One Diffusion to Generate Them All

Supplementary Material

Figure 7. Qualitative comparison between RayDiffusion and `OneDiffusion` on GSO dataset. `OneDiffusion` yields better prediction.

8. Additional quantitative results

8.1. Camera Pose Estimation

We evaluate our model on camera pose estimation using the Google Scanned Object dataset [13]. For this task, we use six rendered images of each synthetic object and estimate the camera poses by denoising the corresponding ray embeddings. Following RayDiffusion [72], we apply least squares optimization to estimate the camera centers and rotations. The camera center accuracy, measured with a threshold of 0.3, is reported in Table 5.

Figure 7 provides a qualitative comparison between our model and RayDiffusion. RayDiffusion consistently predicts camera poses in the upper hemisphere due to the bias in its training data, such as CO3D, which predominantly features upper-hemisphere views. In contrast, thanks to the diversity of our large-scale training dataset, `OneDiffusion` achieves higher accuracy and avoids this limitation.

Method	Accuracy
RayDiffusion [72]	0.20
`OneDiffusion`	0.32

Table 5. Comparison of zero-shot camera pose estimation methods on the GSO dataset, evaluated by Camera Center Accuracy at a threshold of 0.3.

Method	**Background Preservation**		**CLIP Semantics**	
	PSNR ↑	**LPIPS** ↓	**Whole** ↑	**Edited** ↑
Prompt-to-Prompt	17.87	208.80	25.01	22.44
Null-text Inversion	27.03	60.67	24.75	21.86
PnPInversion	27.22	54.55	25.02	22.10
Pix2pix-zero	20.44	172.22	22.80	20.54
MasaCtrl	22.17	106.62	23.96	21.16
InstructPix2Pix	16.69	271.33	23.49	22.20
MGIE	21.20	142.25	24.28	21.79
EditAR	21.32	117.15	24.87	21.87
OneDiffusion	27.49	56.67	25.84	22.34

Table 6. Evaluating image editing on PIE-Bench.

Methods	**DINO** ↑	**CLIP-I** ↑	**CLIP-T** ↑
Real Images (Oracle)	0.774	0.885	–
Fine-Tuning			
Textual Inversion	0.569	0.780	0.255
DreamBooth	0.668	0.803	0.305
BLIP-Diffusion	0.670	0.805	0.302
Tuning Free (Zero-shot)			
Re-Imagen	0.600	0.740	0.270
BLIP-Diffusion	0.594	0.779	0.300
`OneDiffusion`	0.692	0.814	0.297

Table 7. Evaluating subject-driven generation on DreamBench.

8.2. Image Editing and Subject-driven generation

We evaluate the performance of `OneDiffusion` for instruction-based image editing with the PIE-Bench dataset [22] in Table 6; and for subject-driven generation using DreamBench [52] in Table 7. `OneDiffusion` achieves strong performance compared to specialized editing and generation approaches without any fine-tuning.

9. Additional qualitative results

ID Customization. We report ours results for ID Customization tasks in Figure 13 and Figure 14. It can be observed from Figure 13 that `OneDiffusion` well preserve the identity of a person with a single input and highly manipulatable. It can re-contextualize the image (as in

first prompt), change the style from realistic photo to Pixar style (second prompt) and modify the medium to watercolor painting (third prompt).

Moreover, our approach does not relying on face embedding as in previous works [21, 28, 63] making it highly versatile. As illustrated in Figure 14, our model can preserver highly details, intricate structure as armor of the person in 4^{th} row. `OneDiffusion` can also work with non-human subject as the Gundam robot in 3^{rd} row. The model performs well with other style than photorealistic input such as anime (1^{st} row), 3D figure (2^{nd} row), cartoon (5^{th} row). Our model is highly editable where we can control the style, human pose, camera angle, expression.

Multiview generation We report additional results for multiview generation in Figure 10 and Figure 11. The generation process is as follow: we set the azimuth ranges to $[-0.45, 0.6]$ and elevation ranges to $[-15, 45]$, except for the last row of Figure 11. Then we equally slice these ranges to 80 views. We first generate 3 anchor views from the input image and independently synthesize subsequent images based on input image and the nearest anchor. For each generation batch, we generate 3 novels view and condition on 2 images. We report views with index in $[0, 10, 20, \cdots, 70]$ in below figures.

`OneDiffusion` is capable of generating photorealistic results of arbitrary objects or scenes from any number of input views either realistic captured ($2^{nd}, 3^{rd}$ row of Figure 11) or synthesized images (Figure 10). Our model works best for camera trajectory covering front views of a scene.

As mentioned ealier, `OneDiffusion` can also generate consistent multiview images from pure text and without any input images. Specifically, we simply input the azimuth and elevation as input for camera poses and generate all images from Gaussian noises as in Figure 12.

Depth estimation We provide additional qualitative results of `OneDiffusion` and compare it with Marigold-LCM [23] and DepthAnything-v2 [67] in Figure 15 and 16. We can see that our model estimator is more robust than Marigold on open-world test suits and is highly correlated with output of DepthAnything model.

Human Pose estimation We report additional results for pose estimation on COCO dataset in Figure 17. It can be observed that our model can predict multiple people in an image without relying on object detector models.

Semantic Segmentation We report qualitative results of semantic segmentation on COCO dataset in Figure 18. Unlike previous models [45, 75], our semantic-to-image and vice verse does not enforce hard association between colors and the target classes. We provide the color masks and class name as additional input in caption.

Zero-shot task composition `OneDiffusion` demonstrates remarkable generalization capabilities during inference, extending beyond its training on single-condition images to handle multiple conditioning inputs. Notably, it performs zero-shot task composition, such as inpainting with a reference face or generating images based on semantic maps and human pose, as illustrated in Figure 8.

Figure 8. `OneDiffusion` is capable of performing several zero-shot task compositions. Left is inpainting with reference image and right is image generation with human pose and semantic map (butterfly, flower, stone)

10. Summary Datasets

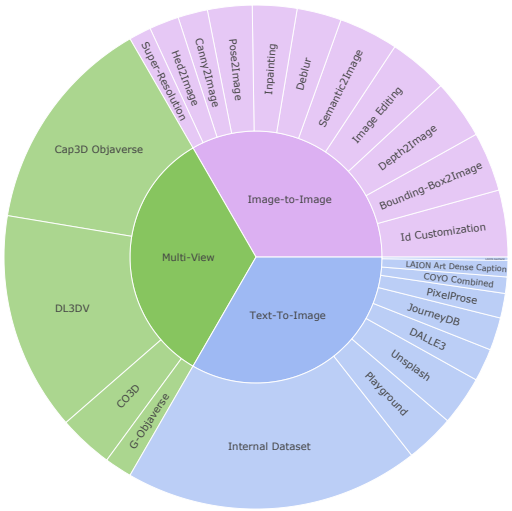

Figure 9. Distribution of training datasets for all tasks. Segments proportional to sampling rates. The inner section shows the super-category of target tasks, it can be observed that we train the model with equal budget for text-to-image, image-to-image and multi-view generation. The outer section shows datasets used for each super-category.

We train the model on multiple datasets reported in Section 4 and illustrated in Figure 9. The pie-chart segment each dataset proportional to the sampling rate of it in **third stage** of the training process. We train the model with equal budget for text-to-image, image-to-image translation (2 frames), and multiview generations ($2-6$ frames). Note that we filter and only use a subset of COYO with $11M$ images in our training. Due to the missing samples during download process, the LAION-aesthetic dataset only has $6M$ images. We recaption the LAION-aesthetic dataset with Molmo [11].

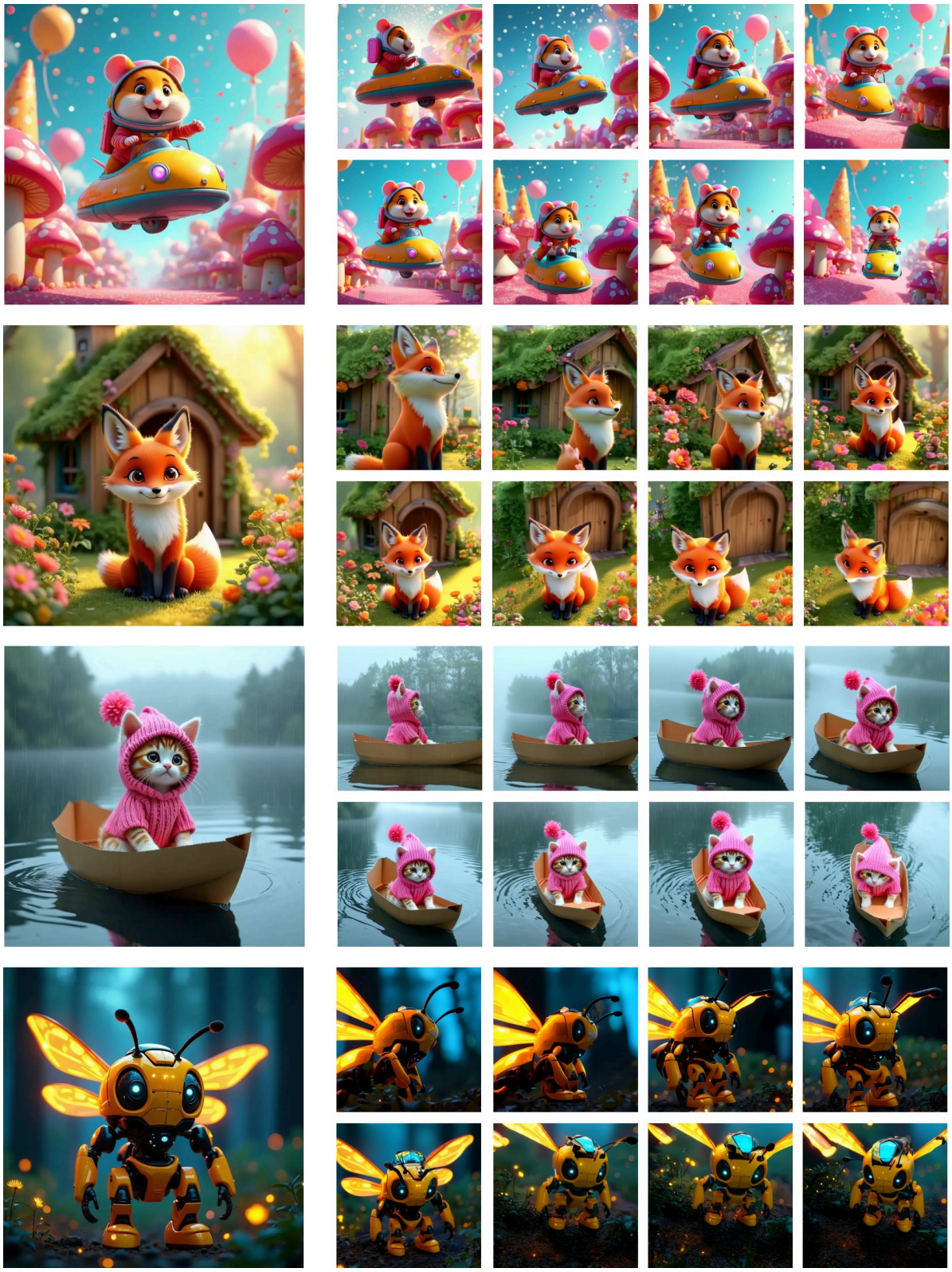

Figure 10. Qualitative results of image-to-multiview generation. The left most images are input. We equally slice the azimuth in range of $[-45, 60]$ and elevation in range of $[-15, 45]$ for all scenes.

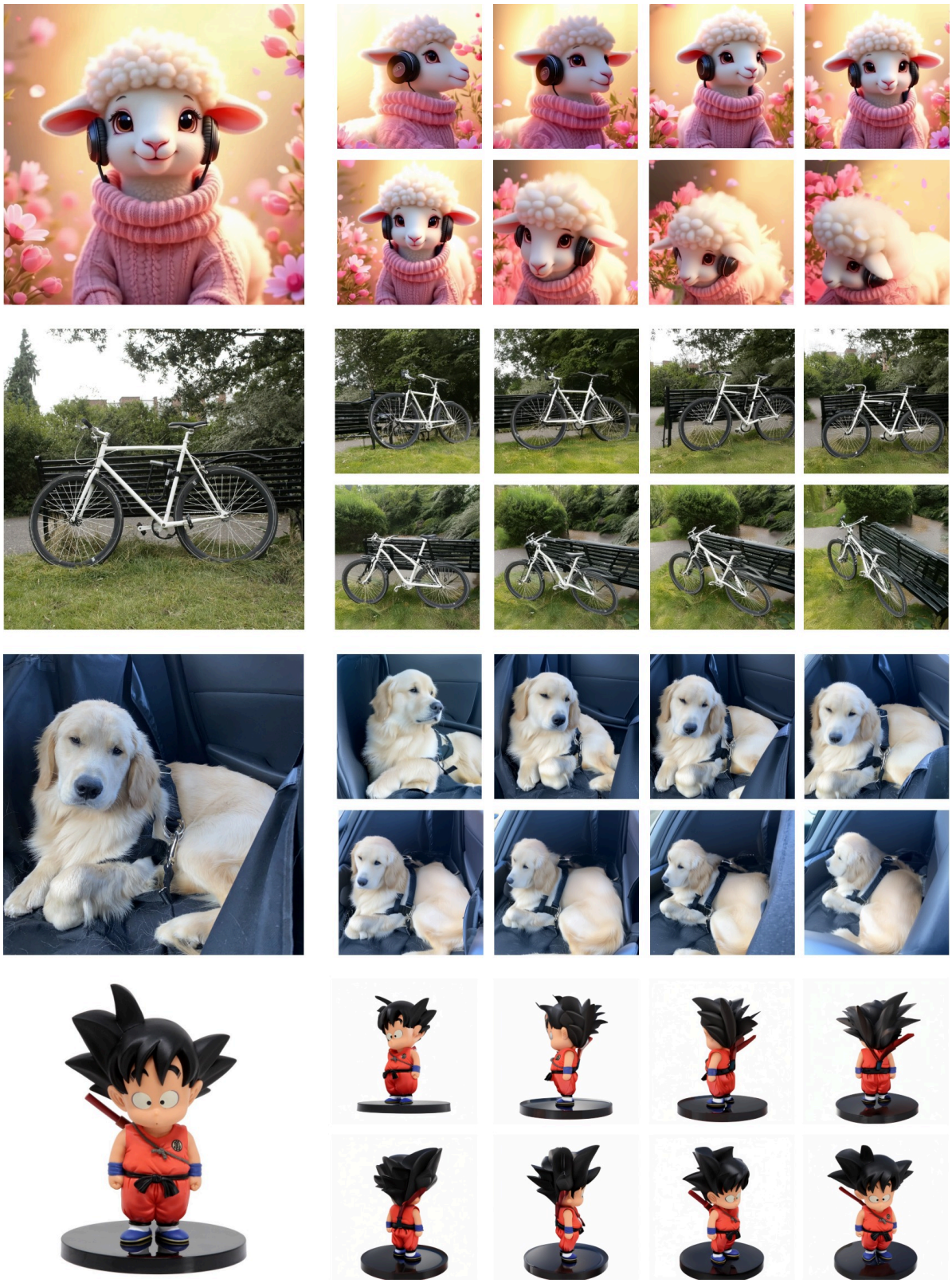

Figure 11. Qualitative results of image-to-multiview generation. We equally slice the azimuth in range of $[-45, 60]$ and elevation in range of $[-15, 45]$ for the first 3 scenes. For the last scene, the azimuth range is set to $[0; 360]$ and elevation range is set to $[-15; 15]$.

- The 3D scene features a striking black raven perched on a weathered rock in a rugged, mountainous landscape. Its glossy feathers shimmer with iridescent highlights, adding depth and realism. The background reveals a misty valley with rolling hills and a solitary stone cottage, exuding a sense of isolation and mystery. The earthy tones of the terrain, scattered with rocks and tufts of grass, contrast beautifully with the raven's dark plumage. The atmosphere feels serene yet haunting, evoking themes of solitude and nature's quiet power.

-The 3D scene portrays a haunting yet whimsical Halloween atmosphere. A ghostly figure, shrouded in glowing white fabric, kneels by a reflective puddle, clutching a carved jack-o'-lantern with a mischievous grin that radiates warm orange light. Behind the figure stands an imposing Victorian-style mansion, its dark silhouette contrasted against a full, luminous moon and bare trees reaching skyward. The dim, eerie blue lighting sets an atmospheric tone, highlighting the spectral glow of the ghost and casting faint shadows across the scene. Other jack-o'-lanterns dot the background, amplifying the festive yet spooky Halloween setting, while subtle reflections ripple in the water.

- The 3D scene showcases an adorable, highly detailed squirrel sitting at a wooden table, indulging in a plate of spaghetti. Its tiny paws grasp strands of the pasta, which is vibrantly orange, topped with fresh herbs and small cherry tomatoes. The squirrel's wide, curious eyes and delicate whiskers bring a sense of playfulness and charm. The background is softly lit, suggesting a cozy indoor setting with blurred greenery visible through a window. The overall tone is warm and whimsical, emphasizing the humorous juxtaposition of a woodland creature enjoying a human-style meal in an almost storybook-like moment.

-The 3D scene depicts a regal polar bear seated peacefully in a serene, snow-covered landscape under a sparkling night sky. The bear wears an elaborate midnight-blue cloak adorned with shimmering jewels and intricate golden embroidery, giving it a majestic and otherworldly appearance. The surrounding terrain features icy rocks and soft snow illuminated by a gentle blue glow, adding depth and mystique. Stars twinkle brightly above, creating a celestial ambiance that complements the bear's dignified presence. The combination of the bear's noble posture and the enchanting environment evokes a sense of calm, power, and magical storytelling.

- The 3D scene features a wise, luminous owl perched on an open book under a starry night sky. The owl's intricate feathers shimmer with hues of soft blue and warm orange, glowing as though illuminated by an unseen magical light. Its large, piercing orange eyes exude intelligence and focus, as if it's deeply immersed in the book beneath its talons. The book's pages are slightly ruffled, adding realism, while the background depicts distant mountains bathed in moonlight. Twinkling stars and glowing orbs punctuate the serene night, blending mysticism and wisdom in a captivating and enchanting atmosphere.

- The 3D scene features a humorous and surreal depiction of a person wearing a shark costume, complete with a cheerful, cartoony face and sharp teeth. The character is standing in a foggy, dystopian urban setting with damaged buildings and rubble lining the street. The shark is holding a smartphone, seemingly taking a selfie, while wearing a brown leather backpack and rope harness, adding an adventurous touch. The blend of playful absurdity and the somber, desolate environment creates a striking and amusing contrast.

- 3d figure of chibi mario, super mario, 3D character model resembling a cartoon plumber with a red cap marked by an "M," blue overalls, white gloves, and brown shoes. The character is in a neutral T-pose, emphasizing its proportions and vibrant colors, with smooth textures and clear details typical of digital rendering.

- An astronaut wearing a full space suit, complete with a helmet and backpack, riding a galloping horse. The astronaut is holding the reins, and the horse is depicted in motion with detailed musculature and dynamic posture. The scene combines futuristic elements with classic equestrian imagery. The background is solid white.

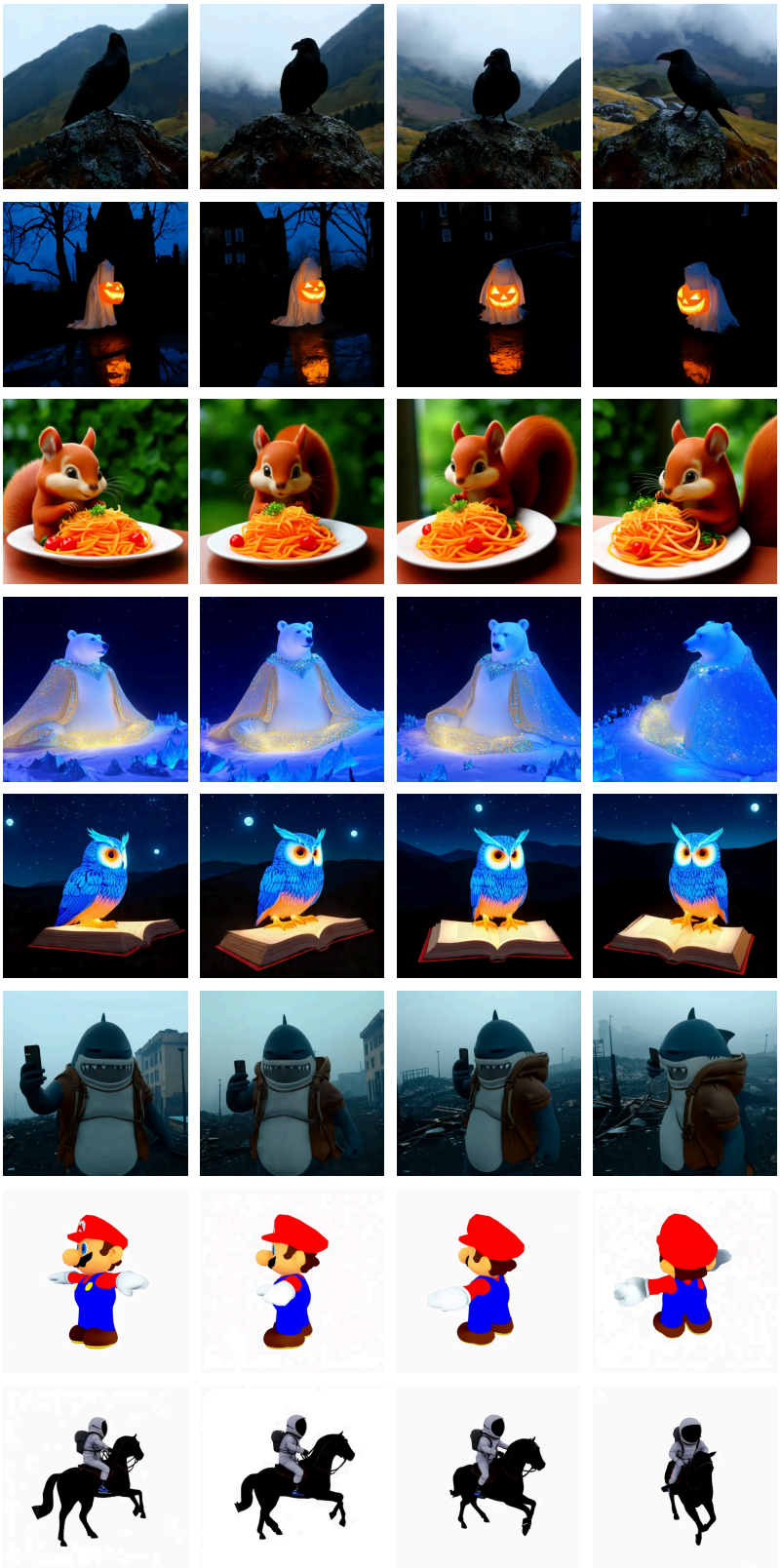

Figure 12. Qualitative results of text-to-multiview generation. The azimuth and elevation of left to right columns are $[0, 30, 60, 90]$ and $[0, 10, 20, 30]$, respectively. We use following prefix for all prompts to improve the quality and realism of generated images: "*photorealistic, masterpiece, highly detail, score_9, score_8_up*".

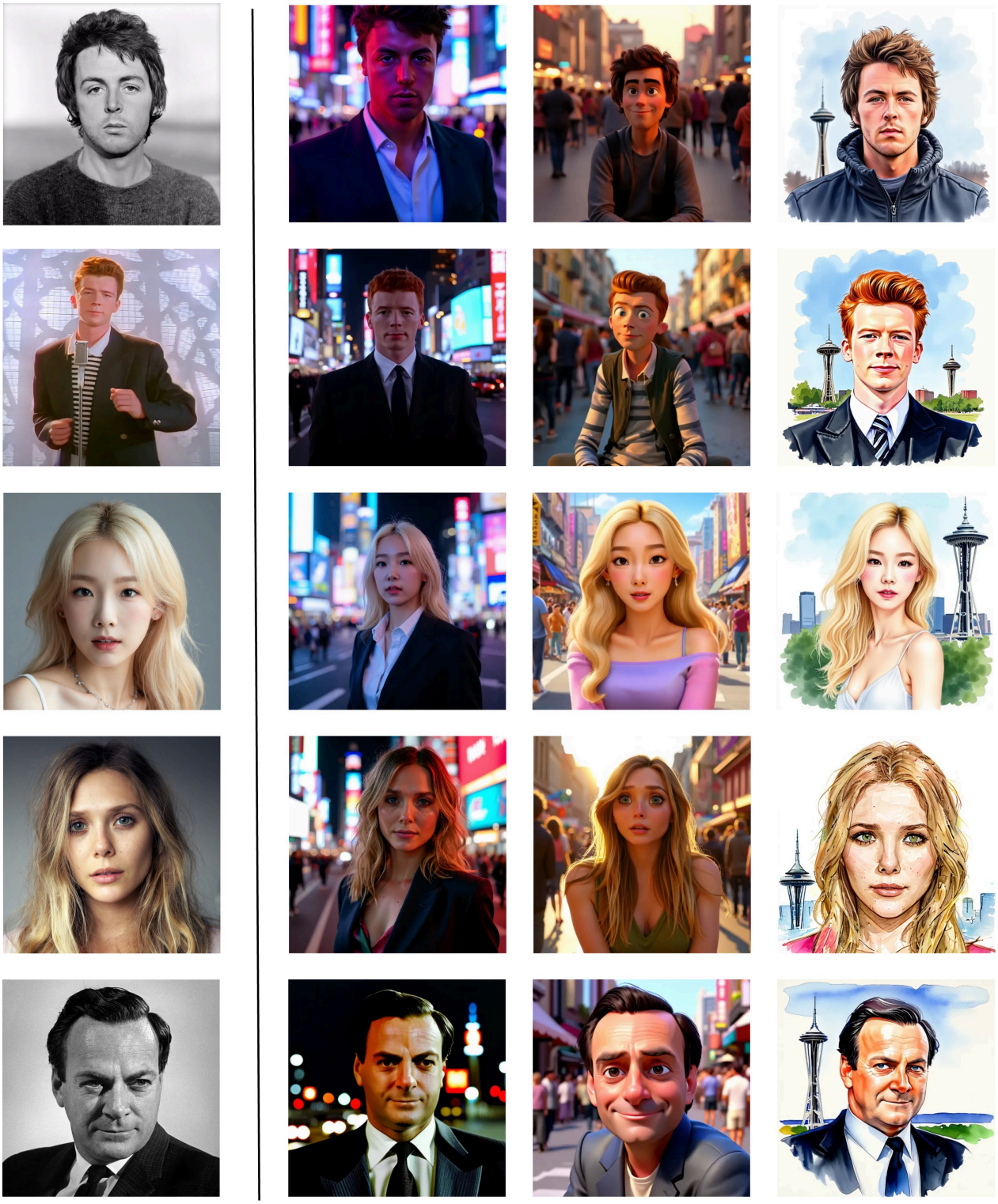

Figure 13. Qualitative results of `OneDiffusion` for (single reference) ID Customization task with photo of human faces. The left most images are input, target prompts for left to right columns are: 1) "*Photo of a man/woman wearing suit at Shibuya at night. He/She is looking at the camera*", 2) "*pixarstyle, cartoon, a person in pixar style sitting on a crowded street*", 3) "*watercolor drawing of a man/woman with Space Needle in background*"

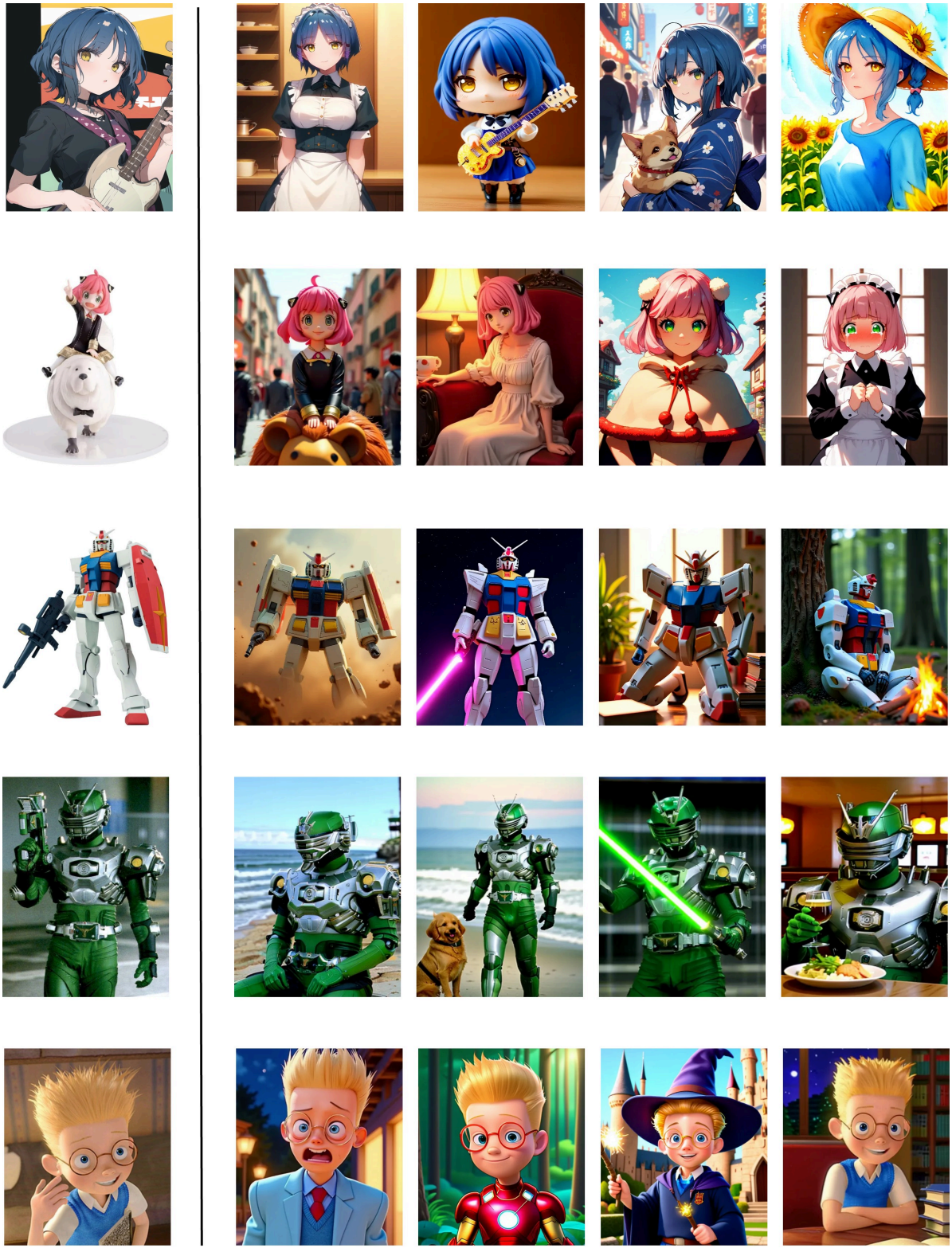

Figure 14. Qualitative results of `OneDiffusion` for (single reference) ID Customization task with photo of of non-human subjects or cartoon style input. `OneDiffusion` is highly versatile and can produce good results for all kind of input and not limited to photorealistic human images. Since we rely on attention, the model can attend to the condition view and preserve intricate details and is not limited by any bottleneck *e.g.* latent representation.

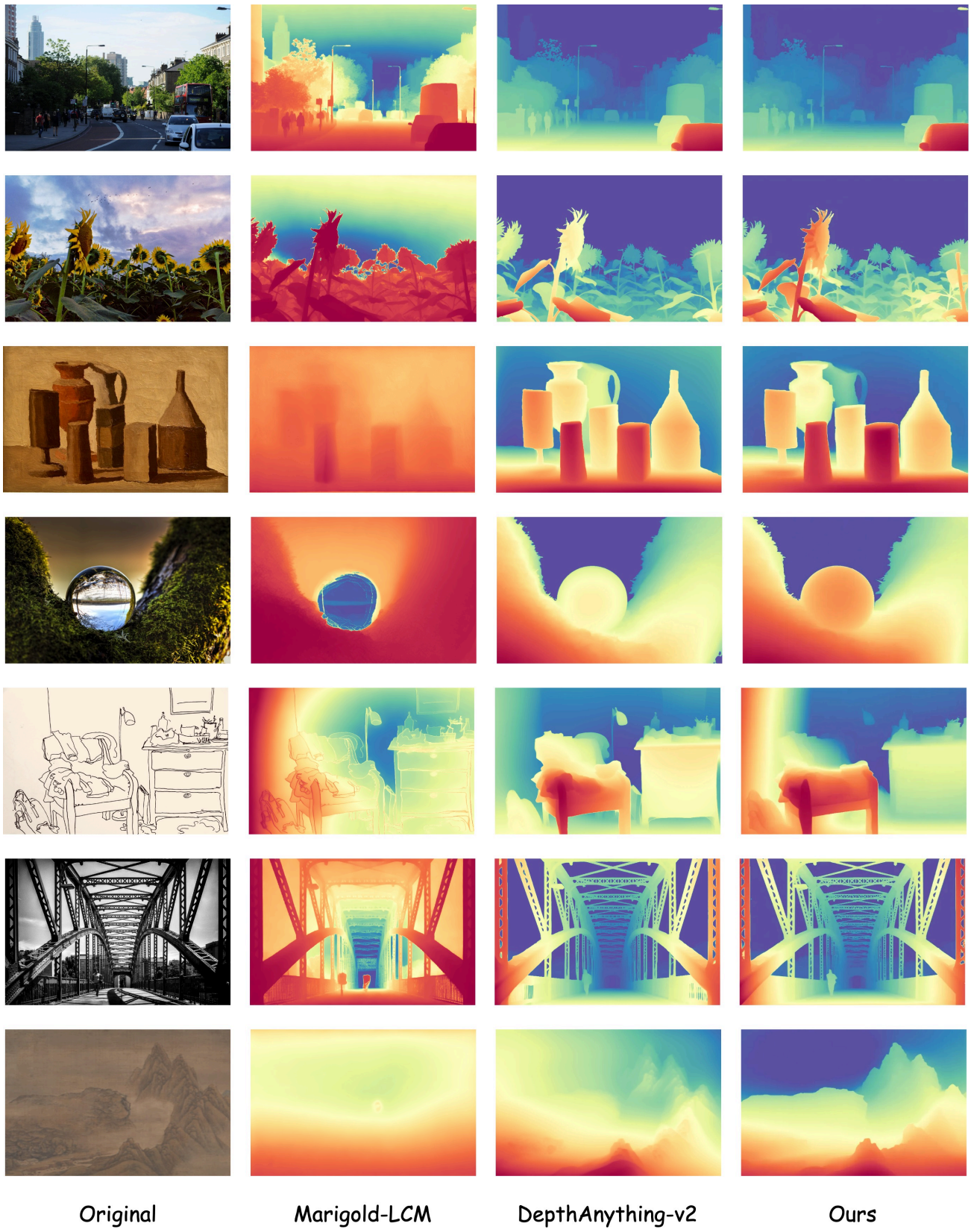

Figure 15. Qualitative comparison for depth estimation between `OneDiffusion`, Marigold [23] and DepthAnything-v2 [67]

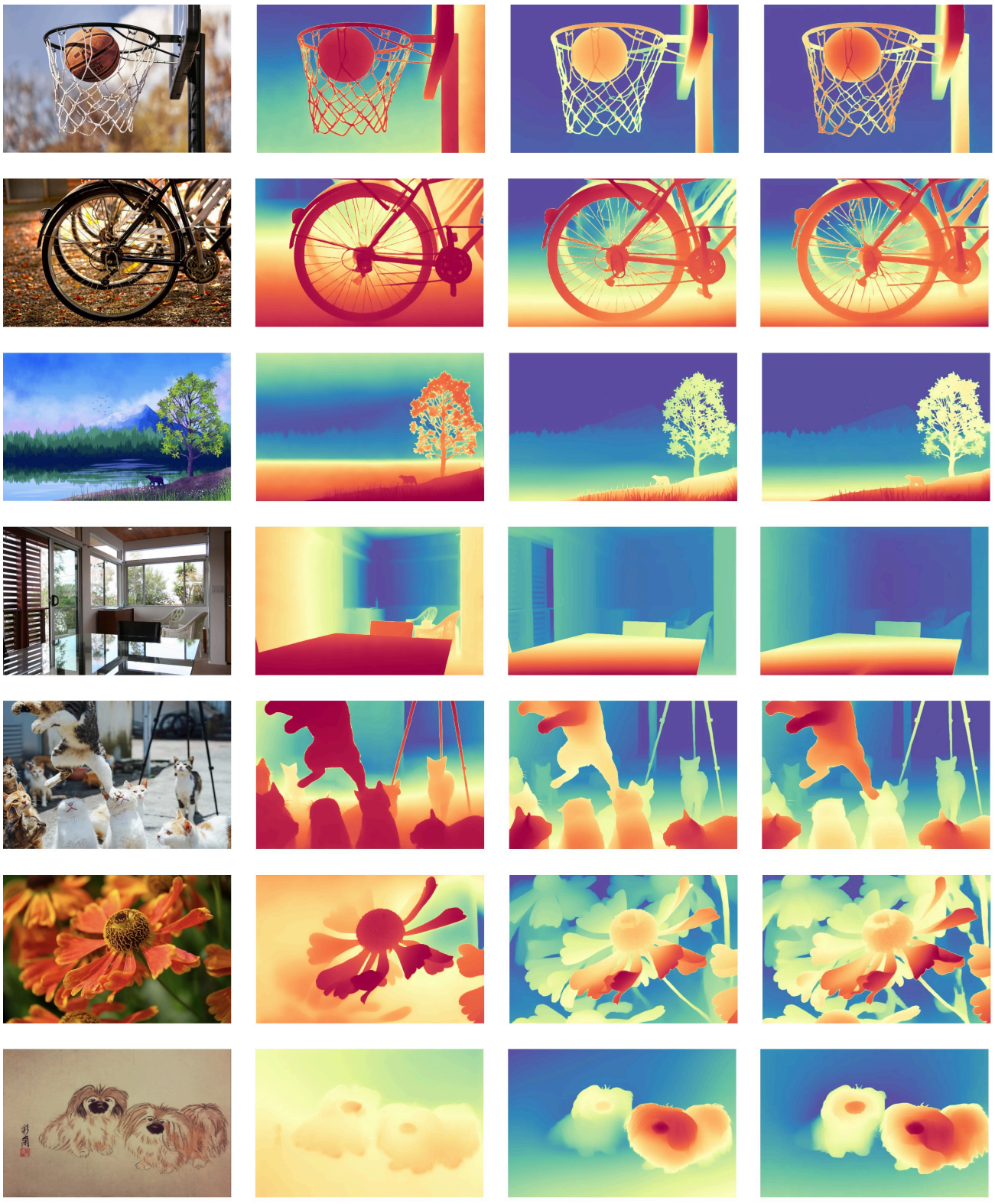

Figure 16. Qualitative comparison for depth estimation between `OneDiffusion`, Marigold [23] and DepthAnything-v2 [67]

Figure 17. Qualitative examples of human pose estimation on COCO datasets.

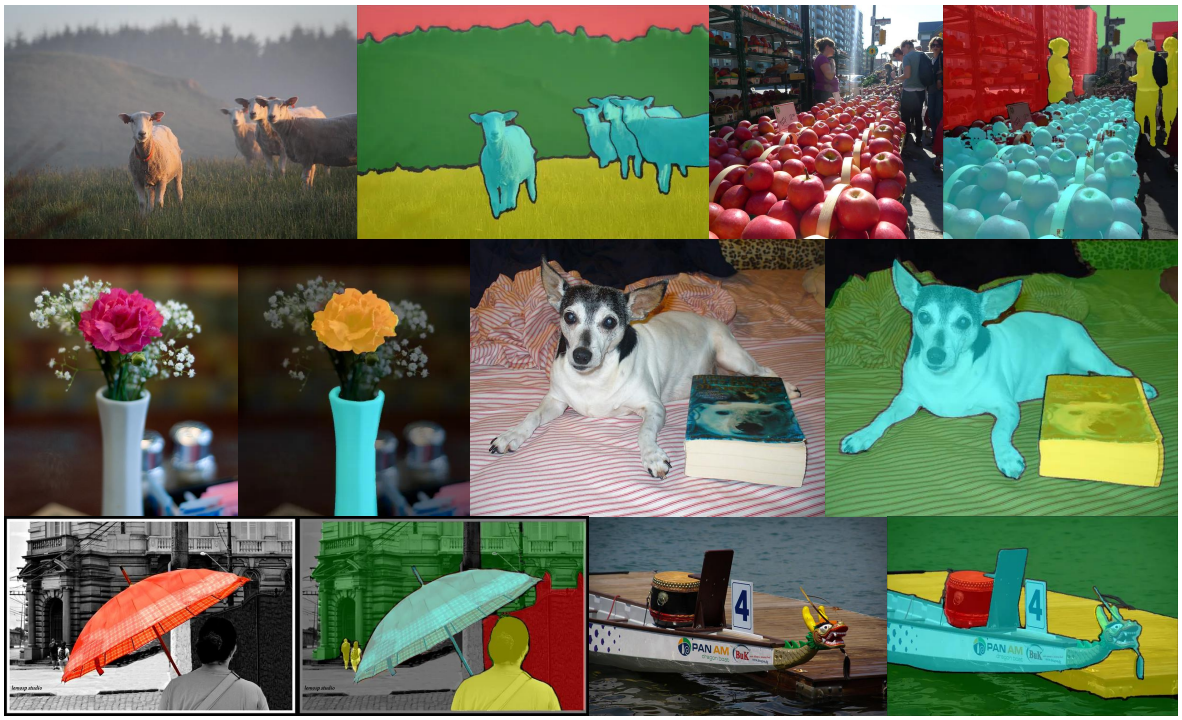

Figure 18. Qualitative examples of semantic segmentation on COCO datasets. The target class for each image (from left to right, from top to bottom) are (sheep, grass, mountain, sky), (apple, person, building), (vase, flower,), (dog, book, sheet), (umbrella, person, building, gate), (boat, dock, drum).

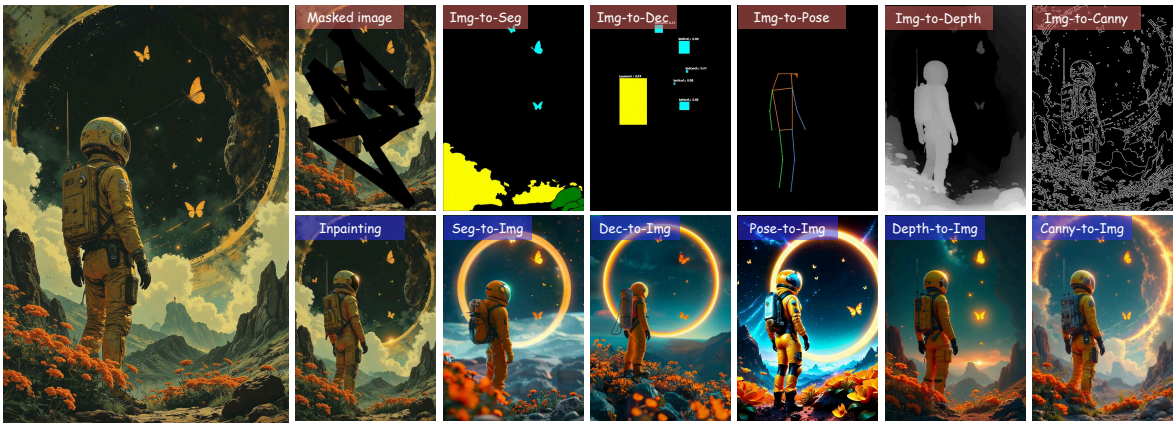

Figure 19. Illustration of our model capability to generate semantic mask, detection, human pose, depth, and canny edge from input image. For semantic segmentation, we segment the flower (highlighted in yellow) and the rock (highlighted in green). For object detection, We localize the backpack (highlighted in yellow) and butterfly (highlighted in cyan). Leveraging these conditions, we can reverse the process to recreate a variant of the input image based on the same caption.