

1 Hyperparameter Sensitivity Analysis

This section details the hyperparameter sensitivity analysis conducted to evaluate the sensitivity of MIRI to various configuration settings.

1.1 Methodology

The sensitivity analysis was performed by systematically varying one hyperparameter at a time while keeping all other parameters fixed to a base configuration. This allows for the isolated assessment of each parameter’s impact on imputation performance. The study was conducted on three synthetic datasets generated using the methods described in the main text: Gaussian, Mixed (Gaussian and Uniform), and Uniform Correlated. Performance was evaluated using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Maximum Mean Discrepancy (MMD) between the imputed values and the ground truth for the initially missing entries. Lower values indicate better performance for all metrics.

1.2 Data Generation

The synthetic datasets used in this study consist of 1000 samples and 20 features. A missing rate of 20% was applied to each dataset. The data types include Gaussian, Uniform Correlated, and Mixed (Gaussian and Uniform) as follows:

- **Gaussian Data:** Multivariate normal distribution $X \sim \mathcal{N}(\mathbf{0}, I)$ where features are independent standard normal variables $X_i \sim \mathcal{N}(0, 1)$.
- **Uniform Correlated Data:** Generated as $X = UL^T$ where $U \sim \text{Uniform}(0, 1)^{N \times D}$ and L is a lower triangular matrix derived from a predefined correlation structure, resulting in uniformly distributed but correlated features.
- **Mixed Data:** First $\lfloor D/2 \rfloor$ features follow standard normal distributions, while the remaining features follow uniform distributions, creating a heterogeneous dataset with mixed distributions.

1.3 Base Configuration

The base configuration used for the sensitivity analysis is as follows:

- Hidden Dimensions: 1024x1024x1024
- Activation Function: SiLU
- Dropout Rate: 0.0
- Learning Rate: 1×10^{-4}
- ODE Solver Steps: 1000
- Automatic Mixed Precision (AMP): False
- Optimizer: Adam
- Gradient Clipping Threshold: 1.0
- Reinitialize Network Per Round: False (i.e., network weights are retained across imputation rounds)

1.4 Results

Tables 1, 2, and 3 present the results for each dataset type. The ‘base config’ row indicates the performance with the default parameters listed above. Each subsequent row shows the result when the specified parameter is changed to the given value, holding all others constant.

Table 1: Sensitivity Analysis Results on Gaussian Synthetic Data

Parameter	Value	RMSE	MAE	MMD
base config	-	1.433	1.142	0.044
activation	gelu	1.417	1.130	0.044
activation	leaky_relu	1.418	1.133	0.044
activation	relu	1.409	1.127	0.044
activation	tanh	1.408	1.123	0.044
amp	True	1.418	1.129	0.044
dropout_rate	0.1	1.426	1.136	0.044
dropout_rate	0.2	1.385	1.102	0.044
gradient_clip	0.0	1.415	1.131	0.044
gradient_clip	0.5	1.408	1.124	0.044
gradient_clip	1.5	1.414	1.124	0.044
gradient_clip	2.0	1.422	1.135	0.044
gradient_clip	5.0	1.413	1.126	0.044
hidden_dims	1024	1.419	1.137	0.044
hidden_dims	1024x1024	1.426	1.138	0.044
hidden_dims	1024x512x1024	1.411	1.124	0.044
hidden_dims	2048x1024x512	1.434	1.142	0.044
hidden_dims	2048x2048	1.448	1.151	0.044
hidden_dims	256x256x256x256	1.415	1.129	0.044
hidden_dims	4096	1.449	1.143	0.044
hidden_dims	512x1024x2048	1.422	1.133	0.044
hidden_dims	512x512	1.455	1.158	0.044
hidden_dims	512x512x512x512	1.418	1.129	0.044
learning_rate	1e-02	1.389	1.112	0.044
learning_rate	1e-03	1.412	1.121	0.044
learning_rate	1e-05	1.411	1.125	0.044
learning_rate	5e-03	1.406	1.120	0.044
learning_rate	5e-04	1.405	1.118	0.044
learning_rate	5e-05	1.422	1.133	0.044
ode_steps	50	1.414	1.126	0.044
ode_steps	100	1.416	1.127	0.044
ode_steps	250	1.402	1.118	0.044
ode_steps	500	1.400	1.112	0.044
ode_steps	750	1.404	1.119	0.044
ode_steps	1500	1.424	1.137	0.044
optimizer	adagrad	1.681	1.336	0.044
optimizer	adamw	1.410	1.125	0.044
optimizer	rmsprop	1.405	1.120	0.044
optimizer	sgd	1.408	1.121	0.044
reinitialize_net_per_round	True	1.489	1.192	0.044

1.4.1 Gaussian Data

Analysis: The model demonstrates robustness to most hyperparameter changes on the Gaussian dataset.

- **Activation:** ReLU and Tanh slightly outperform the base SiLU, while GELU and Leaky ReLU perform similarly.
- **Dropout:** A higher dropout rate (0.2) surprisingly improves performance, suggesting potential overfitting in the base model for this dataset.
- **Learning Rate:** Rates around 5×10^{-4} to 1×10^{-3} show slight improvements. Very high (1×10^{-2}) or very low (1×10^{-5}) rates maintain reasonable performance, though 1×10^{-2} is slightly better than base.
- **ODE Steps:** Fewer steps (250-500) seem slightly beneficial compared to the base 1000 steps.
- **Optimizer:** Adam/AdamW, RMSprop, and SGD perform similarly well and slightly better than the base Adam. Adagrad performs noticeably worse.
- **Hidden Dimensions:** Performance is relatively stable across different architectures, except for very small networks (excluded) or very large single-layer networks (4096). Deeper networks (e.g., 4 layers of 512) perform well.
- **Other:** AMP has minimal impact. Reinitializing the network each round degrades performance. Gradient clipping values between 0.5 and 1.5 yield slight improvements over the base 1.0 or no clipping (0.0).

1.4.2 Mixed Data

Analysis: The mixed dataset shows more sensitivity, particularly to learning rate and optimizer choice.

- **Activation:** GELU performs slightly better than the base SiLU. Leaky ReLU, ReLU, and Tanh perform slightly worse.
- **Learning Rate:** This parameter has a significant impact. The base rate (1×10^{-4}) and lower rates (1×10^{-5} , 5×10^{-5}) perform well. Rates of 5×10^{-4} and higher lead to substantially worse performance, especially 1×10^{-2} and 5×10^{-3} , which drastically increase all error metrics.
- **Optimizer:** Adam/AdamW and RMSprop perform well, close to the base Adam. SGD and Adagrad result in significantly higher errors.
- **Hidden Dimensions:** Larger networks (e.g., 2048x2048, 4096) and specific architectures (2048x1024x512) show improvements over the base 1024x1024x1024. Smaller networks like 512x512 perform worse.
- **Other:** Performance is less sensitive to activation, dropout, ODE steps, AMP, gradient clipping (0.5 slightly better), and reinitialization (True is slightly worse). Dropout shows less benefit than in the Gaussian case.

Table 2: Sensitivity Analysis Results on Mixed Synthetic Data

Parameter	Value	RMSE	MAE	MMD
base config	-	1.040	0.734	0.042
activation	gelu	1.036	0.724	0.042
activation	leaky_relu	1.044	0.731	0.042
activation	relu	1.051	0.739	0.042
activation	tanh	1.055	0.739	0.042
amp	True	1.045	0.730	0.042
dropout_rate	0.1	1.069	0.745	0.042
dropout_rate	0.2	1.048	0.729	0.042
gradient_clip	0.0	1.053	0.735	0.042
gradient_clip	0.5	1.032	0.727	0.042
gradient_clip	1.5	1.048	0.741	0.042
gradient_clip	2.0	1.044	0.732	0.042
gradient_clip	5.0	1.051	0.737	0.042
hidden_dims	1024	1.049	0.737	0.042
hidden_dims	1024x1024	1.045	0.735	0.042
hidden_dims	1024x512x1024	1.045	0.729	0.042
hidden_dims	2048x1024x512	1.029	0.725	0.042
hidden_dims	2048x2048	1.018	0.716	0.042
hidden_dims	256x256x256x256	1.059	0.741	0.042
hidden_dims	4096	1.023	0.716	0.042
hidden_dims	512x1024x2048	1.046	0.737	0.043
hidden_dims	512x512	1.066	0.748	0.042
hidden_dims	512x512x512x512	1.034	0.727	0.042
learning_rate	1e-02	1.281	1.020	0.053
learning_rate	1e-03	1.052	0.740	0.042
learning_rate	1e-05	1.019	0.721	0.042
learning_rate	5e-03	1.298	1.029	0.054
learning_rate	5e-04	1.115	0.829	0.045
learning_rate	5e-05	1.043	0.732	0.042
ode_steps	50	1.042	0.732	0.042
ode_steps	100	1.037	0.724	0.042
ode_steps	250	1.038	0.731	0.042
ode_steps	500	1.049	0.733	0.042
ode_steps	750	1.058	0.741	0.042
ode_steps	1500	1.049	0.738	0.042
optimizer	adagrad	1.248	0.936	0.048
optimizer	adamw	1.042	0.736	0.042
optimizer	rmsprop	1.036	0.730	0.042
optimizer	sgd	1.238	0.977	0.052
reinitialize_net_per_round	True	1.073	0.740	0.043

1.4.3 Uniform Correlated Data

Analysis: Similar to the mixed data, this dataset shows high sensitivity to learning rate and optimizer. MMD values are generally higher than for Gaussian or Mixed data.

- **Activation:** Leaky ReLU provides the best performance, notably improving MMD compared to the base SiLU. Other activations perform similarly to or slightly worse than the base.
- **Learning Rate:** Extremely sensitive. Only the base rate (1×10^{-4}) and nearby lower rates (5×10^{-5} , 1×10^{-5}) or slightly higher (5×10^{-4}) maintain good performance. Rates of 1×10^{-3} and above cause a dramatic increase in all error metrics, especially MMD.
- **Optimizer:** Adam/AdamW perform best. RMSprop is slightly worse. SGD and Adagrad lead to very poor results, particularly in MMD.
- **Dropout:** A rate of 0.2 slightly improves RMSE/MAE and maintains MMD, similar to the Gaussian case.
- **Hidden Dimensions:** A large single layer (4096) performs slightly better than the base. Most other architectures perform similarly or slightly worse.
- **Other:** Performance is relatively insensitive to ODE steps, AMP, gradient clipping, and reinitialization.

1.5 Overall Conclusions

The sensitivity analysis highlights that the model is generally robust, but performance, especially on non-Gaussian or mixed-type data, is highly sensitive to the learning rate and optimizer choice. The base configuration with Adam optimizer and a learning rate of 1×10^{-4} provides a strong starting point. However, for specific data types, adjustments might be beneficial:

- Leaky ReLU activation seems promising for uniform-like data.
- Careful tuning of the learning rate is crucial, especially avoiding rates above 5×10^{-4} for non-Gaussian data.
- Adam/AdamW consistently perform well across datasets, while SGD and Adagrad struggle with more complex distributions.
- Moderate dropout (e.g., 0.2) might be beneficial, potentially dataset-dependent.
- Network architecture variations generally have a smaller impact than learning rate or optimizer, provided the network is sufficiently large.

2 CIFAR-10 Sensitivity Analysis

2.1 Experimental Setup

CIFAR-10 images ($32 \times 32 \times 3$) with 40% Missing Completely At Random (MCAR) were used. To balance computational cost, a subset of 2000 images from the standard CIFAR-10 test set was used for evaluation; this constitutes a potential limitation of the study. We vary activation (SiLU, ReLU, LeakyReLU, Tanh), learning rates (1×10^{-3} , 1×10^{-4} , 1×10^{-6}), and optimizers (Adam, Adagrad). Evaluation metrics:

- FID (Fréchet Inception Distance, lower is better)
- PSNR (Peak Signal-to-Noise Ratio in dB, higher is better)
- SSIM (Structural Similarity Index, higher is better)

Table 3: Sensitivity Analysis Results on Uniform Correlated Synthetic Data

Parameter	Value	RMSE	MAE	MMD
base config	-	0.404	0.326	0.105
activation	gelu	0.409	0.329	0.106
activation	leaky_relu	0.400	0.322	0.100
activation	relu	0.408	0.328	0.104
activation	tanh	0.407	0.328	0.104
amp	True	0.407	0.330	0.108
dropout_rate	0.1	0.414	0.332	0.108
dropout_rate	0.2	0.400	0.322	0.105
gradient_clip	0.0	0.406	0.327	0.105
gradient_clip	0.5	0.410	0.332	0.103
gradient_clip	1.5	0.407	0.327	0.104
gradient_clip	2.0	0.407	0.329	0.108
gradient_clip	5.0	0.413	0.333	0.107
hidden_dims	1024	0.410	0.331	0.108
hidden_dims	1024x1024	0.413	0.333	0.109
hidden_dims	1024x512x1024	0.406	0.329	0.105
hidden_dims	2048x1024x512	0.411	0.329	0.109
hidden_dims	2048x2048	0.407	0.327	0.107
hidden_dims	256x256x256x256	0.409	0.329	0.105
hidden_dims	4096	0.401	0.325	0.103
hidden_dims	512x1024x2048	0.409	0.330	0.107
hidden_dims	512x512	0.413	0.332	0.108
hidden_dims	512x512x512x512	0.406	0.328	0.107
learning_rate	1e-02	1.080	0.856	0.479
learning_rate	1e-03	1.044	0.724	0.431
learning_rate	1e-05	0.409	0.331	0.108
learning_rate	5e-03	1.591	1.325	0.520
learning_rate	5e-04	0.410	0.331	0.107
learning_rate	5e-05	0.409	0.331	0.109
ode_steps	50	0.412	0.333	0.109
ode_steps	100	0.407	0.328	0.103
ode_steps	250	0.413	0.335	0.108
ode_steps	500	0.412	0.332	0.106
ode_steps	750	0.407	0.328	0.105
ode_steps	1500	0.409	0.329	0.109
optimizer	adagrad	1.166	0.809	0.425
optimizer	adamw	0.408	0.329	0.104
optimizer	rmsprop	0.411	0.332	0.109
optimizer	sgd	1.330	1.054	0.503
reinitialize_net_per_round	True	0.412	0.332	0.105

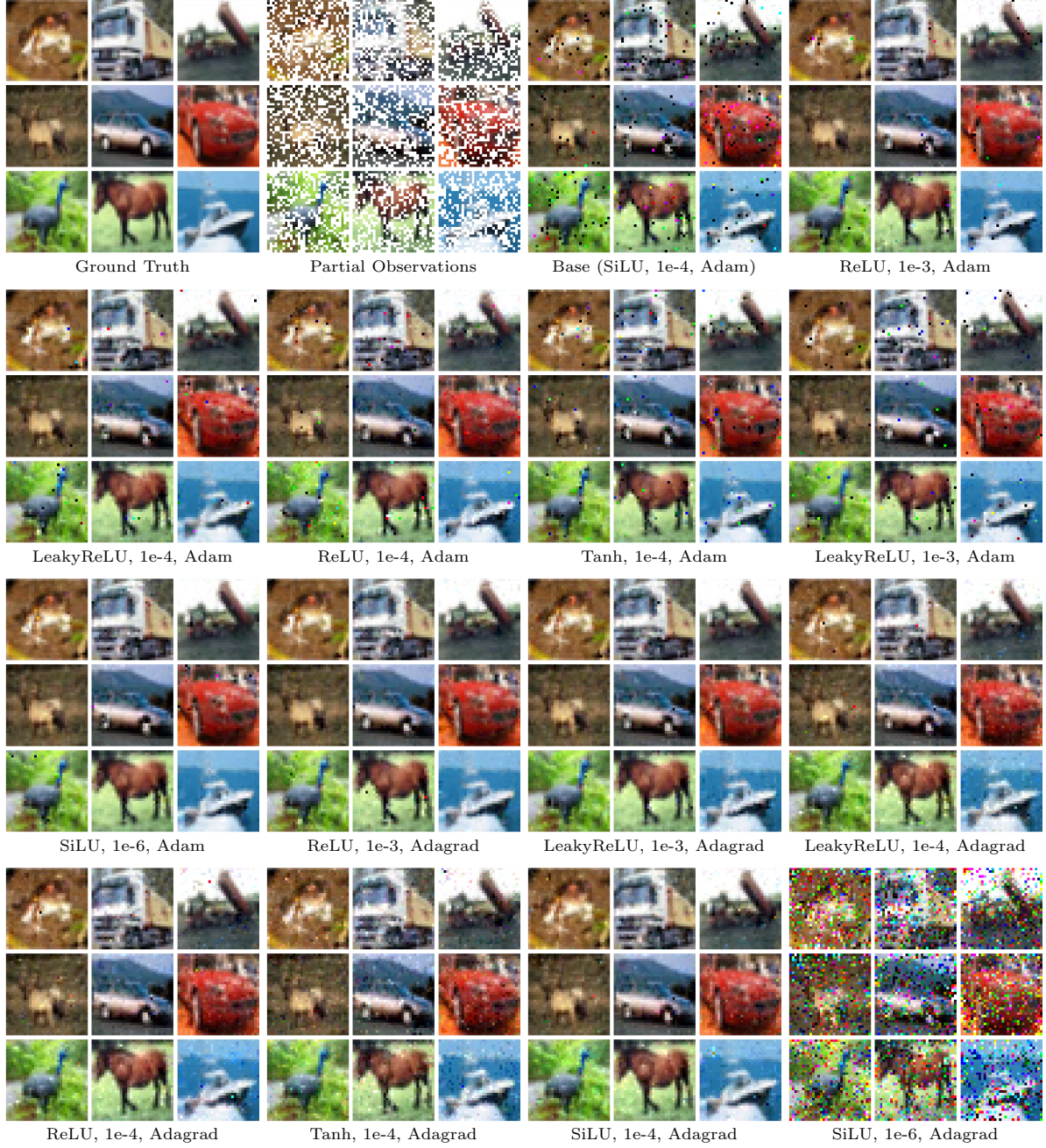


Figure 1: Sample CIFAR-10 images (9 samples per grid). Visual examples are shown for several configurations, including Ground Truth, Partial Observations, Base (SiLU, 1e-4, Adam), and other variants with 40% of missingness.

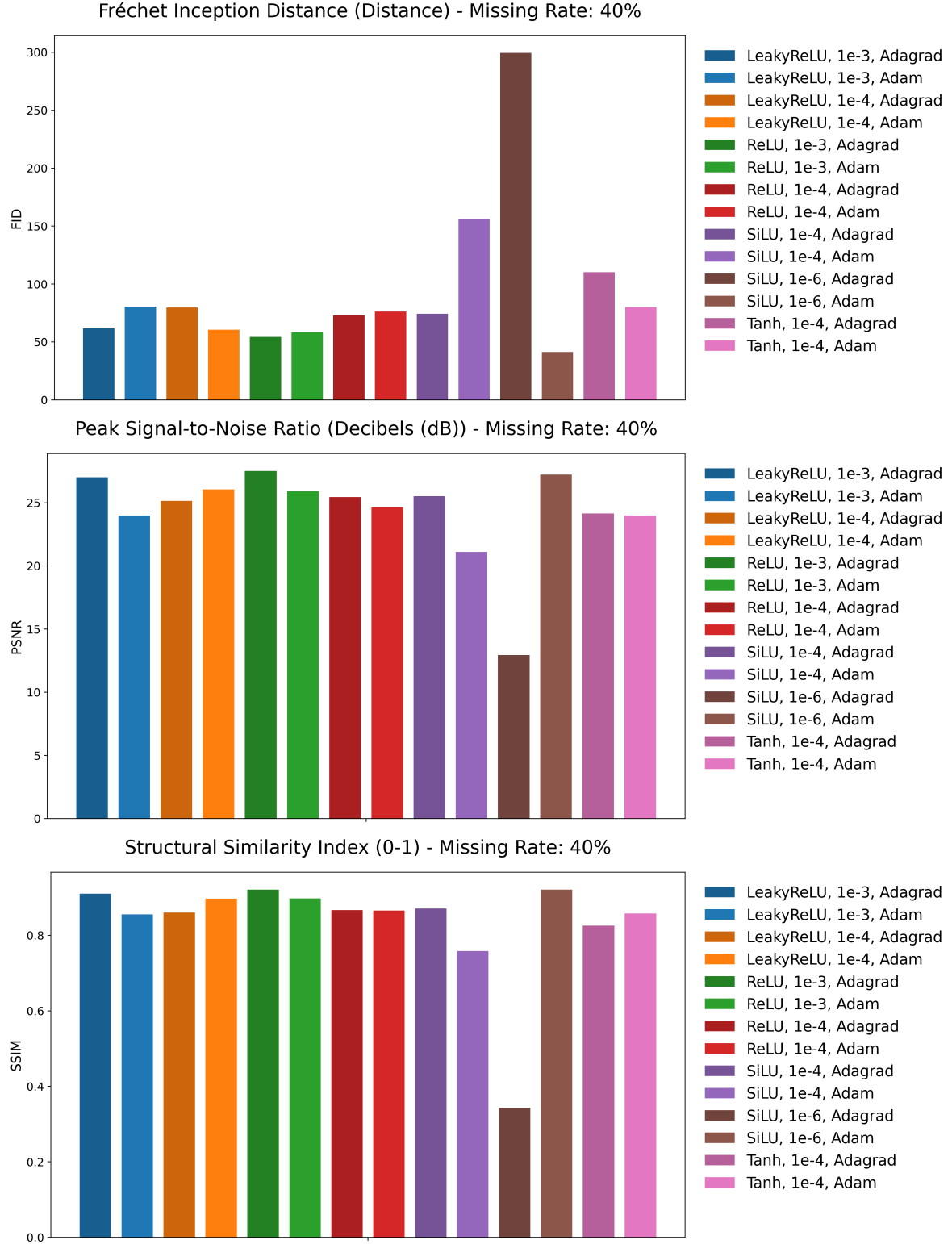


Figure 2: Imputation metrics for CIFAR-10 at 40% of missingness: (top) FID, (middle) PSNR and (bottom) SSIM.

Table 4: Imputation performance on CIFAR-10 at 40% missingness across 11 hyperparameter variants.

Configuration	FID ↓	PSNR ↑	SSIM ↑
SiLU, 1e-6, Adam	41.3386	27.2229	0.9207
ReLU, 1e-3, Adagrad	54.2558	27.5151	0.9211
ReLU, 1e-3, Adam	58.2959	25.9304	0.8979
LeakyReLU, Adam	70.5850	25.5446	0.8808
ReLU, 1e-4, Adagrad	72.9518	25.4514	0.8668
SiLU, 1e-4, Adagrad	74.2793	25.5197	0.8706
ReLU, 1e-4, Adam	76.2365	24.6423	0.8656
Tanh, 1e-4, Adam	80.1353	23.9803	0.8579
Tanh, 1e-4, Adagrad	110.1267	24.1390	0.8257
SiLU, 1e-4, Adam	155.8566	21.1063	0.7582
SiLU, 1e-6, Adagrad	299.3573	12.9417	0.3423

2.2 Results and Analysis

The quantitative results of the CIFAR-10 sensitivity analysis are presented in Table 4. Visual comparisons of imputed samples for all configurations are shown in Figure 1.

Key observations are:

- **Optimizer Impact:** The choice of optimizer is critical, with performance varying between Adam and Adagrad. The (SiLU, 1e-6, Adam) configuration achieved the best FID (41.34). Notably, (ReLU, 1e-3, Adagrad) secured the second-best FID (54.26) and the top PSNR (27.52) and SSIM (0.9211) scores overall. The worst-performing configuration across all metrics remains (SiLU, 1e-6, Adagrad) with an FID of 299.36, indicating a failure to learn meaningful imputations. Out of the 11 configurations, 6 use Adam and 5 use Adagrad. The top 5 configurations by FID include 3 Adam and 2 Adagrad variants, suggesting Adagrad can be competitive with appropriate settings.
- **Activation Function (with Adam):**
 - **SiLU with LR 1e-6:** Achieved the best FID (41.34) and excellent PSNR (27.22, 2nd) and SSIM (0.9207, 2nd).
 - **ReLU with LR 1e-3:** Strong performer with 3rd best FID (58.30), PSNR (25.93), and SSIM (0.8979).
 - **LeakyReLU (unspecified LR in table, assumed 1e-3 or 1e-4):** Good FID (70.59, 4th), PSNR (25.54), and SSIM (0.8808).
 - **ReLU with LR 1e-4:** FID of 76.24 (7th), PSNR (24.64), SSIM (0.8656).
 - **Tanh with LR 1e-4:** FID of 80.14 (8th), PSNR (23.98), SSIM (0.8579).
 - **SiLU with LR 1e-4 (Base):** Poorest Adam performer in this set (FID 155.86, 10th), PSNR (21.11), SSIM (0.7582).
- **Activation Function (with Adagrad):**
 - **ReLU with LR 1e-3:** Exceptional performance, achieving the best PSNR (27.52) and SSIM (0.9211) overall, and the 2nd best FID (54.26).
 - **ReLU with LR 1e-4:** Good FID (72.95, 5th), PSNR (25.45), SSIM (0.8668).
 - **SiLU with LR 1e-4:** Competitive FID (74.28, 6th), PSNR (25.52), SSIM (0.8706).
 - **Tanh with LR 1e-4:** Moderate FID (110.13, 9th), PSNR (24.14), SSIM (0.8257).
 - **SiLU with LR 1e-6:** Worst performer overall (FID 299.36, 11th).
- **Learning Rate Sensitivity:** Performance is highly sensitive to learning rates for both optimizers. With Adam, a very low LR of 1e-6 (for SiLU) yielded the best FID, while 1e-3 (for ReLU) was also strong. The base LR of 1e-4 (for SiLU) was suboptimal among Adam variants. With Adagrad, LR 1e-3 (for ReLU) was optimal for PSNR/SSIM and very good for FID. The 1e-6 LR (for SiLU with Adagrad) was detrimental.

In summary, for CIFAR-10 imputation at 40% missingness, the interplay of optimizer, activation function, and learning rate is crucial and leads to varied outcomes. The (SiLU, 1e-6, Adam) configuration provided the best FID score (41.34), indicating high similarity in feature distributions. However, the (ReLU, 1e-3, Adagrad) configuration achieved the highest PSNR (27.52) and SSIM (0.9211) scores, suggesting better pixel-level accuracy and structural similarity, alongside a very competitive FID (54.26). The base configuration (SiLU, 1e-4, Adam) was significantly outperformed by several other variants. The (SiLU, 1e-6, Adagrad) combination performed very poorly across all metrics, indicating the risk of specific hyperparameter combinations.