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

Can Qin1,2, Shu Zhang1, Ning Yu1, Yihao Feng1, Xinyi Yang1, Yingbo Zhou1, Huan Wang1, Juan Carlos Niebles1, Caiming Xiong1, Silvio Savarese1, Stefano Ermon1, Yun Fu1, 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.

However, most of classical methods (ControlNet, T2I, UniControl) 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 unified spatial-to-image generation models.

**Method**

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.

**Dataset**

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

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

**Pre-training Tasks**

<table>
<thead>
<tr>
<th>Condition</th>
<th>Single-task ControlNet</th>
<th>UniControl</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canny</td>
<td>0.564</td>
<td>0.577</td>
</tr>
<tr>
<td>RED</td>
<td>0.466</td>
<td>0.582</td>
</tr>
<tr>
<td>Normal</td>
<td>0.623</td>
<td>0.778</td>
</tr>
<tr>
<td>Depth</td>
<td>0.654</td>
<td>0.700</td>
</tr>
<tr>
<td>Pose</td>
<td>0.741</td>
<td>0.747</td>
</tr>
<tr>
<td>Segmentation</td>
<td>0.693</td>
<td>0.693</td>
</tr>
</tbody>
</table>

**Zero-shot Tasks**

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

<table>
<thead>
<tr>
<th></th>
<th>Stable Diffusion</th>
<th>ControlNet</th>
<th>MoE-Adapter</th>
<th>TaskHyperNet</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>UniControl</td>
<td>1065.7M</td>
<td>361M</td>
<td>0.008</td>
<td>12.7M</td>
<td>4.32B</td>
</tr>
<tr>
<td>Multi-ControlNet</td>
<td>1065.7M</td>
<td>361M x 9</td>
<td>0.008</td>
<td>12.7M</td>
<td>4.32B</td>
</tr>
</tbody>
</table>

Compared with our direct baseline - Multi-ControlNet, UniControl significantly compresses the model size by 7x overall and achieves comparable and even better performance on each task. It would be beneficial for:

1. **Image Storage**: There is only one checkpoint to save for UniControl whereas ControlNet has nine checkpoints instead.
2. **Inference Efficiency**: For multi-condition tasks the users would not need to load multiple models into memory when dealing with multiple spatial conditions for content generation.

**Table 2: Image Perceptual Distance**

- UniControl: Default, ControlNet: Baseline (361M Parameters, diffusion model).

---

*UniControl: A Unified Spatial-to-image Generative Model*