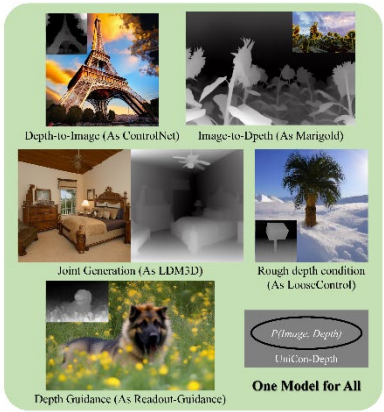
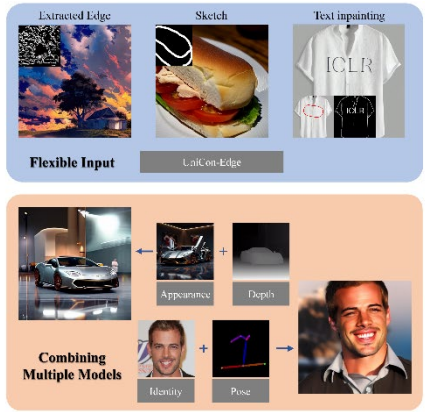
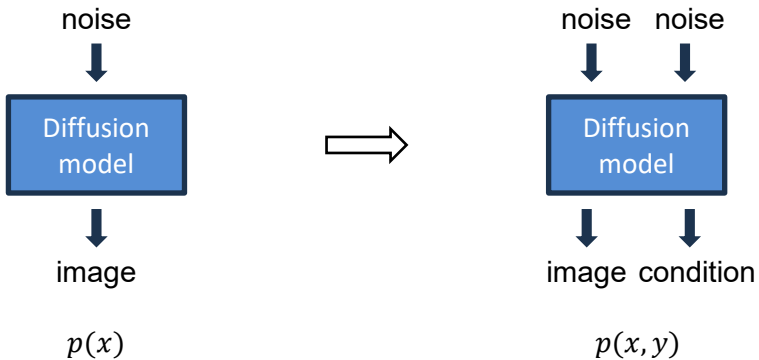
UniCon: A Simple Approach to Unifying Diffusion-based Conditional Generation





Project webpage

Xirui Li, Charles Herrmann, Kelvin C.K. Chan, Yinxiao Li, Deqing Sun, Chao Ma, Ming-Hsuan Yang



Adapt a diffusion model to a correlated (image, condition) pair.

Diffusion models have been extended for different generation tasks and condition paradigms.

Controllable Generation


ControlNet [1]

Estimation

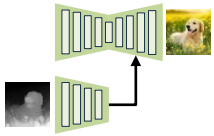

Marigold [2]

Joint Generation

"Some muffins on a table"

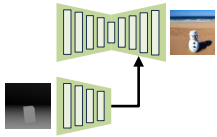

LDM3D [3]

Precise Control



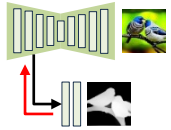
ControlNet [1]

Rough Control



LooseControl [4]

Guidance

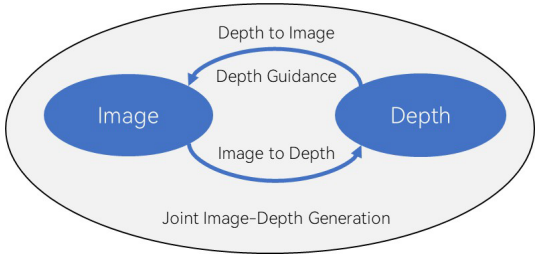


Readout-Guidance [5]

[1] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models". In ICCV. 2023.
[2] Bingxin Ke, Anton Obukhov, et al. "Repurposing diffusion-based image generators for monocular depth estimation". In CVPR. 2024.
[3] Gabriela Ben Melech Stan, Diana Wofk, et al. "Ldm3d: Latent diffusion model for 3d". arXiv preprint arXiv:2305.10853, 2023.
[4] Shariq Farooq Bhat, Niloy Mitra, and Peter Wonka. "Loosecontrol: Lifting controlnet for generalized depth conditioning.". In SIGGRAPH, 2024.
[5] Gabriela Ben Melech Stan, Diana Wofk, et al. "Readout Guidance: Learning Control from Diffusion Features". In CVPR, 2024.

Motivation

- All these tasks are conditional generation involving a specific image-condition correlation (e.g. image and depth).

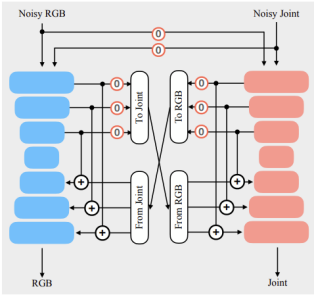


- UniCon enables diverse generation behavior in one model for a target image-condition pair.

Prior Work

JointNet

Learning the joint distribution with a ControlNet-like structure with dual residual connection.



Limitations:

- Heavy training
- Require explicit mask as model input
- Double the parameter number.

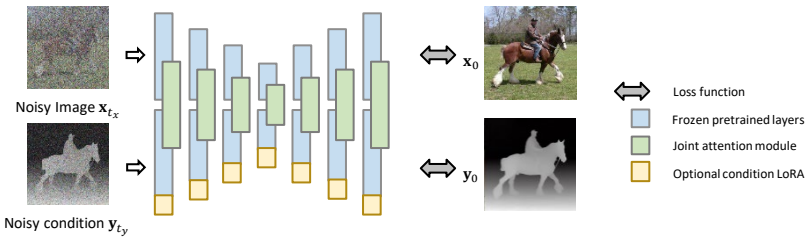
Method

To achieve the target, we aim to

1. Train a diffusion model for joint distribution p(x,y)
2. Develop sampling strategies for diverse generation tasks.

Method

How to adapt an existing model for p(x,y)?

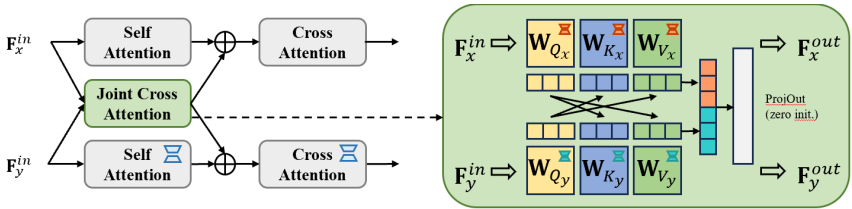


UniCon model structure

- Model structure: a lightweight adaptation on pretrained diffusion model (e.g. SD), by injecting a joint cross attention module.
- With minimal change to the model and about 15% additional parameters.

Method

Joint Cross Attention Module

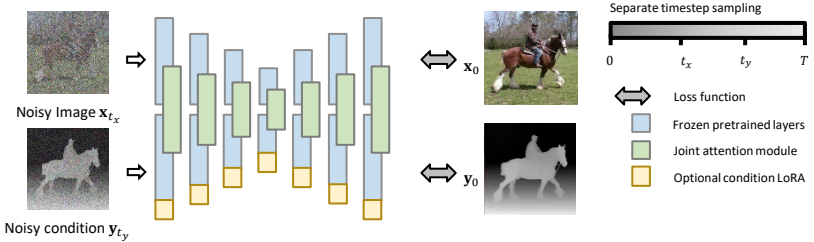


- Parallel to self attention modules.
- x,y attend to each other.
- Zero projection out.

Method

How to support different conditional generation tasks?

- Attempts made: inpainting, inpainting+guidance, adding inpainting mask to input.
- Finally, we train the model using independent noise levels for x and y
- Simple yet surprisingly effective.



Different Generation Behavior

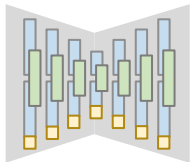
Depth to Image

Noise level: $T \rightarrow 0$

Image Input

Noise level: $0 \rightarrow 0$

Depth Input



UniCon-Depth

Image Output

Depth Output

Different Generation Behavior

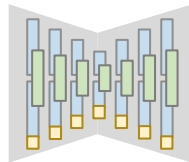
Image to Depth

Noise level: $0 \rightarrow 0$

Image Input

Noise level: $T \rightarrow 0$

Depth Input



UniCon-Depth

Image Output

Depth Output

Different Generation Behavior

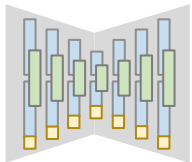
Image and Depth

Noise level: $T \rightarrow 0$

Image Input

Noise level: $T \rightarrow 0$

Depth Input



UniCon-Depth

Image Output

Depth Output

Different Generation Behavior

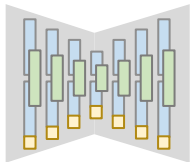
Rough Depth to Image

Noise level: $T \rightarrow 0$

Image Input

Noise level: $\frac{3}{4}T \rightarrow 0$

Depth Input



UniCon-Depth

Image Output

Depth Output

Results

One UniCon model supports diverse generation behavior at inference time.

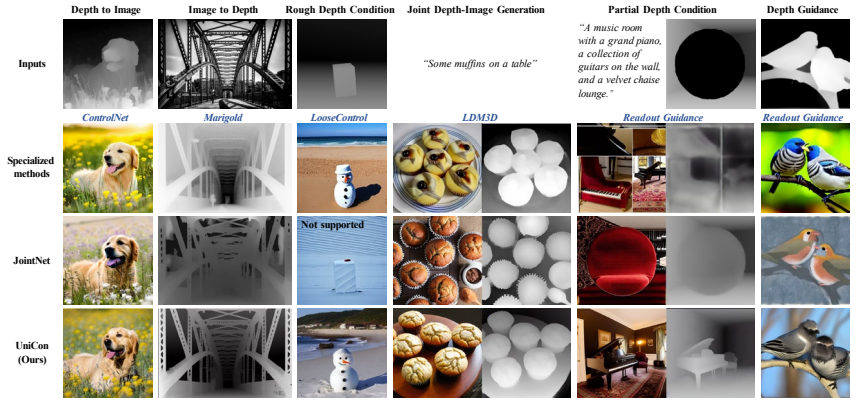
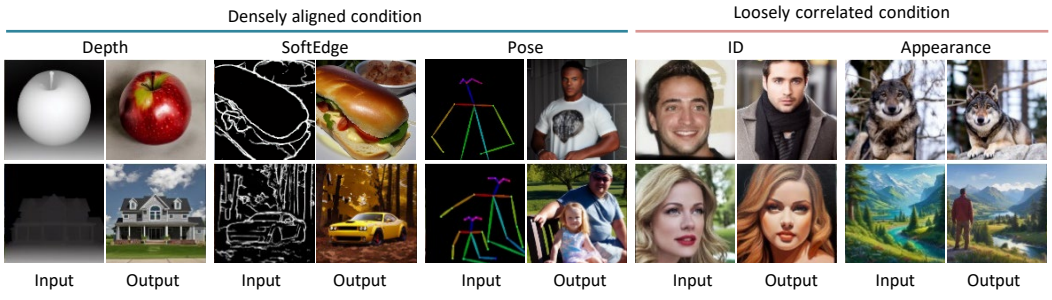


Figure 5: Qualitative comparison of UniCon-Depth and other specialized methods.

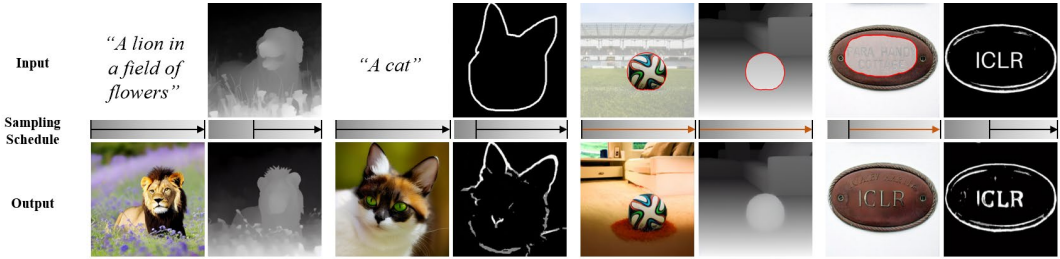
Results

UniCon models can be trained for various conditions, including densely aligned ones (depth, edge, pose) and loosely correlated ones (identity and appearance).



Results

UniCon models enable flexible conditional generation via different sampling schedules.



UniCon: A Simple Approach to Unifying Diffusion-based Conditional Generation

A simple framework that adapts an image diffusion model for versatile conditional generation tasks.

Key advantages:

- Enabling diverse generation behavior in one model
- Efficient training and parameter consumption.
- Achieving comparable results to specialized methods and better results than previous unified methods.

Main technical contribution:

- Proposing a framework including model adaptation, training strategy, and sampling methods allowing flexible conditional generation at inference.

Project webpage →

Thanks