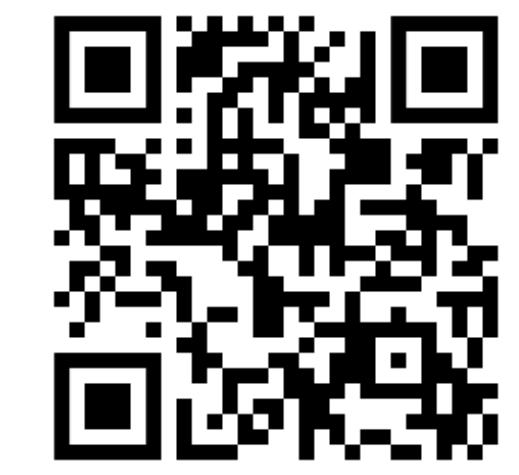
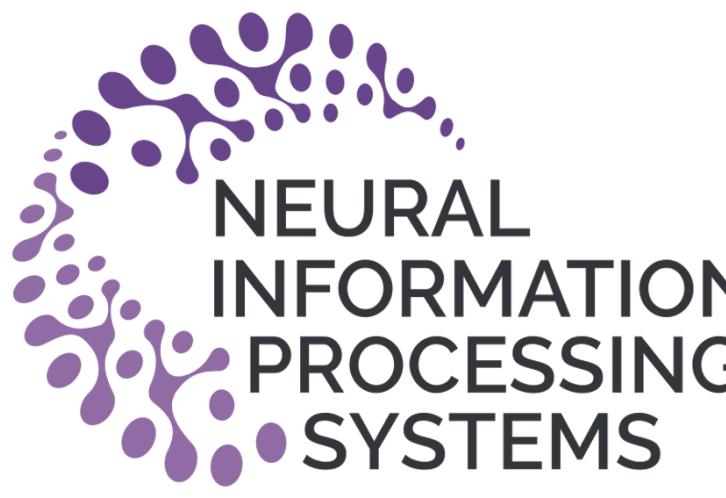




UniControl: A Unified Diffusion Model for Controllable Visual Generation In the Wild

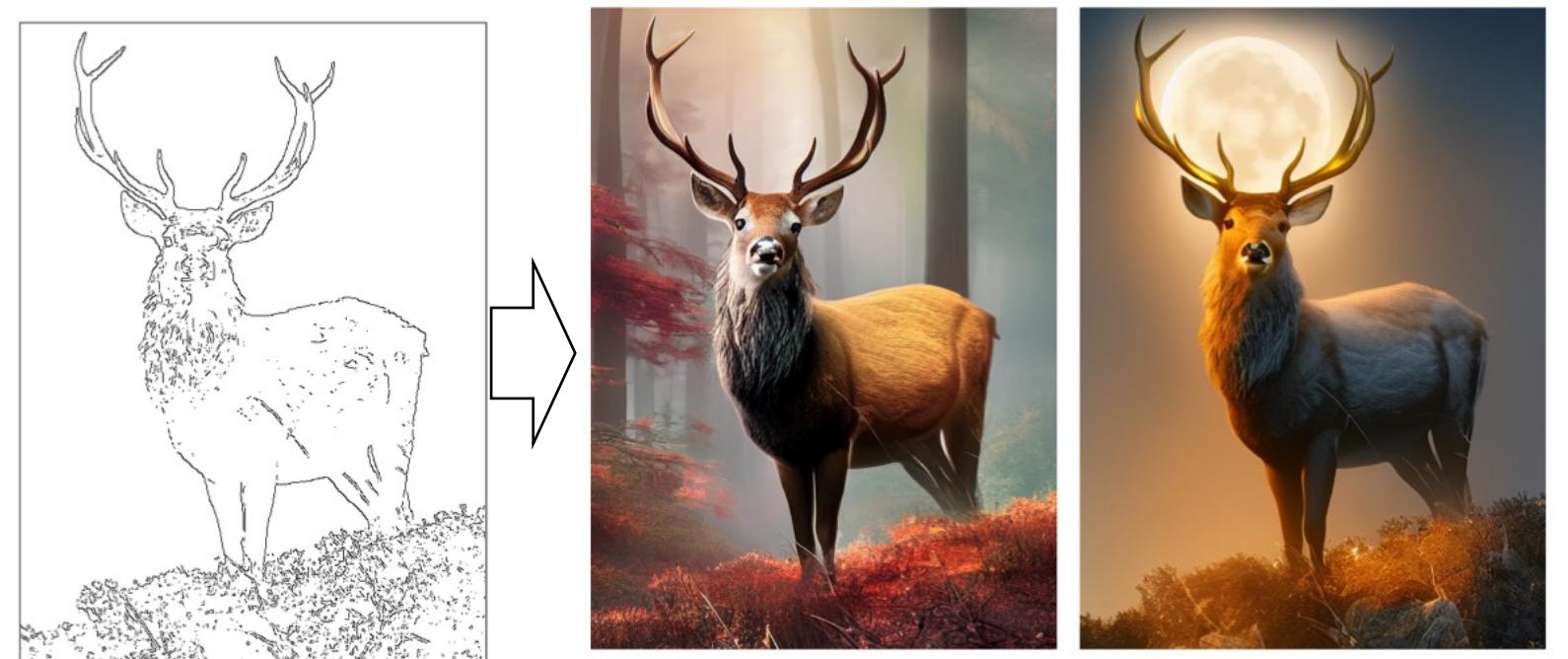
Can Qin^{1,2}, Shu Zhang¹, Ning Yu¹, Yihao Feng¹, Xinyi Yang¹, Yingbo Zhou¹, Huan Wang¹, Juan Carlos Niebles¹, Caiming Xiong¹, Silvio Savarese¹, Stefano Ermon³, Yun Fu², and Ran Xu¹

¹Salesforce Research ²Northeastern University ³Stanford University



Background

- Controllable text-to-image synthesis, generating photorealistic images from text prompts and spatial conditions, has witnessed a tremendous surge in capabilities recently.

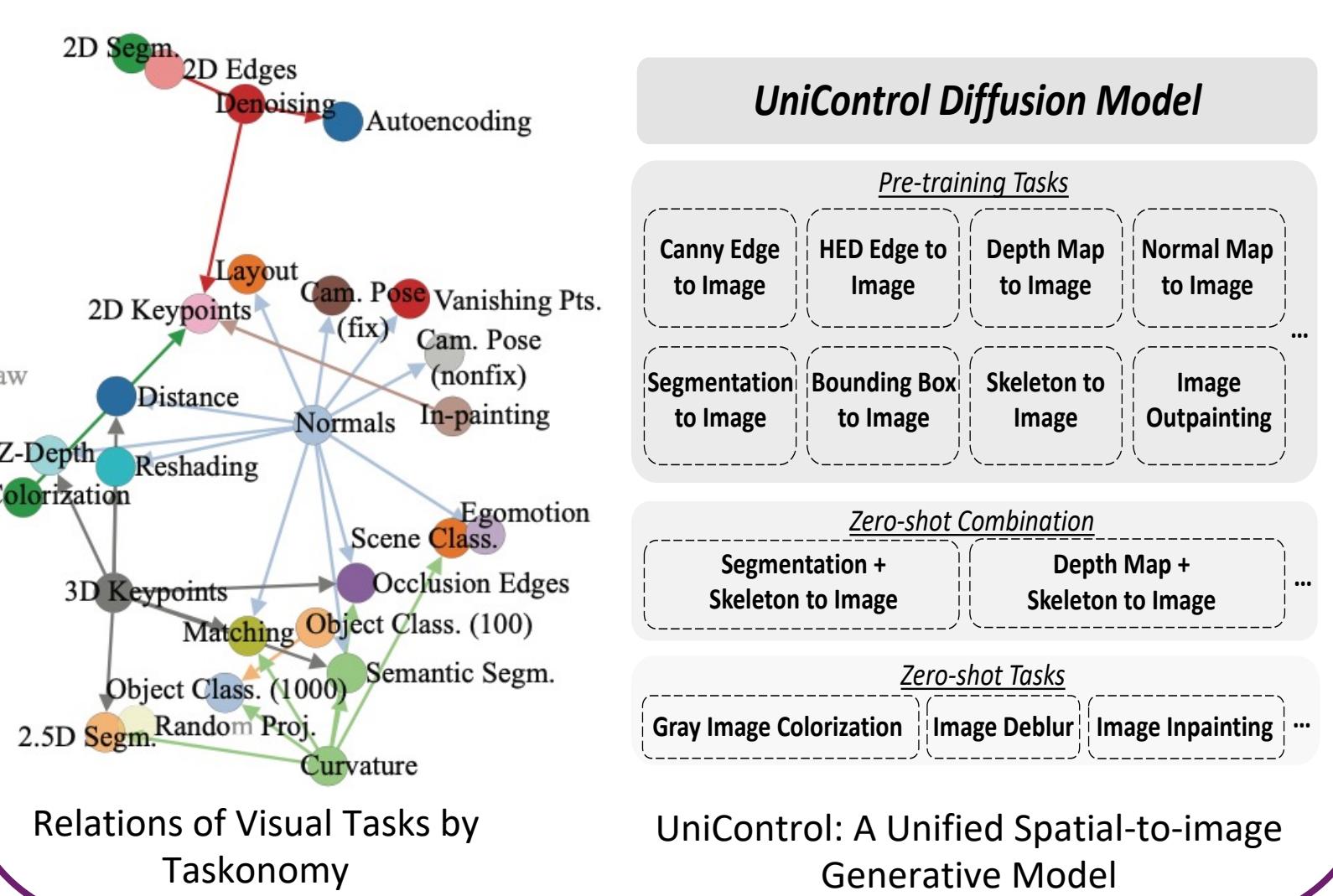


"masterpiece of fairy tale, giant deer, golden antlers"

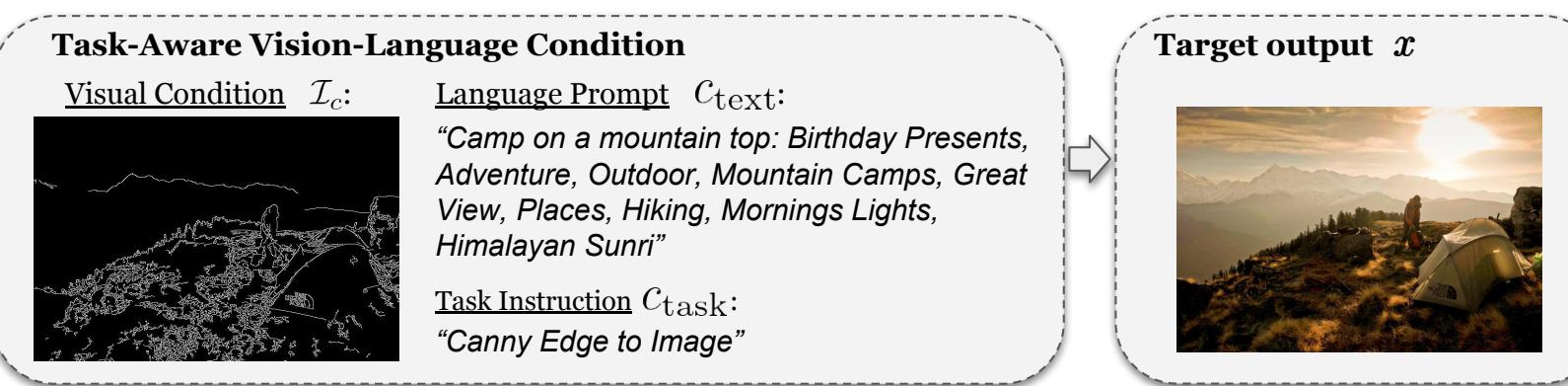
However, most of classical methods (ControlNet, T2I-adapter, Composer, etc.) are domain/task specific which need to train different models for correspondent conditions.

Motivation

- Inspired by the multi-task learning such as Taskonomy, cross-modality visual inputs share common and relational information which is implicitly beneficial for building a unified spatial-to-image generative model.



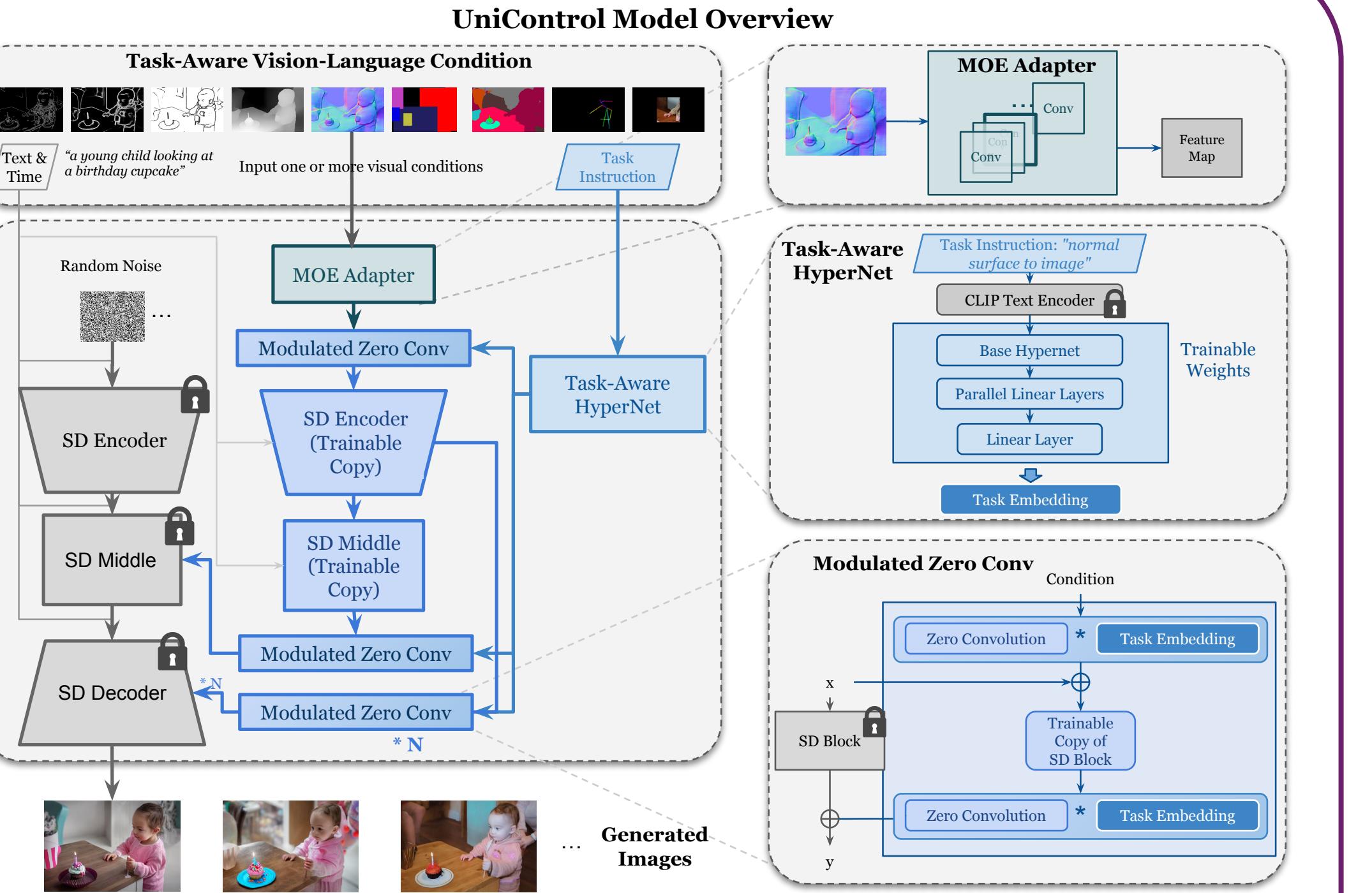
Dataset



The UniControl is trained by MultiGen-20M (opensourced) which includes nine distinct tasks:

- Edges (Canny, HED, User Sketch);
- Region-wise maps (Segmentation Maps, Bounding Boxes);
- Skeletons (Human Pose Skeletons);
- Geometric maps (Depth, Surface Normal);
- Editing (Image Outpainting).

Method



The proposed UniControl introduces three new components to enable unified multi-task controllable generation:

- Mixture-of-Experts Adapters:** Parallel convolutional modules, one per task, that adapt to each condition's visual features.
- Task-Aware HyperNetwork:** Dynamically modulates the convolution kernels of a base model given embeddings of task instructions.
- Modulated Zero-conv:** The weights of zero-conv layers would be modulated by the task embedding by Hypernet to adapt to different tasks/conditions.



The UniControl also shows promising zero-shot new-task generalization capacities including Hybrid Tasks Generalization and New Task Generalization. It achieves the later one by applying multiple MoE adapters with weights ensemble according to relations between the new and pre-training tasks.

Table 1: Architecture and Model Size (#Params): UniControl vs. Multi-ControlNet

	Stable Diffusion	ControlNet	MoE-Adapter	TaskHyperNet	Total
UniControl	1065.7M	361M	0.06M	12.7M	1.44B
Multi-ControlNet	1065.7M	$361M \times 9$	-	-	4.32B

Compared with our direct baseline - Multi-Controlnet, UniControl significantly compresses the model size by ~3X overall and achieves comparable and even better performance on each task. It would be beneficial for:

- Saving Storage:** There is only one checkpoint to save for UniControl whereas ControlNet has nine checkpoints instead.
- Efficient Computing for Multi-condition:** The users would not need to load multiple models into memory when dealing with multiple spatial conditions for content generation.

Pre-training Tasks

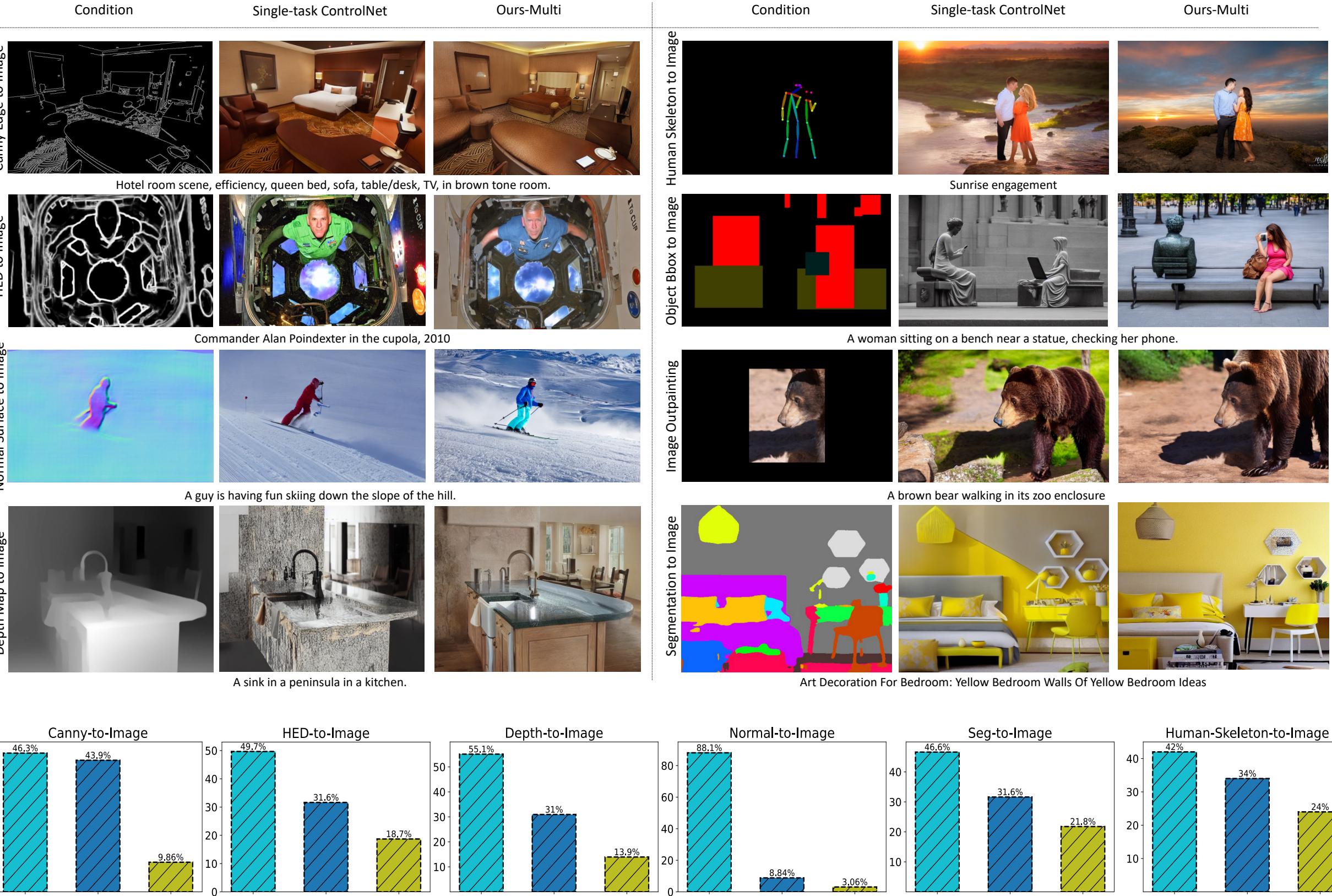


Table 2: Image Perceptual Distance

	Canny ↓	HED ↓	Normal ↓	Depth ↓	Pose ↓	Segmentation ↓
UniControl	0.546	0.466	0.623	0.654	0.741	0.693
ControlNet	0.577	0.582	0.778	0.700	0.747	0.693

Zero-shot Tasks

