Instruções: Responda às questões de múltipla escolha assinalando apenas uma alternativa. Nas dissertativas, apresente o raciocínio passo a passo, incluindo fórmulas e cálculos intermediários quando solicitado. Mantenha a notação dos slides: β_t para o *noise schedule*, $\alpha_t = \prod_{s=1}^{t}(1-\beta_s)$ para o produto acumulado, z_t para o estado em t, ϵ para o ruído gaussiano e $g_t[z_t, \theta_t]$ para a rede que estima o ruído. Questões marcadas com ☠ são consideradas difíceis.

Questões de Múltipla Escolha.

I) Considere o kernel de difusão $z_t = \sqrt{\alpha_t}\, x + \sqrt{1-\alpha_t}\, \epsilon$, com $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. Dado $\alpha_t = 0.64$, $x = 2$ e $\epsilon = 0.5$, qual é o valor amostrado de z_t?

(a) 1.90.

(b) 1.60.

(c) 1.46.

(d) 2.10.

II) Para um *schedule* $\boldsymbol{\beta} = (0.05,\ 0.10,\ 0.20)$ com $T = 3$, qual é o valor de α_3?

(a) 0.684.

(b) 0.350.

(c) 0.650.

(d) 0.001.

III) No processo de difusão, $q(z_t \mid x) = \mathcal{N}_{z_t}[\mu,\ \Sigma]$. Para $\alpha_t = 0.49$ e $x = 4$, quais são, respectivamente, a média μ e a variância Σ (esta última como múltiplo de $\mathbf{I}$)?

(a) $\mu = 2.80$ e $\Sigma = 0.51\,\mathbf{I}$.

(b) $\mu = 1.96$ e $\Sigma = 0.49\,\mathbf{I}$.

(c) $\mu = 2.80$ e $\Sigma = 0.71\,\mathbf{I}$.

(d) $\mu = 2.04$ e $\Sigma = 0.51\,\mathbf{I}$.

IV) Após a reparametrização que faz a rede $g_t[z_t, \theta_t]$ predizer o ruído, a função de custo do modelo difusor (excluindo o termo de reconstrução) assume qual forma?

(a) $\sum_{i,t} \left| g_t[z_t, \theta_t] - x^{(i)} \right|^2$.

(b) $\sum_{i,t} \left| g_t[z_t, \theta_t] - z_{t-1} \right|^2$.

(c) $\sum_{i,t} \left| g_t[z_t, \theta_t] - \epsilon_t \right|^2$.

(d) $\sum_{i,t} -\log g_t[z_t, \theta_t]\, \epsilon_t$.

V) Sobre o processo de difusão (*forward*) definido nos slides, qual afirmação é correta?

(a) A transição $q(z_t \mid z_{t-1})$ depende explicitamente de x e de todos os estados anteriores.

(b) A cadeia é markoviana, com $q(z_t \mid z_{t-1}, x) = q(z_t \mid z_{t-1})$, e a variância da transição é $\beta_t\,\mathbf{I}$.

(c) A transição é determinística e a aleatoriedade entra somente em $z_T \sim \mathcal{N}(0, \mathbf{I})$.

(d) A média de $q(z_t \mid z_{t-1})$ é $\sqrt{\alpha_t}\, z_{t-1}$, com variância $(1-\alpha_t)\,\mathbf{I}$.

VI) ☠️ Considere o *schedule* $\boldsymbol{\beta} = (0.1,\ 0.2,\ 0.3,\ 0.4)$ com $T = 4$, que produz $\alpha = (0.900,\ 0.720,\ 0.504,\ 0.302)$. Definindo a razão sinal-ruído como $\mathrm{SNR}(t) = \alpha_t/(1-\alpha_t)$, em qual passo $\mathrm{SNR}(t)$ é o primeiro a se aproximar de 1 (i.e., $|\mathrm{SNR}(t) - 1|$ é mínimo)?

(a) Em $t = 2$.

(b) Em $t = 3$.

(c) Em $t = 4$.

(d) Em $t = 1$.

VII) Nos slides, a rede g_t é treinada para predizer o ruído ϵ adicionado, e não x ou z_{t-1} diretamente. Qual a justificativa apresentada para essa escolha?

(a) Predizer o ruído evita o uso de *skip connections* na U-Net.

(b) Predizer o ruído elimina a necessidade de *noise schedule* no *forward process*.

(c) Predizer o ruído é uma observação empírica: na prática produz melhores resultados que predizer x diretamente.

(d) Predizer o ruído torna a *loss* numericamente equivalente a uma entropia cruzada.

VIII) A distribuição condicional $q(z_{t-1} \mid z_t, x)$, derivada no curso, é uma gaussiana cuja média é uma combinação linear de z_t e x. Para $\beta_t = 0.2$, $\alpha_{t-1} = 0.9$ e $\alpha_t = 0.72$, qual é, respectivamente, o coeficiente de z_t e o coeficiente de x na média?

(a) Coef. $z_t \approx 0.319$; coef. $x \approx 0.678$.

(b) Coef. $z_t \approx 0.894$; coef. $x \approx 0.200$.

(c) Coef. $z_t \approx 0.357$; coef. $x \approx 0.678$.

(d) Coef. $z_t \approx 0.849$; coef. $x \approx 0.300$.

IX) No trilema apresentado nos slides (qualidade, cobertura, velocidade de amostragem), em qual canto os modelos difusores se posicionam por padrão?

(a) Qualidade alta e cobertura alta, com amostragem lenta.

(b) Qualidade alta e amostragem rápida, com cobertura baixa.

(c) Cobertura alta e amostragem rápida, com qualidade baixa.

(d) Qualidade alta, cobertura alta e amostragem rápida simultaneamente.

X) *Latent Diffusion Models* (base do *Stable Diffusion*) aplicam a difusão no espaço latente de um VAE. Qual é a principal motivação dessa escolha?

(a) Permitir treino sem uso de *classifier-free guidance*.

(b) Reduzir o custo computacional, já que o latente tem dimensionalidade muito menor que a imagem em pixels.

(c) Tornar a *loss* convexa em relação aos parâmetros da U-Net.

(d) Eliminar a necessidade da arquitetura U-Net no estimador de ruído.

Gabarito - Objetivas

I) a

II) a

III) a

IV) c

V) b

VI) b

VII) c

VIII) a

IX) a

X) b

Resoluções - Objetivas

I) $z_t = \sqrt{0.64} \cdot 2 + \sqrt{0.36} \cdot 0.5 = 0.8 \cdot 2 + 0.6 \cdot 0.5 = 1.60 + 0.30 = 1.90$. A (b) corresponde a trocar os papéis dos coeficientes, $\sqrt{1-\alpha_t} \cdot x + \sqrt{\alpha_t} \cdot \epsilon = 0.6 \cdot 2 + 0.8 \cdot 0.5 = 1.60$. A (c) corresponde a usar α_t e $1-\alpha_t$ sem aplicar as raízes, $0.64 \cdot 2 + 0.36 \cdot 0.5 = 1.46$. A (d) corresponde a aplicar $\sqrt{\alpha_t}$ no sinal, mas esquecer de escalar o ruído, $0.8 \cdot 2 + 0.5 = 2.10$.

II) $\alpha_3 = (1-0.05)(1-0.10)(1-0.20) = 0.95 \cdot 0.90 \cdot 0.80 = 0.684$. A (b) corresponde a somar os β_t em vez de aplicar o produto $\prod(1-\beta_s)$. A (c) corresponde a $1 - \sum \beta_t$. A (d) corresponde ao produto direto $\prod \beta_t$.

III) Como $q(z_t \mid x) = \mathcal{N}[\sqrt{\alpha_t}\, x,\ (1-\alpha_t)\, \mathbf{I}]$, temos $\mu = \sqrt{0.49} \cdot 4 = 0.7 \cdot 4 = 2.80$ e $\Sigma = (1-0.49)\, \mathbf{I} = 0.51\, \mathbf{I}$. A (b) usa α_t em vez de $\sqrt{\alpha_t}$ na média. A (c) confunde a variância com o desvio padrão $\sqrt{1-\alpha_t} \approx 0.71$. A (d) usa $(1-\alpha_t)\, x = 0.51 \cdot 4 = 2.04$ na média, trocando o papel de α_t com o de $1-\alpha_t$, embora acerte a variância.

IV) Substituindo a reparametrização da rede no termo quadrático da *loss*, todos os termos em z_t e x se cancelam, restando $|g_t[z_t, \theta_t] - \epsilon_t|^2$, conforme caixa final no slide "Reparametrizando a rede, Loss final". A (a) ignora a reparametrização. A (b) corresponde ao alvo da formulação anterior. A (d) tem forma de log-verossimilhança e não corresponde a um MSE.

V) Pela definição $q(z_t \mid z_{t-1}) = \mathcal{N}[\sqrt{1-\beta_t}\, z_{t-1},\ \beta_t\, \mathbf{I}]$, vale a propriedade de Markov com variância $\beta_t\, \mathbf{I}$. A (a) viola a propriedade de Markov. A (c) descreve um processo determinístico, o que não é o caso. A (d) confunde a média da transição passo a passo (que usa $\sqrt{1-\beta_t}$) com a do kernel acumulado (que usa $\sqrt{\alpha_t}$).

VI) Calculando $\mathrm{SNR}(t) = \alpha_t/(1-\alpha_t)$: $\mathrm{SNR}(1) = 0.900/0.100 = 9.00$; $\mathrm{SNR}(2) = 0.720/0.280 \approx 2.57$; $\mathrm{SNR}(3) = 0.504/0.496 \approx 1.016$; $\mathrm{SNR}(4) = 0.302/0.698 \approx 0.433$. O cruzamento de SNR = 1 acontece em $t = 3$ (onde $\sqrt{\alpha_t} \approx \sqrt{1-\alpha_t}$). A (a) antecipa o cruzamento: em $t = 2$ o SNR ainda vale ≈ 2.57, bem acima de 1. A (c) já passou do cruzamento, com SNR ≈ 0.43. A (d) corresponde ao início, onde o sinal domina (SNR = 9).

VII) No slide “Porém… um detalhe empírico” fica explícito: a escolha de predizer ϵ é uma observação empírica de que treinar a rede para estimar o ruído produz melhores resultados que estimar x diretamente. As alternativas (a), (b) e (d) descrevem consequências falsas dessa reparametrização.

VIII) Pelo slide “Distribuição condicional, Produto de gaussianas”, a média é $\frac{(1-\alpha_{t-1})\sqrt{1-\beta_t}}{1-\alpha_t} z_t + \frac{\sqrt{\alpha_{t-1}}\beta_t}{1-\alpha_t} x$. Substituindo os valores dados, o coeficiente de z_t é $\frac{(1-0.9)\sqrt{1-0.2}}{1-0.72} = \frac{0.1 \cdot 0.8944}{0.28} \approx 0.319$, e o coeficiente de x é $\frac{\sqrt{0.9} \cdot 0.2}{0.28} = \frac{0.9487 \cdot 0.2}{0.28} \approx 0.678$. A (b) confunde o coeficiente de z_t com $\sqrt{1-\beta_t}$ e o de x com β_t puro (esquece a normalização por $1-\alpha_t$). A (c) inflaciona o coeficiente de z_t ao usar $1-\beta_t$ no lugar de $\sqrt{1-\beta_t}$. A (d) confunde com a média de $q(z_{t-1} \mid x)$ derivada do kernel direto, sem a correção bayesiana introduzida pela condição em z_t.

IX) Pelo slide “Onde modelos difusores ficam no trilema”, modelos difusores ocupam o canto qualidade + cobertura, com amostragem tipicamente lenta ($T \approx 1000$ passos pela U-Net). A (b) confunde com GANs. A (c) confunde com modelos autoregressivos rápidos de baixa fidelidade. A (d) descreveria um modelo ideal que ainda não existe.

X) No slide “*Latent Diffusion Models*”, a motivação central é o custo computacional, o latente é $\sim 64 \times 64$ contra 1024×1024 em pixels. A (a) é falsa: *Latent Diffusion* é compatível com *classifier-free guidance*. A (c) é falsa: a *loss* continua não convexa. A (d) é falsa: a U-Net continua sendo a arquitetura padrão para o estimador de ruído no espaço latente.

Questões Dissertativas.

I) Seja o *noise schedule* $\boldsymbol{\beta} = (0.1,\ 0.3)$ com $T = 2$ e $x = 1$.

(a) Calcule α_1 e α_2.

(b) Usando o kernel de difusão, escreva $q(z_2 \mid x)$ explicitamente (média e variância).

(c) Para $\epsilon = -1.0$, calcule o valor amostrado de z_2 via kernel.

(d) Compare z_2 com a média da distribuição $q(z_2 \mid x)$. Em uma frase, interprete o resultado.

II) ☠ Partindo das definições passo a passo $z_1 = \sqrt{1-\beta_1}\,x + \sqrt{\beta_1}\,\epsilon_1$ e $z_2 = \sqrt{1-\beta_2}\,z_1 + \sqrt{\beta_2}\,\epsilon_2$, com $\epsilon_1, \epsilon_2 \sim \mathcal{N}(0, \mathbf{I})$ independentes, mostre algebricamente que existe um único ruído $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ tal que $z_2 = \sqrt{\alpha_2}\,x + \sqrt{1-\alpha_2}\,\epsilon$, onde $\alpha_2 = (1-\beta_1)(1-\beta_2)$. Indique claramente onde a propriedade "soma de gaussianas independentes" é usada.

III) Compare modelos difusores, VAEs e GANs em três eixos: (i) estabilidade do treinamento, (ii) qualidade da amostra, (iii) custo de amostragem. Apresente a comparação como uma tabela ou três parágrafos curtos, justificando cada entrada com base no objetivo de treinamento de cada família (ELBO, jogo adversarial, MSE sobre o ruído).

IV) ☠ Em *classifier-free guidance*, treina-se uma única rede que pode operar em modo condicional, isto é, $g_t[z_t, c]$, ou incondicional, $g_t[z_t, \varnothing]$. Na amostragem, define-se um estimador guiado $\tilde{g}_t = (1+w)\,g_t[z_t, c] - w\,g_t[z_t, \varnothing]$, com $w \geq 0$.

(a) Explique como o treinamento incondicional é obtido a partir do mesmo modelo (sem treinar duas redes separadas).

(b) Para $w = 0$, qual amostrador é recuperado? E para $w \to \infty$ (ou valores muito grandes), qual o efeito esperado nas amostras geradas?

(c) Justifique, em uma ou duas frases, por que valores intermediários de w (e.g., $w \in [1,\ 7]$) costumam produzir amostras com maior aderência à condição, mesmo sem usar um classificador externo.

Gabarito

I) **(a)** $\alpha_1 = 1 - 0.1 = 0.9$; $\alpha_2 = (1-0.1)(1-0.3) = 0.9 \cdot 0.7 = 0.63$.

(b) Pelo kernel, $q(z_2 \mid x) = \mathcal{N}_{z_2}[\sqrt{\alpha_2}\,x,\ (1-\alpha_2)\,\mathbf{I}] = \mathcal{N}_{z_2}[\sqrt{0.63}\cdot 1,\ 0.37\,\mathbf{I}] \approx \mathcal{N}_{z_2}[0.794,\ 0.37\,\mathbf{I}]$.

(c) $z_2 = \sqrt{0.63}\cdot 1 + \sqrt{0.37}\cdot(-1.0) \approx 0.794 + 0.608\cdot(-1.0) \approx 0.794 - 0.608 \approx 0.186$.

(d) O valor amostrado $z_2 \approx 0.186$ está abaixo da média ≈ 0.794 por aproximadamente um desvio-padrão ($\sqrt{0.37} \approx 0.608$), o que reflete a contribuição de um ruído $\epsilon = -1.0$ no kernel. Em uma população grande de amostras com o mesmo x, a média empírica de z_2 convergiria para 0.794.

II) Partindo de $z_2 = \sqrt{1-\beta_2}\,z_1 + \sqrt{\beta_2}\,\epsilon_2$ e substituindo z_1:

$$z_2 = \sqrt{1-\beta_2}\left(\sqrt{1-\beta_1}\,x + \sqrt{\beta_1}\,\epsilon_1\right) + \sqrt{\beta_2}\,\epsilon_2.$$

Distribuindo,

$$z_2 = \sqrt{(1-\beta_1)(1-\beta_2)}\,x + \sqrt{(1-\beta_2)\beta_1}\,\epsilon_1 + \sqrt{\beta_2}\,\epsilon_2.$$

Como ϵ_1 e ϵ_2 são gaussianas independentes de média zero e variância 1, a soma $\sqrt{(1-\beta_2)\beta_1}\,\epsilon_1 + \sqrt{\beta_2}\,\epsilon_2$ é uma gaussiana de média zero e variância

$$(1-\beta_2)\beta_1 + \beta_2 = \beta_1 + \beta_2 - \beta_1\beta_2 = 1 - (1-\beta_1)(1-\beta_2) = 1 - \alpha_2.$$

Portanto existe um único $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ tal que $\sqrt{(1-\beta_2)\beta_1}\,\epsilon_1 + \sqrt{\beta_2}\,\epsilon_2 = \sqrt{1-\alpha_2}\,\epsilon$, e segue

$$z_2 = \sqrt{\alpha_2}\,x + \sqrt{1-\alpha_2}\,\epsilon, \qquad \alpha_2 = (1-\beta_1)(1-\beta_2). \quad \square$$

A propriedade "soma de gaussianas independentes é gaussiana com variâncias somadas" é usada exatamente no passo onde combinamos ϵ_1 e ϵ_2 em um único ϵ.

III) **Tabela comparativa:**

Eixo	Difusores	VAEs	GANs
Estabilidade do treino	alta	alta	baixa
Qualidade da amostra	muito alta	moderada	alta
Custo de amostragem	alto (T passos)	baixo (1 passo)	baixo (1 passo)

Justificativa. Difusores otimizam um MSE sobre o ruído (objetivo bem comportado, sem competição entre redes), o que confere estabilidade. VAEs maximizam o ELBO, também estável, mas o termo de KL e a fatoração gaussiana costumam suavizar as amostras (qualidade moderada, amostras tendem a borrar). GANs jogam um *min-max* adversarial, suscetível a *mode collapse* e oscilações (baixa estabilidade), mas, quando convergem, geram amostras nítidas. No custo de amostragem, difusores precisam de T passos da U-Net (tipicamente $T = 1000$), enquanto VAEs e GANs amostram com uma única passada do decoder ou gerador.

IV) **(a)** Durante o treinamento, com probabilidade p (tipicamente $p \approx 0.1$) o vetor de condicionamento c é substituído por um *token* ou *embedding* especial $\varnothing$, "descondicionando" o exemplo. A mesma rede aprende a operar tanto em modo condicional quanto incondicional, sem necessidade de treinar uma segunda rede separada.

(b) Para $w = 0$, $\tilde{g}_t = g_t[z_t, c]$, ou seja, o amostrador condicional puro. Para $w \to \infty$, o estimador guiado afasta-se cada vez mais da estimativa incondicional na direção da estimativa condicional, exagerando os traços da condição c. Na prática isso aumenta a aderência à condição (e.g., *prompt* de texto) ao preço de menor diversidade e, em valores muito altos, artefatos visuais.

(c) O estimador guiado equivale, na perspectiva de *score matching*, a amostrar de uma distribuição proporcional a $p(z_t \mid c)\,[p(z_t \mid c)/p(z_t)]^w$, que reforça a razão de verossimilhança da condição. Valores intermediários de w amplificam essa razão sem destruir o suporte da distribuição original, produzindo amostras mais alinhadas com c sem precisar de um classificador externo treinado separadamente.