Supplementary Questions

1. Why does mode collapse occur in GANs?

The original GAN objective [1] is

$$\min_G \max_D \mathbb{E}_{x\sim p_{\text{data}}}[\log D(x)] + \mathbb{E}_{z\sim p(z)}[\log(1 - D(G(z)))]. \tag{1}$$

For an optimal discriminator,

$$D^*(x) = \frac{p_{\text{data}}(x)}{p_{\text{data}}(x) + p_\theta(x)}. \tag{2}$$

Plugging back yields

$$\min_G \; 2\,\text{JS}(p_{\text{data}}\|p_\theta) - \log 4. \tag{3}$$

Thus GAN minimizes the Jensen–Shannon divergence, not forward KL.

If the supports are disjoint,

$$\text{JS}(p_{\text{data}}\|p_\theta) = \log 2, \tag{4}$$

a constant. Hence gradients vanish when p_θ misses modes of p_{data}, allowing mode collapse.

By contrast, forward KL,

$$\text{KL}(p_{\text{data}}\|p_\theta) = \mathbb{E}_{p_{\text{data}}}\left[\log \frac{p_{\text{data}}}{p_\theta}\right], \tag{5}$$

diverges whenever $p_\theta(x) \to 0$ on regions where $p_{\text{data}}(x) > 0$, strongly penalizing missing modes (mode-covering).

2. Why diffusion models avoid mode collapse

Diffusion models [2] optimize a variational bound equivalent to maximum likelihood:

$$\min_\theta \text{KL}(p_{\text{data}}\|p_\theta). \tag{6}$$

Since this corresponds to forward KL, missing probability mass incurs infinite cost. Therefore the learned distribution is forced to cover all data modes.

3. Why is Group Normalization used in U-Net DDPM?

Group Normalization (GN) [3] normalizes features per sample by dividing channels into G groups. For $x \in \mathbb{R}^{B\times C\times H\times W}$, define for each group g:

$$\mu_g = \frac{1}{m}\sum x_g, \qquad \sigma_g^2 = \frac{1}{m}\sum (x_g - \mu_g)^2, \tag{7}$$

where $m = (C/G)\cdot H \cdot W$. Then

$$\hat{x} = \frac{x - \mu_g}{\sqrt{\sigma_g^2 + \varepsilon}}, \qquad y = \gamma\hat{x} + \beta. \tag{8}$$

Unlike BatchNorm, GN does not depend on batch statistics. This is crucial in diffusion U-Nets [2] because:

- Training uses small batch sizes.
- Sampling often uses batch size 1.
- Feature statistics vary strongly across timesteps t.

BatchNorm would introduce instability due to batch-dependent normalization. GN instead normalizes each sample independently, ensuring stable activations across noise levels.

Thus GN provides:

- Batch-size invariance,
- Deterministic behavior at inference,
- Stable feature scaling across diffusion timesteps.

References

[1] Goodfellow et al. *Generative Adversarial Nets*, 2014.

[2] Ho et al. *Denoising Diffusion Probabilistic Models*, 2020.

[3] Wu & He. *Group Normalization*, 2018.