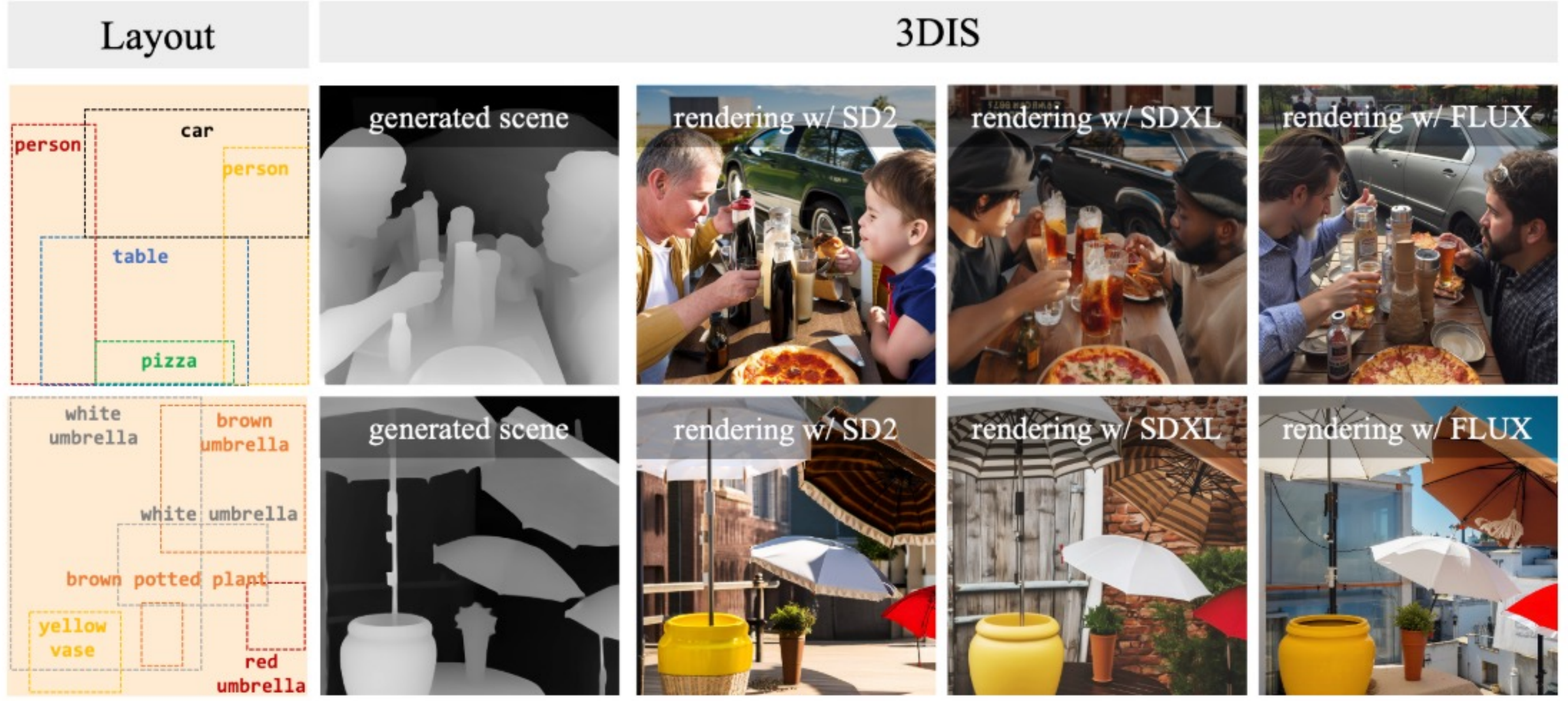
3DIS: DEPTH-DRIVEN DECOUPLED IMAGE SYNTHESIS FOR UNIVERSAL MULTI-INSTANCE GENERATION

Dewei Zhou[1*], Ji Xie[1*], Zongxin Yang[2*], Yi Yang[1]

[1] ReLER, CCAI, Zhejiang University [2] Harvard University * Equal Contribution

Code

Introduction



TL;DR: 3DIS allows users to perform MIG using various foundational models (including SD1.5, SD2, SDXL, FLUX), where they can specify the position and attributes of each instance in one image.

Background:

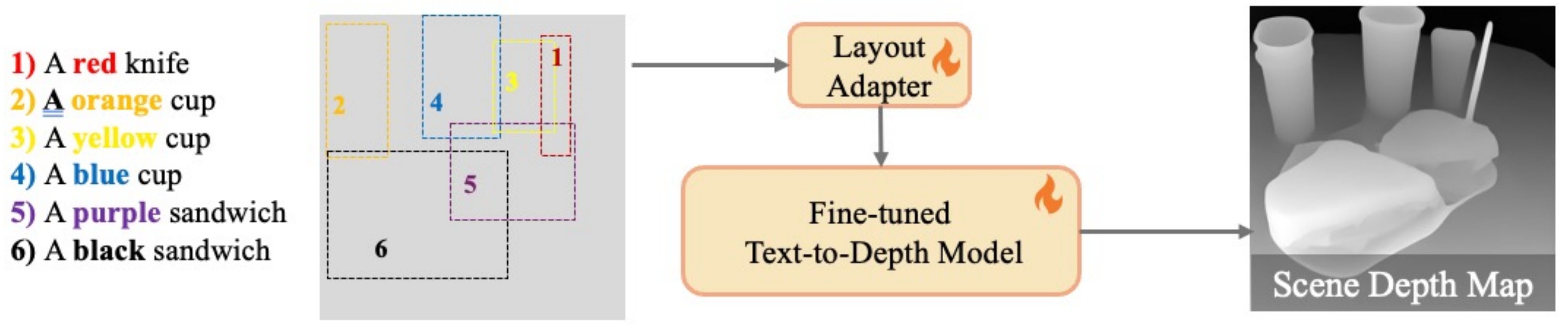
- Multi-Instance Generation (MIG) allows users to define the locations and attribute of multiple instances in the generated image.
- The mainstream MIG methods involve training an adapter directly on the generative model to control both the position and attributes of each instance. Whenever a more powerful base model emerges, the adapter needs to be retrained, which is resource-intensive.

Motivations:

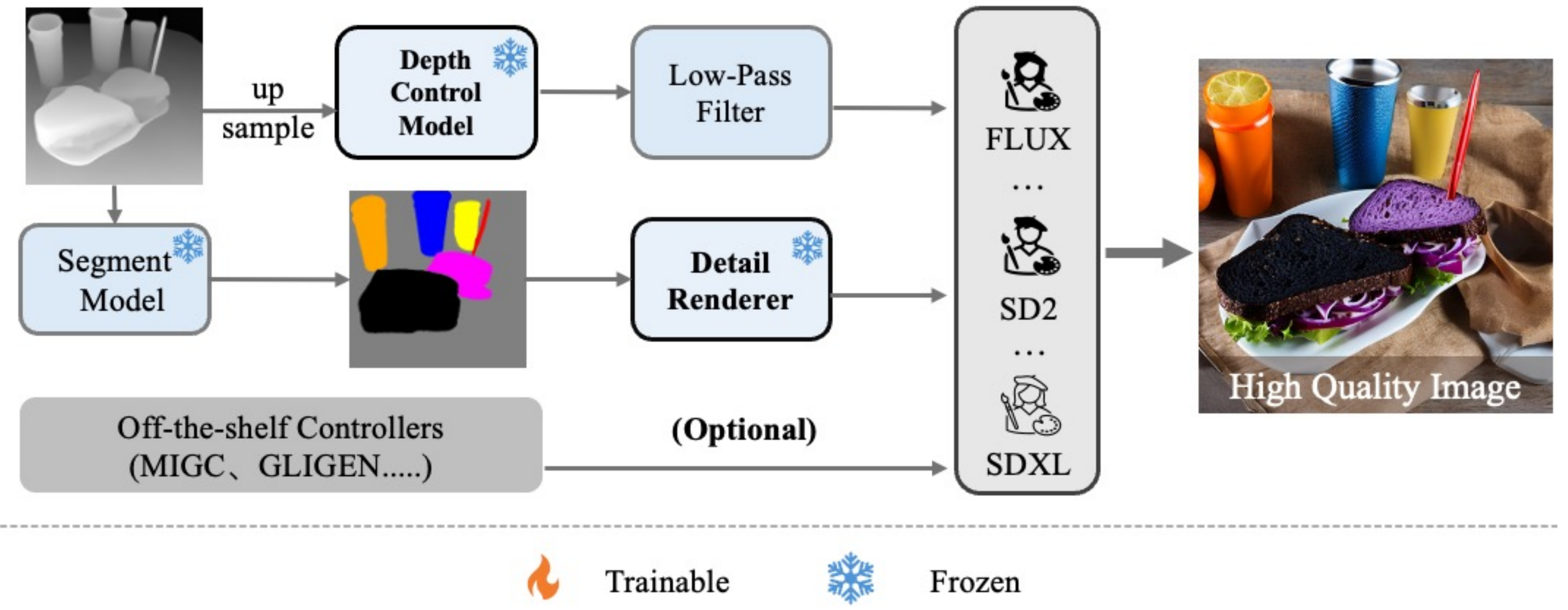
- Decoupling MIG into scene building and detail rendering. **We only train one Layout-to-Depth model** to control instance positions, and then use a training-free approach to **render details using multiple models.**

Methodology: 3DIS

Stage 1 Generate a coarse-grained scene depth map



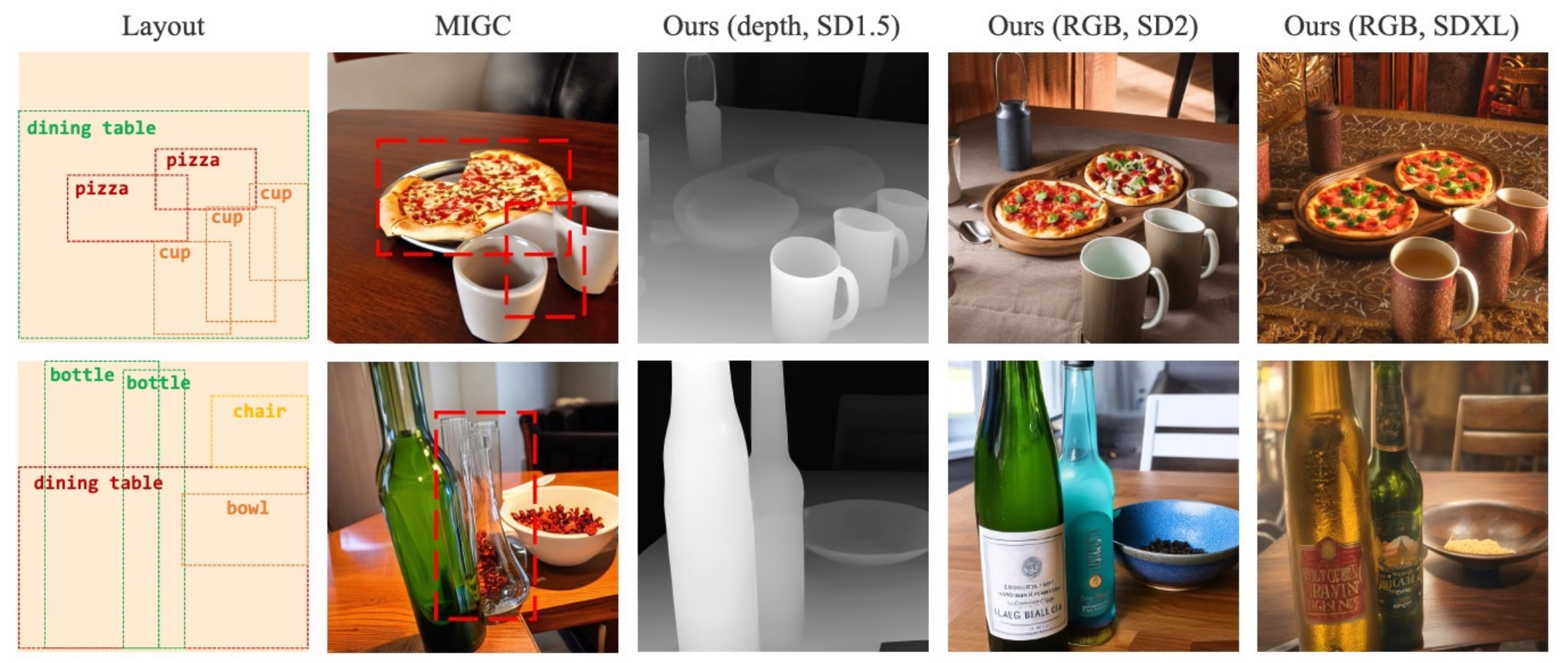
Stage 2 Rendering fine-grained instance details



- This method divides the multi-instance generation process into two stages.
- First, a Layout-to-depth network is trained to generate a scene depth map.
- Secondly, we use Depth-ControlNet to precisely position each instance and a Detail Renderer to accurately render the attributes of each instance without any training.

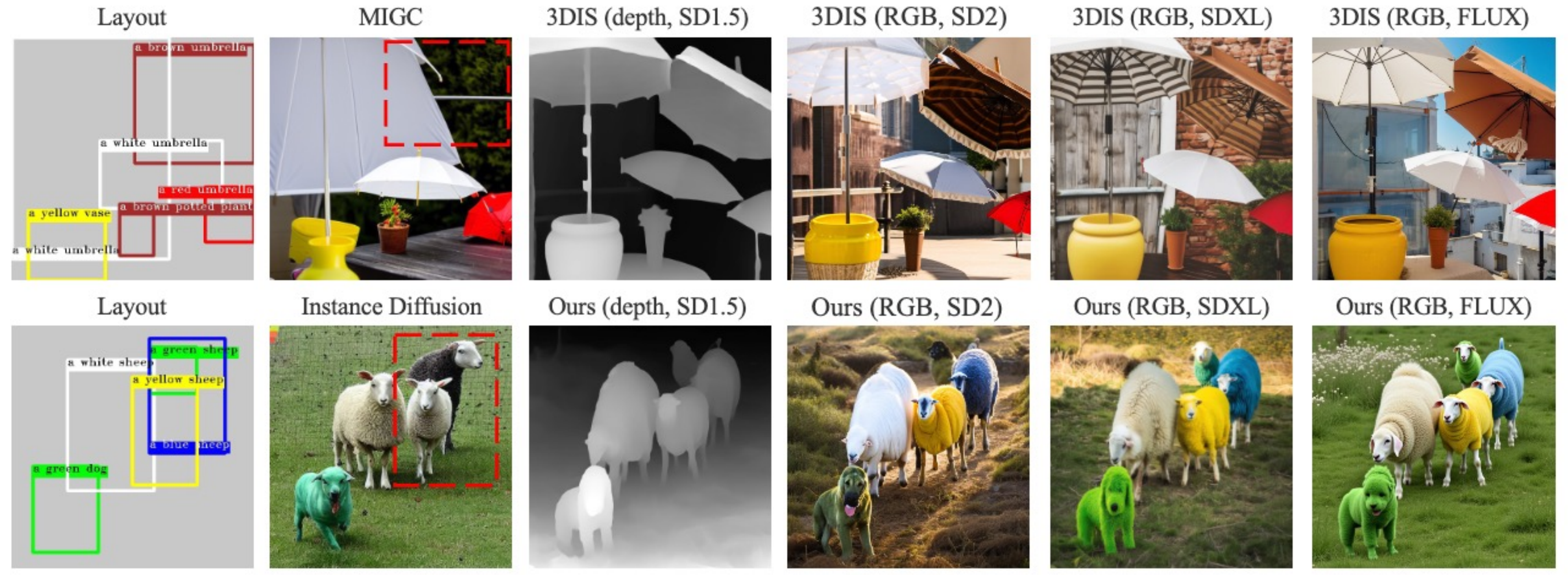
Experiments & Application

Results on COCO-POS

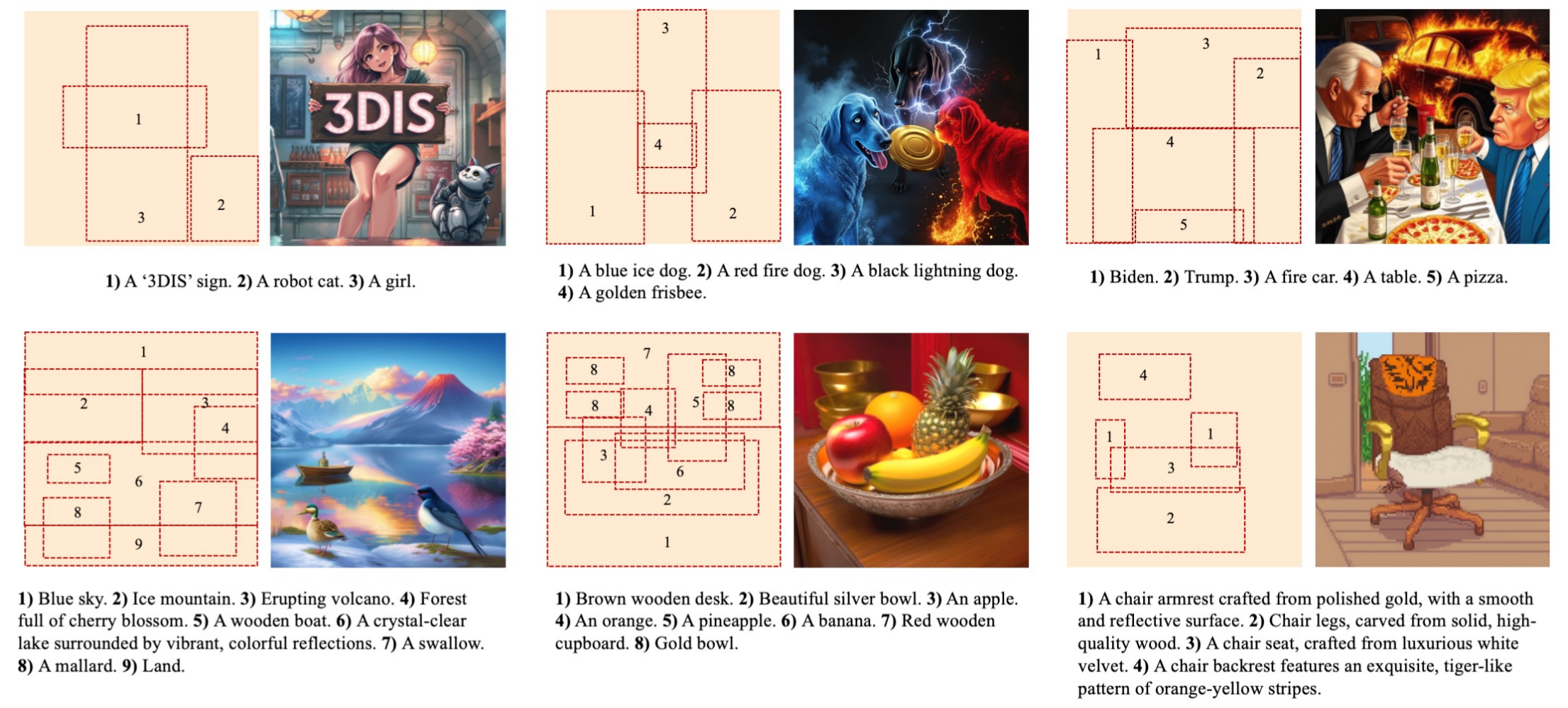


	Layout Accuracy			Instance Accuracy			Image Quality	
Method	$AP\uparrow$	$AP_{75}\uparrow$	$AP_{50}\uparrow$	$SR_{inst}\uparrow$	MIoU	$CLIP\uparrow$	$SR_{img}\uparrow$	$FID\downarrow$
BoxDiff [ICCV23]	3.15	2.12	10.92	22.74	27.28	18.82	0.53	25.15
MultiDiff [ICML23]	6.37	4.24	13.22	28.75	34.17	20.12	0.80	33.20
GLIGEN [CVPR23]	38.49	40.75	63.79	83.31	70.14	19.61	40.13	26.80
MIGC [CVPR24]	45.03	46.15	80.09	83.37	71.92	20.07	43.25	24.52
3DIS	**56.83**	**62.40**	**82.29**	**84.71**	**73.32**	**20.84**	**46.50**	**23.24**
vs. prev. SoTA	**+11.8**	**+16.3**	**+2.2**	**+1.3**	**+1.4**	**+0.8**	**+3.3**	**+1.3**

Results on COCO-MIG



MIG using various models（e.g., FLUX and SDXL）



1) A '3DIS' sign. 2) A robot cat. 3) A girl.

1) A blue ice dog. 2) A red fire dog. 3) A black lightning dog. 4) A golden frisbee.

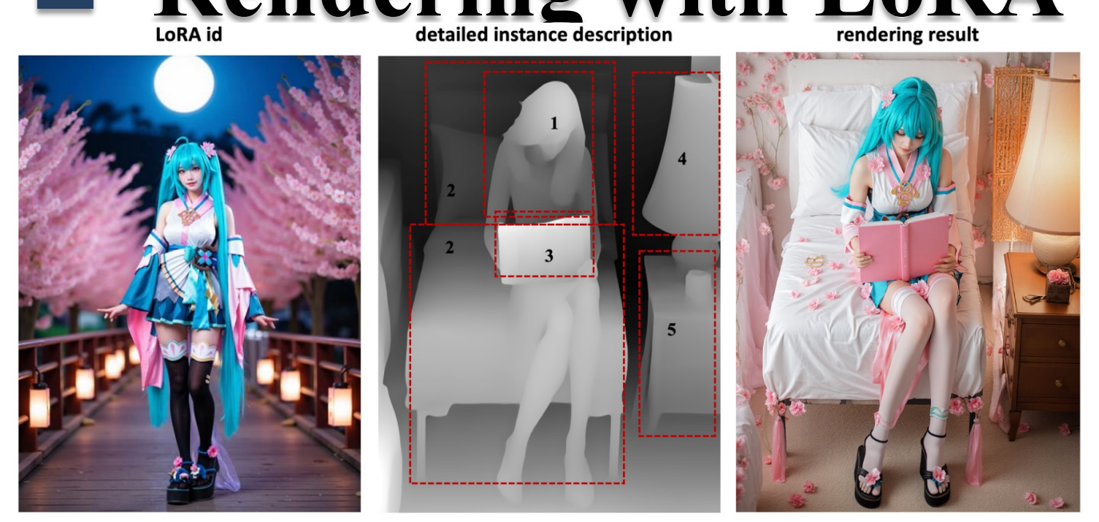
1) Biden. 2) Trump. 3) A fire car. 4) A table. 5) A pizza.

1) Blue sky. 2) Ice mountain. 3) Erupting volcano. 4) Forest full of cherry blossom. 5) A wooden boat. 6) A crystal-clear lake surrounded by vibrant, colorful reflections. 7) A swallow. 8) A mallard. 9) Land.

1) Brown wooden desk. 2) Beautiful silver bowl. 3) An apple. 4) An orange. 5) A pineapple. 6) A banana. 7) Red wooden cupboard. 8) Gold bowl.

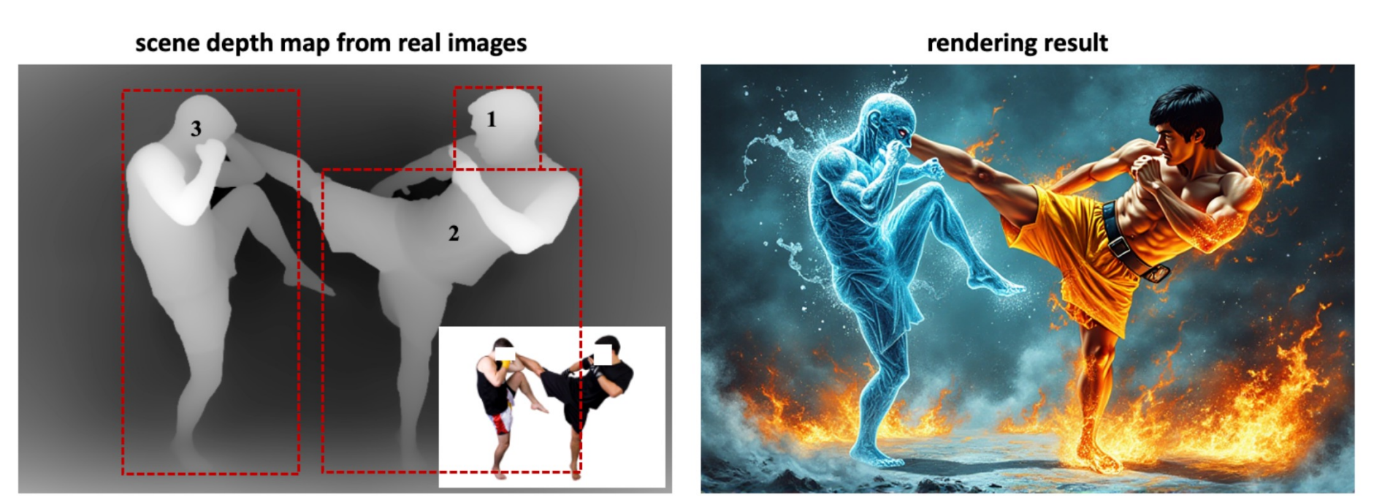
1) A chair armrest crafted from polished gold, with a smooth and reflective surface. 2) Chair legs, carved from solid, high-quality wood. 3) A chair seat, crafted from luxurious white velvet. 4) A chair backrest features an exquisite, tiger-like pattern of orange-yellow stripes.

Rendering with LoRA

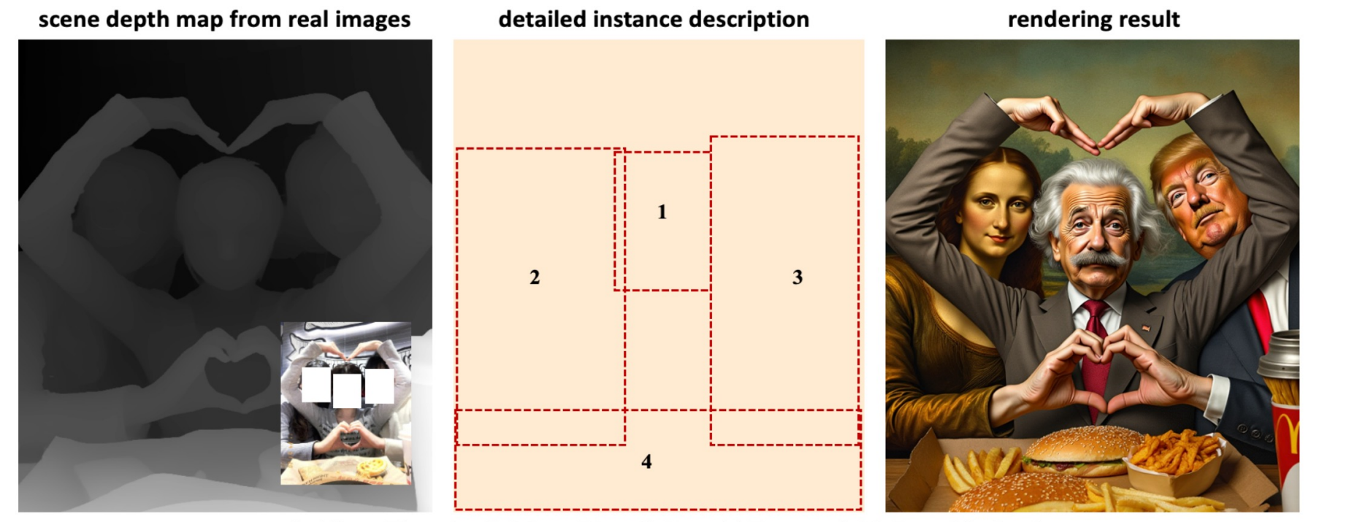


1) A girl. 2) cherry blossoms are placed on the white silk bed. 3) a pink book. 4) a paper table lamp. 5) a wooden table.

Rendering Real Scene Depth Maps

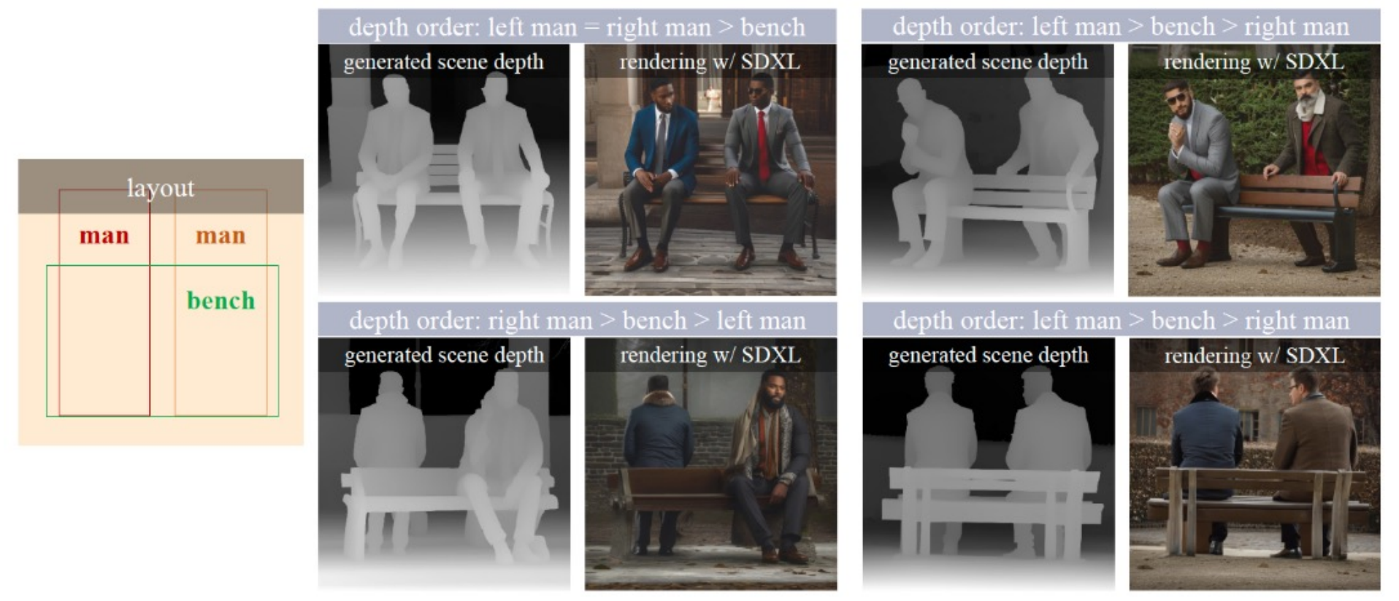


1) Bruce Lee. 2) Naked body surrounded by flames. 3) icy zombie.

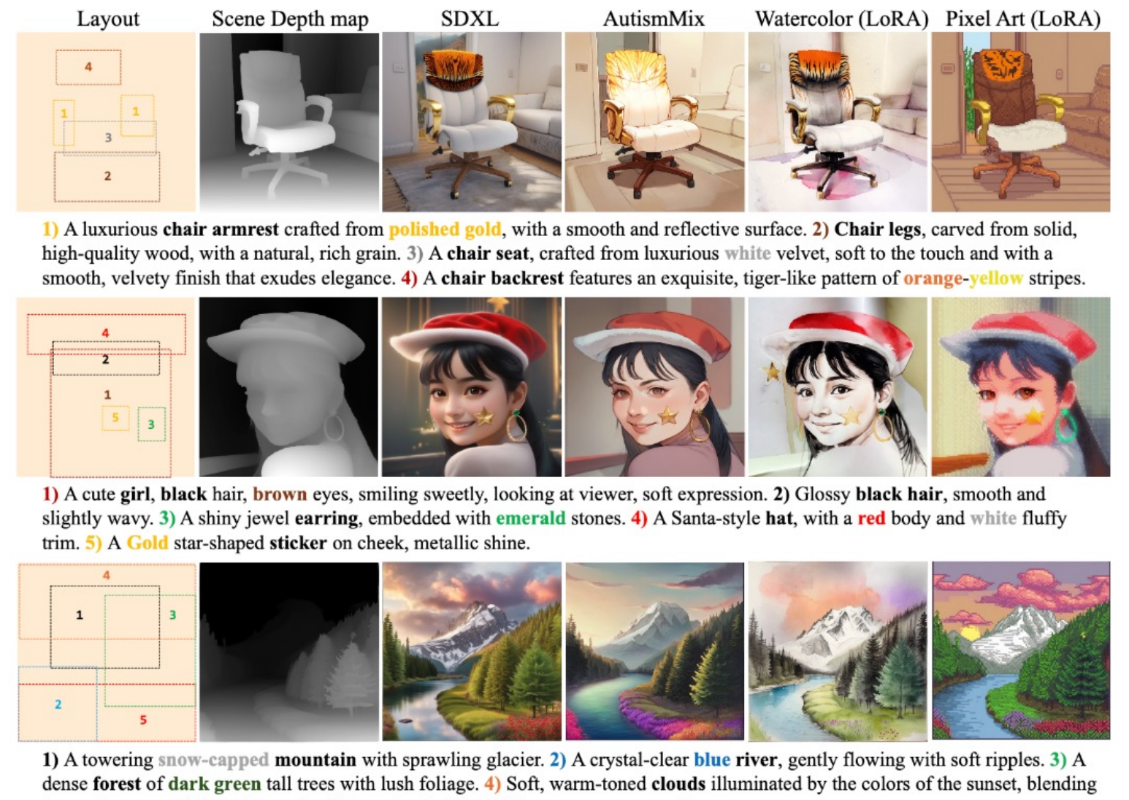

scene depth map from real images
detailed instance description
rendering result

1) Albert Einstein. 2) Mona Lisa. 3) Donald Trump. 4) McDonald's food

Control Depth Order


layout
man
man
bench
depth order: left man = right man > bench
depth order: left man > bench > right man
depth order: right man > bench > left man
depth order: left man > bench > right man
generated scene depth
rendering w/ SDXL

Complex Attribute Rendering



1) A luxurious **chair armrest** crafted from **polished gold**, with a smooth and reflective surface. 2) **Chair legs**, carved from solid, high-quality wood, with a natural, rich grain. 3) A **chair seat**, crafted from luxurious white velvet, soft to the touch and with a smooth, velvety finish that exudes elegance. 4) A **chair backrest** features an exquisite, tiger-like pattern of **orange-yellow** stripes.

1) A cute **girl**, **black** hair, **brown** eyes, smiling sweetly, looking at viewer, soft expression. 2) Glossy **black hair**, smooth and slightly wavy. 3) A shiny jewel **earring**, embedded with **emerald** stones. 4) A Santa-style **hat**, with a **red** body and white fluffy trim. 5) A **Gold** star-shaped **sticker** on cheek, metallic shine.

1) A towering snow-capped **mountain** with sprawling glacier. 2) A crystal-clear **blue river**, gently flowing with soft ripples. 3) A dense **forest** of **dark green** tall trees with lush foliage. 4) Soft, warm-toned **clouds** illuminated by the colors of the sunset, blending hues of **orange**, **pink**, and **purple**. 5) Dotted colorful **wildflowers**, some are **red**, some are **purple**.