LIFT and PLACE: A Simple, Stable, and Effective Knowledge Distillation Framework for Lightweight Diffusion Models

Supplementary Material

Table 6. All of the above models use a fixed size of the student model (1.3M). For OutKD and OutKD+FeatKD, larger teachers result in higher FID means and larger FID variances. When distilled from the strongest teacher, LIFT and PLACE achieve the lowest FID mean and standard deviation. Values in **bold** correspond to those reported in Tab. 1.

Method	Params	Teacher FID	Try 1	Try 2	Try 3	Try 4	Try 5	FID Mean	FID Std
OutKD	78.7M	6.48	210.09	56.67	90.72	**55.41**	70.33	96.64	64.99
	19.7M	5.30	48.84	88.92	128.96	50.85	73.84	78.28	32.87
	16.6M	5.83	35.67	57.66	100.70	59.01	81.37	66.88	24.87
	9.2M	6.32	41.22	50.48	82.36	91.14	40.25	61.09	23.96
OutKD+FeatKD	78.7M	6.48	218.48	204.46	**211.23**	102.11	231.54	193.56	52.10
	19.7M	5.30	53.68	40.72	38.41	42.29	36.94	42.41	6.63
	16.6M	5.83	43.34	34.20	45.88	39.27	46.73	41.89	5.18
	9.2M	6.32	44.76	37.82	43.94	38.44	39.45	40.88	3.23
Ours	78.7M	6.48	19.65	15.80	18.09	15.85	**15.73**	17.03	1.77
	19.7M	5.30	21.69	25.95	24.84	31.72	22.70	25.38	3.93
	16.6M	5.83	23.50	20.87	21.09	29.01	23.24	23.54	3.28
	9.2M	6.32	25.20	29.40	20.05	23.56	22.58	24.16	3.48

8. Experimental Results of Figure 1

Fig. 1 illustrates the challenge of the teacher–student capacity gap in KD for lightweight diffusion models. We fix a 90%-pruned 1.3M-student and distill it from four teachers of varying capacities (78.7M, 19.7M, 16.6M, and 9.2M), evaluating two KD objectives under fixed hyperparameters. Each setting is run five times, and Tab. 6 reports the mean and standard deviation of FID. Conventional KD becomes increasingly unstable as teacher capacity grows: both $\mathcal{L}_{\text{OutKD}}$ and $\mathcal{L}_{\text{OutKD}} + \mathcal{L}_{\text{FeatKD}}$ show worse mean FID and higher variance with larger teachers. The 78.7M-teacher gives the worst results, with most FeatKD runs collapsing.

These results suggest that the degradation of conventional KD is driven more by the teacher–student capacity gap than by the teacher's own generative quality. The difference in teacher FID is relatively small (6.48 vs. 5.30), yet the resulting student performance can collapse dramatically to FIDs of 90–200+, which is difficult to explain solely by teacher quality. Moreover, teacher FID does not reliably predict KD success: OutKD degrades more severely with the better-performing 19.7M teacher than with the 9.2M teacher, and collapses entirely with the 78.7M teacher. Similar trends are also observed in Tab. 7.

In contrast, when OutKD is replaced with LIFT and PLACE under the same setting, the strongest teacher achieves the best mean FID and the smallest standard deviation across five runs. Notably, if the 78.7M teacher provided intrinsically poor supervision, all methods should degrade similarly. Instead, our method remains stable and achieves the best mean/std with this teacher (17.03/1.77), indicating that the teacher signal itself remains useful, while naive KD becomes unstable under a large capacity gap. Overall, these results show that LIFT and PLACE mitigate the optimization difficulty caused by large capacity gaps and enable lightweight students to benefit from stronger teachers.

9. Experiments Details

Across all experiments, we fix PLACE's group size to $K{=}16$, as determined by our ablation study (see Fig. 7). For image space diffusion models, we use the Diff-Pruning base pruned model with varying pruning ratios, where the pruning ratio denotes the fraction of teacher channels removed. We set $\lambda_{\text{diff}}{=}1$ and $\lambda_{\text{FeatKD}}{=}1\text{e}{-6}$ for all such experiments. Our main comparison is between the baseline output-level KD (OutKD) and our LIFT loss; by default, we assign equal weights, $\lambda_{\text{OutKD}}{=}1.0$ and $\lambda_{\text{LIFT}}{=}1.0$. An exception arises on CelebA with 50% and 70% pruning, where the student trained without KD already outperforms the teacher. For these two settings only, we reduce the guidance weight for both methods to 0.1 (i.e., $\lambda_{\text{OutKD}}{=}0.1$ and

Table 7. Effects of teacher capacity in DiT. We compare the TinyFusion baseline with our method when distilling a DiT-D7 student from two teachers of different capacities. As in pixel-space diffusion, larger teachers degrade FID for TinyFusion, whereas our LIFT and PLACE provide consistent improvements, with a slightly larger gain when distilling from the stronger teacher. **Bold** indicates the best performance for each teacher used in distillation, and † denotes metrics reported in [6].

Data	Architecture	Iters	MACs	Params	Method	FID↓	IS↑	Precision↑	Recall↑
ImageNet (256×256)	DiT-D14	500K	59.4G	340.5M	Teacher	2.86	234.5	0.82	0.55
	DiT-D7	500k	29.7G	173.1M	TinyFusion†	5.87	166.9	0.78	0.53
					TinyFusion	5.99	164.0	0.78	0.53
					Ours (K=16)	**5.91**	**164.2**	0.78	0.53
	DiT-XL/2	7000K	118.7G	675.1M	Teacher	2.27	278.2	0.83	0.57
	DiT-D7	500k	29.7G	173.1M	TinyFusion	7.87	145.5	0.75	0.54
					Ours (K=16)	**7.66**	**146.9**	0.75	0.53

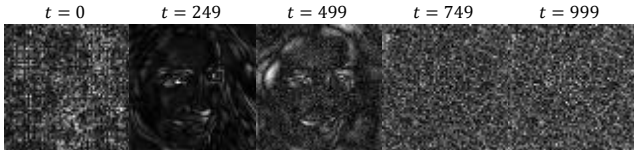
Table 8. Ablation study on the components of Eq. (8). All of student models distilled from the strongest 78.7M-teacher.

Components of Eq. (8)			Student (1.3M)
LIFT	PLACE	FeatKD	FID↓
✓	✓	✓	**15.73**
✓		✓	18.49
✓	✓		16.67
✓			19.07

λ_{LIFT}=0.1) and relax $\mathcal{L}_{\text{coarse}}$ to use an ℓ_2 distance. For latent space diffusion models, we follow the BK-SDM and TinyFusion baselines and simply replace their output-level distillation loss $\mathcal{L}_{\text{OutKD}}$ with our $\mathcal{L}_{\text{LIFT}}$. All other hyperparameters, including their respective λ weights, are kept identical to the original papers for a fair comparison. This also includes adopting TinyFusion's schedule, which anneals the feature-level KD weight $\lambda_{\text{MaskedFeatKD}}$ to 0 during training.

10. Non-uniform Error Across Architectures

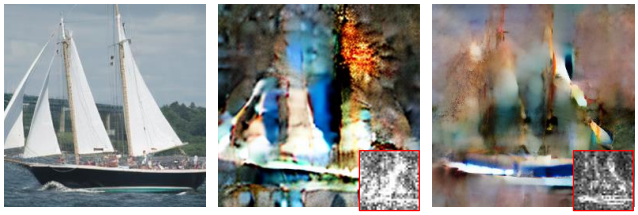
In Sec. 3.2 of the main paper, we showed that the distillation error between teacher and student is spatially non-uniform and exhibits a highly structured pattern that correlates with semantic content. Here, we examine whether this phenomenon persists across diffusion models with substantially different architectures. Fig. 8 presents the student-teacher error map for a pixel space diffusion model. Consistent with Fig. 3, the distillation error is not random or uniformly distributed in other models. Fig. 8 shows that the distillation error concentrates on semantically meaningful regions such as facial features at the middle time step. This confirms that spatially non-uniform distillation difficulty is not specific to latent space models, but also arises in pixel space diffusion models. We further provide the error map of TinyFusion. As shown in Fig. 9, the DiT model is consistent with Sec. 3.2. The error map again forms a distinct, structured pattern that aligns with semantic content.



(a) Error map of OutKD

(b) Error map of OutKD+FeatKD

Figure 8. Error map of lightweight student models after being trained by using conventional KD: (a) using only $\mathcal{L}_{\text{OutKD}}$, (b) using $\mathcal{L}_{\text{OutKD}} + \mathcal{L}_{\text{FeatKD}}$. We observe that the distillation error is non-uniform across diffusion time steps.



(a) Real image (b) Early iteration, 5k (c) After training, 50k

Figure 9. Visualization of error map of TinyFusion (i.e., DiT-D7): (a) real image, (b) early iteration, and (c) after training. Consistent with Fig. 3, the DiT architecture also exhibits spatially non-uniform distillation error.

11. Additional Ablation Studies

We provide detailed ablation studies to validate our LIFT and PLACE. All student models in these experiments are distilled from the strongest 78.7M-teacher model. While the CelebA experiments in Tab. 1 (i.e., 19.7M- and 1.3M-student models) reveal that OutKD+FeatKD occasionally leads to performance degradation compared to FeatKD alone, the results in Tab. 8 demonstrate that LIFT effec-

tively stabilizes this combination. Consistent with prior literature [6, 13, 15], LIFT ensures robust performance gains. Notably, the multi-set coefficient estimation via PLACE enhances the reliability of LIFT's optimization, resulting in a substantial performance gain in FID from 19.07 to 16.67.

11.1. Training Overhead

During training, error-magnitude sorting in PLACE introduces a resolution-dependent cost. However, as shown in Tab. 9, training throughput (iter/s) and VRAM usage on an Intel Xeon 6740P CPU and an NVIDIA RTX PRO 6000 GPU show that this overhead is negligible, since the overall cost is dominated by the model's forward and backward passes. At inference time, our method incurs no additional cost, consistent with prior works [5, 6, 13].

Table 9. Training overhead and peak VRAM usage evaluated on an Intel Xeon 6740P CPU and an NVIDIA RTX PRO 6000 GPU.

Dataset (Resolution)	OutKD+FeatKD		Ours (K=16)	
	Speed (iter/s)	VRAM (GB)	Speed (iter/s)	VRAM (GB)
LSUN (256×256)	4.8935 iter/s	5.615 GB	4.8634 iter/s	5.625 GB
LSUN (512×512)	3.6337 iter/s	18.287 GB	3.6291 iter/s	18.326 GB

12. Related Works

12.1. Efficient Diffusion Model

Although diffusion models [11, 25, 27, 32] demonstrate outstanding performance, their inherent iterative denoising process not only demands substantial computational resources but also makes it challenging to apply existing compression methods designed for feed-forward networks [2, 16, 21, 37, 38]. To address these problems, compression methods were explored that take the unique characteristics of diffusion models. For fast sampling, some prior work [7, 14, 29, 33, 34, 39, 40] studied KD to reduce the number of denoising steps. However, this method did not consider compressing the network architecture. In contrast, pruning has been considered for compressing diffusion models. For instance, Diff-pruning [5] used Taylor scores to assign weight importance for pruning, accounting for varying noise levels across diffusion steps. TinyFusion [6] proposed block pruning and feature-level distillation by learning pruning masks and masking outliers of intermediate features. Moreover, KD for compressed diffusion models, such as random-conditioning [15] and DKDM [36] were proposed for data-efficient KD. In this work, we focus on KD itself for lightweight diffusion model.

12.2. Capacity Gap in Knowledge Distillation

Knowledge distillation, which trains a student model to mimic the teacher model's outputs and intermediate features, is a widely used approach for model compression across various tasks. By distilling the knowledge, the student model can learn the teacher's complex knowledge with a smaller model capacity. In various domains, the capacity gap between student and teacher models hinders the transfer of complex knowledge [1, 12, 24, 31, 35]. To address this problem, DIST [12] relaxed the matching objective by maximizing the Pearson correlation between teacher and student outputs, while TAKD [24] employed a teaching assistant with intermediate capacity to progressively bridge the gap and transfer knowledge across multiple stages.

13. Visualization Results

We provide visualization results of our experiments. The following figures present representative samples for image and latent space diffusion models. Each Figs. 10 to 12 correspond to Tabs. 1 to 3, respectively. The results show that across all model sizes (see Fig. 10), our method produces noticeably more stable and realistic samples than OutKD+FeatKD. Fig. 11 shows that our method captures not only the main subject but also details (e.g., background) compared to OutKD+FeatKD. As shown in Tab. 2, the overall outputs are similar to TinyFusion, while our method sometimes captures finer details more effectively.

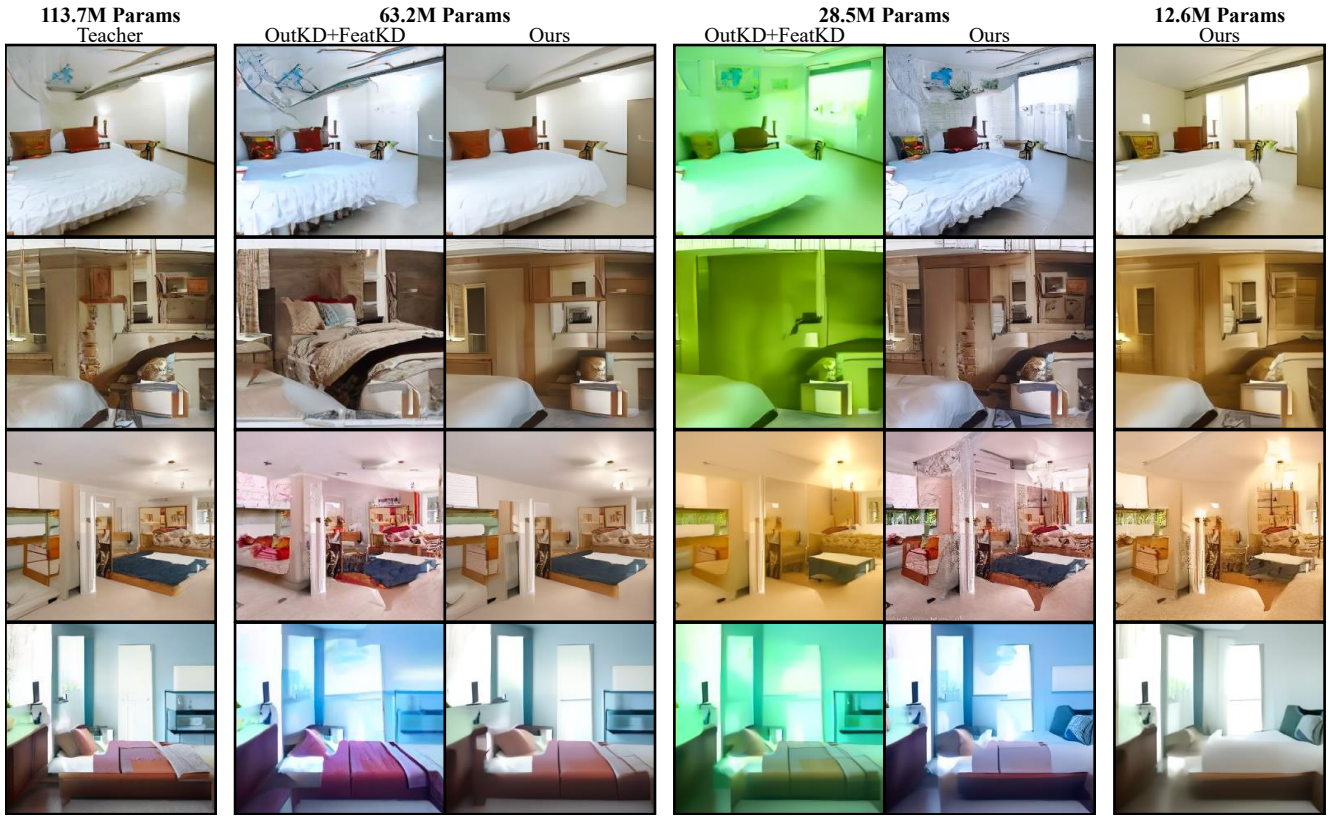


Figure 10. Visualization of pixel space diffusion models with LSUN Bedroom.

Figure 11. Visualization of pruned Stable Diffusion 2.1.

(a) TinyFusion DiT-D14.

(b) DiT-D14 with our method.

(c) TinyFusion DiT-D7.

(d) DiT-D7 with our method.

Figure 12. Visualization results for DiT-D14 and DiT-D7. The top row compares DiT-D14, and the bottom row compares DiT-D7.