

Efficient text-to-video generation with latent diffusion models

Group member: Hongxun Ding, Mingjian Zhu

Supervisor: Dan Zeng, Bo Tang

March 30, 2022

Outline

- Problem introduction
- Challenges
- Related work
 - ◆ CogVideo, Make a video
 - ◆ Latent diffusion model
- Timeline
- References

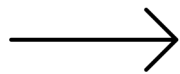
Problem introduction

■ Text-to-video problem

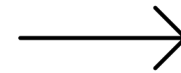
- ◆ Generate a video from a given textual information
- ◆ Extension of text-to-image problem

Input text: a description sentence

A dog wearing a
Superhero outfit with
red cape flying
through the sky



T2V Model



Output video: high-quality
video with coherent motion
(e.g., a 3-second clip of 32
frames)



Outline

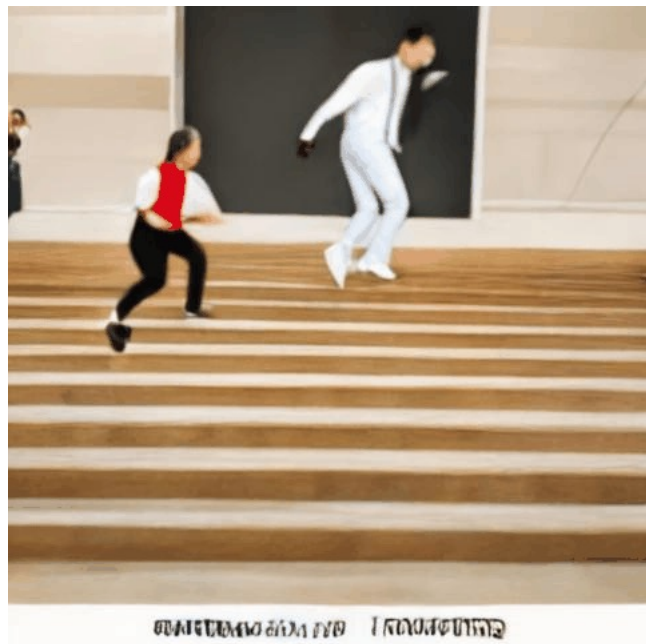
- Problem introduction
 - Challenges
 - Related work
 - ◆ CogVideo, Make a video
 - ◆ Latent diffusion model
 - Timeline
 - References
-

Challenges

- Scarcity and weak relevance of text-video data
 - ◆ Massive text-image pairs, yet biggest text-video dataset VATEX has only 41,250 videos
 - ◆ Most pairs in Howto100M only describe the scene without the temporal information
- Misalignment between the text and its temporal counterparts
 - ◆ Confused to learn the accurate meaning of complex action
 - ◆ Not coherent

Bad cases

Originated from model in CogVideo.



A male student is chasing a female student.



一个男人在公交车上为一个老奶奶让座。

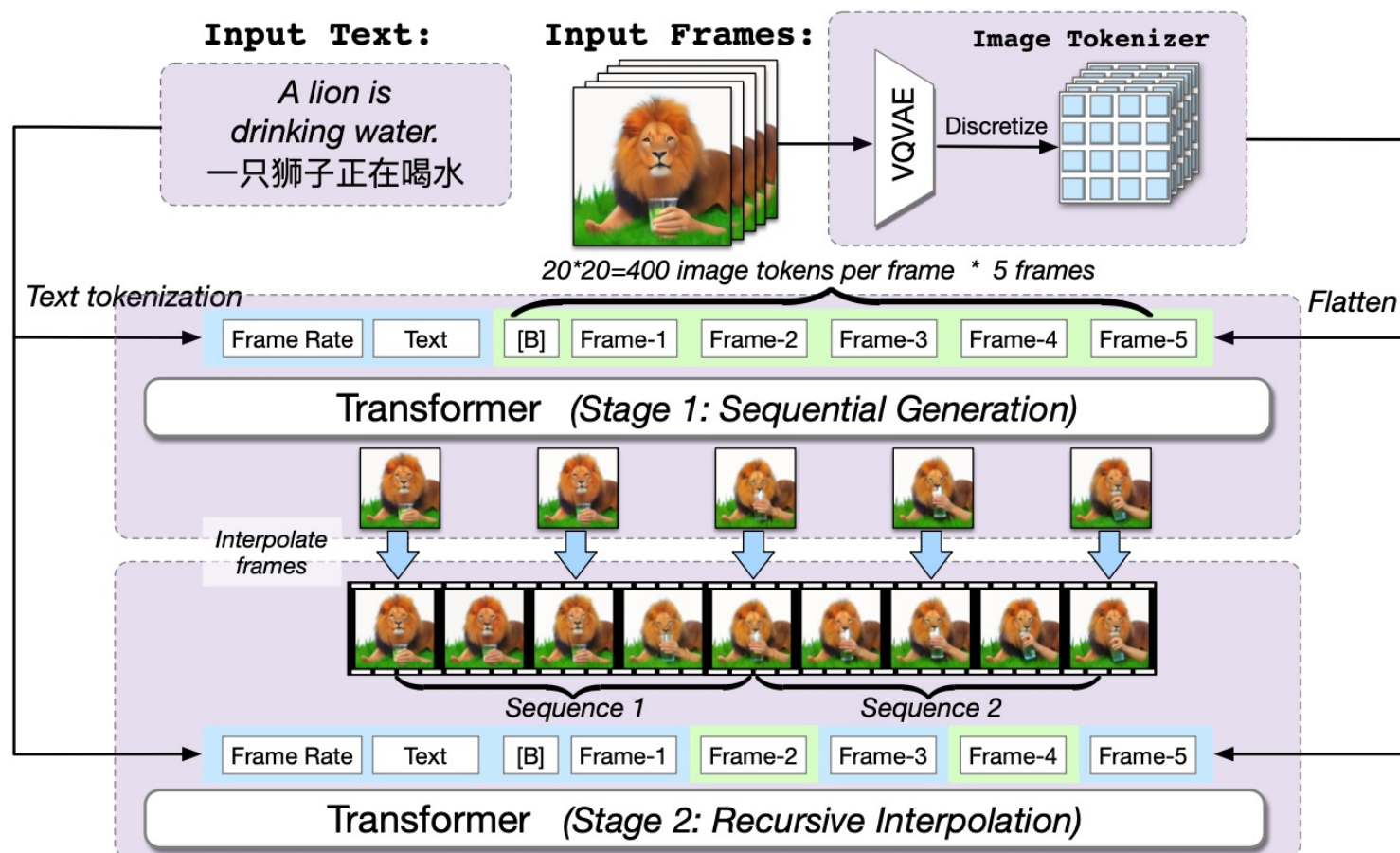


弘扬社会正能量。

Outline

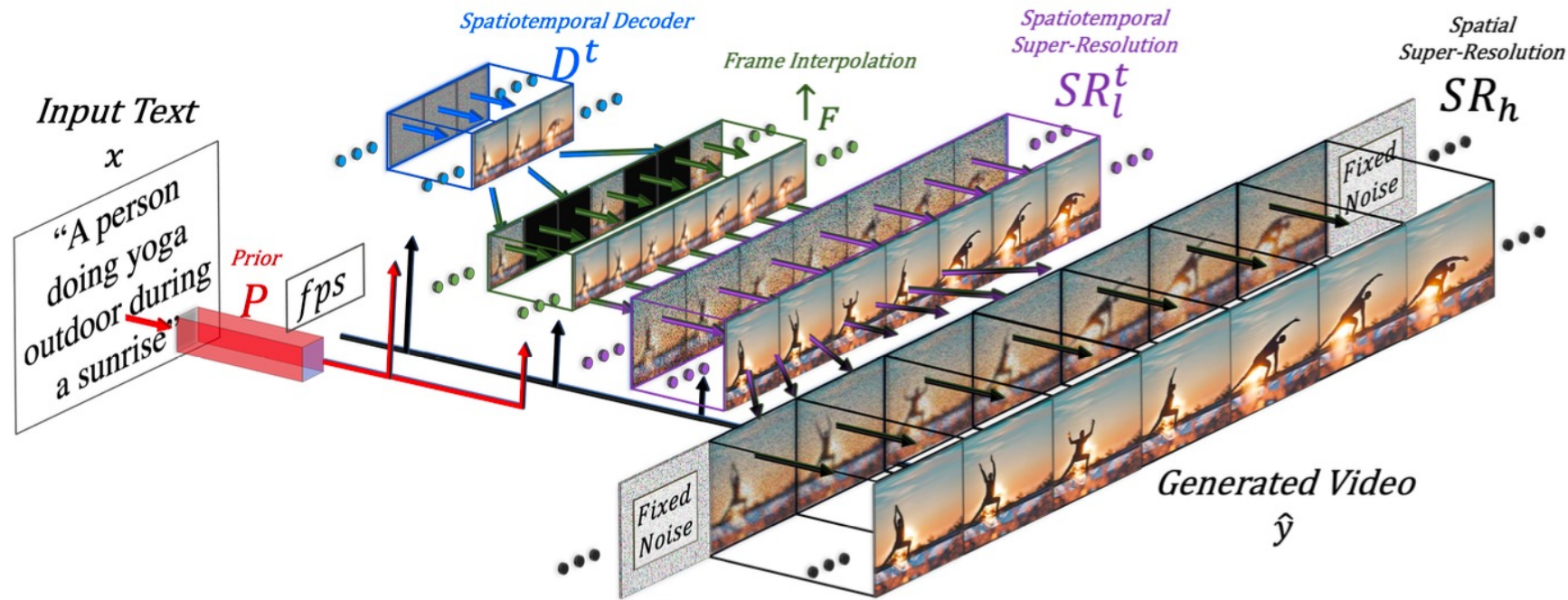
- Problem introduction
 - Challenges
 - Related work
 - ◆ CogVideo, Make a video
 - ◆ Latent diffusion model
 - Timeline
 - References
-

CogVideo (arXiv 2022)



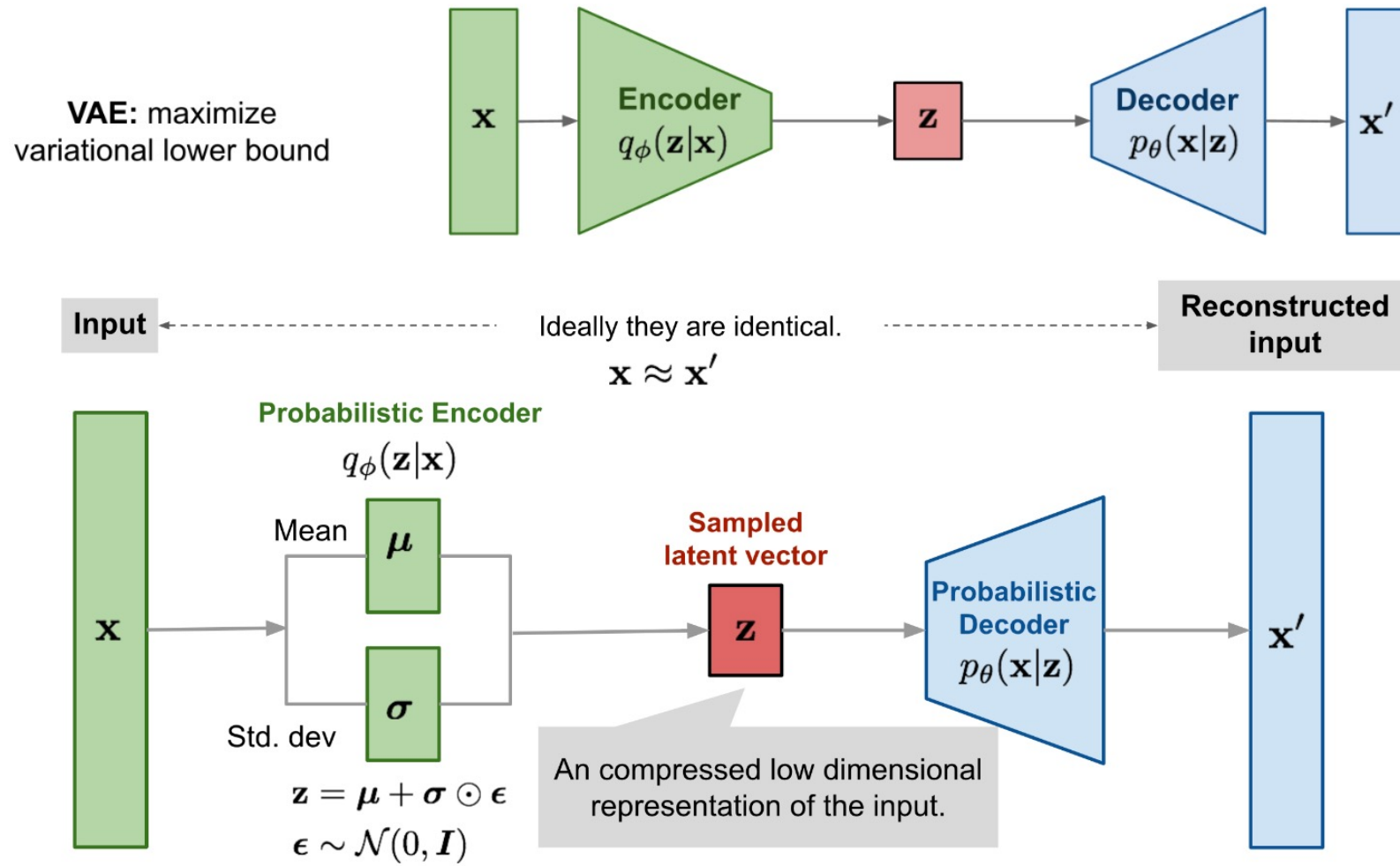
- Sequentially generate 5 key frames based on a low frame rate and text.
- Recursively interpolate frames based on the text, frame rate and known frames.

Make-a-video (arXiv 2022)



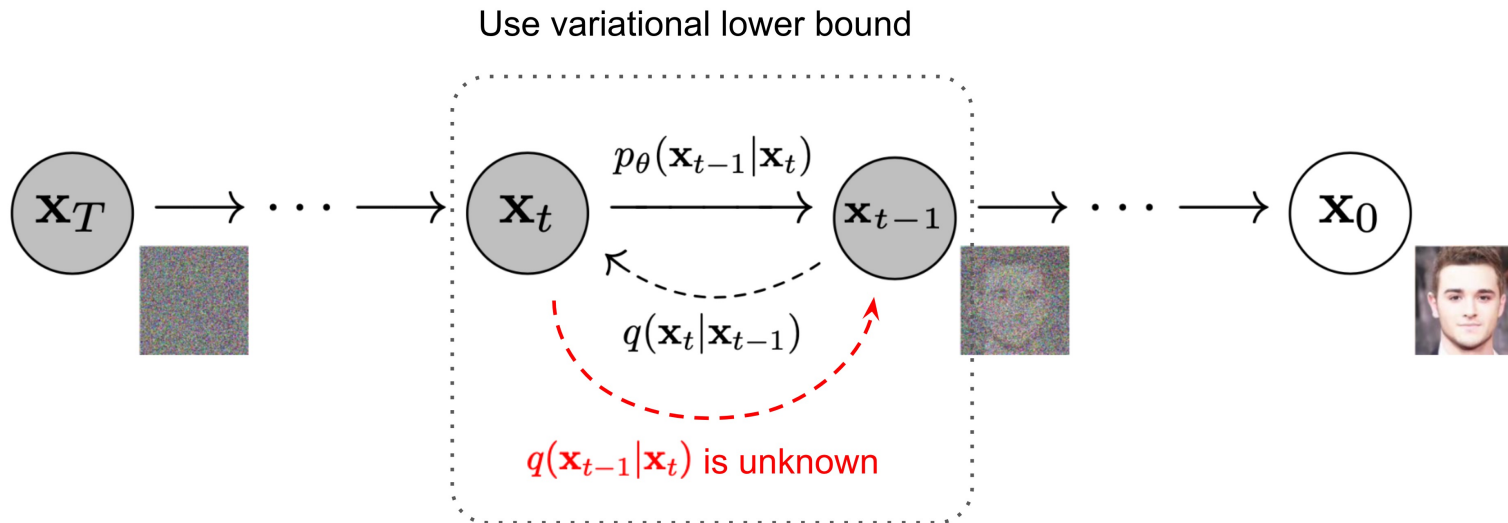
- Based on pretrained text-to-image model (priors)
- Spatiotemporal decoder and frame interpolation
- Super resolution

Variational autoencoder (VAE)



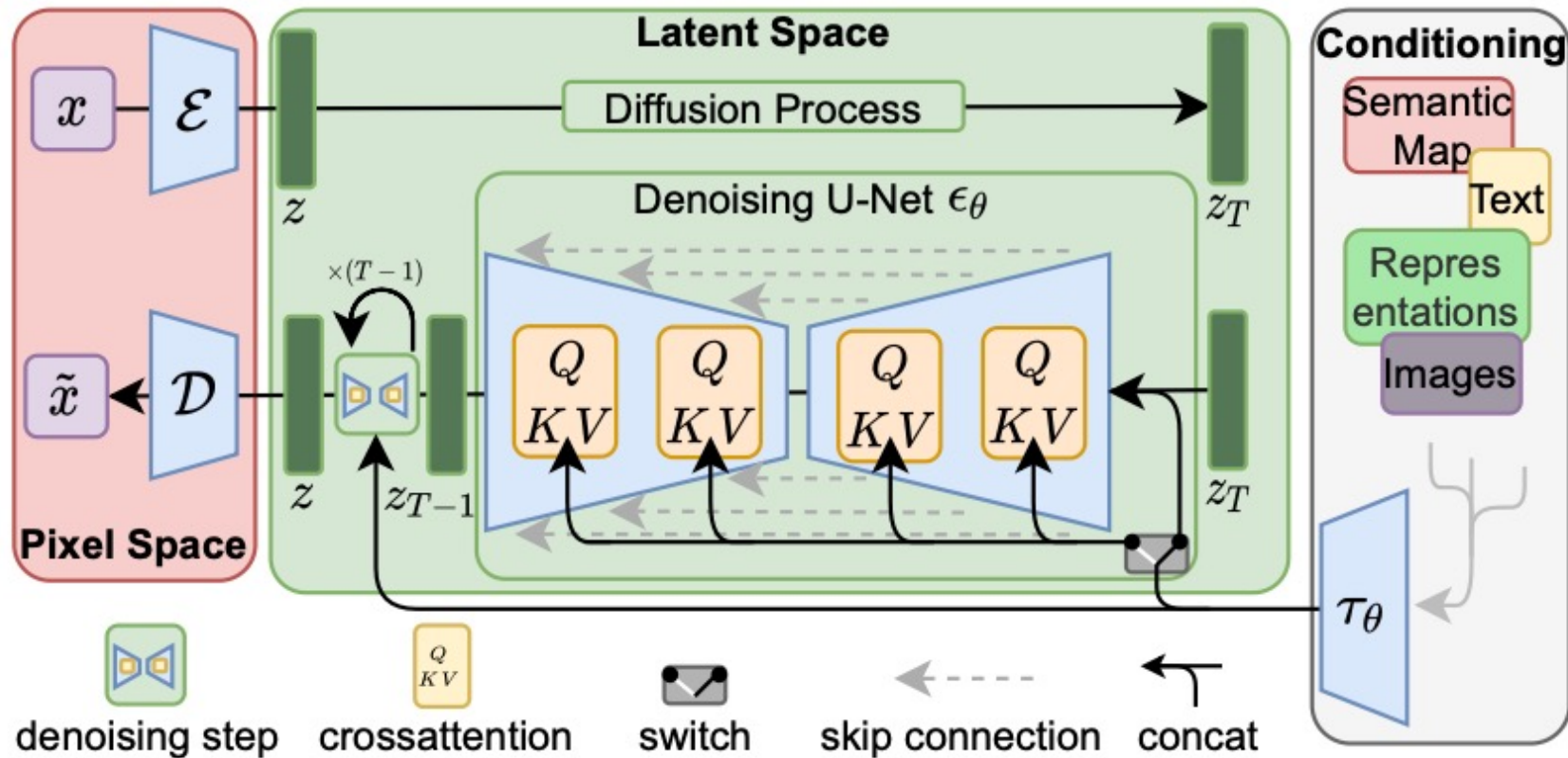
Diffusion model

Diffusion models:
Gradually add Gaussian
noise and then reverse



- Forward (diffusion process): add small amount of Gaussian noise to the sample in T steps
- Backward: reverse the above process
- Reconstruct the true sample from a Gaussian noise input

Latent diffusion model (Image generation)



- Diffusion process on latent space
 - ◆ Efficient
 - ◆ Faithful reconstruction
- Conditioning mechanism based on cross attention
 - ◆ Multi-modal training
 - ◆ Text-to-image

Outline

- Problem introduction
 - Challenges
 - Related work
 - ◆ CogVideo, Make a video
 - ◆ Latent diffusion model
 - Timeline
 - References
-

Timeline

- Current progress
 - ◆ Finish survey (Mid-February)
 - ◆ Finish paper reading (Late February)
 - ◆ Inference of SOTA (Mid-March)
 - Before Midterm presentation
 - ◆ Simple demo of **latent diffusion model** in t2v generation
 - Before Final presentation
 - ◆ Complete and bug-free demo of **latent diffusion model** in t2v generation
-

Outline

- Problem introduction
 - Challenges
 - Related work
 - ◆ CogVideo, Make a video
 - ◆ Latent diffusion model
 - Timeline
 - References
-

Reference

- [1] Hong W, Ding M, Zheng W, et al. Cogvideo: Large-scale pretraining for text-to-video generation via transformers[J]. arXiv preprint arXiv:2205.15868, 2022.
- [2] Singer U, Polyak A, Hayes T, et al. Make-a-video: Text-to-video generation without text-video data[J]. arXiv preprint arXiv:2209.14792, 2022.
- [3] Ho J, Salimans T, Gritsenko A, et al. Video diffusion models[J]. arXiv preprint arXiv:2204.03458, 2022.
- [4] Ding M, Yang Z, Hong W, et al. Cogview: Mastering text-to-image generation via transformers[J]. Advances in Neural Information Processing Systems, 2021, 34: 19822-19835.
- [5] Ding M, Zheng W, Hong W, et al. Cogview2: Faster and better text-to-image generation via hierarchical transformers[J]. arXiv preprint arXiv:2204.14217, 2022.
- [6] Weng, Lilian. (Jul 2021). What are diffusion models? Lil'Log. <https://lilianweng.github.io/posts/2021-07-11-diffusion-models/>.
- [7] Kingma D P, Welling M. Auto-encoding variational bayes[J]. arXiv preprint arXiv:1312.6114, 2013.
- [8] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 10684-10695.

Thanks for your attention!