Part 4/6: Key Challenges

Evaluation Difficulties

⚠ **No single metric captures generation quality completely**

Metric	What It Measures	Quality	Diversity
Inception Score (IS)	Quality and diversity via classifier confidence	High	High
Fréchet Inception Distance (FID)	Distribution similarity between real and generated	High	High
Precision	Quality of generated samples (realism)	High	Low
Recall	Coverage of real data modes	Low	High
Human Evaluation	Subjective assessment by human raters	High	High

🏆 Gold Standard

Human evaluation remains the most reliable method despite being time-

⚠ Key Challenge

Mode coverage assessment is difficult - metrics may miss partial collapse

💻 Computational Cost

Most metrics require running large pre-trained models, making evaluation

consuming and subjective

expensive