AI+研发数字峰会
2024 AI+研发数字峰会
AI+ Development Digital summit
AI驱动研发变革 促进企业降本增效
北京站 08/16-17

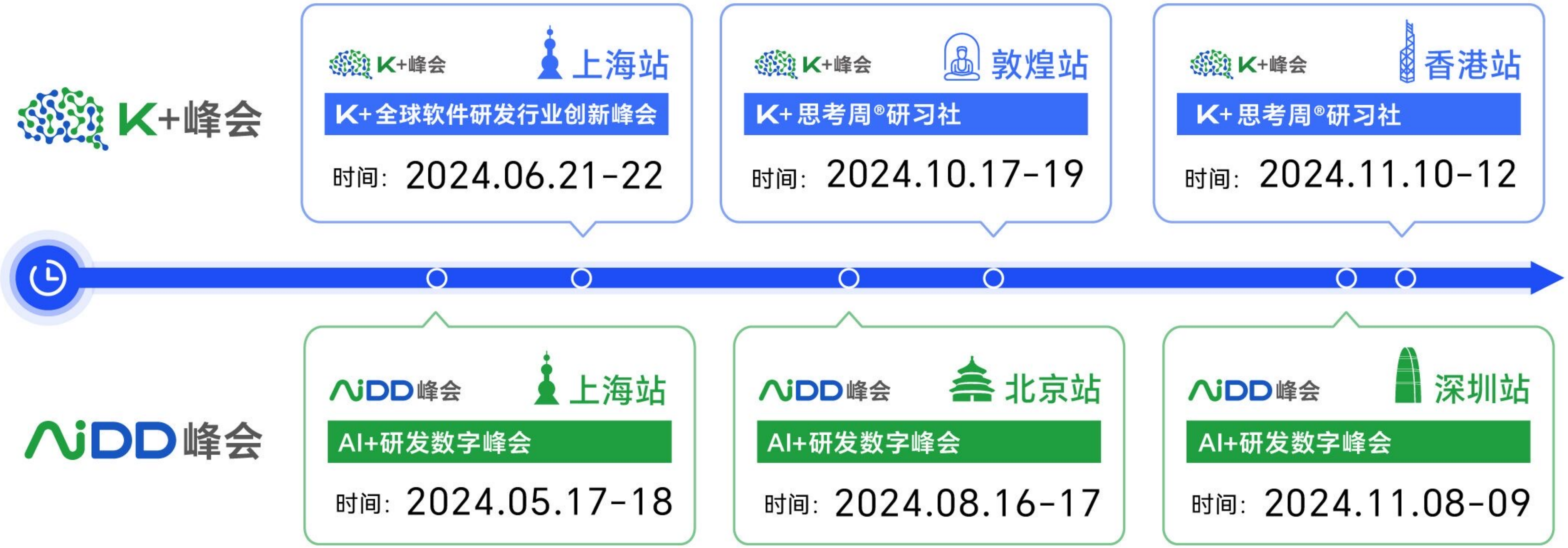
基于物理条件约束的可信视觉生成大模型

朱思语 复旦大学

科技生态圈峰会 + 深度研习

——1000 + 技术团队的共同选择

KEYLINK ing



K+峰会

K+峰会 上海站

K+全球软件研发行业创新峰会

时间：2024.06.21-22

K+峰会 敦煌站

K+思考周®研习社

时间：2024.10.17-19

K+峰会 香港站

K+思考周®研习社

时间：2024.11.10-12

AiDD峰会

AiDD峰会 上海站

AI+研发数字峰会

时间：2024.05.17-18

AiDD峰会 北京站

AI+研发数字峰会

时间：2024.08.16-17

AiDD峰会 深圳站

AI+研发数字峰会

时间：2024.11.08-09

K+峰会详情

AiDD峰会详情

AiDD

2024 AI+研发数字峰会

AI+ Development Digital summit

深圳站 11/08-09

AI驱动研发变革 促进企业降本增效

2024深圳站-议题设置

AI+产品线	LLM驱动产品创新	LLM驱动需求与业务分析	AI驱动设计与用户体验
AI+开发线	AI 原生应用开发框架与技术	AI Agents在研发落地实践	LLM驱动编程与单测
AI+测试线	LLM驱动测试分析与设计	基于LLM生成测试脚本与数据	LLM和AI应用的评测
AI+工程线	AI+DevOps 与工具 （LLM 时代的平台工程）	大模型对齐与安全	端侧大模型与云端协同
AI+领域线	领域大模型 SFT 与优化	知识增强与数据智能	大厂专场

扫描右侧二维码
查看更多会议详情

早鸟票限时抢购中 (截止到9月30日)

¥3680
早鸟票

¥2800
学生票

演讲嘉宾

朱思语

复旦大学教授

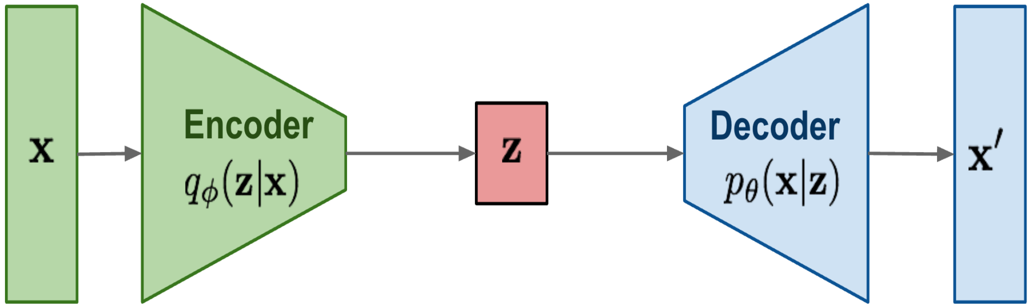
复旦大学人工智能创新与产业研究院研究员，长聘正教授，博士生导师。朱思语本科毕业于浙江大学，博士毕业于香港科技大学。在博士阶段，作为联合创始人创立了3D视觉公司Altizure，并后来被苹果公司收购。2017年至2023年，在阿里云人工智能实验室担任总监。2023年起，任职于复旦大学人工智能创新与产业研究院，担任研究员和博士生导师。朱思语的主要研究方向包括视频和三维生成式模型，涉及基于视觉的三维和视频的重建、生成、理解、方针和模拟。他发表了60余篇高水平会议和期刊论文，包括CVPR、ICCV、ICLR和TPAMI等计算机视觉和机器学习领域，包括Hallo， Champ， AnimateAnything等有一定行业影响力的视频生成大模型。在40余个计算机视觉国际比赛和榜单上取得第一名。

Visual generative model

Input

Output

VAE: maximize variational lower bound

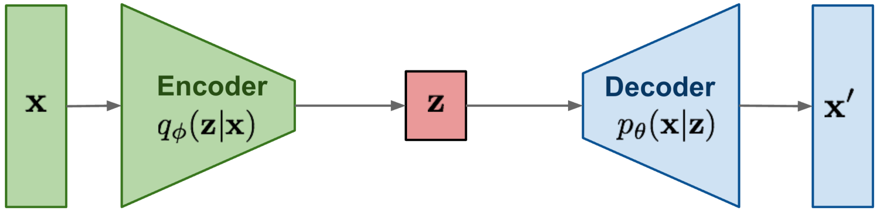


x
Encoder
$q_\phi(\mathbf{z}|\mathbf{x})$
z
Decoder
$p_\theta(\mathbf{x}|\mathbf{z})$
x'

Video generative methods

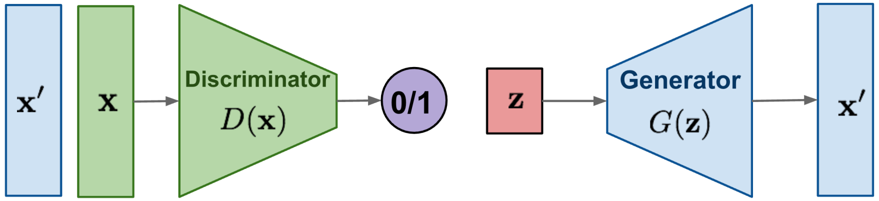
- The field of video generation has seen rapid development, reaching several milestones...

VAE: maximize variational lower bound

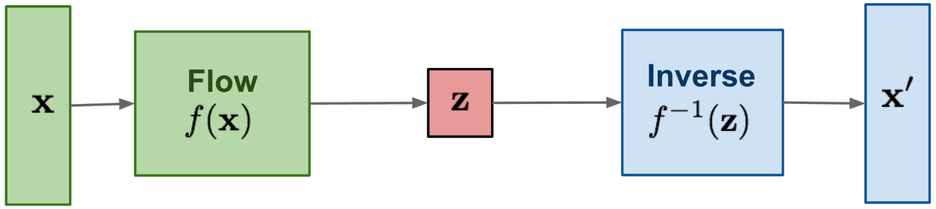


$\mathbf{x}$
Encoder
$q_\phi(\mathbf{z}|\mathbf{x})$
$\mathbf{z}$
Decoder
$p_\theta(\mathbf{x}|\mathbf{z})$
$\mathbf{x}'$

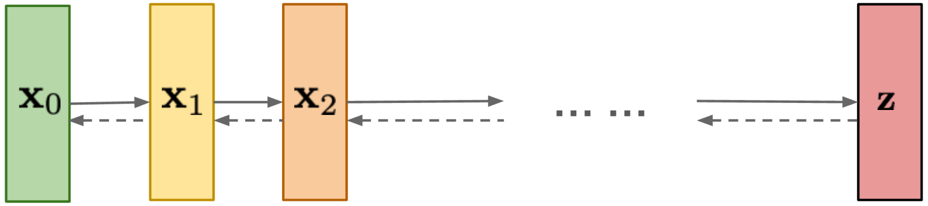
GAN: Adversarial training



Flow-based models: Invertible transform of distributions

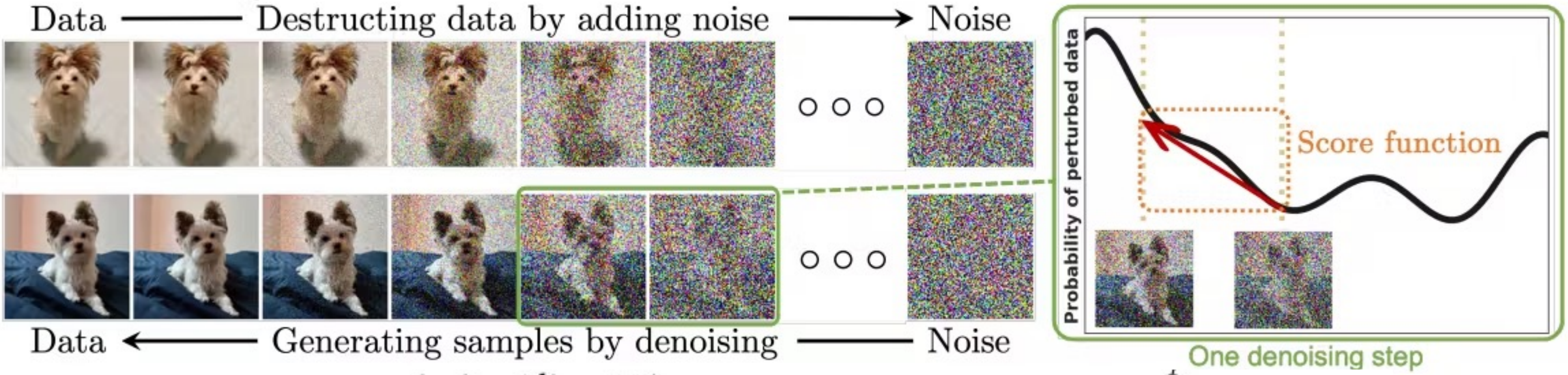


Diffusion models: Gradually add Gaussian noise and then reverse



Diffusion for visual generation (1)

- Denoising Diffusion Probabilistic Models (DDPMs)


Data
Destructing data by adding noise
Noise
Data
Generating samples by denoising
Noise
Probability of perturbed data
Score function
One denoising step

$$p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$$

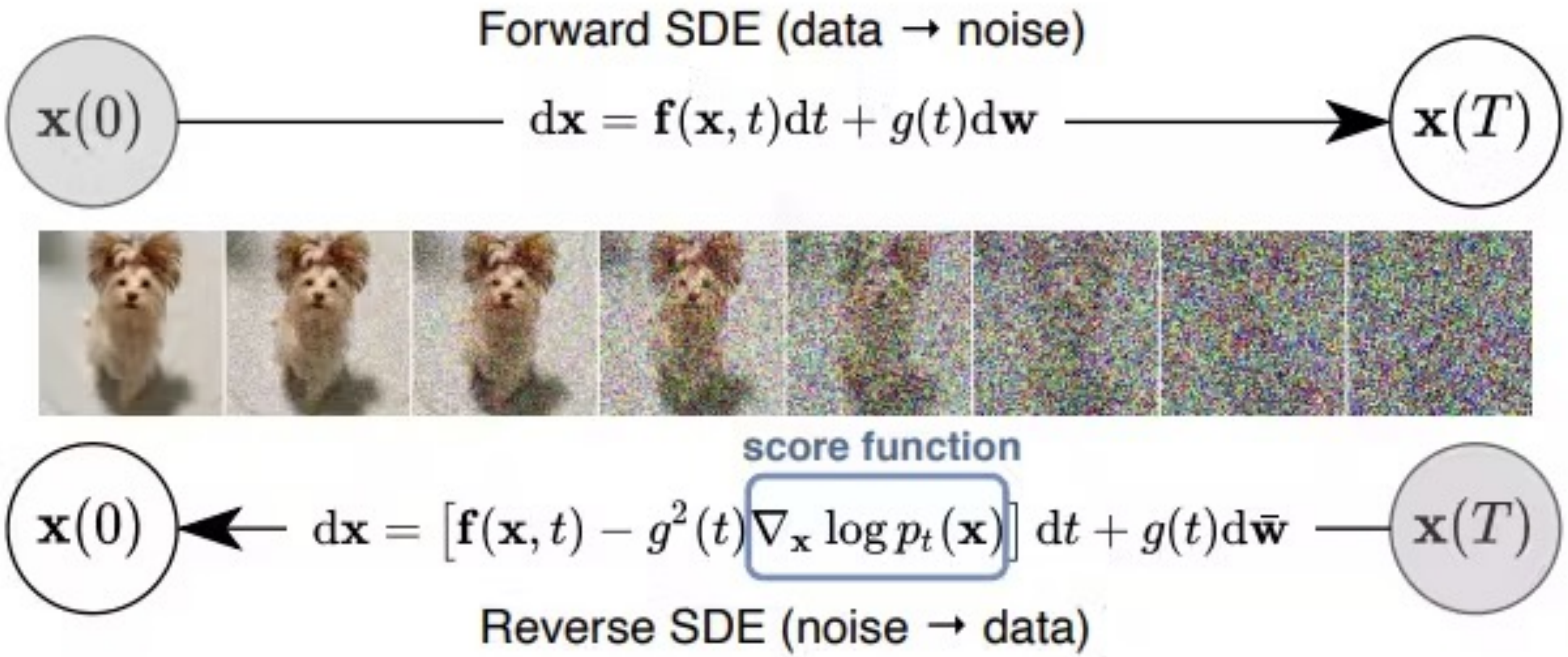
$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \underbrace{\mu_\theta(\mathbf{x}_t, t)}_{\text{Trainable network (U-net, Denoising Autoencoder)}}, \sigma_t^2\mathbf{I})$$

➡

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T)\prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$$

Diffusion for visual generation (2)

- Stochastic Differential Equations (Score SDEs)



Forward SDE (data → noise)

$$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}$$

score function

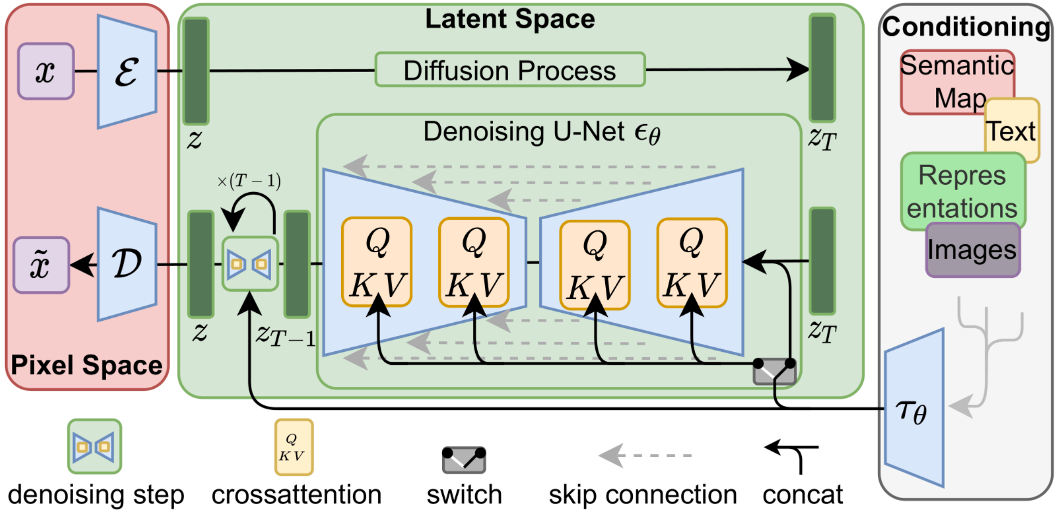
$$\mathrm{d}\mathbf{x} = [\mathbf{f}(\mathbf{x}, t) - g^2(t)\boxed{\nabla_{\mathbf{x}} \log p_t(\mathbf{x})}]\, \mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}$$

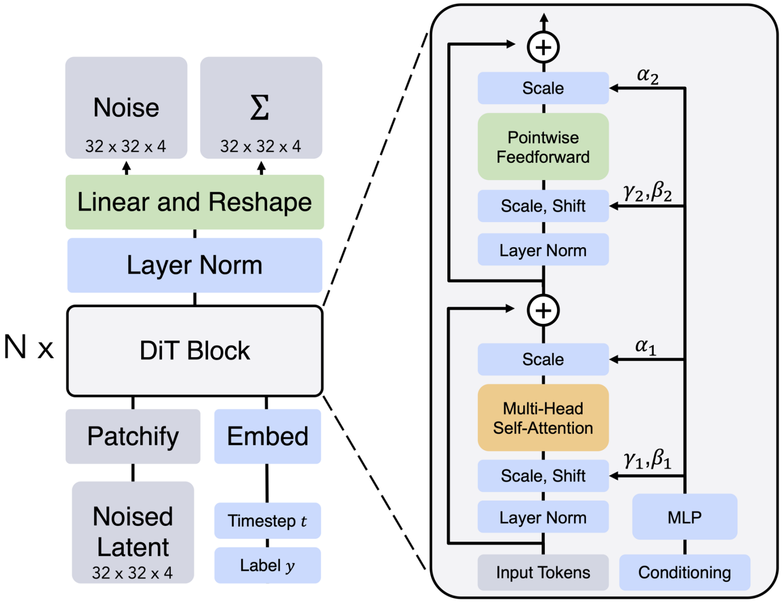
Reverse SDE (noise → data)

Key Elements of visual Diffusion Models

- Pixel diffusion (original input)
- Latent space diffusion

- Unet
- Transformer





Latent Diffusion Transformer

DiT Block with adaLN-Zero

Sora, breakthrough

- **Consistency**: consistency in 3D rendering, long-range coherence, and object permanence.
- **High fidelity**.
- **Surprising length**: extended video length capability (Sora: 1 minute vs. previous systems: seconds).
- **Flexible resolution**: generation of videos across various durations, aspect ratios, and resolutions.

Sora, key technologies

- The **DiT** framework by Meta (2022.12) is designed for video processing.
- Google's **MAGViT** (2022.12) focuses on Video Tokenization.
- Google DeepMind introduced **NaViT** (2023.07) to support various resolutions and aspect ratios.
- OpenAI's **DALL-E 3** (2023.09) enhances Video Caption generation for improved conditioned video creation.

Modeling the physical world

- We know that it is very complicated real physical model.

probabilistic

- bayesian inference;
- probabilistic graphical models.

deterministic

- mathematical equations;
- physics based simulation;
- control theory.

Modeling the physical world

- We know that it is very complicated real physical model.

probabilistic

- bayesian inference;
- probabilistic graphical models.

deterministic

- mathematical equations;
- physics based simulation;
- control theory.

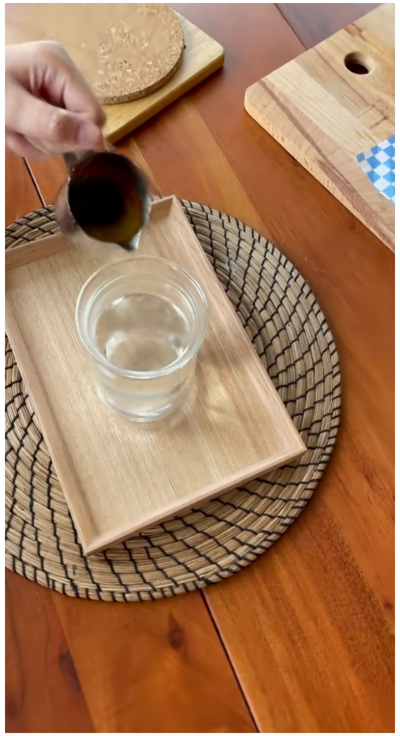
Key elements of a physical world

- Given a Sora demo (the walking woman in the Tokyo street), the key elements of a physical world, in the graphical way...

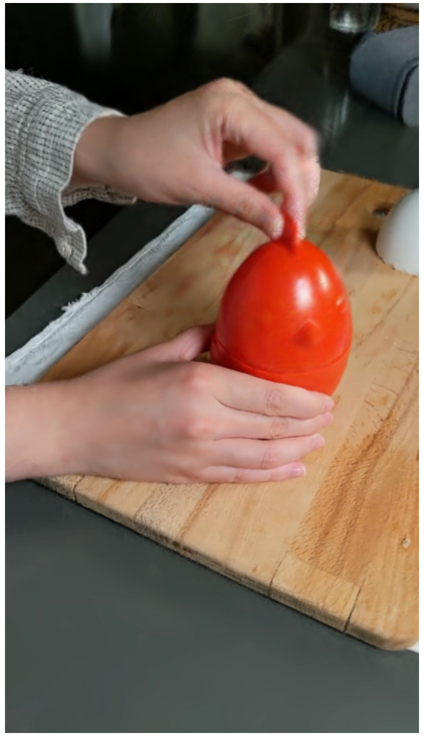
- Appearance
- Geometry
- Lighting
- Motion & Animation
- Audio

Modeling the physical world

- [CVPR] Gaussian-Flow: 4D Reconstruction with Dynamic 3D Gaussian Particle



Espresso



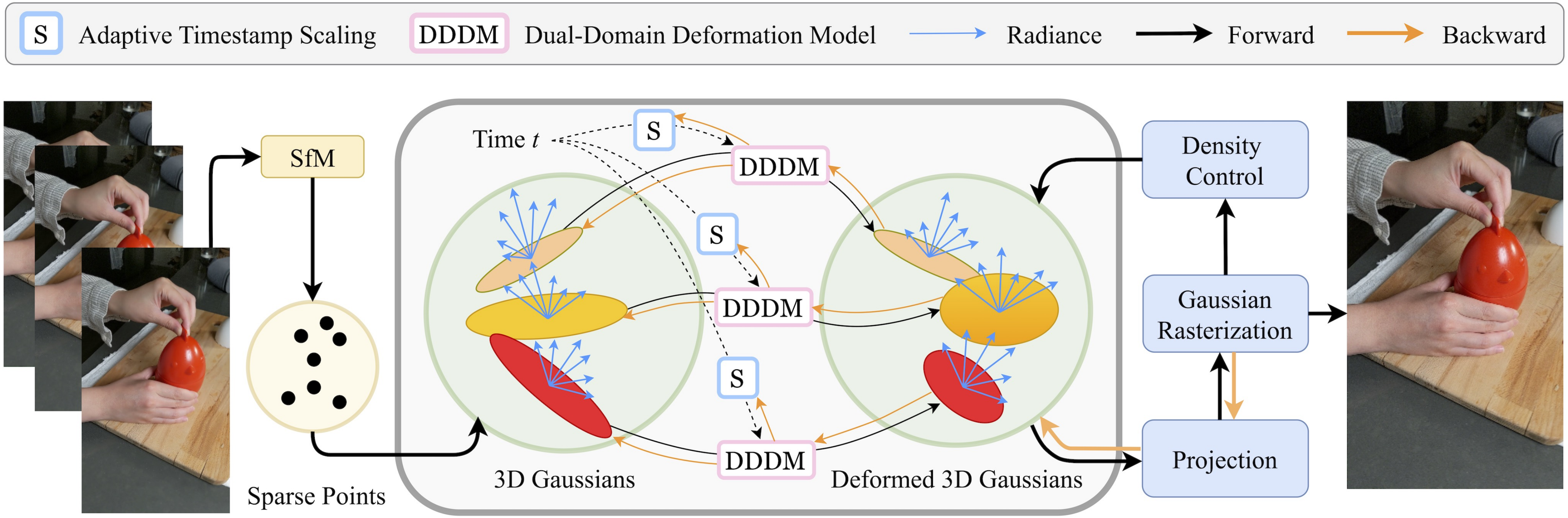
Chick-Chicken

Split-Cookie

Flame-Steak

Modeling the physical world

- [CVPR] Gaussian-Flow: 4D Reconstruction with Dynamic 3D Gaussian Particle



It is hard to model the physical world

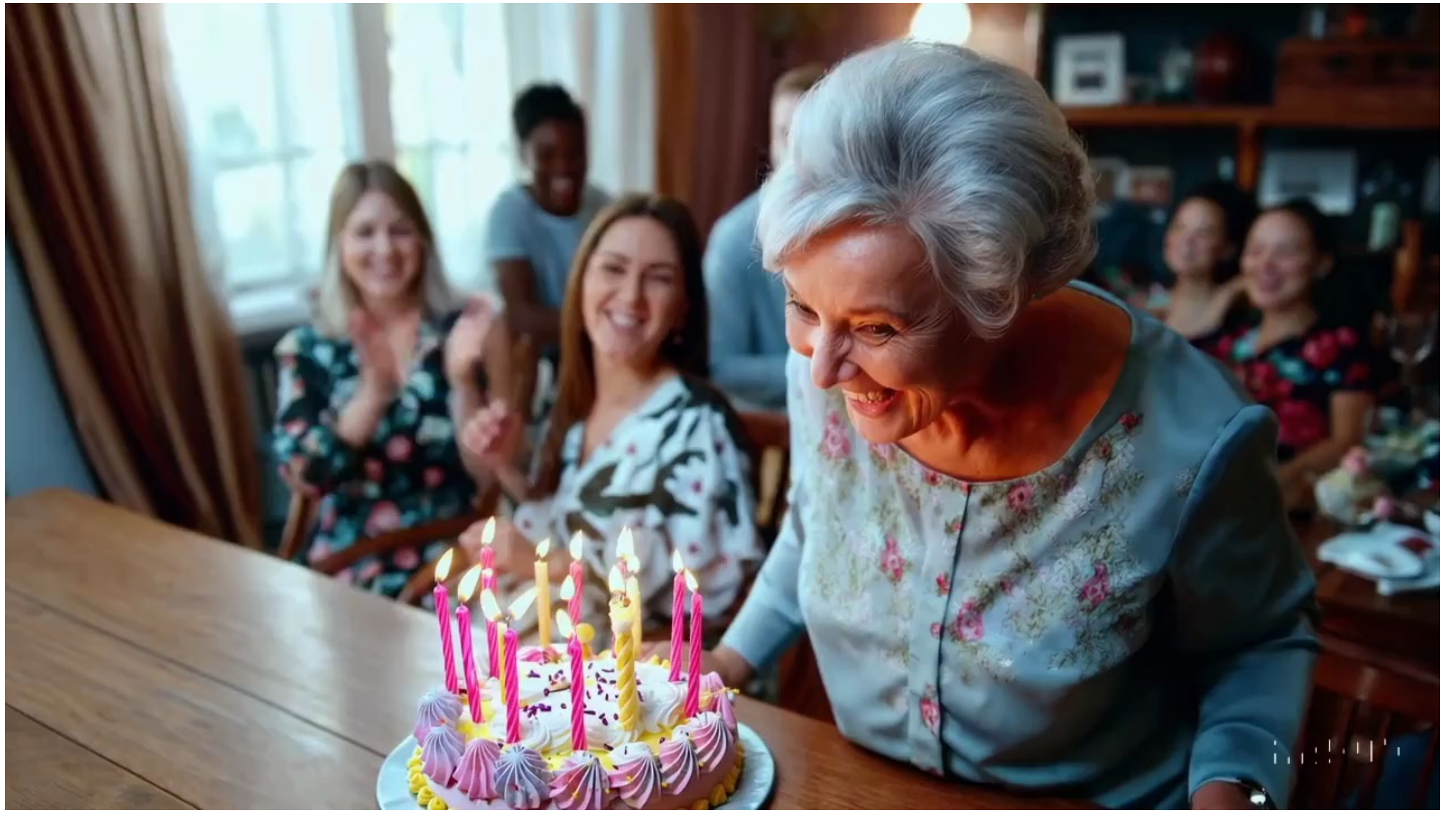
- In fact, the world is hard to model in a **probablistic** way.
- Sora resource consumption...
  - 1 billions of images;
  - 1 millions of hours of video data;
  - 10 trillions tokens after tokenizing images and videos
  - Training with ~5,000 A100s in parallel.

It is hard to model the physical world

- Sora failure case in geometry and appearance.

It is hard to model the physical world

- Sora failure case in lighting.

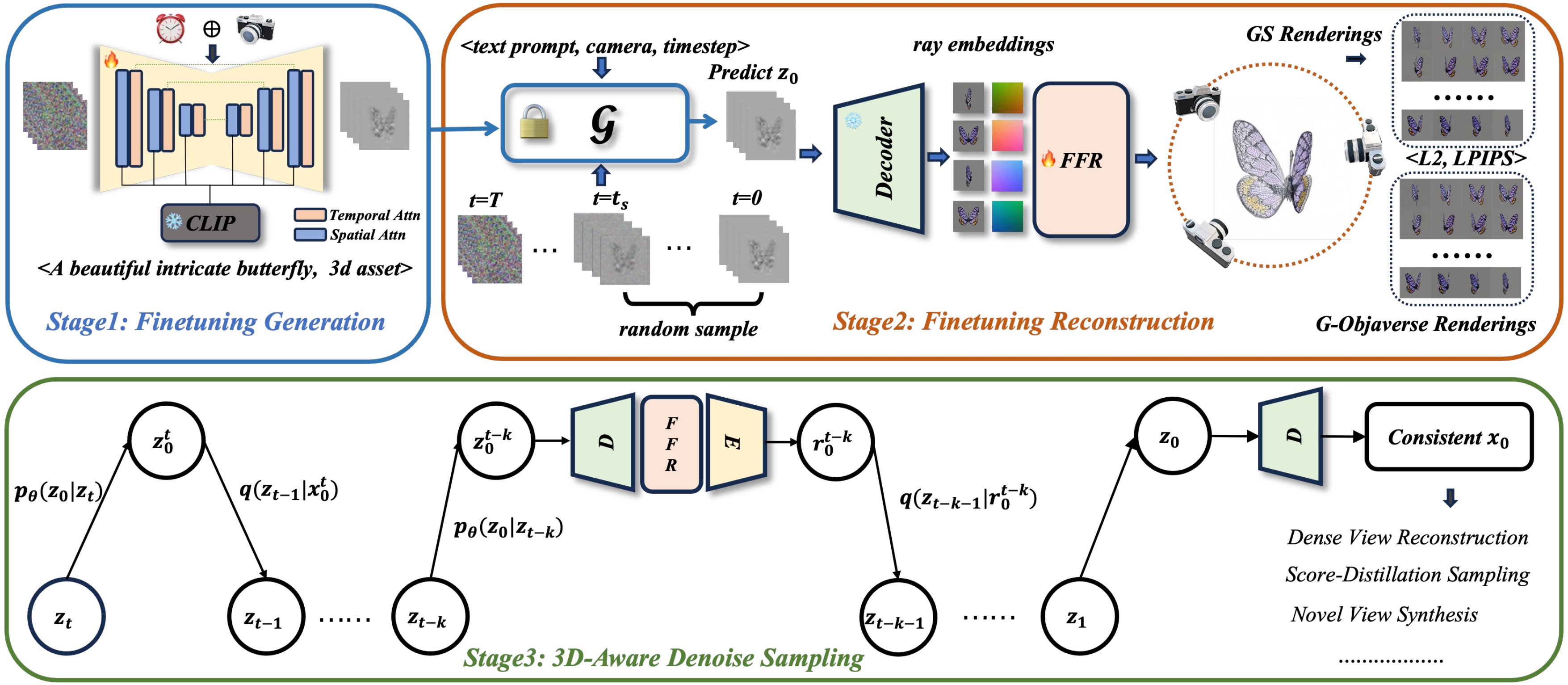


It is hard to model the physical world

- Sora failure case in motion and animation.

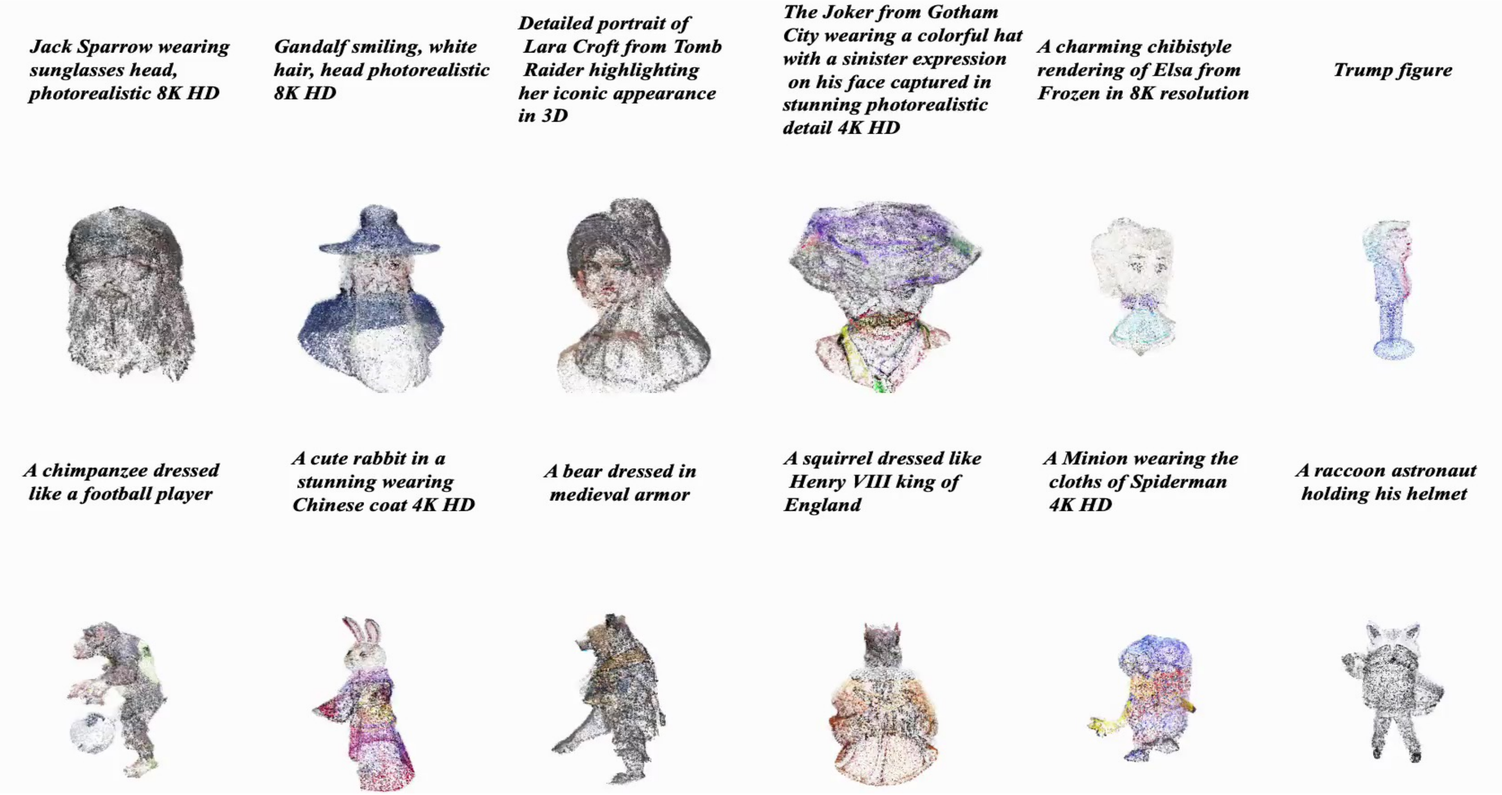
It is hard to model the physical world

- VideoMV: Consistent Multi-View Generation Based on Large Video Generative Model
- Geometric enhancement is still needed for multi-view images.



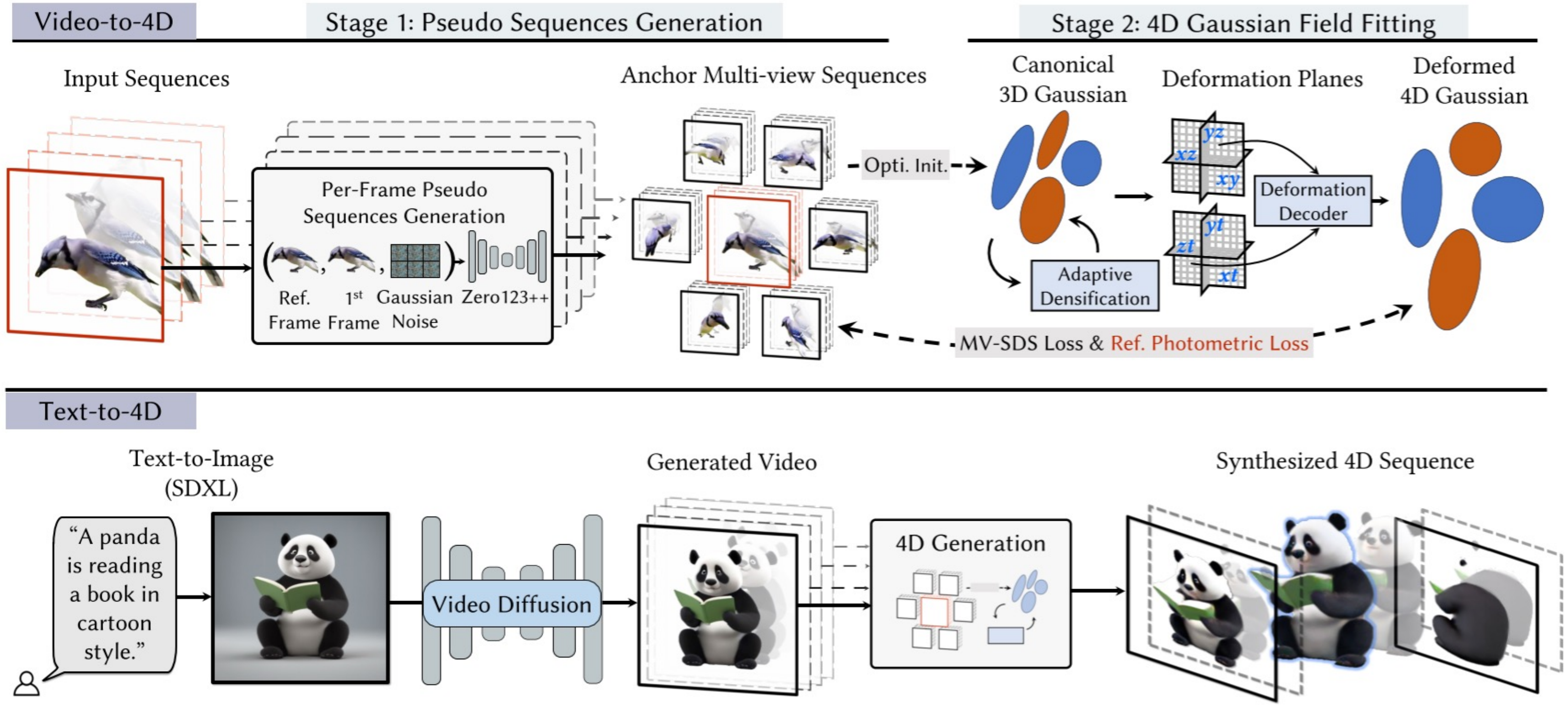
It is hard to model the physical world

- VideoMV: Consistent Multi-View Generation Based on Large Video Generative Model
- From a **static** aspects, SVD is able to model multi-view images.


Jack Sparrow wearing sunglasses head, photorealistic 8K HD
Gandalf smiling, white hair, head photorealistic 8K HD
Detailed portrait of Lara Croft from Tomb Raider highlighting her iconic appearance in 3D
The Joker from Gotham City wearing a colorful hat with a sinister expression on his face captured in stunning photorealistic detail 4K HD
A charming chibistyle rendering of Elsa from Frozen in 8K resolution
Trump figure
A chimpanzee dressed like a football player
A cute rabbit in a stunning wearing Chinese coat 4K HD
A bear dressed in medieval armor
A squirrel dressed like Henry VIII king of England
A Minion wearing the cloths of Spiderman 4K HD
A raccoon astronaut holding his helmet

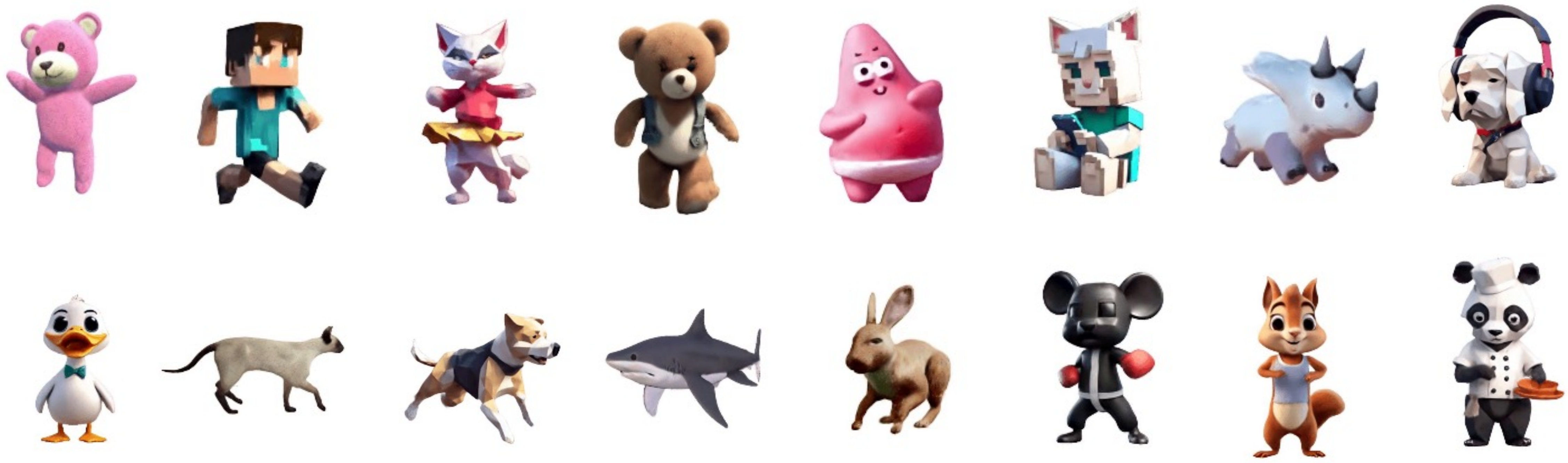
It is hard to model the physical world

- Stag4D: Spatial-Temporal Anchored Generative 4D Gaussians
- From a temporal aspects...



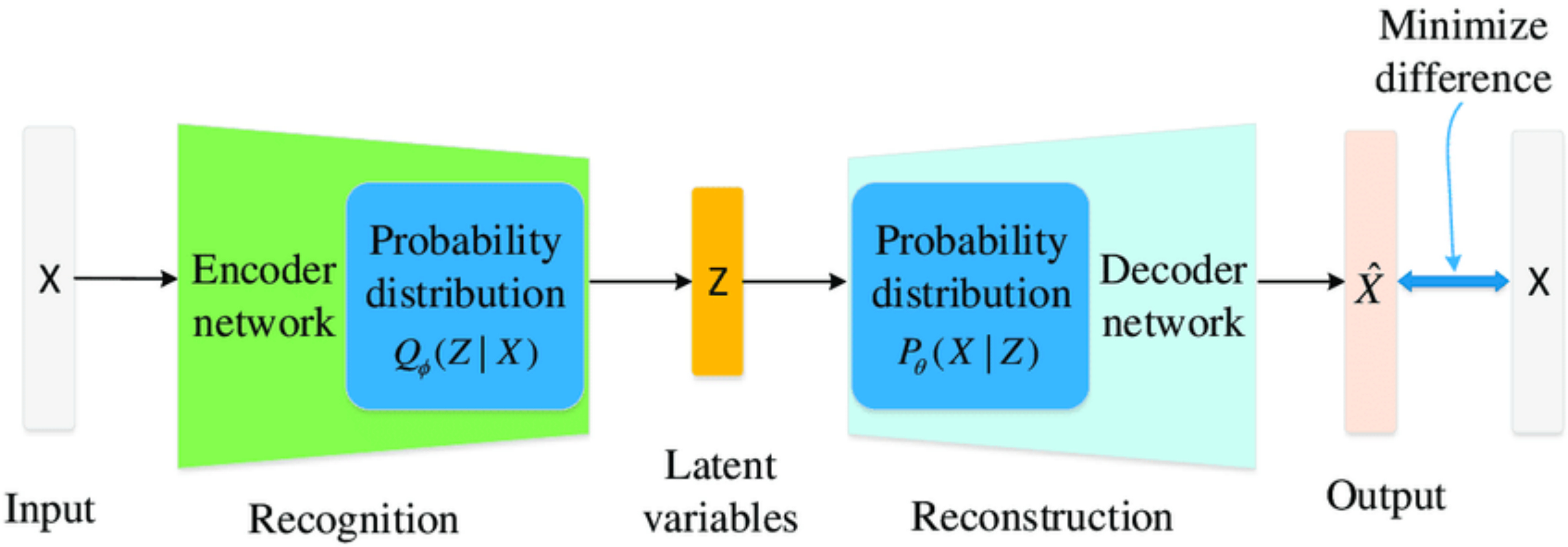
It is hard to model the physical world

- STAG4D: Spatial-Temporal Anchored Generative 4D Gaussians
- From a **temporal** aspects...



It is hard to model the physical world

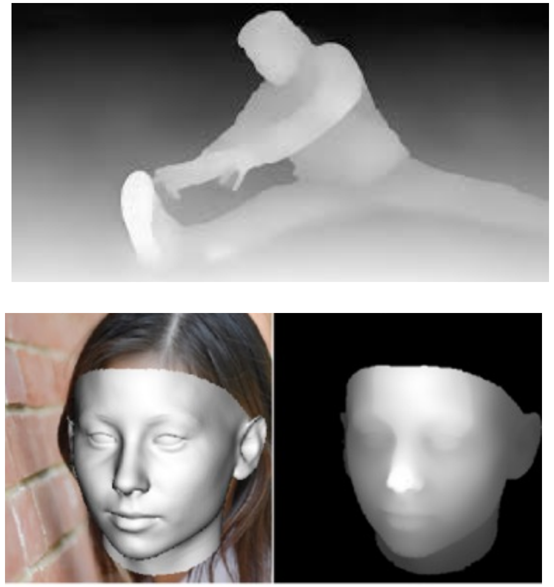
- Ilya Sutskever: compression is generalization.
- The best lossless compression for a dataset is the best generalization for data outside the dataset.



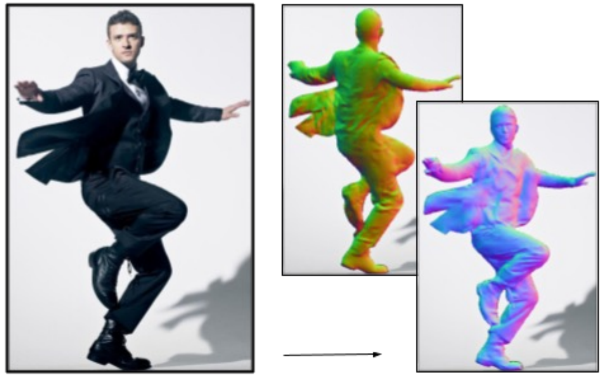
Apply the deterministic conditions

- Different representations of deterministic conditions in the physical world.
- Much less data and parameters!

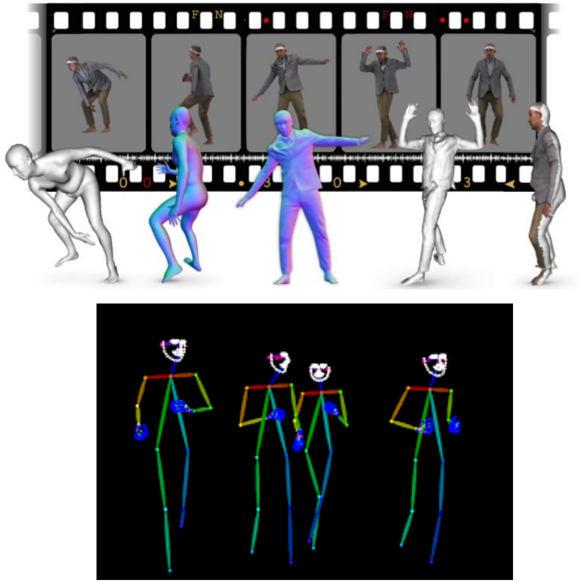
Geometry



Lighting



Motion & Animation



Apply the deterministic conditions

- There are two ways to inject deterministic information.

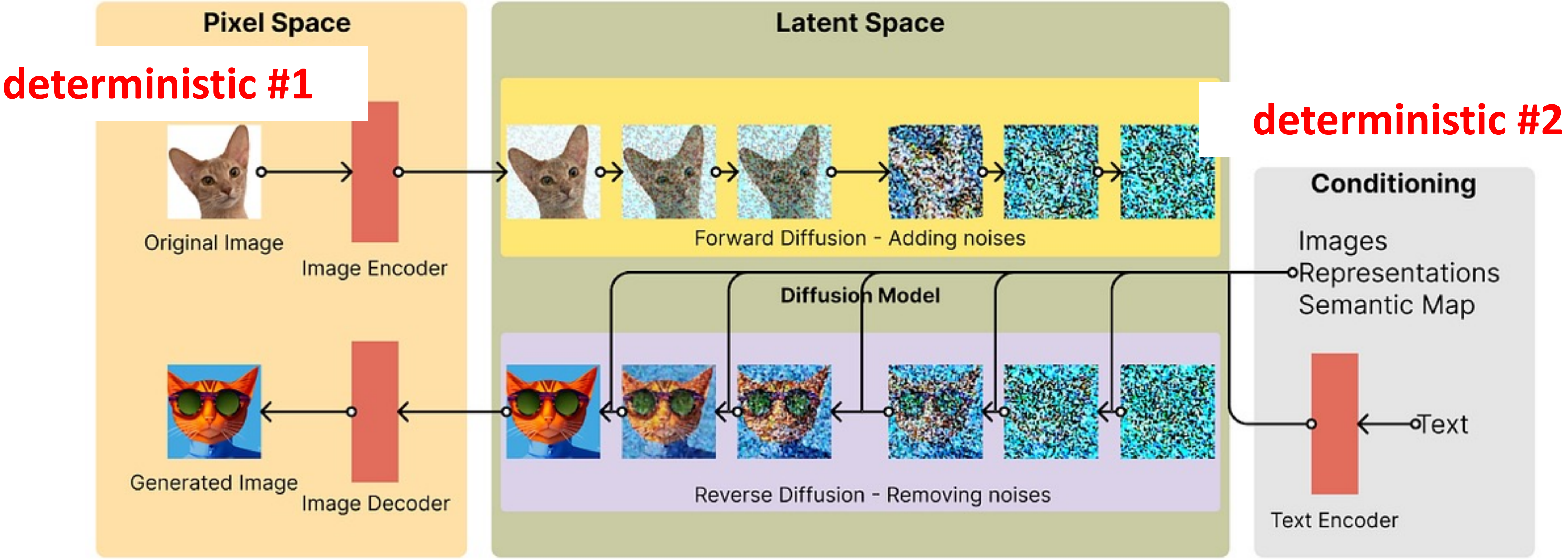
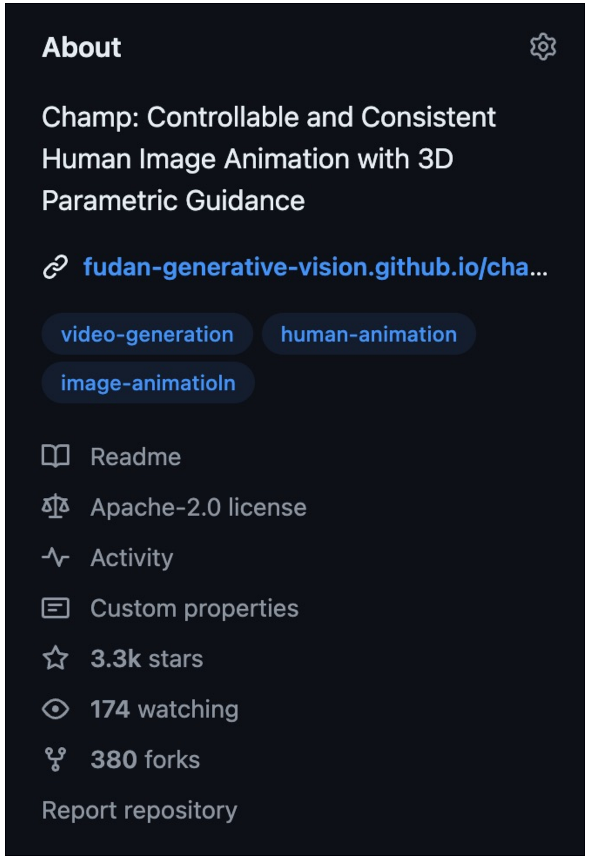


Image Human Animation

- Champ: Controllable and Consistent Human Image Animation with 3D Parametric Guidance



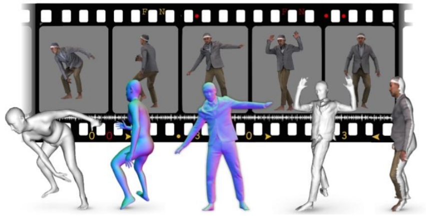
About

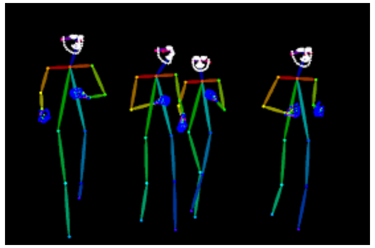
Champ: Controllable and Consistent Human Image Animation with 3D Parametric Guidance

fudan-generative-vision.github.io/cha...

Image Human Animation

- Champ: Controllable and Consistent Human Image Animation with 3D Parametric Guidance





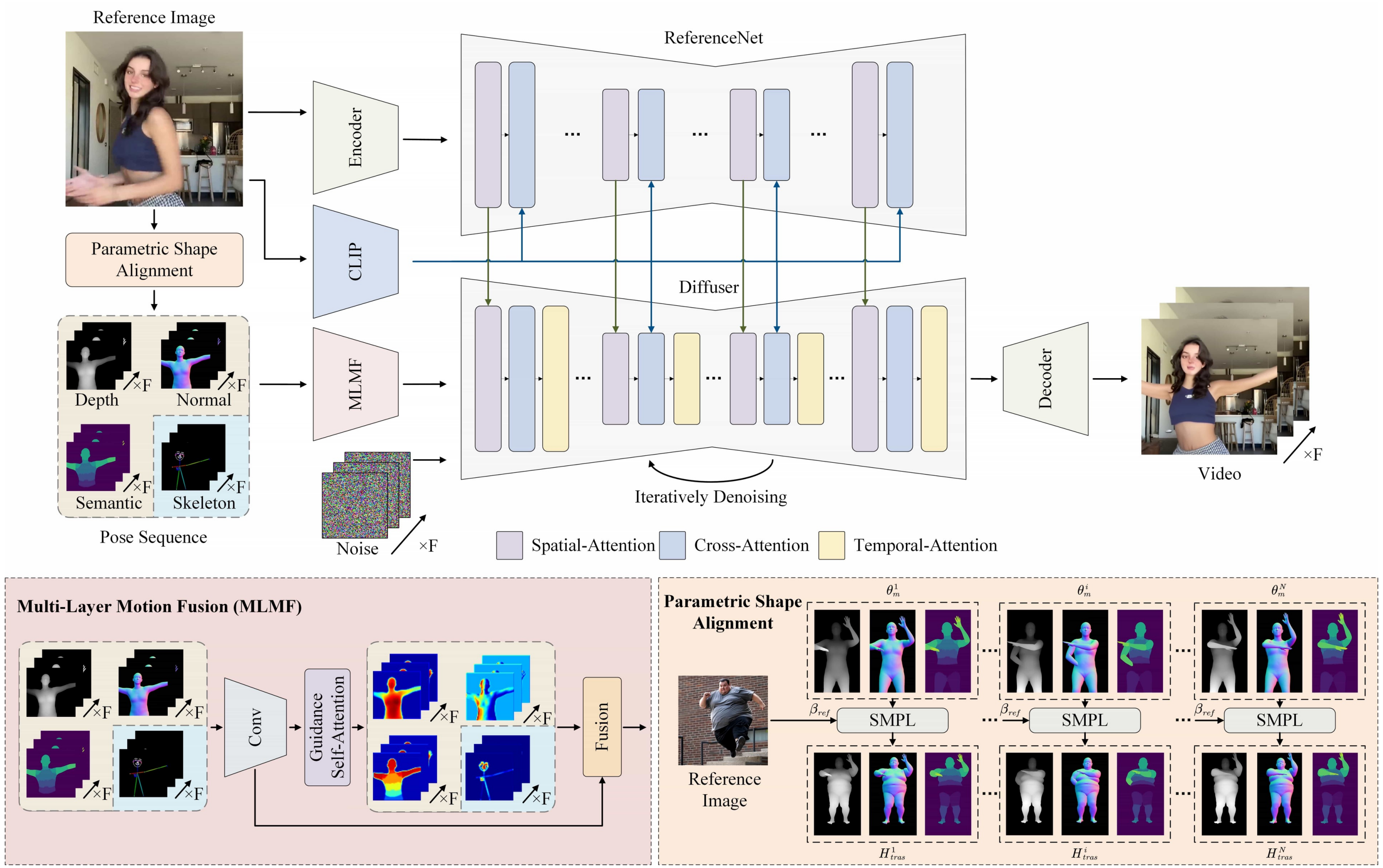
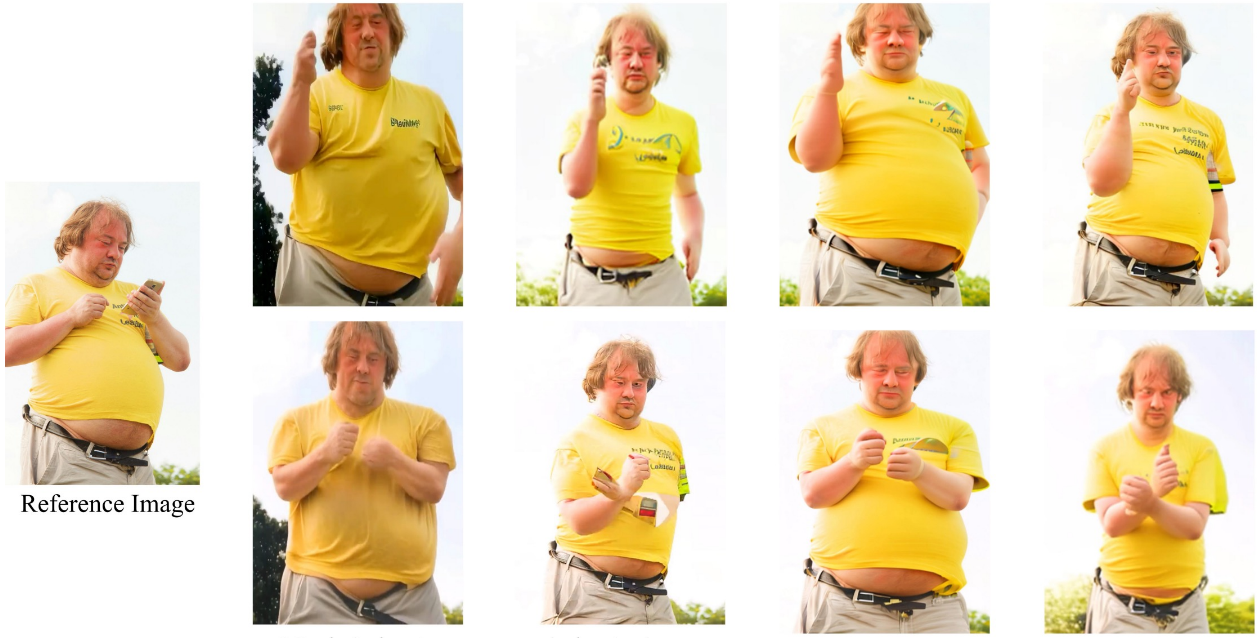


Image Human Animation

- Champ: Controllable and Consistent Human Image Animation with 3D Parametric Guidance



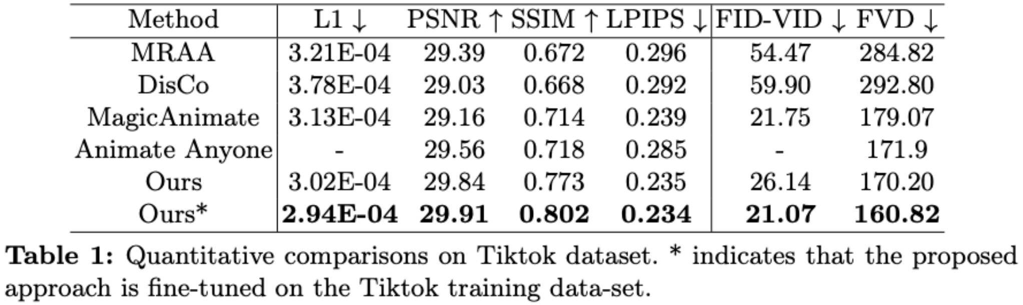
Reference Image

MagicAnimate

Animate Anyone

Ours with PST

Ours without PST

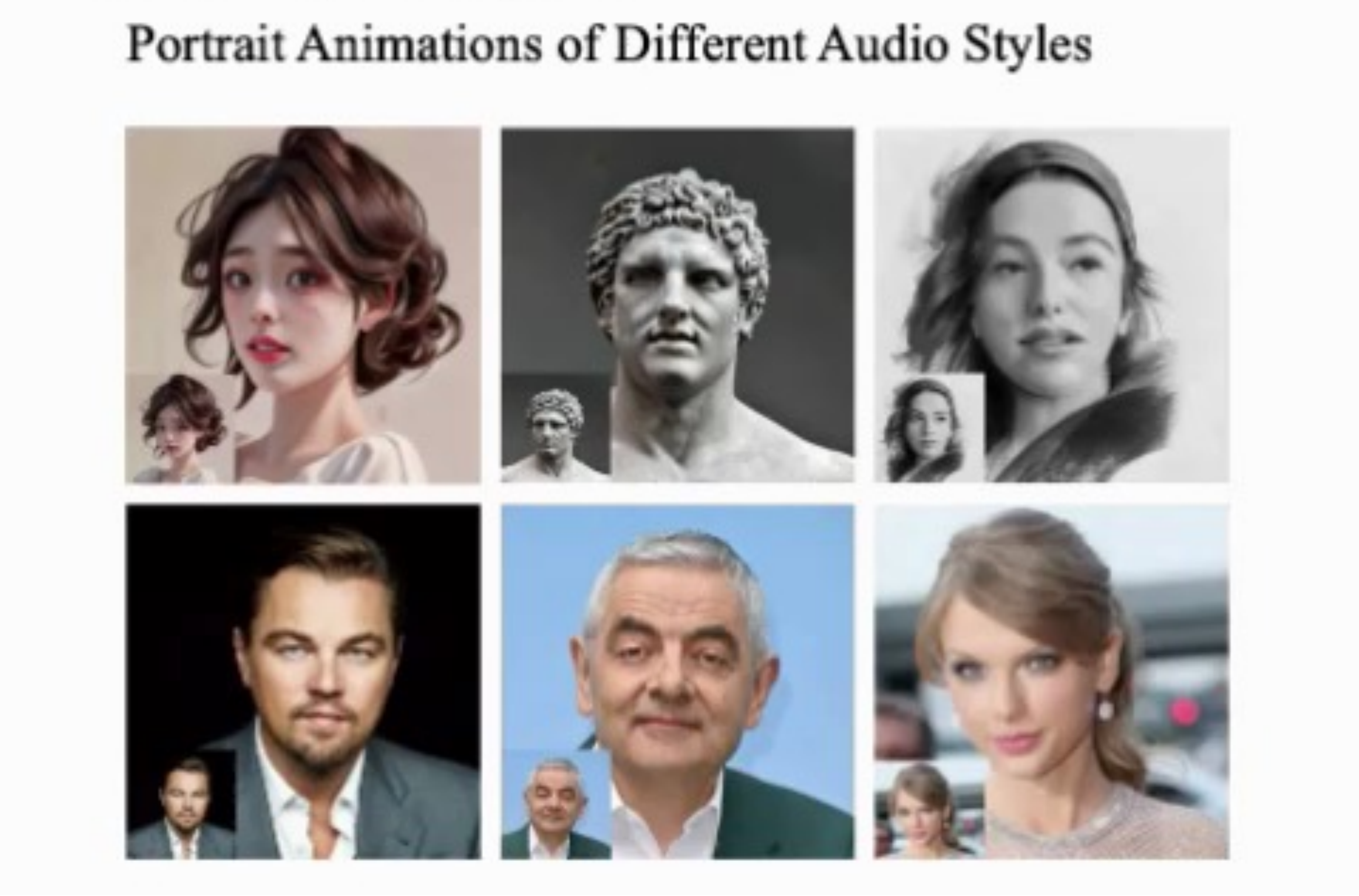


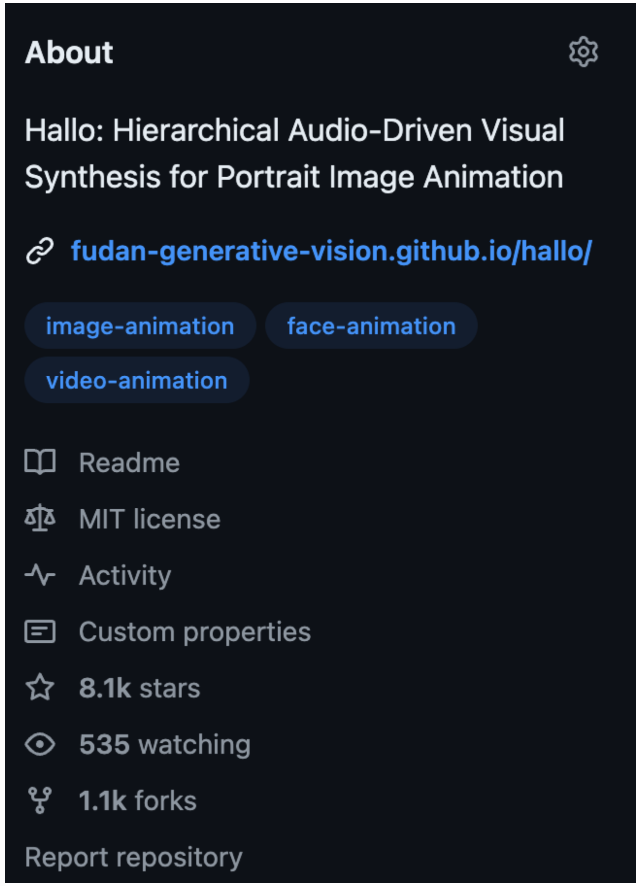
Method	L1 ↓	PSNR ↑	SSIM ↑	LPIPS ↓	FID-VID ↓	FVD ↓
MRAA	3.21E-04	29.39	0.672	0.296	54.47	284.82
DisCo	3.78E-04	29.03	0.668	0.292	59.90	292.80
MagicAnimate	3.13E-04	29.16	0.714	0.239	21.75	179.07
Animate Anyone	-	29.56	0.718	0.285	-	171.9
Ours	3.02E-04	29.84	0.773	0.235	26.14	170.20
Ours*	**2.94E-04**	**29.91**	**0.802**	**0.234**	**21.07**	**160.82**

Table 1: Quantitative comparisons on Tiktok dataset. * indicates that the proposed approach is fine-tuned on the Tiktok training data-set.

Image Portrait Animation

- Hallo: Hierarchical Audio-Driven Visual Synthesis for Portrait Image Animation




About

Hallo: Hierarchical Audio-Driven Visual Synthesis for Portrait Image Animation

fudan-generative-vision.github.io/hallo/

image-animation face-animation video-animation

- Readme
- MIT license
- Activity
- Custom properties
- 8.1k stars
- 535 watching
- 1.1k forks

Report repository

Image Portrait Animation

- Hallo: Hierarchical Audio-Driven Visual Synthesis for Portrait Image Animation

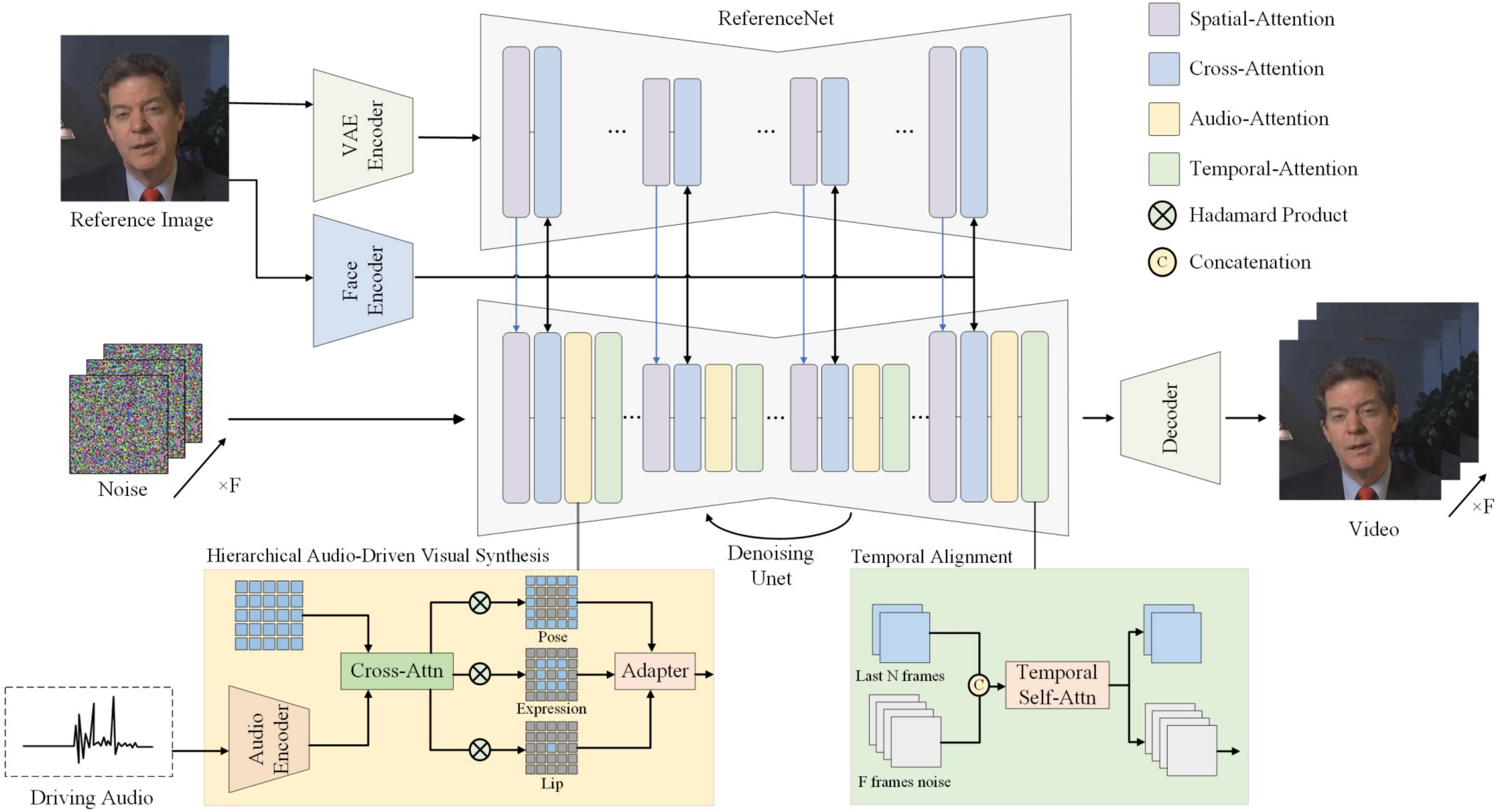
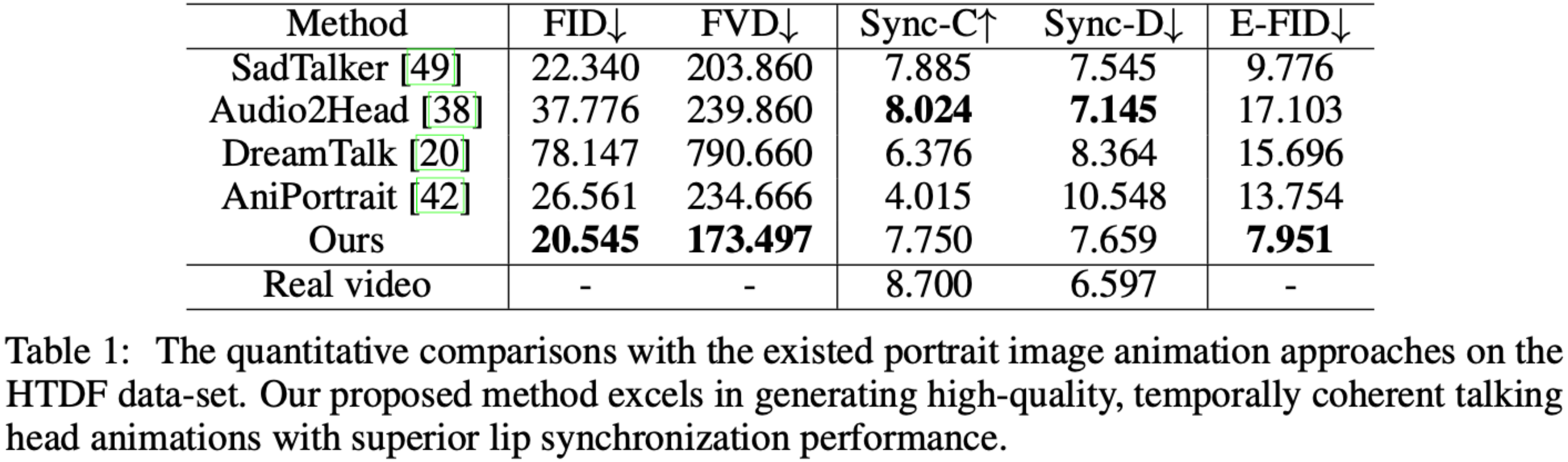


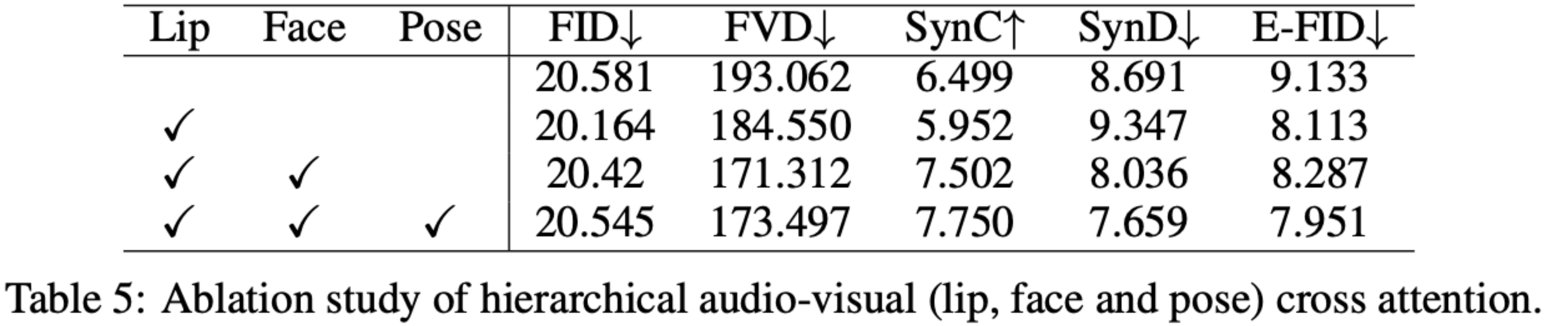
Image Portrait Animation

- Hallo: Hierarchical Audio-Driven Visual Synthesis for Portrait Image Animation



Method	FID↓	FVD↓	Sync-C↑	Sync-D↓	E-FID↓
SadTalker [49]	22.340	203.860	7.885	7.545	9.776
Audio2Head [38]	37.776	239.860	**8.024**	**7.145**	17.103
DreamTalk [20]	78.147	790.660	6.376	8.364	15.696
AniPortrait [42]	26.561	234.666	4.015	10.548	13.754
Ours	**20.545**	**173.497**	7.750	7.659	**7.951**
Real video	-	-	8.700	6.597	-

Table 1: The quantitative comparisons with the existed portrait image animation approaches on the HTDF data-set. Our proposed method excels in generating high-quality, temporally coherent talking head animations with superior lip synchronization performance.

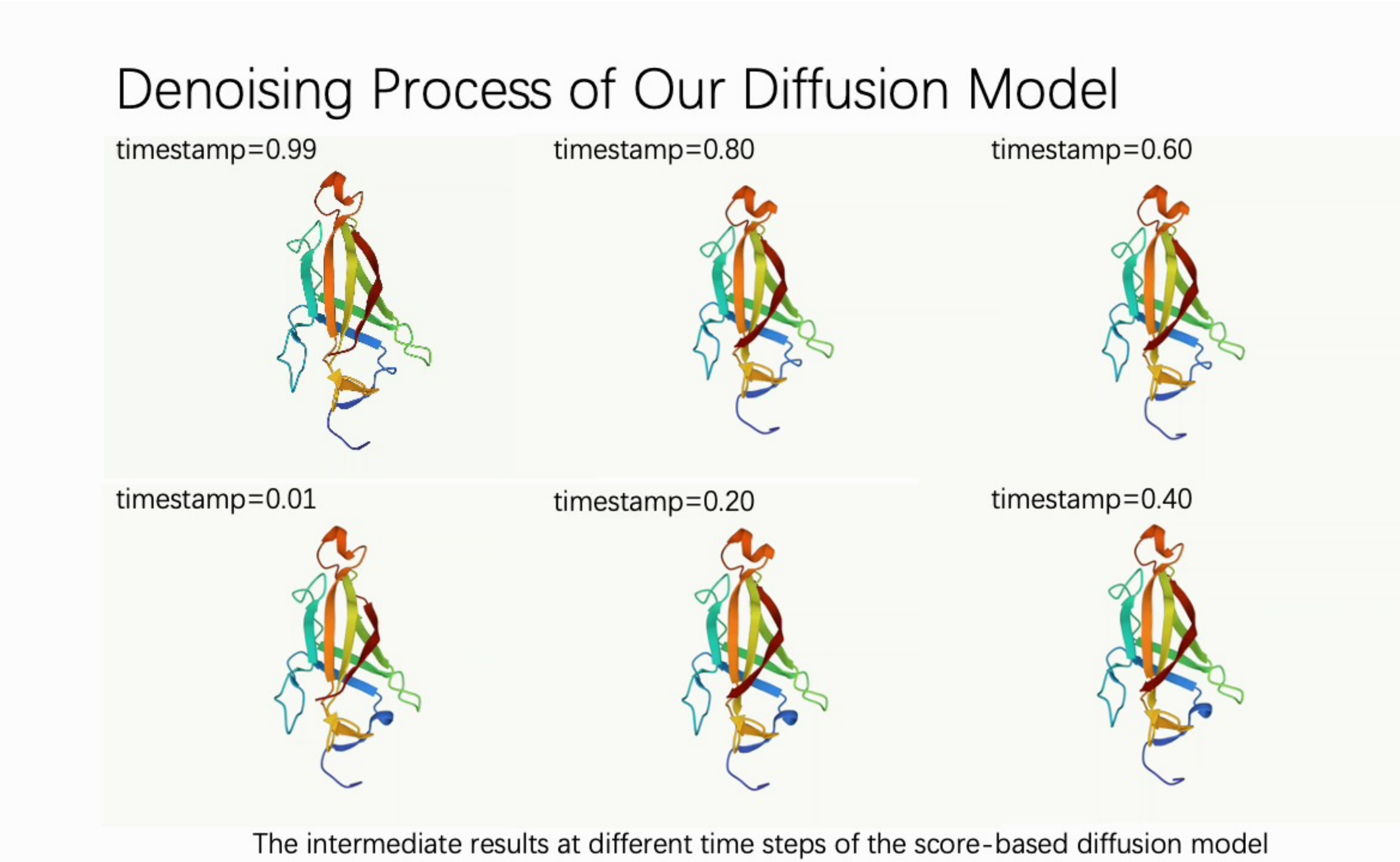


Lip	Face	Pose	FID↓	FVD↓	SynC↑	SynD↓	E-FID↓
			20.581	193.062	6.499	8.691	9.133
✓			20.164	184.550	5.952	9.347	8.113
✓	✓		20.42	171.312	7.502	8.036	8.287
✓	✓	✓	20.545	173.497	7.750	7.659	7.951

Table 5: Ablation study of hierarchical audio-visual (lip, face and pose) cross attention.

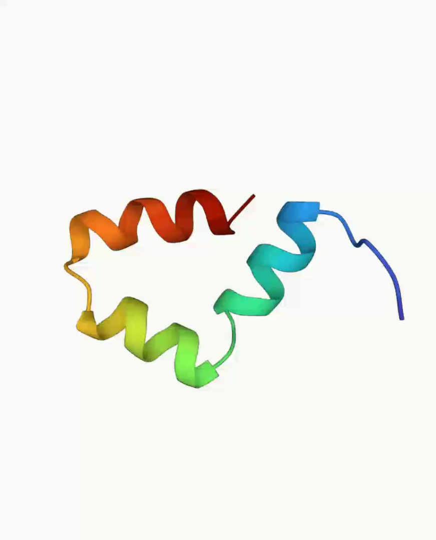
Dynamic Protein Structure Prediction

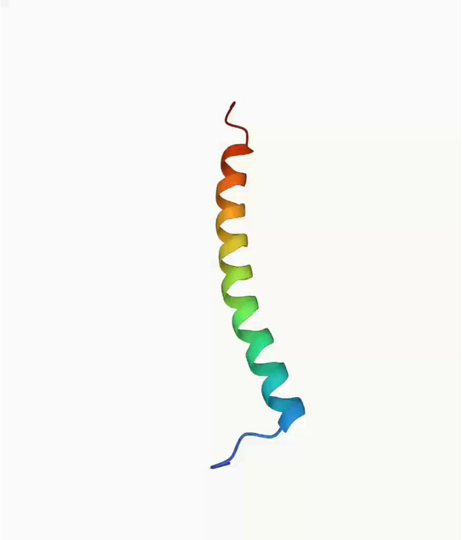
- 4D Diffusion for Dynamic Protein Structure Prediction with Reference Guided Temporal Alignment



Denoising Process of Our Diffusion Model

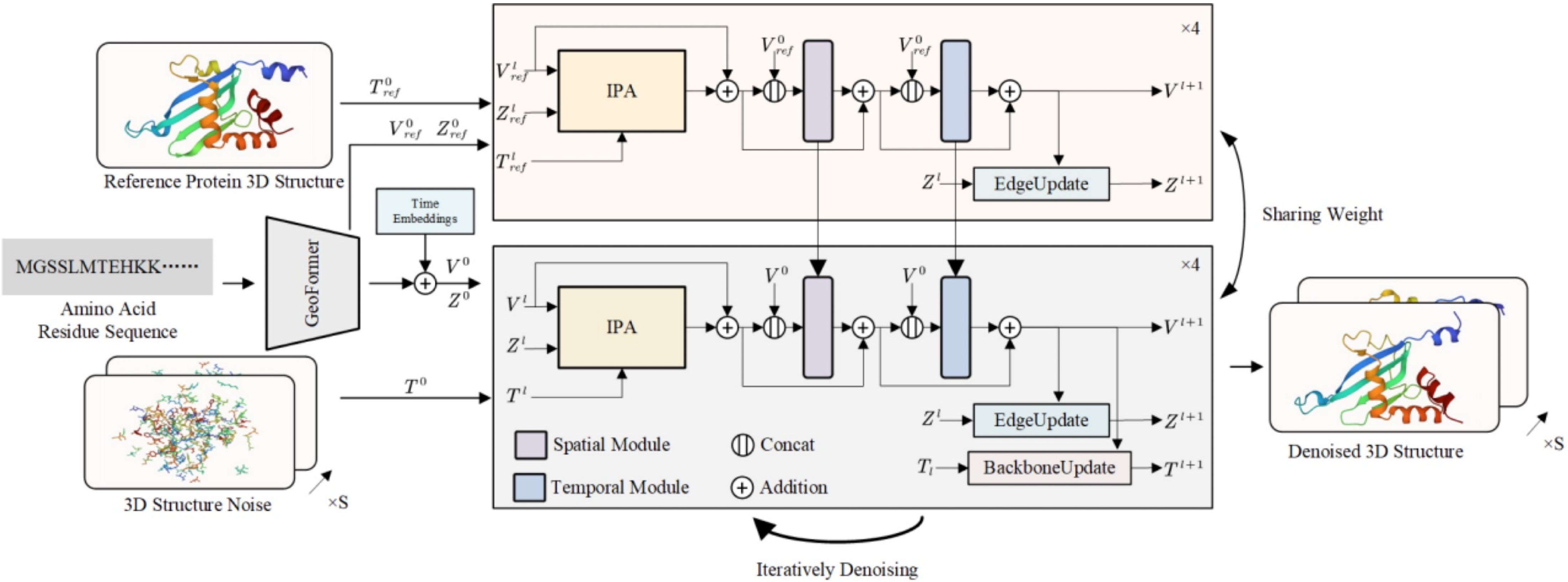
The intermediate results at different time steps of the score-based diffusion model





Dynamic Protein Structure Prediction

- 4D Diffusion for Dynamic Protein Structure Prediction with Reference Guided Temporal Alignment

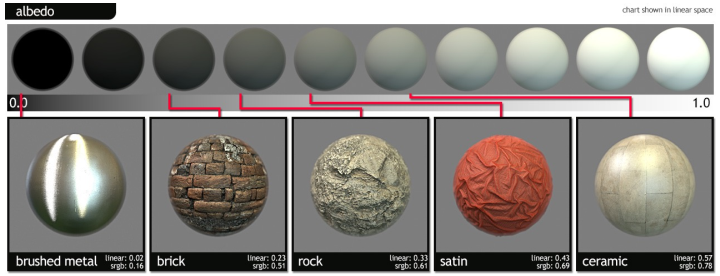


Future work

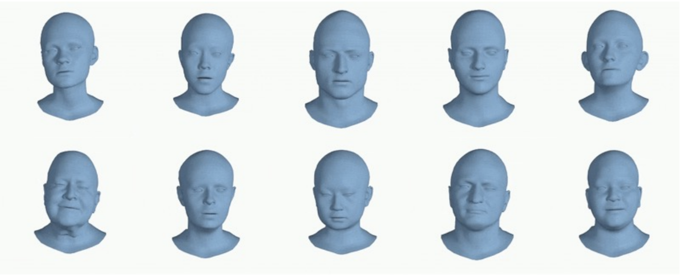
- Apply deterministic conditions to probabilistic diffusion.
- Less data and paramters!

Geometry

Lighting



Motion & Animation



科技生态圈峰会 + 深度研习

——1000 + 技术团队的共同选择

KEYLINK ing

K+峰会

K+峰会 上海站
K+全球软件研发行业创新峰会
时间：2024.06.21-22

K+峰会 敦煌站
K+思考周®研习社
时间：2024.10.17-19

K+峰会 香港站
K+思考周®研习社
时间：2024.11.10-12

K+峰会详情

AiDD峰会

AiDD峰会 上海站
AI+研发数字峰会
时间：2024.05.17-18

AiDD峰会 北京站
AI+研发数字峰会
时间：2024.08.16-17

AiDD峰会 深圳站
AI+研发数字峰会
时间：2024.11.08-09

AiDD峰会详情

AiDD

2024 AI+研发数字峰会

AI+ Development Digital summit

深圳站 11/08-09

AI驱动研发变革 促进企业降本增效

2024深圳站-议题设置

AI+产品线	LLM驱动产品创新	LLM驱动需求与业务分析	AI驱动设计与用户体验
AI+开发线	AI 原生应用开发框架与技术	AI Agents在研发落地实践	LLM驱动编程与单测
AI+测试线	LLM驱动测试分析与设计	基于LLM生成测试脚本与数据	LLM和AI应用的评测
AI+工程线	AI+DevOps 与工具 （LLM 时代的平台工程）	大模型对齐与安全	端侧大模型与云端协同
AI+领域线	领域大模型 SFT 与优化	知识增强与数据智能	大厂专场

扫描右侧二维码
查看更多会议详情

早鸟票限时抢购中 (截止到9月30日)

¥3680 早鸟票

¥2800 学生票

THANKS