InstantID

Instant Zero-shot Identity-Preserving Generation in Seconds

Qixun Wang[12], Xu Bai[12], Haofan Wang[12*], Zekui Qin[12], Anthony Chen[123], Huaxia Li[2], Xu Tang[2], and Yao Hu[2]

InstantX Team[1], Xiaohongshu Inc[2], Peking University[3]
{haofanwang.ai@gmail.com}
https://instantid.github.io

Zhuosheng Liu
May 7th
ECS 289G

Index of contents

Background knowledge

Image generative models brief overview

Generative adversarial network (GAN)

Pros

1. Simple neural network based architecture
2. High fidelity of generated images

Cons

1. Training instability due to battling of discriminator and generator
2. Diversity of generated images is limited (not distribution sampling model)
3. GAN is hard to interpret

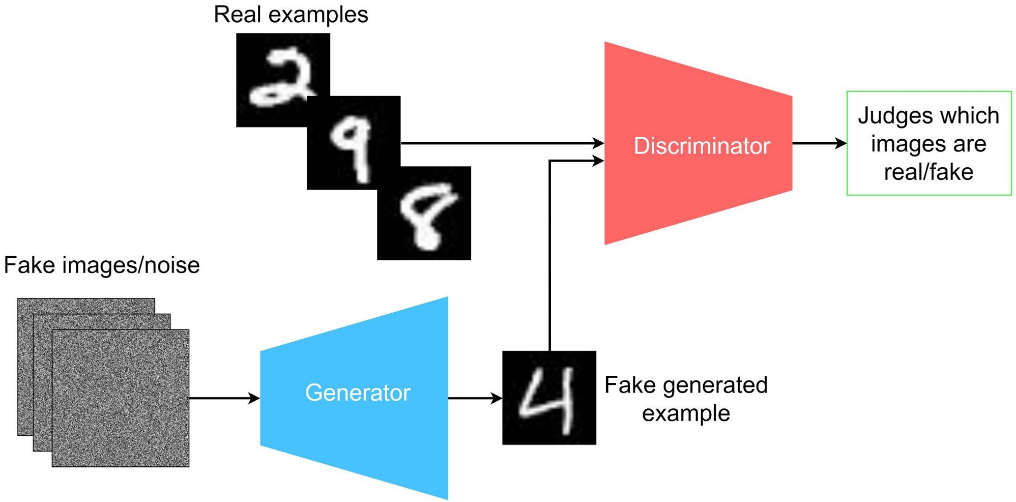


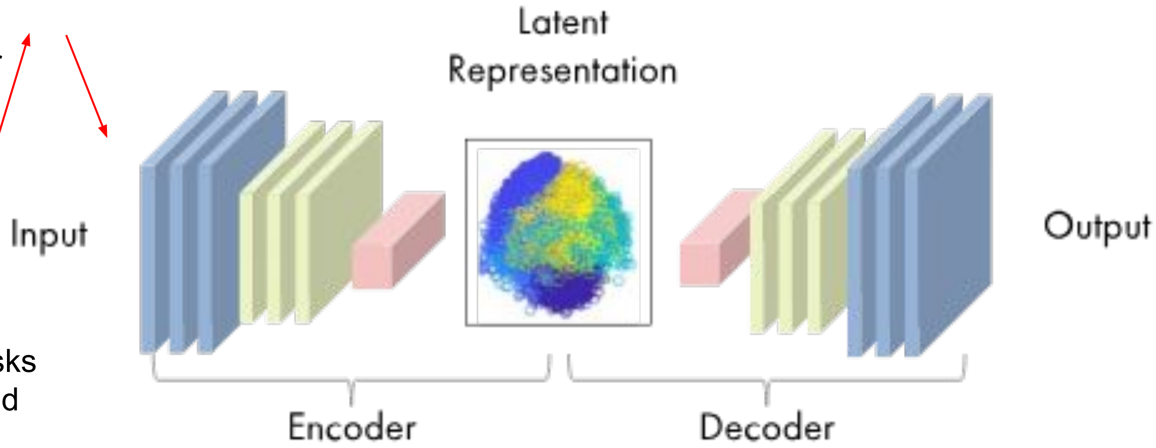
Image source: IBM

Autoencoder (AE) / Denoising auto-encoder (DAE)

Pros

1. Efficient feature learning via encoder
2. High quality reconstruction of input image using bottleneck

corrupted input (DAE)



Cons

This architecture was first designed for tasks such as classification, object detection, and segmentation using bottleneck (learned features from input).

Using bottleneck in generative tasks is not appropriate idea (no probability distribution information in bottleneck)

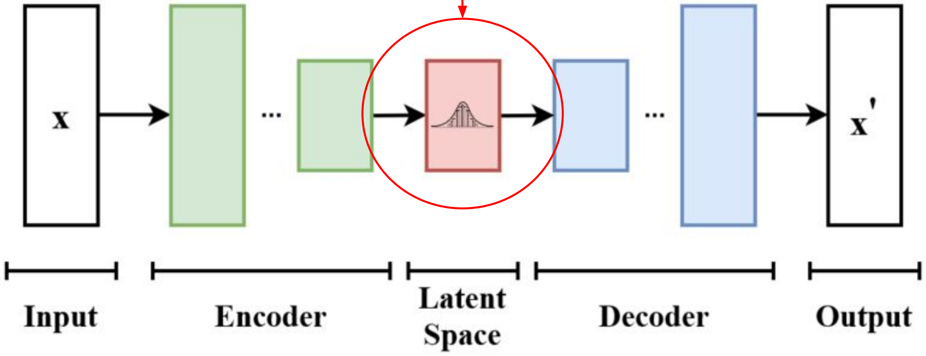
Goal: using output to reconstruct input

Image source: mathworks

Variational Autoencoder (VAE)

Instead of fixed bottleneck feature map, here it is learned distribution of input features

Compared with AE/DAE:
A learned distribution of input features was learned



This makes encoder-decoder architecture suitable for generation task!

Directly using sampled distribution for downstream generation tasks

Image source: wikipedia

Text-to-Image Diffusion Models

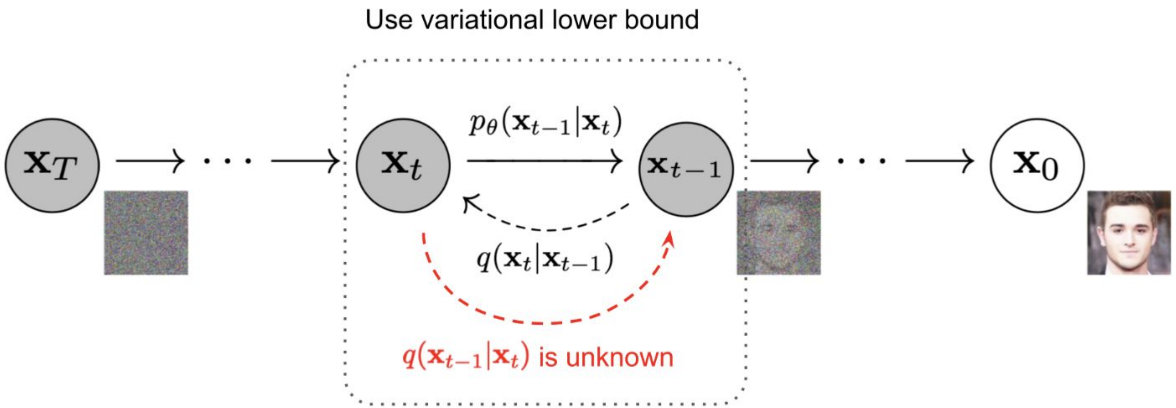
Text-to-image diffusion models are a type of generative model that can create detailed images from textual descriptions.

GLIDE; DALLE2

Imagen

IPAdaptor

Diffusion model



Forward diffusion process:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I})$$

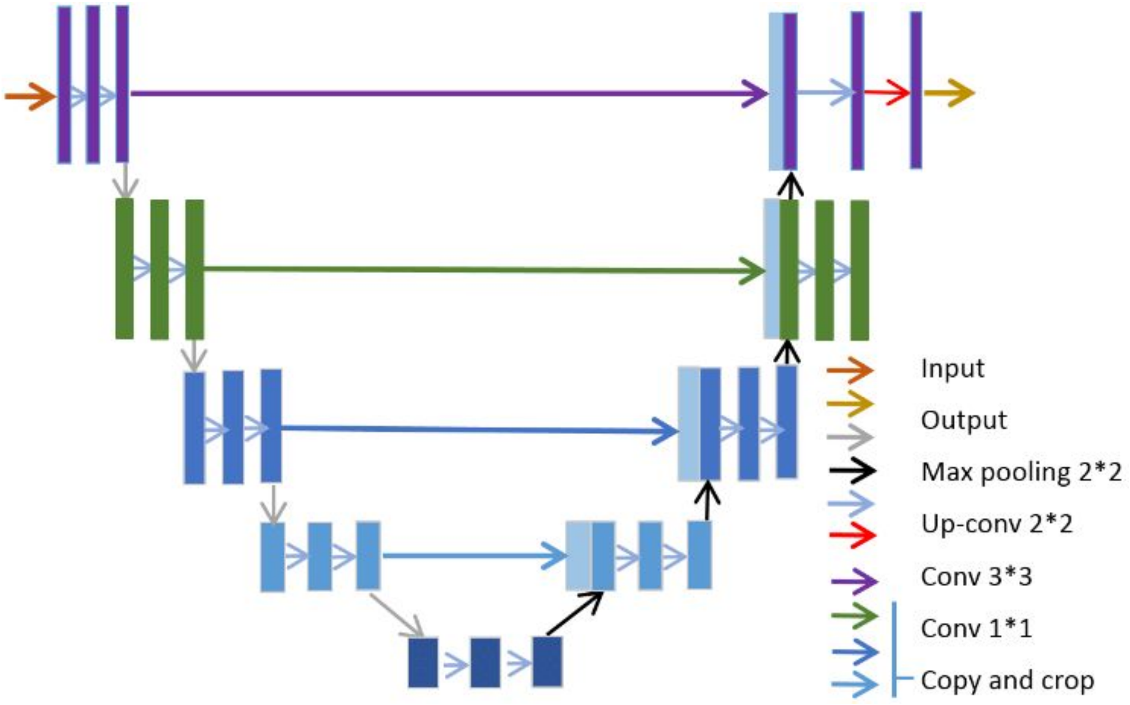
Reverse diffusion process:

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T)\prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) \quad p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t))$$

U Net (model used in reverse diffusion)

Why U-net?

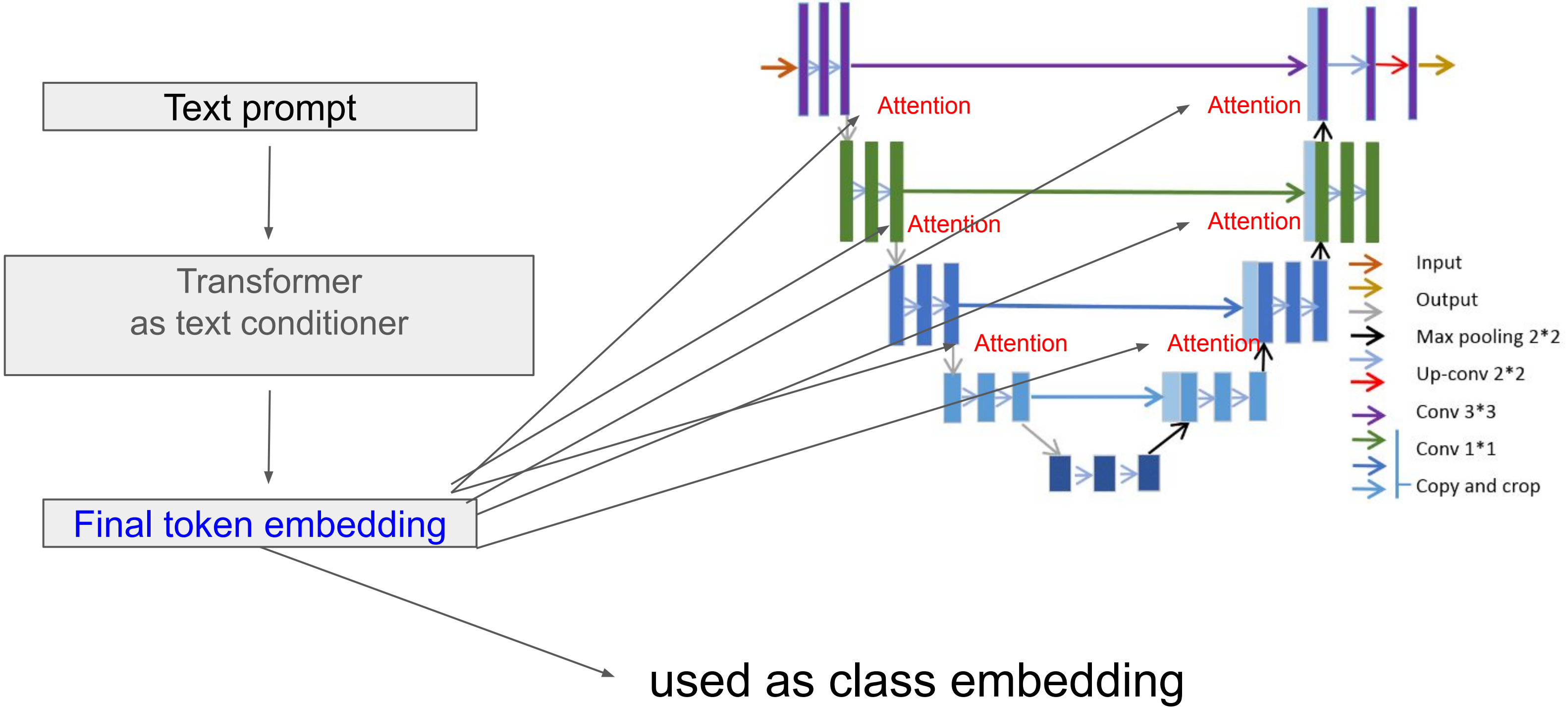
1. U-net gradually predicts and subtracts the noise from the noisy image step-by-step to recover the original image.
2. U-net preserves the data dimensionality



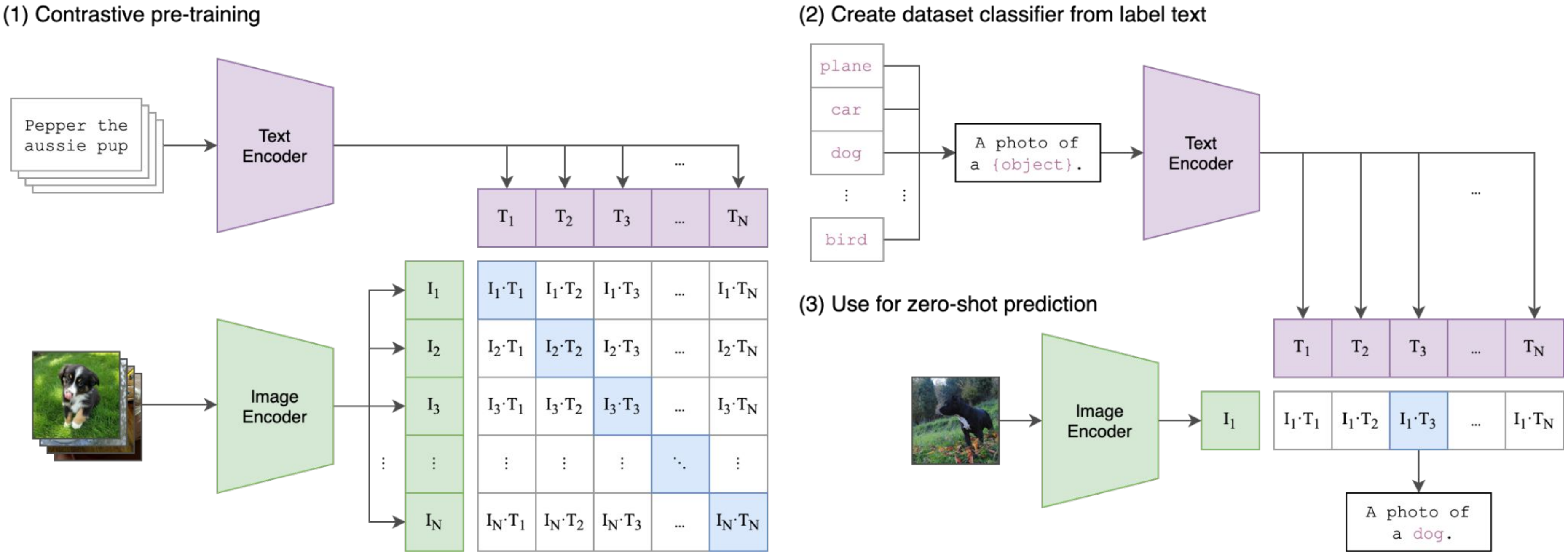
Downsampling first

Upsampling later

GLIDE (Guided Language to Image Diffusion for Generation and Editing)



CLIP (Contrastive Language-Image Pre-Training)

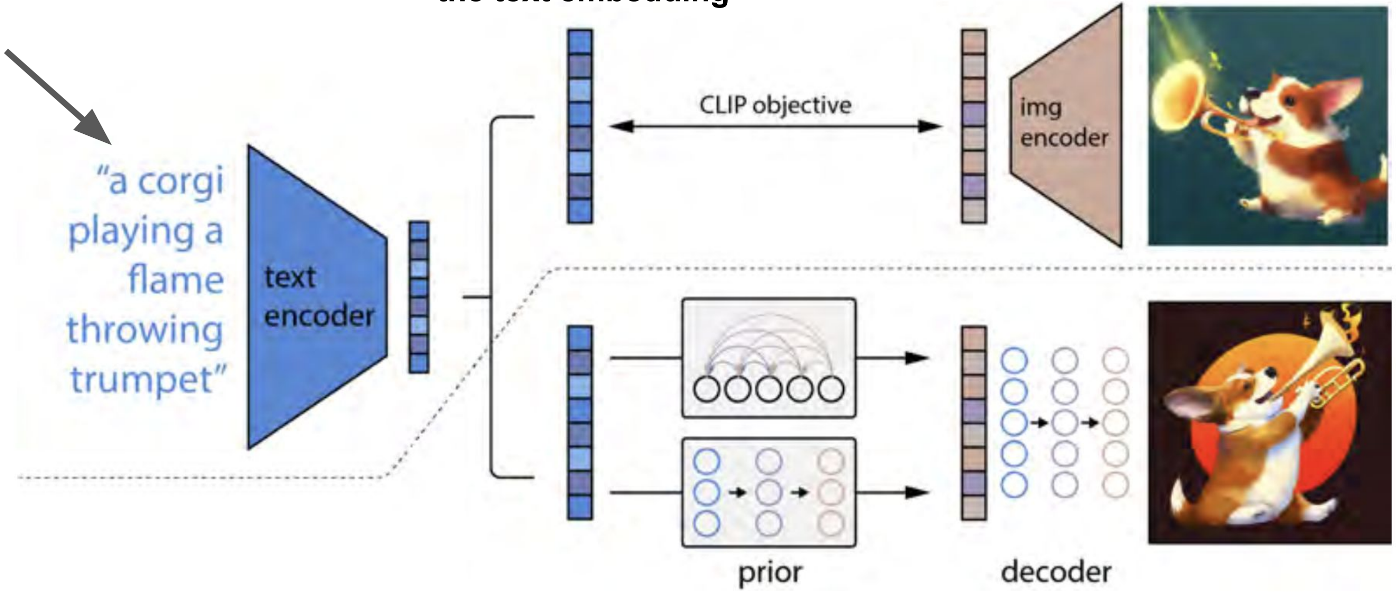


DALLE-2 (CLIP+GLIDE)

CLIP can capture much of information based prompt in the text embedding

Img encoder from frozen CLIP serves as the groundtruth

Frozen CLIP



Q&A: this part of model in DALLE-2 is also called unCLIP. Any ideas why?

prior: generate CLIP image embedding (diffusion based)

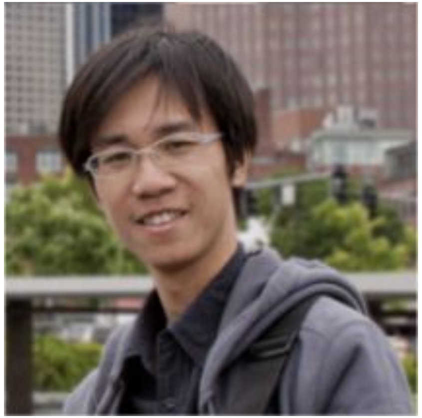
decoder: generate an image conditioned on image embedding

Motivation

Limitations using CLIP guided image generation

Current issues regarding CLIP guided image generation

1. CLIP embedding is relatively coarse-aligned
    a. CLIP tends to produce only weakly aligned signals, falling short in creating high-fidelity, customized images



Reference image

CLIP guided stable diffusion

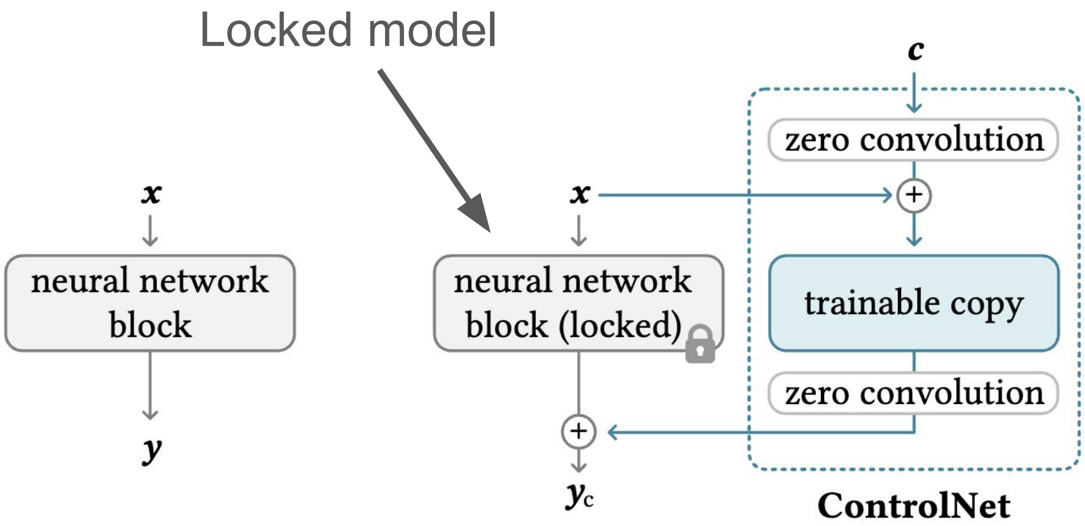
Works related to this study

Controlnet and IPAdaptpr

ControlNet

ControlNet is a neural network structure to control diffusion models by adding extra conditions.

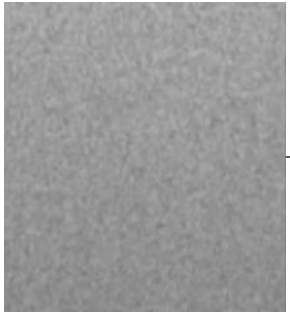
Hidden motivation: How to train a model to take additional conditioning inputs



(a) Before (b) After

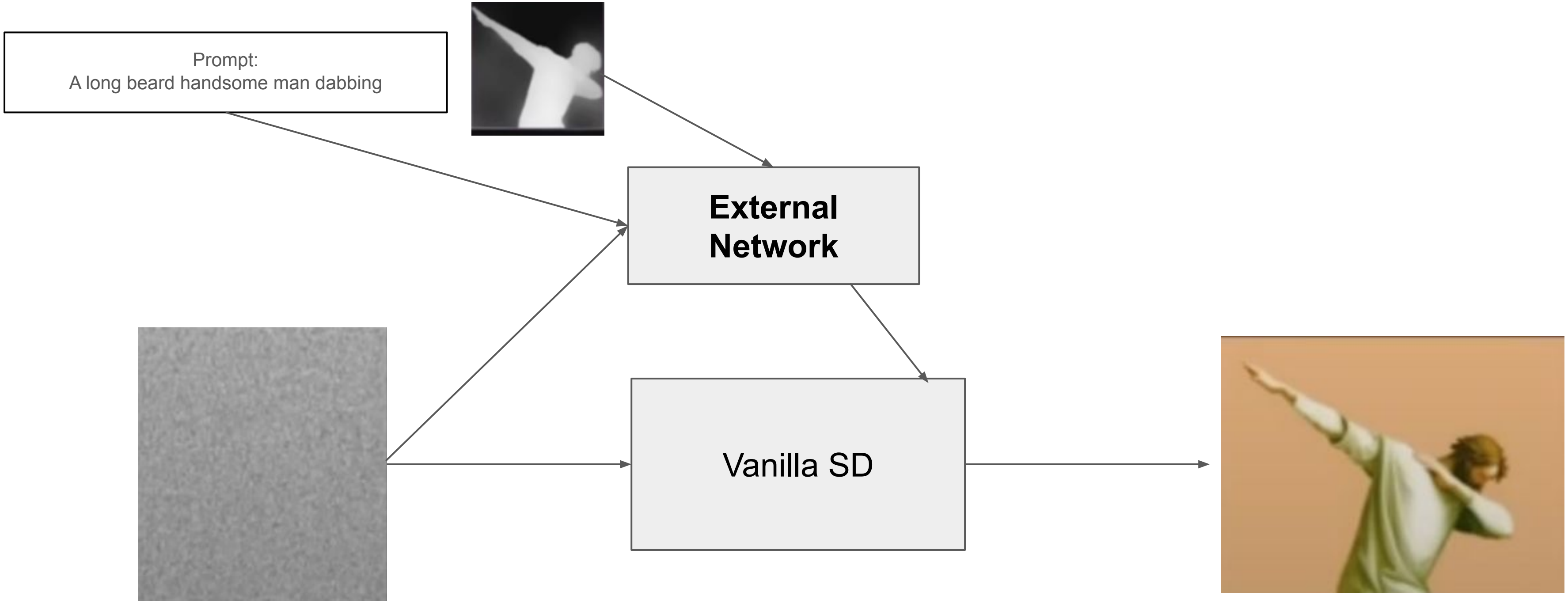
ControlNet

Prompt:
A long beard handsome man dabbing



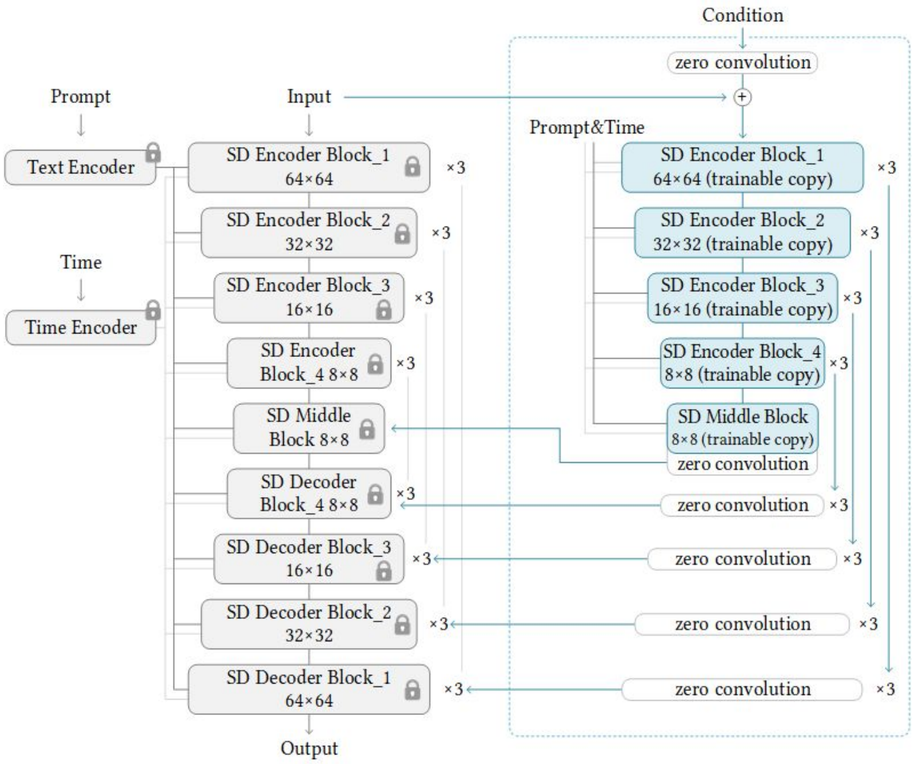
Vanilla SD

ControlNet


Prompt:
A long beard handsome man dabbing
External Network
Vanilla SD

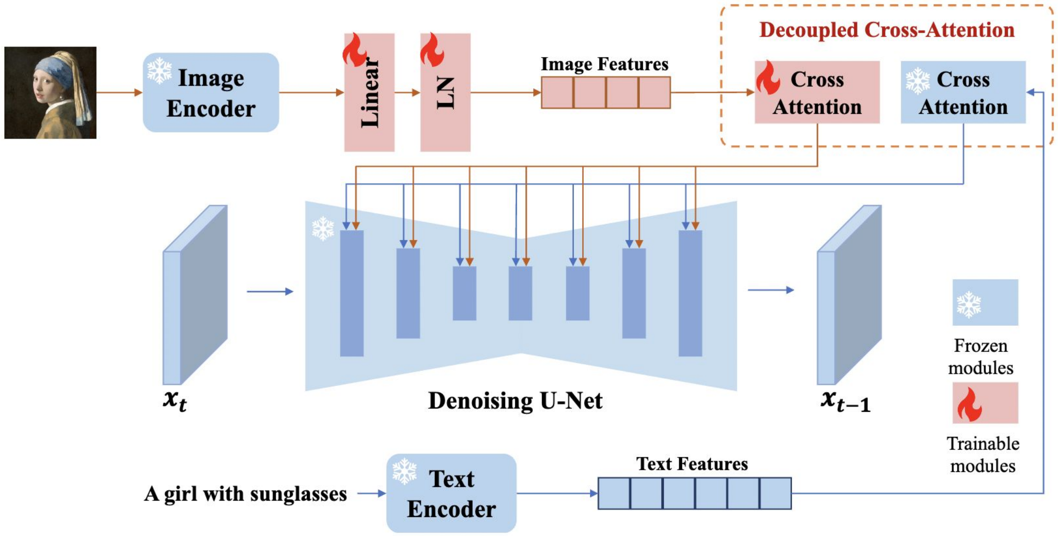
Leveraging the repetition of a simple structure in stable diffusion

ControlNet harnesses the SD encoder as a resilient backbone, facilitating stable diffusion and enabling versatile control learning.



(a) Stable Diffusion (b) ControlNet

IPAdaptor



Text-driven cross-attention + Image-driven cross-attention = Decoupled cross-attention

$$\mathbf{Z}' = \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V},$$

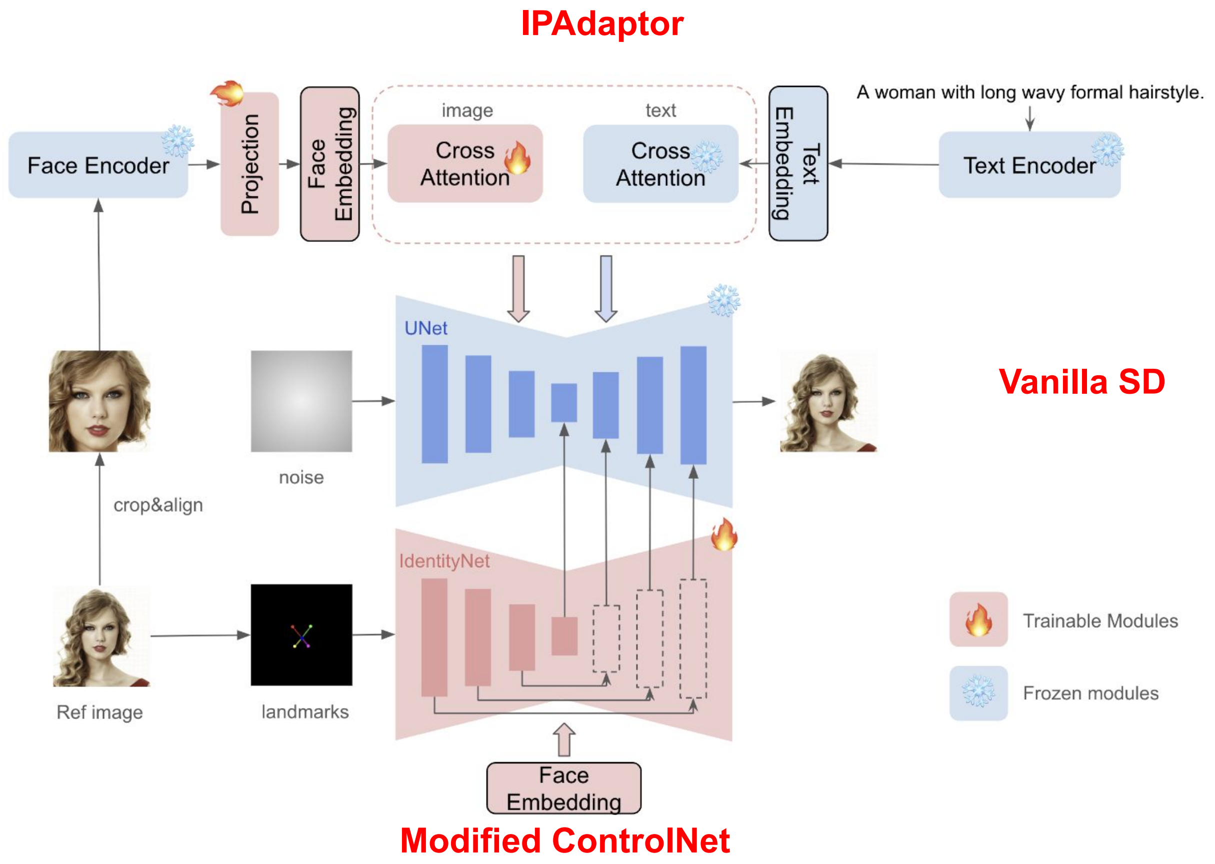
$$\mathbf{Z}'' = \text{Attention}(\mathbf{Q}, \mathbf{K}', \mathbf{V}') = \text{Softmax}\left(\frac{\mathbf{Q}(\mathbf{K}')^\top}{\sqrt{d}}\right)\mathbf{V}',$$

$$\mathbf{Z}^{new} = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V} + \text{Softmax}\left(\frac{\mathbf{Q}(\mathbf{K}')^\top}{\sqrt{d}}\right)\mathbf{V}'$$

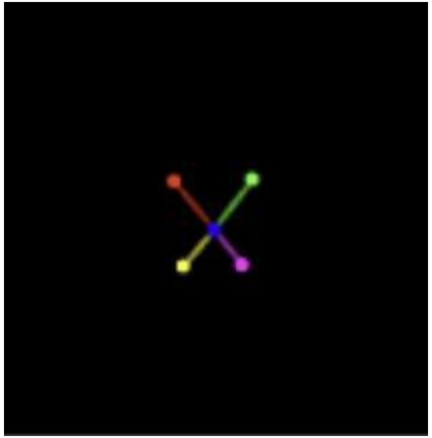
where $\mathbf{Q} = \mathbf{Z}\mathbf{W}_q, \mathbf{K} = \boldsymbol{c}_t\mathbf{W}_k, \mathbf{V} = \boldsymbol{c}_t\mathbf{W}_v, \mathbf{K}' = \boldsymbol{c}_i\mathbf{W}'_k, \mathbf{V}' = \boldsymbol{c}_i\mathbf{W}'_v$

InstantID Framework

Instant ID architecture



IdentityNet



landmark

Extracting facial features

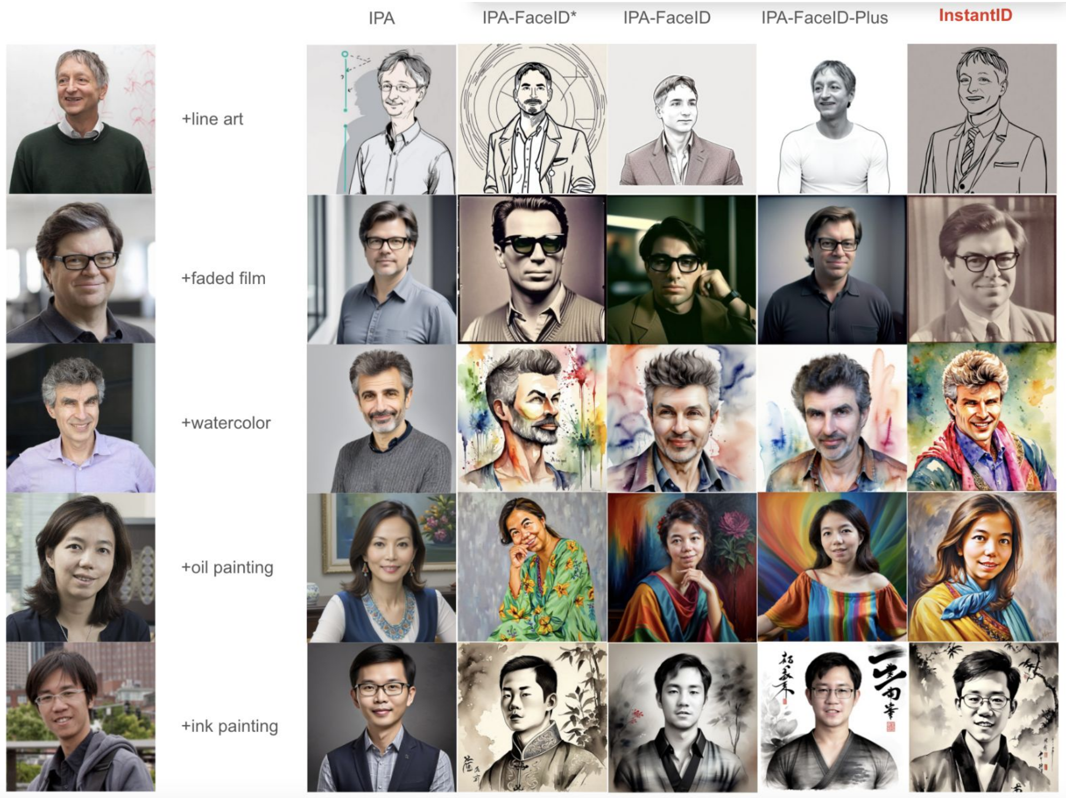
Ref image

Using facial landmarks (two for the eyes, one for the nose, and two for the mouths with elimination of text prompts)

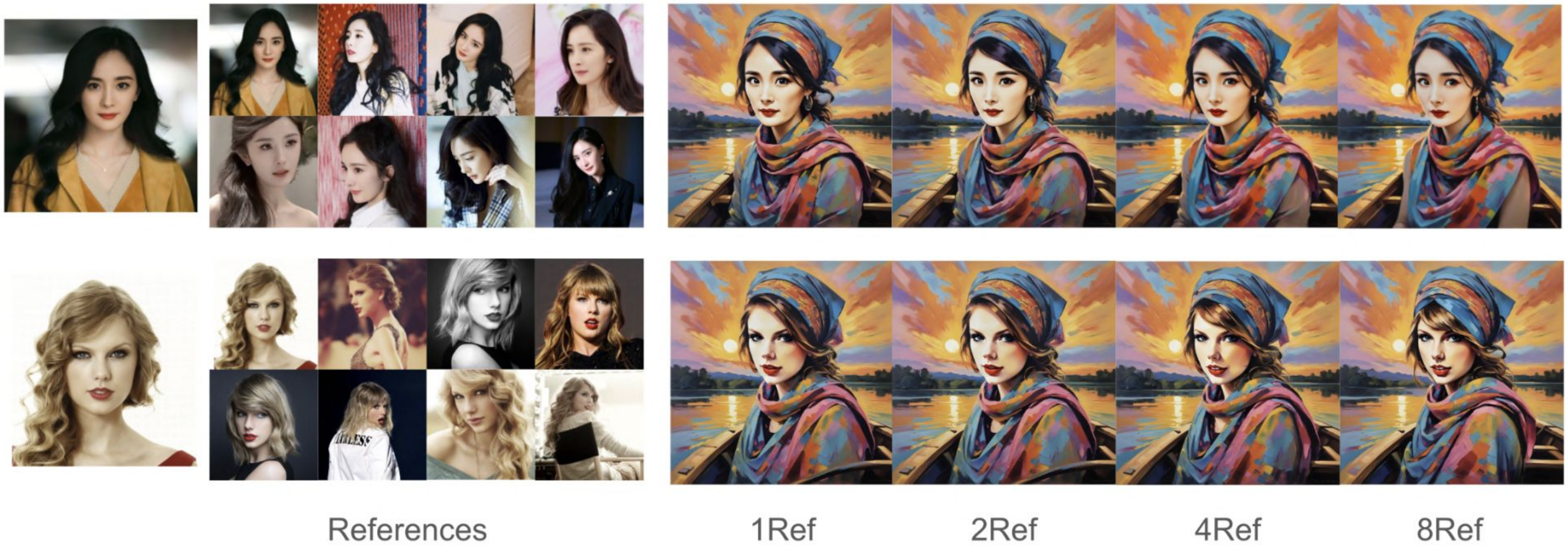
Results

Performance of InstantID

Comparison of InstantID with other methods conditioned on different characters and styles.



Effect of the number of reference images



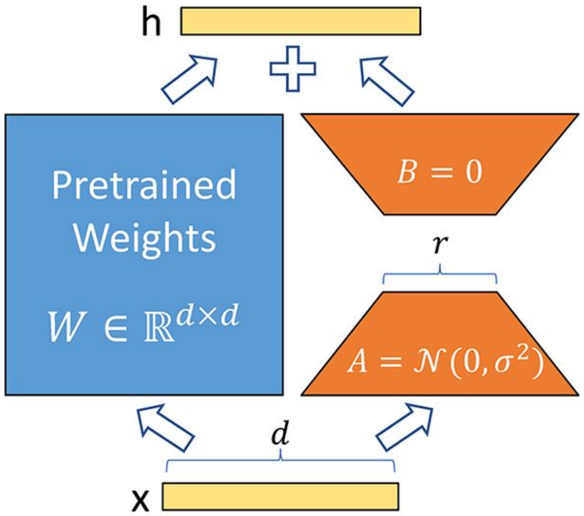
Demonstration of the robustness, editability, and compatibility of InstantID

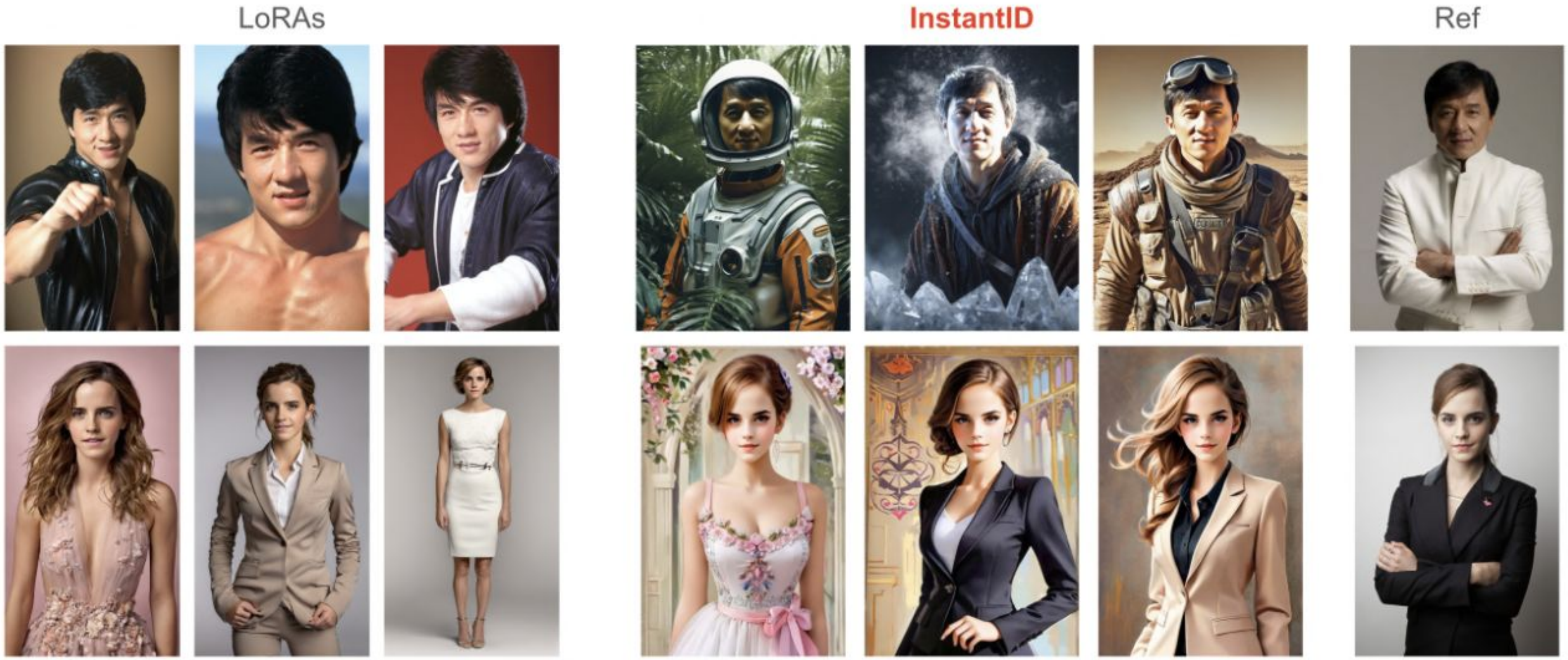
Results

Comparison with pre-trained character LoRA models

Low-Rank Adaptation (LoRA)

- LoRA is a technique specific for fine tuning large language models initially. More studies use LoRA to finetune vision-language model in a trend.
- Motivation: Downstream fine-tunings have low intrinsic dimension.
- Fine-tuned weight = W_0+W_Δ, where W_Δ is based on low intrinsic rank





Comparison of InstantID with pre-trained character LoRAs

Applications

Application case 1: Image generation based on both ref image and pose image

Novel View Synthesis under any given pose

Applications

Application case 2: Image generation using single ref image for multiple identity

Multi-identity synthesis with regional control

Limitations

1. ID embedding in our model, while rich in semantic information like gender and age, has highly coupled facial attributes, which poses a challenge for face editing.
    a. The authors in this paper mentioning decouple facial attribute features to enhance further flexibility
2. The biases inherent in the face models used in this study

References

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., ... & Chen, M. (2021). Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., ... & Wu, J. (2020, May). Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 1055-1059). IEEE.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, *1*(2), 3.

Wang, Q., Bai, X., Wang, H., Qin, Z., & Chen, A. (2024). Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*.

Ye, H., Zhang, J., Liu, S., Han, X., & Yang, W. (2023). Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

Zhang, L., Rao, A., & Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 3836-3847).

Q&A