MMG: Mutual Information Estimation via the MMSE Gap in Diffusion

Longxuan Yu[*1] **Xing Shi**[*1] **Xianghao Kong**[1] **Tong Jia**[2] **Greg Ver Steeg**[1]

[1]Computer Science Dept., University of California, Riverside, Riverside, California, USA
[2]Computer Science Dept., University of California, San Diego, La Jolla, California, USA

Abstract

Mutual information is one of the most general ways to measure relationships between random variables, but estimating this quantity for complex systems is challenging. Denoising diffusion models have recently set a new bar for density estimation, so it is natural to consider whether these methods could also be used to improve mutual information estimation. Using the recently introduced information-theoretic formulation of denoising diffusion models, we show that diffusion models can be used in a straightforward way to estimate mutual information. In particular, the mutual information corresponds to half the gap in the minimum mean square error between conditional and unconditional diffusion, integrated over all signal-to-noise ratios in the noising process. Our approach not only passes self-consistency tests but also outperforms traditional and score-based diffusion estimators. Furthermore, our method leverages adaptive importance sampling to achieve scalable estimation, while maintaining strong performance even when the mutual information is high. Code to reproduce our experiments is available at `https://github.com/fengsxy/Diffusion_MI`, and the unified mutual information estimation library is available at `https://github.com/fengsxy/Diffusion-MI`.

1 INTRODUCTION

Estimating Mutual Information (MI) from samples is a fundamental problem with widespread applications, including representation learning [van den Oord et al., 2018, Hjelm et al., 2019], domain adaptation [Gholami et al., 2020], the information bottleneck principle [Tishby et al., 2000, Alemi et al., 2016], and scientific discovery and data analysis [Kinney and Atwal, 2014, Kraskov et al., 2004]. Methods based on local density estimation [Kraskov et al., 2004, Pál et al., 2010, Gao et al., 2015] have been displaced by variational approaches using neural networks [Poole et al., 2019] to estimate lower bounds on MI [Belghazi et al., 2018, Nguyen et al., 2010]. Unfortunately, these approaches may have sample complexity or variance which scale exponentially with the true MI [Gao et al., 2015, McAllester and Stratos, 2020, van den Oord et al., 2018]. In practical scenarios, the MI between two variables is usually unknown, making reliable estimation challenging. To tackle this, Song and Ermon [2019] introduced three self-consistency experiments designed to evaluate the robustness of MI estimators. For a more standardized and comprehensive evaluation, Czyż et al. [2023] developed a benchmark consisting of 40 synthetic datasets derived from diverse distributions, providing a consistent basis for assessing the performance of different MI estimators.

Denoising diffusion models [Sohl-Dickstein et al., 2015, Ho et al., 2020, Kingma et al., 2021, Kong et al., 2023] have dramatically improved the modeling of complex distributions, igniting a new industry for AI art generation [Rombach et al., 2022, Ramesh et al., 2022, Saharia et al., 2022]. Recently, Franzese et al. [2024] introduced an MI estimator, MINDE, based on score-based diffusion models. This estimator not only achieved an 87.5% estimation success rate on the benchmark of Czyż et al. [2023] but also passed the self-consistency tests. However, its reliance on accurately approximating the log-density gradient can be a challenging intermediate step. Although the score and the optimal denoiser are related by an affine transformation at the population optimum via Tweedie's formula, this equivalence does not imply that finite-capacity neural networks approximate the two training targets equally well: the conditional-mean target remains bounded within the data range, whereas the score target is amplified at low noise levels and in concentrated, high-MI distributions, making it harder to fit due to spectral bias [Rahaman et al., 2019] (see Appendix C). This

[*]Equal contribution.

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:7913–7927.

motivates a more direct formulation, connecting MI to the denoising objective itself rather than its gradient.

In this paper, we exploit the recently discovered connections between information theory and diffusion models to derive an elegant and effective approach to MI estimation [Guo et al., 2005, Kong et al., 2023]. Our contributions are as follows:

- We show that conditional density estimation and mutual information estimation can both be written exactly in terms of the global optimum of a denoising objective. Conditional density estimation corresponds to a gap between MMSEs [Kong et al., 2023] for conditional and unconditional denoising diffusion models, and mutual information is the expected gap over all data, leading to the **M**utual Information Estimation via the **M**MSE **G**ap in Diffusion (MMG) estimator.
- We develop an adaptive importance sampling scheme that tailors the integration to the specific data distribution. By dynamically fitting a sampling distribution to the MMSE gap, this technique significantly improves the precision and efficiency of the final MI estimate.
- MMG has passed all self-consistency tests and achieved state-of-the-art results across multiple tasks, particularly excelling in high MI estimation, and has outperformed the current leading estimator, MINDE, in these scenarios.
- We release a unified PyTorch library that, for the first time, brings diffusion-based and established neural MI estimators into a single, consistent framework. Its simple API is designed to streamline their future side-by-side evaluation.

2 BACKGROUND

Let $p(\boldsymbol{z}_\gamma|\boldsymbol{x})$ be a Gaussian noise channel with $\boldsymbol{z}_\gamma = \sqrt{\gamma/(1+\gamma)}\boldsymbol{x} + \sqrt{1/(1+\gamma)}\boldsymbol{\epsilon}$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbb{I})$, where γ represents the Signal-to-Noise Ratio (SNR). The unknown data distribution is $p(\boldsymbol{x})$. The MMSE refers to the minimum mean square error for recovering $\boldsymbol{x}$ in this noisy channel [Kong et al., 2023].

$$\mathrm{mmse}_x(\gamma) \equiv \min_{\hat{\boldsymbol{x}}(\boldsymbol{z}_\gamma,\gamma)} \mathbb{E}_{p(\boldsymbol{z}_\gamma,\boldsymbol{x})}[(\boldsymbol{x} - \hat{\boldsymbol{x}}(\boldsymbol{z}_\gamma,\gamma))^2] \quad (1)$$

The optimal estimator, $\hat{\boldsymbol{x}}^*$, can be derived via variational calculus and written analytically.

$$\begin{aligned}\hat{\boldsymbol{x}}^*(\boldsymbol{z}_\gamma,\gamma) &\equiv \arg\min_{\hat{\boldsymbol{x}}(\boldsymbol{z}_\gamma,\gamma)} \mathbb{E}_{p(\boldsymbol{z}_\gamma,\boldsymbol{x})}\left[(\boldsymbol{x} - \hat{\boldsymbol{x}}(\boldsymbol{z}_\gamma,\gamma))^2\right] \\ &= \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x}|\boldsymbol{z}_\gamma)}[\boldsymbol{x}]\end{aligned} \quad (2)$$

Sampling from the true posterior is typically intractable, but by using a neural network to solve Eq. 1 we can get an approximation for $\hat{\boldsymbol{x}}^*$. We also introduce pointwise MMSE which is just the MMSE evaluated at a single point $\boldsymbol{x}$, and $\mathbb{E}_{p(\boldsymbol{x})}[\mathrm{mmse}(\boldsymbol{x}|\gamma)] = \mathrm{mmse}_x(\gamma)$.

$$\mathrm{mmse}(\boldsymbol{x}|\gamma) \equiv \mathbb{E}_{p(\boldsymbol{z}_\gamma|\boldsymbol{x})}[(\boldsymbol{x} - \hat{\boldsymbol{x}}^*(\boldsymbol{z}_\gamma,\gamma))^2] \quad (3)$$

From Kong et al. [2023] we see that log likelihood can be written *exactly* in terms of an expression that depends only on the MMSE solution to the Gaussian denoising problem, where d denotes the dimensionality of $\boldsymbol{x}$. This information-estimation connection has also been leveraged to construct perceptual distance metrics [Ohayon et al., 2025].

$$\boxed{\begin{aligned}-\log p(\boldsymbol{x}) = &\frac{d}{2}\log(2\pi e) \\ &- \frac{1}{2}\int_0^\infty \left(\frac{d}{1+\gamma} - \mathrm{mmse}(\boldsymbol{x} \mid \gamma)\right) d\gamma\end{aligned}} \quad (4)$$

We can use this result to derive elegant formulations for supervised learning and mutual information estimation.

3 METHOD

3.1 MUTUAL INFORMATION ESTIMATION

Consider that our data is drawn from some unknown joint distribution, $p(\boldsymbol{x}, y)$. At this point we don't specify the domain of y, it could be discrete or continuous, vector or scalar. The pointwise denoising relation, Eq. 4, holds for *any* input distribution, so a valid choice would be $p(\boldsymbol{x}|y)$. Therefore we can write a conditional version that holds for each y.

$$\begin{aligned}-\log p(\boldsymbol{x} \mid y) = &\frac{d}{2}\log(2\pi e) \\ &- \frac{1}{2}\int_0^\infty d\gamma \left(\frac{d}{1+\gamma} - \mathrm{mmse}(\boldsymbol{x} \mid \gamma, y)\right)\end{aligned} \quad (5)$$

We use the following definitions for conditional MMSE.

$$\begin{aligned}\hat{\boldsymbol{x}}^*(\boldsymbol{z}_\gamma,\gamma,y) &\triangleq \arg\min_{\hat{\boldsymbol{x}}} \mathbb{E}_{p(\boldsymbol{x},y,\boldsymbol{z}_\gamma)}\left[\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2\right], \\ \mathrm{mmse}(\boldsymbol{x} \mid \gamma, y) &\triangleq \mathbb{E}_{p(\boldsymbol{z}_\gamma|\boldsymbol{x})}\left[\|\boldsymbol{x} - \hat{\boldsymbol{x}}^*(\boldsymbol{z}_\gamma,\gamma,y)\|^2\right].\end{aligned} \quad (6)$$

We write the expected conditional MMSE as $\mathrm{mmse}_{x|y}(\gamma) = \mathbb{E}_{p(\boldsymbol{x},y)}[\mathrm{mmse}(\boldsymbol{x}|\gamma,y)]$.

Now we can subtract Eq. 4 and Eq. 5 to get the following.

$$\begin{aligned}\log p(\boldsymbol{x} \mid y) - \log p(\boldsymbol{x}) &= \frac{1}{2}\int_0^\infty \Delta(\boldsymbol{x},\gamma,y)\, d\gamma, \\ \Delta(\boldsymbol{x},\gamma,y) &\coloneqq \mathrm{mmse}(\boldsymbol{x} \mid \gamma) - \mathrm{mmse}(\boldsymbol{x} \mid \gamma, y).\end{aligned} \quad (7)$$

The expression on the left is sometimes called *pointwise mutual information*, because its expectation, $\mathbb{E}[\log p(\boldsymbol{x}|y)/p(\boldsymbol{x})] = I(\boldsymbol{x};y)$ is equal to the mutual information.

$$\boxed{I(\boldsymbol{x};y) = {}^1\!/_2 \int_0^\infty d\gamma \left(\mathrm{mmse}_x(\gamma) - \mathrm{mmse}_{x|y}(\gamma)\right)} \quad (8)$$

Conditioning on y can only decrease the MMSE [Wu and Verdú, 2011], so the mutual information is non-negative as expected. While Eq. 7 appears to be novel, a version of Eq. 8 appeared in Guo et al. [2005]. This result holds for discrete or continuous data [Guo et al., 2005], or even mixed continuous and discrete, a particularly challenging problem [Gao et al., 2017]. We propose to use recent advances in denoising diffusion modeling to approximate the right-hand side to achieve better mutual information estimates.

To efficiently estimate the integral in Eq. 8, we applied importance sampling in the integral on the right-hand side of Eq. 8 with the importance weights $q(\gamma)$. Note that the expectations over the data are already contained in the expected MMSE terms, so the only remaining expectation is over the SNR.

$$\boxed{I(\boldsymbol{x};y) = {}^1\!/_2 \mathbb{E}_{\gamma\sim q}\left[\left(\mathrm{mmse}_x(\gamma) - \mathrm{mmse}_{x|y}(\gamma)\right)/q(\gamma)\right]} \quad (9)$$

Therefore, we can train expert denoisers to estimate the MMSE for various distributions of $q(\gamma)$, as different densities consistently lead to distinct distributions of the MMSE gap in Figure 1.

3.2 MODEL TRAINING

In practice, we parametrize the denoising function in terms of a neural network. At training time, we have to solve two minimization problems, or actually a continuum of minimization problems for each SNR level.

$$\theta^* = \arg\min_\theta \mathbb{E}_{p(\boldsymbol{z}_\gamma|\boldsymbol{x})p(\boldsymbol{x},y)}[(\boldsymbol{x} - \hat{\boldsymbol{x}}_\theta(\boldsymbol{z}_\gamma,\gamma,y))^2] \quad (10)$$
$$\phi^* = \arg\min_\phi \mathbb{E}_{p(\boldsymbol{z}_\gamma|\boldsymbol{x})p(\boldsymbol{x})}[(\boldsymbol{x} - \hat{\boldsymbol{x}}_\phi(\boldsymbol{z}_\gamma,\gamma))^2]$$

We parametrize $\hat{\boldsymbol{x}}_\theta(\boldsymbol{z}_\gamma,\gamma,y)$ as some neural network that is trained with a mean square error loss to recover the signal, $\boldsymbol{x}$, from noise and the auxiliary signal, y, across different SNRs, γ (see Figure 2). In principle we should train a separate neural network that recovers data from noise without conditioning on y. However, we can borrow from existing architectures which train a single network for both conditional and unconditional denoising [Ho and Salimans, 2021], so MMG does not require training two separate networks. Conditional diffusion models have shown compelling results in Nichol et al. [2021], Ramesh et al. [2022], Saharia et al. [2022]. During training we simply replace y with a null value, $y = \emptyset$, some fraction of the time so that the same (conditional) network can learn to denoise in the absence of conditioning. We set this fraction to 50%. This differs from the 10–15% dropout typically used for classifier-free guidance [Ho and Salimans, 2021], where the unconditional branch plays only an auxiliary role. In MMG, by contrast, the conditional and unconditional denoising errors enter the MI estimate symmetrically (Eq. 8), so we allocate comparable training signal to both branches; 50% is the simplest balanced choice, and nearby ratios perform similarly.

In practice, we can't guarantee that the neural network finds the true global optimum, the MMSEs in Eq. 8. Because MMSEs appear with both signs in that expression, we can't even guarantee that our network gives an upper or lower bound. While this is unfortunate, conventional wisdom suggests that sufficiently expressive neural networks do converge to global optima [Du et al., 2019, Jacot et al., 2018].

4 IMPLEMENTATION

We emphasize that, although we build on the machinery of denoising diffusion models, MMG only uses the *denoising diffusion objective*: it requires accurate denoisers (equivalently, MMSE estimates) across all SNRs, and never runs a generative sampler. Unlike generative diffusion models, which weight the MSE loss across SNRs because low-SNR accuracy matters little for sampling, our estimator relies on the unweighted MMSE at every noise level. The core of the experiment is therefore training a denoiser capable of effectively performing denoising across different noise levels, characterized by varying SNRs, to simulate an MMSE curve. The MI is then computed from the gap between the conditional and unconditional MMSE curves (Eq. 8). This MMSE-based theoretical approach offers greater flexibility to enhance the accuracy and robustness of MI estimation. To leverage this flexibility for enhanced accuracy and robustness, we introduce the core components of our implementation below.

4.1 ADAPTIVE IMPORTANCE SAMPLING

As in Kong et al. [2023], we estimate integrals using importance weighted Monte Carlo estimators, with log-SNR values chosen from a logistic distribution with some location, μ and scale, σ. Since the shape of the integration area (Δ area in Figure 1) is data-dependent, a fixed sampling distribution can be inefficient. For our adaptive variants, we therefore optimize this distribution for each specific task.

To find the data-adapted parameters (μ, σ), we employ a two-stage procedure. First, we train a preliminary model and analyze its conditional MMSE curve, $\mathrm{MMSE}_{x|y}$. Inspired by the use of error landmarks in density estimation [Kong et al., 2023], our heuristic identifies the critical transition region of this curve, where the denoiser becomes effective.

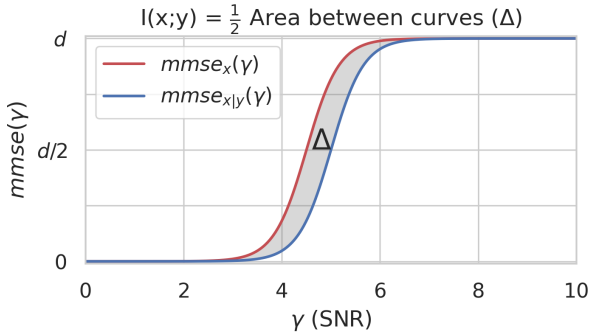


Figure 1: The mutual information is exactly half the area between MMSE curves for conditional and unconditional denoising. We use denoising diffusion models to approximate the MMSE curves, then numerically integrate to get an estimate of the mutual information.

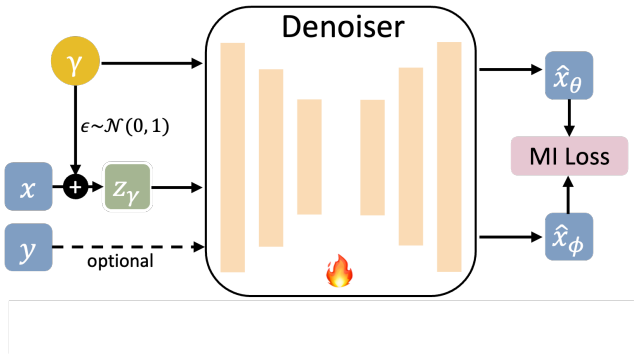


Figure 2: The schematic of the MMG training process. Noise at a given γ level is added to the data $\boldsymbol{x}$, and a denoiser is used to recover it, with or without conditioning on y at a 50% probability. Finally, the MI loss, as defined in Eq. 10, is computed to backpropagate the gradient.

Specifically, we define the parameters as follows (recall that d is the dimensionality of $\boldsymbol{x}$, defined in Section 2):

- The location μ is set to the log-SNR where the MMSE curve crosses the $d/2$ error threshold. This centers our sampling distribution on the midpoint of the denoiser's transition from high to low error.
- The scale σ is derived from the log-SNR where the curve crosses the $d/4$ threshold, capturing the steepness of this transition. For example, it can be set as the difference between the log-SNRs at the $d/2$ and $d/4$ crossings.

The intuition behind these thresholds is that the conditional MMSE curve decreases from roughly d at pure noise (the variance of standardized data) to 0 at high SNR. The $d/2$ crossing is therefore a natural proxy for the midpoint of the transition region, where the MMSE gap tends to have most of its mass, so we center the proposal there; the $d/4$ crossing provides a coarse estimate of the transition width. These constants are not meant to be special: the ablation in Table 2 shows that this data-adaptive calibration reduces integration variance relative to a fixed SNR sampling distribution, and the estimator is not sensitive to the precise threshold values.

The final adaptive estimators are then trained with this optimized distribution, targeting the critical SNR range to yield more accurate and efficient MI estimates.

4.2 THE ORTHOGONAL PRINCIPLE

To further enhance estimator stability, we incorporate the orthogonality principle from Kong et al. [2024]. The principle states that for optimal denoisers, the gap in MMSE is precisely the expected squared distance between the conditional and unconditional estimators. This provides an alternative way to express the MI integrand. Formally, let $\hat{\boldsymbol{x}}(\boldsymbol{z}_\gamma) = \mathbb{E}[\boldsymbol{x}|\boldsymbol{z}_\gamma]$ and $\hat{\boldsymbol{x}}(\boldsymbol{z}_\gamma, y) = \mathbb{E}[\boldsymbol{x}|\boldsymbol{z}_\gamma, y]$. The identity at a fixed SNR is:

$$\begin{aligned}\Delta_{\text{MMSE}} &:= \mathbb{E}\left[\|\boldsymbol{x} - \hat{\boldsymbol{x}}(\boldsymbol{z}_\gamma)\|^2\right] - \mathbb{E}\left[\|\boldsymbol{x} - \hat{\boldsymbol{x}}(\boldsymbol{z}_\gamma, y)\|^2\right] \\ &= \mathbb{E}\left[\|\hat{\boldsymbol{x}}(\boldsymbol{z}_\gamma, y) - \hat{\boldsymbol{x}}(\boldsymbol{z}_\gamma)\|^2\right].\end{aligned} \tag{11}$$

After expanding the left-hand side, the cross terms vanish due to the orthogonality principle [Kong et al., 2024], resulting in the expression on the right-hand side. This allows us to substitute the original integrand in our MI formula (Eq. 8) with the term on the right-hand side. The orthogonal MI estimator, in its practical importance-sampled form, is therefore:

$$I(\boldsymbol{x}; y) = \frac{1}{2}\mathbb{E}_{\boldsymbol{x},y,\boldsymbol{z}_\gamma,\gamma\sim q}\left[\frac{\|\hat{\boldsymbol{x}}(\boldsymbol{z}_\gamma, y) - \hat{\boldsymbol{x}}(\boldsymbol{z}_\gamma)\|^2}{q(\gamma)}\right] \tag{12}$$

In practice, this is implemented as a *training-free, inference-time plug-in*. This formulation guarantees a non-negative integrand and improves the stability of MI estimates, as detailed in Appendix A.

The practical advantage of this formulation is its numerical stability. The standard MMSE gap is computed as a difference between two large, separately estimated MSE values. Small approximation errors in each term can lead to a noisy, high-variance result for their difference, which can even become negative. In contrast, the orthogonal form computes the integrand as a single, squared term, which is guaranteed to be non-negative and is empirically much smoother.

5 EXPERIMENTS

We evaluate four variants of our estimator in the experiments for ablation comparison:

	Asinh @ *St* 1 × 1 (dof=1)	Asinh @ *St* 2 × 2 (dof=1)	Asinh @ *St* 3 × 3 (dof=2)	Asinh @ *St* 5 × 5 (dof=2)	Bimodal 1 × 1	Bivariate *Nm* 1 × 1	*Hc* @ Bivariate *Nm* 1 × 1	*Hc* @ *Mn* 25 × 25 (2-pair)	*Hc* @ *Mn* 3 × 3 (2-pair)	*Hc* @ *Mn* 5 × 5 (2-pair)	*Mn* 2 × 2 (2-pair)	*Mn* 2 × 2 (dense)	*Mn* 25 × 25 (2-pair)	*Mn* 25 × 25 (dense)	*Mn* 3 × 3 (2-pair)	*Mn* 3 × 3 (dense)	*Mn* 5 × 5 (2-pair)	*Mn* 5 × 5 (dense)	*Mn* 50 × 50 (dense)	*Nm* CDF @ Bivariate *Nm* 1 × 1	*Nm* CDF @ *Mn* 25 × 25 (2-pair)	*Nm* CDF @ *Mn* 3 × 3 (2-pair)	*Nm* CDF @ *Mn* 5 × 5 (2-pair)	*Sp* @ *Mn* 25 × 25 (2-pair)	*Sp* @ *Mn* 3 × 3 (2-pair)	*Sp* @ *Mn* 5 × 5 (2-pair)	*Sp* @ *Nm* CDF @ *Mn* 25 × 25 (2-pair)	*Sp* @ *Nm* CDF @ *Mn* 3 × 3 (2-pair)	*Sp* @ *Nm* CDF @ *Mn* 5 × 5 (2-pair)	*St* 1 × 1 (dof=1)	*St* 2 × 2 (dof=1)	*St* 2 × 2 (dof=2)	*St* 3 × 3 (dof=2)	*St* 3 × 3 (dof=3)	*St* 5 × 5 (dof=2)	*St* 5 × 5 (dof=3)	Swiss roll 2 × 1	Uniform 1 × 1 (additive noise=0.1)	Uniform 1 × 1 (additive noise=0.75)	Wiggly @ Bivariate *Nm* 1 × 1
GT	0.2	0.4	0.3	0.4	0.4	0.4	0.4	1.0	1.0	1.0	1.0	0.3	1.0	1.3	1.0	0.4	1.0	0.6	1.6	0.4	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.2	0.4	0.2	0.3	0.2	0.4	0.3	0.4	1.7	0.3	0.4
MINE	0.2	0.4	0.2	0.4	0.4	0.4	0.4	1.0	1.0	1.0	1.0	0.3	1.0	1.3	1.0	0.4	1.0	0.6	1.6	0.4	0.9	0.9	0.9	0.8	0.7	0.6	0.9	0.9	0.9	0.0	0.0	0.1	0.1	0.1	0.2	0.2	0.4	1.7	0.3	0.4
InfoNCE	0.2	0.4	0.3	0.4	0.4	0.4	0.4	1.0	1.0	1.0	1.0	0.3	1.0	1.3	1.0	0.4	1.0	0.6	1.6	0.4	0.9	1.0	1.0	0.8	0.8	0.8	0.9	1.0	1.0	0.2	0.3	0.2	0.3	0.2	0.4	0.3	0.4	1.7	0.3	0.4
D-V	0.2	0.4	0.3	0.4	0.4	0.4	0.4	1.0	1.0	1.0	1.0	0.3	1.0	1.3	1.0	0.4	1.0	0.6	1.6	0.4	0.9	1.0	1.0	0.8	0.8	0.8	0.9	1.0	1.0	0.0	0.0	0.1	0.1	0.2	0.2	0.2	0.4	1.7	0.3	0.4
NWJ	0.2	0.4	0.3	0.4	0.4	0.4	0.4	1.0	1.0	1.0	1.0	0.3	1.0	1.3	1.0	0.4	1.0	0.6	1.6	0.4	0.9	1.0	1.0	0.8	0.8	0.8	0.9	1.0	1.0	0.0	0.0	0.0	-0.6	0.1	0.1	0.2	0.4	1.7	0.3	0.4
DoE(Gaussian)	0.2	0.5	0.3	0.6	0.4	0.4	0.4	0.7	1.0	1.0	1.0	0.4	0.7	7.8	1.0	0.6	0.9	1.3		0.4	0.7	1.0	1.0	0.5	0.6	0.6	0.6	0.7	0.8	6.7	7.9	1.8	2.5	0.6	4.2	1.2	1.6	0.1	0.4	
DoE(Logistic)	0.1	0.4	0.2	0.4	0.4	0.4	0.4	0.6	0.9	0.9	1.0	0.3	0.7	7.8	1.0	0.6	0.9	1.3		0.4	0.8	1.1	1.0	0.5	0.6	0.6	0.7	0.8	0.8		2.0	0.5	0.8	0.3	1.5	0.6	1.6	0.1	0.4	
MINDE–J ($\sigma = 1$)	0.2	0.4	0.3	0.4	0.4	0.4	0.4	1.1	1.0	1.0	1.0	0.3	0.9	1.2	1.0	0.4	1.0	0.6	1.7	0.4	1.0	1.0	1.0	0.9	0.9	0.9	1.0	0.9	1.0	0.2	0.4	0.2	0.3	0.2	0.5	0.3	0.5	1.6	0.3	0.4
MINDE–J	0.2	0.4	0.3	0.4	0.4	0.4	0.4	1.2	1.0	1.0	1.0	0.3	1.0	1.3	1.0	0.4	1.0	0.6	1.7	0.4	1.1	1.0	1.0	1.0	0.9	0.9	1.1	1.0	1.0	0.1	0.2	0.2	0.3	0.2	0.5	0.3	0.4	1.7	0.3	0.4
MINDE–C ($\sigma = 1$)	0.2	0.4	0.3	0.4	0.4	0.4	0.4	1.0	1.0	1.0	1.0	0.3	1.0	1.3	1.0	0.4	1.0	0.6	1.6	0.4	0.9	1.0	1.0	0.9	0.9	0.9	0.9	1.0	0.9	0.1	0.3	0.2	0.3	0.2	0.4	0.3	0.3	1.7	0.3	0.4
MINDE–C	0.2	0.4	0.3	0.4	0.4	0.4	0.4	1.0	1.0	1.0	1.0	0.3	1.0	1.3	1.0	0.4	1.0	0.6	1.6	0.4	1.0	1.0	1.0	0.9	0.9	0.9	1.0	1.0	1.0	0.1	0.3	0.2	0.3	0.2	0.4	0.3	0.4	1.7	0.3	0.4
MMG	0.2	0.4	0.3	0.5	0.4	0.4	0.4	0.9	1.1	0.9	1.0	0.2	0.9	1.2	1.0	0.4	1.0	0.6	1.4	0.5	1.0	1.0	1.0	1.1	1.0	1.0	1.2	1.0	1.0	0.2	0.3	0.2	0.3	0.2	0.4	0.2	0.3	1.7	0.3	0.4
MMG-adaptive	0.2	0.4	0.3	0.4	0.4	0.4	0.4	0.9	1.0	0.9	1.0	0.3	1.0	1.2	1.0	0.4	1.0	0.6	1.4	0.4	1.0	1.0	1.1	1.0	1.0	1.0	1.1	1.0	1.0	0.2	0.3	0.2	0.3	0.2	0.5	0.2	0.3	1.7	0.3	0.4
MMG-orthogonal	0.2	0.4	0.3	0.4	0.4	0.4	0.4	1.0	1.0	1.0	1.0	0.3	1.0	1.3	1.0	0.4	1.0	0.6	1.6	0.4	1.0	1.0	1.0	0.9	1.0	1.0	1.0	1.0	1.0	0.1	0.3	0.2	0.3	0.2	0.4	0.3	0.4	1.7	0.3	0.4
MMG-orthogonal-adaptive	0.2	0.4	0.3	0.4	0.4	0.4	0.4	1.0	1.0	1.0	1.0	0.3	1.0	1.3	1.0	0.4	1.0	0.6	1.6	0.4	1.0	1.0	1.0	0.9	1.0	1.0	1.0	1.0	1.0	0.2	0.4	0.2	0.3	0.2	0.4	0.3	0.4	1.7	0.3	0.4

Table 1: Low MI estimation over 10 seeds using N = 10k test samples against ground truth (GT) [Czyż et al., 2023]. All methods were trained with 100k samples. **Color indicates relative negative (red) and positive bias (blue)**. Blank entries indicate that an estimator experienced numerical instabilities. List of abbreviations (*Mn*: Multinormal, *St*: Student-t, *Nm*: Normal, *Hc*: Half-cube, *Sp*: Spiral)

- **MMG**: The baseline estimator using a fixed, default importance sampling distribution.
- **MMG-adaptive**: Employs adaptive importance sampling for more accurate integration.
- **MMG-orthogonal**: Applies the orthogonal principle at inference time with baseline sampling.
- **MMG-orthogonal-adaptive**: Combines both adaptive sampling and the orthogonal principle.

5.1 MI ESTIMATION BENCHMARK

We evaluate on the benchmark of Czyż et al. [2023], which combines base distributions (Uniform, Normal with dense or sparse correlation, and long-tailed Student-t) with MI-preserving nonlinear transformations (Half-Cube, Asinh, Swiss-roll, Spiral). This setup introduces high dimensionality, heavy tails, sparsity, and non-linear geometry. We compare MMG against neural estimators, such as MINE [Belghazi et al., 2018], INFONCE [van den Oord et al., 2018], NWJ [Nguyen et al., 2007], DOE [McAllester and Stratos, 2020] and MINDE [Franzese et al., 2024].

The results in Table 1 establish our method's state-of-the-art performance. Our **MMG-orthogonal-adaptive** and **MMG-orthogonal** variants succeed on 39/40 and 37/40 tasks respectively, surpassing the MINDE (35/40). This robustness is rooted in our two main contributions: adaptive importance sampling ensures accuracy by focusing the integral on critical SNR regions, while the orthogonal principle guarantees a low-variance integrand for stability. This combination allows our method to excel on complex non-linear datasets where traditional methods fail.

5.2 HIGH MI BENCHMARK

To test the robustness and limits of our estimator, we extend the high-MI study from MINDE [Franzese et al., 2024], pushing their experimental setup from a range of MI ≤ 5 into the significantly more challenging regime of MI $\in [10, 15]$. Following their protocol, we use a sparse 3×3 Multinormal setup and its two MI-preserving transforms (Half-cube, Spiral), comparing all four MMG variants against both MINDE-C and MINDE-σ. Figure 3 reports the results; the complete per-task absolute errors and the experimental setup are given in Appendix F.

On the *original* and *Half-cube* cases (Fig. 3a&b), **MMG-adaptive** remains the most accurate as MI grows, achieving the lowest absolute error in 7 out of 9 settings with a mean absolute error of 1.75 nats, compared to 4.12 nats for MINDE. Other estimators falter due to distinct sources of error. **MINDE** saturates once the true MI exceeds roughly 13 nats: its score-matching objective requires approximating

the sharp score functions of concentrated high-MI distributions, a target where neural networks' spectral bias leads to significant underestimation [Rahaman et al., 2019] (see Appendix C). In contrast, our orthogonal variants exhibit a different limitation: a systematic conservative bias. This occurs because the orthogonal estimator measures the distance between neural network approximations of the denoisers, and the distance between these "smoothed-out" approximations is inherently smaller than the true distance between the optimal denoisers. This underestimation bias becomes more pronounced as the true MI gap grows larger.

This reveals a clear bias-variance trade-off dictating the optimal estimator. For the broad low-MI benchmark, the main challenge is variance, making the stable **MMG-orthogonal-adaptive** the superior choice. In this high-MI setting, however, this systematic bias becomes the dominant error, rendering the less-biased **MMG-adaptive** more accurate. This relative performance ranking holds even on the highly nonlinear *Spiral* transform (Fig. 3c), where all methods are challenged.

5.3 CONSISTENCY TEST

We conduct self-consistency tests inspired by Song and Ermon [2019] to evaluate the properties of MMG using high-dimensional real-world data, specifically samples from the MNIST dataset (28×28 resolution) [Deng, 2012]. Let A represent an image and B_r denote the image consisting of the top r rows of A. These tests are designed to verify whether the estimators adhere to fundamental properties of MI through three subtasks: Baseline Test, Data-Processing Test, and Additivity Test for two independent images sampled from dataset. Ideally, as r increases, $\frac{I(A;B_r)}{I(A;B)}$ should monotonically approach 1, $\frac{I(A;[B_{r+k},B_r])}{I(A;B_{r+k})}$ should consistently equal 1, and $\frac{I([A^1,A^2];[B_r^1,B_r^2])}{I(A^1;B_r^1)}$ should consistently equal 2. To ensure a fair comparison with MINDE [Franzese et al., 2024], we aligned our experimental settings and parameters and tested MMG using five random seeds. The results of the three tests are presented in Figure 4. Overall, MMG performed well and successfully passed all tests. We further repeat the baseline consistency test at higher dimensionality on CIFAR-10 (3072 dimensions) for all four MMG variants in Appendix E.

6 ABLATION STUDY ON ADAPTIVE SAMPLING

In this section, we conduct an ablation study to specifically evaluate the impact of our adaptive importance sampling strategy. We compare the performance of two non-orthogonal estimator variants:

- **MMG-adaptive ("Adaptive"):** Employs the adaptive importance sampling scheme described in Section 4.
- **MMG ("Baseline"):** Uses a fixed, default importance sampling distribution.

This comparison is designed to isolate the effect of the sampling strategy on estimation accuracy and stability, particularly across a diverse set of distributions. The results are presented in Table 2.

The results in Table 2 demonstrate the consistent benefits of the adaptive sampling strategy. The "Adaptive" method frequently achieves a lower standard deviation, indicating improved estimator stability. This is particularly evident in high-dimensional and complex transformation tasks, such as `multinormal-dense-50-50-0.5` and `spiral-multinormal-sparse-25-25-2.0`, where the reduction in variance is substantial. While the accuracy of the point estimate is competitive across both methods, the adaptive approach often provides estimates closer to the ground truth in the more challenging settings. Overall, this ablation study validates that tailoring the sampling distribution to the specific data leads to a more robust and reliable MI estimator.

7 VISUAL ANALYSIS OF THE MMG ESTIMATOR

This section provides an intuitive, qualitative analysis of the MMG estimator's integrand to support the core components of our method. By visualizing the integrand on a challenging three-dimensional, Spiral-transformed sparse Multinormal distribution (GT MI = 9.90), we can clearly see the complementary roles of the orthogonal principle for stability and adaptive importance sampling for accuracy. The plots below were generated by densely sampling 10,000 log SNR points and then binning the results (bin=50) to illustrate the underlying trends.

Figure 5 directly compares the integrand calculated via direct MMSE subtraction against the one derived from our orthogonal principle.

Analysis of the Orthogonal Principle's Contribution The orthogonal principle's primary contribution is enhancing integrand stability. As shown in Figure 5b, it replaces the highly volatile direct subtraction method with an exceptionally smooth and non-negative integrand, crucial for reliable numerical integration. This stability, however, introduces a trade-off that becomes particularly evident in **high mutual information estimation**: the orthogonal integrand's peak is lower, which can lead to a conservative bias by underestimating the true distance between optimal denoisers. While the dramatic reduction in variance makes it a more robust choice for general cases, this systematic underestimation can become a limiting factor when the true MI is large.

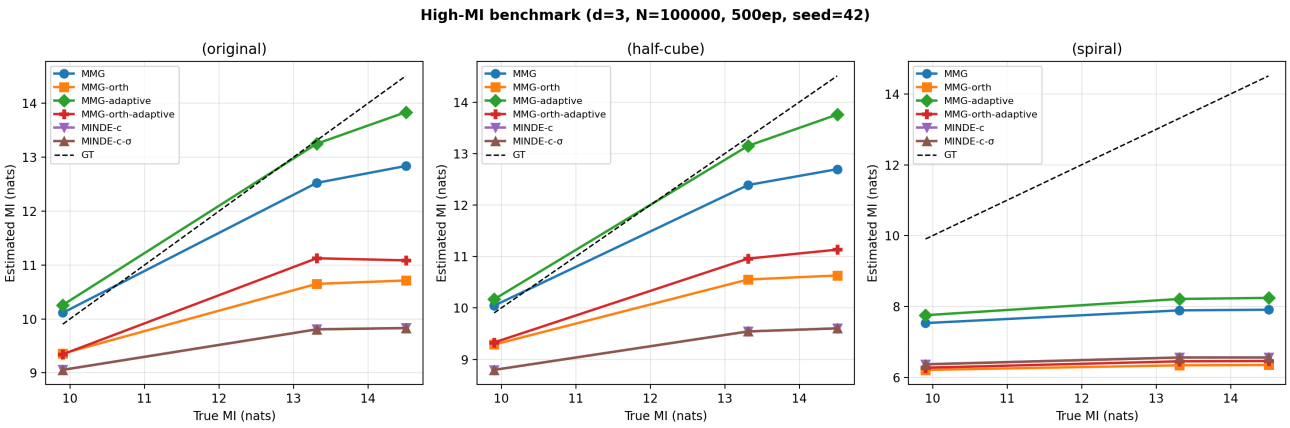
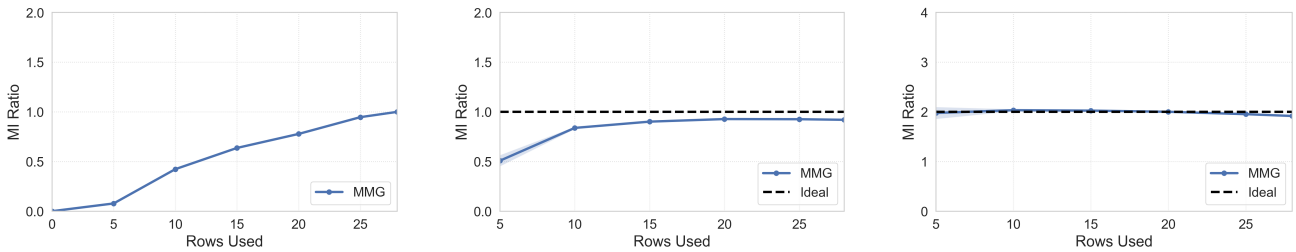


Figure 3: High MI benchmark on the sparse 3×3 Multinormal task (true MI ≈ 9.9, 13.3, and 14.5 nats) and its two MI-preserving transforms: (a) original, (b) Half-cube, and (c) Spiral. An accurate estimator should track the dashed ground-truth line. MMG-adaptive stays closest to the ground truth as MI grows, while MINDE saturates and the orthogonal variants underestimate due to their conservative bias; on the Spiral transform all methods degrade, but MMG-adaptive retains the lowest error.



(a) Baseline test (b) Data processing test (c) Additivity test

Figure 4: Consistency tests over the MNIST dataset. A well-behaved estimator should satisfy: (a) $\frac{I(A;B_r)}{I(A;B)}$ increases monotonically toward 1 as more rows r are revealed; (b) $\frac{I(A;[B_{r+k},B_r])}{I(A;B_{r+k})}$ stays at 1 for $k > 0$ (data-processing); and (c) $\frac{I([A^1,A^2];[B_r^1,B_r^2])}{I(A^1;B_r^1)}$ stays at 2 (additivity for two independent images). MMG matches all three expected trends.

Analysis of Adaptive Sampling's Contribution Adaptive sampling boosts estimation accuracy and efficiency. As illustrated in Figure 5b, while a fixed sampler may be misaligned with the integrand's peak, our adaptive method dynamically concentrates samples in this most informative SNR region. This targeted strategy leads to a demonstrably more accurate MI estimate (**8.075 nats** vs. **6.933 nats** for the direct method), confirming the value of focusing the integration where it matters most.

8 RELATED WORK

Estimating mutual information (MI) from samples is a central challenge in machine learning. Traditional non-parametric approaches, such as methods based on data binning or kernel density estimation (KDE), struggle with the curse of dimensionality. More advanced estimators based on k-nearest neighbors have shown significant improvements [Kraskov et al., 2004, Pál et al., 2010], but can still be challenged by complex, high-dimensional data [Gao et al., 2015]. Consequently, the modern paradigm is dominated by neural network-based approaches that optimize a variational bound on the MI.

Variational Bounds on Mutual Information. Most recent neural MI estimators are built upon variational lower bounds derived from f-divergences. The pioneering work, MINE [Belghazi et al., 2018], leverages the Donsker-Varadhan representation of the KL-divergence, spurring related estimators like the variance-reduced SMILE [Song and Ermon, 2019] and DoE [McAllester and Stratos, 2020]. A particularly successful family is based on contrastive learning, where estimators like InfoNCE [van den Oord et al., 2018] train a critic to distinguish between joint and marginal samples [Poole et al., 2019]. However, these methods face significant limitations: their sample complexity can scale exponentially with the true MI, often being capped by the

Table 2: Ablation study comparing MMG-adaptive ("Adaptive") with adaptive importance sampling against the baseline MMG ("Baseline") with a fixed sampling distribution. This analysis uses the non-orthogonal variants to isolate the impact of the sampling strategy. For each task, the estimate closer to the Ground Truth and the lower standard deviation (Std) are independently bolded.

Task	Ground Truth	Adaptive		Baseline	
		Estimate	Std	Estimate	Std
Basic Distribution Tasks					
1v1-normal-0.75	0.4133	**0.4160**	**0.0447**	0.4208	0.0754
1v1-additive-0.1	1.7094	1.6929	**0.0494**	**1.6984**	0.0677
1v1-bimodal-0.75	0.4133	**0.4133**	**0.0612**	0.4201	0.0766
High Dimensional Tasks					
multinormal-dense-25-25-0.5	1.2922	**1.2063**	**0.1636**	1.1603	0.2526
multinormal-dense-50-50-0.5	1.6243	**1.7926**	**0.3053**	1.8493	0.4398
Non-Gaussian Distribution Tasks					
student-identity-1-1-1	0.2242	0.3644	**0.2472**	**0.3583**	0.2884
student-identity-3-3-2	0.2909	**0.4901**	**0.6502**	0.5123	0.7055
student-identity-5-5-2	0.4482	0.7747	1.0550	**0.7683**	**1.0492**
Complex Transformation Tasks					
spiral-multinormal-sparse-3-3-2.0	1.0217	**1.0012**	**0.0805**	1.0152	0.1187
spiral-multinormal-sparse-25-25-2.0	1.0217	**0.8713**	**0.1557**	0.8215	0.2683

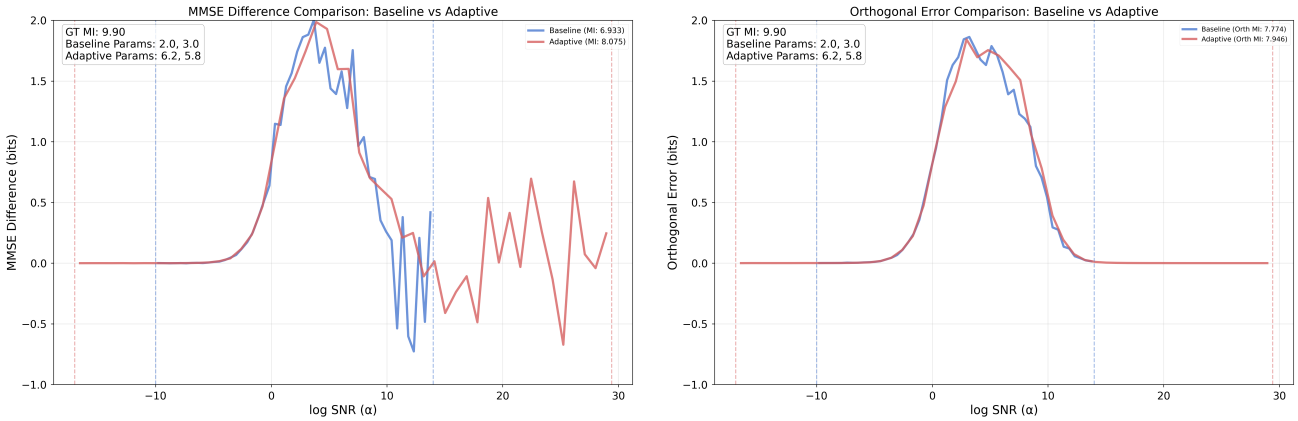
logarithm of the batch size [McAllester and Stratos, 2020]. This difficulty can be viewed as a "density chasm": the marginal distribution $p(x)p(y)$ is often a poor proposal for the joint $p(x, y)$, leading to high-variance estimates, especially in high-MI settings [Rhodes et al., 2020, Brekelmans et al., 2021]. Several lines of work address these limitations from complementary directions: Choi and Lee [2022] regularize MINE-style losses to combat their training instability, Butakov et al. [2024] estimate MI through normalizing flows with closed-form pointwise MI, and InfoNet [Hu et al., 2024] amortizes estimation in a feed-forward network to avoid test-time optimization. MMG is complementary to these approaches: it is neither a variational-bound regularizer nor an amortized estimator, but a denoising-based estimator built on the MMSE-gap identity.

Mutual Information Estimation with Diffusion Models. Denoising diffusion models offer a natural and powerful framework to bridge this density chasm. They define a continuous process that transforms a complex data distribution into a simple tractable one, providing a path of intermediate distributions. The potential of this framework for MI estimation was recently demonstrated by **MINDE** [Franzese et al., 2024], which connects MI to the difference between conditional and unconditional score functions ($\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x})$). Our work, **MMG**, builds on a different and more direct connection [Guo et al., 2005, Kong et al., 2023]. Instead of relying on learned score functions, we show that MI corresponds exactly to the integrated gap between the Minimum Mean Square Error (MMSE) of conditional and unconditional denoising. This formulation connects MI directly to the denoising objective itself, rather than its gradient, providing an elegant and potentially more robust pathway for estimation. Concurrently, Ohayon et al. [2025] exploit the same information-estimation geometry to construct a perceptual distance metric (IEM) from denoiser errors across noise levels, further demonstrating the versatility of this fundamental connection.

9 CONCLUSION

In this work, we introduced MMG, a principled and robust estimator for mutual information derived from the information-theoretic properties of denoising diffusion models. Our method directly connects MI to the integrated Minimum Mean Square Error (MMSE) gap between a conditional and an unconditional denoiser. We further enhanced this framework with two key techniques: an adaptive importance sampling scheme to improve integration accuracy and an orthogonal principle to increase estimator stability.

Through extensive experiments, we demonstrated that MMG achieves exceptional accuracy and robustness, successfully providing stable estimates on 39 out of 40 tasks in a comprehensive benchmark, and successfully passes all self-consistency tests. Notably, our method excels in the challenging high-MI regime, significantly outperforming current score-based diffusion estimators. Our analysis also uncovered a fundamental bias-variance trade-off, revealing that the optimal estimator configuration—either with



(a) MMSE Difference

(b) Orthogonal Error

Figure 5: **Integrand analysis on a Spiral-transformed task (GT MI = 9.90).** Comparison between (a) the volatile direct MMSE subtraction method and (b) the stable orthogonal principle. In both plots, the adaptive sampler (red) outperforms the baseline (blue) by focusing on the most informative region.

or without the orthogonal principle—depends on the MI magnitude of the problem.

10 LIMITATIONS AND FUTURE WORK

Unlike variational methods such as MINE [Belghazi et al., 2018] or InfoNCE [van den Oord et al., 2018] that optimize explicit lower bounds, MMG provides no formal bound guarantee: because the MMSE terms appear with opposite signs in Eq. 8, approximation errors do not cancel in a controlled way. We note that the closest diffusion-based baseline, MINDE [Franzese et al., 2024], similarly does not provide a formal MI bound; it too relies on neural approximation, in its case of the score functions rather than the denoisers. Our empirical validation is also largely restricted to synthetic benchmarks and image consistency tests up to 3072 dimensions (Appendix E); full validation on high-resolution images or 3D data remains open. Additionally, the adaptive variant requires a two-stage training procedure, and scaling to very high-dimensional data would likely require replacing the MLP architecture with more expressive models such as U-Nets. Our experiments also revealed that the optimal variant depends on the unknown MI magnitude, with no automatic selection mechanism currently available.

Future work could address these limitations by developing adaptive criteria to select the best variant at inference time, and by leveraging pre-trained diffusion models to reduce training costs. Because MMG requires only denoising/MMSE predictions across SNRs, rather than a generative sampler, a pre-trained diffusion model whose conditioning variable and noising process already match the task could in principle be reused after light calibration; when the conditioning variable differs, a small conditional adapter or head, together with an emptied condition for the unconditional branch, may suffice. More broadly, the MMSE-gap framework naturally extends to estimating conditional mutual information, transfer entropy, and directed information, broadening its applicability to causal inference and time-series analysis. Beyond mutual information estimation, our results suggest a deeper connection between denoising diffusion geometry and information-theoretic quantities: by interpreting estimation accuracy through MMSE gaps across noise scales, diffusion models may provide a unified computational framework for measuring statistical dependence without explicit density estimation, opening opportunities in representation learning, causal discovery, and scientific data analysis.

References

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International Conference on Machine Learning*, pages 531–540. PMLR, 2018.

Rob Brekelmans, Sicong Huang, Marzyeh Ghassemi, Greg Ver Steeg, Roger Baker Grosse, and Alireza Makhzani. Improving mutual information estimation with annealed and energy-based bounds. In *International Conference on Learning Representations*, 2021.

Ivan Butakov, Alexander Tolmachev, Sofia Malanchuk, Anna Neopryatnaya, and Alexey Frolov. Mutual information estimation via normalizing flows. *Advances in Neural Information Processing Systems*, 37:78242–78266, 2024.

Kwanghee Choi and Siyeong Lee. Combating the instability of mutual information-based losses via regularization. In *Uncertainty in Artificial Intelligence*, pages 411–421. PMLR, 2022.

Paweł Czyż, Frederic Grabowski, Julia E Vogt, Niko Beerenwinkel, and Alexander Marx. Beyond normal: On the evaluation of mutual information estimators. *Advances in Neural Information Processing Systems*, 36:16957–16990, 2023.

Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012. doi: 10.1109/MSP.2012.2211477.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, pages 1675–1685. PMLR, 2019.

Giulio Franzese, Mustapha Bounoua, and Pietro Michiardi. MINDE: Mutual information neural diffusion estimation. In *International Conference on Learning Representations*, 2024.

Shuyang Gao, Greg Ver Steeg, and Aram Galstyan. Efficient estimation of mutual information for strongly dependent variables. In *Artificial intelligence and statistics*, pages 277–286. PMLR, 2015.

Weihao Gao, Sreeram Kannan, Sewoong Oh, and Pramod Viswanath. Estimating mutual information for discrete-continuous mixtures. *Advances in neural information processing systems*, 30, 2017.

Behnam Gholami, Pritish Sahu, Ognjen Rudovic, Konstantinos Bousmalis, and Vladimir Pavlovic. Unsupervised multi-target domain adaptation: An information theoretic approach. *IEEE Transactions on Image Processing*, 29: 3993–4002, 2020.

Dongning Guo, Shlomo Shamai, and Sergio Verdú. Mutual information and minimum mean-square error in gaussian channels. *IEEE transactions on information theory*, 51 (4):1261–1282, 2005.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2019.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Zhengyang Hu, Song Kang, Qunsong Zeng, Kaibin Huang, and Yanchao Yang. Infonet: Neural estimation of mutual information without test-time optimization. In *International Conference on Machine Learning*. PMLR, 2024.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Diederik P Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *arXiv preprint arXiv:2107.00630*, 2021.

Justin B Kinney and Gurinder S Atwal. Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences*, 111 (9):3354–3359, 2014.

Xianghao Kong, Rob Brekelmans, and Greg Ver Steeg. Information-theoretic diffusion. In *International Conference on Learning Representations*, 2023. URL `https://openreview.net/pdf?id=UvmDCdSPDOW`.

Xianghao Kong, Ollie Liu, Han Li, Dani Yogatama, and Greg Ver Steeg. Interpretable diffusion via information decomposition. In *International Conference on Learning Representations*, 2024.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Phys. Rev. E*, 69:066138, Jun 2004. doi: 10.1103/PhysRevE.69.066138. URL https://link.aps.org/doi/10.1103/PhysRevE.69.066138.

David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. In *International Conference on Artificial Intelligence and Statistics*, pages 875–884. PMLR, 2020.

XuanLong Nguyen, Martin J Wainwright, and Michael Jordan. Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In *Advances in Neural Information Processing Systems*, 2007.

XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Guy Ohayon, Pierre-Etienne H Fiquet, Florentin Guth, Jona Ballé, and Eero P Simoncelli. Learning a distance measure from the information-estimation geometry of data. *arXiv preprint arXiv:2510.02514*, 2025.

Dávid Pál, Barnabás Póczos, and Csaba Szepesvári. Estimation of rényi entropy and mutual information based on generalized nearest-neighbor graphs. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems-Volume 2*, pages 1849–1857, 2010.

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International conference on machine learning*, pages 5301–5310. PMLR, 2019.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Benjamin Rhodes, Kai Xu, and Michael U Gutmann. Telescoping density-ratio estimation. *Advances in Neural Information Processing Systems*, 2020.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

Jascha Sohl-Dickstein, Eric A Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. *arXiv preprint arXiv:1503.03585*, 2015.

Jiaming Song and Stefano Ermon. Understanding the limitations of variational mutual information estimators. In *International Conference on Learning Representations*, 2019.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Yihong Wu and Sergio Verdú. Functional properties of minimum mean-square error and mutual information. *IEEE Transactions on Information Theory*, 58(3):1289–1301, 2011.

A DERIVATION OF THE ORTHOGONAL PRINCIPLE

This section provides a brief derivation for the orthogonal principle (Equation 11) used in our estimator, following the original work of Kong et al. [2024].

The principle is based on the following identity. Let $\hat{\boldsymbol{x}}(\boldsymbol{z}_\gamma) \equiv \hat{\boldsymbol{x}} = \mathbb{E}[\boldsymbol{x}|\boldsymbol{z}_\gamma]$ be the unconditional MMSE denoiser and $\hat{\boldsymbol{x}}(\boldsymbol{z}_\gamma, y) \equiv \hat{\boldsymbol{x}}_y = \mathbb{E}[\boldsymbol{x}|\boldsymbol{z}_\gamma, y]$ be the conditional one. The identity states:

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2] - \mathbb{E}[\|\boldsymbol{x} - \hat{\boldsymbol{x}}_y\|^2] \\ = \mathbb{E}[\|\hat{\boldsymbol{x}}_y - \hat{\boldsymbol{x}}\|^2]\end{aligned} \tag{13}$$

Below is a brief proof sketch for Equation 13.

Proof Sketch. For brevity, define the conditional error $\boldsymbol{e}_y \coloneqq \boldsymbol{x} - \hat{\boldsymbol{x}}_y$, the unconditional error $\boldsymbol{e} \coloneqq \boldsymbol{x} - \hat{\boldsymbol{x}}$, and the denoiser gap $\boldsymbol{\delta} \coloneqq \hat{\boldsymbol{x}}_y - \hat{\boldsymbol{x}}$, so that $\boldsymbol{e} = \boldsymbol{e}_y + \boldsymbol{\delta}$.

We first show the cross-term vanishes:

$$\mathbb{E}[\boldsymbol{e}_y \cdot \boldsymbol{\delta}] = \mathbb{E}\big[\mathbb{E}[\boldsymbol{e}_y \cdot \boldsymbol{\delta} \mid \boldsymbol{z}_\gamma, y]\big] \tag{14}$$

$$= \mathbb{E}\big[\underbrace{(\mathbb{E}[\boldsymbol{x}|\boldsymbol{z}_\gamma, y] - \hat{\boldsymbol{x}}_y)}_{=\mathbf{0}}\boldsymbol{\delta}\big]$$

$$= 0 \tag{15}$$

where equation 14 applies the tower property, and equation 15 uses $\hat{\boldsymbol{x}}_y = \mathbb{E}[\boldsymbol{x}|\boldsymbol{z}_\gamma, y]$.

Since the cross-term is zero, expanding $\|\boldsymbol{e}\|^2$ gives:

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{e}\|^2] &= \mathbb{E}[\|\boldsymbol{e}_y + \boldsymbol{\delta}\|^2] \\ &= \mathbb{E}[\|\boldsymbol{e}_y\|^2] + \mathbb{E}[\|\boldsymbol{\delta}\|^2] \\ &\quad + 2\underbrace{\mathbb{E}[\boldsymbol{e}_y \cdot \boldsymbol{\delta}]}_{=0} \\ &= \mathbb{E}[\|\boldsymbol{e}_y\|^2] + \mathbb{E}[\|\boldsymbol{\delta}\|^2]\end{aligned}$$

Rearranging yields Equation 13.

Table 3: MMG network training hyper-parameters. The task dimension (Dim) corresponds to the sum of the dimensions of the two variables, and p_{drop} denotes the conditioning dropout probability.

Benchmark Dim	p_{drop}	Width	Time Emb. Size	Batch Size	lr	Iterations	# Params
≤ 10	0.5	64	64	128	1e-3	390k	55425
50	0.5	128	128	256	2e-3	290k	220810
100	0.5	256	256	256	2e-3	290k	898354
Consistency Tests	0.5	256	256	64	1e-3	390k	1597968

Table 4: Sampling Hyperparameters

Sampling Param.	LogSNR Loc	LogSNR Scale	Clip	EMA Decay	Inf. Times	N_Points
-	2.0	3.0	4.0	0.999	10	10000

B IMPLEMENTATION DETAILS

We follow the implementation of Franzese et al. [2024] which uses stacked multi-layer perceptron (MLP) with skip connections. We adopt a simplified version of the same network architecture: this involves three Residual MLP blocks. We use the *Adam optimizer* [Kingma and Ba, 2015] for training and Exponential moving average (EMA) with a momentum parameter $m = 0.999$. We use the *ReduceLROnPlateau* scheduler with a patience of 200 epochs, reducing the learning rate by half if the training loss does not improve after 200 epochs. We returned the mean estimate on the test data set over 10 runs. All experiments are run on NVIDIA RTX A6000 GPUs. The hyper-parameters are presented in Table 3 for MMG. Concerning the consistency tests, we independently train an autoencoder for each version of the MNIST dataset with r rows available.

C SCORE TARGET VS. DENOISING TARGET: APPROXIMATION DIFFICULTY

Although the score and the optimal denoiser are related by an affine transformation at the population optimum via Tweedie's formula, finite-capacity neural networks need not approximate the two training targets equally well. In this section we support this approximation-difficulty explanation with two complementary experiments.

C.1 SPECTRAL-BIAS DIAGNOSTIC ON A BIMODAL MIXTURE

We use a one-dimensional bimodal Gaussian mixture as a non-Gaussian diagnostic and vary the separation between the two modes. Larger separation makes the score target sharper around the low-density transition region between the modes, while the denoiser target remains smoother and bounded within the data range. We train the same 2-layer MLP (64 hidden units) to fit each target and report the resulting function-fitting MSE in Table 5 and Figure 6.

Table 5: Function-fitting MSE for the score and denoiser targets using the same 2-layer MLP (lower is better). As mode separation grows, the score target becomes relatively harder to fit.

Mode separation	Score MSE	Denoiser MSE	Ratio
1.0 (easy)	6.6×10^{-7}	7.1×10^{-7}	0.9×
3.0 (medium)	1.4×10^{-6}	4.7×10^{-7}	2.9×
5.0 (hard)	1.0×10^{-5}	3.1×10^{-6}	3.2×

This behavior is consistent with the spectral bias of neural networks [Rahaman et al., 2019]: networks learn smoother, lower-frequency functions more easily, so the smoother denoising target is easier to approximate when the score target contains sharp, high-frequency structure.

C.2 SAME-NETWORK CONTROLLED COMPARISON

To isolate the effect of the training target from all other design choices, we train MMG (standard and orthogonal forms) and MINDE-C with *identical* networks (a simplified U-Net-style MLP with hidden width 64, roughly 43K parameters), the same optimizer, data ($N = 100$K samples of a 3-dimensional Gaussian task), and 500 training epochs, changing only the objective. Table 6 sweeps the true MI from 0.14 to 9.32 nats; Figure 7 visualizes the results.

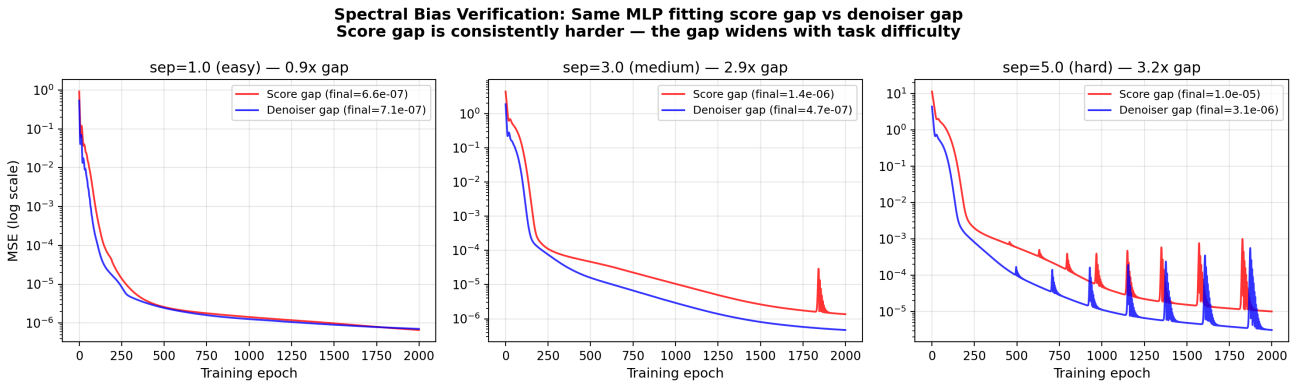


Figure 6: Spectral-bias diagnostic on a 1D bimodal Gaussian mixture. As the separation between modes increases, the score target develops sharp variations around the low-density transition region, while the denoiser target remains smooth; the same small MLP fits the denoiser target with substantially lower error.

Table 6: MI estimation absolute error under the same-network controlled comparison. The best estimator per row is marked in bold.

True MI	MMG	MMG-orth	MINDE-C
0.14	0.133	**0.006**	**0.006**
0.43	0.129	**0.013**	**0.013**
1.01	0.122	**0.024**	**0.024**
1.92	0.114	**0.039**	**0.038**
3.49	0.106	**0.066**	0.073
5.88	**0.106**	0.109	0.111
9.32	**0.112**	0.231	0.161

The results reflect the bias-variance trade-off discussed in the main text from a complementary angle. In the low-MI regime, MMG-orth and MINDE-C are nearly indistinguishable, while standard MMG carries a roughly constant positive offset: the orthogonal identity of Eq. 11 holds exactly only for Bayes-optimal denoisers, and for finite-capacity networks the direct MMSE subtraction picks up a non-zero approximation cross term that the orthogonal form avoids by sharing the same noise realization. At high MI, the ranking reverses: MINDE-C starts to underestimate, consistent with the spectral-bias effect above, while standard MMG remains stable. This confirms that MMG is not a single fixed estimator that only helps at high MI: the orthogonal form is preferable when the MMSE gap is small, while the standard form is preferable when it is large.

D CONTROLLED RUNTIME COMPARISON WITH MINDE

MMG trains a single shared denoising network via conditioning dropout, not two separate networks, so its computational profile is comparable to score-based diffusion estimators. To quantify this, we ran a controlled CPU comparison against MINDE-C using the same lightweight backbone, matched parameter counts, the same data, batch size, and optimizer steps, and 10 Monte Carlo samples at inference. Table 7 reports mean absolute estimation error and wall-clock time on the same ground-truth MI tasks.

Table 7: Controlled runtime comparison under matched backbones and training settings.

Method	Mean abs. error	Train time	Inference time
MMG	**0.056**	**9.73s**	0.173s
MINDE-C	0.105	10.68s	**0.162s**

Under this controlled setting, MMG and MINDE-C have comparable runtime; MMG's contribution is the denoising/MMSE training target, not a wall-clock speedup.

E CIFAR-10 CONSISTENCY TESTS

To evaluate MMG beyond low-dimensional synthetic benchmarks and the 784-dimensional MNIST tests in the main text, we repeat the baseline consistency test on CIFAR-10 ($3 \times 32 \times 32 = 3072$ dimensions). Since ground-truth MI is unavailable for natural images, we again use the row-revealing protocol: with A an image and B_r its top r rows, the ratio $I(A; B_r)/I(A; B)$ should increase monotonically toward 1 as more rows are revealed.

Table 8: CIFAR-10 baseline consistency test with a convolutional U-Net denoiser.

r	4	8	12	16	20	24	28	32
ratio	0.063	0.158	0.269	0.391	0.522	0.652	0.805	1.000

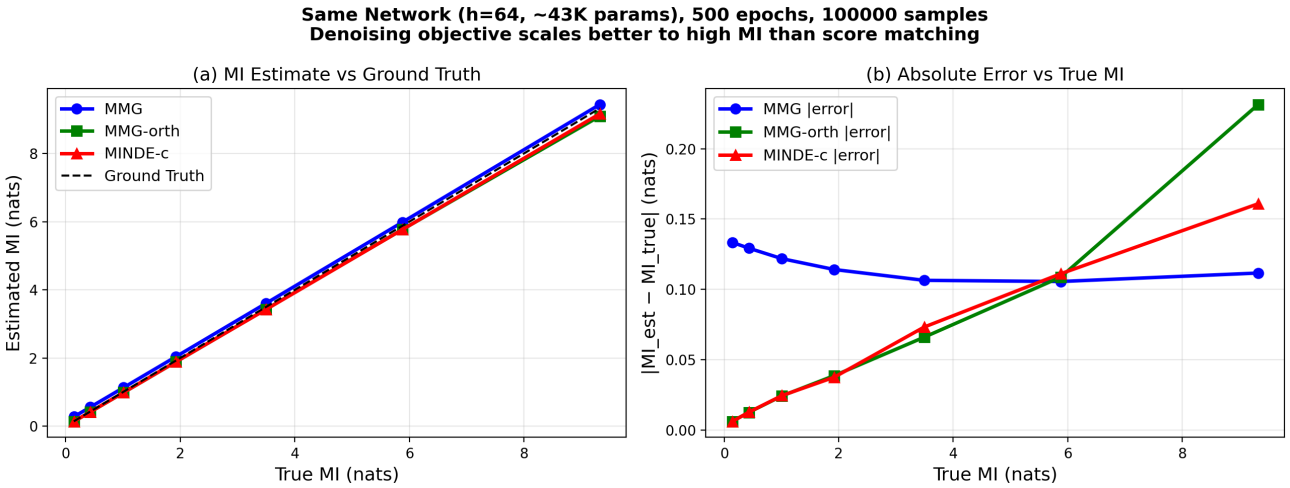


Figure 7: Same-network controlled comparison: identical architectures, optimizers, and data; only the training objective differs. MMG-orth matches MINDE-C in the low-MI, variance-dominated regime, while standard MMG becomes the most accurate at high MI, where MINDE-C begins to underestimate.

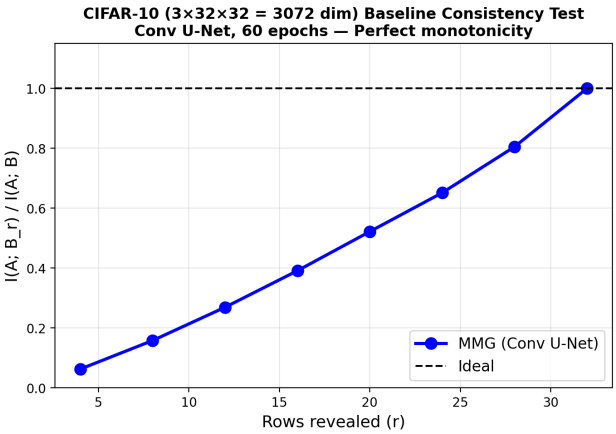


Figure 8: CIFAR-10 baseline consistency test (3072 dimensions). The ratio $I(A;B_r)/I(A;B)$ increases monotonically and reaches 1 at full resolution, as expected.

Table 8 shows that MMG passes the test with perfect monotonicity. We further extend the test to all four MMG variants using the same convolutional U-Net backbone; results are shown in Table 9 and Figure 9. All variants exhibit the expected monotone behavior converging to 1. For the adaptive variants, the calibration of Section 4 is applied per task and is well-defined for all r. While these tests do not fully settle scaling to high-resolution or 3D data, they demonstrate that the estimator behaves consistently well beyond low-dimensional synthetic settings.

Table 9: CIFAR-10 consistency test for all MMG variants: ratio $I(A;B_r)/I(A;B)$ as a function of revealed rows r.

r	MMG	MMG-orth	MMG-adapt.	MMG-orth-adapt.
4	0.060	0.047	0.032	0.045
8	0.156	0.132	0.120	0.120
12	0.266	0.232	0.225	0.212
16	0.388	0.346	0.338	0.321
20	0.519	0.481	0.465	0.443
24	0.649	0.640	0.603	0.581
28	0.804	0.799	0.749	0.735
32	1.000	1.000	1.000	1.000

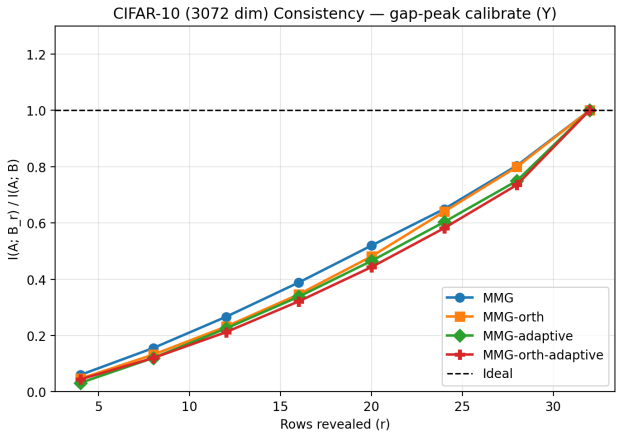


Figure 9: CIFAR-10 consistency test for all four MMG variants. Each curve should increase monotonically toward 1 as more rows are revealed; all variants pass.

F HIGH-MI BENCHMARK: FULL RESULTS

Table 10 reports the complete absolute errors underlying Figure 3, including all four MMG variants and both MINDE variants. The setup follows the MINDE protocol: a 3-dimensional sparse Multinormal task with correlation strengths chosen so that the true MI is approximately 9.9, 13.3, and 14.5 nats, together with its two MI-preserving transforms (Half-cube, Spiral); all estimators are trained with $N = 100$K samples for 500 epochs under matched settings. MMG-adaptive achieves the lowest error in 7 of 9 settings; baseline MMG wins the two lowest-MI cases, where variance dominates and the adaptive calibration can over-tune. MINDE saturates around 9.6–9.8 nats once the true MI exceeds 13 nats.

Table 10: High-MI benchmark: absolute errors $|\text{estimate} - \text{GT}|$ in nats. The best method per row is marked in bold.

Transform	GT	MMG	MMG-orth	MMG-adapt.	MMG-orth-adapt.	MINDE-C	MINDE-σ
original	9.90	**0.21**	0.55	0.36	0.56	0.85	0.85
original	13.31	0.79	2.66	**0.06**	2.19	3.50	3.51
original	14.51	1.67	3.79	**0.67**	3.42	4.68	4.68
half-cube	9.90	**0.14**	0.61	0.26	0.57	1.11	1.11
half-cube	13.31	0.92	2.76	**0.16**	2.35	3.77	3.77
half-cube	14.51	1.81	3.88	**0.75**	3.38	4.91	4.90
spiral	9.90	2.37	3.70	**2.15**	3.63	3.54	3.54
spiral	13.31	5.42	6.98	**5.10**	6.86	6.75	6.76
spiral	14.51	6.60	8.16	**6.27**	8.05	7.95	7.95
mean	–	2.22	3.68	**1.75**	3.45	4.12	4.12