# Appendix

## A. Proof of Theorem 1.

*Proof.* Since we implement our ***MosaicDiff*** on well-pretrained diffusion models, we can assume that the distribution of generated images $\hat{x}_0, \hat{x}_t$ is converge to training data $x_0, x_t$:

$$p_\theta(\hat{x}_0) \to q(x_0), \quad p_\theta(\hat{x}_t) \to q(x_t) \tag{1}$$

Thus, we can get similar relation between $\hat{x}_0$ and $\hat{x}_t$:

$$\hat{x}_t(\hat{x}_0, \epsilon) = \sqrt{\bar{\alpha}_t}\hat{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{2}$$

The expectation of MSE can be derived as:

$$\begin{aligned}
\mathbb{E}\Big[\text{MSE}(t)\Big] &= \frac{1}{d}\mathbb{E}\Big[\|\hat{x}_t - \hat{x}_0\|_2^2\Big] \\
&= \frac{1}{d}\mathbb{E}\Big[\|\sqrt{\bar{\alpha}_t}\hat{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon - \hat{x}_0\|_2^2\Big] \\
&= \frac{1}{d}\mathbb{E}\Big[\|(\sqrt{\bar{\alpha}_t}-1)\hat{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon\|_2^2\Big] \\
&= \frac{1}{d}\mathbb{E}\Big[(\sqrt{\bar{\alpha}_t}-1)^2\|\hat{x}_0\|_2^2 + (1-\bar{\alpha}_t)\|\epsilon\|_2^2 \\
&\qquad +2(\sqrt{\bar{\alpha}_t}-1)(1-\bar{\alpha}_t)\hat{x}_0\epsilon\Big],
\end{aligned} \tag{3}$$

Since $\mathbb{E}[\epsilon] = 0, \mathbb{E}[\|\epsilon\|_2^2] = \|\mathbf{I}\|_2^2$:

$$\mathbb{E}\Big[\text{MSE}(t)\Big] = \frac{1}{d}\Big[(1-\sqrt{\bar{\alpha}_t})^2\|\hat{x}_0\|_2^2 + (1-\bar{\alpha}_t)\|\mathbf{I}\|_2^2\Big]. \tag{4}$$

Then, we can calculate gradient $\text{Grad}(t)(t > 0)$ as :

$$\begin{aligned}
\mathbb{E}\Big[\text{Grad}(t)\Big] &= \mathbb{E}\Big[\text{MSE}(t)\Big] - \mathbb{E}\Big[\text{MSE}(t-1)\Big] \\
&= \frac{1}{d}\Big[((\bar{\alpha}_t - \bar{\alpha}_{t-1}) + 2(\sqrt{\bar{\alpha}_{t-1}} - \sqrt{\bar{\alpha}_t}))\|\hat{x}_0\|_2^2 \\
&\qquad -(\bar{\alpha}_t - \bar{\alpha}_{t-1}))\|\mathbf{I}\|_2^2\Big].
\end{aligned} \tag{5}$$

Define $\delta_t := \bar{\alpha}_t - \bar{\alpha}_{t-1}$. Thus,

$$\mathbb{E}\Big[\text{Grad}(t)\Big] = \frac{1}{d}\Big[(\delta_t + 2(\sqrt{\bar{\alpha}_{t-1}} - \sqrt{\bar{\alpha}_t}))\|\hat{x}_0\|_2^2 - \delta_t\|\mathbf{I}\|_2^2\Big]. \tag{6}$$

## B. Additional Experimental Results.

### B.1. Comparison with Small Models Trained from Scratch.

We evaluate our pruned large-scale model in comparison to the smaller DiT-L/2 model [**?** **?** ] trained from scratch, which contains 458 million parameters. Both models are sampled using DDIM with 50 and 20 steps. As shown in Table 1, our pruned model consistently outperforms the smaller DiT-L/2 across all evaluated metrics, including FID, IS, and Precision, while requiring comparable or fewer MACs. This demonstrates that even after pruning, our large-scale model retains significant performance advantages over smaller models trained from scratch, highlighting the effectiveness of our approach in balancing efficiency and generative quality.

Table 1. Comparison between ***MosaicDiff*** at sparsity 0.35 and the smaller DiT-L/2 model trained from scratch.

| Model | Steps | MACs(T) | IS ↑ | FID ↓ | Precision ↑ | Recall ↑ |
|---|---|---|---|---|---|---|
| DiT-L/2 | 50 | 3.88 | 167.6 | 4.82 | 78.72 | 54.66 |
| Ours | 50 | 3.88 | **265.9** | **2.26** | **81.76** | **57.21** |
| DiT-L/2 | 20 | 1.55 | 160.2 | 6.45 | 77.13 | 53.65 |
| Ours | 20 | 1.51 | **264.5** | **3.33** | **80.37** | **53.72** |

### B.2. Sparsity Allocation

We provide the sparsity allocation for each stage and the corresponding performance, as shown in Table 2 and 3. These results demonstrate that our method maintains strong performance even at higher sparsity levels. In Table 11, our approach achieves an FID of 3.65 at 40% sparsity, showing minimal degradation. While extreme pruning (50% sparsity) impacts performance, our method remains effective by strategically allocating sparsity across stages. Table 12 further confirms this trend for SDXL, where our method achieves an FID of 23.79 at 20% sparsity, maintaining competitive quality. Even at 30% sparsity, the model still produces reasonable results. These findings highlight that our method successfully balances compression and generation quality, outperforming conventional pruning techniques, especially at higher sparsity levels.

Table 2. Sparsity allocation of DiT when $M = 0.55$, stage divided at Step $T = 450$ and $T = 900$.

| Sparsity | Stage 1 | Stage 2 | Stage 3 | FID |
|---|---|---|---|---|
| 0.25 | 0.50 | 0.02 | 0.06 | 3.14 |
| 0.30 | 0.60 | 0.04 | 0.10 | 3.20 |
| 0.35 | 0.70 | 0.06 | 0.20 | 3.33 |
| 0.40 | 0.80 | 0.08 | 0.30 | 3.65 |
| 0.45 | 0.90 | 0.10 | 0.40 | 4.33 |
| 0.50 | 0.90 | 0.15 | 0.40 | 5.27 |

Table 3. Sparsity allocation of SDXL when $M = 0.55$, stage divided at Step $T = 250$ and $T = 900$.

| Sparsity | Stage 1 | Stage 2 | Stage 3 | FID |
|---|---|---|---|---|
| 0.10 | 0.30 | 0.03 | 0.15 | 23.18 |
| 0.15 | 0.40 | 0.04 | 0.20 | 23.73 |
| 0.20 | 0.60 | 0.06 | 0.30 | 23.79 |
| 0.30 | 0.80 | 0.08 | 0.40 | 28.37 |

### B.3. Usability on Step-distilled Models

***MosaicDiff*** is fully compatible with step-distilled models. We use SDXL-Turbo, a distilled variant of SDXL-Base-1.0,

for evaluation. Experiments use 4 steps sampling. As in Table 4, with 0.15 average sparsity, ***MosaicDiff*** surpasses vanilla model and uniform pruning by FID margins of 0.85 and 0.69. In contrast, mismatched sparsity patterns degrade performance noticeably, validating our scoring strategy. We also show changes in image MSE over sampling steps, aligning well with the teacher (Figure 1).

Table 4. Performance of ***MosaicDiff*** on step-distilled model SDXL-turbo with 4 steps of sampling.

| Strategy | Sparsity | | | | FID↓ |
|---|---|---|---|---|---|
| | Step 1 | Step 2 | Step 3 | Step 4 | |
| Vanilla SDXL-turbo | 0 | 0 | 0 | 0 | 30.93 |
| Uniform pruning | 0.15 | 0.15 | 0.15 | 0.15 | 30.77 |
| Reverse *MosaicDiff* | 0.05 | 0.1 | 0.3 | 0.15 | 31.86 |
| *MosaicDiff* | 0.3 | 0.15 | 0.05 | 0.1 | **30.08** |

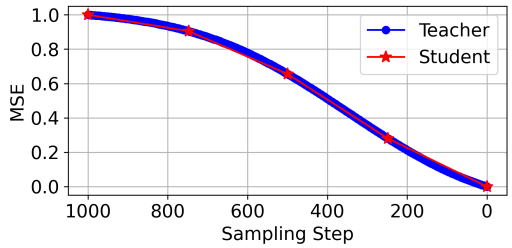



Figure 1. Change in image MSE over sampling steps. Student SDXL-turbo aligns well with teacher SDXL.

### B.4. Relationship between CFG and Sparsity

We observe that as pruning sparsity increases, the optimal CFG required to achieve the best FID also rises. Specifically, as illustrated in Figure 2, the optimal CFG value for the vanilla DiT-XL/2 model is approximately 1.5. At a pruning sparsity of 0.3, the optimal CFG increases to 2.1, and further increases to 3.5 at a sparsity level of 0.45. These results highlight a strong interplay between model compression and guidance strength.

## C. Additional Visualization of *MosaicDiff*

We provide the visualization of MSE and gradient on SDXL, as shown in Figure 3 and 4b. The results are similar as the figure we obtained in the method section.

Moreover, we add more visualization of images generated by MosaicDiff in Figure 5.

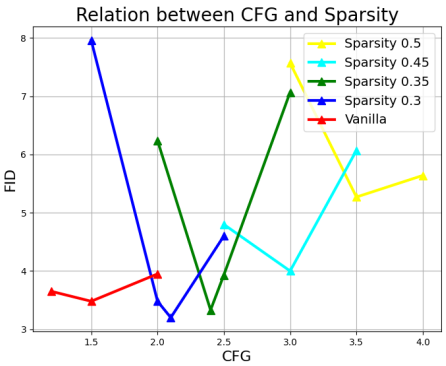



Figure 2. Relationship between CFG and Sparsity.

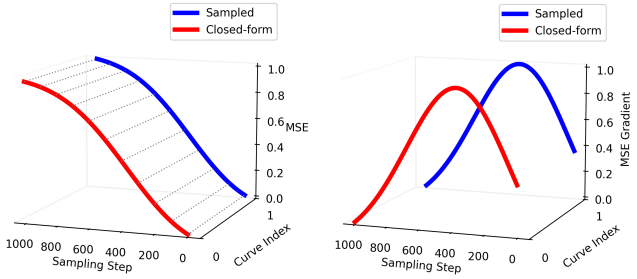
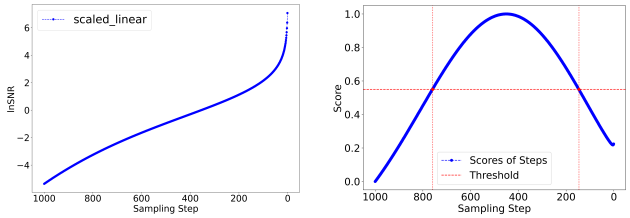



Figure 3. MSE and gradient curves comparison under Scaled-Linear Schedule. *Left*: MSE calculated from our closed-form approximation closely matches the sampled results. *Right*: Gradients derived from our closed-form expression align with empirically sampled gradients.




(a) SNR trend of Linear schedule. (b) Final scores of sampling steps.

Figure 4. Influence of SNR on Final Scores. (a) Change in SNR across sampling steps, showing a sharp increase during the final steps. (b) Final scores computed combining SNR. A threshold of $M = 0.55$ clearly divides the curve into three stages.


Figure 5. Generation Case from MosaicDiff on DiT and SDXL.