SwiftBrush🎨: One-Step Text-to-Image Diffusion Model with Variational Score Distillation
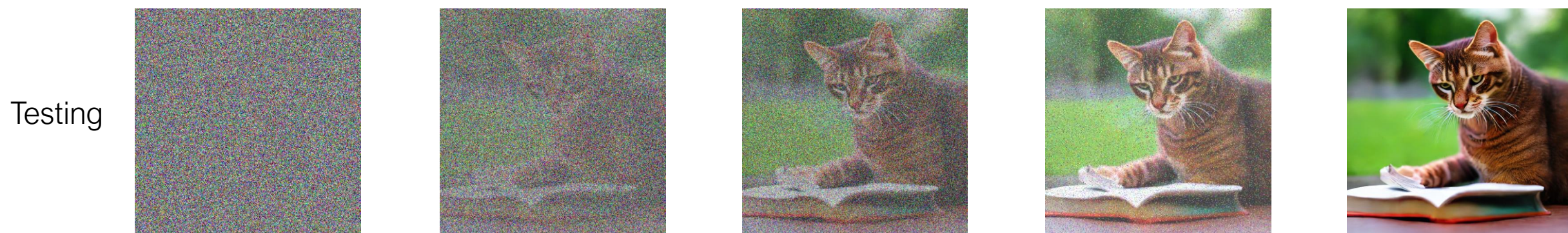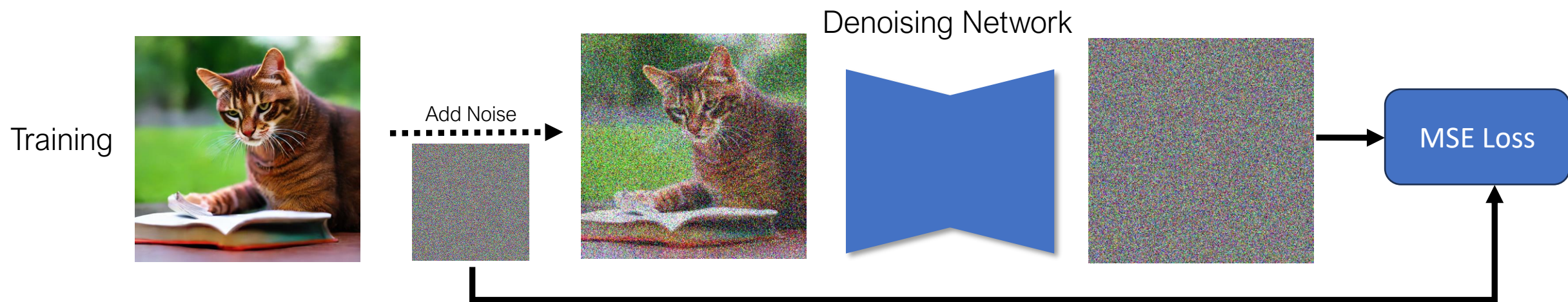
Thuan Hoang Nguyen
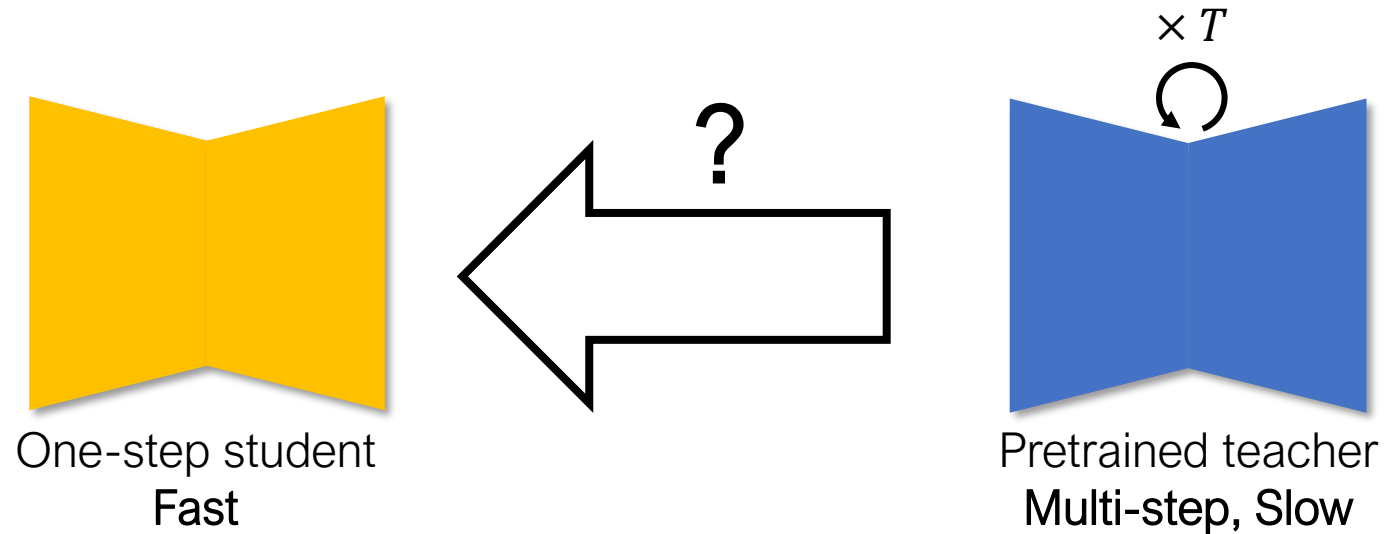
Anh Tran

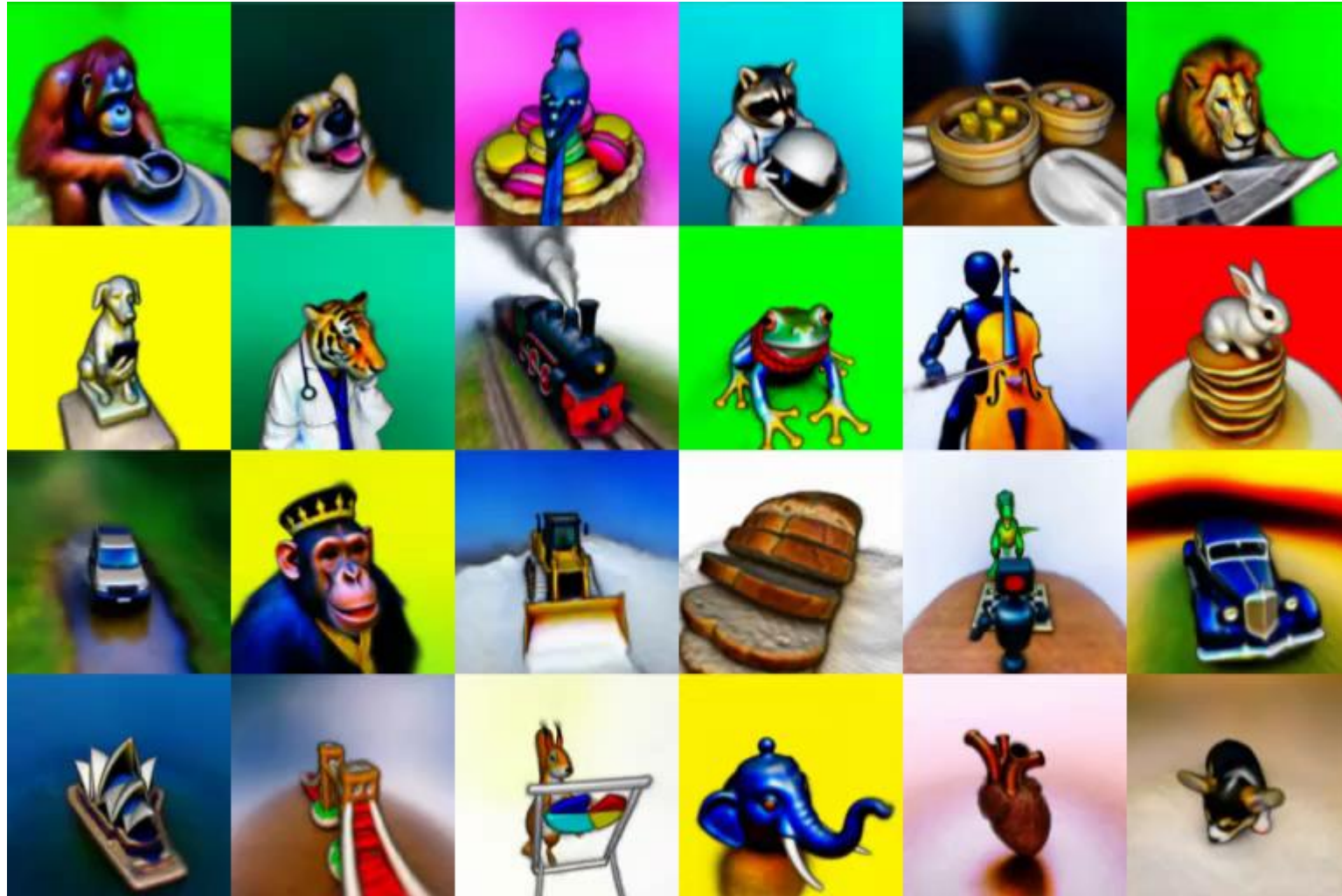# Diffusion Model Basics

Denoising Network

Training

Add Noise

MSE Loss

Testing

Problem Statement

One-step student
Fast

$\times T$

Pretrained teacher
Multi-step, Slow

How do we distill diffusion model into a
one-step generator?

Motivation

Poole, Ben, et al. "DreamFusion: Text-to-3D using 2D Diffusion." ICLR 2023

Score Distillation Sampling (SDS)

SDS Loss

3D NeRF

+Noise

Multi-step pretrained 2D prior
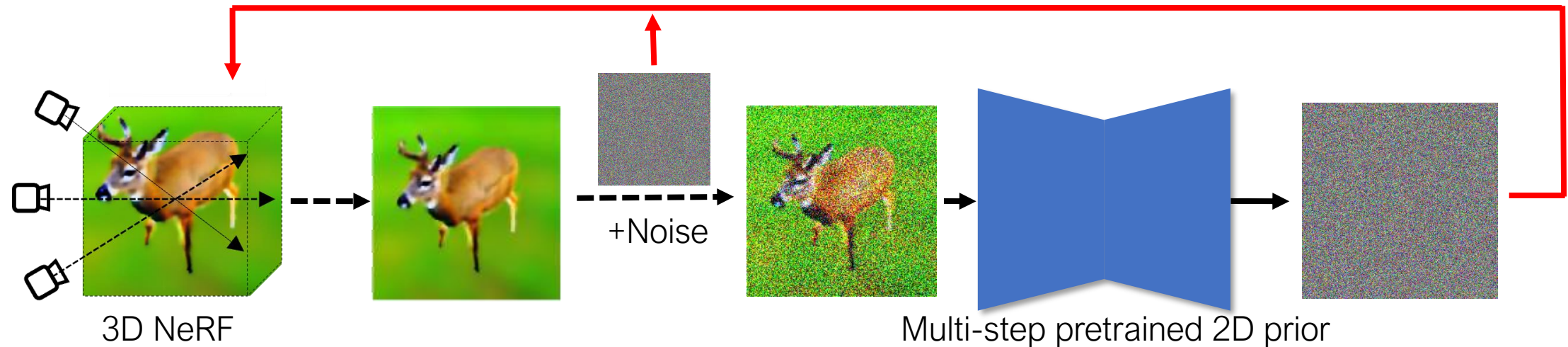
Text-to-3D techniques like SDS can produce 3D NeRF without 3D ground truth!

Poole, Ben, et al. "DreamFusion: Text-to-3D using 2D Diffusion." ICLR 2023

Score Distillation Sampling (SDS)

SDS Loss

One-step student   +Noise   Multi-step pretrained 2D prior
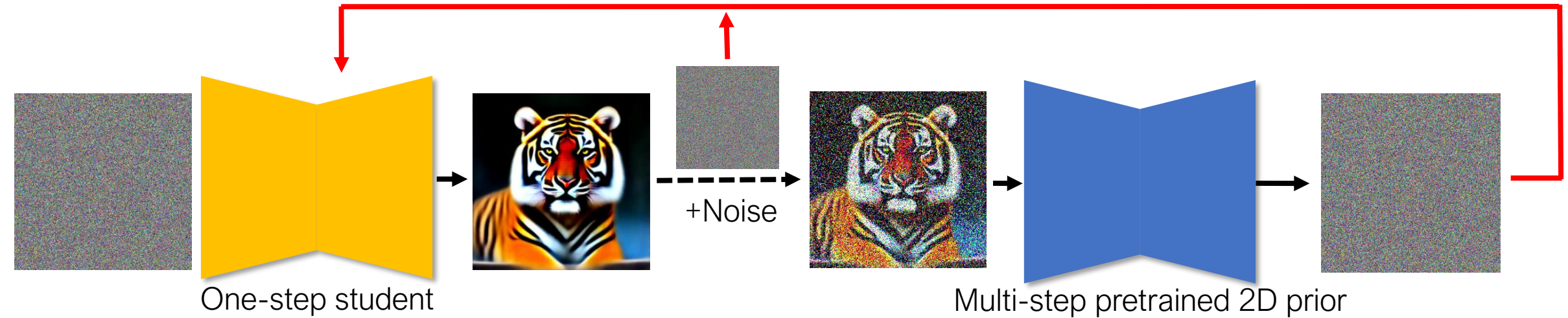
What if we replace the 3D NeRF with a 2D one-step generator?

Poole, Ben, et al. "DreamFusion: Text-to-3D using 2D Diffusion." ICLR 2023

Score Distillation Sampling (SDS)

Naively use SDS loss leads to mode collapse and poor-quality output!

Poole, Ben, et al. "DreamFusion: Text-to-3D using 2D Diffusion." ICLR 2023

Variational Score Distillation (VSD)

Michelangelo style statue of dog reading news on a cellphone.

A pineapple.

A chimpanzee dressed like Henry VIII king of England.

An elephant skull.

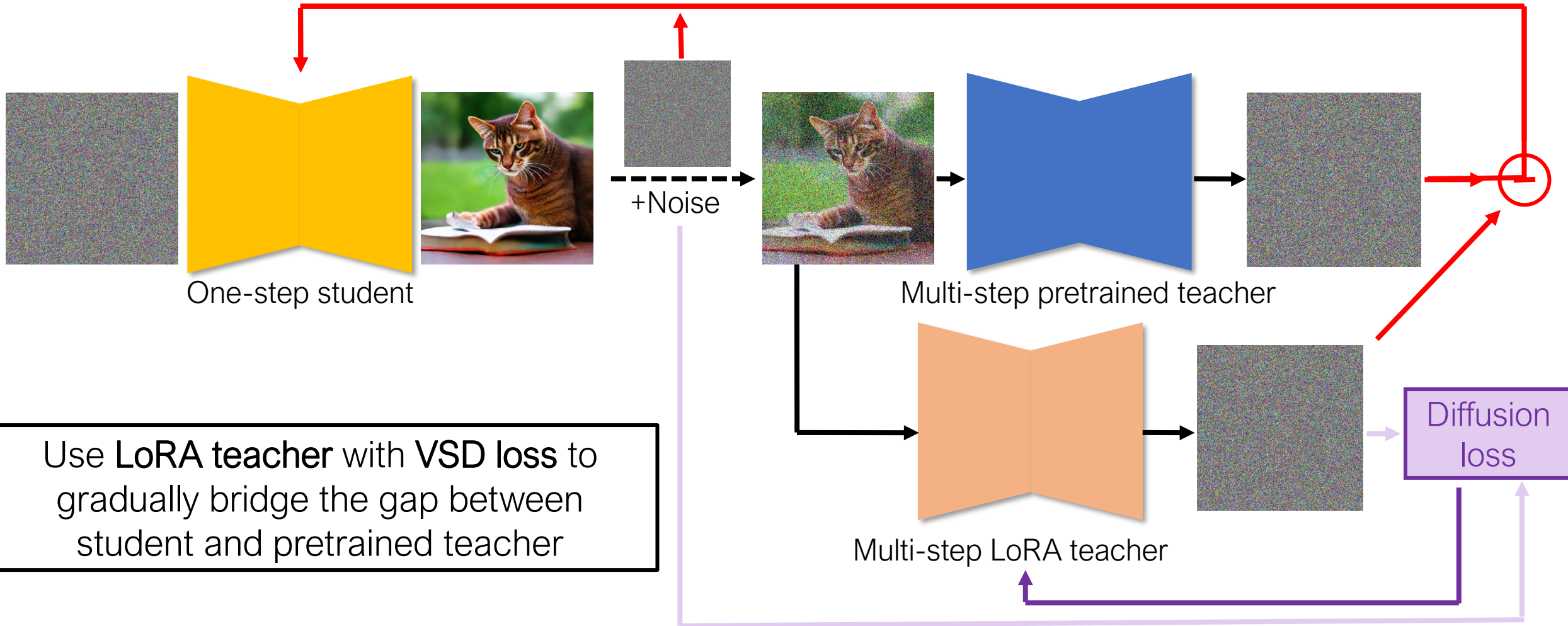A model of a house in Tudor style.

A tarantula, highly detailed.

A snail on a leaf.

An astronaut is riding a horse.

Wang, Zhengyi, et al. "ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation" NeuRIPS 2023

Variational Score Distillation (VSD)

VSD Loss

One-step student

+Noise

Multi-step pretrained teacher

Multi-step LoRA teacher

Use **LoRA teacher** with **VSD loss** to gradually bridge the gap between student and pretrained teacher

Diffusion loss

Wang, Zhengyi, et al. "ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation" NeuRIPS 2023

Qualitative Results

InstaFlow - 199.2 A100 days – Requiring paired image-text training data

Qualitative Results

SwiftBrush - 4.1 A100 days – Image-free training that requires only prompt data

Quantitative Results

MS COCO 2014

Method	Steps	FID-30K ↓	CLIP-30K ↑
Guided Distillation[†]	1	37.3	0.27
LCM[†]	1	35.56	0.24
Instaflow	1	**13.10**[†]	0.28[§]
BOOT[‡]	1	48.20	0.26
Ours	1	16.67	**0.29**
SD 2.1*	25	13.45	0.30
SD 2.1*	1	202.14	0.08

HPS v2

Models	Human Preference Score v2 ↑			
	Anime	Photo	Concept Art	Paintings
LCM[†]	22.61	22.71	22.74	22.91
InstaFlow[†]	25.98	26.32	25.79	25.93
BOOT[‡]	25.29	25.16	24.40	24.61
Ours	**26.91**	**27.21**	**26.32**	**26.37**
SD 2.1*	27.48	26.89	26.86	27.46

Time and Memory

Method	Training		Inference	
	Time (GPU days)	Memory (GB)	Time (ms)	Memory (GB)
Guided Distillation	108.8[†]	-	-	-
LCM	1.3[†]	33.6*	118*	11.3*
InstaFlow	199.2[†]	-	116*	12.3*
BOOT[‡]	5.6	30.2	115	8.3
Ours	4.1	26.4	110	8.2

Additional Qualitative Results

THANK YOU

Github
https://github.com/VinAIResearch/SwiftBrush