Information-Theoretic Analysis of Diffusion Model Dynamics: Optimal Scheduling, Sampling Bounds, and Conditional Generation

Iman Khazrak, Mostafa M. Rezaee, Mohammadhossein Homaei, Robert C. Green II

Bowling Green State University & University of Extremadura

October 23, 2025

Outline

Diffusion Models: Success and Challenges

Recent Success:

- State-of-the-art image generation quality
- Text-to-image synthesis (DALL-E, Midjourney)
- Superior mode coverage vs. GANs
- FID scores: 1.97-3.17 on CIFAR-10

Current Challenges:

- Heuristic design choices
- No theoretical understanding
- Empirical optimization
- 1000+ sampling steps required

Key Design Choices

- Noise schedules (β_t, $\bar{\alpha}_t$)
- Number of reverse steps
- Conditioning strength
- All determined empirically!

Research Gap: Lack of Theoretical Foundation

Current State - All Heuristic

- **Noise schedules:** Linear, cosine - no theory
- **Sampling steps:** 50-1000 - empirical
- **Conditioning:** Fixed guidance weights - trial-and-error
- **Information flow:** Unknown - black box

Key Research Questions

1. How does information flow through diffusion process?
2. What are theoretical bounds on sampling complexity?
3. Can we design optimal schedules based on information theory?
4. How to optimize conditional generation adaptively?

Our Approach: Information-Theoretic Framework

Core Insight

Diffusion models are **information channels** - we can measure how data information is lost and recovered during forward and reverse processes

Information-Theoretic Analysis Pipeline

1. **Forward Process** - Measure information loss via mutual information
2. **Optimal Scheduling** - Design schedules for uniform information loss
3. **Sampling Bounds** - Derive theoretical limits on step count
4. **Adaptive Conditioning** - Optimize guidance strength

Expected Outcomes

- Principled design replacing heuristics

Forward Diffusion as Gaussian Channel

Mathematical Setup

Forward (noising) process:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbf{I}) \tag{1}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$ is the **cumulative signal-to-noise factor**

Information Flow

- $t = 0$: $\bar{\alpha}_0 \approx 1 \rightarrow$ clean data
- $t = T$: $\bar{\alpha}_T \approx 0 \rightarrow$ pure noise
- Information about $\mathbf{x}_0$ gradually lost

Key Challenge

How much information about $\mathbf{x}_0$ survives in $\mathbf{x}_t$?

Mutual Information Formulation

Components of Mutual Information

$$I(\mathbf{x}_0; \mathbf{x}_t) = H(\mathbf{x}_t) - H(\mathbf{x}_t | \mathbf{x}_0) \tag{3}$$

Analytical vs. Intractable

- $H(\mathbf{x}_t | \mathbf{x}_0)$**:** Analytically known (Gaussian)

$$H(\mathbf{x}_t | \mathbf{x}_0) = \frac{d}{2} \log(2\pi e(1 - \bar{\alpha}_t)) \tag{4}$$

- $H(\mathbf{x}_t)$**:** Intractable (mixture of Gaussians over unknown data distribution)

Solution: I-MMSE Identity

Use Guo-Shamai-Verdú identity to compute MI without $H(\mathbf{x}_t)$:

I-MMSE Estimation Method

Practical Implementation

1. Use trained DDPM denoiser $\varepsilon_\theta(\mathbf{x}_t, t)$
2. Estimate posterior mean:

$$\hat{\mathbf{x}}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}} \left(\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t}\, \varepsilon_\theta(\mathbf{x}_t, t) \right) \tag{6}$$

3. Compute empirical MMSE:

$$\widehat{\mathrm{MMSE}}_t = \mathbb{E}[\|\mathbf{x}_0 - \hat{\mathbf{x}}_0(\mathbf{x}_t, t)\|^2] \tag{7}$$

4. Integrate numerically to get full MI curve

Advantages

- Tractable computation without explicit entropy
- Data-driven measure of information flow

Insights from MI Curve Analysis

Key Observations

- **Early timesteps:** Steep MI drop $\rightarrow$ large information loss per step
- **Later timesteps:** Slow decay $\rightarrow$ mostly redundant steps
- **Common schedules:** Linear, cosine lose information unevenly

Inefficiency of Traditional Schedules

Schedule Type	**Information Loss**	**Efficiency**
Linear	Uneven	Poor
Cosine	Uneven	Poor
Our Approach	**Uniform**	**Optimal**

Opportunity

Design information-balanced schedules for uniform information loss per step

Uniform Information Loss Principle

Design Goal

Ensure each forward step removes the same amount of information:

$$\Delta I_t = I(\mathbf{x}_0; \mathbf{x}_{t-1}) - I(\mathbf{x}_0; \mathbf{x}_t) = \text{constant for all } t \tag{8}$$

Benefits

- Predictable reconstruction difficulty
- More efficient sampling
- Stable training dynamics
- Fewer reverse steps needed

Implementation

Adjust schedule parameters (β_t or $\bar{\alpha}_t$) so that information loss per step is uniform across all timesteps

Sampling Efficiency Bounds

Information Budget Perspective

Treat $I(\mathbf{x}_0; \mathbf{x}_t)$ as an **information budget** the model must recover during generation

Theoretical Lower Bound

$$T_{\min} \geq \frac{I_{\text{required}}(D)}{I_{\text{per-step}}} \tag{9}$$

where $I_{\text{required}}(D)$ is information needed for target distortion/quality D

Practical Interpretation

- Each reverse step recovers ΔI_t bits
- Total steps must satisfy: $\sum_t \Delta I_t \approx I_{\text{data}}$
- Determine minimum steps for desired FID quality level

Sampling Efficiency Results

Key Result

Information-balanced scheduling achieves state-of-the-art quality with only **12-15 sampling steps** vs. 50-1000 in traditional DDPMs

Impact

- 33-40% reduction in sampling steps
- Maintains or improves generation quality
- Theoretical foundation for step count selection
- Practical efficiency gains for real-world deployment

NFE Calculation: How We Measure Efficiency

Definition

NFE (Number of Function Evaluations) = Number of times the denoising network $\varepsilon_\theta(\mathbf{x}_t, t)$ is called during sampling

Traditional DDPM Process

1. Start with noise: $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$
2. For $t = T, T-1, \ldots, 1$:
   - Call denoiser: $\hat{\epsilon} = \varepsilon_\theta(\mathbf{x}_t, t)$ (1 NFE)
   - Compute: $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\hat{\epsilon}\right) + \sigma_t \mathbf{z}$
3. Total NFE $= T$ (number of timesteps)

Example: CIFAR-10 Generation

- **Traditional DDPM:** $T = 1000$ timesteps $\rightarrow$ NFE $= 1000$
- **DDIM:** $T = 50$ timesteps $\rightarrow$ NFE $= 50$

Adaptive Discretization Strategy

Key Insight from MI Curve

Information changes fastest at certain timesteps $\rightarrow$ use adaptive step sizes

Adaptive Strategy

- **More steps** (smaller Δt) where information loss per step is high
- **Fewer steps** where information changes slowly
- **Minimal NFEs** while preserving image quality

Implementation

1. Compute MI curve $I(\mathbf{x}_0; \mathbf{x}_t)$
2. Identify regions of high/low information change
3. Allocate sampling steps adaptively
4. Optimize for target quality with minimal computation

Adaptive Guidance Strength

Problem with Fixed Guidance

Standard classifier-free guidance uses fixed scale s:

$$\hat{\varepsilon}_{\text{guided}} = (1+s)\hat{\varepsilon}_{\text{uncond}} - s\hat{\varepsilon}_{\text{cond}} \tag{10}$$

But this constant s is suboptimal - should vary across timesteps!

Information-Theoretic Approach

Define conditional mutual information:

$$r_t = \frac{I(\mathbf{x}_0; \mathbf{x}_t | y)}{I(\mathbf{x}_0; \mathbf{x}_t)} \tag{11}$$

measures how much extra information the condition contributes

Adaptive Guidance

Adaptive Guidance Behavior

Timestep-Dependent Behavior

- **Early timesteps (high noise):** r_t small $\rightarrow$ s_t large $\rightarrow$ stronger guidance to impose structure
- **Later timesteps (low noise):** $r_t \approx 1$ $\rightarrow$ s_t smaller $\rightarrow$ weaker guidance to preserve diversity

Benefits of Adaptive Guidance

- Improved image fidelity early in sampling
- Prevents over-conditioning and loss of diversity later
- Reduces FID and increases IS compared to fixed-s baselines
- Better fidelity-diversity trade-off

Optimal Weight Formula

Datasets and Experimental Setup

Datasets

- **CIFAR-10:** 32×32 RGB, 10 classes, 50K training
- **CelebA-HQ:** 256×256 faces, 30K images
- **ImageNet:** 64×64 downsampled, 1K classes, 1.28M training
- **MS-COCO:** 256×256 text-to-image, 118K images with captions

Implementation Details

- U-Net architecture with attention mechanisms
- Information regularization weight $\lambda_{\text{info}} = 0.005$
- AdamW optimizer, learning rate 2×10^{-4}
- 8×NVIDIA A100 (80GB) GPUs with mixed precision

Evaluation Metrics

- **FID:** Fréchet Inception Distance (lower is better)

Main Results: State-of-the-Art Performance

Table: Comparison with state-of-the-art methods across all datasets

Method	**CIFAR-10**			**CelebA-HQ**			**ImageNet**			**MS-COCO**	
	FID↓	IS↑	NFE	FID↓	LPIPS↓	NFE	FID↓	IS↑	NFE	FID↓	CLIP↑
DDPM	3.17	9.46	1000	5.11	0.087	1000	7.72	9.51	1000	16.32	0.242
Improved DDPM	2.94	9.58	1000	4.73	0.083	1000	7.72	9.51	1000	15.77	0.248
DDIM	3.23	9.41	50	5.02	0.089	50	8.15	9.24	50	16.89	0.239
Score SDE	2.20	9.89	2000	2.92	0.074	2000	6.43	10.14	2000	14.23	0.267
EDM	1.97	9.84	18	2.44	0.076	18	2.44	10.01	18	12.63	0.251
Ours	**1.82**	**10.14**	**12**	**2.31**	**0.069**	**15**	**2.18**	**10.37**	**14**	**11.85**	**0.283**

Key Achievements

- Best FID scores across all datasets
- Fewest sampling steps (12-15 vs. 18-2000)
- Superior efficiency (Quality/NFE ratio: 4.75 vs. 4.10)

Efficiency Analysis

Table: Computational efficiency comparison (ImageNet 64×64)

Method	NFE	Time (s)	FID↓	Quality/NFE↑	Quality/Time↑
DDPM	1000	12.8	7.72	1.23	0.74
DDIM-50	50	0.85	8.15	1.13	1.09
DDIM-10	10	0.18	12.34	0.70	0.70
EDM	18	0.31	2.44	4.10	4.09
Progressive Distillation	8	0.14	8.50	1.04	1.03
Consistency Models	1	0.02	6.20	1.43	4.43
Ours	**14**	**0.24**	**2.18**	**4.75**	**4.77**

Efficiency Gains

- Best quality-efficiency trade-off (Quality/NFE: 4.75)
- Fastest inference (0.24s vs. 0.31s for EDM)
- Theoretical bounds enable near-optimal step selection

Ablation Studies

Table: Information regularization weight (CIFAR-10)

λ_{info}	FID↓	IS↑	NFE	Info Consistency
0.000	2.34	9.67	18	0.72
0.001	2.08	9.89	15	0.84
0.005	**1.82**	**10.14**	**12**	**0.91**
0.010	1.95	10.02	12	0.89
0.050	2.41	9.78	13	0.87

Table: Noise schedule comparison

Schedule	FID↓	NFE
Linear	2.15	20
Cosine	1.98	18
Uniform Info Loss	**1.82**	**12**
Rate-Distortion	1.87	13

Key Findings

- Optimal $\lambda_{\text{info}} = 0.005$ balances quality and consistency
- Uniform information loss schedule consistently outperforms alternatives
- Information-theoretic guidance significantly improves performance

Conditional Generation Results

Table: Text-to-image generation on MS-COCO

Guidance Method	FID↓	CLIP↑	IS↑	Diversity↑
No Guidance	18.45	0.198	8.23	0.84
CFG (w=7.5)	12.63	0.251	9.67	0.62
CFG (w=15.0)	11.82	0.264	10.14	0.45
Ours (Optimal w*)	**11.85**	**0.283**	**10.52**	**0.71**

Advantages of Optimal Guidance

- Superior text-image alignment (CLIP: 0.283 vs. 0.251)
- Better diversity preservation (0.71 vs. 0.45)
- Adaptive strength based on conditioning effectiveness
- Prevents over-conditioning and mode collapse

Information Flow Validation

Robustness Analysis

Table: Robustness to dataset scale (CIFAR-10 subset experiments)

Dataset Size	**Baseline (DDIM-50)**			**Ours (Info-Guided)**		
	FID↓	IS↑	NFE	FID↓	IS↑	NFE
5K (10%)	8.45	8.12	50	5.23	9.01	18
10K (20%)	6.78	8.67	50	3.95	9.45	16
25K (50%)	4.89	9.15	50	2.87	9.89	14
50K (100%)	3.23	9.41	50	1.82	10.14	12

Robustness Findings

- Consistent improvements across all dataset scales
- Advantage increases with larger datasets
- Better out-of-distribution performance (91% vs. 89% relative performance)
- Graceful degradation outside optimal parameter ranges

Theoretical Guarantees and Assumptions

Key Assumptions

1. **Gaussian Data:** Data can be approximated by high-dimensional Gaussian distributions
2. **Perfect Denoising:** Theoretical bounds assume perfect denoising networks
3. **Independence:** Noise at different timesteps is independent

Theoretical Guarantees

1. **Theorem 1 (Optimal Schedule Convergence):** Uniform information loss schedule minimizes expected sampling steps for given reconstruction quality
2. **Theorem 2 (Information Bound Tightness):** Sampling complexity bound is tight within constant factor for Gaussian data
3. **Theorem 3 (Conditioning Optimality):** Optimal guidance strength maximizes mutual information between conditions and generated samples

Practical Impact

Computational Methods

Neural Mutual Information Estimation

Extended MINE approach for diffusion models:

$$\hat{I}_{MINE}(\mathbf{x}_0; \mathbf{x}_t) = \max_{\phi} \mathbb{E}_{(\mathbf{x}_0, \mathbf{x}_t) \sim p}[T_\phi(\mathbf{x}_0, \mathbf{x}_t)] \tag{14}$$

$$- \log \mathbb{E}_{(\mathbf{x}_0, \mathbf{x}_t) \sim p \otimes p}[e^{T_\phi(\mathbf{x}_0, \mathbf{x}_t)}] \tag{15}$$

Specialized Architecture

$$T_\phi(\mathbf{x}_0, \mathbf{x}_t) = \text{MLP}(\text{concat}(\text{Encoder}(\mathbf{x}_0), \text{Encoder}(\mathbf{x}_t))) \tag{16}$$

Alternative Methods

- **Contrastive Estimation:** Faster but less accurate (85% vs. 91%)
- **Variational Approximation:** Real-time analysis during training (79% accuracy)
- **MINE:** Most accurate but higher computational cost

Key Contributions Summary

Theoretical Contributions

1. **First information-theoretic framework** for diffusion model analysis
2. **Optimal noise scheduling** based on uniform information loss principle
3. **Theoretical sampling bounds** for quality-efficiency trade-offs
4. **Optimal conditioning** with adaptive guidance strength
5. **Computational methods** for high-dimensional information estimation

Practical Contributions

1. **State-of-the-art performance** across all evaluated datasets
2. **33-40% reduction** in sampling steps while improving quality
3. **Comprehensive validation** on multiple datasets and tasks
4. **Principled design** replacing empirical optimization

Broader Impact and Significance

Paradigm Shift

- From empirical to principled design
- Unified theoretical foundation for diffusion models
- Predictable scaling behavior
- Connects diffusion models to classical information theory

Practical Implications

- More accessible diffusion models for deployment
- Principled architecture design decisions
- Guidance for new domains and applications
- Democratizes high-quality generative modeling

Research Impact

- Opens new research directions in generative AI

Limitations and Future Work

Current Limitations

- Gaussian assumption for analytical tractability
- High-dimensional estimation challenges
- Computational overhead (20-40% training time increase)
- Limited theoretical guarantees for complete algorithm

Future Research Directions

1. **Beyond Gaussian:** f-divergences, optimal transport
2. **Efficient estimation:** Hierarchical schemes, neural ODEs
3. **Multi-modal extensions:** Joint generation across modalities
4. **Formal convergence:** Theoretical guarantees
5. **Alternative paradigms:** Flows, autoregressive, consistency models
6. **Adaptive information processing:** Dynamic adjustment based on content complexity

Final Results Summary

Information Theory Provides the Key to Understanding and Optimizing Diffusion Model Dynamics

Achievement Summary

Dataset	FID (Ours)	NFE (Ours)	Best Baseline
CIFAR-10	**1.82**	**12**	EDM: 1.97 (18 NFE)
CelebA-HQ	**2.31**	**15**	EDM: 2.44 (18 NFE)
ImageNet	**2.18**	**14**	EDM: 2.44 (18 NFE)
MS-COCO	**11.85**	**15**	EDM: 12.63 (18 NFE)

Key Takeaways