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

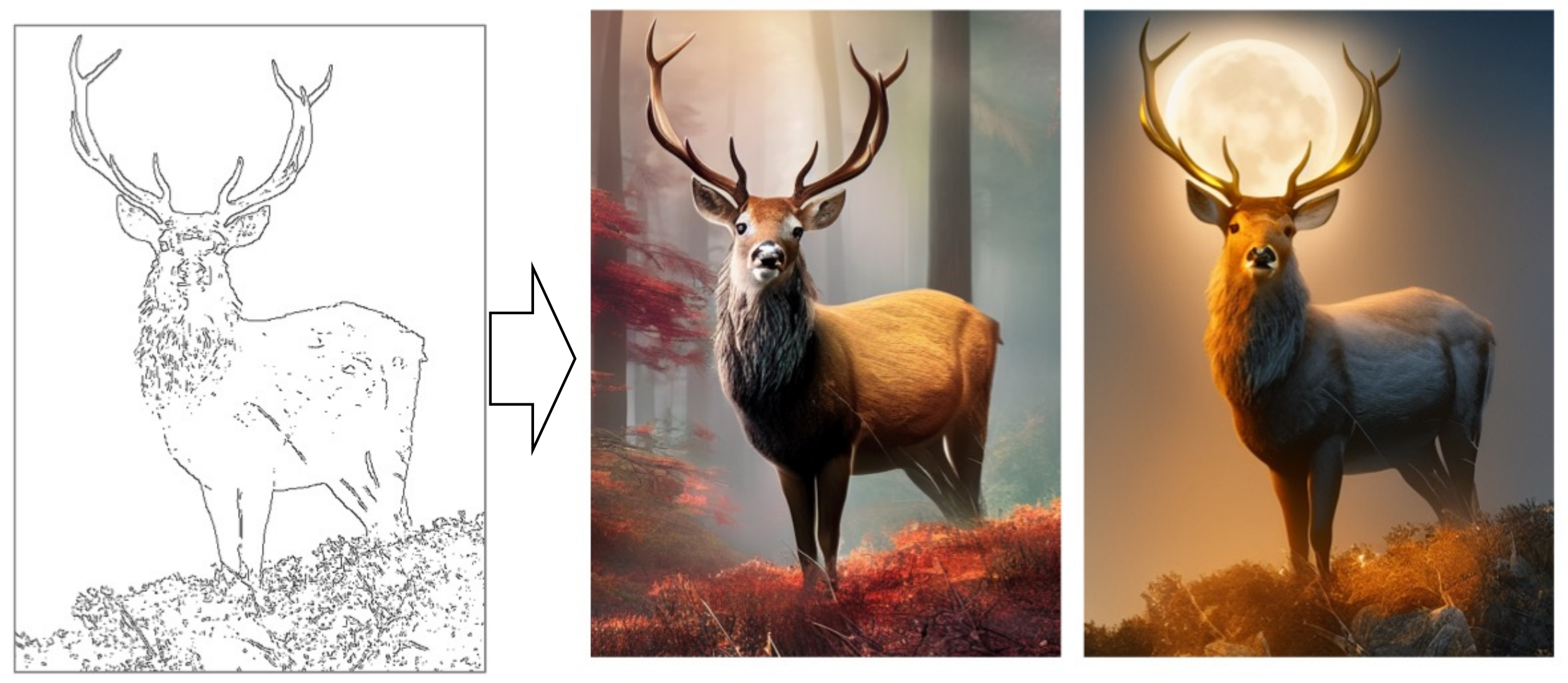
Can Qin[1,2], Shu Zhang[1], Ning Yu[1], Yihao Feng[1], Xinyi Yang[1], Yingbo Zhou[1], Huan Wang[1], Juan Carlos Niebles[1], Caiming Xiong[1], Silvio Savarese[1], Stefano Ermon[3], Yun Fu[2], Ran Xu[1]

[1]Salesforce Research [2]Northeastern University [3]Stanford University

Code and Data

Background

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

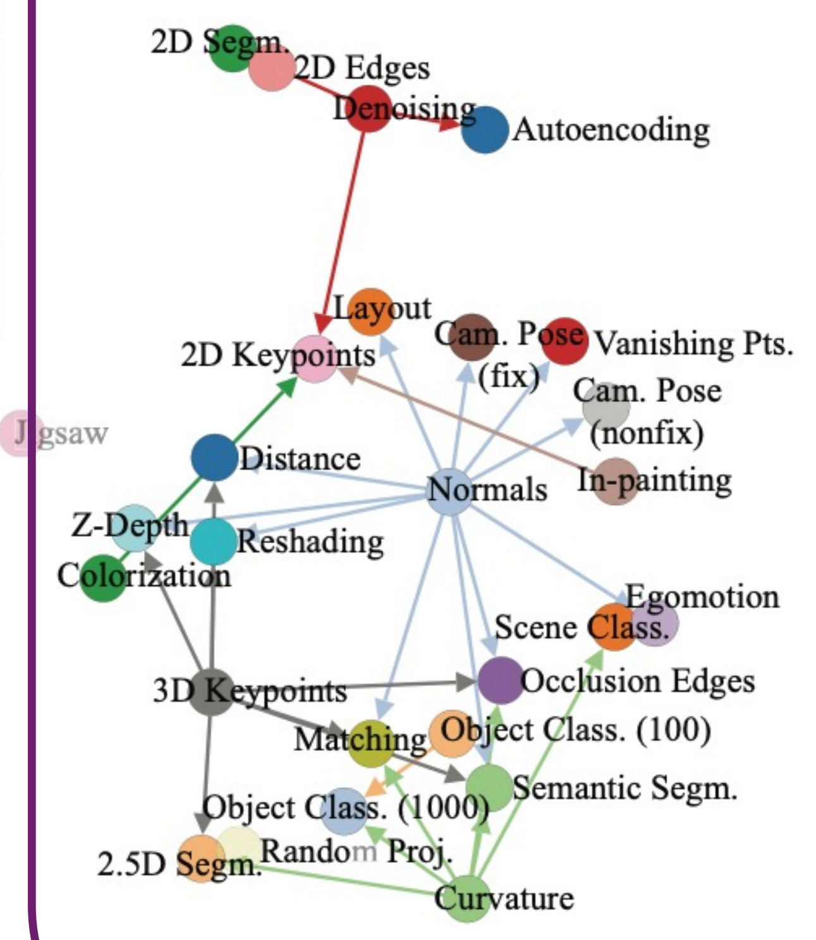


"masterpiece of fairy tale, giant deer, golden antlers"

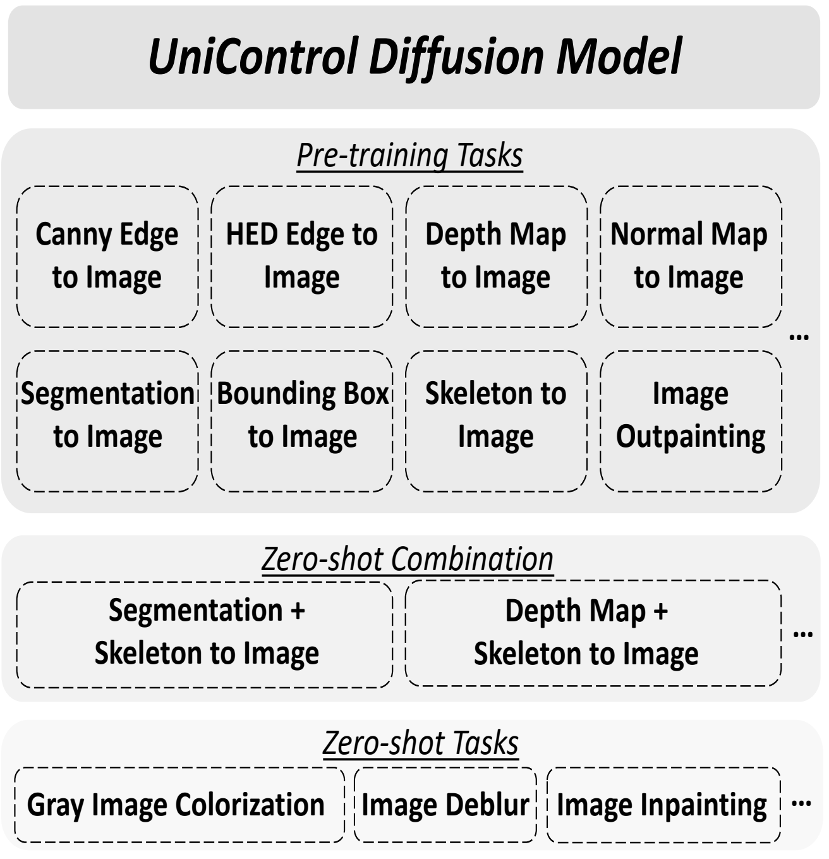
However, most of classical methods (ControlNet, T2I-adapter, Composer, etc.) are domain/task specific which need to train different models for correspondent conditions.

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**.



Relations of Visual Tasks by Taskonomy



UniControl Diffusion Model

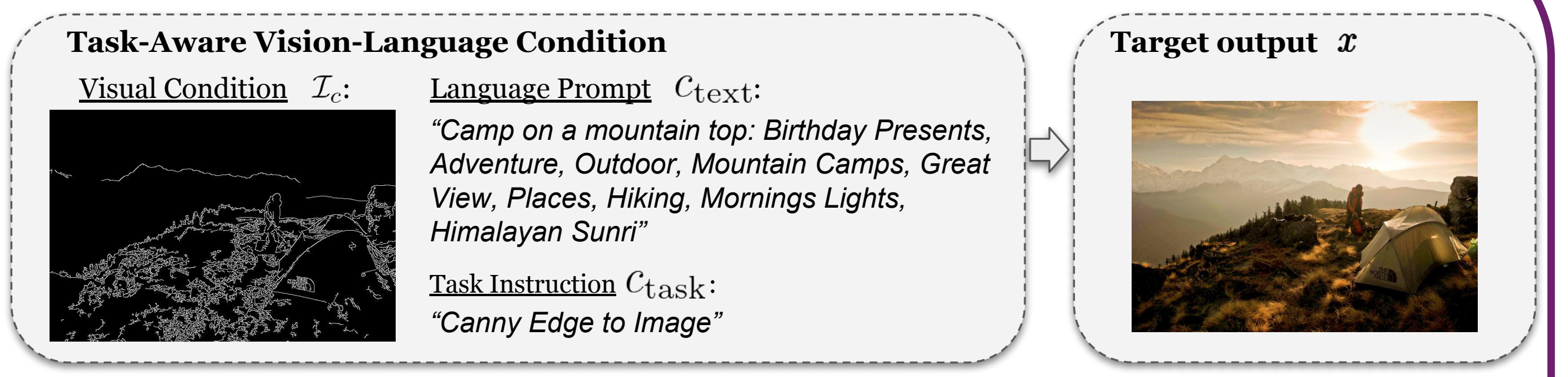
Pre-training Tasks: Canny Edge to Image; HED Edge to Image; Depth Map to Image; Normal Map to Image; Segmentation to Image; Bounding Box to Image; Skeleton to Image; Image Outpainting; …

Zero-shot Combination: Segmentation + Skeleton to Image; Depth Map + Skeleton to Image; …

Zero-shot Tasks: Gray Image Colorization; Image Deblur; Image Inpainting; …

UniControl: A Unified Spatial-to-image Generative Model

Dataset



Task-Aware Vision-Language Condition

Visual Condition $\mathcal{I}_c$:

Language Prompt c_{text}: *"Camp on a mountain top: Birthday Presents, Adventure, Outdoor, Mountain Camps, Great View, Places, Hiking, Mornings Lights, Himalayan Sunri"*

Task Instruction c_{task}: *"Canny Edge to Image"*

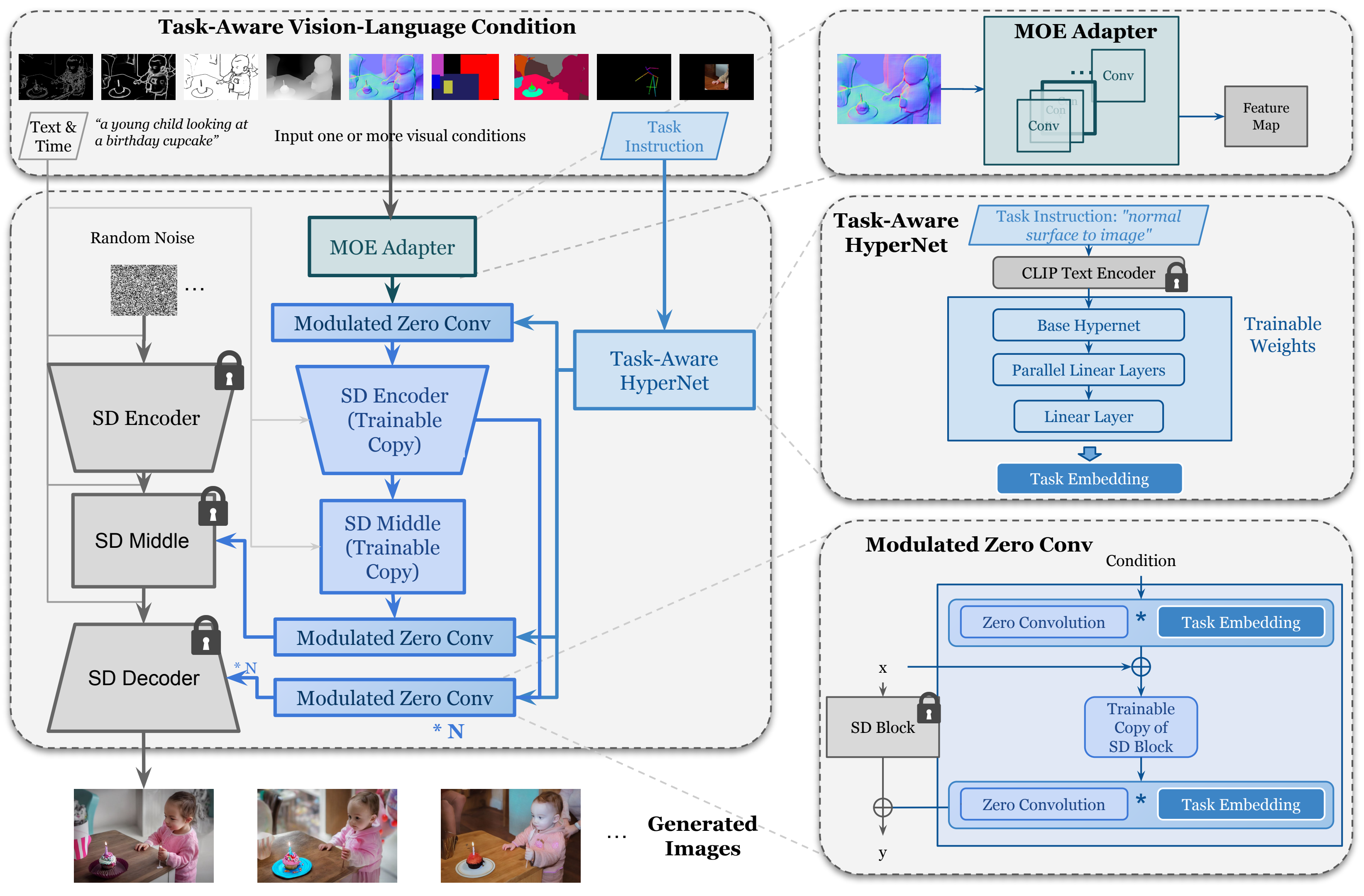
Target output x

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).

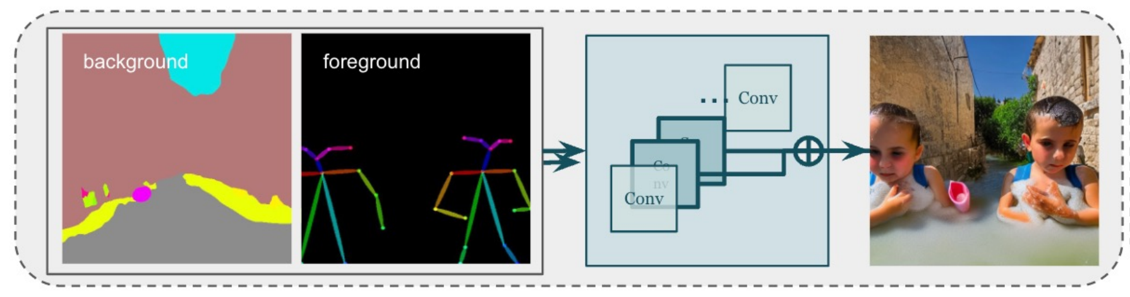
Method

UniControl Model Overview

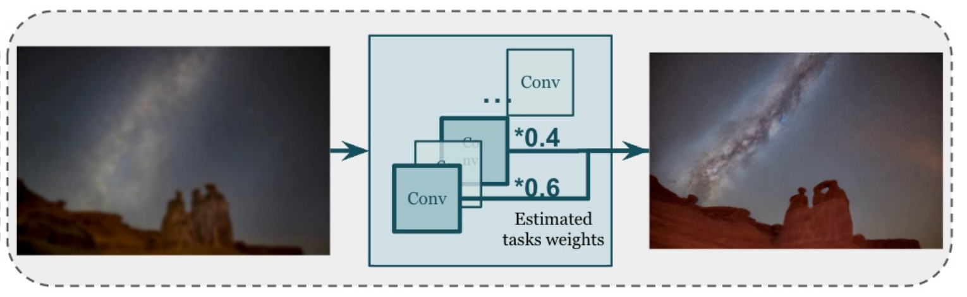


The proposed UniControl introduces three new components to enable unified multi-task controllable generation:

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



MOE Hybrid Tasks Generalization



MOE Zero-Shot New Task Generalization

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.

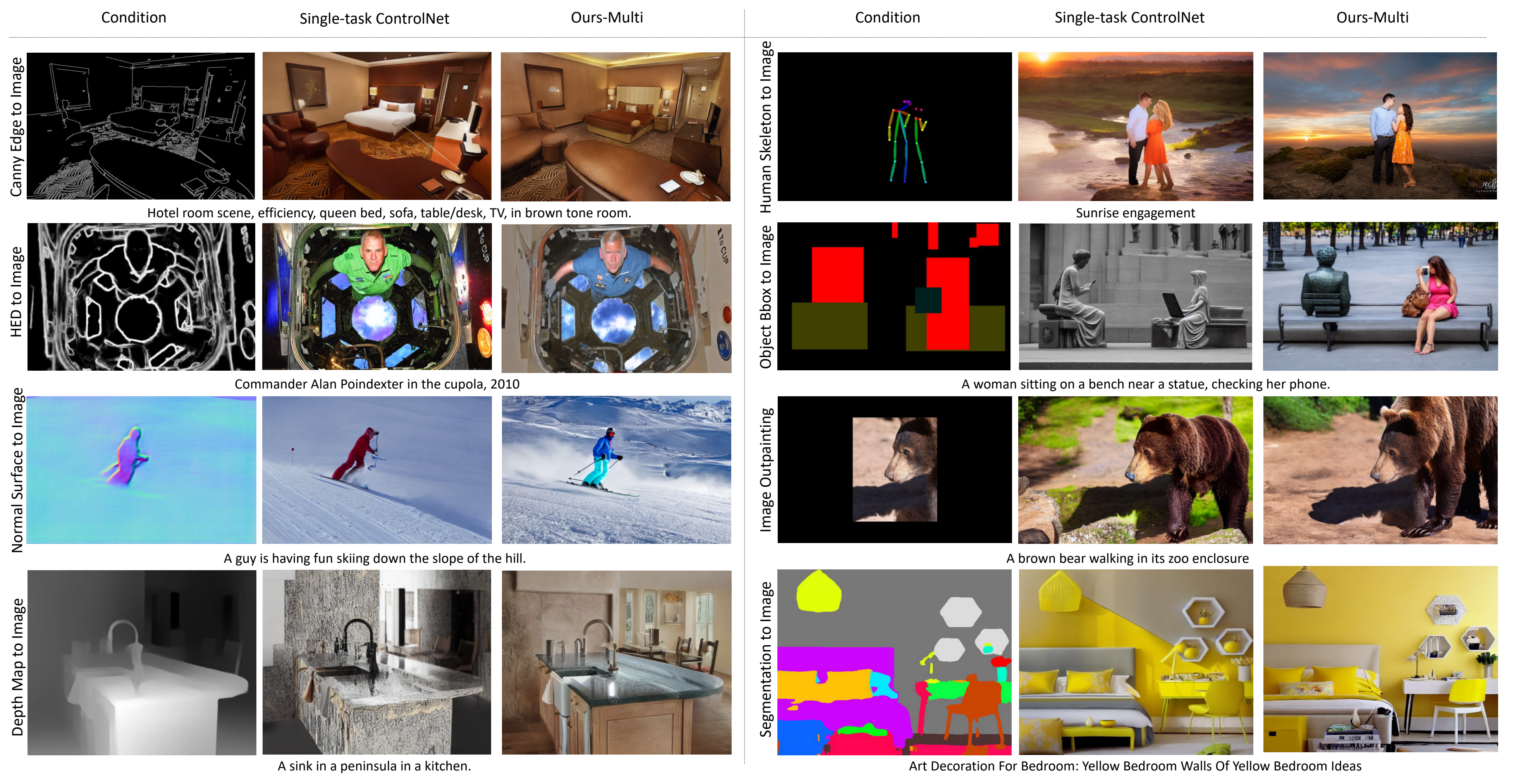
Table 1: Architecture and Model Size (#Params): UniControl *vs.* Multi-ControlNet

	Stable Diffusion	ControlNet	MoE-Adapter	TaskHyperNet	Total
UniControl	1065.7M	361M	0.06M	12.7M	**1.44B**
Multi-ControlNet	1065.7M	361M × 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:

1. Saving Storage: There is only one checkpoint to save for UniControl whereas ControlNet has nine checkpoints instead.
2. 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



Hotel room scene, efficiency, queen bed, sofa, table/desk, TV, in brown tone room.

Sunrise engagement

Commander Alan Poindexter in the cupola, 2010

A woman sitting on a bench near a statue, checking her phone.

A guy is having fun skiing down the slope of the hill.

A brown bear walking in its zoo enclosure

A sink in a peninsula in a kitchen.

Art Decoration For Bedroom: Yellow Bedroom Walls Of Yellow Bedroom Ideas

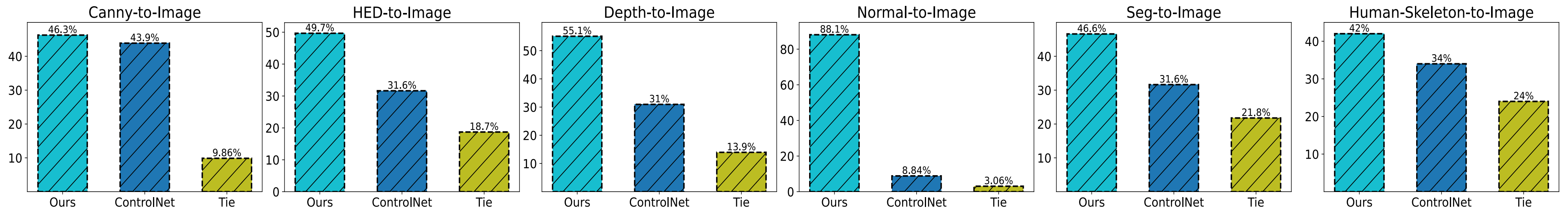
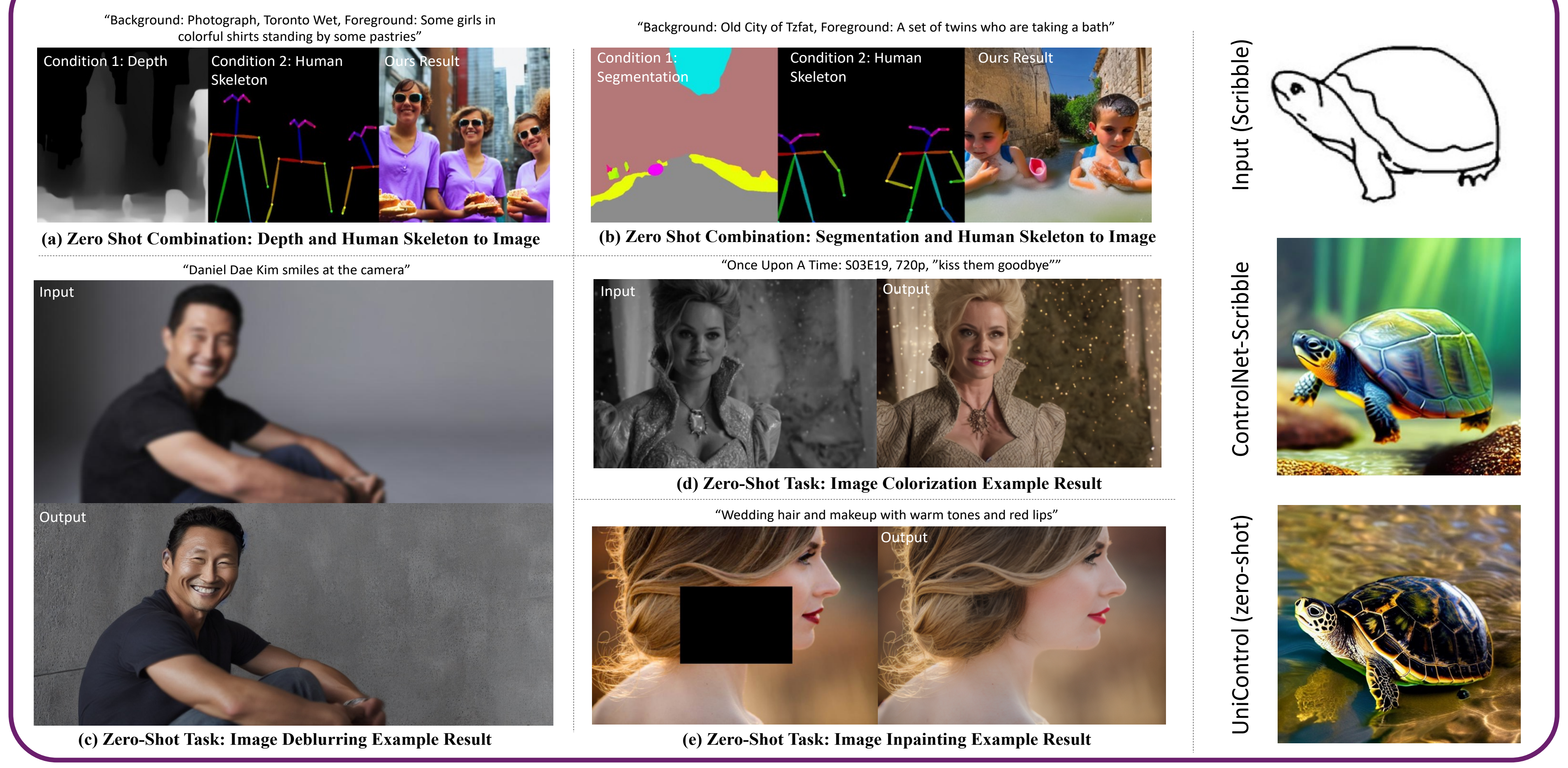


Table 2: Image Perceptual Distance

	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



"Background: Photograph, Toronto Wet, Foreground: Some girls in colorful shirts standing by some pastries"

(a) Zero Shot Combination: Depth and Human Skeleton to Image

"Background: Old City of Tzfat, Foreground: A set of twins who are taking a bath"

(b) Zero Shot Combination: Segmentation and Human Skeleton to Image

"Daniel Dae Kim smiles at the camera"

(c) Zero-Shot Task: Image Deblurring Example Result

"Once Upon A Time: S03E19, 720p, "kiss them goodbye""

(d) Zero-Shot Task: Image Colorization Example Result

"Wedding hair and makeup with warm tones and red lips"

(e) Zero-Shot Task: Image Inpainting Example Result