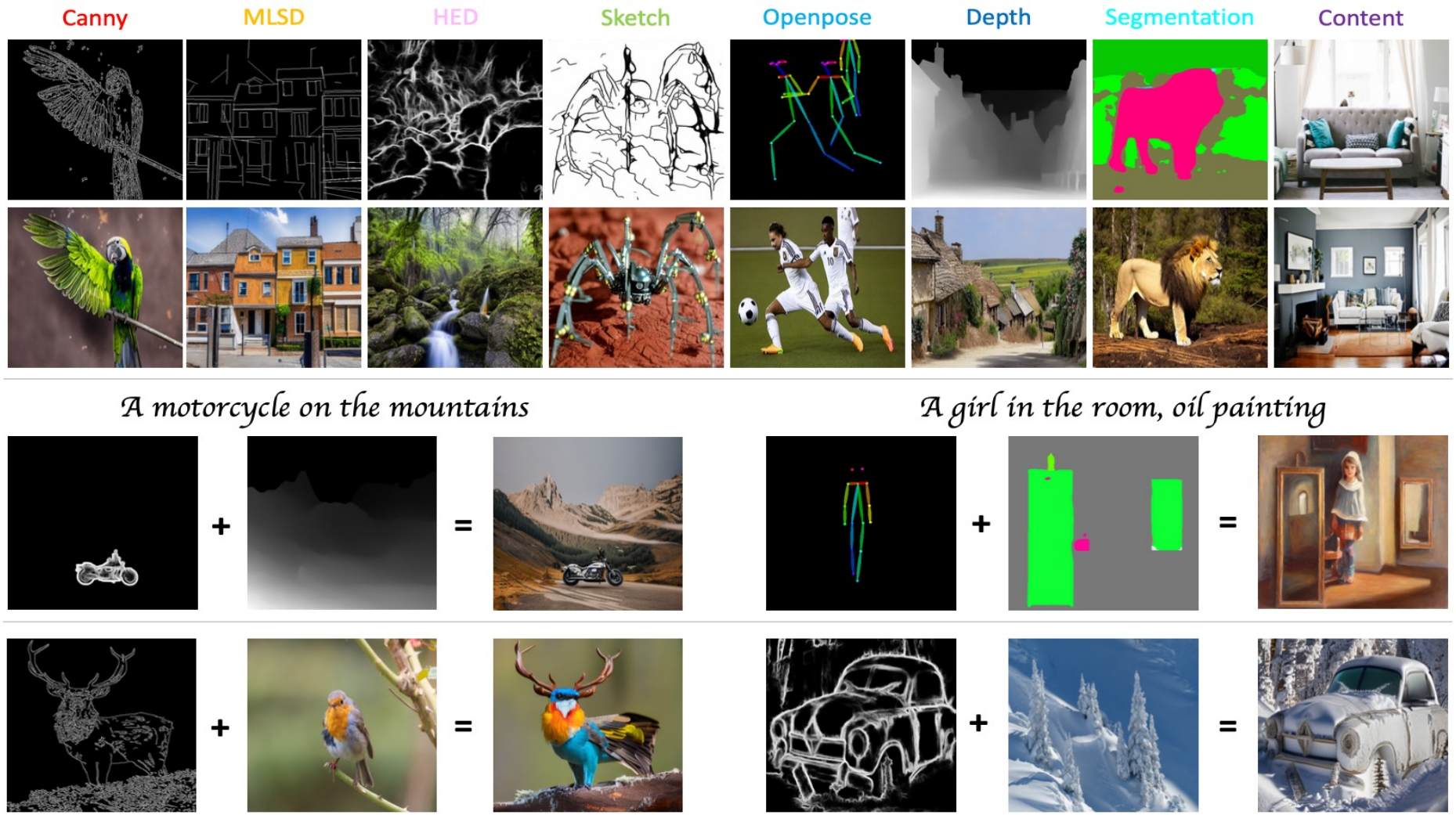
香 港 大 學
THE UNIVERSITY OF HONG KONG

Uni-ControlNet: All-in-One Control to Text-to-Image Diffusion Models

Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, Kwan-Yee K. Wong

Introduction - Overview

Uni-ControlNet: A **controllable diffusion model** that allows for the simultaneous utilization of different local controls and global controls in a **composable** manner within **one single model**



Introduction - Comparison

Comparisons of different controllable diffusion models

	Fine-tuning	Composable Control	Fine-tuning Cost	Adapter Number
Composer	✗	✔	-	-
ControlNet	✔	✔	N	N
GLIGEN	✔	✗	N	N
T2I-Adapter	✔	✔	$N(+1)$	$N(+1)$
Uni-ControlNet (Ours)	✔	✔	2	2

Uni-ControlNet

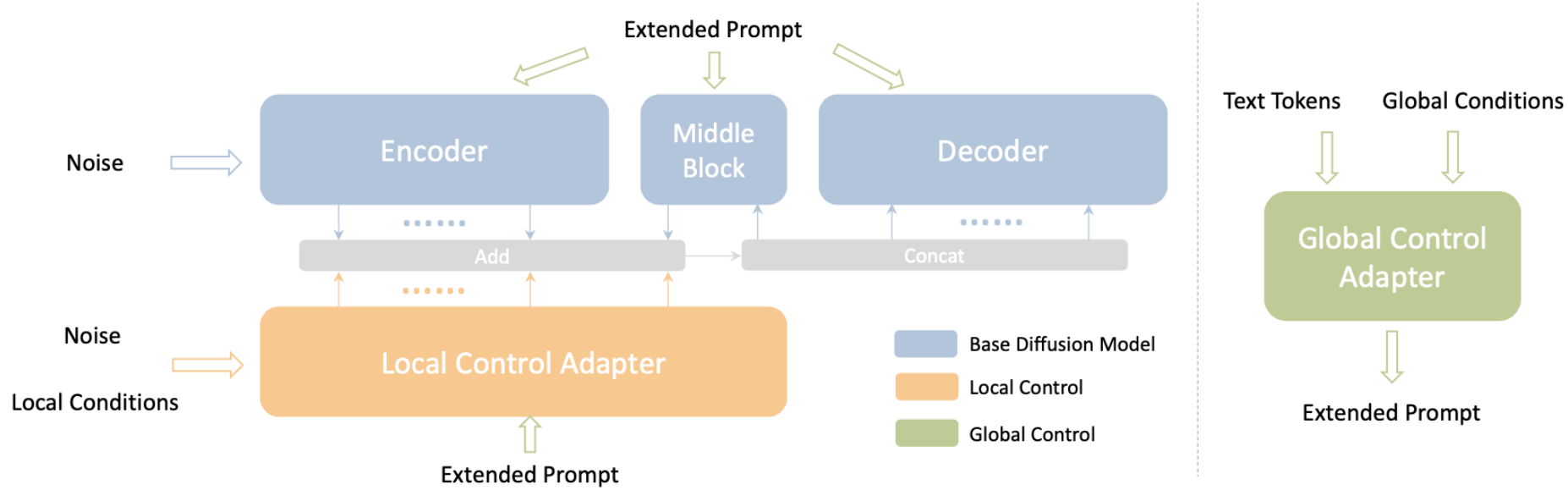
1. Only requires the fine-tuning of adapters upon frozen pre-trained models

2. Only requires 2 adapters, regardless of the number of local or global controls used

3. Great composable control of different conditions

Method - Framework

We design two adapters for local and global controls respectively



The input of the i- th block in the decoder in the base model

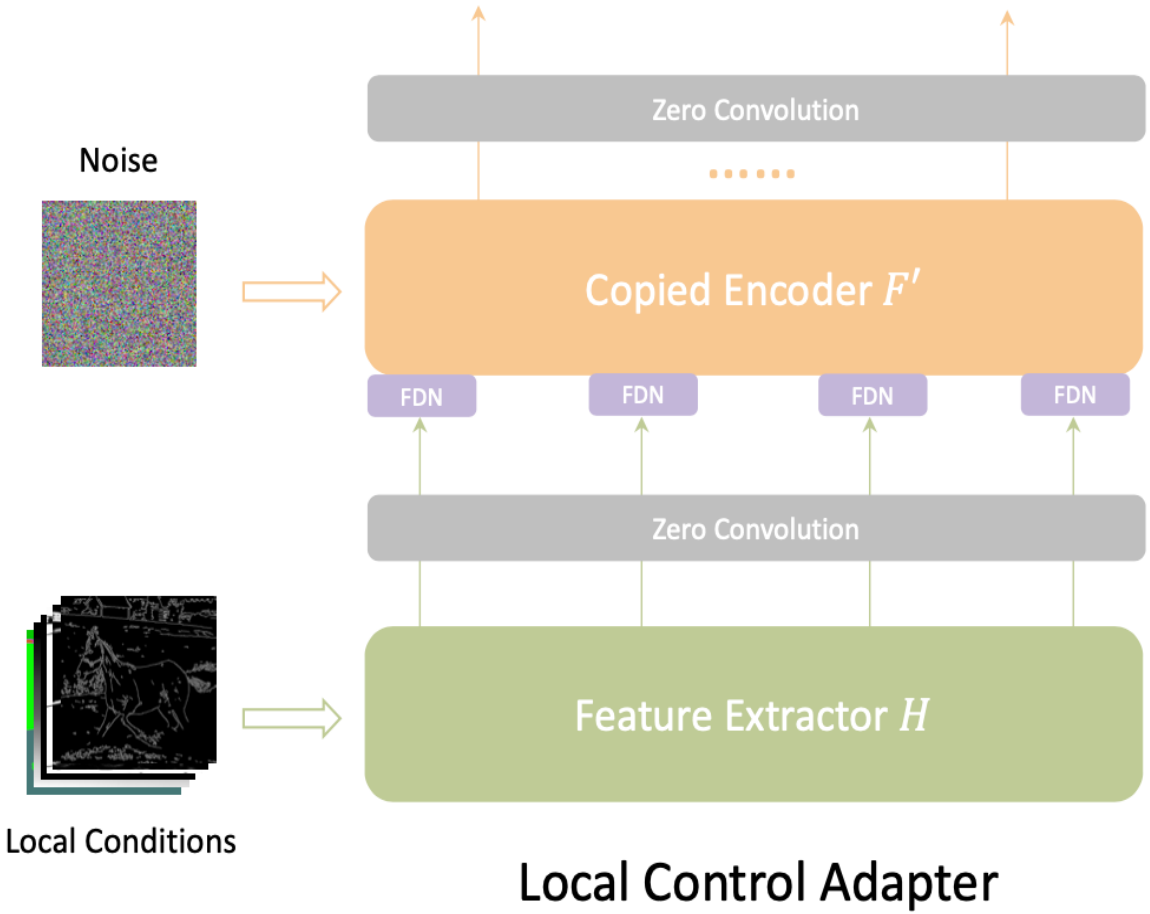
$$\begin{cases} concat(m, f_j) \\ concat(g_{i-1}, f_j) \end{cases} \Rightarrow \begin{cases} concat(m + m', f_j + zero(f_j')) & where \quad i = 1, \quad i + j = 13. \\ concat(g_{i-1}, f_j + zero(f_j')) & where \quad 2 \leq i \leq 12, \quad i + j = 13. \end{cases}$$

m is the output of the middle block, f_i and g_i denote the output of the i- th block in the encoder and decoder of base model

m', f' represents the output of the corresponding local control adapter

Method - Local Control Adapter

Multi-Scale Condition Injection



Local Control Adapter

4 Steps

1. Fix the weights of main SD and copy the structures and weights of the encoder and middle block
2. Concatenate all the local conditions in the channel dimension
3. Use a feature extractor to extract condition features at different resolutions
4. Inject the extracted features into the copied encoder in different scale through FDN

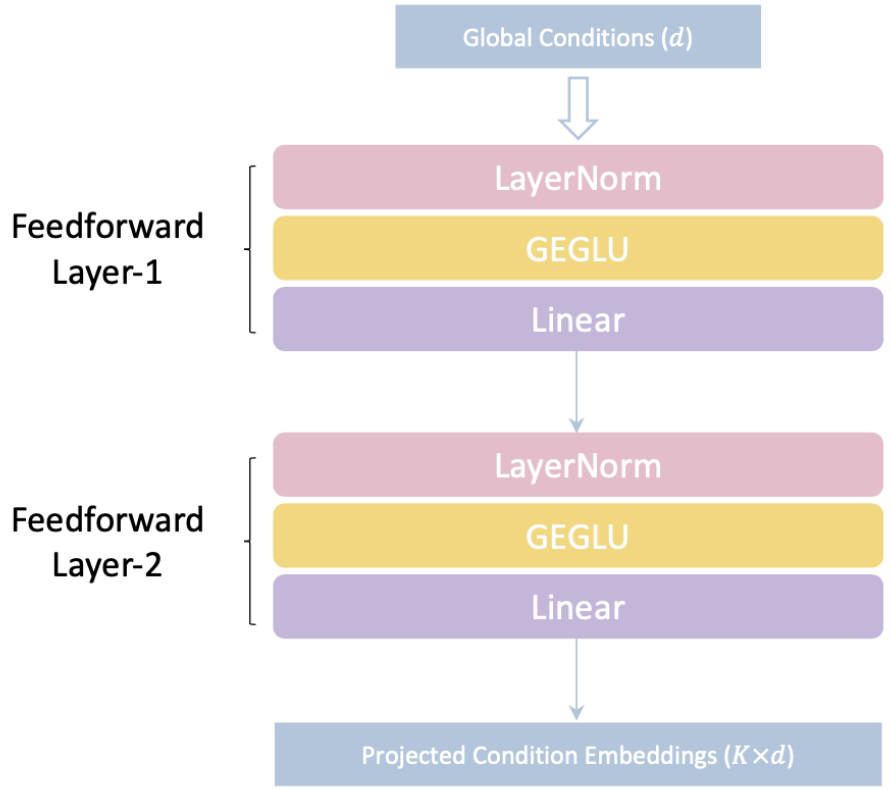
Method - Global Control Adapter

Global conditions: **1-dimensional embeddings**

Use an encoder to align the global controls with the text embeddings in SD

Reshape the projected condition embeddings into K global tokens

Concatenate them with the original text tokens to create an **extended prompt**



Serve as the input to all cross-attention layers in both main SD model and control adapters

Method - Training and Inference

Training: Fine-tune the local adapter and global adapter separately

Inference: Directly combine the 2 adapters together without further joint fine-tuning

Dropout Strategy:

Use a predefined probability to randomly dropout each condition

Meanwhile, keep an additional probability to deliberately keep or drop all conditions

For the dropped conditions, the value of the corresponding input channels is set to 0

Experiment - Settings

Conditions:

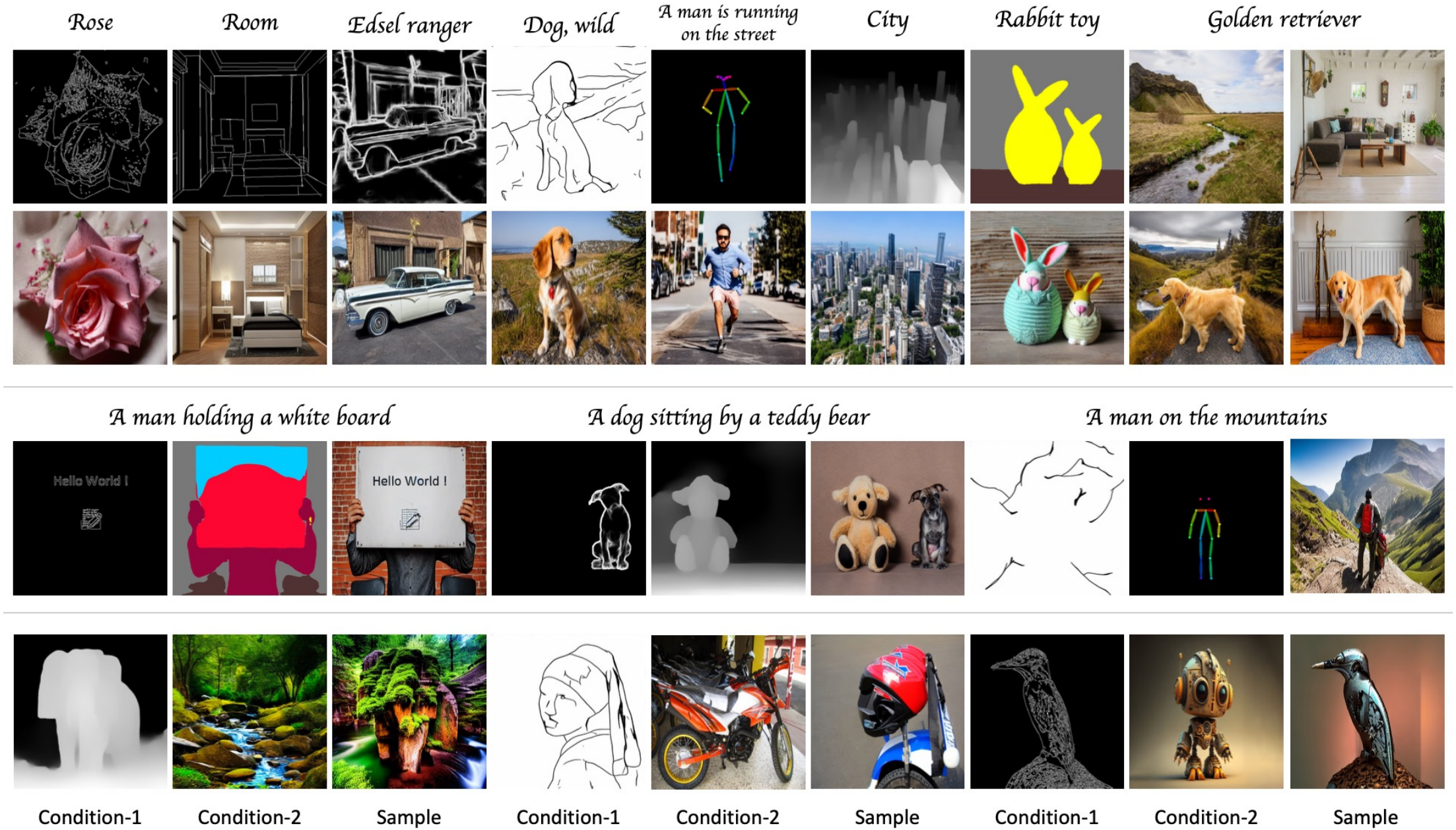
Canny edge, MLSD edge, HED boundary, Sketch, Openpose, Midas depth, Segmentation mask

Content (CLIP image embedding)

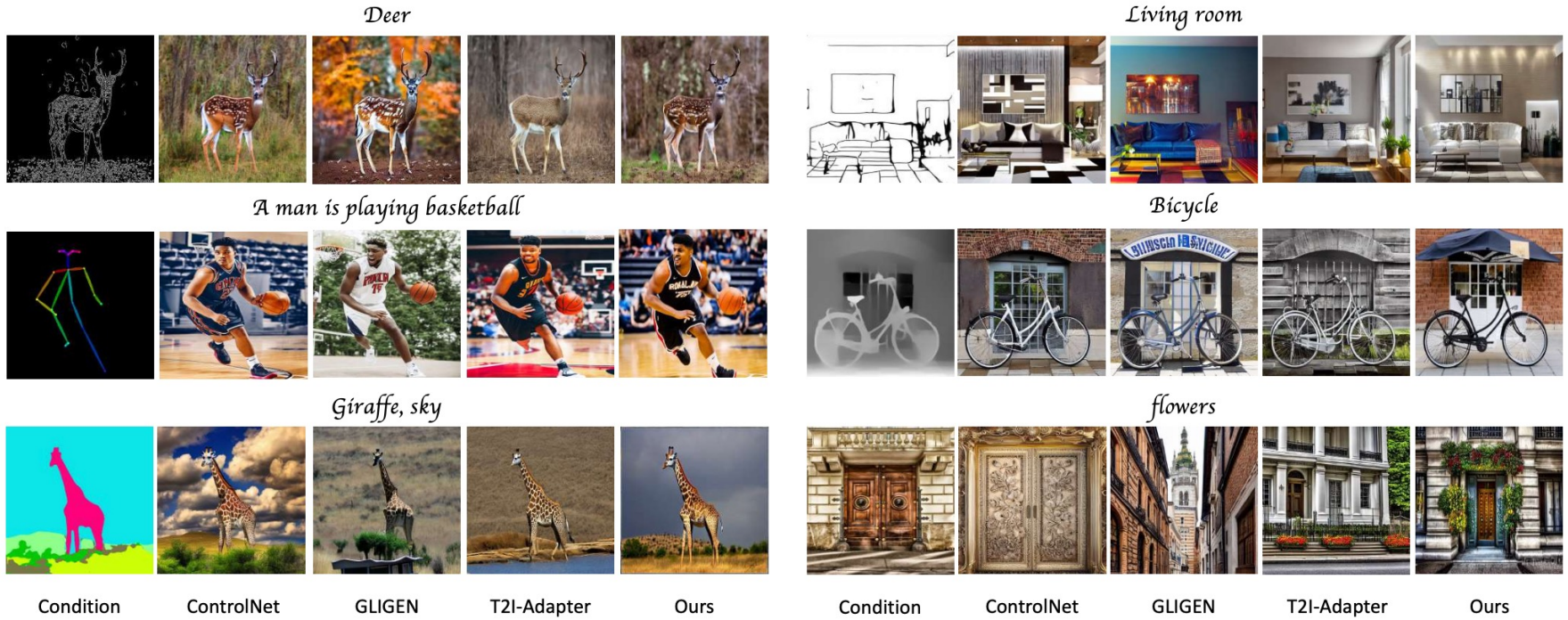
Training: Randomly sample 10 million text-image pairs from the LAION dataset and fine-tune the adapters for 1 epoch. We use the AdamW optimizer with a learning rate of 10^{-5}.

Inference: Use DDIM for sampling, with the number of timesteps set to 50 and the CFG scale set to 7.5

Experiment - Generation Results



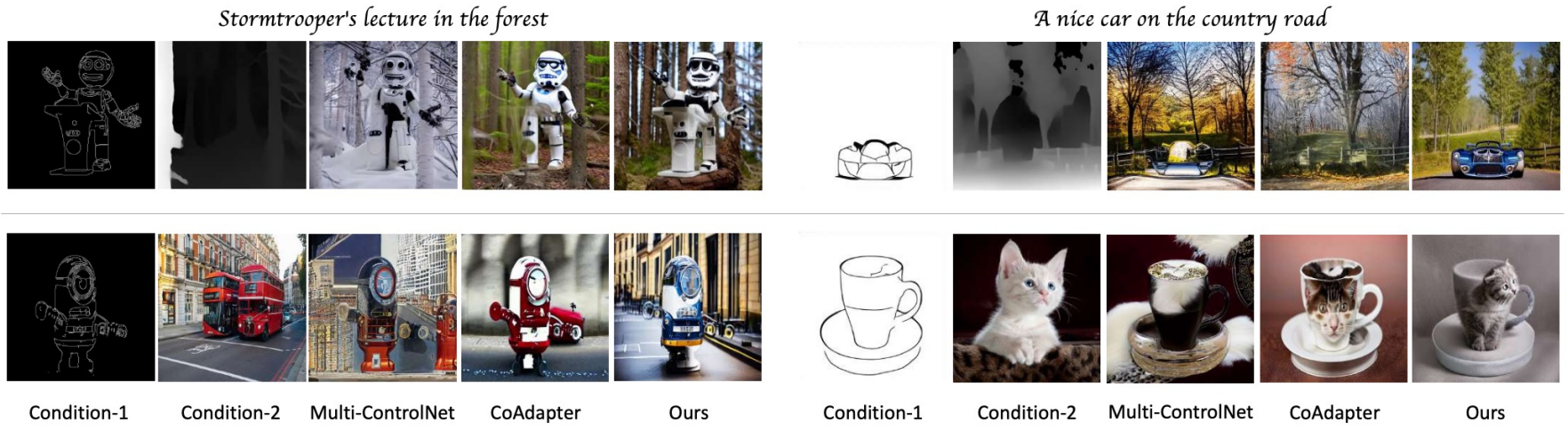
Experiment - Qualitative Comparison



Single Condition

ControlNet, GLIGEN, T2I-Adapter

Uni-ControlNet



Multi-Conditions

Multi-ControlNet, CoAdapter

Uni-ControlNet

Experiment - Quantitative Comparison

Validation Set: COCO2017, **Resolution:** 512 × 512, **Number of Generated Samples:** 5k

As each image has multiple captions in COCO2017, we randomly select one caption per image

Table 2: FID on different controllable diffusion models. The best results are in **bold**.

	Canny	MLSD	HED	Sketch	Pose	Depth	Segmentation	Style\Content
ControlNet	18.90	31.36	26.59	22.19	27.84	21.25	**23.08**	31.17
GLIGEN	24.74	-	28.57	-	**24.57**	21.46	27.39	25.12
T2I-Adapter	18.98	-	-	**18.83**	29.57	21.35	23.84	28.86
Ours	**17.79**	**26.18**	**17.86**	20.11	26.61	**21.20**	23.40	**23.98**

Table 3: Quantitative evaluation of the controllability. The best results are in **bold**.

	Canny (SSIM)	MLSD (SSIM)	HED (SSIM)	Sketch (SSIM)	Pose (mAP)	Depth (MSE)	Segmentation (mIoU)	Style\Content (CLIP Score)
ControlNet	0.4828	**0.7455**	0.4719	0.3657	0.4359	**87.57**	**0.4431**	0.6765
GLIGEN	0.4226	-	0.4015	-	0.1677	88.22	0.2557	0.7458
T2I-Adapter	0.4422	-	-	0.5148	**0.5283**	89.82	0.2406	0.7078
Ours	**0.4911**	0.6773	**0.5197**	**0.5923**	0.2164	91.05	0.3160	**0.7753**

香 港 大 學
THE UNIVERSITY OF HONG KONG

Thank you !

https://arxiv.org/abs/2305.16322

https://github.com/ShihaoZhaoZSH/Uni-ControlNet