Generating Images with 3D Annotations Using Diffusion Models

Wufei Ma

Apr 05, 2024

Generating Images with 3D Annotations Using Diffusion Models

ICLR, 2024 (Spotlight)

Authors

Wufei Ma*, Qihao Liu*, Jiahao Wang*, Angtian Wang, Xiaoding Yuan, Yi Zhang, Zihao Xiao, Guofeng Zhang, Beijia Lu, Ruxiao Duan, Yongrui Qi, Adam Kortylewski, Yaoyao Liu+, Alan Yuille

* Equal contribution

+ Corresponding author

Published as a conference paper at ICLR 2024

Generating Images with 3D Annotations Using Diffusion Models

Wufei Ma[1,*], Qihao Liu[1,*], Jiahao Wang[1,*], Angtian Wang[1], Xiaoding Yuan[1], Yi Zhang[1], Zihao Xiao[1], Guofeng Zhang[1], Beijia Lu[1], Ruxiao Duan[1], Yongrui Qi[1], Adam Kortylewski[2,3], Yaoyao Liu[1,✉], Alan Yuille[1]

[1]Johns Hopkins University [2]University of Freiburg
[3]Max Planck Institute for Informatics, Saarland Informatics Campus

Abstract

Diffusion models have emerged as a powerful generative method, capable of producing stunning photo-realistic images from natural language descriptions. However, these models lack explicit control over the 3D structure in the generated images. Consequently, this hinders our ability to obtain detailed 3D annotations for the generated images or to craft instances with specific poses and distances. In this paper, we propose 3D Diffusion Style Transfer (3D-DST), which incorporates 3D geometry control into diffusion models. Our method exploits ControlNet, which extends diffusion models by using visual prompts in addition to text prompts. We generate images of the 3D objects taken from 3D shape repositories (e.g., ShapeNet and Objaverse), render them from a variety of poses and viewing directions, compute the edge maps of the rendered images, and use these edge maps as visual prompts to generate realistic images. With explicit 3D geometry control, we can easily change the 3D structures of the objects in the generated images and obtain ground-truth 3D annotations automatically. This allows us to improve a wide range of vision tasks, e.g., classification and 3D pose estimation, in both in-distribution (ID) and out-of-distribution (OOD) settings. We demonstrate the effectiveness of our method through extensive experiments on ImageNet-100/200, ImageNet-R, PASCAL3D+, ObjectNet3D, and OOD-CV. The results show that our method significantly outperforms existing methods, e.g., 3.8 percentage points on ImageNet-100 using DeiT-B. Our code is available at https://ccvl.jhu.edu/3D-DST/

1 Introduction

Understanding the underlying 3D world of 2D images is essential to numerous computer vision tasks. The utilization of 3D modeling opens up the possibility of addressing a significant portion of the variability inherent in natural images, which could potentially enhance the overall understanding and interpretation of images (Wu et al., 2020). For example, 3D-aware models show high robustness and generalization ability under occlusion or environmental changes (Liu et al., 2022a). However, it is expensive and time-consuming to obtain ground-truth 3D annotations for 2D images. This training data shortage becomes a main obstacle to training large-scale 3D-aware models.

Recently, diffusion models (Ho et al., 2020) have shown impressive performance in generating photo-realistic images, which can be used to solve the training data shortage. These models allow us to produce high-quality images from various conditional inputs, e.g., natural language descriptions, segmentation maps, and keypoints (Zhang et al., 2023). This facilitates generative data augmentation, e.g., He et al. (2023) use diffusion models to augment ImageNet (Deng et al., 2009) and significantly improve the classification results.

Despite their success, diffusion models still lack explicit control over the underlying 3D world during the generation process. As a result, they still face two challenges that hinder their use in augmenting data for 3D tasks. The first challenge is the inability to control the 3D properties of the object in

* Equal contribution.
✉ Corresponding author: Yaoyao Liu (yyliu@cs.jhu.edu).

Impressive Diffusion Models

DALL·E 3
“A 2D animation of a fold music band composed of anthropomorphic autumn leaves.”

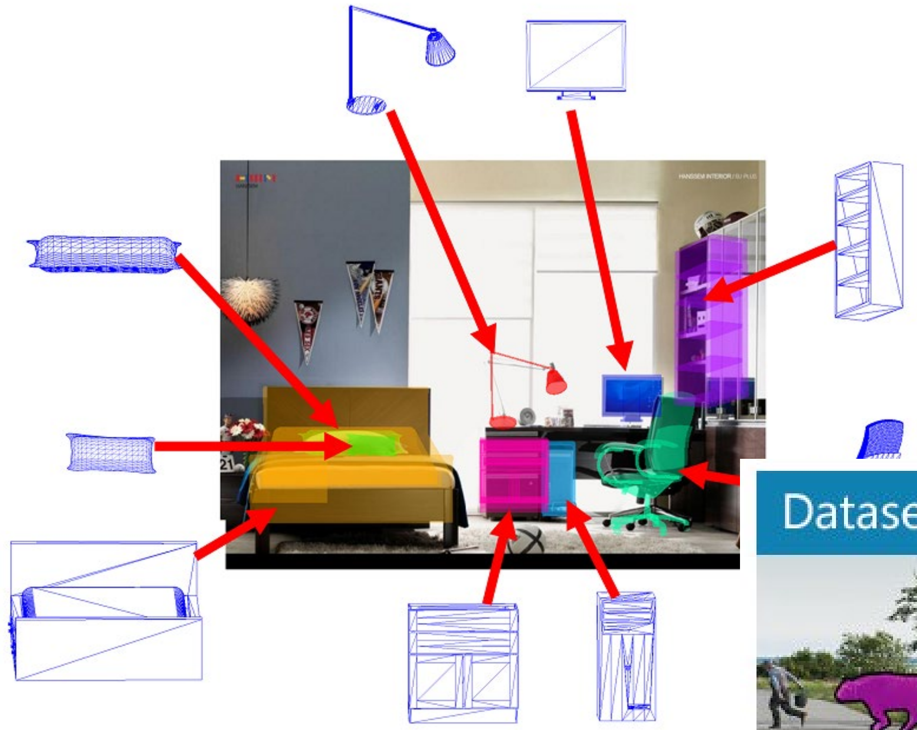
Stable Diffusion XL
“A capybara made of lego sitting in a realistic, natural field.”

Diffusion Model Applications

Character Reference "cref"

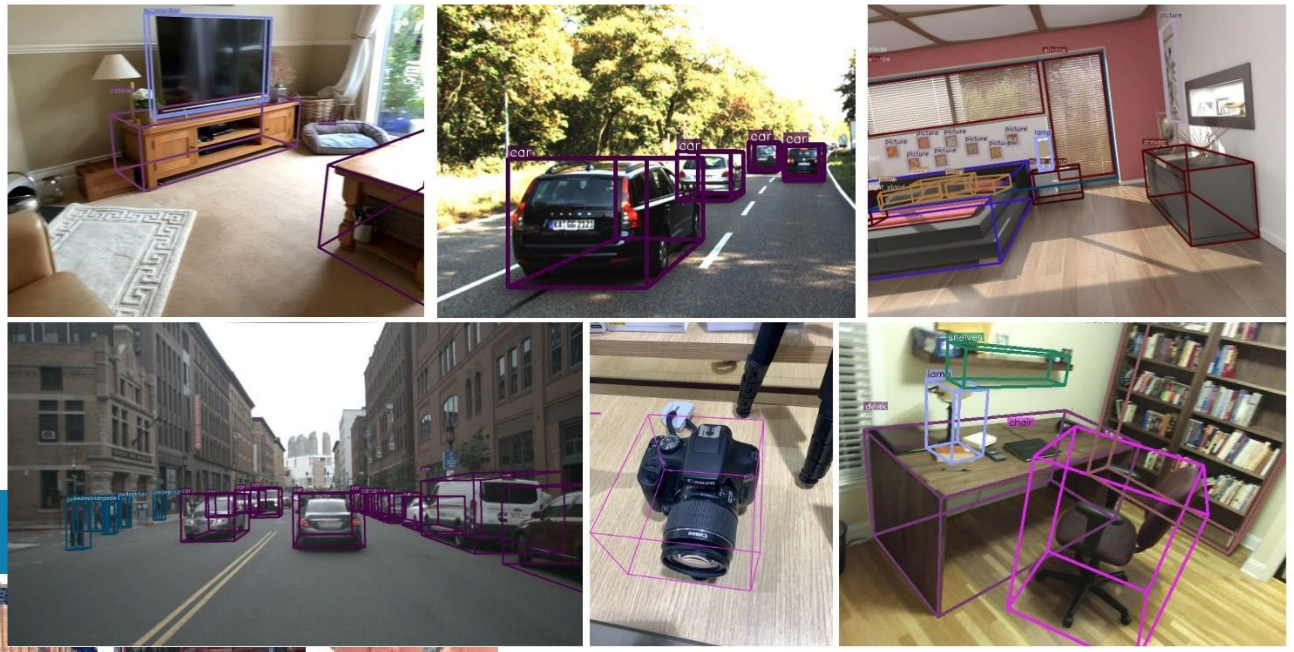
Image Editing

Synthetic Data for Better Recognition

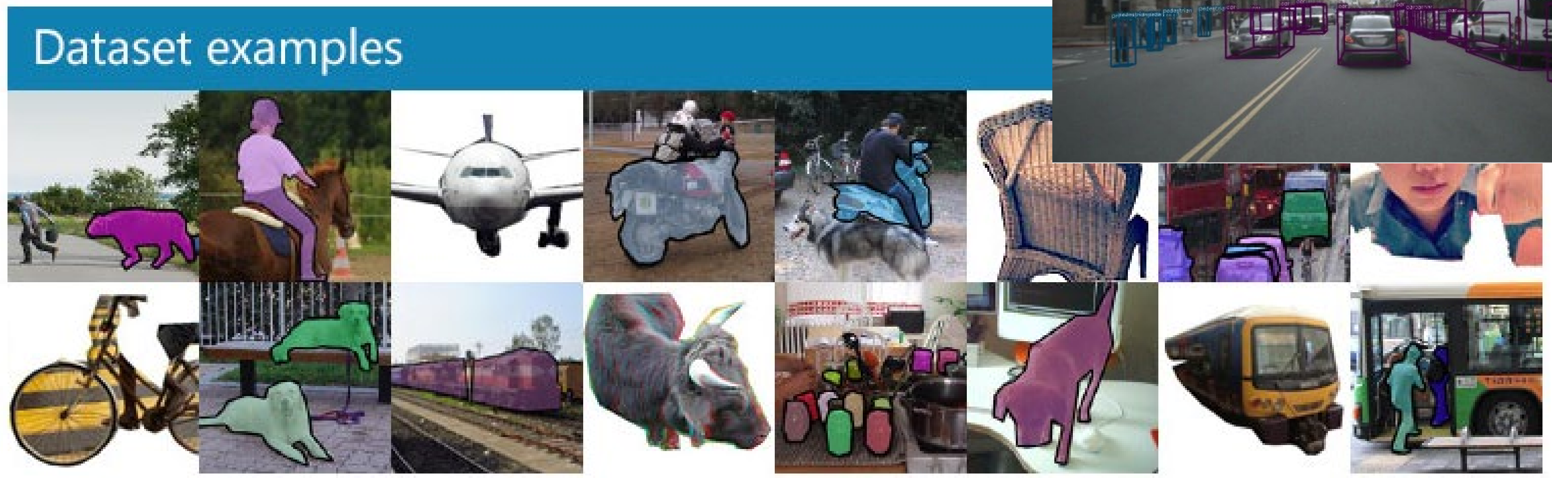


ObjectNet3D

Omni3D



COCO



Synthetic Data for Image Classification

Dataset	Task	CLIP-RN50	CLIP-RN50+SYN	CLIP-ViT-B/16	CLIP-ViT-B/16+SYN
CIFAR-10	o	70.31	80.06 (+9.75)	90.80	92.37 (+1.57)
CIFAR-100	o	35.35	45.69 (+10.34)	68.22	70.71 (+2.49)
Caltech101	o	86.09	87.74 (+1.65)	92.98	94.16 (+1.18)
Caltech256	o	73.36	75.74 (+2.38)	80.14	81.43 (+1.29)
ImageNet	o	60.33	60.78 (+0.45)	68.75	69.16 (+0.41)
SUN397	s	58.51	60.07 (+1.56)	62.51	63.79 (+1.28)
Aircraft	f	17.34	21.94 (+4.60)	24.81	30.78 (+5.97)
Birdsnap	f	34.33	38.05 (+3.72)	41.90	46.84 (+4.94)
Cars	f	55.63	56.93 (+1.30)	65.23	66.86 (+1.63)
CUB	f	46.69	56.94 (+10.25)	55.23	63.79 (+8.56)
Flower	f	66.08	67.05 (+0.97)	71.30	72.60 (+1.30)
Food	f	80.34	80.35 (+0.01)	88.75	88.83 (+0.08)
Pets	f	85.80	86.81 (+1.01)	89.10	90.41 (+1.31)
DTD	t	42.23	43.19 (+0.96)	44.39	44.92 (+0.53)
EuroSAT	si	37.51	55.37 (+17.86)	47.77	59.86 (+12.09)
ImageNet-Sketch	r	33.29	36.55 (+3.26)	46.20	48.47 (+2.27)
ImageNet-R	r	56.16	59.37 (+3.21)	74.01	76.41 (+2.40)
Average	/	55.13	59.47 (+4.31)	65.42	68.32 (+2.90)

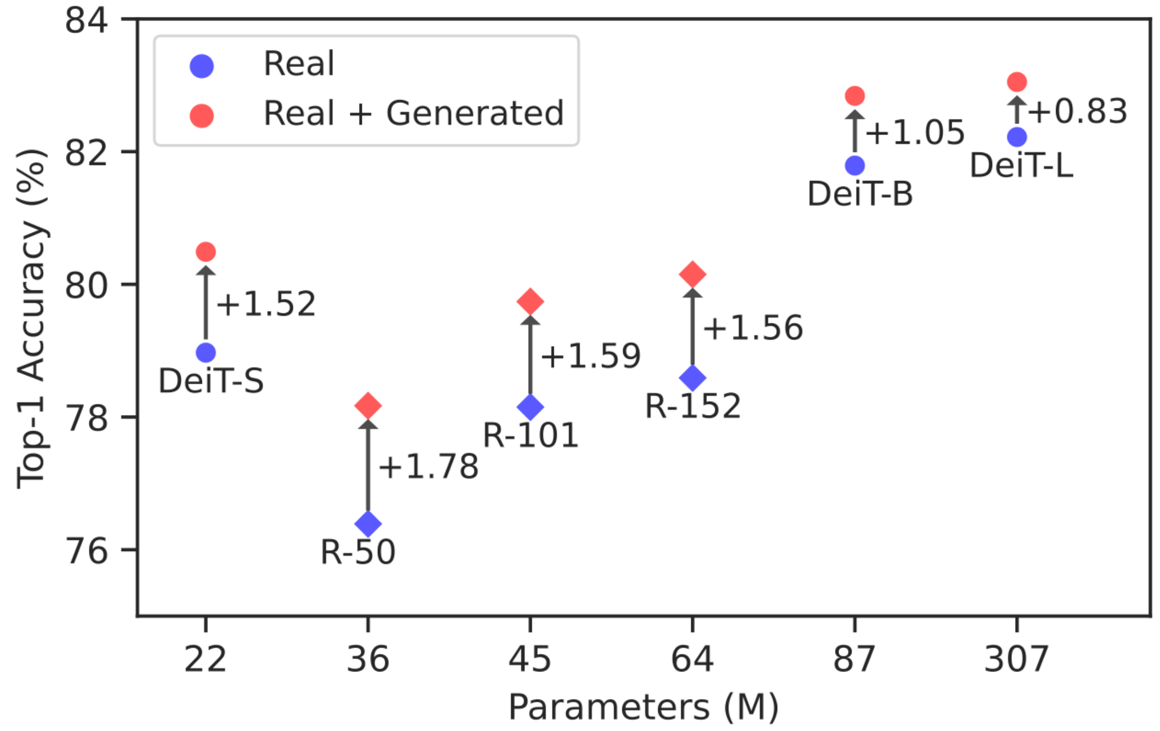
Table 1: **Main Results on Zero-shot Image Recognition.** All results are top-1 accuracy on test set. o: object-level. s: scene-level. f: fine-grained. t: textures. si: satellite images. r: robustness.

Data	pre-trained on IN-1k?	Syn. images amount			
		0	1.2M	2.4M	4.0M
(None)		66.08	-	-	-
IN-1K Syn		-	79.00	80.00	-
IN-2K Syn		-	-	80.54	80.72
(None)	✓	81.30	-	-	-
IN-1K Syn	✓	-	-	**81.78**	-
IN-2K Syn	✓	-	-	**81.87**	**81.91**

Table 10: Results for object detection on PASCAL VOC with **downstream-agnostic supervised pre-training**, all results are reported in AP_{50}.

[1] He et al. Is synthetic data from generative models ready for image recognition? In ICLR, 2023.

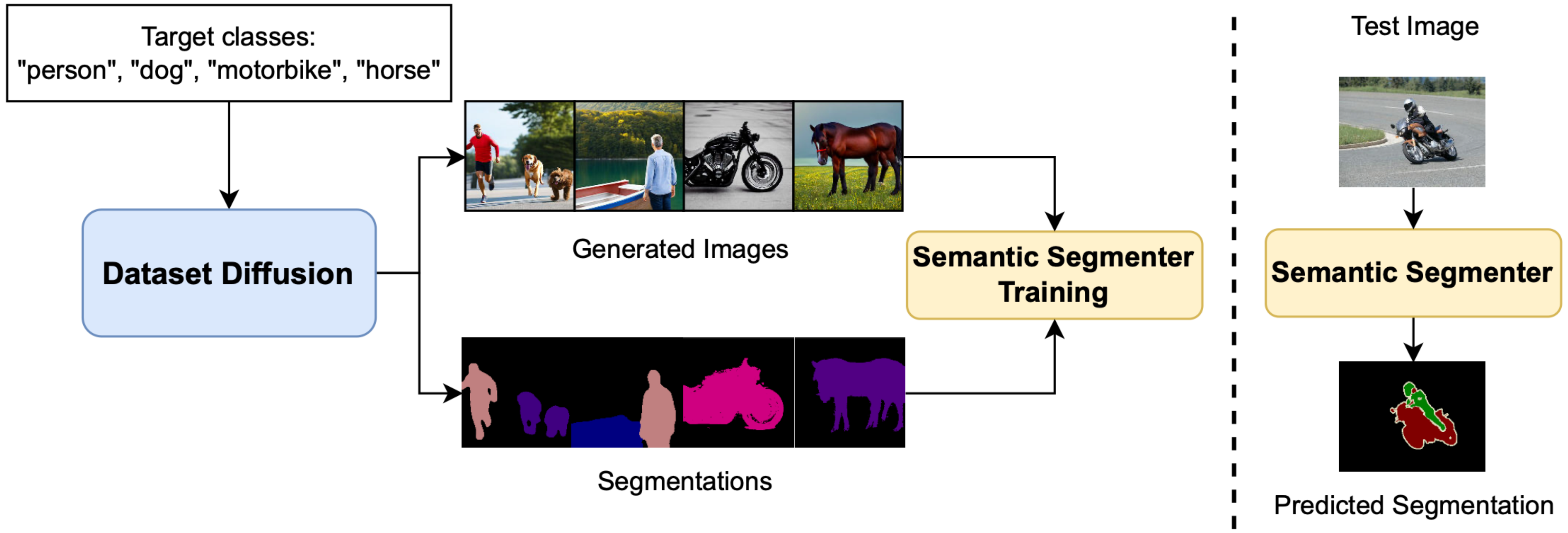
Synthetic Data for Image Classification

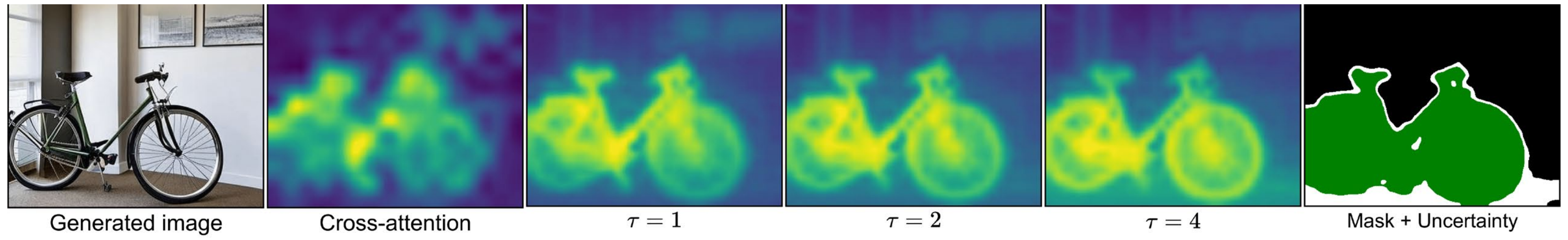


Train Set (M)	256×256	1024×1024
1.2	76.39 ± 0.21	76.39 ± 0.21
2.4	77.61 ± 0.08 (+1.22)	78.12 ± 0.05 (+1.73)
3.6	77.16 ± 0.04 (+0.77)	77.48 ± 0.04 (+1.09)
4.8	76.52 ± 0.04 (+0.13)	76.75 ± 0.07 (+0.36)
6.0	76.09 ± 0.08 (-0.30)	76.34 ± 0.13 (-0.05)
7.2	75.81 ± 0.08 (-0.58)	75.87 ± 0.09 (-0.52)
8.4	75.44 ± 0.06 (-0.95)	75.49 ± 0.07 (-0.90)
9.6	75.28 ± 0.10 (-1.11)	74.72 ± 0.20 (-1.67)
10.8	75.11 ± 0.12 (-1.28)	74.14 ± 0.13 (-2.25)
12.0	75.04 ± 0.05 (-1.35)	73.70 ± 0.09 (-2.69)

[1] Azizi et al. Synthetic data from diffusion models improves ImageNet classification. In TMLR, 2023.

Synthetic Data for Segmentation

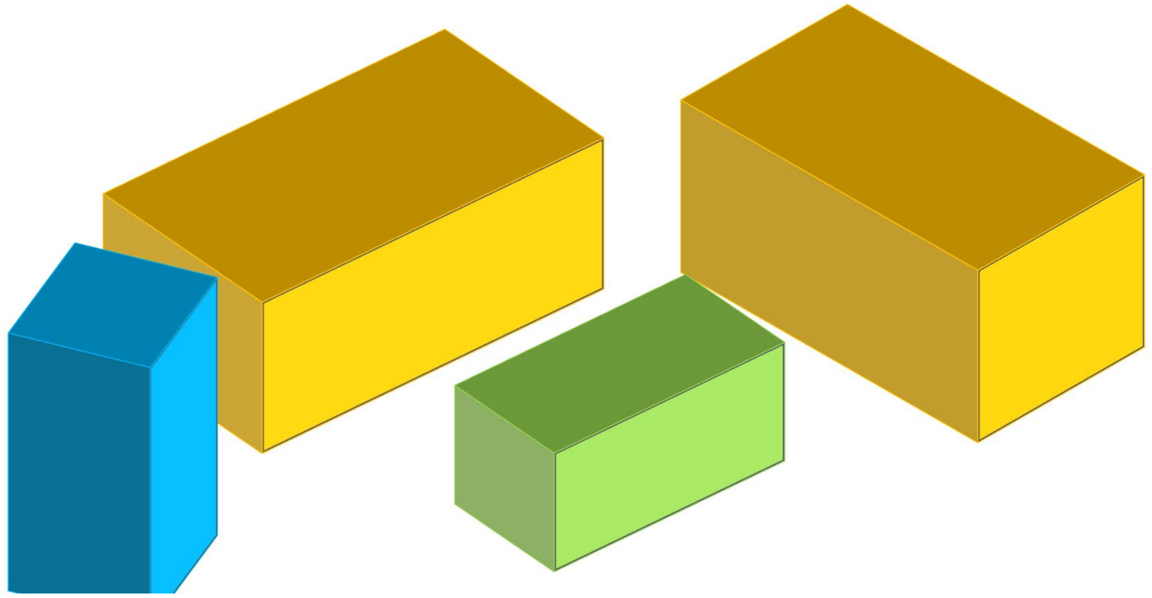




[1] Nguyen et al. Dataset diffusion: Diffusion-based synthetic dataset generation for pixel-level semantic segmentation. In NeurIPS, 2023.

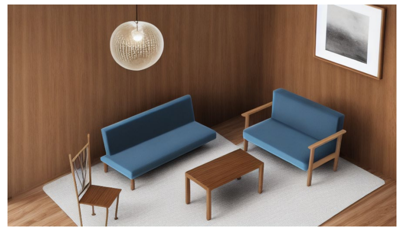
Synthetic Data with 3D Ground Truth

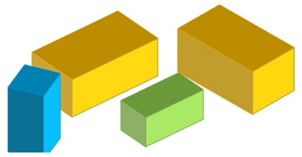
We are also interested in synthetic data with 3D ground truth such as 3D viewpoint, 3D location, object shape, object depth, etc.



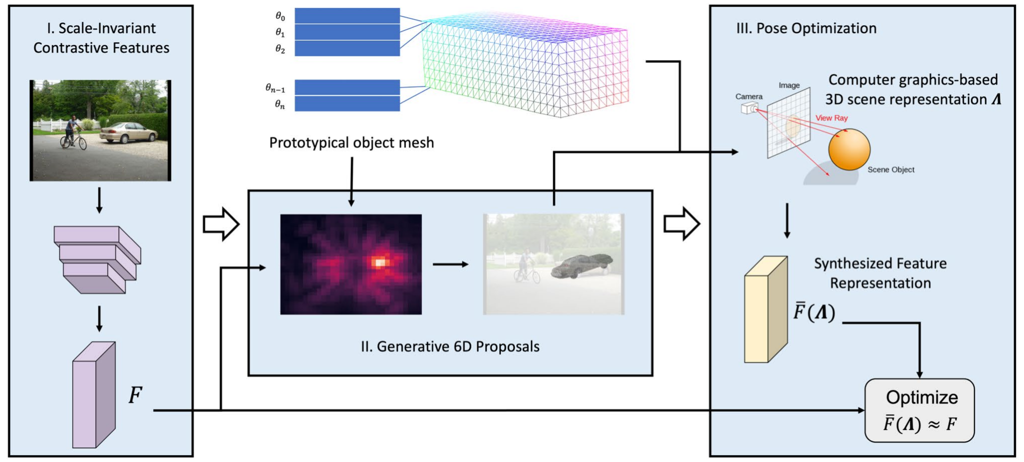
Synthetic Data with 3D Ground Truth

We are also interested in synthetic data with 3D ground truth such as 3D viewpoint, 3D location, object shape, object depth, etc.





I. NeMo models

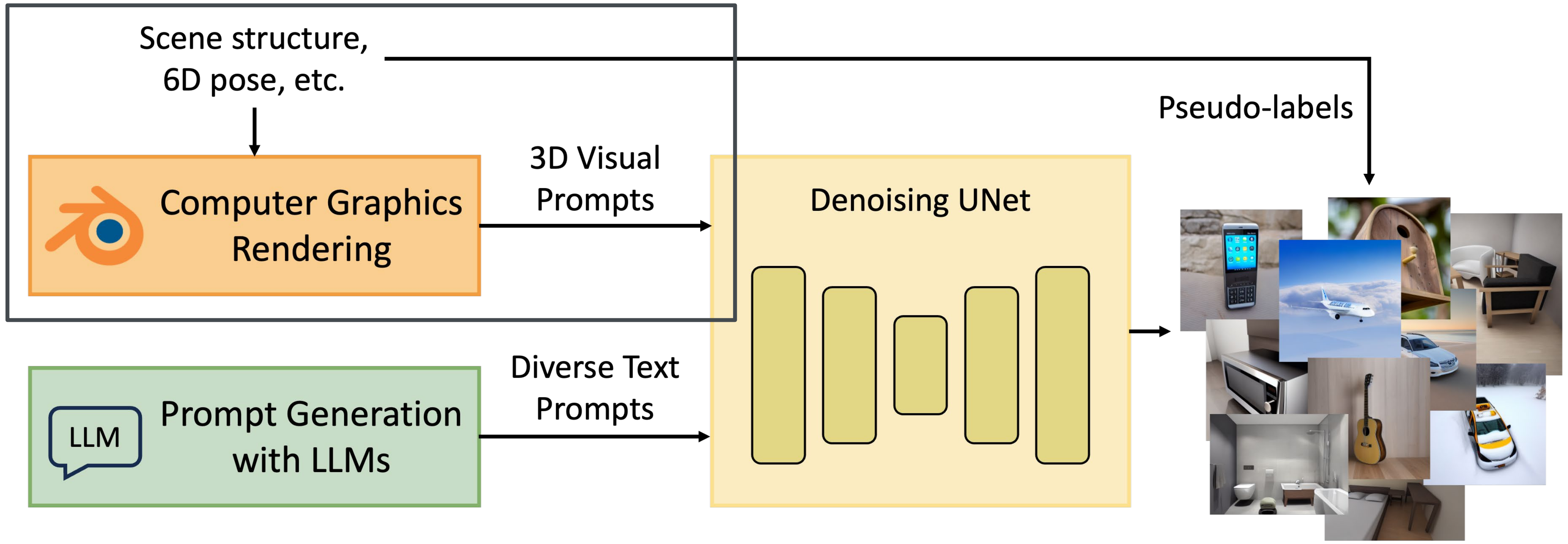


II. Ego/Exo-centric video understanding

III. General purpose AI

Q: Which object is closer to the camera, the washing machine or the microwave? Answer the washing machine or the microwave only.
GPT−4v: the washing machine
GT: the microwave

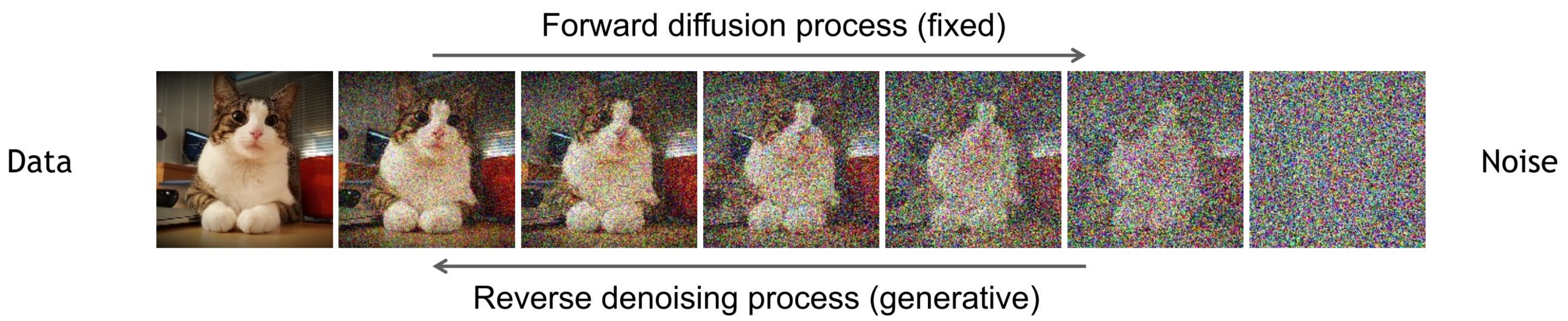
Synthetic Data with 3D Ground Truth



Standard Diffusion Model

Denoising diffusion models consist of two processes:

- Forward diffusion process that gradually adds noise to input
- Reverse denoising process that learns to generate data by denoising



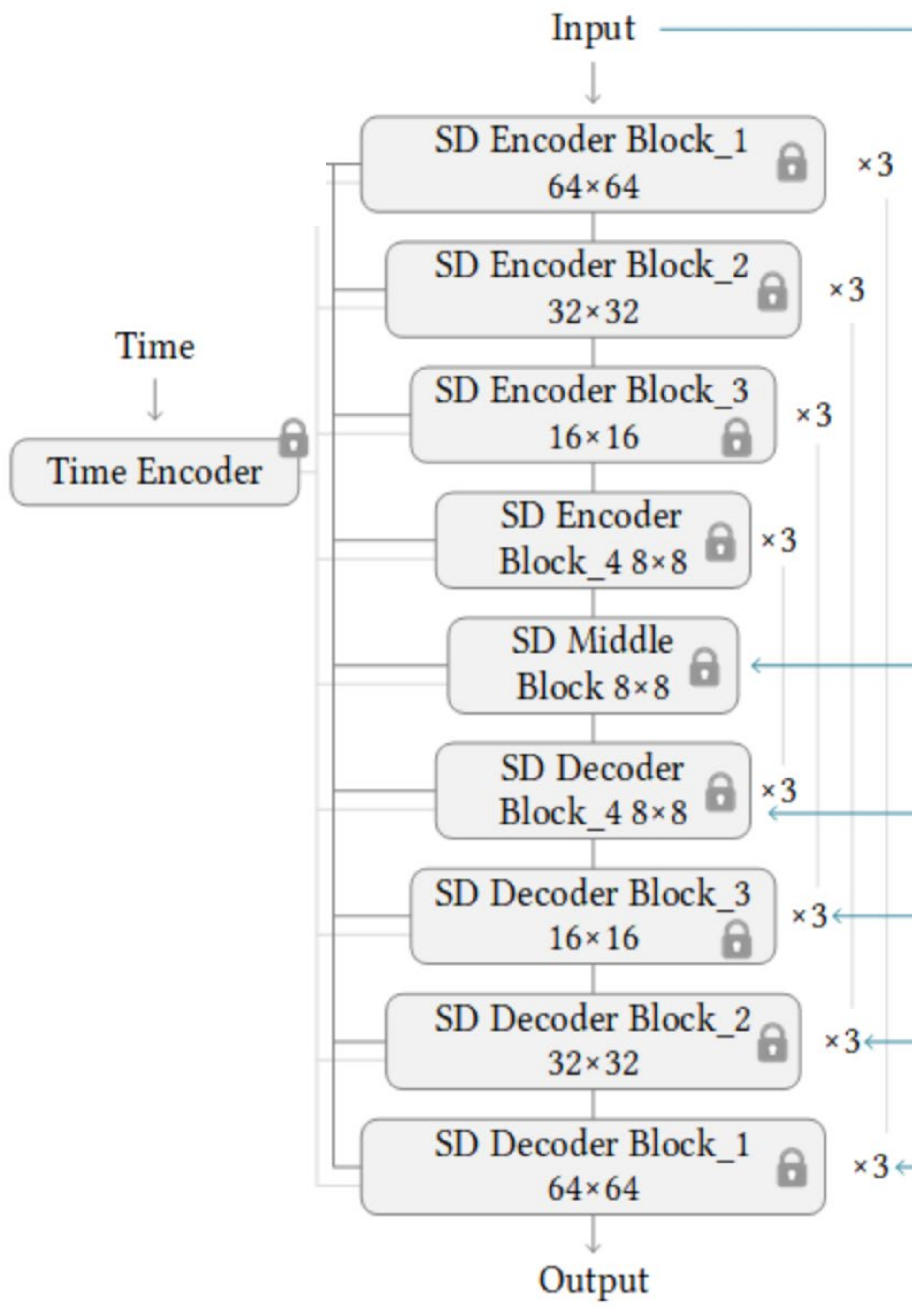
Credit: CVPR 2022 Tutorial on diffusion models: link.

Standard Diffusion Model

During generation, we gradually denoise the latent with an iterative process.

$$z_{t-1} = \epsilon(z_t, t), \quad t = T, \ldots, 1$$

Here ϵ is often modeled with a trainable network with U-Net architecture.



Standard Diffusion Model

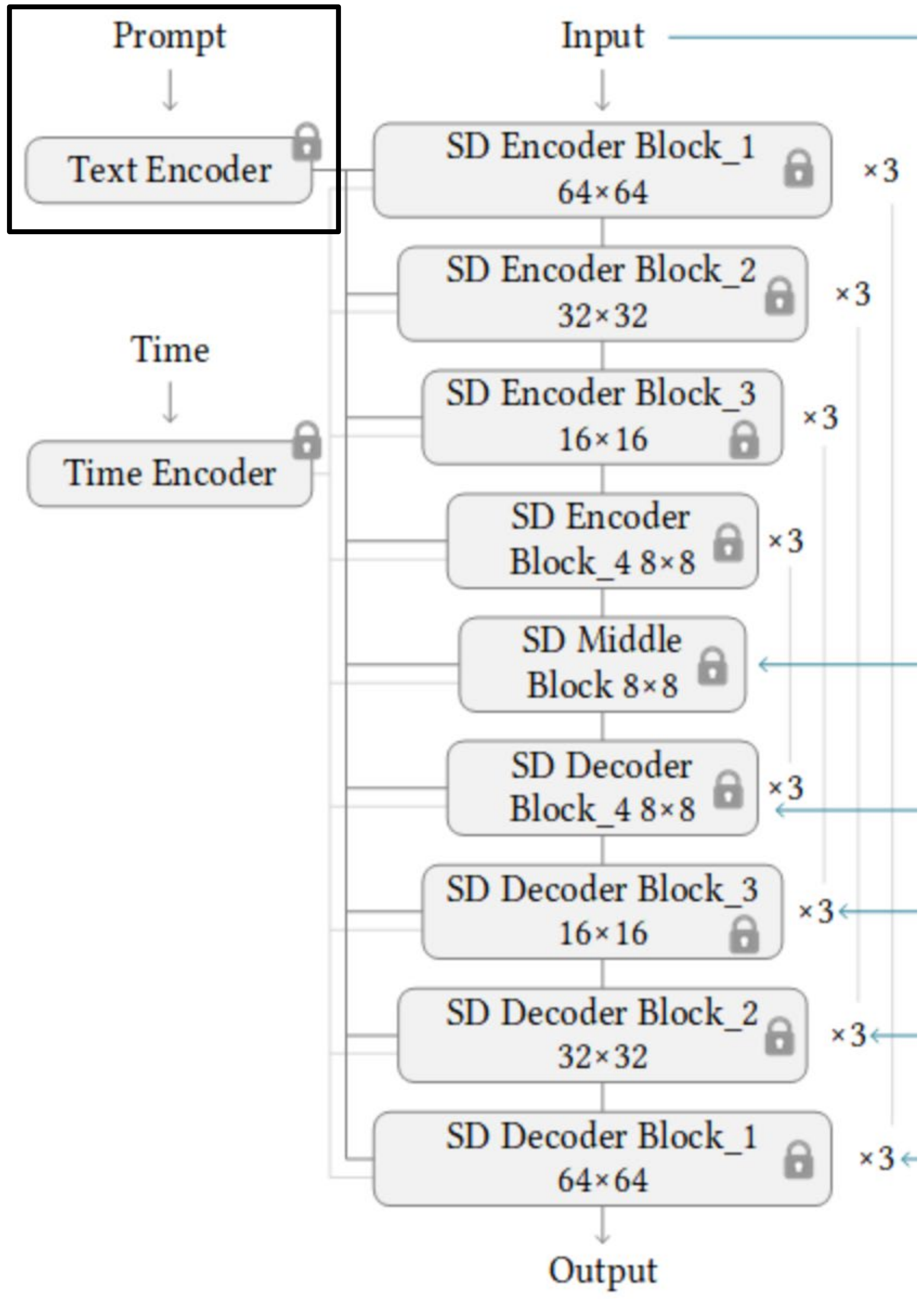
During generation, we gradually denoise the latent with an iterative process.

$$z_{t-1} = \epsilon(z_t, t), \quad t = T, \ldots, 1$$

Here ϵ is often modeled with a trainable network with U-Net architecture.

To generate images with desired contents, we add **text-conditioning** with cross-attention layers.

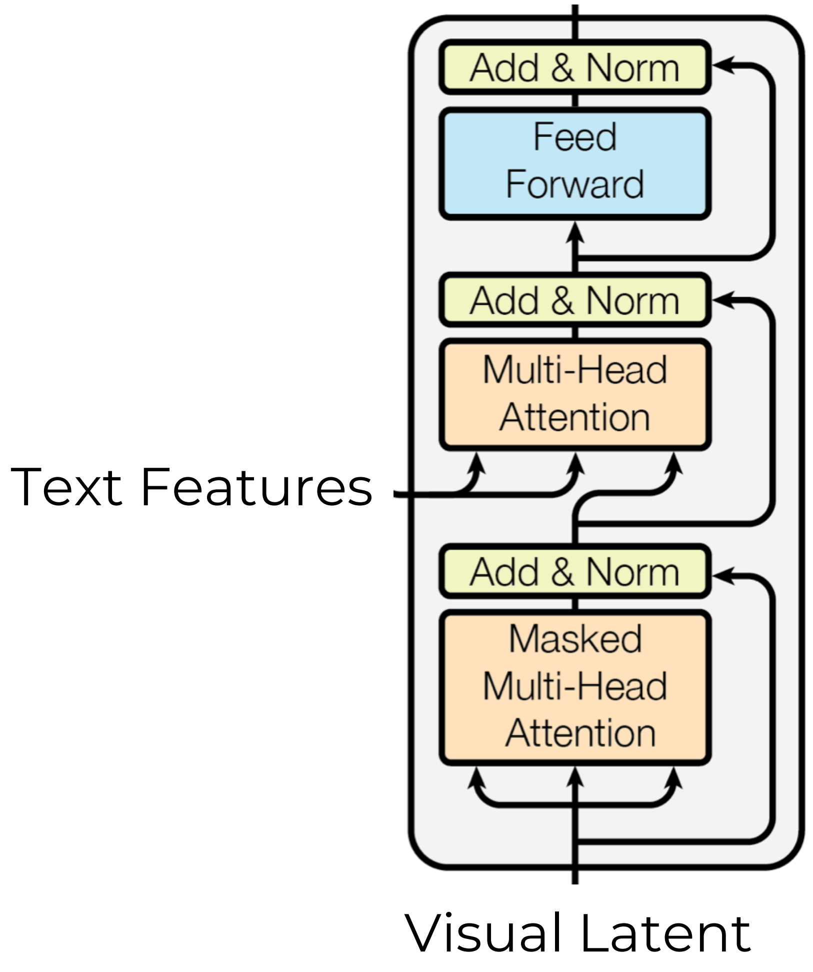
$$z_{t-1} = \epsilon(z_t, \mathcal{T}, t), \quad t = T, \ldots, 1$$

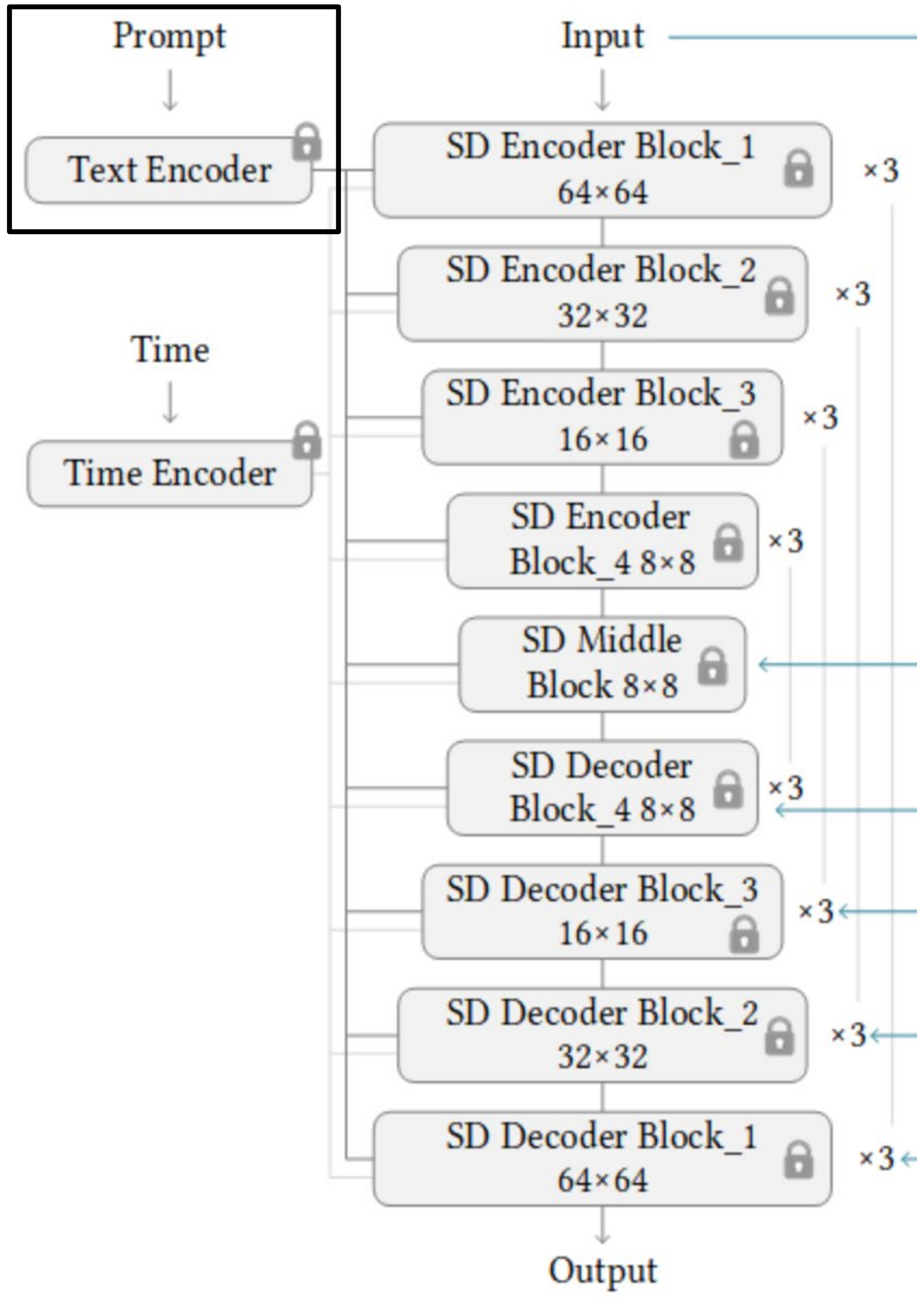


Text Conditioning

To generate images with desired contents, we add **text-conditioning** with cross-attention layers.

$$z_{t-1} = \epsilon(z_t, \mathcal{T}, t), \quad t = T, \ldots, 1$$

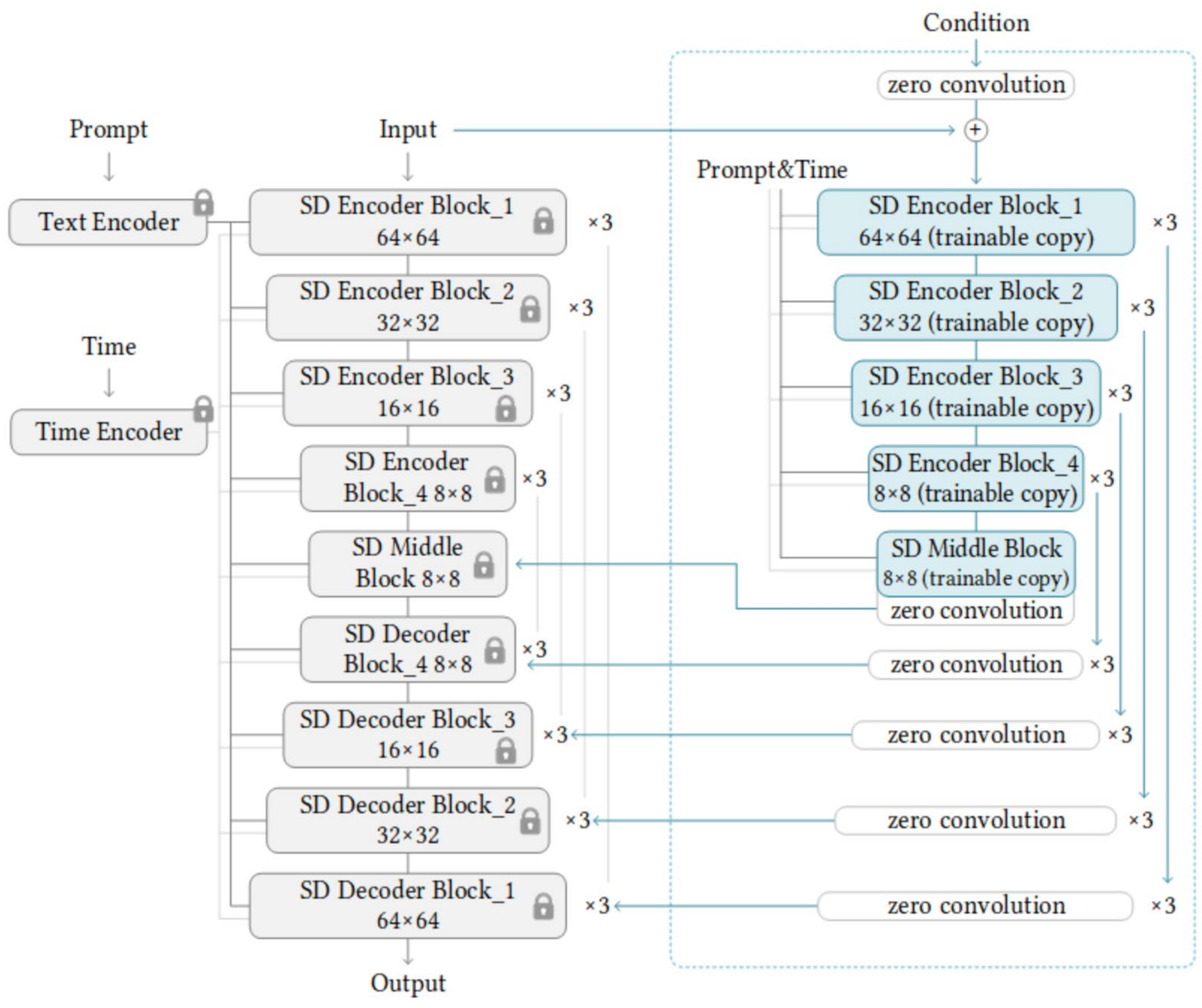




Visual Conditioning

Furthermore, ControlNet adds visual conditioning to text-to-image diffusion models by adding visual features via **zero convolutions** (layers initialized with zero weights).

The advantage is that the base text-to-image diffusion model is fixed, and we **achieve various visual conditioning** by training only an adapter on top of the large pretrained diffusion model.

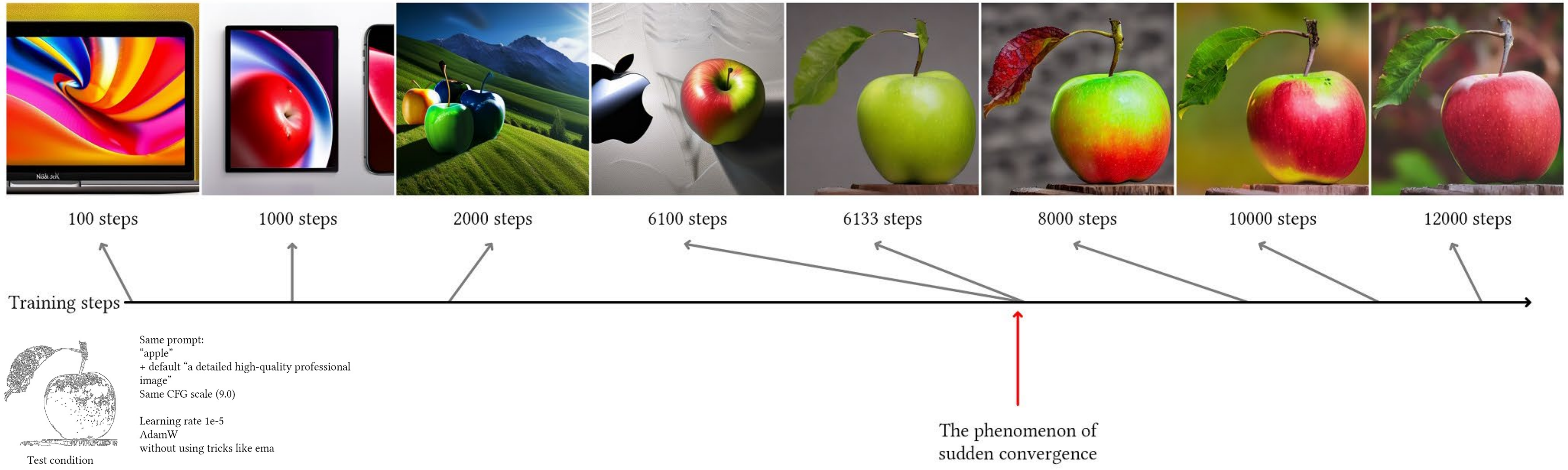


[1] Zhang et al. Adding conditional control to text-to-image diffusion models. In ICCV, 2023.

ControlNet

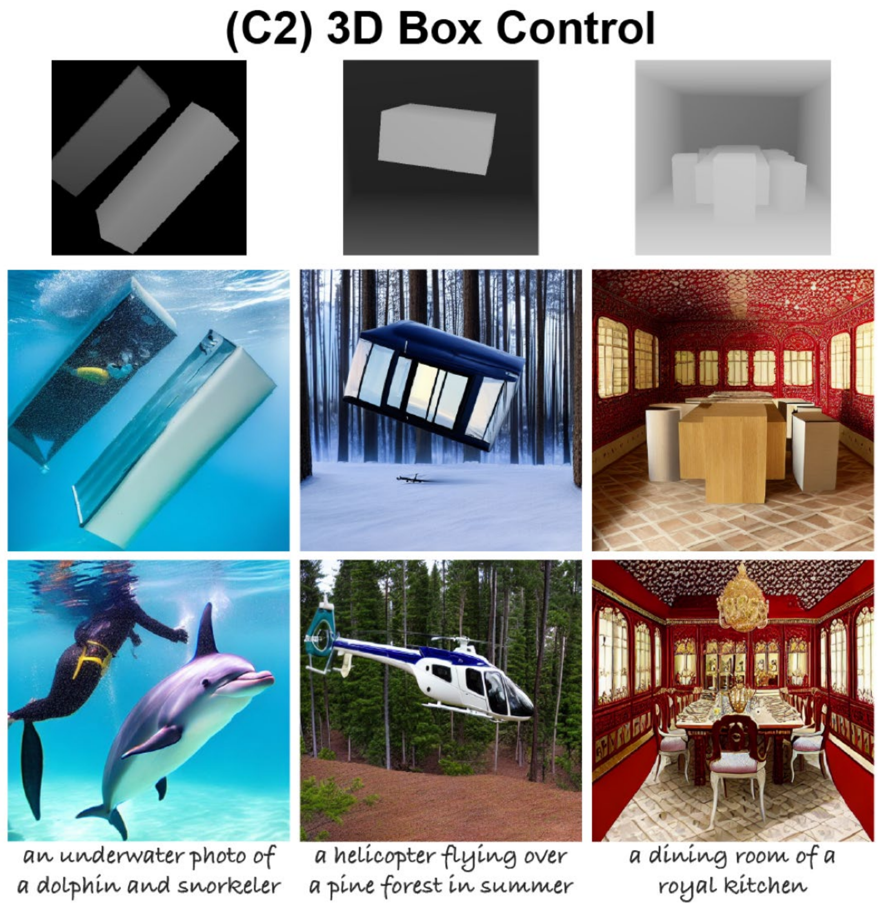
[1] Zhang et al. Adding conditional control to text-to-image diffusion models. In ICCV, 2023.

ControlNet



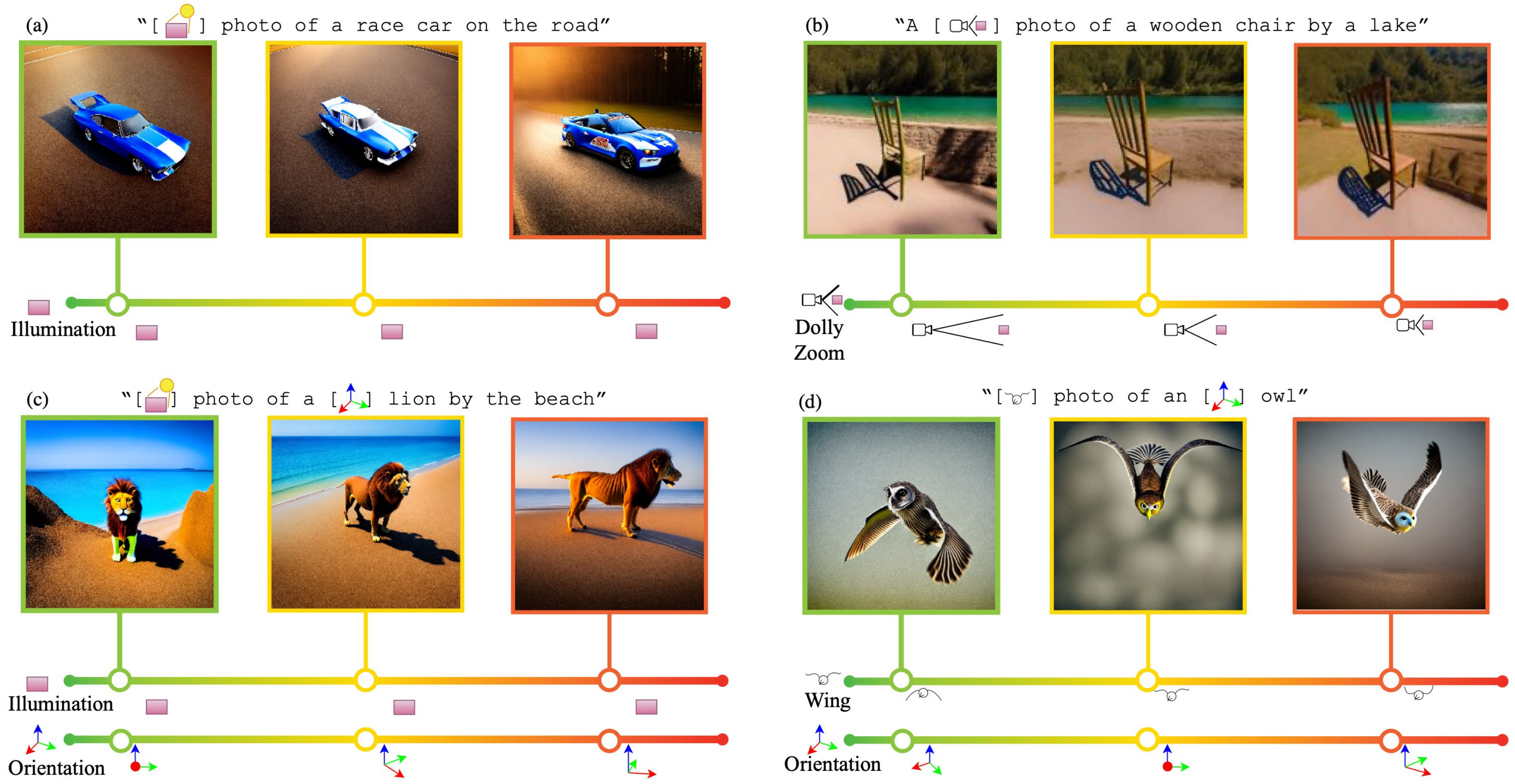
[1] Zhang et al. Adding conditional control to text-to-image diffusion models. In ICCV, 2023.

LooseControl



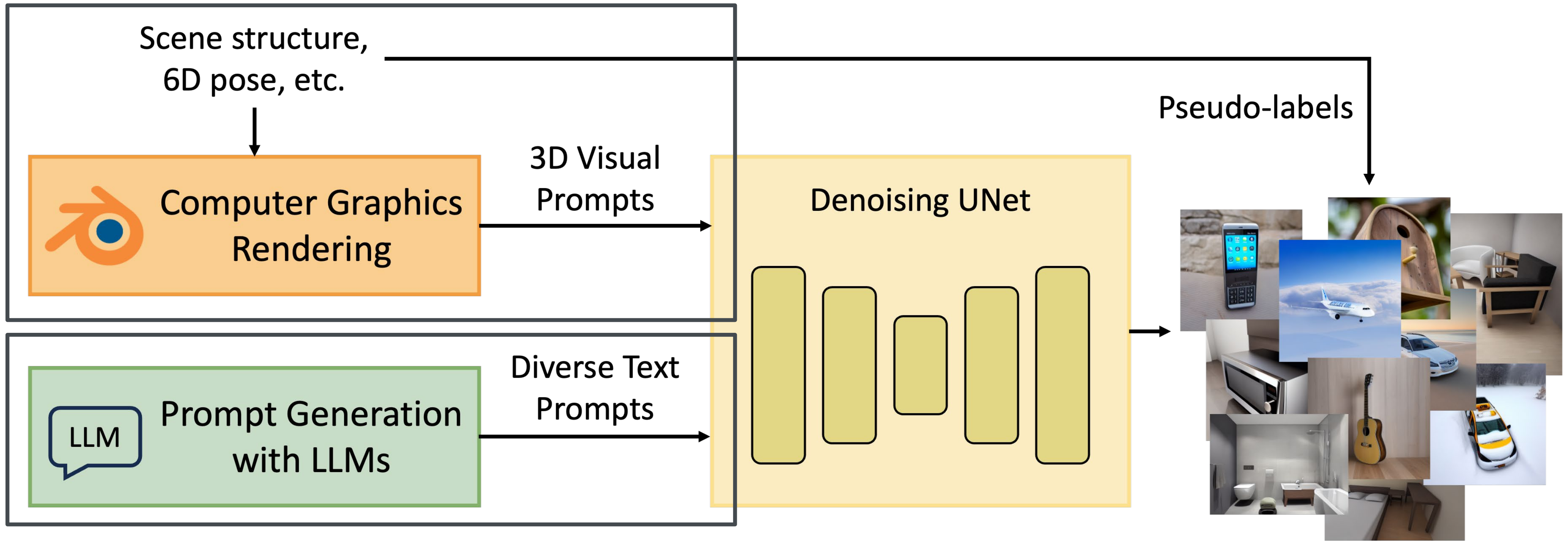
[1] Bhat et al. LooseControl: Lifting ControlNet for generalized depth conditioning. Preprint.
Figure credit: Olaf Dünkel.

3D Words



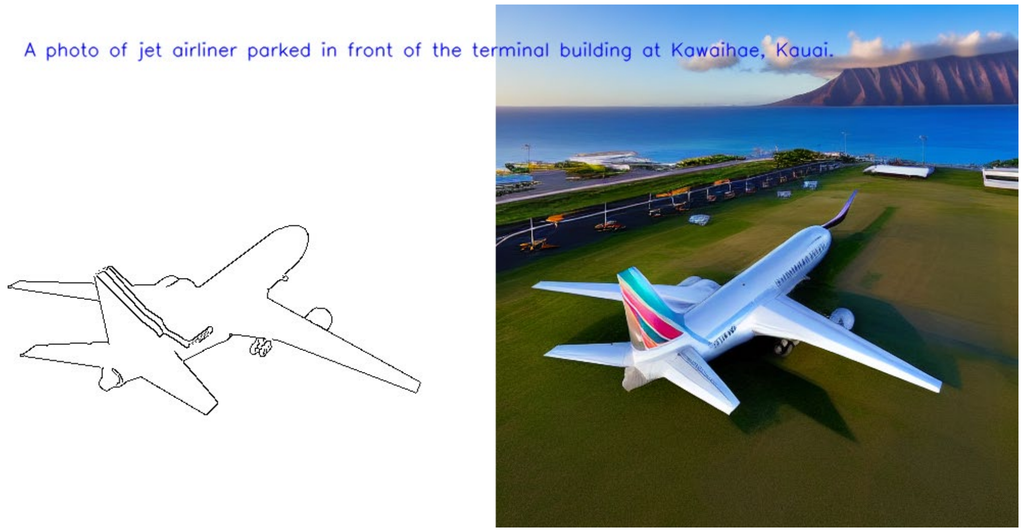
[1] Cheng et al. Learning continuous 3D words for text-to-image generation. Preprint.

Synthetic Data with 3D Ground Truth



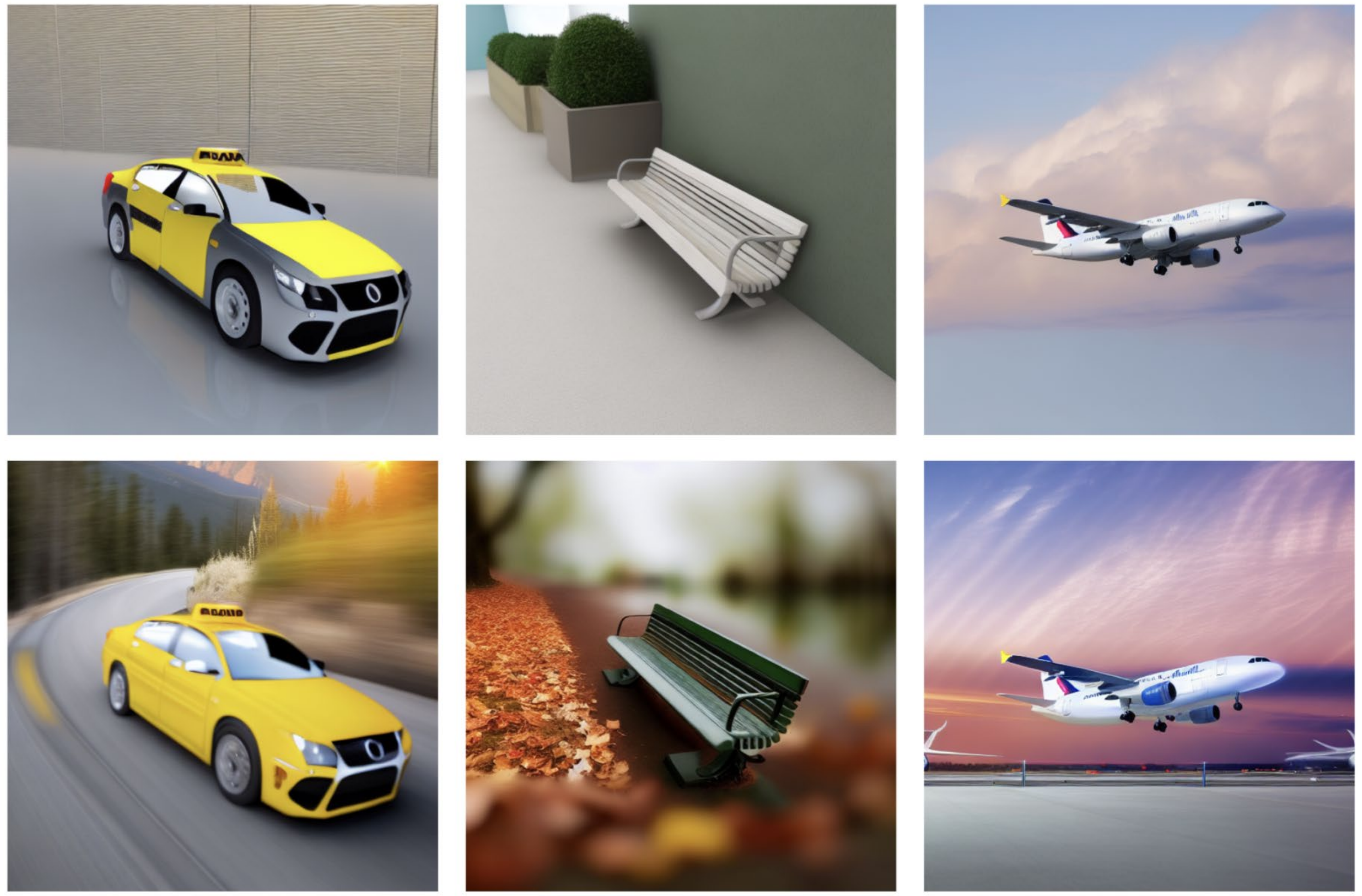
Diverse Prompt Generation with LLMs

Diverse prompts improve the realism and diversity of the synthetic images. Models trained on such images are found to be more robust.



1. Keywords from ShapeNet / Objaverse
2. Class name
3. LLM generation

A photo of jet airliner parked in front of the terminal building at Kawaihae, Kauai

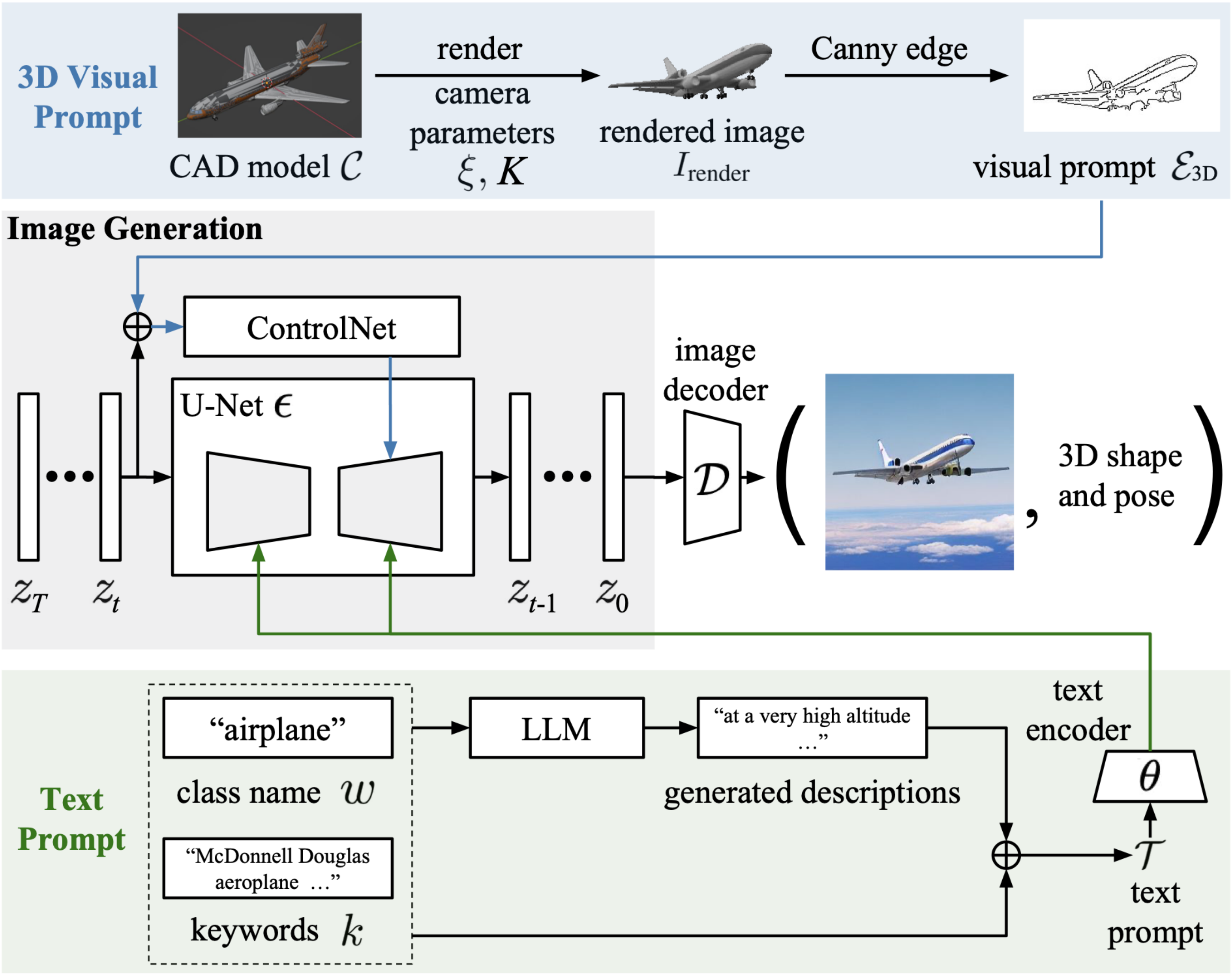


A photo of a taxicab speeding down a winding mountain road

A photo of a park bench covered in fallen leaves

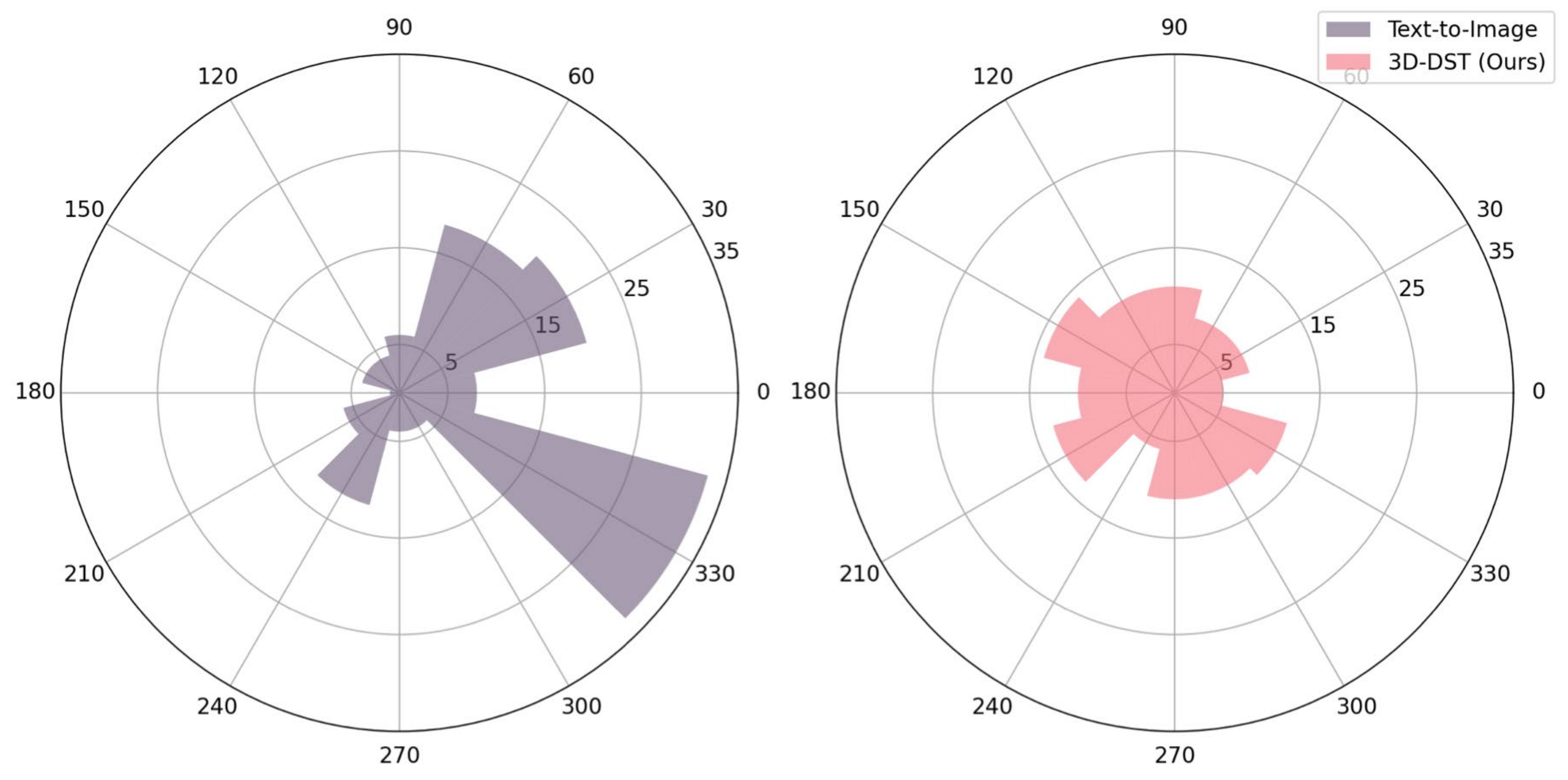
A photo of a airliner parking in the airport ready for departure

Our 3D-DST

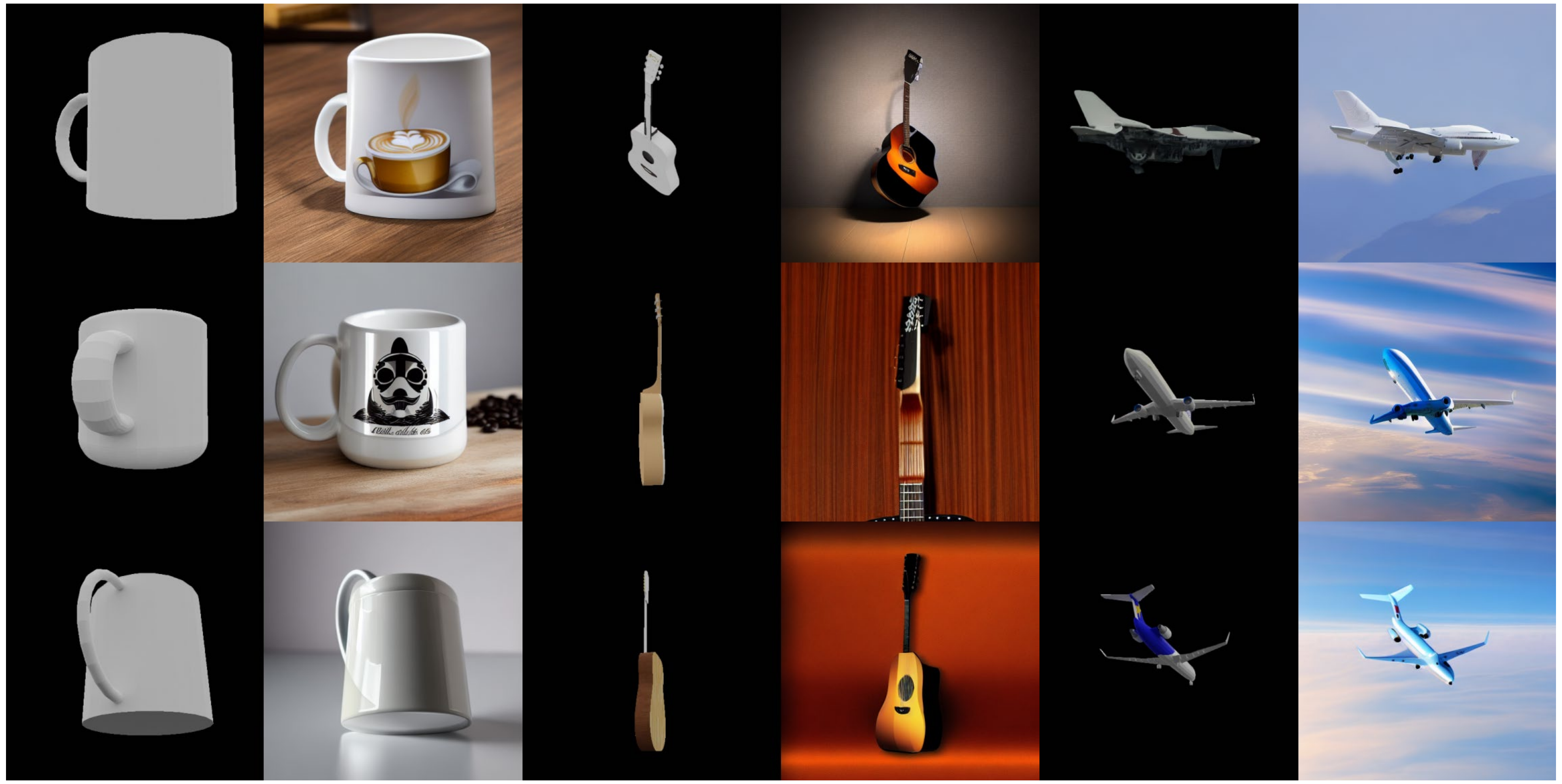

3D Visual Prompt
CAD model $\mathcal{C}$
render
camera parameters
ξ, K
rendered image
I_{render}
Canny edge
visual prompt $\mathcal{E}_{3D}$
Image Generation
ControlNet
U-Net ϵ
z_T
z_t
$z_{t\text{-}1}$
z_0
image decoder
$\mathcal{D}$
3D shape and pose
Text Prompt
"airplane"
class name w
"McDonnell Douglas aeroplane ..."
keywords k
LLM
"at a very high altitude ..."
generated descriptions
text encoder
θ
$\mathcal{T}$
text prompt

Removing Biases in Object Viewpoints

Viewpoint distribution of cars and buses from synthetic images generated by a text-to-image diffusion model and our 3D-DST.

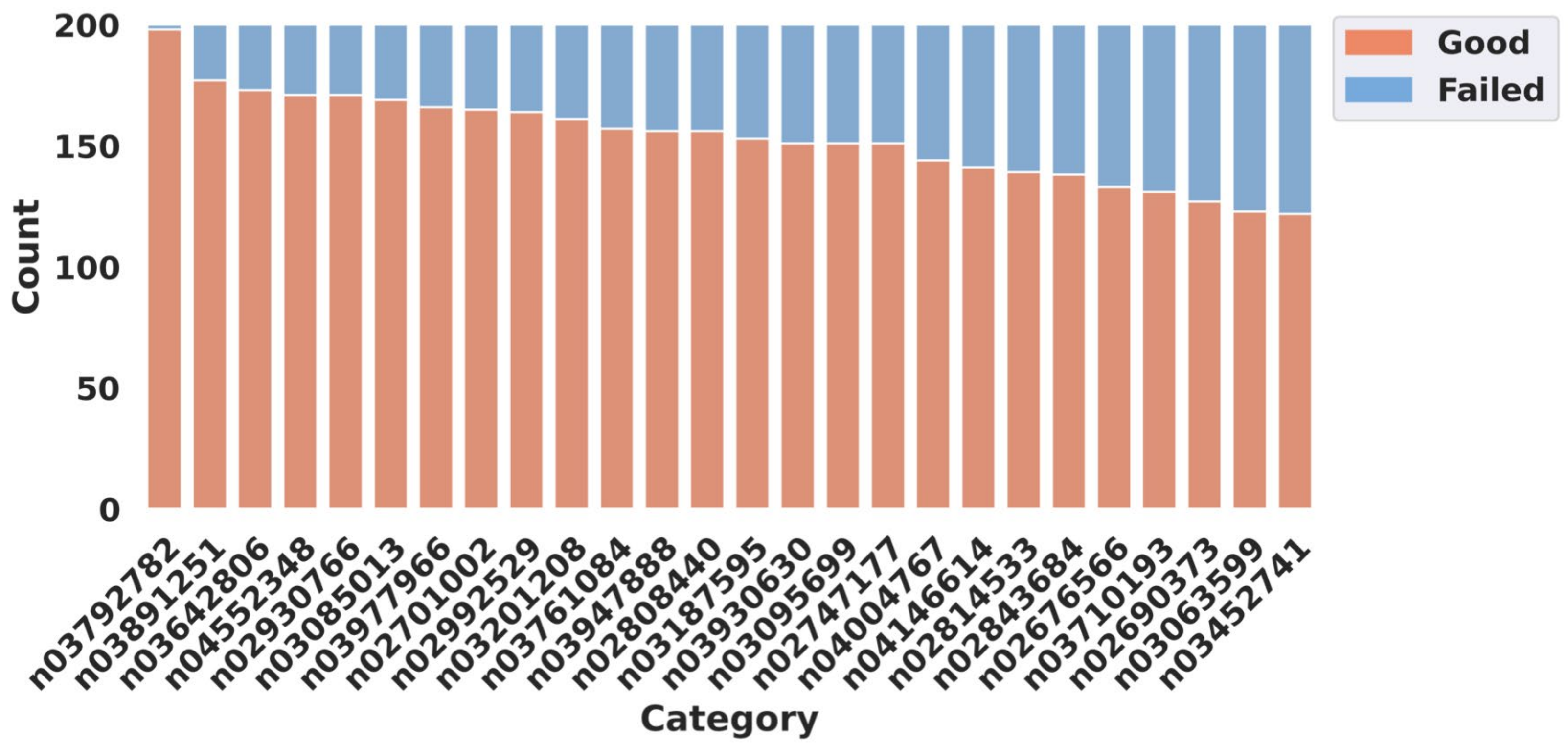


Analyzing 3D Consistencies



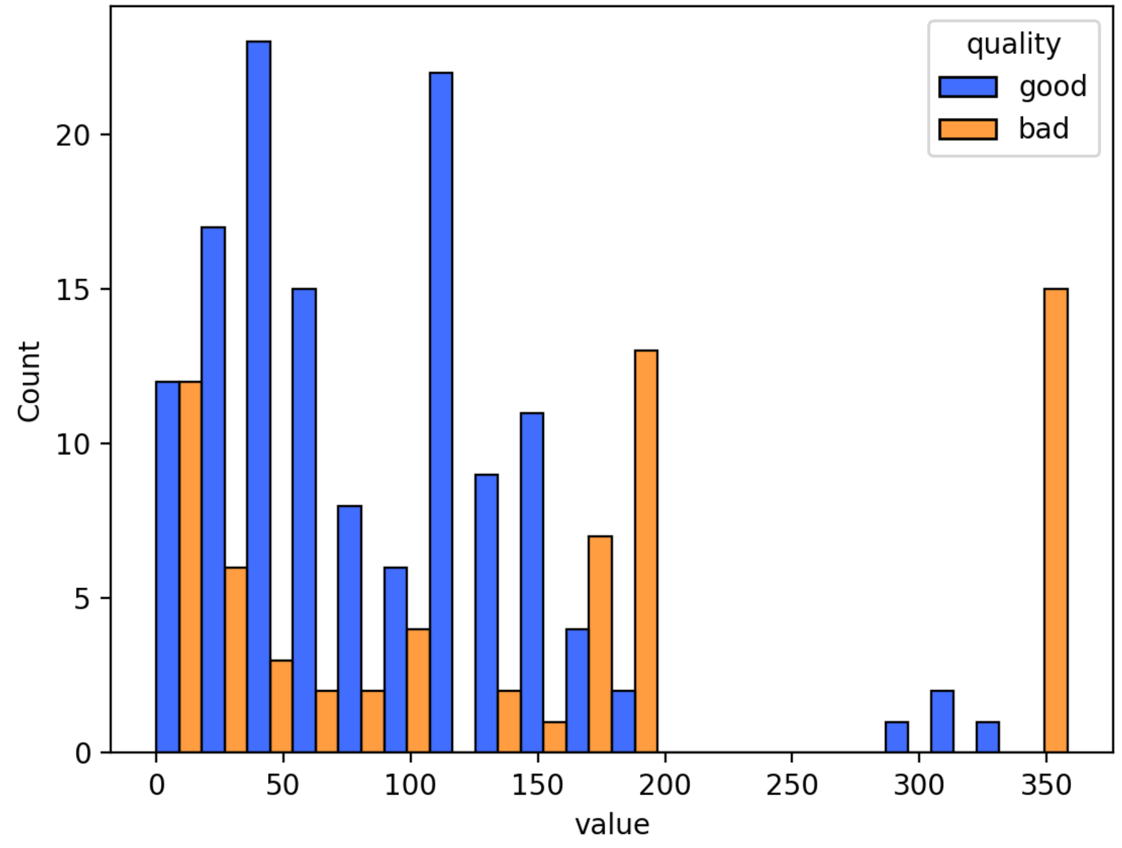
Analyzing 3D Consistencies

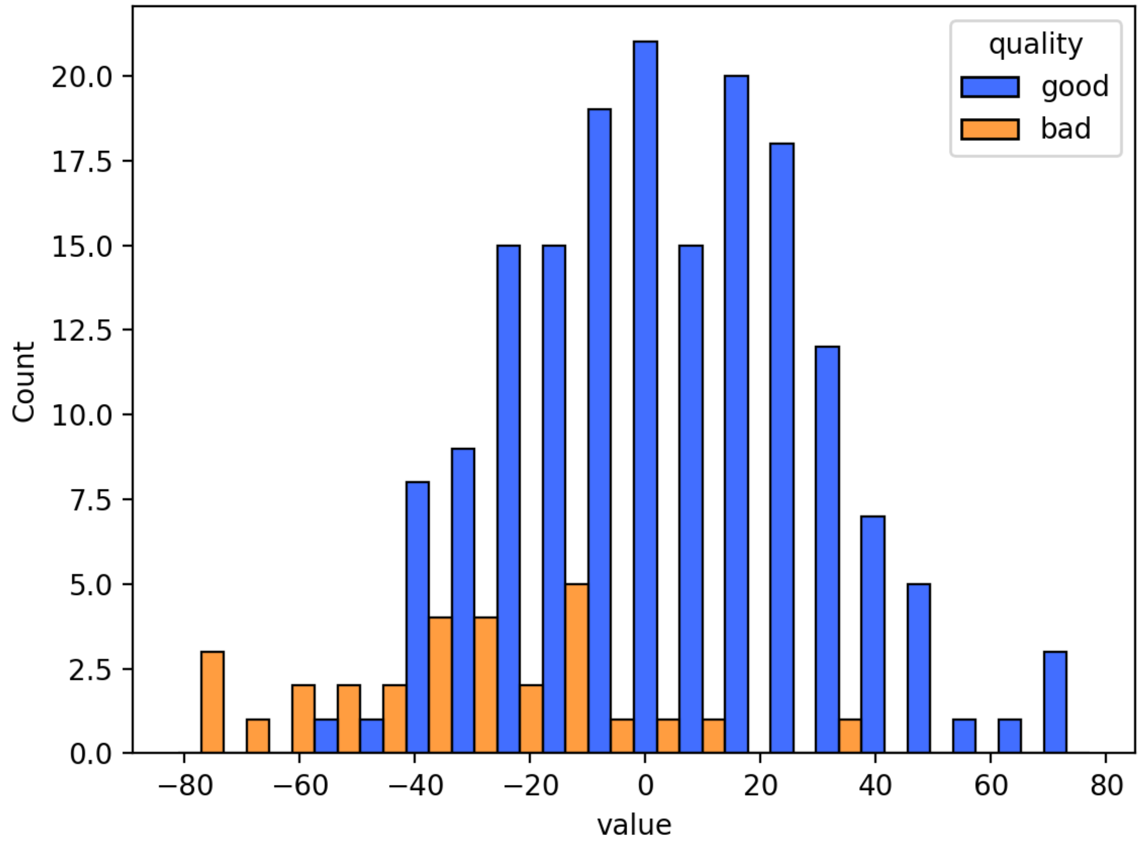
Human evaluation on the consistencies of 3D viewpoints show that about 75% of the images produced by our 3D-DST model have correct 3D annotations for downstream training.



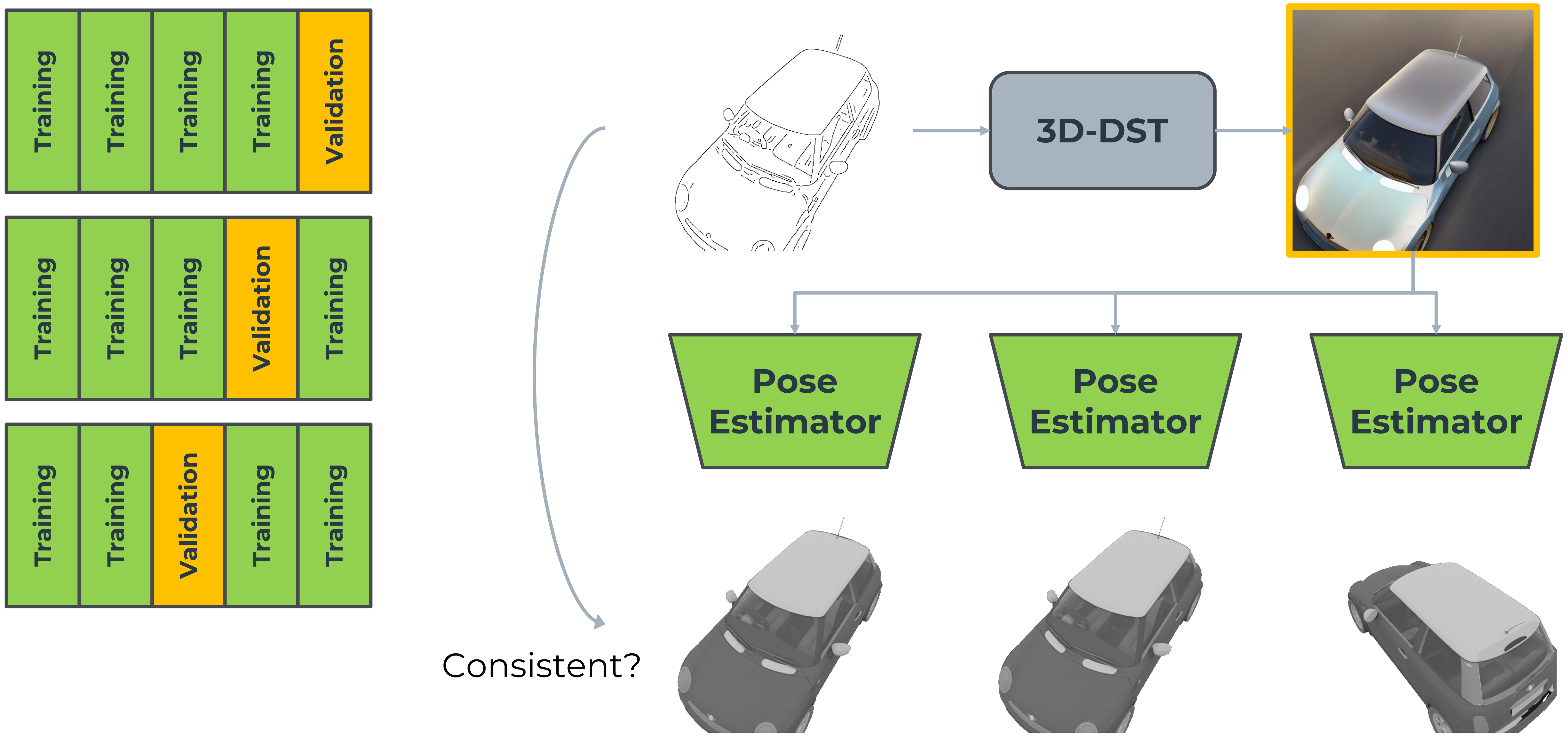
Failure Modes

Failure models such as guitars (top) from side view and taxi cabs (bottom) from bottom view.





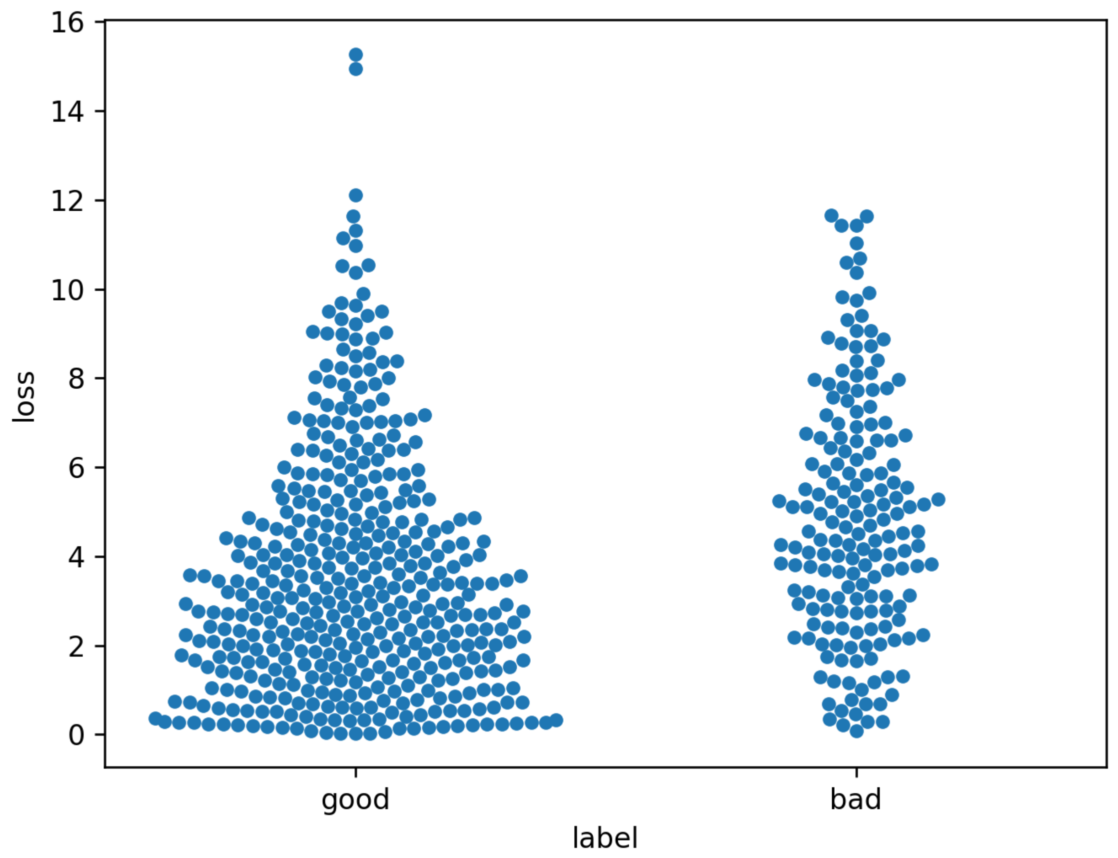
K-Fold Consistency Filter (KCF)



K-Fold Consistency Filter (KCF)

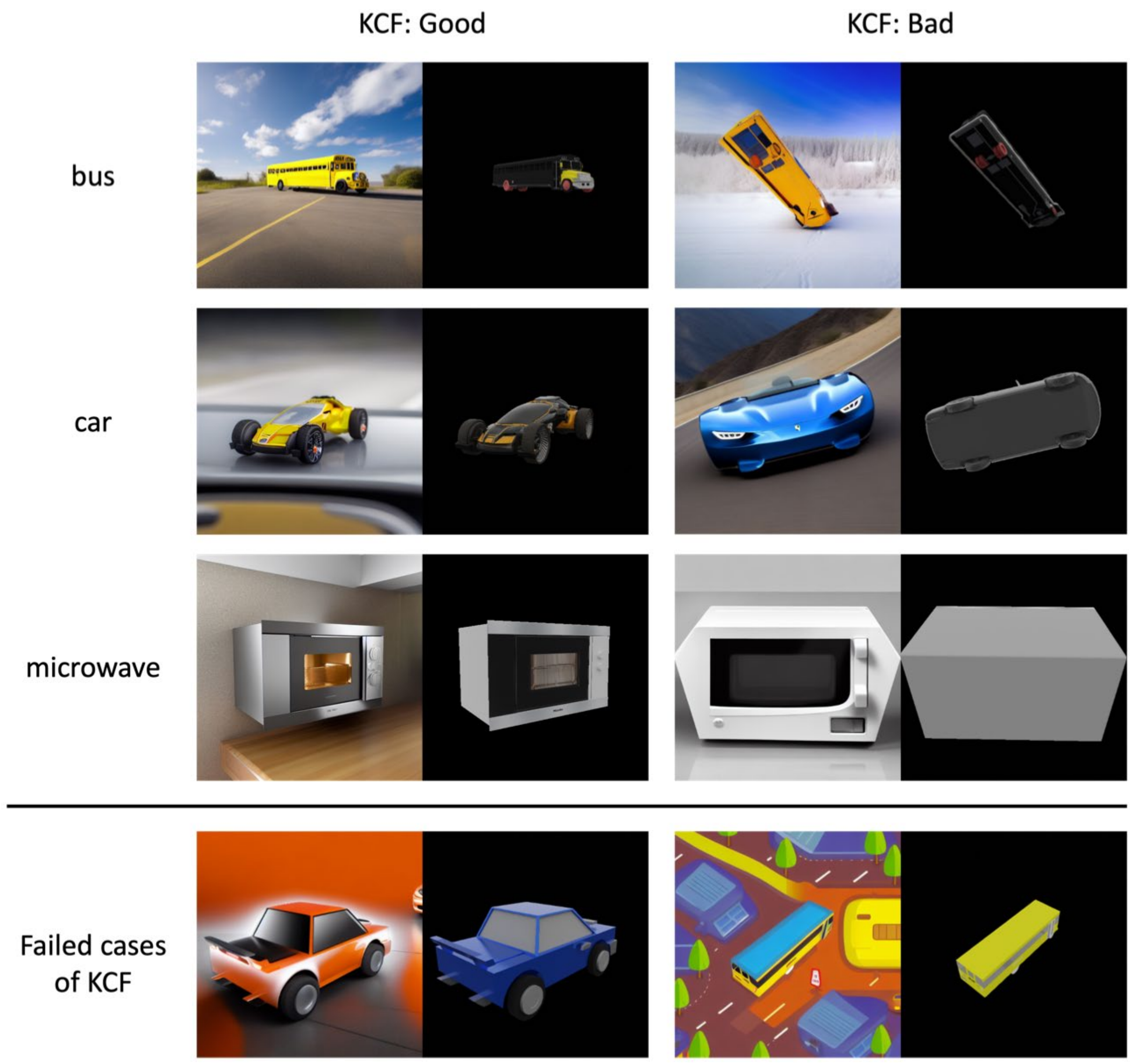
With KCF we can increase the success rate of our 3D-DST model by around 6%.

- School bus: 70% to 80%
- Guitar: 66% to 76%
- Taxi: 86% to 89%



K-Fold Consistency Filter (KCF)

The pose estimators are not robust. 🤦‍♀️



Main Results

- **Classification on ImageNet-100 and ImageNet-R.**
- 3D pose estimation on PASCAL3D+ and OOD-CV.
- 3D object detection on Omni3D.
- Ablation study on image generation.

Methods	In-distribution (ID)						
	DeiT-Ti	DeiT-S	DeiT-B	ConvNeXt-S	ConvNeXt-B	Swin-S	MAE-S
Baseline	83.84	84.56	84.28	91.34	91.50	90.78	88.69
w / Text2Img (2023)	84.58 (↑0.74)	84.80 (↑0.24)	84.10 (↓0.18)	91.26 (↓0.08)	91.28 (↓0.22)	90.42 (↓0.36)	88.78 (↑0.09)
w / 3D-DST (ours)	**87.36** (↑3.52)	**88.96** (↑4.40)	**88.11** (↑3.83)	**92.18** (↑0.84)	**92.44** (↑0.94)	**91.50** (↑0.72)	**90.47** (↑1.78)
Methods	**Out-of-distribution (OOD)**						
	DeiT-Ti	DeiT-S	DeiT-B	ConvNeXt-S	ConvNeXt-B	Swin-S	MAE-S
Baseline	49.96	49.61	50.53	67.19	66.40	54.99	65.18
w / Text2Img (2023)	52.58 (↑2.62)	50.83 (↑1.22)	49.61 (↓0.92)	67.50 (↑0.31)	67.15 (↑0.75)	56.34 (↑1.35)	67.28 (↑2.10)
w / 3D-DST (ours)	**56.12** (↑6.16)	**56.65** (↑7.04)	**56.74** (↑6.21)	**69.69** (↑2.50)	**69.21** (↑2.81)	**59.97** (↑4.98)	**68.42** (↑3.24)

Table 1: Image classification accuracy (%) on ImageNet-100 (ID) and ImageNet-R (OOD) using representative network architectures, ResNet and ViT. We compare the performances when models are (1) trained purely on the target dataset, (2) pre-trained on Text2Img (He et al., 2023) data, which does not have 3D control, and then finetuned on the target dataset, (3) pre-trained on 3D-DST data, and finetuned on the target dataset. Experiments show that our 3D-DST data can help boost the classification accuracy of both models on both ID and OOD cases by a large margin.

Main Results

- Classification on ImageNet-100 and ImageNet-R.
- **3D pose estimation on PASCAL3D+ and OOD-CV.**
- 3D object detection on Omni3D.
- Ablation study on image generation.

Methods	In-distribution (ID)		Out-of-distribution (OOD)	
	Acc@$\pi/6$	Acc@$\pi/18$	Acc@$\pi/6$	Acc@$\pi/18$
ResNet	82.33	52.60	50.38	23.38
ResNet *w/* AugMix (2020)	82.72 (↑0.39)	53.89 (↑1.29)	51.77 (↑1.39)	24.57 (↑1.19)
ResNet *w/* 3D-DST (ours)	84.22 (↑1.89)	56.52 (↑3.92)	52.75 (↑2.37)	25.70 (↑2.32)
NeMo (2021)	82.23	57.12	55.31	26.57
NeMo *w/* AugMix (2020)	83.11 (↑0.88)	58.22 (↑1.10)	56.38 (↑1.07)	26.63 (↑0.06)
NeMo *w/* 3D-DST (ours)	85.70 (↑3.47)	62.51 (↑5.39)	58.81 (↑3.50)	26.44 (↓0.13)

Table 4: Robust 3D pose estimation on ID (PASCAL3D+ & ObjectNet3D (Xiang et al., 2016; 2014)) and OOD (OOD-CV (Zhao et al., 2022)). We experiment with a classification-based pose estimation method, ResNet, and a 3D compositional model, NeMo (Wang et al., 2021). Experimental results demonstrate that our DST synthetic data can effectively improve 3D pose estimation performance on both ID and OOD benchmarks.

Main Results

- Classification on ImageNet-100 and ImageNet-R.
- 3D pose estimation on PASCAL3D+ and OOD-CV.
- **3D object detection on Omni3D.**
- Ablation study on image generation.

Methods	AP2D	AP3D
CubeRCNN (Brazil et al., 2023)	41.50	41.65
w / DST-3D (ours)	**42.34** (↑0.84)	**42.74** (↑1.09)
w / DST-3D + camera aug (ours)	**42.86** (↑1.36)	**43.19** (↑1.54)

Main Results

- Classification on ImageNet-100 and ImageNet-R.
- 3D pose estimation on PASCAL3D+ and OOD-CV.
- 3D object detection on Omni3D.
- **Ablation study on image generation.**

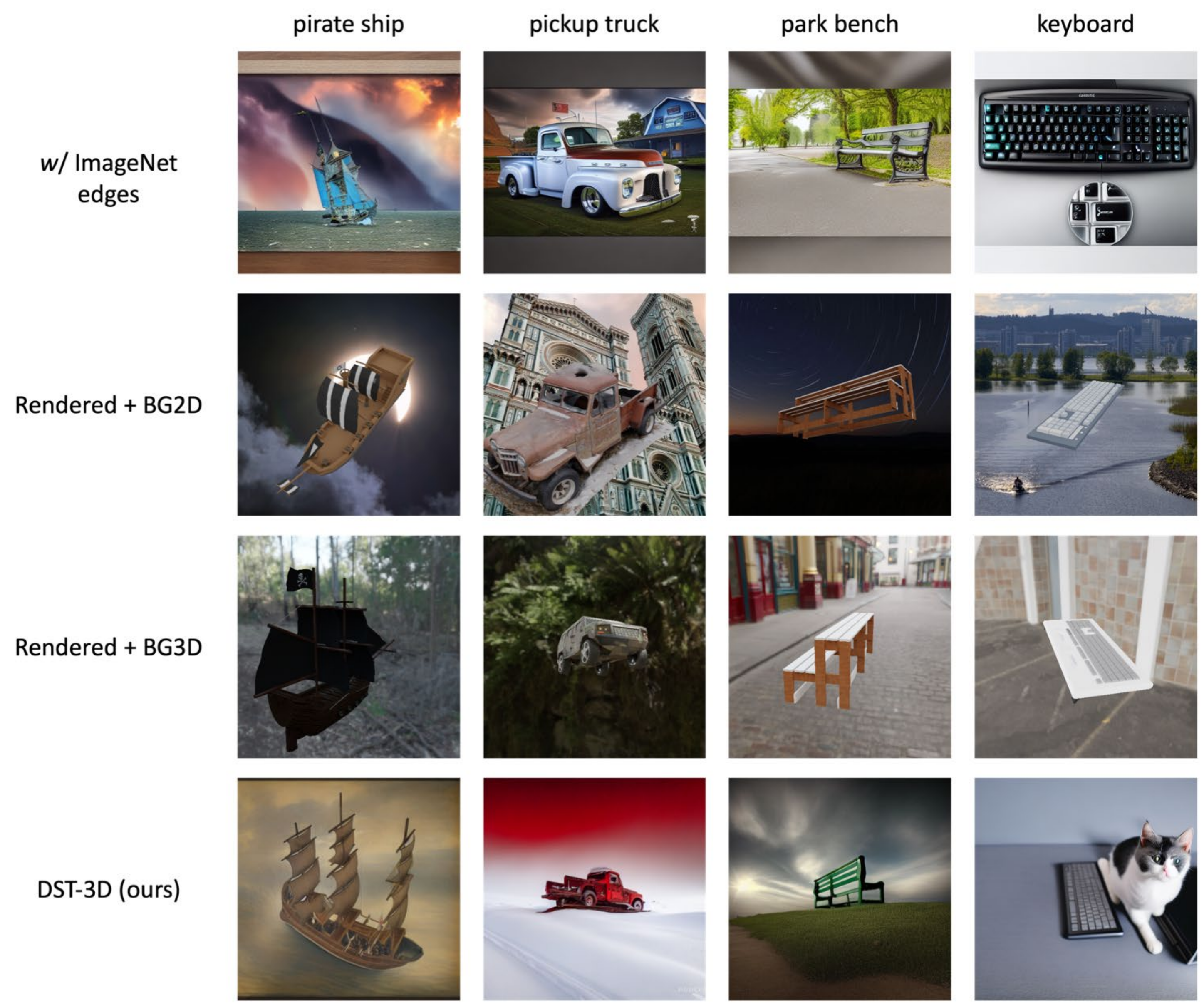


Figure 8: Qualitative examples of different data generation methods for ablation study. Please refer to Section D for specifics of different methods.

Code & Data Release

- **Code:**
  - Synthetic data rendering and generation
  - Prompt completion with LLM
  - K-fold consistency filtering (KCF)

- **Dataset:**
  - DST data for image classification
  - DST data for pose estimation
  - Aligned 3D models for each category

Future Work

3D-DST for animals.

1. 3D consistency?
2. SMAL consistency?
3. Background and foreground diversity

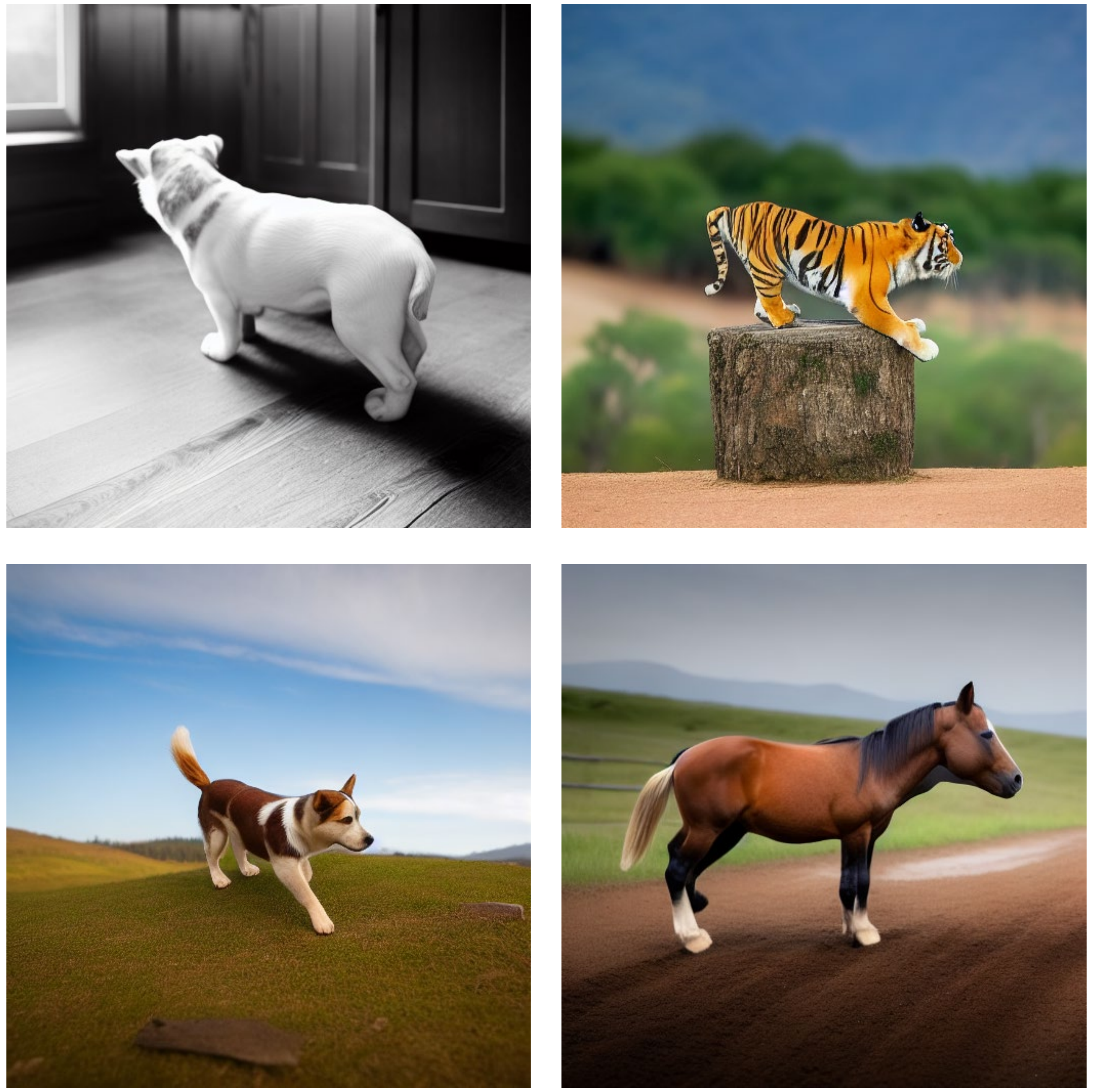


Figure credit: Jiawei Peng.

Future Work

3D-DST for OOD robustness evaluation.

1. Evaluating OOD robustness to snow, rain, fog, etc.

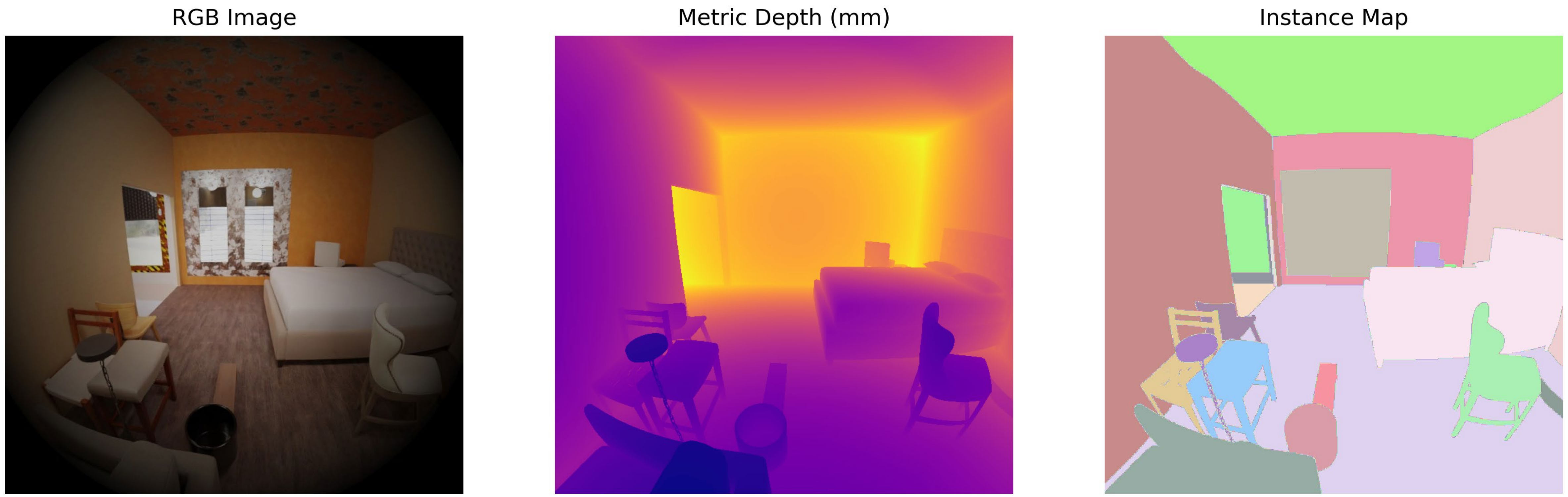
2. Continuous “sliders”

[1] Gandikota et al. Concept sliders: LoRA adapters for precise control in diffusion models. Preprint.

Future Work

3D-DST for multi-category multi-object scenes.

1. Broader applications
2. 3D consistency
3. Temporal consistency



[1] Avetisyan et al. SceneScript: Reconstructing scenes with an autoregressive structured language model. Preprint.

Thanks