HybridDeform4D: Joint Refinement of Mesh and Gaussian Splatting for Video-to-4D Object Generation

Duotun Wang
dwang866@connect.hkust-gz.edu.cn
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China

Yuhang Li
liyh82@lenovo.com
Lenovo CTOO
Beijing, China

Zhijing Shao
zshao@connect.hkust-gz.edu.cn
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China

Wanjun Lv
lvwj1@lenovo.com
Lenovo CTOO
Beijing, China

Liuxin Zhang
zhanglx2@lenovo.com
Lenovo CTOO
Beijing, China

Mingming Fan
mingmingfan@ust.hk
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
The Hong Kong University of Science and Technology
Hong Kong, China

Zeyu Wang
zeyuwang@ust.hk
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
The Hong Kong University of Science and Technology
Hong Kong, China

Abstract

Generating dynamic, spatiotemporally consistent 3D content from monocular video is a significant challenge. Existing methods leveraging Gaussian Splatting (GS) often struggle to maintain such coherence without strong regularization. We introduce *HybridDeform4D*, a novel framework that enhances 4D object generation by jointly optimizing a deformable mesh with its surface-bound GS attributes. This framework supports starting with a coarse mesh and capturing both complex motion and appearance variations without requiring high-quality initialization. To ensure coherent animation, we introduce a coarse-to-fine optimization scheme. This strategy first focuses on learning motion around a key reference frame before progressively expanding to optimize the entire video sequence, ensuring stable and accurate temporal dynamics. Experimental results show that *HybridDeform4D* achieves high rendering quality and spatial-temporal consistency. Furthermore, we demonstrate the versatility of our approach by extending it to text-to-3D synthesis.

ACM Reference Format:
Duotun Wang, Yuhang Li, Zhijing Shao, Wanjun Lv, Liuxin Zhang, Mingming Fan, and Zeyu Wang. 2025. HybridDeform4D: Joint Refinement of Mesh and Gaussian Splatting for Video-to-4D Object Generation. In *SIGGRAPH Asia 2025 Posters (SA Posters '25), December 15-18, 2025*. ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3757374.3771441

SA Posters '25, Hong Kong, China

ACM ISBN 979-8-4007-2134-2/2025/12
https://doi.org/10.1145/3757374.3771441

1 Introduction

Recent advancements in 4D synthesis build upon score distillation sampling (SDS) [Poole et al. 2022] from 2D diffusion models to generate dynamic 3D content. While recent methods using GS [Kerbl et al. 2023] representation are quite popular, they often model motion by optimizing each Gaussian independently [Ren et al. 2022; Zeng et al. 2024], creating an unstructured and inefficient space for learning deformations. DreamMesh4D [Li et al. 2024] leverages a deformable mesh to govern the motion of surface-bound Gaussians (i.e., SuGaR [Guédon and Lepetit 2024]). This approach enforces spatial coherence through the mesh's connectivity. However, DreamMesh4D struggles to capture intricate appearance changes or topological variations like object-part overlaps (Figure 3), which are not easily represented by mesh transformations alone.

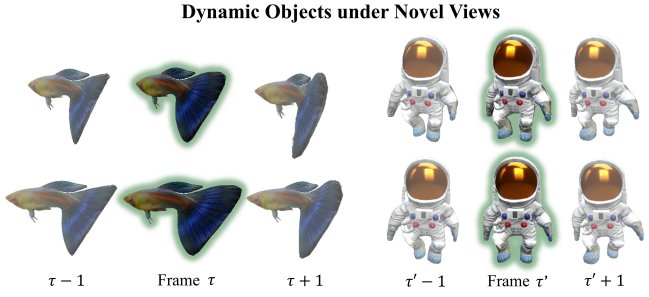

Figure 1: Generated 4D assets through HybridDeform4D.

We introduce *HybridDeform4D*, a coarse-to-fine framework for video-to-4D generation. As illustrated in Figure 2, our method first generates a textured mesh from a key reference frame using image-to-3D techniques [Liu et al. 2023; Xiang et al. 2025]. Then, a dynamic learning stage refines both geometry and appearance through the

joint optimization of the mesh vertices and attributes of bound surface Gaussians. To achieve high-quality mesh deformations, we reparameterize the vertices ($V \rightarrow V^* = (I + \lambda L)$) based on the Laplace-Beltrami operator L [Nicolet et al. 2021].

2 Method

HybridDeform4D starts by attaching 6 Gaussians to each face of an initial mesh, adopting the differentiable rendering pipeline and loss designs from DreamMesh4D. After rendering its RGB $\hat{C}^*_\tau$ and alpha $\hat{S}^*_\tau$ under reference view, we compute reconstruction loss $\mathcal{L}_{\text{ref}} = ||\hat{C}^*_\tau - C^*_\tau||$ and mask loss $\mathcal{L}_{\text{mask}} = ||\hat{S}^*_\tau - S^*_\tau||$, where C^*_τ and S^*_τ are the ground-truth data from input video at timestamp τ.

$$\mathcal{L}_{\text{4D}} = \lambda_{\text{SDS}}\mathcal{L}_{\text{SDS}} + \lambda_{\text{ref}}\mathcal{L}_{\text{ref}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}} + \lambda_{\text{ARAP}}\mathcal{L}_{\text{ARAP}} \quad (1)$$

The hybrid representation is optimized with the above composite weighted loss function, where an SDS loss is utilized for supervision under other randomly sampled views. As-Rigid-As-Possible (ARAP) regularization is incorporated to preserve geometric integrity.

A core component of our method is a Multi-Layer Perceptron (MLP) that explicitly models the joint optimization of time-varying geometry and appearance. Conditioned on the embedding of frame τ, this network predicts the deformed mesh vertex positions (V^*) as well as the dynamic attributes of bound Gaussians (i.e., opacity and Spherical Harmonic (SH) coefficients).

To ensure temporally coherent results, we employ a coarse-to-fine optimization strategy. The training process initially focuses on a narrow window of frames around the reference view (e.g., $\tau_0 \pm 1$). This temporal window then linearly expands throughout the first 70% of the optimization steps until it encompasses the entire video sequence. Frame sampling within this window is hybrid: 70% of frames are chosen randomly to capture diverse states, while 30% are selected sequentially to maintain local motion dynamics.

3 Evaluation

For quantitative analysis, we adopted the metrics from Stag4D and DreamMesh4D. We evaluated image-level fidelity using the CLIP score and reference-view PSNR. Temporal coherence was assessed with the FID-VID. To measure mesh quality during deformation, the self-intersection ratio was calculated. As detailed in Table 1, our framework consistently outperforms baseline approaches.

An ablation study was conducted to validate our joint optimization and coarse-to-fine strategy. As shown in Figure 3, joint optimization is critical for capturing fine details, such as the movement of the elephant's eye, while the coarse-to-fine approach successfully models large-scale motions, like the frog's arm lifting. Furthermore, we demonstrate the versatility of our joint refinement process, which can be extended to other tasks like optimization-based image-to-3D generation (Figure 4).

Table 1: Quantitative comparison with baseline methods.

Method	PSNR (ref) ↑	FVD ↓	SSIM (ref) ↑	Self-Intersection ↓
DreamMesh4D [Li et al. 2024]	36.68	498.47	0.858	0.69%
Stag4D [Zeng et al. 2024]	33.55	610.77	0.844	-
Ours	**38.03**	**427.02**	**0.865**	**0.18%**

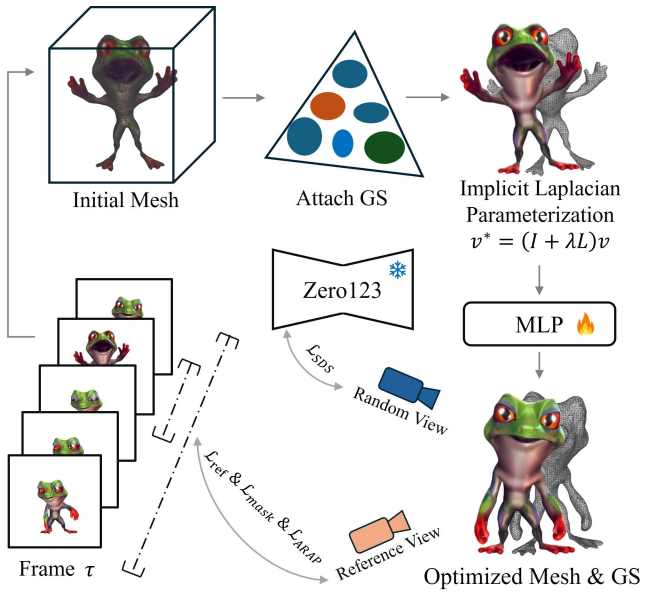

Figure 2: Overview of the HybridDeform4D pipeline. Our pipeline begins by generating a hybrid mesh-Gaussian representation from a single reference frame. We then jointly optimize the mesh deformation and the GS attributes for dynamic appearance.

(a) w/ and w/o Joint Refinement (b) w/ and w/o Coarse-To-Fine

Figure 3: Ablation study results.

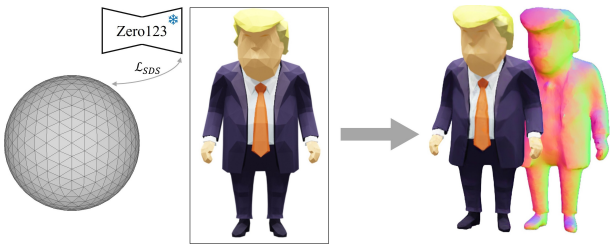

Figure 4: Applying our joint refinement to 3D generation. Mesh deformation serves as an effective geometric prior to improve Gaussian optimizations.

4 Discussion

HybridDeform4D enhances 4D object generation by adopting joint refinement optimizations for both mesh deformation and GS appearance. This strategy effectively supports the spatial-temporal consistency during dynamic generation. Future work can expand our approach to feed-forward training and scene-level generation.

References

Antoine Guédon and Vincent Lepetit. 2024. SuGaR: Surface-Aligned Gaussian Splatting for Efficient 3D Mesh Reconstruction and High-Quality Mesh Rendering. In *CVPR*.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.* 42, 4 (2023).

Zhiqi Li, Yiming Chen, and Peidong Liu. 2024. Dreammesh4D: Video-to-4D Generation with Sparse-Controlled Gaussian-Mesh Hybrid Representation. *Advances in Neural Information Processing Systems* 37 (2024), 21377–21400.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. 2023. Zero-1-to-3: Zero-shot One Image to 3D Object. *arXiv preprint arXiv:2303.11328* (2023).

Baptiste Nicolet, Alec Jacobson, and Wenzel Jakob. 2021. Large Steps in Inverse Rendering of Geometry. *ACM Trans. Graph.* 40, 6, Article 248 (2021), 13 pages.

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. 2022. DreamFusion: Text-to-3D Using 2D Diffusion. *arXiv preprint arXiv:2209.14988* (2022).

Jiawei Ren, Liang Pan, Jiaxiang Tang, Chi Zhang, Ang Cao, Gang Zeng, and Ziwei Liu. 2022. DreamGaussian4D: Generative 4D Gaussian Splatting. *arXiv preprint arXiv:2312.17142* (2022).

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. 2025. Structured 3D Latents for Scalable and Versatile 3D Generation. In *CVPR*.

Yifei Zeng, Yanqin Jiang, Siyu Zhu, Yuanxun Lu, Youtian Lin, Hao Zhu, Weiming Hu, Xun Cao, and Yao Yao. 2024. STAG4D: Spatial-Temporal Anchored Generative 4D Gaussians. In *ECCV*. Berlin, Heidelberg, 163–179.