What are the advantages of operating in latent space rather than pixel space for diffusion models?

Operating in latent space reduces the input dimensionality, which lowers memory and computational cost. It enables training on high-resolution images that would be infeasible in pixel space

3. Consider the Input vector $x = [4, 6, 8]$, Style scaling vector $y_s = [2.0, 1.5, 1.0]$ and Style bias vector $y_b = [1.0, -1, 0.5]$.

Calculate the Adaptive Instance Normalization (AdaIN) output using the formula:

$$\text{AdaIN}(x, y) = y_s \cdot \frac{x - \mu(x)}{\sigma(x)} + y_b$$

Note: show all steps and in the last line write the final output.

Solution:

1. Calculate the mean $\mu(x)$:

$$\mu(x) = \frac{4+6+8}{3} = 6$$

2. Calculate the standard deviation $\sigma(x)$:

$$\sigma(x) = \sqrt{\frac{(4-6)^2 + (6-6)^2 + (8-6)^2}{3}} = \sqrt{\frac{4+0+4}{3}} = \sqrt{\frac{8}{3}} \approx 1.63$$

3. Normalize (x) :

$$\frac{x - \mu(x)}{\sigma(x)} = \left[\frac{4-6}{1.63}, \frac{6-6}{1.63}, \frac{8-6}{1.63} \right]$$

$$= [-1.23, 0, 1.23]$$

National University of Computer and Emerging Sciences

Islamabad Campus

4. Apply AdaIN formula:

$$\text{AdaIN}(x, y) = y_s \cdot \frac{x - \mu(x)}{\sigma(x)} + y_b$$

$$= [2, 1.5, 1] \cdot [-1.23, 0, 1.23] + [1, -1, 0.5]$$

$$= [-2.46, 0, 1.23] + [1, -1, 0.5]$$

$$= [-1.46, -1, 1.73]$$

$$\text{AdaIN} \approx [-1.46, -1, 1.73].$$

4. In a transformer model for language translation, the decoder is trained using embeddings of both source and target language. However, during testing, we do not have access to the target sentence. Since the decoder still expects an input at each step, how can we handle this situation to enable the decoder to generate the translation correctly without having the true target embeddings?

During testing the decoder starts with a special START token. Using this token it actually generates the desired first output token. This process continues in an autoregressive manner where START token and generated tokens are feed back into the decoder to predict next token. The process stops when model predicts the END token.

5. What is Mixing Regularization as used in StyleGAN?

Mixing regularization in StyleGAN involves using two different latent codes. A random crossover point is selected among the layers of the model. Up to this crossover point, the model applies the first latent code, while for the remaining layers, it applies the second latent code. This process helps introduce diversity and improve the quality of generated images.

Question No 3. [10]

Prepare the input for the transformer model. First generate the positional embeddings for the given input sequence and add with input embeddings. Using these final embeddings, calculate the attention scores according to the attention formula given below. Assume that all linear layer weights are all set to 0.5 and biases are set to 0 for each required layer.

$$\text{Input Embeddings} = \begin{bmatrix} 0.1 & 0.2 \\ 0.5 & 0.6 \\ 0.9 & 1.0 \end{bmatrix}$$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

National University of Computer and Emerging Sciences Islamabad Campus

Step 1: Input Embeddings

$$\text{Input Embeddings} = \begin{bmatrix} 0.1 & 0.2 \\ 0.5 & 0.6 \\ 0.9 & 1.0 \end{bmatrix}$$

Positional Encodings

$$\text{PE}_{(pos, 2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d}}}\right)$$

$$\text{PE}_{(pos, 2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d}}}\right)$$

$$\text{Positional Encodings} = \begin{bmatrix} 0 & 1 \\ 0.8415 & 0.5403 \\ 0.9093 & -0.4161 \end{bmatrix}$$

$$\text{Final Embeddings} = \begin{bmatrix} 0.1 & 0.2 \\ 0.5 & 0.6 \\ 0.9 & 1.0 \end{bmatrix} + \begin{bmatrix} 0 & 1 \\ 0.8415 & 0.5403 \\ 0.9093 & -0.4161 \end{bmatrix}$$

$$\text{Final Embeddings} = \begin{bmatrix} 0.1 & 1.2 \\ 1.3415 & 1.1403 \\ 1.8093 & 0.5839 \end{bmatrix}$$

Compute Q , K , and V Matrices, using the weights (0.5) and biases (0), we calculate Q , K , and V from the final embeddings.

$$Q = K = V = 0.5 \times \text{Final Embeddings} = 0.5 \times \begin{bmatrix} 0.1 & 1.2 \\ 1.3415 & 1.1403 \\ 1.8093 & 0.5839 \end{bmatrix}$$

$$Q = K = V = \begin{bmatrix} 0.05 & 0.6 \\ 0.67075 & 0.57015 \\ 0.90465 & 0.29195 \end{bmatrix}$$

$$QK^T = \begin{bmatrix} 0.3625 & 0.3756275 & 0.2204025 \\ 0.3756275 & 0.7749765 & 0.773154 \\ 0.2204025 & 0.773154 & 0.903617 \end{bmatrix}$$

National University of Computer and Emerging Sciences Islamabad Campus

Approximating each division:

$$\frac{QK^T}{\sqrt{d_k}} \approx \begin{bmatrix} 0.2562 & 0.2656 & 0.1558 \\ 0.2656 & 0.5479 & 0.5467 \\ 0.1558 & 0.5467 & 0.6388 \end{bmatrix}$$

Apply Softmax

Next, we apply the softmax function row-wise to get the attention score

Example for the first row:

$$\text{Softmax}(0.2562, 0.2656, 0.1558) = \left(\frac{e^{0.2562}}{\sum e^{\text{row}}}, \frac{e^{0.2656}}{\sum e^{\text{row}}}, \frac{e^{0.1558}}{\sum e^{\text{row}}} \right)$$

$$\text{Softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) \approx \begin{bmatrix} 0.343 & 0.346 & 0.310 \\ 0.346 & 0.424 & 0.389 \\ 0.310 & 0.389 & 0.443 \end{bmatrix}$$

$$\text{Attention Output} = \text{Softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V$$



$$\text{Attention Output} = \begin{bmatrix} 0.343 & 0.346 & 0.310 \\ 0.346 & 0.424 & 0.389 \\ 0.310 & 0.389 & 0.443 \end{bmatrix} \times \begin{bmatrix} 0.05 & 0.6 \\ 0.67075 & 0.57015 \\ 0.90465 & 0.29195 \end{bmatrix}$$

$$\text{Attention Output} \approx \begin{bmatrix} 0.5297 & 0.4936 \\ 0.6536 & 0.5629 \\ 0.6772 & 0.5371 \end{bmatrix}$$

National University of Computer and Emerging Sciences Islamabad Campus

Answer Sheet MCQs

Mark (X) for the correct option. Only one option must be selected. Selection of multiple options or overwriting will result in ZERO marks.

Instruction for filling the sheet 1. This sheet should not be folded or crushed 2. Use only blue/black ball pen	
CORRECT METHOD 	WRONG METHOD 
NAME:	ROLL NO:

<div style="text-align: center;">Roll No</div> <table style="margin: auto;"> <tr><td style="border: 1px solid black; width: 20px; height: 20px;"></td><td style="border: 1px solid black; width: 20px; height: 20px;"></td><td style="border: 1px solid black; width: 20px; height: 20px;"></td><td style="border: 1px solid black; width: 20px; height: 20px;"></td></tr> </table> <div style="display: flex; justify-content: space-between;"> <div style="width: 45%;"> <div style="display: flex; justify-content: space-between; margin-bottom: 5px;"> <div style="width: 10px;"></div> <div style="width: 10px;"></div> <div style="width: 10px;"></div> <div style="width: 10px;"></div> </div> <div style="display: flex; justify-content: space-between;"> <div style="width: 10px;"></div> <div style="width: 10px;"></div> <div style="width: 10px;"></div> <div style="width: 10px;"></div> </div> </div> <div style="display: flex; justify-content: space-between;"> <div style="width: 45%;"> <div style="display: flex; justify-content: space-between; margin-bottom: 5px;"> <div style="width: 10px;"></div> <div style="width: 10px;"></div> <div style="width: 10px;"></div> <div style="width: 10px;"></div> </div> <div style="display: flex; justify-content: space-between;"> <div style="width: 10px;"></div> <div style="width: 10px;"></div> <div style="width: 10px;"></div> <div style="width: 10px;"></div> </div> </div> </div> <div style="width: 45%;"> <div style="display: flex; justify-content: space-between; margin-bottom: 5px;"> <div style="width: 10px;"></div> <div style="width: 10px;"></div> <div style="width: 10px;"></div> <div style="width: 10px;"></div> </div> <div style="display: flex; justify-content: space-between;"> <div style="width: 10px;"></div> <div style="width: 10px;"></div> <div style="width: 10px;"></div> <div style="width: 10px;"></div> </div> </div> </div>				

MCQs

Section1