Decouple-Then-Merge: Finetune Diffusion Models as Multi-Task Learning

Supplementary Material

1. Proof

In Sec. 1.1 and Sec. 1.2, we demonstrate that DeMe can be formally transformed into a loss reweighting framework, just like previous works [3, 6, 15].

1.1. The Derivation of Some Loss Reweighting Strategies

The standard diffusion loss can be formulated as follows:

$$\mathcal{L}_{\text{standard}} = \mathbb{E}_{t,x_0,\epsilon}\left[\left\|\epsilon - \epsilon_\theta\left(x_t, t\right)\right\|^2\right]. \tag{1}$$

It is worth noting that Equation 1 is identical to $\mathcal{L}_\theta$ in Equation 2 in main paper. For the convenience of subsequent explanations, it has been restated here.

Actually, Equation 1 uses ϵ as the prediction target, but we can equivalently transform it into a loss function where x_0 is the prediction target:

$$\begin{aligned}
\mathcal{L}_{\text{standard}} &= \mathbb{E}_{t,x_0,\epsilon}\left[\left\|\epsilon - \epsilon_\theta(x_t, t)\right\|^2\right] \\
&= \mathbb{E}_{t,x_0,x_t}\left[\left\| \frac{1}{\sqrt{1-\bar{\alpha}_t}}\left(x_t - \sqrt{\bar{\alpha}_t}x_0\right)\right.\right. \\
&\quad \left.\left. - \frac{1}{\sqrt{1-\bar{\alpha}_t}}\left(x_t - \sqrt{\bar{\alpha}_t}x_\theta(x_t,t)\right)\right\|^2\right] \\
&= \mathbb{E}_{t,x_0,x_t}\left[\frac{\bar{\alpha}_t}{1-\bar{\alpha}_t}\left\|x_0 - x_\theta(x_t,t)\right\|^2\right] \\
&= \mathbb{E}_{t,x_0,x_t}\left[\text{SNR}(t)\left\|x_0 - x_\theta(x_t,t)\right\|^2\right],
\end{aligned}$$

where $\text{SNR}(t) = \frac{\bar{\alpha}_t}{1-\bar{\alpha}_t}$. Salimans and Ho [15] propose a loss reweighting strategy named Truncated SNR:

$$\mathcal{L}_{\text{Trun-SNR}} = \mathbb{E}_{t,x_0,x_t}\left[\max\left(\text{SNR}(t), 1\right)\left\|x_0 - x_\theta(x_t,t)\right\|^2\right],$$

which is primarily designed to prevent the weight coefficient from reaching zero as the SNR approaches zero. Additionally, Salimans and Ho [15] propose a new prediction target:

$$v = \sqrt{\bar{\alpha}_t}\epsilon - \sqrt{1-\bar{\alpha}_t}x_0. \tag{2}$$

Similarly, the objective function that uses v as the prediction target can also be equivalently transformed into an objective function where x_0 is the prediction target:

$$\begin{aligned}
\mathcal{L}_{\text{SNR+1}} &= \mathbb{E}_{t,x_0,v}\left[\left\|v - v_\theta(x_t,t)\right\|^2\right] \\
&= \mathbb{E}_{t,x_0,x_t}\left[\left\|\sqrt{\bar{\alpha}_t}\frac{1}{\sqrt{1-\bar{\alpha}_t}}\left(x_t - \sqrt{\bar{\alpha}_t}x_0\right)\right.\right. \\
&\quad - \sqrt{1-\bar{\alpha}_t}x_0 - \left(\sqrt{\bar{\alpha}_t}\frac{1}{\sqrt{1-\bar{\alpha}_t}}\left(x_t - \sqrt{\bar{\alpha}_t}x_\theta(x_t,t)\right)\right. \\
&\quad \left.\left.\left. - \sqrt{1-\bar{\alpha}_t}x_\theta(x_t,t)\right)\right\|^2\right] \\
&= \mathbb{E}_{t,x_0,x_t}\left[\frac{1}{1-\bar{\alpha}_t}\left\|x_0 - x_\theta(x_t,t)\right\|^2\right] \\
&= \mathbb{E}_{t,x_0,x_t}\left[\left(\text{SNR}(t)+1\right)\left\|x_0 - x_\theta(x_t,t)\right\|^2\right].
\end{aligned}$$

Furthermore, a new reweighting strategy [6] has been proposed to achieve accelerated convergence during the training process, named Min-SNR-γ:

$$\mathcal{L}_{\text{Min-SNR-}\gamma} = \mathbb{E}_{t,x_0,x_t}\left[\min\left(\text{SNR}(t), \gamma\right)\left\|x_0 - x_\theta(x_t,t)\right\|^2\right].$$

Additionally, P2 Weighting [3] proposes to assign minimal weights to the unnecessary clean-up stage thereby assigning relatively higher weights to the rest, the weighting term is:

$$\begin{aligned}
\mathcal{L}_{\text{P2}} &= \mathbb{E}_{t,x_0,\epsilon}\left[\frac{1}{\left(k + \text{SNR}(t)\right)^\gamma}\left\|\epsilon - \epsilon_\theta\left(x_t,t\right)\right\|^2\right] \\
&= \mathbb{E}_{t,x_0,x_t}\left[\frac{\text{SNR}(t)}{\left(k + \text{SNR}(t)\right)^\gamma}\left\|x_0 - x_\theta\left(x_t,t\right)\right\|^2\right],
\end{aligned} \tag{3}$$

and the author recommends using $k = 1$ and $\gamma = 1$.

In a word, if the prediction target is x_0, the reweighting strategies can be written as follows:

- Standard diffusion loss [7]:

$$\mathcal{L}_{\text{standard}} = \mathbb{E}_{t,x_0,x_t}\left[\text{SNR}(t)\left\|x_0 - x_\theta(x_t,t)\right\|^2\right] \tag{4}$$

- SNR+1 [15]:

$$\mathcal{L}_{\text{SNR+1}} = \mathbb{E}_{t,x_0,x_t}\left[\left(\text{SNR}(t)+1\right)\left\|x_0 - x_\theta(x_t,t)\right\|^2\right] \tag{5}$$

- Truncated SNR [15]:

$$\mathcal{L}_{\text{Trun-SNR}} = \mathbb{E}_{t,x_0,x_t}\left[\max\left(\text{SNR}(t), 1\right)\left\|x_0 - x_\theta(x_t,t)\right\|^2\right] \tag{6}$$

- Min-SNR-γ [6]:

$$\mathcal{L}_{\text{Min-SNR-}\gamma} = \mathbb{E}_{t,x_0,x_t}\left[\min\left(\text{SNR}(t), \gamma\right)\left\|x_0 - x_\theta(x_t,t)\right\|^2\right] \tag{7}$$

- P2 Weighting [3]:

$$\mathcal{L}_{\text{P2}} = \mathbb{E}_{t,x_0,x_t}\left[\frac{\text{SNR}(t)}{(k+\text{SNR}(t))^{\gamma}}\left\|x_0 - x_\theta\left(x_t, t\right)\right\|^2\right] \tag{8}$$

1.2. Transform DeMe Framework to Loss Reweighting Framework

In Sec. 3.2, we divide the overall timesteps $[0, T)$ into N multiple continuous and non-overlapped timesteps ranges, which can be formulated as $\{{}^{(i-1)T}/_{N}, {}^{iT}/_{N}\}_{i=1}^{N}$. For each range, we finetune a diffusion model ϵ_{θ_i}, the training objective of ϵ_{θ_i} can be formulated as follows:

$$\begin{aligned}\mathcal{L}_i = \mathbb{E}_{t\sim U\left[\frac{(i-1)T}{N}, \frac{iT}{N}\right], x_0, \epsilon} \\ \left[\left\|\epsilon - \epsilon_{\theta_i}\left(x_t, t\right)\right\|^2 + \left\|\epsilon_\theta\left(x_t, t\right) - \epsilon_{\theta_i}\left(x_t, t\right)\right\|^2\right].\end{aligned} \tag{9}$$

In Equation 9, the first term is the standard diffusion loss over the subrange, and the second term is the consistency loss, ensuring that the finetuned model ϵ_{θ_i} stays close to the original model ϵ_θ.

In Sec. 3.3, we compute task vector $\tau_i = \theta_i - \theta$ after finetuning ϵ_{θ_i}, and merge N post-finetuned diffusion models by

$$\theta_{\text{merged}} = \theta + \sum_{i=1}^{N} w_i \tau_i, \tag{10}$$

where w_i are the merging weights determined(via grid search).

The update in parameters τ_i on due to finetuning on timestep range i is:

$$\tau_i = \theta_i - \theta = -\eta \nabla_\theta L_i, \tag{11}$$

where η is the learning rate. The merged model's parameters in Equation 11 could be rewritten as:

$$\theta_{\text{merged}} = \theta - \eta \sum_{i=1}^{N} w_i \nabla_\theta \mathcal{L}_i, \tag{12}$$

which implies θ_{merged} minimizes the combined loss

$$\mathcal{L}_{\text{merged}} = \sum_{i=1}^{N} w_i \mathcal{L}_i. \tag{13}$$

L_i is computed over its respective timestep range, which menas L_{merged} can be viewed as an integration over the entire timestep range with a piecewise constant weighting function $w\left(t\right)$. We rewrite L_{merged} as:

$$\begin{aligned}\mathcal{L}_{\text{merged}} = \mathbb{E}_{t\sim U[0,T], x_0, \epsilon} \\ \left[w(t)\cdot\left\|\epsilon - \epsilon_\theta(x_t,t)\right\|^2 + \left\|\epsilon_\theta(x_t,t) - \epsilon_{\theta_i}(x_t,t)\right\|^2\right],\end{aligned} \tag{14}$$

where

$$w(t) = \begin{cases} w_i, & \text{if } t \in \left[\frac{(i-1)T}{N}, \frac{iT}{N}\right) \\ 0, & \text{otherwise}\end{cases}. \tag{15}$$

In Sec. 3.2, we propose to use θ to initialize θ_i and to utlize consistency loss to unsure $\epsilon_\theta(x_t, t) \approx \epsilon_{\theta_i}(x_t, t)$, which means that the second term in Equation 14 becomes negligible. The merged loss simplifies to

$$\begin{aligned}\mathcal{L}_{\text{merged}} &= \mathbb{E}_{t\sim U[0,T], x_0, \epsilon}\left[w(t)\cdot\left\|\epsilon - \epsilon_\theta(x_t,t)\right\|^2\right] \\ &= \mathbb{E}_{t\sim U[0,T], x_0, \epsilon}\left[w(t)\cdot\text{SNR}(t)\cdot\left\|x_0 - x_\theta(x_t,t)\right\|^2\right].\end{aligned} \tag{16}$$

Equation 16 is exactly the form of a reweighted loss function over timesteps, similar to Equations 4–8.

2. Experimental Details

Implementation Details. During the finetuning process, we set $N = 4$ for all four datasets, $p = 0.4$ for CIFAR10 dataset, $p = 0.3$ for LSUN-Church, LSUN-Bedroom, and L-Aes 6.5+ datasets. For CIFAR10, each model is trained for 20K iterations with a batch size of 64 and a learning rate of 2e-4. For LSUN-Church, LSUN-Bedroom, and L-Aes 6.5+ datasets, each model is trained for 20K iterations with a batch size of 16, a learning rate of 5e-5, and gradient accumulation is set to 4. We employ a 50-step DDIM sampler for DDPM and a 50-step PNDM sampler for Stable Diffusion. For model merging, we use grid search to explore all possible combinations of coefficients. At the same time, we use FID as the objective for the grid search to evaluate the generative quality of the merged model. Considering that the computational cost grows exponentially with N, it is impractical to generate 50K images for each combination of coefficients to compute the FID. Therefore, we generate 5K images to calculate the FID for each merged model to expedite the search process (results reported in the paper are FID-50K). Additionally, we consider variable step size search, where we first use a larger step size to identify a range that may contain the optimal combination of coefficients. Then refine the search within this range using a smaller step size to pinpoint the optimal combination. For drawing Fig. 1a, we compute 1K timesteps' pairwise gradient similarities per 2K training iterations with 512 samples. Batchsize is set to be 128 with 4-step gradient accumulation. A similar implementation is ANT GitHub[1]. All experiments are implemented on NVIDIA A100 80GB PCIe GPU and NVIDIA GeForce RTX 4090.

Dataset Details. For unconditional image generation datasets CIFAR10, LSUN-Church, and LSUN-Bedroom,

[1] https://github.com/gohyojun15/ANT_diffusion

we generated 50K images to obtain the Fréchet Inception Distance (FID)for evaluation. For zero-shot text-to-image generation, we finetune each model on a subset of LAION-Aesthetics V2 (L-Aes) 6.5+, containing 0.22M image-text pairs. We use 30K prompts from the MS-COCO validation set, downsample the 512×512 generated images to 256×256, and compare the generated results with the whole validation set. We also use class names from ImageNet1K and 1.6K prompts from PartiPrompts to generate 2K images (2 images per class for ImageNet1k) and 1.6K images, individually. Fréchet Inception Distance (FID) and CLIP score are used to evaluate the quality of generated images.

3. Related Works on Model Merging

Merging models in parameter space emerged as a trending research field in recent years, aiming at enhancing performance on a single target task via merging multiple task-specific models [9, 13, 16, 18]. In contrast to multi-task learning, model merging fuses model parameters by performing arithmetic operations directly in the parameter space [8, 17], allowing the merged model to retain task-specific knowledge from various tasks. Diffusion Soup [2] suggests the feasibility of model merging in diffusion models by linearly merging diffusion models that are finetuned on different datasets, leading to a mixed-style text-to-image zero-shot generation. MaxFusion [14] fuses multiple diffusion models by merging intermediate features given the same input noisy image. LCSC [12] searches the optimal linear combination for a set of checkpoints in the training process, leading to a considerable training speedups and FID reduction. Unlike Diffusion Soup [2], MaxFusion [14] and LCSC [12], DeMe leverages model merging to *fuse models finetuned at different timesteps*, combines the knowledge acquired at different timesteps, and resulting in improved model performance.

4. Related Works on Timestep-wise Model Ensemble

Previous works [1, 10, 11] have revealed that the performance of diffusion models varies across different timesteps, suggesting that diffusion models may excel at certain timesteps while underperforming at others. Inspired by this observation, several works [1, 10, 20] explore the idea of proposing an ensemble of diffusion experts, each specialized for different timesteps, to achieve better overall performance. MEME [10] propose a multi-architecture and multi-expert diffusion models, which assign distinct architectures to different time-step intervals based on the frequency characteristics observed during the diffusion process. Zhang et al. [20] introduce a multi-stage framework and tailored multi-decoder architectures to enhance the efficiency of diffusion models. eDiff-I [1] propose training an ensemble of expert denoisers, each specialized for different stages of the iterative text-to-image generation process. Spectral Diffusion [19] can also be viewed as an ensemble of experts, each specialized in processing particular frequency components during the iterative image synthesis. Go et al. [5] leverages multiple guidance models, each specialized in handling a specific noise range, called Multi-Experts Strategy. OMS-DPM [11] propose a predictor-based search algorithm that optimizes the model schedule given a set of pretrained diffusion models.

5. Additional Experiment: Comparison with Mixture of Experts Methods

DeMe improves the performance of pretrained diffusion by decoupling the training process and then merging the finetuned models in the parameter space. Notably, although DeMe finetunes multiple models, it ultimately obtains a single model through the model merging method, which is used during the inference stage. *Although not directly related*, we nonetheless compare several timestep-wise model ensemble methods, also referred to as mixture-of-experts methods, for diffusion models, as they share a similar motivation with our approach. Considering the relevance of the experimental settings and the accessibility of the codebase, we compare DeMe with OMS-DPM [10] and DiffPruning [4], highlighting the efficiency and competitive performance of DeMe compared to mixture-of-experts methods. OMS-DPM [4] trains a zoo of models with varying sizes and optimizes a model schedule tailored to a specified computation budget. DiffPruning [4] finetunes pruned diffusion models on different timestep intervals separately to obtain a mixture of efficient experts.

As shown in Table 1, DeMe achieves better performance than other mixture-of-experts methods with only a single model by utilizing the decouple-then-merge mechanism. For example, on the CIFAR-10 dataset, OMS-DPM achieved an FID of 3.80 with a time budget of 9.0×10^3 and a model zoo size of 6, whereas DeMe achieved an FID of 3.51 with only a single model, demonstrating the effectiveness of DeMe. Mixture-of-experts methods tackles different denoising tasks across timesteps during inference by utilizing multiple models. In contrast, DeMe achieves comparable or even better performance while maintaining a single model through its decouple-then-merge mechanism.

6. Similarity Between Task Vectors

In Fig. 1, we analyze the cosine similarity between task vectors across different timestep ranges to explore how multiple finetuned diffusion models can be merged into a unified diffusion model through additive combination. We observe that task vectors from different timestep ranges are generally close to orthogonal, with cosine similarities remaining

Table 1. Comparison results of DeMe *vs.* mixture-of-experts methods for diffusion models. The number in brackets following OMS-DPM [11] means the time budget(ms). Percentage in bracket following DiffPruning [4] means the pruning ratio. #Models means the number of models used in the mixture-of-experts method. #Params refers to the total number of model parameters used during the inference process. †: improved performance to the DDPM model. Mixture-of-experts methods achieve better performance by leveraging the combination of multiple models in different timesteps, whereas DeMe achieves superior performance with only a single model through its decouple-then-merge mechanism.

CIFAR10 (32×32)

Model	#Models	#Params	FID (↓)
DDPM [7]	1	35.75M	4.42
OMS-DPM(9.0×10^3) [11]	6	-	3.80
OMS-DPM(6.0×10^3) [11]	6	-	4.07
OMS-DPM(3.0×10^3) [11]	6	-	5.20
DeMe (Before Merge)	4	36.80M × 4	3.79 $_{(-0.63)}$ †
DeMe (After Merge)	1	36.80M	**3.51** $_{(-0.91)}$ †

(a)

LSUN-Church (256×256)

Model	#Models	#Params	FID (↓)
DDPM [7]	1	113.67M	10.69
OMS-DPM(55×10^3) [11]	6	-	10.95
OMS-DPM(25×10^3) [11]	6	-	11.10
OMS-DPM(10×10^3) [11]	6	-	13.70
DiffPruning (70%) [4]	2	188.09M	9.39
DiffPruning (50%) [4]	2	112.60M	10.89
DeMe (Before Merge)	4	115.31M × 4	9.57 $_{(-1.12)}$ †
DeMe (After Merge)	1	115.31M	**7.27** $_{(-3.42)}$ †

(b)

LSUN-Bedroom (256×256)

Model	#Models	#Params	FID (↓)
DDPM [7]	1	113.67M	6.46
DiffPruning (70%) [4]	2	162.06M	5.90
DiffPruning (50%) [4]	2	100.87M	6.73
DeMe (Before Merge)	4	115.31M × 4	5.87 $_{(-0.59)}$ †
DeMe (After Merge)	1	115.31M	**5.84** $_{(-0.62)}$ †

(c)

low, often near zero. We speculate that this orthogonality facilitates the additive merging of multiple finetuned diffusion models into a unified model with minimal interference, allowing for effective combination without conflicting gradients between the different timestep ranges. For instance, timestep ranges $t \in [0, 250)$ and $t \in [500, 750)$ on LSUN-Church exhibit a cosine similarity of 0.07, this relatively low value indicates that the task vectors for these two non-adjacent ranges are close to orthogonal, allowing for more effective combination during model merging with minimal interference between different denoising tasks.

Additionally, the slight deviations from orthogonality within different timestep ranges suggest some shared information between neighboring denoising tasks, reflecting a degree of continuity in the model's learning across these ranges. These deviations also highlight the effectiveness of the *Probabilistic Sampling Strategy* introduced in Sec. **??**, which ensures a balance between specialization in the range and generalization across all timesteps, effectively preserving knowledge across different stages of denoising task training.

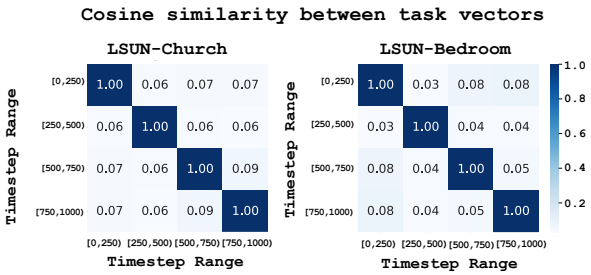


Figure 1. The cosine similarity between task vectors at different timestep ranges on two datasets: **Task vectors are nearly orthogonal between different timestep ranges**. This orthogonality suggests that knowledge from different timesteps is largely independent, allowing for effective additive combination of task vectors with minimal interference, thereby facilitating the merging of finetuned models.

7. Sensitive Study

DeMe decouples the training of diffusion models by finetuning multiple diffusion models in N different timestep ranges. A larger N indicates that the timesteps are divided into finer ranges, further reducing gradient conflicts and potentially enhancing the model's performance. Meanwhile, the probability p determines the tradeoff between learning from specific and global timesteps, thereby influencing the model's performance. Therefore, we do some sensitive study on the influence of *number of ranges N* and *possibility p* on CIFAR10.

Influence on Number of ranges N. A larger N implies each diffusion model is finetuned on a narrower timestep range, leading to less gradient conflicts. As illustrated in Fig. 2, it is observed that: (i) Training diffusion model across the entire timestep range results in the poorest performance. With $N = 1$, i.e., training diffusion on the overall timesteps, a minor improvement is achieved, with a FID of 4.34. We posit that severe gradient conflicts occurred, negatively impacting the overall training process. (ii) The finer the division of the overall timesteps into N non-overlapping ranges, the more effectively it mitigates gradient conflicts, leading to a notable reduction in FID. For example, dividing the timesteps into 4 ranges can result in a 0.63 FID reduction, whereas dividing them into only 2 ranges leads to

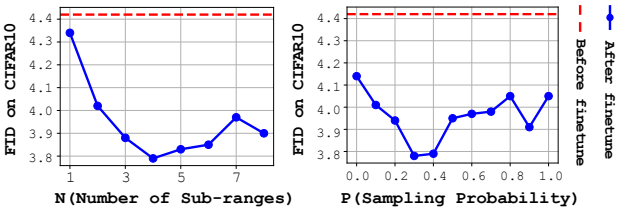


Figure 2. Sensitive study of the influence on the *number of ranges* N and *possibility* p of training of all timesteps on CIFAR10

a reduction of just 0.4 FID. A larger N is associated with improved model performance, indicating reduced gradient conflicts. (iii) As N increases, the model's training exhibits marginal utility. For instance, when N exceeds 4, the FID no longer follows the decreasing trend observed when N smaller than 4. This suggests that the model's gains notably diminish as N increases. Considering the finetuning overhead and the complexity of model merging, we recommend $N = 4$ as a trade-off in practice.

Influence on Probability p. Probability p means a sampling probability p for a diffusion model beyond its specific timespte range, which indicates a trade-off between specific knowledge and general knowledge. Varying choices of probability p can enhance model performance to different extents. As shown in Fig. 2, it is observed that: (i) Training solely on either the full timestep range or specific subranges limits knowledge sharing, resulting in only minor improvements. $P = 0$ corresponds to training across all timesteps, while $p = 1$ focuses exclusively on a specific range. Both of these settings restrict knowledge transfer between the overall and specific timestep ranges, leading to modest FID reductions of 0.28 and 0.37, respectively. (ii) Our method achieves varying degrees of improvement across the range $p \in [0, 1]$. When $p > 0.5$, sampling occurs more frequently over the overall timestep, while for $p < 0.3$, sampling is more concentrated in a specific timestep range. Both cases restrict knowledge transfer between the overall and specific timestep ranges, leading to minor FID improvements, shown in Fig. 2. To maximize the effectiveness of the method, we recommend using $p = 0.3$ or $p = 0.4$ in practice.

8. Additional Qualitative Experiments

8.1. Additional Qualitative Results on LSUN for DDPM

In Fig. 3, Fig. 4 and Fig. 5, generated images of LSUN are presented. DeMe has more **effectively captured the underlying patterns in the images, specifically the church and bedroom scenes**, allowing for more detailed and accurate generation of these structures. While diffusion before finetuning fails to generate churches or bedrooms, diffusion after finetuning successfully generates them with finer details. The finetuned diffusion demonstrates an improved ability to generate coherent and realistic representations of the target objects, as evidenced by the success in producing church-style buildings and bedroom-style interiors.

8.2. Additional qualitative Results for Stable Diffusion

In Fig. 6, Fig. 7 and Fig. 8, additional qualitative results are presented based on various detailed text prompts. DeMe more **effectively generates images that align with the provided text descriptions**, producing results that are both more detailed and photorealistic. The finetuned Stable Diffusion model demonstrates an improved ability to generate visually coherent and contextually accurate images that closely match the nuances of the prompts, as highlighted in the comparison between before- and after-finetuning results, showcasing its enhanced capacity for text-to-image synthesis.

References

[1] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 3

[2] Benjamin Biggs, Arjun Seshadri, Yang Zou, Achin Jain, Aditya Golatkar, Yusheng Xie, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Diffusion soup: Model merging for text-to-image diffusion models. *arXiv preprint arXiv:2406.08431*, 2024. 3

[3] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11472–11481, 2022. 1, 2

[4] Alireza Ganjdanesh, Yan Kang, Yuchen Liu, Richard Zhang, Zhe Lin, and Heng Huang. Mixture of efficient diffusion experts through automatic interval and sub-network selection. *arXiv preprint arXiv:2409.15557*, 2024. 3, 4

[5] Hyojun Go, Yunsung Lee, Jin-Young Kim, Seunghyun Lee, Myeongho Jeong, Hyun Seung Lee, and Seungtaek Choi. Towards practical plug-and-play diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1962–1971, 2023. 3

[6] Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng, and Baining Guo. Efficient diffusion training via min-snr weighting strategy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7441–7451, 2023. 1

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 4

Figure 3. Qualitative comparison between our method and original DDPM on LSUN.

Figure 4. Additional qualitative results on LSUN-Church. The top row shows images generated by DDPM before finetuning, while the bottom row displays images generated by DDPM after finetuning using our training framework. In the bottom row, church-style buildings are successfully generated, whereas the top row fails to produce similar structures.

[8] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022. 3

[9] Xisen Jin, Xiang Ren, Daniel Preotiuc-Pietro, and Pengxiang Cheng. Dataless knowledge fusion by merging weights of language models. *arXiv preprint arXiv:2212.09849*, 2022. 3

[10] Yunsung Lee, JinYoung Kim, Hyojun Go, Myeongho Jeong, Shinhyeok Oh, and Seungtaek Choi. Multi-architecture multi-expert diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13427–13436, 2024. 3

[11] Enshu Liu, Xuefei Ning, Zinan Lin, Huazhong Yang, and Yu Wang. Oms-dpm: Optimizing the model schedule for diffusion probabilistic models. In *International Conference on Machine Learning*, pages 21915–21936. PMLR, 2023. 3, 4

[12] Enshu Liu, Junyi Zhu, Zinan Lin, Xuefei Ning, Matthew B Blaschko, Sergey Yekhanin, Shengen Yan, Guohao Dai, Huazhong Yang, and Yu Wang. Linear combination of saved checkpoints makes consistency and diffusion models better. *arXiv preprint arXiv:2404.02241*, 2024. 3

[13] Michael S Matena and Colin A Raffel. Merging models with fisher-weighted averaging. *Advances in Neural Information Processing Systems*, 35:17703–17716, 2022. 3

[14] Nithin Gopalakrishnan Nair, Jeya Maria Jose Valanarasu, and Vishal M Patel. Maxfusion: Plug&play multi-modal generation in text-to-image diffusion models. *arXiv preprint arXiv:2404.09977*, 2024. 3

[15] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022. 1

[16] Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos,

Figure 5. Additional qualitative results on LSUN-Bedroom. The top row shows images generated by DDPM before finetuning, while the bottom row displays images generated by DDPM after finetuning using our training framework. In the bottom row, bedroom scenes are successfully generated, whereas the top row fails to produce coherent structures.

Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pages 23965–23998. PMLR, 2022. 3

[17] Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. *Advances in Neural Information Processing Systems*, 36, 2024. 3

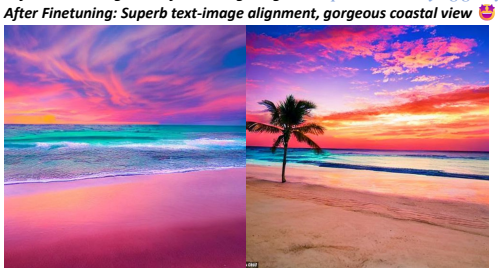
[18] Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. Model merging in llms, mllms, and beyond: Methods, theories, applications and opportunities. *arXiv preprint arXiv:2408.07666*, 2024. 3

[19] Xingyi Yang, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Diffusion probabilistic model made slim. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22552–22562, 2023. 3

[20] Huijie Zhang, Yifu Lu, Ismail Alkhouri, Saiprasad Ravishankar, Dogyoon Song, and Qing Qu. Improving efficiency of diffusion models via multi-stage framework and tailored multi-decoder architectures. *arXiv preprint arXiv:2312.09181*, 2023. 3

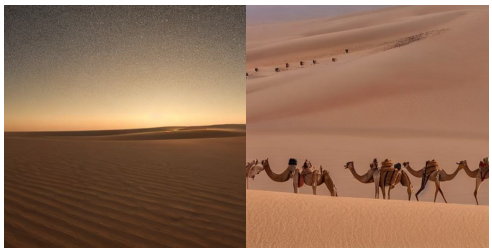
Prompt V: "*A tranquil beach at sunrise, with soft waves lapping at the shore, palm trees swaying gently in the breeze, and a vibrant sky painted in shades of pink, orange, and purple.***"**

Before Finetuning: Loss of text-image alignment: *palm trees swaying gently* 😞
After Finetuning: Superb text-image alignment, gorgeous coastal view 🤩



Prompt VI: "*A vast desert with rolling sand dunes, the golden sands stretching endlessly under a cloudless sky, a caravan of camels making its way across the horizon.***"**

Before Finetuning: Loss of text-image alignment: *a caravan of camels* 😞
After Finetune: Superb text-image alignment 🤩



Prompt VII: "*A herd of wild horses galloping across a wide open field, their manes and tails flowing in the wind, dust kicking up beneath their hooves as they run.***"**

Before Finetuning: Loss of text-image alignment: *dust kicking up..* 😞
After Finetuning: Superb text-image alignment, vivid graphic depiction 🤩

Before Finetuning **After Finetuning**

Prompt VIII: "*A scientist in a modern laboratory, wearing a white coat and safety goggles, examining a vial of glowing blue liquid with curiosity and focus.***"**

Before Finetuning: Loss details in scientist's body 😞
After Finetuning: Detailed generation in scientist's body 🤩

Before Finetuning **After Finetuning**

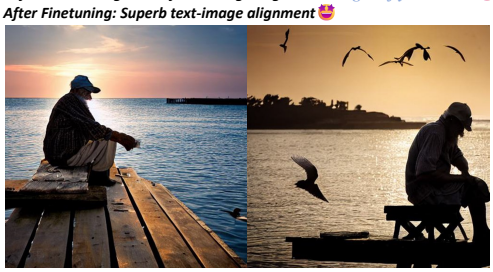
Figure 6. Additional qualitative results based on various text prompts for Stable Diffusion

Prompt IX: "*A majestic eagle soaring high above the mountains, its wings spread wide, gliding effortlessly through the sky with the distant peaks below.***"**

Before Finetuning: Loss of realistic graphic depiction 😞
After Finetuning: Superb text-image alignment, lifelike figure 🤩

Prompt X: "*An old fisherman with weathered hands, sitting on a wooden dock by the sea, seagulls fly overhead in the golden light of sunset.***"**

Before Finetuning: Loss of text-image alignment: *seagulls fly overhead* 😞
After Finetuning: Superb text-image alignment 🤩



Prompt XI: "*A dense, mystical forest in autumn, with rays of golden sunlight filtering through the vibrant orange and red leaves, a narrow path leading deeper into the trees.***"**

Before Finetuning: Loss of text-image alignment: *sunlight filtering through..* 😞
After Finetuning: Superb text-image alignment, photorealistic 🤩

Before Finetuning **After Finetuning**

Prompt XII: "*A snowy winter landscape, a small village nestled in a valley, smoke rising from chimneys, snow-covered trees, and twinkling lights glowing warmly in the dusk.***"**

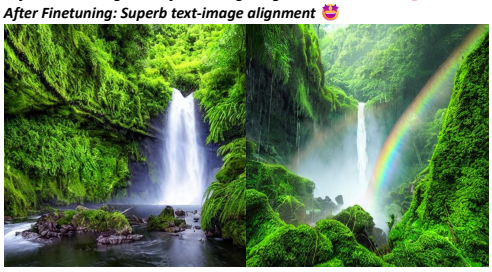
Before Finetuning : Loss of text-image alignment: *smoke .. chimneys* 😞
After Finetuning: Superb text-image alignment 🤩

Before Finetuning **After Finetuning**

Figure 7. Additional qualitative results based on various text prompts for Stable Diffusion

Prompt XIII: "*A breathtaking view of a waterfall cascading down into a lush jungle, with vibrant green foliage, moss-covered rocks, and* ***a rainbow forming in the mist****.***"**

Before Finetuning: Loss of text-image alignment: *a rainbow* 😞
After Finetuning: Superb text-image alignment 😍



Prompt XIV: "*A grand library with towering bookshelves filled with ancient tomes, a spiral staircase winding up to a high ceiling adorned with intricate frescoes, and soft light streaming through stained glass windows.***"**

Before Finetuning: Loss details: twisted staircase 😞
After Finetuning: Excellent details generation 😍

Prompt XV: "*A majestic lion* ***standing proudly*** *on a rocky ledge, its golden mane blowing in the wind, the vast African savannah stretching out behind it at sunset****.***"**

Before Finetuning: Loss of text-image alignment: *standing proudly..* 😞
After Finetuning: Superb text-image alignment 😍

Prompt XVI: "*A curious red fox* ***sitting in a snowy forest****, its bright fur contrasting against the white snow, its ears perked up and eyes focused on something in the distance****.***"**

Before Finetuning : Loss of text-image alignment: *sitting* 😞
After Finetuning: Superb text-image alignment 😍

Before Finetuning **After Finetuning** **Before Finetuning** **After Finetuning**

Figure 8. Additional qualitative results based on various text prompts for Stable Diffusion