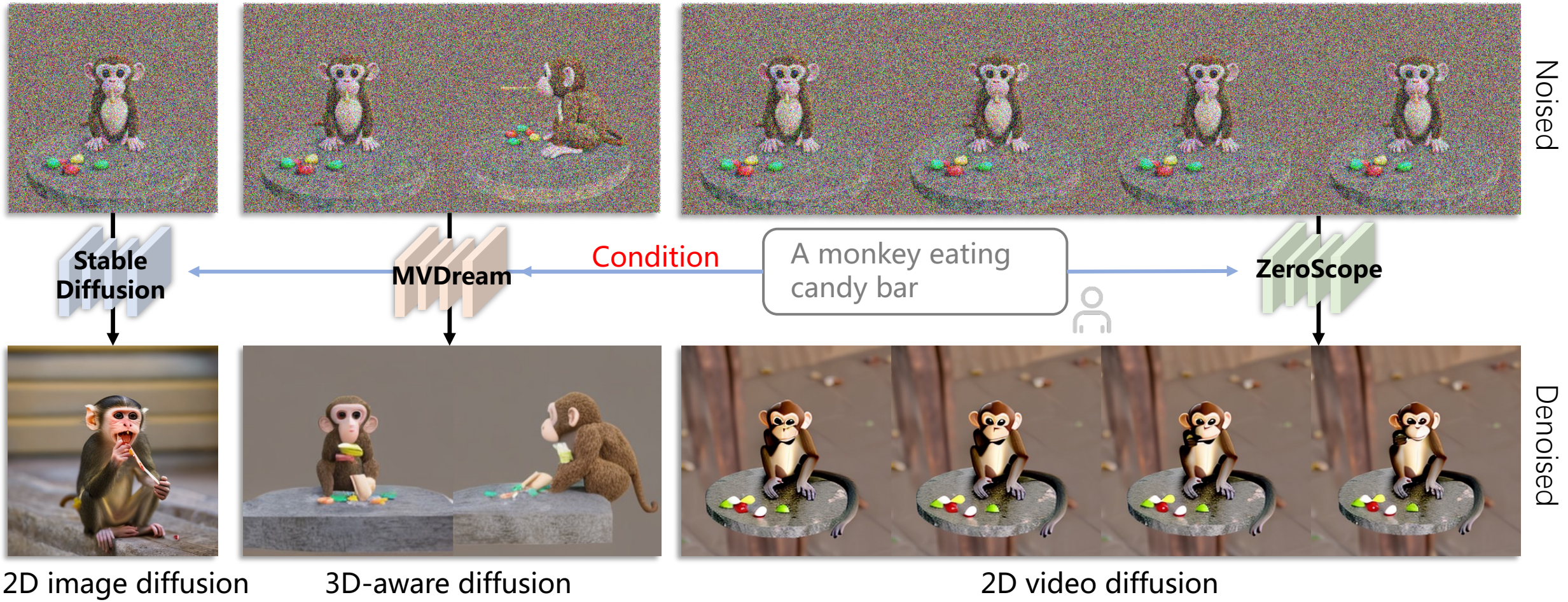
4Diffusion: Multi-view Video Diffusion Model for 4D Generation

Haiyu Zhang[1,2] Xinyuan Chen[2] Yaohui Wang[2] Xihui Liu[3] Yunhong Wang[1] Yu Qiao[2]

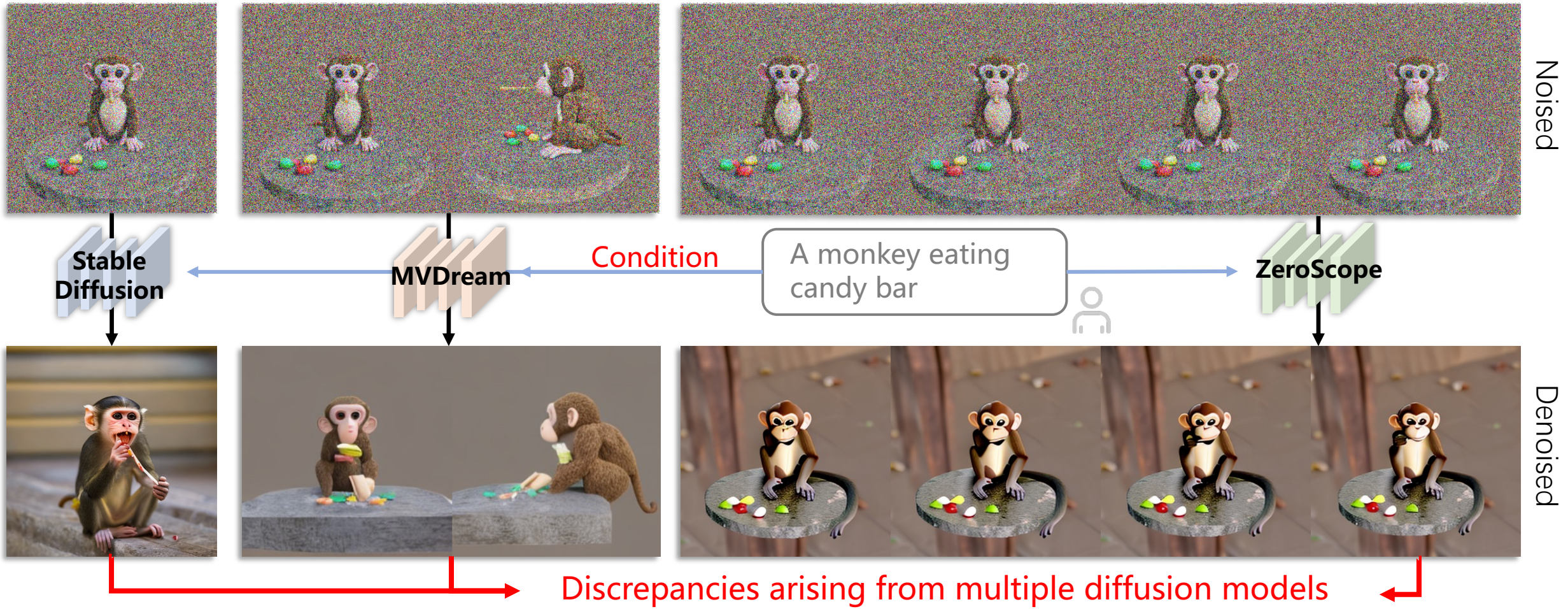
[1]Beihang University [2]Shanghai AI Laboratory [3]The University of Hong Kong

Motivation



- Most previous methods rely on hybrid diffusion models for 4D generation.

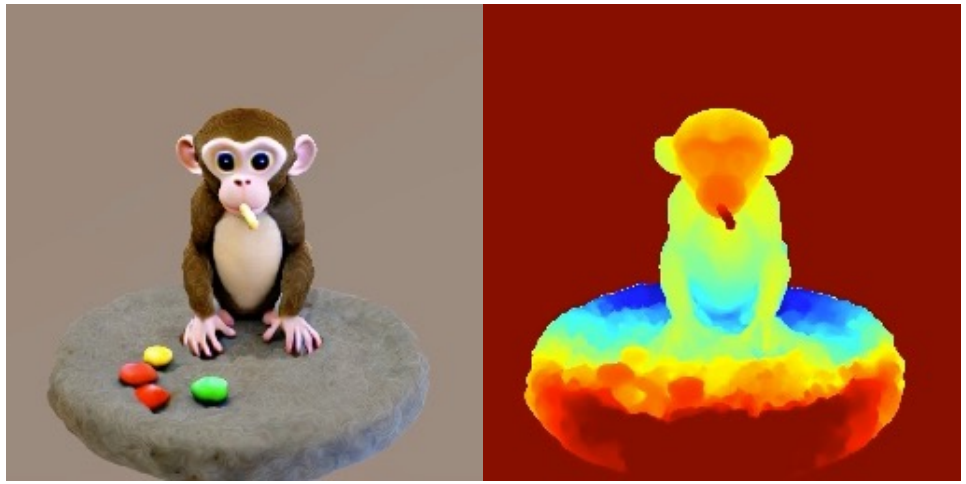
Motivation



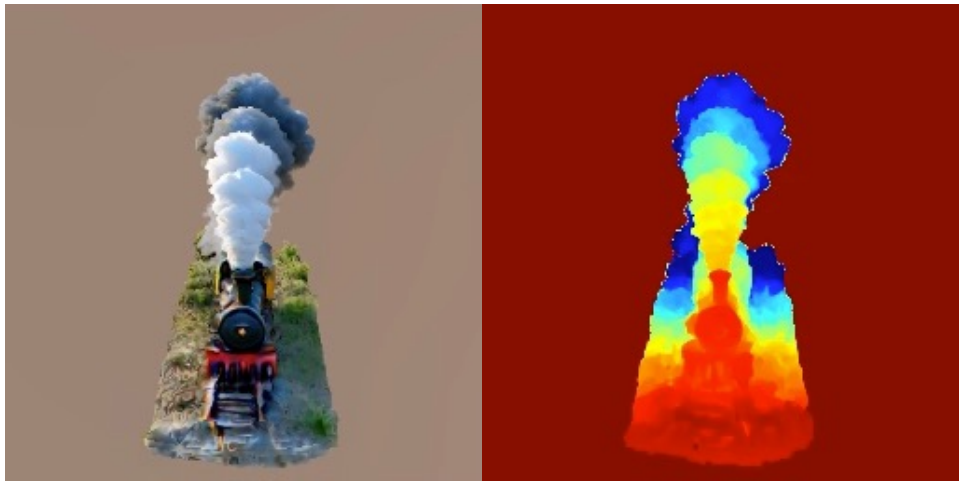
- Most previous methods rely on hybrid diffusion models for 4D generation.
- These diffusion models lack **multi-view spatial-temporal guidance** and exhibit **discrepancies**, making their integration challenging.

Motivation

Stage 1
3D Generation



A monkey eating a candy bar



A steam engine train is emitting steam into the air

Stage 2
4D Generation

Lavie

4D-fy(Zeroscope)

Lavie

4D-fy(Zeroscope)

- They may result in inconsistent temporal appearance and flickers.

[1] Wang Y, et al. Lavie: High-quality video generation with cascaded latent diffusion models. 2023.
[2]Bahmani S, et al. 4d-fy: Text-to-4d generation using hybrid score distillation sampling. CVPR, 2024.

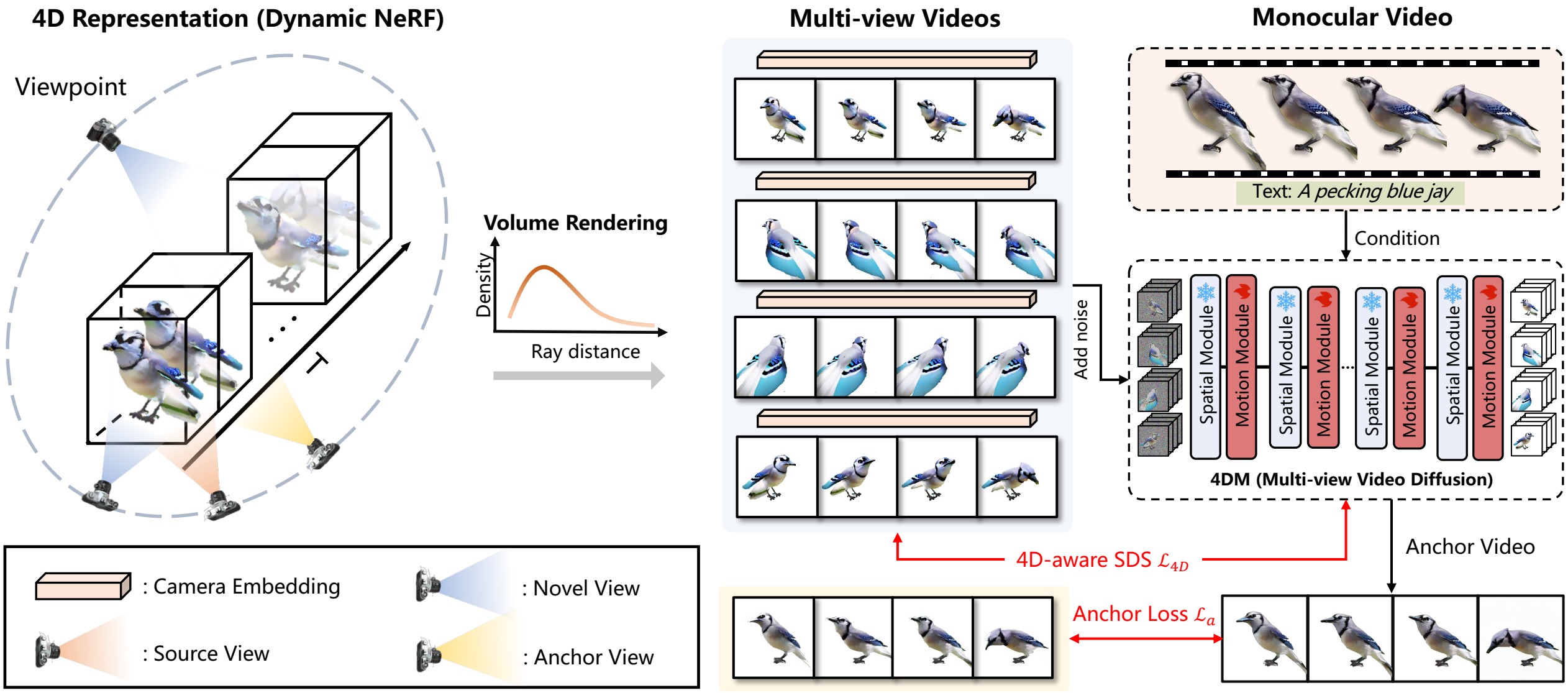
Motivation

How to generate high-quality **spatial-temporally consistent** 4D contents in a unified manner?

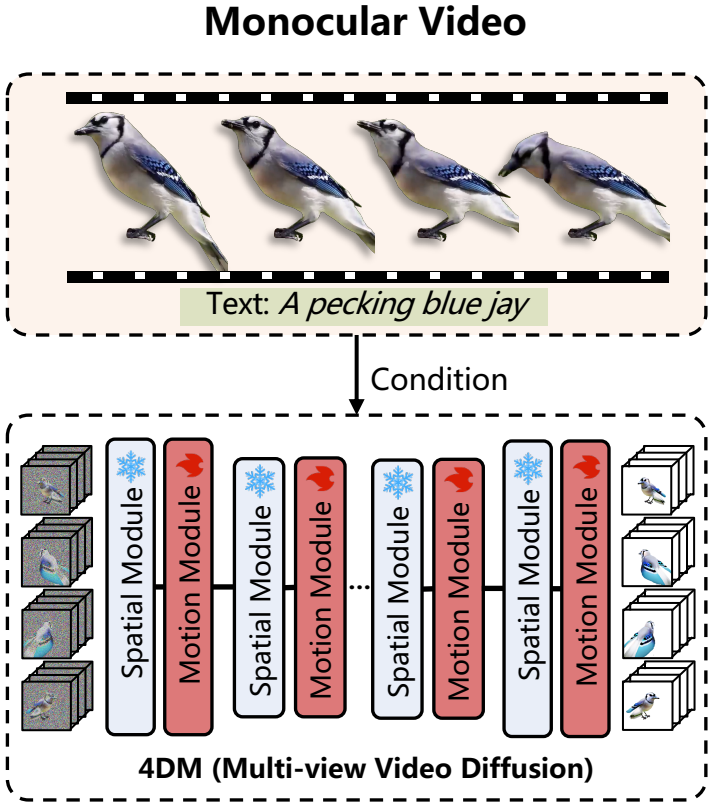
➢ They may result in inconsistent temporal appearance and flickers.

[1] Wang Y, et al. Lavie: High-quality video generation with cascaded latent diffusion models. 2023.
[2]Bahmani S, et al. 4d-fy: Text-to-4d generation using hybrid score distillation sampling. CVPR. 2024.

4Diffusion Overview

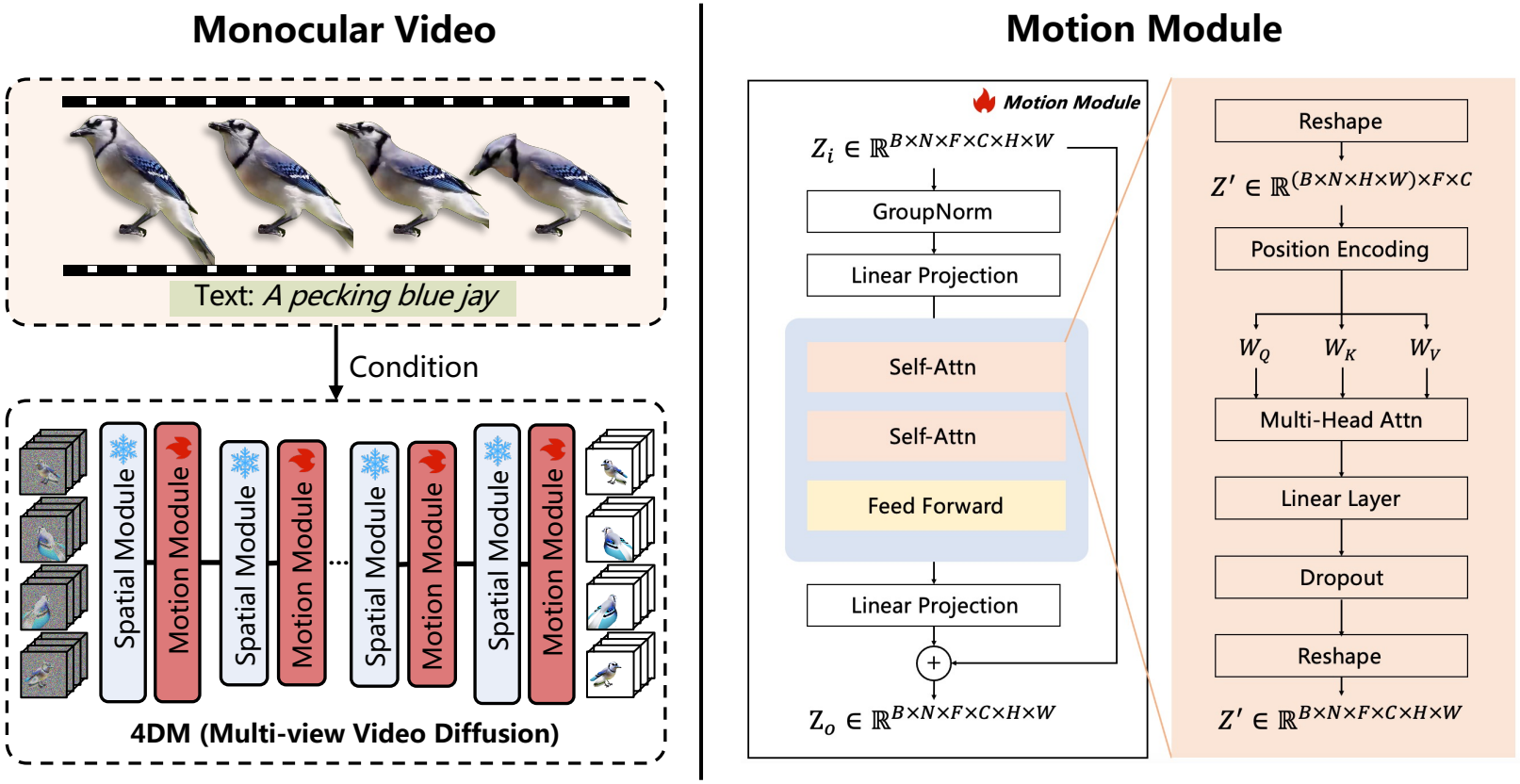


4DM (Multi-view Video Diffusion)



- We design 4DM based on a **frozen** pre-trained 3D-aware diffusion model (ImageDream), which has learned spatial relationships.

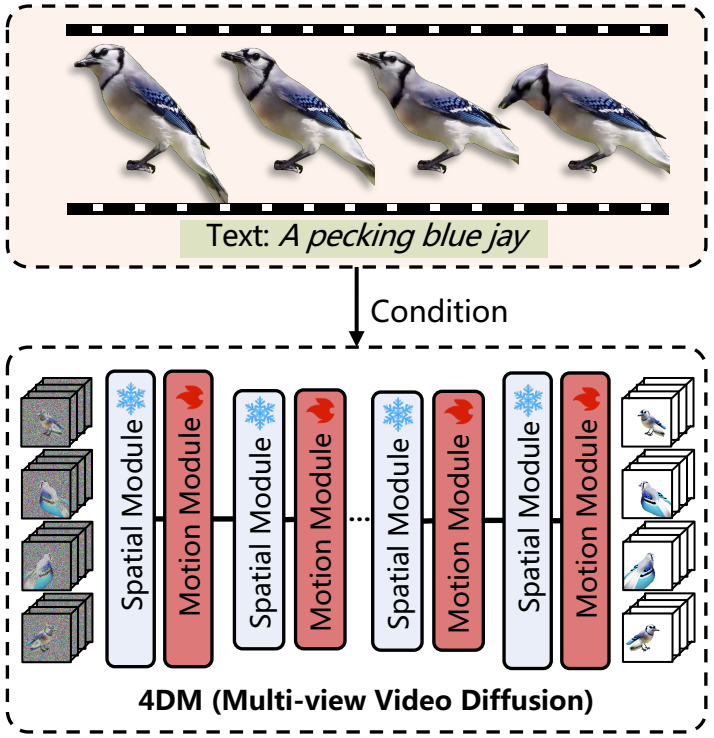
4DM (Multi-view Video Diffusion)



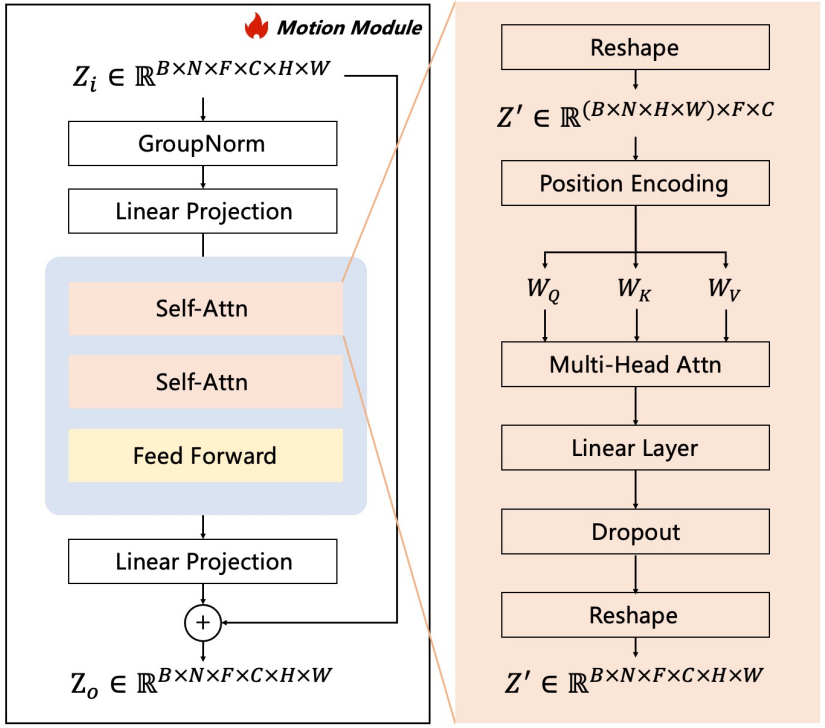
- We design 4DM based on a **frozen** pre-trained 3D-aware diffusion model (ImageDream), which has learned spatial relationships.
- We seamlessly add a **zero-initialized motion module** at the end of each block of the UViT network of ImageDream for **temporal modeling.**

4DM (Multi-view Video Diffusion)

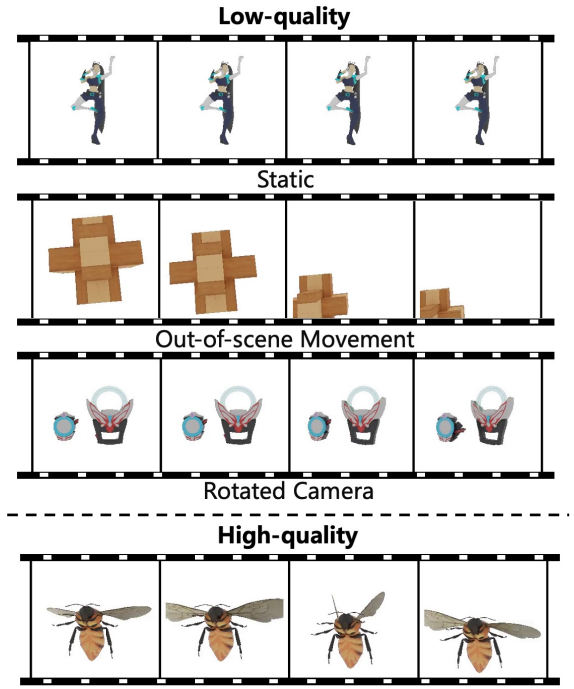
Monocular Video



Motion Module

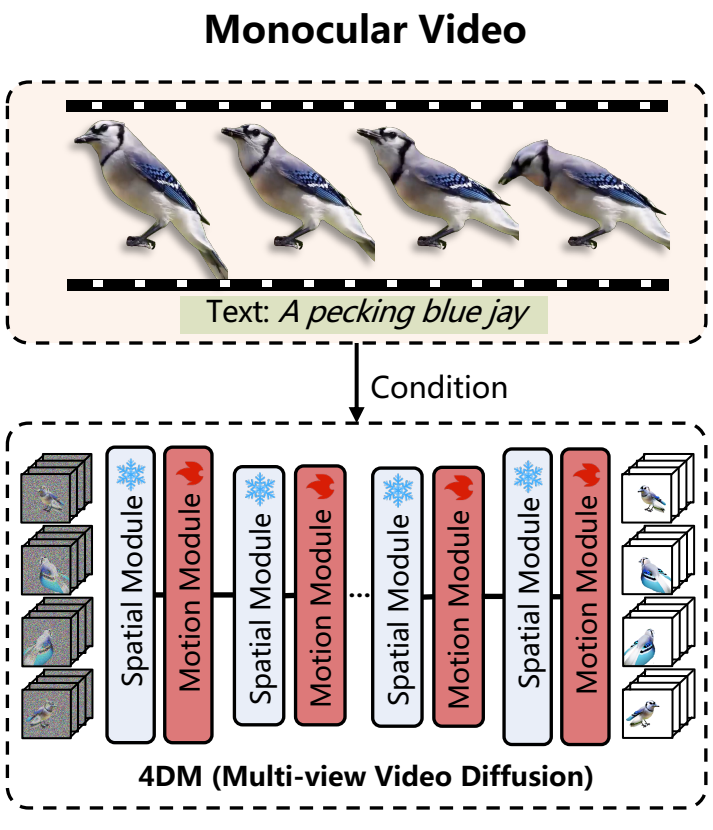


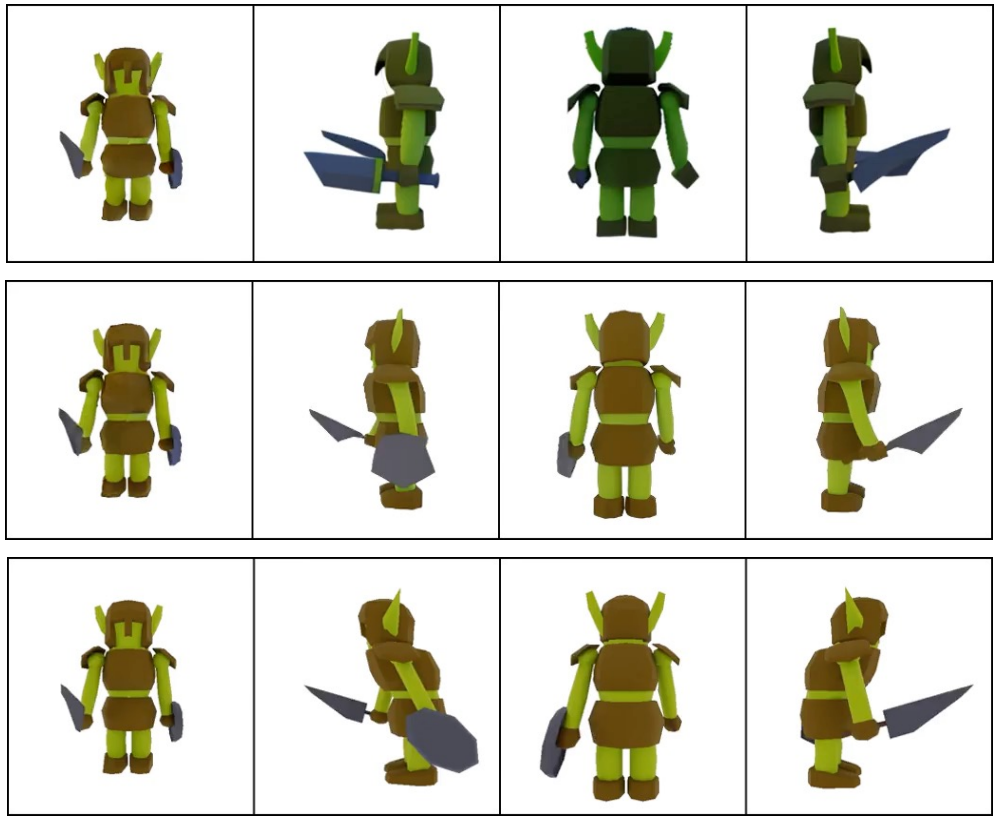
Training Data



- We design 4DM based on a **frozen** pre-trained 3D-aware diffusion model (ImageDream), which has learned spatial relationships.
- We seamlessly add a **zero-initialized motion module** at the end of each block of the UViT network of ImageDream for **temporal modeling.**
- We manually select a curated subset of 926 high-quality animated 3D shapes from Objaverse dataset and render multi-view videos to tune motion modules while holding the parameters of the origin ImageDream frozen.

4DM (Multi-view Video Diffusion)





ImageDream

Temporal inconsistent

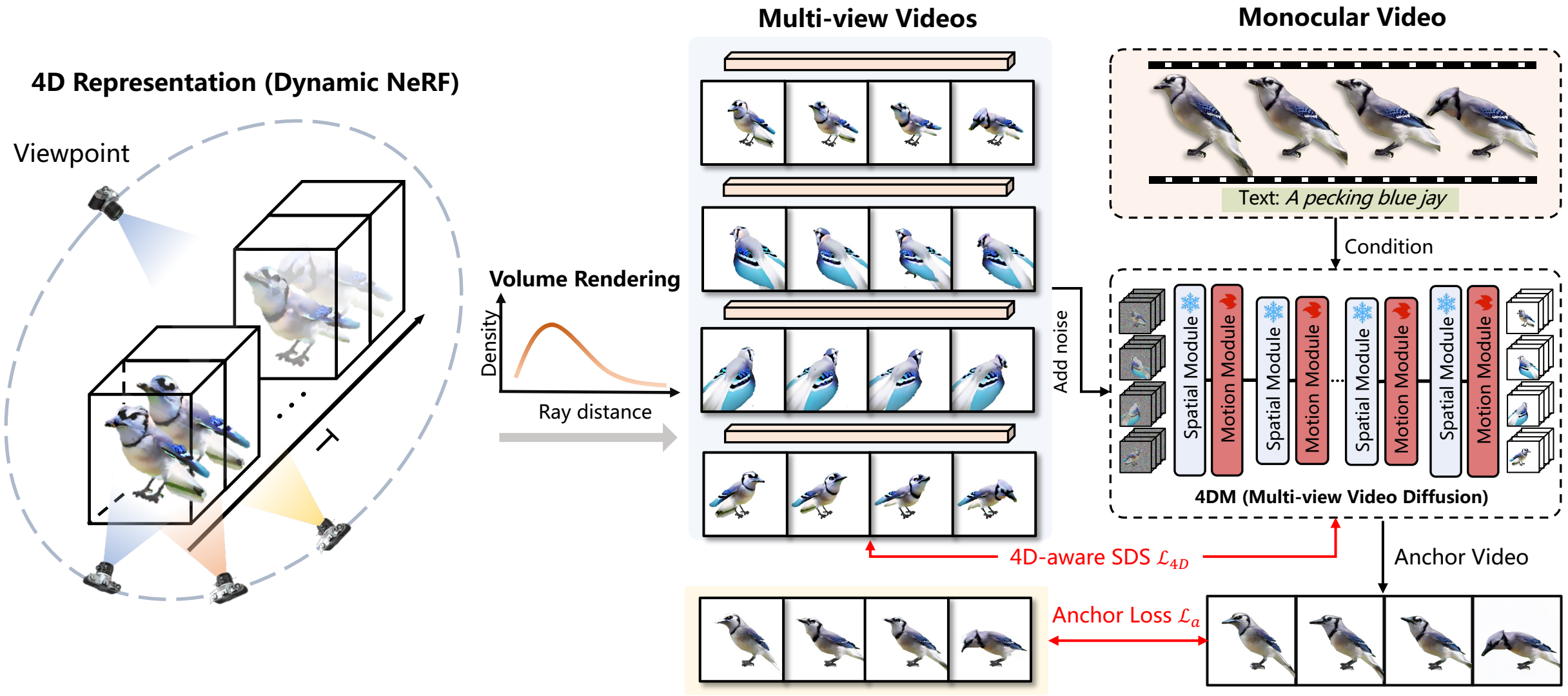
Ours(4DM)

Temporal consistent

GT

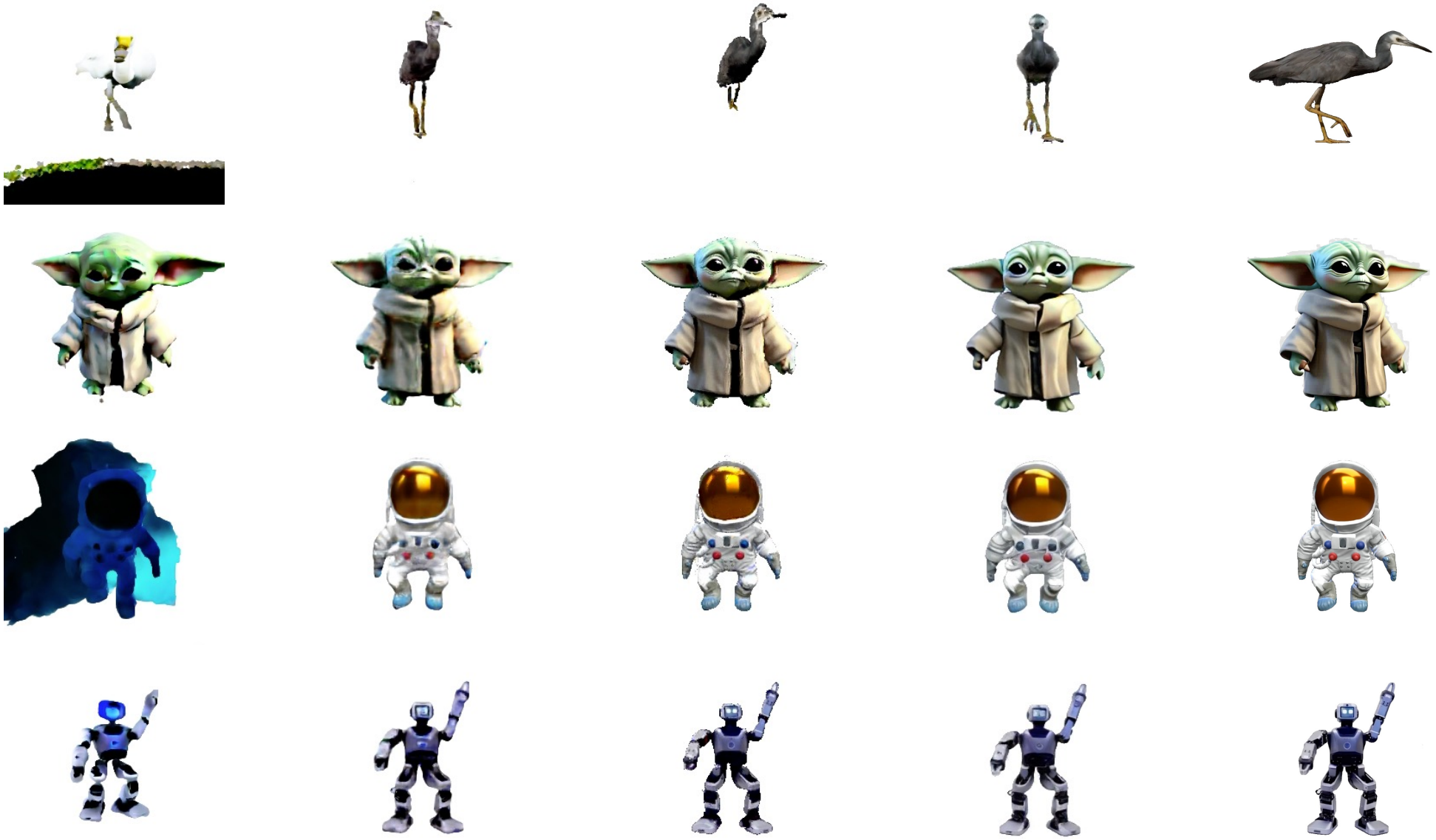
- 4DM successfully learns **reasonable temporal dynamics** and preserves the characteristics of ImageDream, including **generalization ability and spatial consistency**, even when trained on a small curated dataset.
- Finally, 4DM can generate four spatial-temporally consistent multi-view videos. Moreover, it provides **multi-view spatial-temporal guidance** for 4D generation.

4D Generation



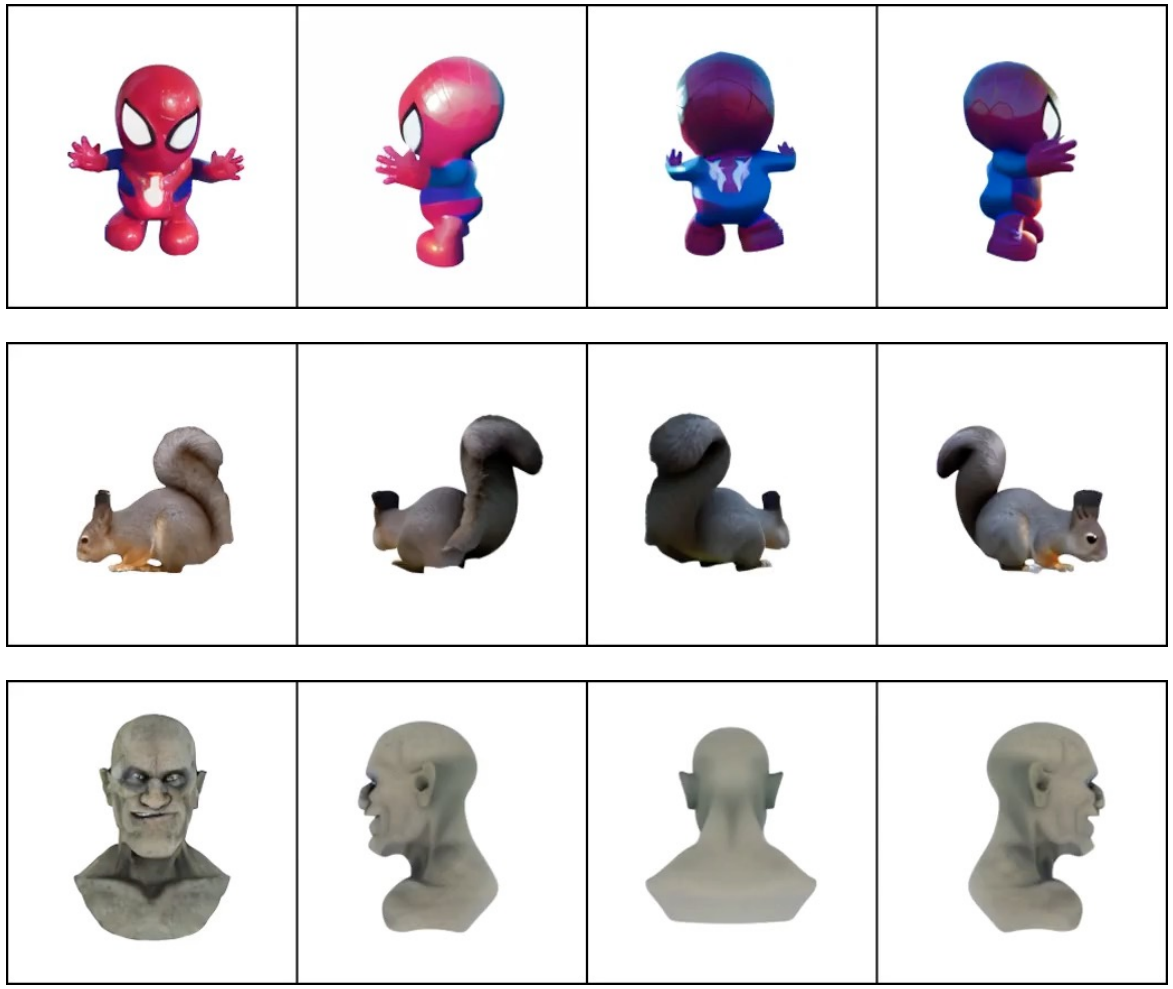
- We leverage **4D-aware SDS** to optimize the dynamic NeRF, enabling effective rendering from novel viewpoints across the temporal dimension.
- Moreover, we devise an **anchor loss** to enhance the appearance details and facilitate the learning.

Comparisons on 4D Generation

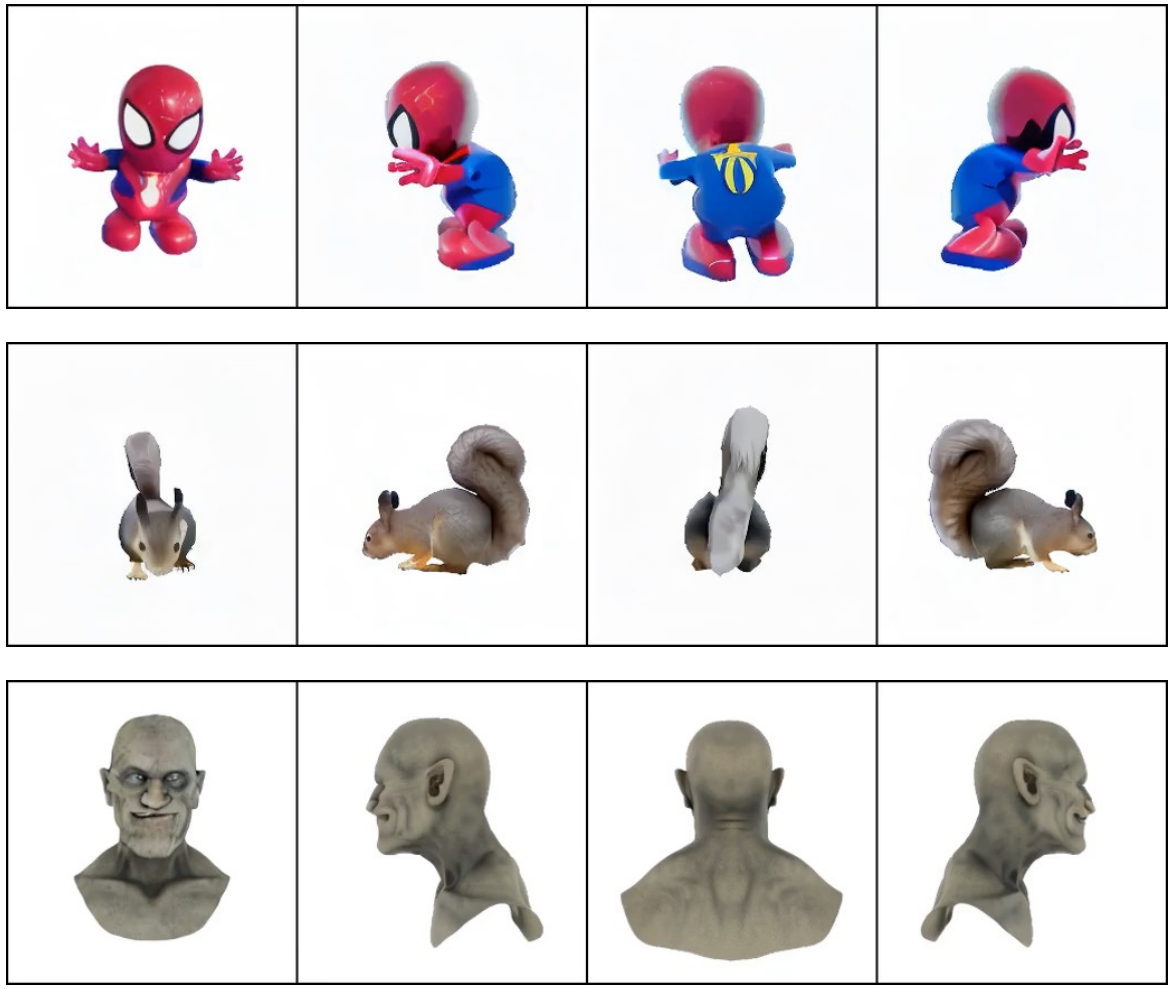


4D-fy | Consistent4D | Dream Gaussian4D | Ours | Reference

Comparisons on Multi-view Video Generation



ImageDream



Ours(4DM)

4DM effectively captures spatial-temporal correlations.

IRIP Laboratory
https://irip.buaa.edu.cn

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

Project Page :
https://aejion.github.io/4diffusion/

Thank you !

Haiyu Zhang

Xinyuan Chen

Yaohui Wang

Xihui Liu

Yunhong Wang

Yu Qiao