MMG: Mutual Information Estimation via the MMSE Gap in Diffusion

Anonymous Author(s)
Affiliation
Address
email

Abstract

Mutual information (MI) 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 MI estimation. Using the recently introduced information-theoretic formulation of denoising diffusion models, we show the diffusion models can be used in a straightforward way to estimate MI. In particular, the MI corresponds to half the gap in the Minimum Mean Square Error (MMSE) between conditional and unconditional diffusion, integrated over all Signal-to-Noise-Ratios (SNRs) in the noising process. Our approach not only passes self-consistency tests but also outperforms traditional and score-based diffusion MI estimators. Furthermore, our method leverages adaptive importance sampling to achieve scalable MI estimation, while maintaining strong performance even when the MI is high.

1 Introduction

Estimating Mutual Information (MI) from samples is a fundamental problem with widespread applications. 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., 2018b; 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 & 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 & 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. (2022) 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 a 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. This 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., 2022). Our contributions are as follows:

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

- 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., 2022) 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. Our code can be found at: `https://anonymous.4open.science/r/DMI-50D8`

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., 2022).

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

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

$$\hat{\boldsymbol{x}}^*(\boldsymbol{z}_\gamma,\gamma) \equiv \arg\mathrm{mmse}_x(\gamma) = \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x}|\boldsymbol{z}_\gamma)}[\boldsymbol{x}] \tag{2}$$

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 $\mathbb{E}_{p(\boldsymbol{x})}[\mathrm{mmse}(\boldsymbol{x}|\gamma)] = \mathrm{mmse}_x(\gamma)$.

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

From Kong et al. (2022) 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.

$$\boxed{-\log p(\boldsymbol{x}) = d/2\log(2\pi e) - {}^1\!/_2 \int_0^\infty d\gamma \left(\frac{d}{1+\gamma} - \mathrm{mmse}(\boldsymbol{x}|\gamma)\right)} \tag{4}$$

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.

$$-\log p(\boldsymbol{x}|y) = d/2\log(2\pi e) - {}^1\!/_2 \int_0^\infty d\gamma \left(\frac{d}{1+\gamma} - \mathrm{mmse}(\boldsymbol{x}|\gamma, y)\right) \tag{5}$$

We use the following definitions for conditional MMSE.

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

$$\mathrm{mmse}(\boldsymbol{x}|\gamma,y) \equiv \mathbb{E}_{p(\boldsymbol{z}_\gamma|\boldsymbol{x})}[(\boldsymbol{x} - \hat{\boldsymbol{x}}^*(\boldsymbol{z}_\gamma,\gamma,y))^2] \tag{6}$$

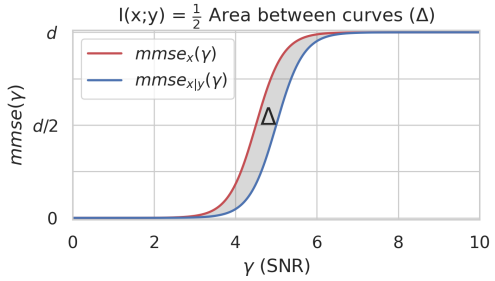


Figure 1: The mutual information is exactly half the area between MMSE curves for conditional and unconditional denoising. We use denoising diffusion models to approximate the MMSE curves, then numerically integrate to get an estimate of the mutual information.

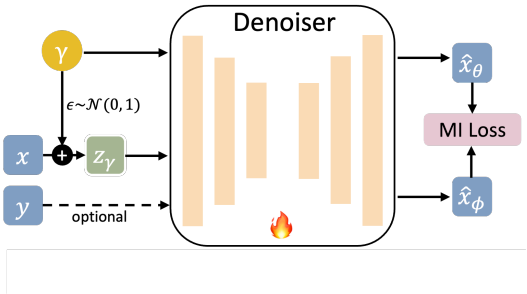


Figure 2: The schematic of the MMG training process. Noise at a given γ level is added to the data $\boldsymbol{x}$, and a denoiser is used to recover it, with or without conditioning on y at a 50% probability. Finally, the MI loss, as defined in Eq. 10, is computed to backpropagate the gradient.

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

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

$$\log p(\boldsymbol{x}|y) - \log p(\boldsymbol{x}) = {}^1\!/_2 \int_0^\infty d\gamma \left(\text{mmse}(\boldsymbol{x}|\gamma) - \text{mmse}(\boldsymbol{x}|\gamma, y)\right) \tag{7}$$

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.

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

Conditioning on y can only decrease the MMSE Wu & 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 $q(\gamma)$.

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

Therefore, we can train expert denoisers to estimate the MMSE for various distributions of $q(\gamma)$, as different density 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.

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

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 & Salimans (2021). 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.

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

The core of the experiment is 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. (2022), 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 Fig. 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, $\text{MMSE}_{x|y}$. Inspired by the use of error landmarks in density estimation (Kong et al., 2022), our heuristic identifies the critical transition region of this curve, where the denoiser becomes effective. Specifically, we define the parameters as follows (where d is the data dimensionality):

- 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 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. (2023). 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:

$$\underbrace{\mathbb{E}[\|\boldsymbol{x} - \hat{\boldsymbol{x}}(\boldsymbol{z}_\gamma)\|^2] - \mathbb{E}[\|\boldsymbol{x} - \hat{\boldsymbol{x}}(\boldsymbol{z}_\gamma, y)\|^2]}_{\text{MMSE Gap}} = \underbrace{\mathbb{E}[\|\hat{\boldsymbol{x}}(\boldsymbol{z}_\gamma, y) - \hat{\boldsymbol{x}}(\boldsymbol{z}_\gamma)\|^2]}_{\text{Orthogonal Form of the Gap}} \tag{11}$$

After expanding the left-hand side, the cross terms vanish due to the orthogonality principle Kong et al. (2023), 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:

- **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., 2018a), INFONCE (Oord et al., 2018),NWJ (Nguyen et al., 2007), DOE (McAllester & Stratos, 2020) and MINDE Franzese et al. (2023).

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.

	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	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)

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). Figure 3 reports the mean $\pm$ one standard deviation over 10 seeds.

On the *original* and *Half-cube* cases (Fig. 3a&b), **MMG-adaptive** remains the most accurate as MI grows, while other estimators falter due to distinct sources of error. **MINDE** struggles significantly as its score-matching objective requires approximating the sharp, high-frequency score functions of high-MI distributions, a task where neural networks' spectral bias leads to significant underestimation Rahaman et al. (2019). 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 non-linear *Spiral* transform (Fig. 3c), where all methods are challenged.

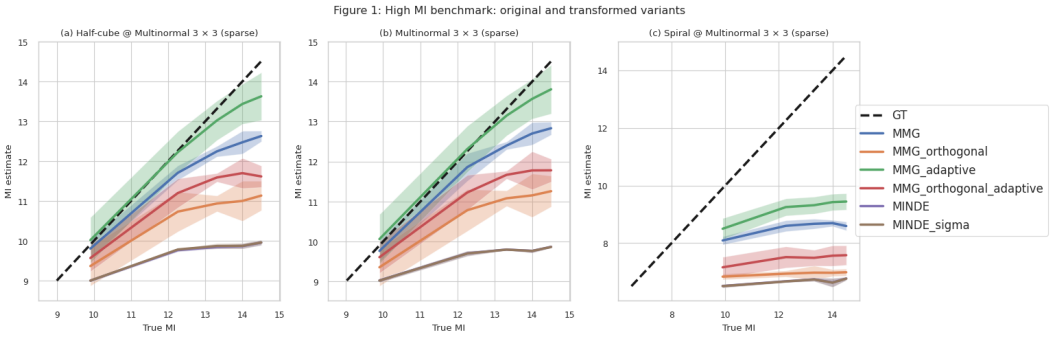
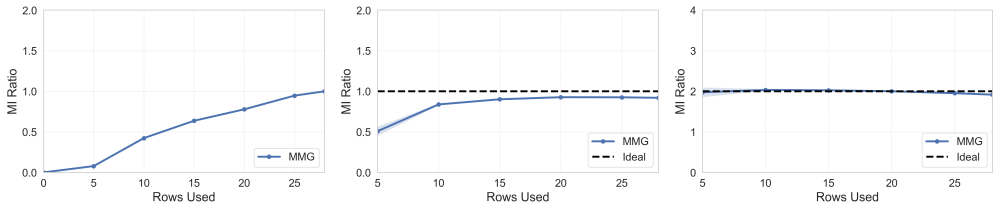


Figure 3: High MI benchmark: original (column (b)) and transformed variants (columns (a) and (c)).

5.3 Consistency Test



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

Figure 4: Consistency Tests over MNIST dataset: (a) evaluation of $\frac{I(A;B_r)}{I(A;B)}$; (b) evaluation of $\frac{I(A;[B_{r+k},B_r])}{I(A;B_{r+k})}$ for $k > 0$; (c) evaluation of $\frac{I([A^1,A^2];[B_r^1,B_r^2])}{I(A^1;B_r^1)}$.

We conduct self-consistency tests inspired by Song & 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.

6 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., 2018b), leverages the Donsker-Varadhan representation of the KL-divergence, spurring related estimators like the variance-reduced SMILE (Song & Ermon, 2019) and DoE (McAllester & 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 logarithm of the batch size (McAllester & 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).

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., 2022). 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.

7 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 or without the orthogonal principle—depends on the MI magnitude of the problem. Future work could explore strategies to automatically navigate this bias-variance trade-off or apply the MMSE-gap framework to other information-theoretic quantities.

References

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 531–540, 2018a.

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*, pp. 531–540. PMLR, 2018b.

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.

Paweł Czyż, Frederic Grabowski, Julia E. Vogt, Niko Beerenwinkel, and Alexander Marx. Beyond normal: On the evaluation of mutual information estimators, 2023. URL `https://arxiv.org/abs/2306.11078`.

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*, pp. 1675–1685. PMLR, 2019.

Giulio Franzese, Mustapha Bounoua, and Pietro Michiardi. Minde: Mutual information neural diffusion estimation. *arXiv preprint arXiv:2310.09031*, 2023.

Giulio Franzese, Mustapha Bounoua, and Pietro Michiardi. Minde: Mutual information neural diffusion estimation, 2024. URL `https://arxiv.org/abs/2310.09031`.

Shuyang Gao, Greg Ver Steeg, and Aram Galstyan. Efficient estimation of mutual information for strongly dependent variables. In *Artificial intelligence and statistics*, pp. 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.

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.

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.

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. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

Xianghao Kong, Rob Brekelmans, and Greg Ver Steeg. Information-theoretic diffusion. *ICLR*, 2022. URL `https://openreview.net/pdf?id=UvmDCdSPDOW`.

Xianghao Kong, Ollie Liu, Han Li, Dani Yogatama, and Greg Ver Steeg. Interpretable diffusion via information decomposition. *arXiv preprint arXiv:2310.07972*, 2023.

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*, pp. 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.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *Advances in neural information processing systems*, 2018.

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*, pp. 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*, pp. 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*, pp. 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.

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. (2023).

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:

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

Below is a brief proof sketch for Equation 13.

Proof Sketch. The proof relies on the law of total expectation and the orthogonality property of MMSE estimation (i.e., the estimation error is orthogonal to any function of the conditioning variables). We begin by showing a key cross-term is zero:

$$\begin{aligned}
\mathbb{E}[(\boldsymbol{x} - \hat{\boldsymbol{x}}_y) \cdot (\hat{\boldsymbol{x}}_y - \hat{\boldsymbol{x}})] &= \mathbb{E}\left[\mathbb{E}[(\boldsymbol{x} - \hat{\boldsymbol{x}}_y) \cdot (\hat{\boldsymbol{x}}_y - \hat{\boldsymbol{x}})|\boldsymbol{z}_\gamma, y]\right] && \text{(Law of Total Expectation)} \\
&= \mathbb{E}\left[(\mathbb{E}[\boldsymbol{x}|\boldsymbol{z}_\gamma, y] - \hat{\boldsymbol{x}}_y) \cdot (\hat{\boldsymbol{x}}_y - \hat{\boldsymbol{x}})\right] && \text{(Pulling out known terms)} \\
&= \mathbb{E}\left[(\hat{\boldsymbol{x}}_y - \hat{\boldsymbol{x}}_y) \cdot (\hat{\boldsymbol{x}}_y - \hat{\boldsymbol{x}})\right] = 0 && \text{(Definition of } \hat{\boldsymbol{x}}_y\text{)}
\end{aligned}$$

Because this cross-term is zero, we can expand the unconditional error $\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2$ as follows:

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2] &= \mathbb{E}[\|(\boldsymbol{x} - \hat{\boldsymbol{x}}_y) + (\hat{\boldsymbol{x}}_y - \hat{\boldsymbol{x}})\|^2] \\
&= \mathbb{E}[\|\boldsymbol{x} - \hat{\boldsymbol{x}}_y\|^2] + \mathbb{E}[\|\hat{\boldsymbol{x}}_y - \hat{\boldsymbol{x}}\|^2] + 2 \cdot \underbrace{\mathbb{E}[(\boldsymbol{x} - \hat{\boldsymbol{x}}_y) \cdot (\hat{\boldsymbol{x}}_y - \hat{\boldsymbol{x}})]}_{0} \\
&= \mathbb{E}[\|\boldsymbol{x} - \hat{\boldsymbol{x}}_y\|^2] + \mathbb{E}[\|\hat{\boldsymbol{x}}_y - \hat{\boldsymbol{x}}\|^2]
\end{aligned}$$

Rearranging the terms yields the identity in Equation 13.

Table 2: MMG network training hyper-parameters. Dim of the task correspond the sum of the two variables dimensions, and d corresponds to the randomization probability.

Benchmark Dim	d	Width	Time Embedding Size	Batch Size	lr	Iterations	# of 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

Sampling Parameter	LogSNR Loc	LogSNR Scale	Clip	EMA Decay	Inference Times	N_Points
-	2.0	3.0	4.0	0.999	10	10000

Table 3: Sampling Hyperparameters

B 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 4.

Table 4: 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

The results in Table 4 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.

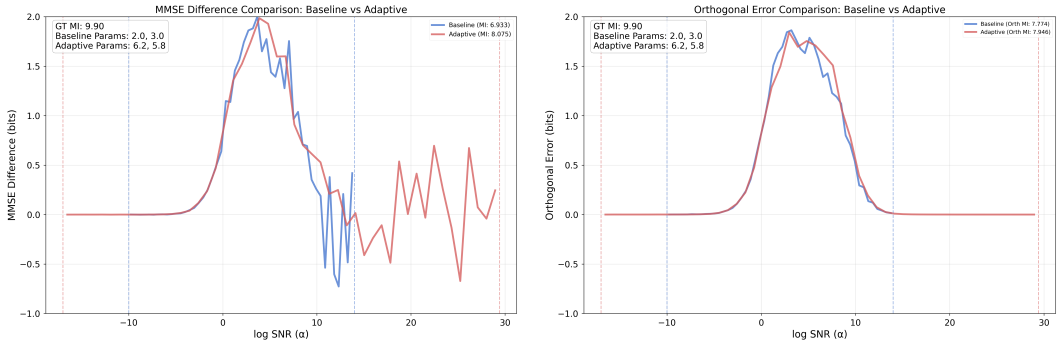
C Implementation Details

We follow the implementation of Franzese et al. (2023) which uses stacked multi-layer perception (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, 2014) 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 2 for MMG. Concerning the consistency tests, we independently train an autoencoder for each version of the MNIST dataset with r rows available.

D 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.



(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.

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.

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 bits** vs. **6.933 bits** for the direct method), confirming the value of focusing the integration where it matters most.