Gaussian Variation Field Diffusion for High-fidelity Video-to-4D Synthesis

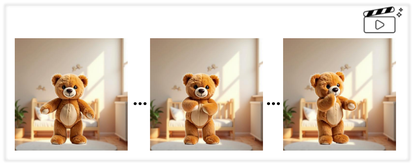
Bowen Zhang[1], Sicheng Xu[2], Chuxin Wang[1], Jiaolong Yang[2],
Feng Zhao[1], Dong Chen[2], Baining Guo[2].
[1]University of Science and Technology of China, [2]Microsoft Research Asia

Scan Project Page for More Video Results

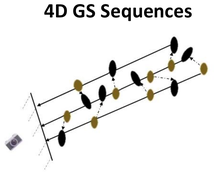
ICCV OCT 19-23. 2025 HONOLULU HAWAII

Looking for full-time research roles in Generative AI, starting 2026. Scan to learn more about me.

Introduction



115K Latent Tokens for 720p, 32 frames
✅ Manageable for Spatial Temporal Transformer



Typical 100K GS × 32 frames ≈ 3.2M Tokens
❌ 28x more tokens
❌ Existing architectures fail at this scale

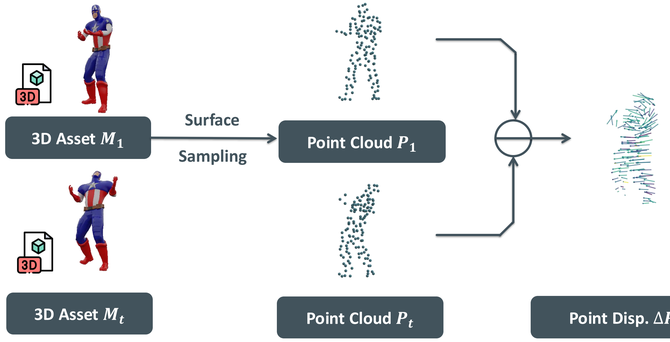
Goal: Generate high-quality **dynamic 3D content (4D)** from a single video.

Challenge: Direct 4D diffusion is costly — building data + jointly modeling **shape**, **appearance**, and **motion** is intractable.

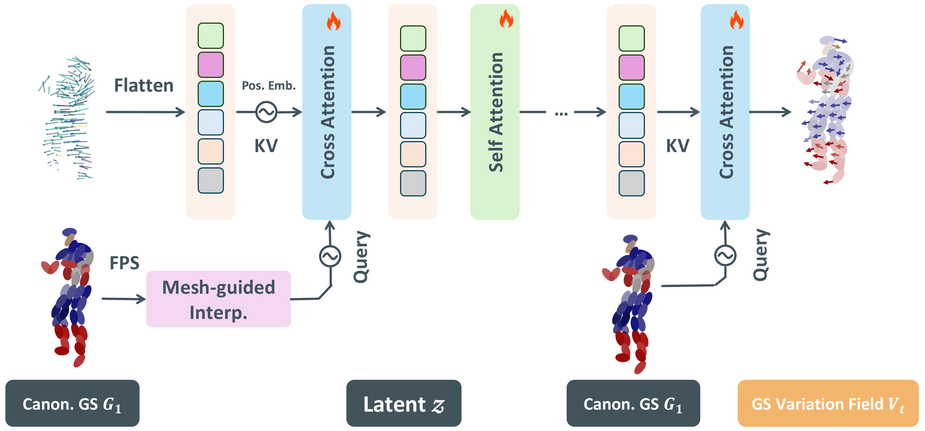
Our Approach:

- **4DMesh-to-GS Variation Field VAE** encodes canonical Gaussians and temporal variations from 4D data and compresses them into a **compact latent space**.
- **Gaussian Variation Field Diffusion** with **temporal-aware transformers** synthesizes 4D assets from input videos.

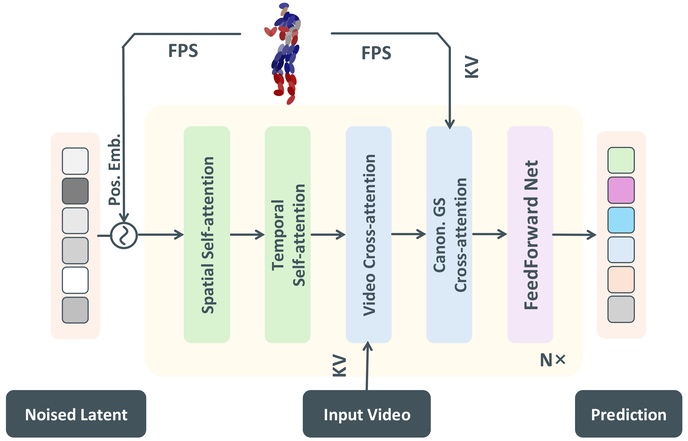
Directly Extract Motion from 4D Mesh Sequences



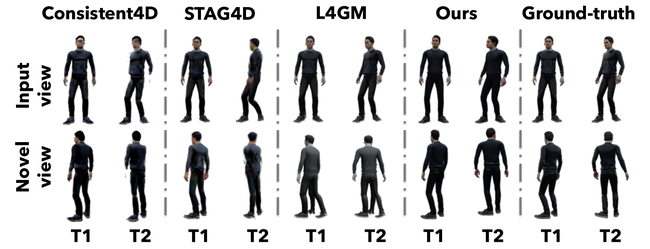
Direct 4DMesh-to-GS Variation Field VAE



Gaussian Variation Field Diffusion



Results



Method	PSNR↑	LPIPS↓	SSIM↑	CLIP↑	FVD↓	Time↓
Consistent4D [28]	16.20	0.146	0.880	0.910	935.19	∼1.5 hr
SC4D [79]	15.93	0.164	0.872	0.870	833.15	∼20 min
STAG4D [88]	16.85	0.144	0.887	0.893	1008.40	∼1 hr
DreamGaussian4D [56]	15.24	0.162	0.868	0.904	799.56	∼15 min
L4GM [57]	17.03	0.128	0.891	0.930	529.10	**3.5 s**
Ours	**18.47**	**0.114**	**0.901**	**0.935**	**476.83**	4.5s

Quantitative comparison of video-to-4D generation results.