Multi-Student Diffusion Distillation for Better One-step Generators

Yanke Song, Jonathan Lorraine, Weili Nie, Karsten Kreis, James Lucas

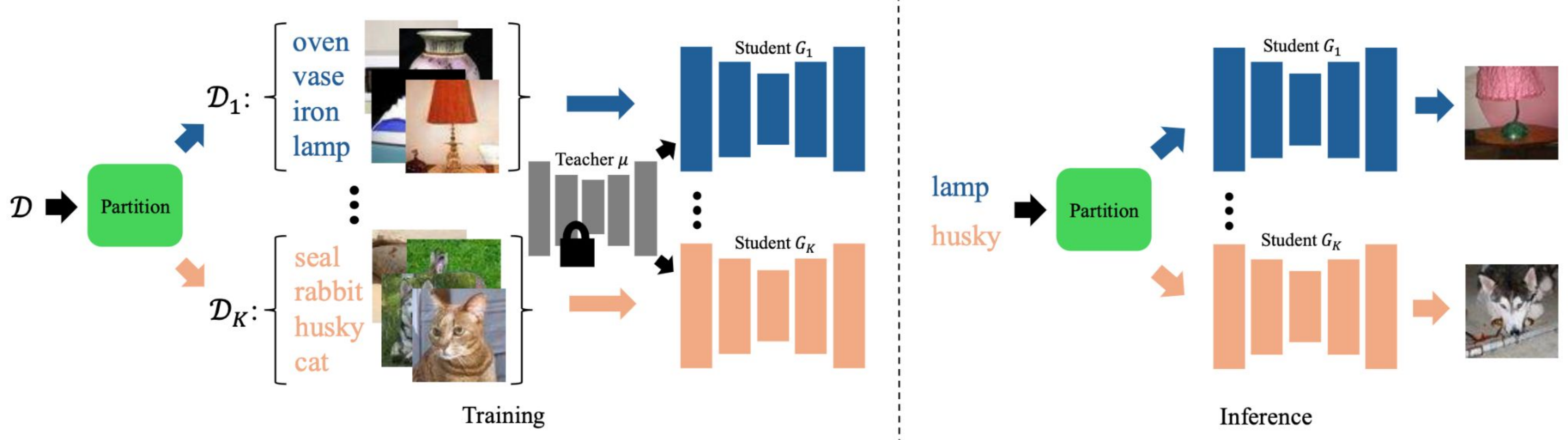
Introduction

- **Goal:** We seek real-time high-quality AR video generation from video-diffusion backbones.
- **Problem:** Even distilling diffusion models to 1-step is not fast enough - we need smaller students that maintain high-quality.
- **Our Solution:** Multi-Student Distillation (MSD), distills diffusion models into 1-step students, for:
  - Improved quality by specializing in subsets
  - Improved latency by distilling to small students
- Our strategy circumvents the capacity-latency tradeoff of existing diffusion distillation.

Our Method - Overview

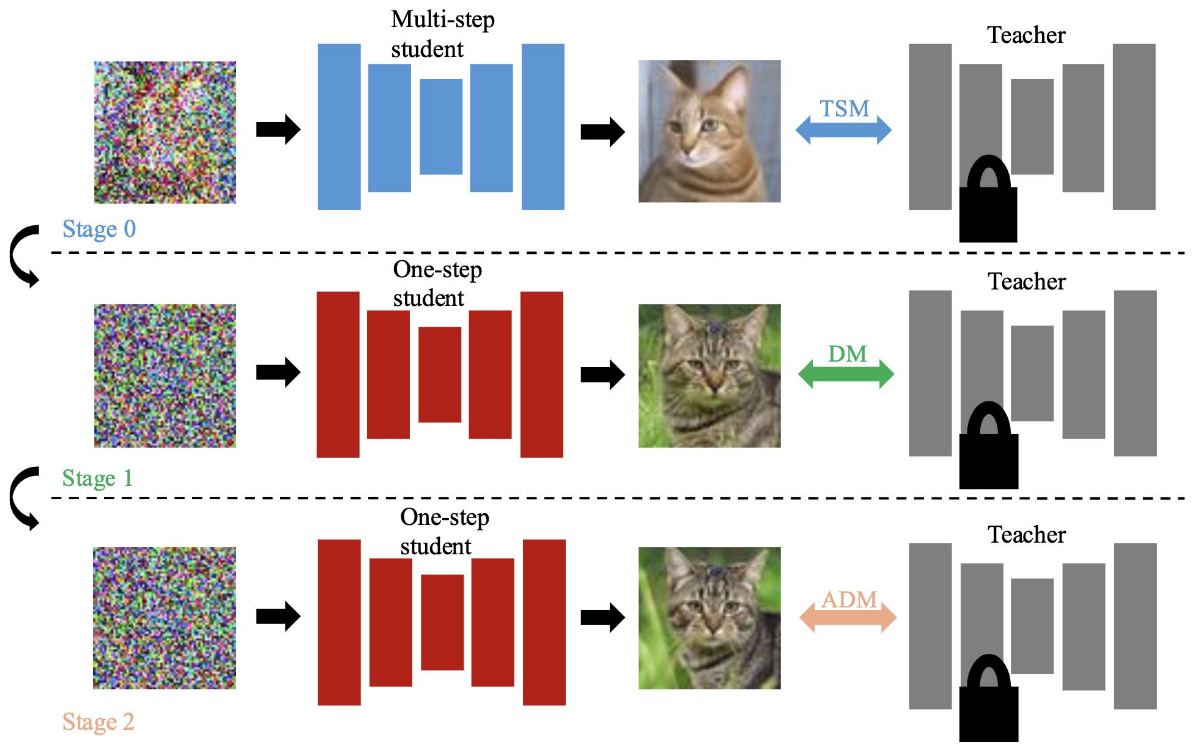
- Recent work [1-3] distills models to as few as 1-step.
- We build on these for generation quality and speed by:

1. During training, MSD partitions the dataset and assigns them to different students.
2. During inference, MSD uses only one student.



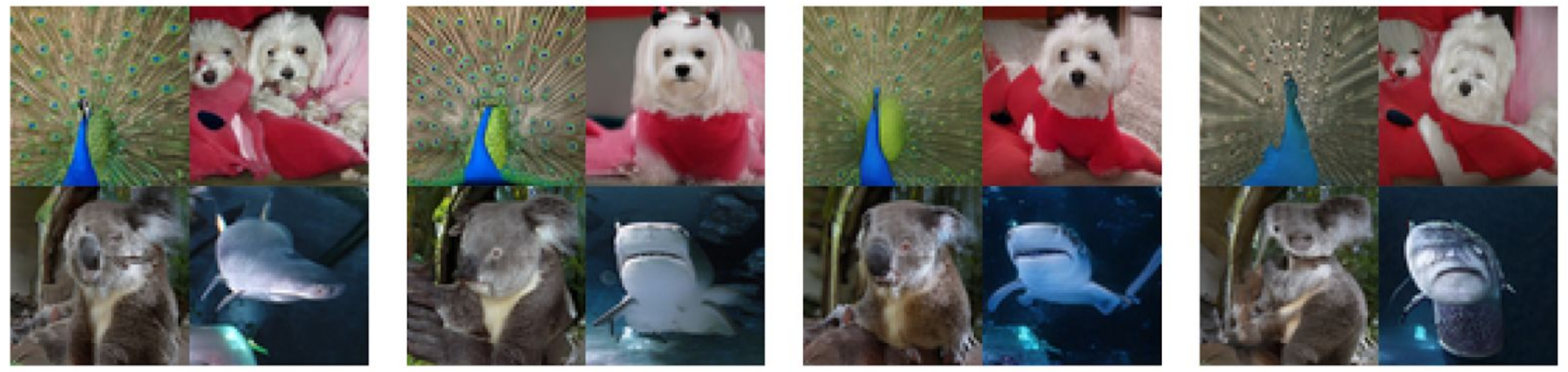
Our Method - Using Small Students

- Existing distillations do not work for smaller students.
- We use a teacher score matching (TSM) stage for this.
- TSM trains small students to emulate teacher scores.



Experimental Results

(a) Teacher (multistep) (b) Same-sized student (c) 83% smaller student



(a) Teacher (multistep) (b) Same-sized students (c) 42% smaller students (d) 71% smaller students

Experimental Results

- MSD boosts FID on class-conditional ImageNet-64x64 and text-to-image zero-shot COCO2014 generation, improving on single-student counterparts using only 4 students.

Method	NFE (↓)	FID (↓)
Multiple Steps		
RIN [23]	1000	1.23
ADM [12]	250	2.07
DPM Solver [39]	10	7.93
Multistep CD [17]	2	2.0
Single Step, w/o GAN		
PD [55]	1	15.39
DSNO [83]	1	7.83
Diff-Instruct [43]	1	5.57
iCT-deep [63]	1	3.25
Moment Matching [56]	1	3.0
DMD [76]	1	2.62
MSD (ours): 4 students, DM only	1	2.37
EMD [70]	1	2.20
SiD [88]	1	1.52
Single Step, w/ GAN		
Post-distillation, 4, 42% smaller students	1	11.67
MSD (ours): 4, 42% smaller students, ADM	1	2.88
StyleGAN-XL [57]	1	1.52
CTM [27]	1	1.92
DMD2 [75]	1	1.28
MSD (ours): 4 students, ADM	1	1.20
SiDA [87]	1	1.11
teacher		
EDM (teacher, ODE) [25]	511	2.32
EDM (teacher, SDE) [25]	511	1.36

Method	Latency (↓)	FID (↓)
Unaccelerated		
DALL·E 2 [52]	-	10.39
LDM [54]	3.7s	12.63
eDiff-I [3]	32.0s	6.95
GANs		
StyleGAN-T [58]	0.10s	12.90
GigaGAN [77]	0.13s	9.09
Accelerated		
DPM++ (4 step) [40]	0.26s	22.36
InstaFlow-0.9B [37]	0.09s	12.10
UFOGen [72]	0.09s	12.78
DMD [76]	0.09s	11.49
EMD [70]	0.09s	9.66
DMD2 (w/o GAN)	0.09s	9.28
MSD (ours): 4 students, DM only	0.09s	8.80
DMD2 [75]	0.09s	8.35
MSD (ours): 4 students, ADM	0.09s	8.20
SiD-LSG [86]	0.09s	8.15
teacher		
SDv1.5 (50 step, CFG=3, ODE)	2.59s	8.59
SDv1.5 (200 step, CFG=2, SDE)	10.25s	7.21

- We provide ablations and more results in the paper.

Future Work

- Limitations to improve upon:
  - More sophisticated routing schemes between students for quality
  - Better methods to reduce student size for latency and quality
  - Weight-sharing or hierarchical branching strategies among students for training efficiency

Links

- Webpage in QR-code: research.nvidia.com/labs/toronto-ai/MSD/

- [1] Yin, Tianwei, et al. "One-step diffusion with distribution matching distillation. 2024 IEEE." CVPR 2023.
- [2] Song, Yang, et al. "Consistency models." ICML 2023
- [3] Yin, Tianwei, et al. "Improved distribution matching distillation for fast image synthesis." NeurIPS 2024