Helix4D: Complex 4D Mesh Generation

Jiraphon Yenphraphai[1,2*] **Jianqi Chen**[1,3*] **Jian Wang**[1] **Gordon Qian**[1]
Sergey Tulyakov[1] **Rameen Abdal**[1] **Raymond A. Yeh**[2] **Peter Wonka**[1,3] **Chaoyang Wang**[1]
[1]Snap [2]Purdue University [3]KAUST

https://snap-research.github.io/helix4d/

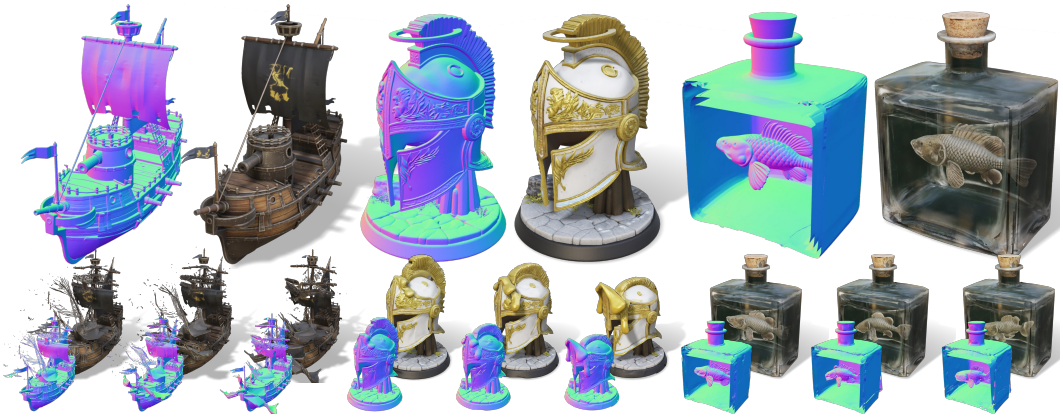


Figure 1: We propose Helix4D, a framework that can generate dynamic 3D shapes, including their materials, from an input video. **Left:** accurate reconstruction of thin structures and changes of fine details such as sails, ropes, railings, and flags breaking. **Middle:** detailed ornaments in the geometry and texture while melting. **Right:** transparent and inner-surface structures, such as a fish moving inside a glass bottle. Note, we cut the mesh open to reveal the fish inside. The bottom row shows generated dynamic sequences, demonstrating that Helix4D can model challenging 4D scenarios, including topology changes, deformation, shattering, melting, transparency, and thin structures.

Abstract

Current video-to-4D methods struggle with complex topology changes, transparent materials, thin structures, and inner surfaces. We present Helix4D, a dynamic mesh generation framework by inheriting the expressive representation of Trellis2, adapting it from image-to-3D to video-conditioned 4D generation. Our design arises from two key questions: (a) how to enable Trellis2's frame-local attention to share information across frames while preserving its pretrained quality on rare cases such as transparent objects and inner surfaces, and (b) how to inject temporal information into a purely 3D positional encoding without breaking pretrained capabilities. We address (a) with a sliding-window cross-frame attention and anchor on the first frame. The first frame is generated by the base Trellis2 model and injected into our model, letting it inherit Trellis2's quality in rare cases through cross-frame attention. We address (b) with a 4D temporal encoding that repurposes redundant low-frequency spatial RoPE bands for time, extending the encoding from 3D with no additional parameters. Extensive experiments show the effectiveness of Helix4D for high-quality dynamic mesh generation on ActionBench and our own challenging complex dynamics set.

*Work done during an internship at Snap.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1 Introduction

We propose a new framework for video to dynamic 3D (also called 4D) shape generation, a fundamental problem in vision and graphics, with applications in animation, virtual reality, and robotics. Compared to static 3D generation, 4D shape generation requires not only accurate geometry and materials, but also consistent modeling of motion, topology changes, and temporal coherence.

Recent approaches to video-to-4D generation have made progress using inference-time optimization [1, 14, 39, 46], multi-view video generation and reconstruction [8, 15, 25, 29, 35], or separate shape and motion generation [2, 7, 42]. More recently, feedforward methods that extend image-to-3D diffusion models [13, 22, 36] have demonstrated improved quality and generalization. However, despite strong performance on rigid and simple objects, these approaches struggle with complex topology variations, material modeling, transparency, and the reconstruction of inner surfaces.

In contrast, strong priors over geometry and materials, including the ability to represent thin structures, non-watertight surfaces, and complex appearance properties, only became available in recent large-scale foundational 3D models, *e.g.*, Trellis2 [31]. In this work, we present a novel framework (called Helix4D) for *dynamic mesh generation from video* by systematically extending Trellis2 to 4D while preserving its pretrained strengths. Our approach enables high-quality 4D generation with significantly improved handling of challenging cases, including transparent and semi-transparent objects, complex materials, topological changes, and inner-surface reconstruction (See Fig. 1).

To achieve this, we address four technical challenges: enabling efficient cross-frame interaction at scale, retaining the generative ability of Trellis for challenging geometry (*e.g.*, transparency, inner surfaces) despite limited training data, incorporating temporal information into a spatial-only positional encoding, and preserving the generative quality of the pretrained model. First, we design a sliding-window cross-frame attention mechanism augmented with an anchor frame and first-frame conditioning. This combines efficient local temporal interaction with a global reference signal, allowing the model to share information across frames while maintaining high-quality geometry and material quality similar to full attention. Further, conditioning on an anchor frame enables the model to retain capabilities of Trellis (transparent surfaces, inner geometry) despite very limited 4D training data of such challenging cases. Second, we introduce a spatiotemporal rotary embedding inspired by ReRoPE [10], which repurposes low-frequency spatial RoPE bands to encode time. This yields a parameter-free extension to 4D that preserves RoPE's relative-position properties and maintains compatibility with pretrained weights. A natural alternative, adopted by SS4D [13], adds a temporal RoPE [24] on top of the backbone's existing positional encoding at every attention layer. This is suboptimal for a pretrained backbone: the extra rotation introduces phases the pretrained key/query projections never saw, disrupting the learned positional signal. We confirm this in Sec. 5.2: applying the SS4D recipe in our architecture underperforms our proposed embedding.

Overall, our method converts a state-of-the-art image-to-3D generator into a video-conditioned 4D model that produces temporally consistent dynamic meshes while retaining strong geometric and material fidelity. On ActionBench [22], Helix4D improves CD-3D by 3.8% over ActionMesh [22]. On our harder 52-video benchmark, it outperforms all baselines on every metric, improving ULIP-2 [34] and Uni3D [47] by 5.7% and 7.8% over the strongest baseline, and is preferred over the best-performing baseline in 67.9% of user-study comparisons.

Our main contributions are:

- **4D generation with advanced geometric and material capabilities.** We extend a strong image-to-3D generative model to video, enabling dynamic mesh generation that for the first time handles challenging cases including transparency, complex materials, complex topology changes, and inner surface reconstruction.
- **Efficient cross-frame modeling with reference conditioning.** We propose a sliding-window attention mechanism with an anchor frame and first-frame conditioning. This enables efficient training and inference, as well as overcoming the data scarcity of challenging 4D training data.
- **Spatiotemporal RoPE via frequency repurposing.** We introduce a parameter-free extension of rotary position embeddings based on ReRoPE, which encodes time by repurposing low-frequency spatial bands while preserving relative-position properties and pretrained initialization.

2 Related Work

Optimization-based 4D generation. Early 4D methods optimize a per-instance representation against pretrained diffusion priors. Text-conditioned variants distill motion from video diffusion via score distillation sampling [1, 18, 23, 39, 46]; video-conditioned variants lift a monocular reference clip by combining photometric reconstruction with SDS, frame-interpolation, or non-rigid warping losses [6, 20, 28, 38, 40, 43]; two-stage methods replace distillation with multi-view video diffusion followed by 4D reconstruction [8, 15, 25, 29, 35]; and V2M4 [3] registers 3D meshes into a shared topology. All four are slow (hours per asset) and have artifacts, motivating feed-forward generation.

Feed-forward 3D generation. 3D feed-forward generators differ mainly in their latent representation. Voxel-based methods such as Trellis [32] and Direct3D-S2 [30] attach features to a sparse grid intersecting the surface, while vecset-based methods originating from 3DShape2VecSet [41] encode shapes as unordered latent sets decoded into an implicit field [5, 9, 11, 12, 45]. Both lines decode signed distance or occupancy, which requires watertight, manifold training data and cannot represent open surfaces, non-manifold topology, or inner surfaces. Trellis2 [31] resolves this with O-Voxels, a sparse near-surface representation that supports such structure together with PBR materials. We adopt Trellis2 as a 3D prior and adapt it to 4D.

Feed-forward 4D generation. Existing feed-forward 4D methods split along a representation versus quality tradeoff. Deformation-centric approaches, including Mesh4D [7], Motion 3-to-4 [2], ActionMesh [22], and GVFD [42], reconstruct a canonical asset from the first frame and predict a warp field; this yields smooth motion but inherits the canonical asset's topology, which prohibits them from modeling topology changes. Other generators such as L4GM [21], SS4D [13], Sculpt4D [37], and ShapeGen4D [36] learn spacetime latents end-to-end and avoid the topology constraint, but their per-frame geometry and materials qualities are limited. By building on the O-Voxel representation, Helix4D is the first feed-forward 4D generator to support non-manifold geometry, topology changes, and transparent materials at high quality.

3 Background

Trellis2 [31] is a 3D asset generation model that takes an input image and predicts a textured mesh. The approach comprises three main components: (i) an O-Voxel representation that converts a 3D asset into sparse voxel features, (ii) a Sparse Compression VAE that encodes these features into a latent space, and (iii) three flow-matching models that generate sparse structure, geometry, and material latents conditioned on an input image, respectively. Given a 3D asset, Trellis2 converts it into a sparse set of active voxels: $\mathcal{F} = \{(f_i^{\text{shape}}, f_i^{\text{mat}}, p_i)\}_{i=1}^{L}$, where p_i is the i^{th} active O-Voxels, f_i^{shape} stores local geometry information, and f_i^{mat} stores material information. Empty voxels are discarded, giving a sparse representation that is efficient for high-resolution 3D generation. Unlike SDF-based representations [11, 33, 45], this representation does not require watertight geometry, allowing it to represent open surfaces, thin structures, and interior surfaces. With this representation, Trellis2 generates a 3D asset from a *single input image* through three flow-matching stages built on the same DiT-style backbone: a *sparse-structure* stage that predicts active voxel locations, a *geometry* stage that predicts per-voxel dual vertices and connectivity, and a final stage that predicts the *material*. As this work improves the foundation of the backbone that is applied across all three stages, we describe our method generically over a single stage.

4 Dynamic Mesh Generation

The goal is to generate a dynamic textured mesh sequence from an object-centric input video. Building on the Trellis2 architecture reviewed in Sec. 3, we generate dynamic mesh sequences from an object-centric input video by converting Trellis2 from image-to-3D into video-conditioned 4D generation while reusing its pretrained weights as in Fig. 2. Doing so lets us tackle cases that prior video-to-4D approaches [2, 7, 13, 22, 36] struggle with: complex topology changes, transparent or semi-transparent objects, and inner-surface reconstructions.

Our design aims to address the following questions: (a) How to enable Trellis2's frame-local attention layers to share information across frames, while preserving its pretrained generation quality on rare cases such as transparent objects and inner surfaces that are barely seen in 4D datasets? **(b)**

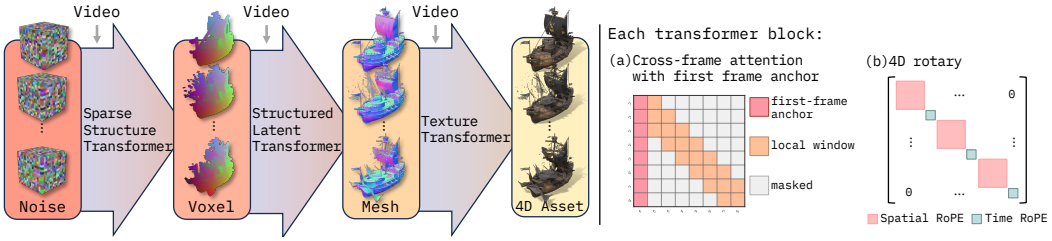


Figure 2: **Helix4D pipeline.** Given an input video, Helix4D generates a 4D asset through three flow-matching stages: *sparse-structure* generation, *geometry* generation, and *material* generation. Each self-attention layer in the transformer block uses (a) sliding-window cross-frame attention with a first-frame anchor and (b) repurposes low-frequency spatial RoPE bands for temporal encoding.

How to inject temporal information into a model whose positional encoding is purely 3D, without breaking pretrained capabilities? We address **(a)** in Sec. 4.1 with sliding-window cross-frame attention augmented by an anchor frame: the first frame is generated by the base Trellis2 model and injected into our model as the anchor, letting our model inherit Trellis2's capabilities on rare cases through cross-frame attention. We address **(b)** with a ReRoPE-inspired [10] temporal positional encoding in Sec. 4.2, which repurposes low-frequency spatial RoPE bands for time domains, keeping the model dimension fixed while extending the encoding from 3D to 4D.

4.1 Cross-frame attention with first-frame anchor

Each Trellis2 stage operates on a sparse per-frame token sequence (reviewed in Sec. 3), and its pretrained attention layers are limited to only within a single per-frame sequence. To extend this to 4D, we treat the full video token stream as a *sequence of sequences*, and apply cross-frame attention at every self-attention layer of the pretrained backbone. Each frame in the input video, indexed by $f \in \{0, \dots, F-1\}$, contributes S_f tokens represented by features $\{\mathbf{x}_{f,s}\}_{s=1}^{S_f}$ at voxel coordinates $\{\mathbf{p}_{f,s}\}_{s=1}^{S_f}$. The per-frame tokens are then concatenated across frames, yielding a total token sequence of length $S = \sum_{f=0}^{F-1} S_f$. Each token is identified by the pair (f, s); $f = 0$ denotes the first frame.

Within the transformer model, attention layers compute a query $\mathbf{q}_{f,s}$ and a key $\mathbf{k}_{f,s}$ from its feature $\mathbf{x}_{f,s}$ for each token at (f, s) via learned projections. Next, how information is aggregated across tokens is based on a binary mask $\mathbf{M} \in \{0,1\}^{S \times S}$. When $M_{(f,s),(f',s')} = 1$, it allows the query at (f, s) to attend to the key at (f', s') and 0 blocks it. In other words, the final attention weights are computed as

$$A_{(f,s),(f',s')} = M_{(f,s),(f',s')} \exp\left(\mathbf{q}_{f,s}^{\top} \mathbf{k}_{f',s'} / \sqrt{d}\right), \tag{1}$$

normalized over dimensions of (f', s'). As illustrated in Fig. 3, different attention designs correspond to different choices of $\mathbf{M}$. Naively using a full attention $\mathbf{M} \equiv \boldsymbol{I}$ is too costly, as a single 4D reconstruction in our setting contains up to 10^5 tokens. To keep the cost low, it is important to design a sparse attention pattern while allowing sufficient information to be shared across frames.

Sliding window with anchor. We propose to restrict the attention to a *sliding window with an anchor frame*. That is, a token at frame f attends to tokens within a temporal window of half-width w around f, plus all tokens in the first frame:

$$M_{(f,s),(f',s')} := \begin{cases} 1, & |f - f'| \le w \text{ or } f' = 0, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

The local window captures short-range motion while the anchor (first-frame, $f' = 0$) provides a shortcut for the shape context that is the most accurate in the first frame. This avoids the need to propagate shape information throughout the sequence. Empirically, sliding window with anchor attention matches full attention quality at $2\times$ lower computation cost; more discussion in Tab. 5.

First-frame conditioning. A key challenge in 4D generation is data scarcity for transparent objects, semi-transparent materials, or inner surfaces' motion. A model trained from scratch on these datasets

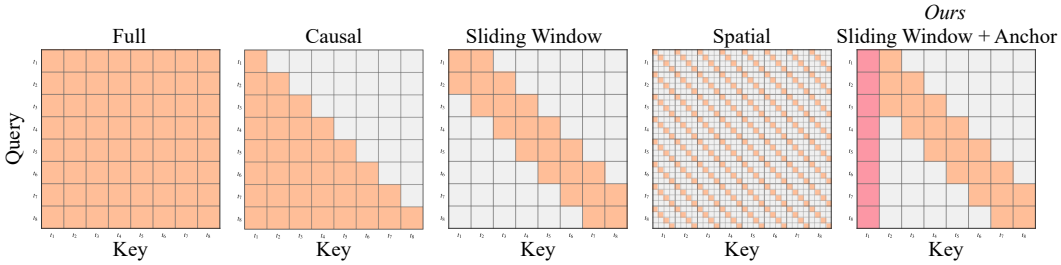


Figure 3: **Cross-frame attention patterns.** Orange cells indicate allowed attention, gray cells indicate masked attention, and red cells indicate attention to the first-frame anchor. Full attention allows every frame to attend to each other, but it is computationally expensive. Causal attention restricts each frame to previous frames, while sliding-window attention limits attention to nearby frames for efficiency. Spatial attention attends only within corresponding spatial positions across frames. Our design combines sliding-window attention with a first-frame anchor (right), enabling efficient temporal information sharing while preserving the static 3D prior from the Trellis2 reconstruction.

struggles to generate such properties. The pretrained Trellis2 model, by contrast, handles these objects well on static 3D assets. We feed the Trellis2-generated first frame as a clean reference, so the noisy frames $f \geq 1$ can attend to it and inherit its representations, making this task easier.

This is implemented by placing the first-frame tokens at frame index $f = 0$ of the cross-frame sequence. The denoising timestep is set to zero, and frames $f \geq 1$ use the standard noisy latents. During training, the flow-matching loss is computed only on frames $f \geq 1$, and we use the ground-truth first-frame latent. At sampling time, we generate the anchor by running a frozen Trellis2.

4.2 Repurposing Spatial RoPE for Time

Recap of RoPE. Rotary Position Embedding (RoPE) [24] encodes position into the key and query through a position-dependent rotation. Let d be the feature dimension per attention head. Trellis2 [31] partitions these features into three equal axes for x, y, z, each carrying $M = d/6$ rotary frequency pairs ($d/2$ pairs total). Voxel coordinates $\mathbf{p} = (p^x, p^y, p^z)$ are integers in $[0, N-1]^3$ with grid resolution N, and time indices t are integers from 0 to $T-1$. Each token m has an input feature $\mathbf{x}_m \in \mathbb{R}^d$ at voxel coordinate $\mathbf{p}_m$, and $\mathbf{W}_k, \mathbf{W}_q \in \mathbb{R}^{d\times d}$ denote the learned key and query projection matrices. The rotary frequencies along each axis are: $\omega_i = \theta^{-2(i-1)/M}$, where $i \in \{1, \ldots, M, \}$, $\theta = 10000$. We write $\mathbf{R}(\phi) = \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix} \in SO(2)$ for the 2×2 rotation matrix.

Spatial RoPE in Trellis2. For a scalar position p along a single axis, the per-axis rotary block is

$$\mathbf{R}_{\Omega,p} = \bigoplus_{i=1}^{M} \mathbf{R}(\omega_i p) = \begin{bmatrix} \mathbf{R}(\omega_1 p) & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{R}(\omega_M p) \end{bmatrix} \in \mathbb{R}^{2M\times 2M}. \tag{3}$$

where the notation $\oplus$ means the block diagonal concatenation, and the off-diagonals are zeros.

The 3D rotary at voxel coordinate $\mathbf{p} = (p^x, p^y, p^z)$ is the block-diagonal concatenation across axes:

$$\mathbf{R}_{\Omega,\mathbf{p}} = \mathbf{R}_{\Omega,p^x} \oplus \mathbf{R}_{\Omega,p^y} \oplus \mathbf{R}_{\Omega,p^z} \in \mathbb{R}^{d\times d}. \tag{4}$$

Queries and keys at token m are $\mathbf{q}_m = \mathbf{R}_{\Omega,\mathbf{p}_m}\mathbf{W}_q\mathbf{x}_m$ and $\mathbf{k}_m = \mathbf{R}_{\Omega,\mathbf{p}_m}\mathbf{W}_k\mathbf{x}_m$, giving relative-position attention:

$$\mathbf{q}_m^\mathsf{T}\mathbf{k}_n = (\mathbf{R}_{\Omega,\mathbf{p}_m}\mathbf{W}_q\mathbf{x}_m)^\mathsf{T}(\mathbf{R}_{\Omega,\mathbf{p}_n}\mathbf{W}_k\mathbf{x}_n) = \mathbf{x}_m^\mathsf{T}\mathbf{W}_q^\mathsf{T}\,\mathbf{R}_{\Omega,\,\mathbf{p}_n-\mathbf{p}_m}\,\mathbf{W}_k\mathbf{x}_n. \tag{5}$$

For 4D generation each token has a coordinate $(\mathbf{p}, t)$, where $\mathbf{p} \in \{0, \ldots, N-1\}^3$ is the voxel coordinate and $t \in \{0, \ldots, T-1\}$ is the frame index. We want a 4D rotary $\mathbf{R}_{\Omega,(\mathbf{p},t)}$ that (a) reuses Trellis2's pretrained weights, (b) keeps d unchanged, (c) preserves RoPE's relative-position property. A separate 1D temporal RoPE applied on top of $\mathbf{R}_{\Omega,\mathbf{p}}$ would entangle spatial and temporal phases in the same feature pairs, violating (c). Inspired by ReRoPE [10], we instead repurpose existing information inside $\mathbf{R}_{\Omega,\mathbf{p}}$.

Proposed re-purposing. Our observation is that the high-frequency part of the spatial rotary matrix is sufficient to distinguish voxel coordinates, while the lower-frequency part varies slowly over the

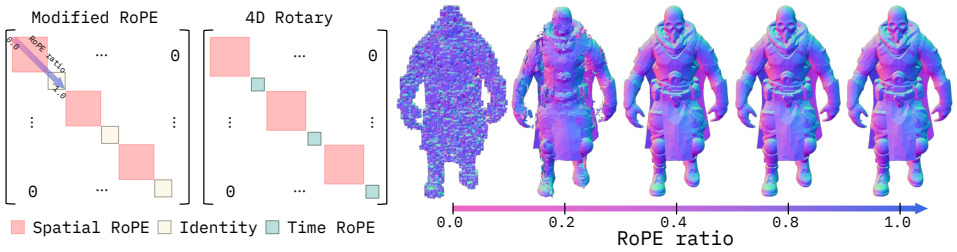


Figure 4: **Effect of different RoPE ratios.** We test how many low-frequency spatial RoPE bands can be removed from a pretrained Trellis2 model by replacing them with identity matrices at inference time (Left). Too small RoPE ratios degrade geometry. In contrast, the results (Right) at ratios 0.4 and 1.0 are visually indistinguishable, suggesting that the removed low-frequency bands contribute little to the output and can be repurposed for temporal encoding without quality drop (Middle).

voxel grid and contributes less to spatial localization. These low-frequency channels can therefore be reused to encode time without sacrificing the generation quality.

We test how many low-frequency bands per axis can be replaced. We split the per-axis rotary into a high-frequency block (top αM pairs) and a low-frequency block (bottom $(1-\alpha)M$ pairs):

$$\mathbf{R}_{\Omega,p} = \mathbf{R}^{\text{high}}_{\Omega,p} \oplus \mathbf{R}^{\text{low}}_{\Omega,p} \quad \longrightarrow \quad \mathbf{R}^{\text{high}}_{\Omega,p} \oplus \mathbf{I}. \tag{6}$$

We replace the low-frequency block with identity and run inference on the Trellis2 pretrained model on a held-out 32-object validation set. As shown in Fig. 4, generations is visually indistinguishable from the original ($\alpha = 1$) with $\alpha \geq 0.4$, and quantitative metrics saturate across $\alpha \in [0.4, 1.0]$ (see Tab. A1). When $\alpha < 0.4$, the quality drops. Therefore, any $\alpha \geq 0.4$ preserves spatial quality, and we can repurpose this low-frequency part for time embedding. Within this range, we set $\alpha = 0.75$ to allocate rotary bands proportionally to each axis's length: since the spatial extent $N = 64$ and temporal extent $T = 16$ are at comparable scale, the $T/N = 1/4$ ratio gives a balanced $75\%/25\%$ split between space and time.

4D rotary. Combining the truncated spatial RoPE with the phase-matched temporal block gives the full 4D rotary at (p, t):

$$\mathbf{R}_{\Omega,(p,t)} = \mathbf{R}^{\text{spatial}}_{\Omega,p} \oplus \mathbf{R}^{\text{temporal}}_{\Omega,t}, \tag{7}$$

where $\mathbf{R}^{\text{spatial}}_{\Omega,p}$ uses the top αM frequency pairs along each spatial axis ($3\alpha = 3d/8$ pairs total) and $\mathbf{R}^{\text{temporal}}_{\Omega,t}$ uses the high-frequency part N/T-scaled frequencies on the remaining features $d/8$ pairs.

The resulting RoPE construction gives a parameter-free extension of Trellis2's 3D RoPE to 4D space-time coordinates. Rather than adding new temporal channels or increasing the attention dimension, we allocate only the redundant low-frequency spatial bands to time. Importantly, the spatial and temporal RoPE remain separated as block-diagonal $SO(2)$ rotations; the attention map depends only on relative space-time distances.

5 Experiments

We evaluate our method on 4D generation against five video-to-4D baselines [2, 7, 13, 22, 36], on both our newly introduced 52-video benchmark, covering topology change, transparency, and volumetric phenomena, and ActionBench [22]. We further evaluate our method on real-world videos from DAVIS [17]. We then ablate each of our three core design choices and analyze alternative cross-frame attention patterns (Sec. 5.2).

Data curation. We curate our training data from the subset of animated TexVerse-1K [44], which has about 55k objects. For each asset, we extract 16 animation frames and convert every frame into an O-voxel representation at a resolution of 1024^3, following Trellis2 [31]. To ensure consistent scale and position across time, each object is rescaled using the union of per-frame bounding boxes so that the entire animation lies within $[-0.5, 0.5]$. For each animation, we render 16 views from randomly



Figure 5: **Qualitative comparison on Helix4DBench.** The left column shows input video frames, and each method is shown with geometry/normal renderings and, when available, shading. Compared with ActionMesh [22], Motion 3-to-4 [2], and Mesh4D [7], our method better preserves fine geometry, topology changes, emerging structures, and transparent objects.

sampled camera viewpoints on the sphere looking at the origin at a resolution of 1024×1024, with randomized focal length, radius, azimuth, and elevation.

Architecture and training details. We apply our method (4D rotary, our proposed attention, and first-frame conditioning) uniformly to all three Trellis2 stages: sparse-structure, geometry, and material. Each stage generates 16 frames jointly. 4D rotary keeps $\alpha = 0.75$ of the spatial frequencies, and the rest are repurposed for time. Cross-frame attention uses a window size of 5 plus the first frame as an anchor. Each stage fine-tunes only self-attention layers from the pretrained Trellis2 with a batch size of 32 on 32 A100 GPUs for 20K iterations, using AdamW [16] with learning rate 2×10^{-5}.

5.1 Comparisons

Test set. As no public benchmark focuses on complex dynamic 4D generation, we construct *Helix4DBench*, a 52-video test set covering morphing, emerging objects, shattering, transparent and translucent objects, and volumetric phenomena such as smoke and fire. We source still images from publicly available Trellis2 [31] examples, and animate each into a 16-frame video using Wan2.2 [27], and remove backgrounds with rembg; full construction details are provided in the supplementary. To compare against prior video-to-4D approaches on their benchmark, we evaluate on ActionBench [22] and DAVIS [17].

Baselines. We compare against five recent methods for dynamic 3D generation on 16-frame videos: SS4D [13], ShapeGen4D [36], Mesh4D [7], Motion 3-to-4 [2] and ActionMesh [22]. Some baselines do not output texture (ActionMesh, ShapeGen4D), and we mark texture-dependent metrics as '–' for

Table 1: **Quantitative comparison on our Helix4DBench.** Ours outperforms prior work across all reported metrics, including 3D alignment, video quality, temporal consistency, and pairwise user preference. Preference reports the 1-on-1 win rate of ours against each baseline.

Method	Alignment				Video Quality		Preference
	CLIP↑	CLIP-N↑	ULIP-2↑	Uni3D↑	DreamSim↓	FVD↓	Ours win rate↑
SS4D [13]	0.8468	–	0.4191	0.3739	0.2745	1046	**71.2%**
ShapeGen4D [36]	–	0.6876	0.4169	0.3687	–	–	**67.9%**
Mesh4D* [7]	0.8533	0.7399	0.4353	0.3691	0.3025	979	**79.8%**
Motion 3-to-4 [2]	0.8674	0.7442	0.4331	0.3768	0.2697	955	**80.7%**
ActionMesh [22]	–	0.7156	0.4226	0.3761	–	–	**89.8%**
Ours	**0.8723**	**0.7569**	**0.4600**	**0.4063**	**0.2593**	**950**	Ref.

these methods. Mesh4D* supports only six output frames; therefore, for a fair comparison with our 16-frame setting, we uniformly sample six frames from each generated sequence.

Evaluation metrics. Our Helix4DBench is constructed by animating still images with Wan2.2 [27] and therefore has no ground-truth geometry. We follow Trellis2 [31] for static-frame quality and add temporal-consistency metrics for the 4D setting. When ground-truth geometry is available (*e.g.*, on ActionBench [22]), we report Chamfer distance following ActionMesh [22], as introduced later in this section. CLIP [19] and CLIP-N measure appearance and geometry quality, respectively, computed as the similarity between rendered views (or normal maps) at azimuths $\{0°, 90°, 180°, 270°\}$ and the ground-truth video frames. ULIP-2 [34] and Uni-3D [47] measure 3D-image alignment by sampling 10K points from the mesh surface via farthest-point sampling and computing similarity to the groundtruth image. Baselines without textures are assigned a uniform gray color for fairness. DreamSim [4] is a learned perceptual similarity metric that aligns with human judgments on pose, layout, color, and semantics differences. We apply it per-frame between rendered and ground-truth images and average across frames. FVD [26] measures the temporal consistency of the rendered images. We also conduct a user study. Participants compare our result against one baseline in a pairwise A/B setting and select the better result in terms of overall visual quality, geometric detail, and temporal consistency. We report the win rate of ours against each baseline.

Quantitative comparisons on Helix4DBench. Tab. 1 reports results on our benchmark. Our method outperforms all baselines across every metric, indicating both higher per-frame quality and stronger temporal consistency.

These quantitative gains are consistent with the qualitative results in Fig. 5, which shows three modes that distinguish our approach from ActionMesh [22], Motion 3-to-4 [2], and Mesh4D [7]:

(i) Emerging objects. They all treat the first frame as an anchor and build motion on top of an initial mesh. They therefore fail whenever new content appears in later frames. The first example in Fig. 5 shows paint splashing onto and partially melting the figure. The content that is absent in the first frame is thus unrecoverable for anchor-based baselines.

(ii) Vertex sticking. Even when an object is present in the first frame, anchor-based methods reconstruct it as a single mesh and propagate vertices through time. When the object needs to be separated later, this leads to a vertex-sticking problem as visible in the farmer-with-vegetable-basket example: the basket fuses with the hands and distorts as the figure moves.

(iii) Inner surfaces. Baselines rely on a watertight-mesh assumption and therefore cannot represent inner surfaces. The fish-in-bottle example illustrates this: only our method reconstructs the transparent shell together with the fish swimming inside.

Quantitative comparisons on ActionBench and TexVerse test set. We follow ActionMesh [22] for geometric accuracy, given a predicted mesh sequence and the ground-truth point clouds,

we sample dense point clouds from the predicted meshes and align them to the ground truth via Iterative Closest Point (ICP). CD-3D computes the average per-frame Chamfer distance after per-frame ICP alignment, measuring shape accuracy. CD-4D estimates a single ICP alignment from the first frame and applies it uniformly to every subsequent frame, measuring both shape accuracy and temporal consistency.

Note, ActionBench contains mainly simple motion, non-topology-changing sequences, the settings

Table 2: **Quantitative comparison on ActionBench and TexVerse test set.** Ours achieves the lowest CD-3D on both benchmarks, and the lowest CD-4D on the TexVerse set.

Method	ActionBench		TexVerse Test Set	
	CD-3D ↓	CD-4D ↓	CD-3D ↓	CD-4D ↓
SS4D [13]	0.105	0.120	0.052	0.103
ShapeGen4D [36]	0.056	0.170	0.075	0.129
Mesh4D* [7]	0.096	0.141	0.058	0.100
Motion 3-to-4 [2]	0.068	0.114	0.056	0.108
ActionMesh [22]	0.053	**0.081**	0.056	0.111
Ours	**0.051**	0.093	**0.045**	**0.097**

Table 3: **Quantitative comparison on real-world videos from DAVIS.** We evaluate on alignment and video quality metrics. Our method achieves the best performance on five out of six metrics.

Method	Alignment				Video Quality	
	CLIP↑	CLIP-N↑	ULIP-2↑	Uni3D↑	DreamSim↓	FVD↓
Mesh4D* [7]	0.8673	0.7656	0.3630	0.3144	0.3115	995
Motion 3-to-4 [2]	0.8689	0.7707	0.3633	0.3150	**0.2968**	856
ActionMesh [22]	–	0.7649	0.3461	0.3055	–	–
Ours	**0.8691**	**0.7784**	**0.3677**	**0.3274**	0.2994	**747**

ActionMesh is specifically designed for. Hence, we also evaluate on a held-out subset of TexVerse-1K, which has broader dynamic scenarios. We randomly sample 32 objects from this held-out subset and use the ground-truth meshes for CD evaluation. Tab. 2 reports CD-3D and CD-4D on both benchmarks. On ActionBench, our method achieves the lowest CD-3D, indicating the highest per-frame geometry accuracy. On our harder TexVerse, our method achieves the best results on *both* CD-3D and CD-4D, confirming that the gains generalize to sequences with topological and material complexity that existing methods struggle to handle. Baselines exhibit vertex sticking and fail to follow the motion when topology changes (*e.g.*, the opening door fuses shut); qualitative comparisons provided in Fig. A1.

Quantitative comparisons on DAVIS. DAVIS is a video object segmentation benchmark consisting of real-world footage with relatively simple object motions (*e.g.*, cars moving, animals walking). We segment the object using the DAVIS ground-truth segmentation masks and select the first 16 frames per clip. We then evaluate Mesh4D [7], Motion 3-to-4 [2], ActionMesh [22], and our method using the same inputs and evaluation protocol as in Tab. 1 except the user study. As shown in Tab. 3, our method achieves the best performance on five of the six metrics and is comparable to the strongest baseline on DreamSim (0.2994 vs. 0.2968). The improvements are smaller than on *Helix4DBench* because DAVIS mostly contains simple motions with largely fixed topology. This is the setting in which deformation-based baselines are designed to perform well.

5.2 Ablation

We analyze each design choice by removing one component at a time from the second stage (mesh) model. Results on our held-out TexVerse-1k benchmark are reported in Tab. 4. We observe that all three components contribute to the generation quality. The drop is largest when the first-frame conditioning is removed, consistent with our motivation that the first-frame anchor makes the 4D generation task easier. Removing 4D rotary embedding by replacing it with 1D temporal RoPE applied on top of the 3D spatial RoPE causes the model to confuse tokens at different time steps, and this hurts both geometric and semantic accuracy.

The numerical gap on this row is modest because we believe the metrics saturate, but visual inspection in Fig. 6 shows the cost clearly: without the 4D rotary embedding, reconstructions exhibit an extra arm and a rough back surface, whereas our model produces clean geometry and correct articulation. The *w/o anchor attention* row replaces our sliding-window-plus-anchor pattern with full self-attention. Full attention has too much unnecessary context, while restricting attention to a local window plus a

Table 4: **Component ablation.** The full model is best across all metrics.

Variant	CD-3D ↓	ULIP-2 ↑	Uni3D ↑
w/o first-frame cond.	0.0488	0.2809	0.2554
w/o 4D rotary	0.0478	0.2819	0.2584
w/o our attention	0.0467	0.2819	0.2530
Ours	**0.0464**	**0.2881**	**0.2589**

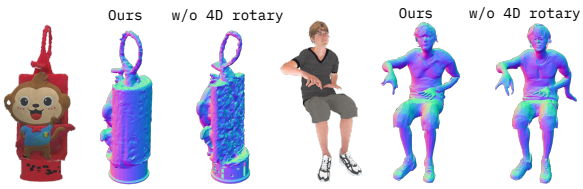


Figure 6: **Qualitative comparisons of the 4D rotary embedding.** Compared with our full model, removing 4D rotary leads to degraded geometry, including rough back surfaces and extra arms.

global first-frame anchor matches the structure of the task: short-range motion is captured locally, and the anchor maintains a consistent global identity.

Next, we compare five attention patterns: *full* attention across all F frames, *causal* attention, where frame f attends only to frame $1, \ldots, f$, *sliding* window, where frame f attends to $[f - w/2, f + w/2]$, spatial, where N^3 voxel grid is partitioned into $8 \times 8 \times 8$ blocks and tokens attend across time only within their block, and our is *sliding window + anchor*, which augments the local window with all tokens from the first frame. Tab. 5 shows that our pattern achieves the best score on all four quality metrics while remaining competitive in cost. Full attention is the most expensive yet underperforms ours, indicating that full attention injects noise rather than than a useful signal. A plain sliding window gives the worst CD-4D, since each frame can drift away from the initial geometry without a stable global reference. Spatial attention is the cheapest overall but caps quality because tokens cannot exchange information across blocks. Our sliding window + anchor combines the locality of sliding window with a single global anchor, recovering temporal consistency.

Table 5: **Cross-frame attention pattern.** Time is wall-clock normalized to ours. Sliding window + anchor (ours) is best on every metric while running $2.3\times$ faster than full attention.

	CD-3D ↓	CD-4D ↓	ULIP-2 ↑	Uni3D ↑	Time ↓
Full attention	0.0470	0.1009	0.2835	0.2447	$2.3\times$
Causal attention	0.0491	0.1038	0.2838	0.2564	$1.5\times$
Sliding window	0.0468	0.1085	0.2870	0.2538	$0.9\times$
Spatial attention	0.0471	0.1011	0.2791	0.2548	**0.4**$\times$
Ours	**0.0454**	**0.0970**	**0.2879**	**0.2576**	$1.0\times$

6 Conclusion

We presented Helix4D, an approach that adapts a pretrained image-to-3D generator into a video-to-4D mesh generator. We introduce a lightweight conversion of Trellis2 to dynamic generation through temporal ReRoPE, sliding-window cross-frame attention with an anchor frame, and first-frame conditioning. This design preserves the strong geometry and material prior of O-Voxels while adding temporal consistency, allowing the model to handle topology changes, transparent objects, thin structures, and inner surfaces. Experiments show strong geometric accuracy on ActionBench and consistent gains over recent video-to-4D baselines. This is important for the community because it shows that strong static 3D foundation models can be lifted to 4D without training a new model from scratch. Future work aims to extend Helix4D to longer videos, scene-level interactions, stronger camera motion, and more accurate physical dynamics.

7 Acknowledgment

We thank Ashkan Mirzaei for valuable discussions, and Vladislav Shakhray and Gleb Dmukhin for their help with Blender.

References

[1] Sherwin Bahmani, Ivan Skorokhodov, Victor Rong, Gordon Wetzstein, Leonidas Guibas, Peter Wonka, Sergey Tulyakov, Jeong Joon Park, Andrea Tagliasacchi, and David B. Lindell. 4D-fy: Text-to-4D generation using hybrid score distillation sampling. In *Proc. CVPR*, 2024. 2, 3

[2] Hongyuan Chen, Xingyu Chen, Youjia Zhang, Zexiang Xu, and Anpei Chen. Motion 3-to-4: 3D motion reconstruction for 4D synthesis. *arXiv preprint arXiv:2601.14253*, 2026. 2, 3, 6, 7, 8, 9, 15, 19, 25

[3] Jianqi Chen, Biao Zhang, Xiangjun Tang, and Peter Wonka. V2M4: 4D mesh animation reconstruction from a single monocular video. In *Proc. ICCV*, 2025. 3

[4] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. DreamSim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023. 8

[5] Team Hunyuan3D, Shuhui Yang, Mingxin Yang, Yifei Feng, Xin Huang, Sheng Zhang, Zebin He, Di Luo, Haolin Liu, Yunfei Zhao, et al. Hunyuan3D 2.1: From images to high-fidelity 3D assets with production-ready PBR material. *arXiv preprint arXiv:2506.15442*, 2025. 3

[6] Yanqin Jiang, Li Zhang, Jin Gao, Weimin Hu, and Yao Yao. Consistent4D: Consistent 360° dynamic object generation from monocular video. *arXiv preprint arXiv:2311.02848*, 2023. 3

[7] Zeren Jiang, Chuanxia Zheng, Iro Laina, Diane Larlus, and Andrea Vedaldi. Mesh4D: 4D mesh reconstruction and tracking from monocular video. *arXiv preprint arXiv:2601.05251*, 2026. 2, 3, 6, 7, 8, 9, 15, 19, 25

[8] Yudong Jin, Sida Peng, Xuan Wang, Tao Xie, Zhen Xu, Yifan Yang, Yujun Shen, Hujun Bao, and Xiaowei Zhou. Diffuman4D: 4D consistent human view synthesis from sparse-view videos with spatio-temporal diffusion models. In *Proc. ICCV*, 2025. 2, 3

[9] Zeqiang Lai, Yunfei Zhao, Haolin Liu, Zibo Zhao, Qingxiang Lin, Huiwen Shi, Xianghui Yang, Mingxin Yang, Shuhui Yang, Yifei Feng, et al. Hunyuan3D 2.5: Towards high-fidelity 3D assets generation with ultimate details. *arXiv preprint arXiv:2506.16504*, 2025. 3

[10] Chunyang Li, Yuanbo Yang, Jiahao Shao, Hongyu Zhou, Katja Schwarz, and Yiyi Liao. ReRoPE: Repurposing RoPE for relative camera control. *arXiv preprint arXiv:2602.08068*, 2026. 2, 4, 5

[11] Weiyu Li, Xuanyang Zhang, Zheng Sun, Di Qi, Hao Li, Wei Cheng, Weiwei Cai, Shihao Wu, Jiarui Liu, Zihao Wang, et al. Step1X-3D: Towards high-fidelity and controllable generation of textured 3D assets. *arXiv preprint arXiv:2505.07747*, 2025. 3

[12] Yangguang Li, Zi-Xin Zou, Zexiang Liu, Dehu Wang, Yuan Liang, Zhipeng Yu, Xingchao Liu, Yuan-Chen Guo, Ding Liang, Wanli Ouyang, et al. TripoSG: High-fidelity 3D shape synthesis using large-scale rectified flow models. *TPAMI*, 2025. 3

[13] Zhibing Li, Mengchen Zhang, Tong Wu, Jing Tan, Jiaqi Wang, and Dahua Lin. SS4D: Native 4D generative model via structured spacetime latents. *TOG*, 2025. 2, 3, 6, 7, 8, 9, 15, 25

[14] Huan Ling, Seung Wook Kim, Antonio Torralba, Sanja Fidler, and Karsten Kreis. Align your Gaussians: Text-to-4D with dynamic 3D Gaussians and composed diffusion models. In *Proc. CVPR*, 2024. 2

[15] Tianqi Liu, Zihao Huang, Zhaoxi Chen, Guangcong Wang, Shoukang Hu, Liao Shen, Huiqiang Sun, Zhiguo Cao, Wei Li, and Ziwei Liu. Free4D: Tuning-free 4D scene generation with spatial-temporal consistency. In *Proc. ICCV*, 2025. 2, 3

[16] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7

[17] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 724–732, 2016. 6, 7

[18] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 3

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. ICML*, 2021. 8

[20] Jiawei Ren, Liang Pan, Jiaxiang Tang, Chi Zhang, Ang Cao, Gang Zeng, and Ziwei Liu. DreamGaussian4D: Generative 4D gaussian splatting. *arXiv preprint arXiv:2312.17142*, 2023. 3

[21] Jiawei Ren, Kevin Xie, Ashkan Mirzaei, Hanxue Liang, Xiaohui Zeng, Karsten Kreis, Ziwei Liu, Antonio Torralba, Sanja Fidler, Seung W Kim, et al. L4GM: Large 4D Gaussian reconstruction model. In *Proc. NeurIPS*, 2024. 3

[22] Remy Sabathier, David Novotny, Niloy J Mitra, and Tom Monnier. ActionMesh: Animated 3D mesh generation with temporal 3D diffusion. In *Proc. CVPR*, 2026. 2, 3, 6, 7, 8, 9, 15, 19, 25

[23] Uriel Singer, Shelly Sheynin, Adam Polyak, Oron Ashual, Iurii Makarov, Filippos Kokkinos, Naman Goyal, Andrea Vedaldi, Devi Parikh, Justin Johnson, et al. Text-to-4D dynamic scene generation. *arXiv preprint arXiv:2301.11280*, 2023. 3

[24] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 2024. 2, 5

[25] Wenqiang Sun, Shuo Chen, Fangfu Liu, Zilong Chen, Yueqi Duan, Jun Zhu, Jun Zhang, and Yikai Wang. DimensionX: Create any 3D and 4D scenes from a single image with decoupled video diffusion. In *Proc. ICCV*, 2025. 2, 3

[26] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. FVD: A new metric for video generation. 2019. 8

[27] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 7, 8, 15, 25

[28] Yikai Wang, Xinzhou Wang, Zilong Chen, Zhengyi Wang, Fuchun Sun, and Jun Zhu. Vidu4D: Single generated video to high-fidelity 4D reconstruction with dynamic gaussian surfels. In *Proc. NeurIPS*, 2024. 3

[29] Rundi Wu, Ruiqi Gao, Ben Poole, Alex Trevithick, Changxi Zheng, Jonathan T. Barron, and Aleksander Holynski. CAT4D: Create anything in 4D with multi-view video diffusion models. In *Proc. CVPR*, 2025. 2, 3

[30] Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Yikang Yang, Yajie Bao, Jiachen Qian, Siyu Zhu, Xun Cao, Philip Torr, et al. Direct3D-S2: Gigascale 3D generation made easy with spatial sparse attention. *arXiv preprint arXiv:2505.17412*, 2025. 3

[31] Jianfeng Xiang, Xiaoxue Chen, Sicheng Xu, Ruicheng Wang, Zelong Lv, Yu Deng, Hongyuan Zhu, Yue Dong, Hao Zhao, Nicholas Jing Yuan, et al. Native and compact structured latents for 3D generation. *arXiv preprint arXiv:2512.14692*, 2025. 2, 3, 5, 6, 7, 8, 25

[32] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3D latents for scalable and versatile 3D generation. In *Proc. CVPR*, 2025. 3

[33] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3D latents for scalable and versatile 3D generation. In *Proc. CVPR*, 2025. 3

[34] Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. ULIP: Learning a unified representation of language, images, and point clouds for 3D understanding. In *Proc. CVPR*, 2023. 2, 8

[35] Chun-Han Yao, Yiming Xie, Vikram Voleti, Huaizu Jiang, and Varun Jampani. SV4D 2.0: Enhancing spatio-temporal consistency in multi-view video diffusion for high-quality 4D generation. In *Proc. ICCV*, 2025. 2, 3

[36] Jiraphon Yenphraphai, Ashkan Mirzaei, Jianqi Chen, Jiaxu Zou, Sergey Tulyakov, Raymond A Yeh, Peter Wonka, and Chaoyang Wang. ShapeGen4D: Towards high quality 4D shape generation from videos. In *Proc. ICLR*, 2026. 2, 3, 6, 7, 8, 9, 15, 25

[37] Minghao Yin, Wenbo Hu, Jiale Xu, Ying Shan, and Kai Han. Sculpt4D: Generating 4D shapes via sparse-attention diffusion transformers. *arXiv preprint arXiv:2604.21592*, 2026. 3

[38] Yuyang Yin, Dejia Xu, Zhangyang Wang, Yao Zhao, and Yunchao Wei. 4DGen: Grounded 4D content generation with spatial-temporal consistency. *arXiv preprint arXiv:2312.17225*, 2023. 3

[39] Ye Yuan, Xueting Li, Yangyi Huang, Shalini De Mello, Koki Nagano, Jan Kautz, and Umar Iqbal. GAvatar: Animatable 3D Gaussian avatars with implicit mesh learning. In *Proc. CVPR*, 2024. 2, 3

[40] Yifei Zeng, Yanqin Jiang, Siyu Zhu, Yuanxun Lu, Youtian Lin, Hao Zhu, Weiming Hu, Xun Cao, and Yao Yao. STAG4D: Spatial-temporal anchored generative 4D gaussians. In *Proc. ECCV*, 2024. 3

[41] Biao Zhang, Jiapeng Tang, Matthias Niessner, and Peter Wonka. 3DShape2VecSet: A 3D shape representation for neural fields and generative diffusion models. *TOG*, 2023. 3

[42] Bowen Zhang, Sicheng Xu, Chuxin Wang, Jiaolong Yang, Feng Zhao, Dong Chen, and Baining Guo. Gaussian variation field diffusion for high-fidelity video-to-4D synthesis. *arXiv preprint arXiv:2507.23785*, 2025. 2, 3

[43] Haiyu Zhang, Xinyuan Chen, Yaohui Wang, Xihui Liu, Yunhong Wang, and Yu Qiao. 4Diffusion: Multi-view video diffusion model for 4D generation. In *Proc. NeurIPS*, 2024. 3

[44] Yibo Zhang, Li Zhang, Rui Ma, and Nan Cao. Texverse: A universe of 3D objects with high-resolution textures. *arXiv preprint arXiv:2508.10868*, 2025. 6, 14, 25

[45] Zibo Zhao, Zeqiang Lai, Qingxiang Lin, Yunfei Zhao, Haolin Liu, Shuhui Yang, Yifei Feng, Mingxin Yang, Sheng Zhang, Xianghui Yang, et al. Hunyuan3D 2.0: Scaling diffusion models for high resolution textured 3D assets generation. *arXiv preprint arXiv:2501.12202*, 2025. 3

[46] Yufeng Zheng, Xueting Li, Koki Nagano, Sifei Liu, Otmar Hilliges, and Shalini De Mello. A unified approach for text- and image-guided 4D scene generation. In *Proc. CVPR*, 2024. 2, 3

[47] Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3D: Exploring unified 3D representation at scale. *arXiv preprint arXiv:2310.06773*, 2023. 2, 8

Appendix

Table A1: **Ablation on the RoPE ratio.** We evaluate on 32 held-out objects from TexVerse [44]. We use the full spatial RoPE setting (ratio $= 1.0$) as the reference and report the CD-3D between the mesh generated by each ratio and this reference.

	0.0	**0.2**	**0.4**	**0.6**	**0.8**	**1.0**
CD-3D ↓	0.0181	0.0098	0.0079	0.0079	0.0079	0

A1 Limitations or failure cases

Our method inherits two limitations from the Trellis2 backbone we built upon. First, the generated meshes occasionally contain holes, since Trellis2 does not enforce a watertight-mesh assumption and provides no guarantee that the output surface will be closed. Second, the textures produced by Trellis2 exhibit a color-shifting artifact, which could result in the generation of metallic surfaces for transparent objects.

Our own design introduces another limitation: because we output a mesh sequence rather than a single static mesh, the results sometimes lack temporal consistency, particularly in regions with high-frequency geometric details.

Although our sliding-window attention with a first-frame anchor scales linearly with the number of frames, attention remains quadratic in the number of O-Voxel tokens per frame, making spatial resolution the primary computational bottleneck. On a single A100 GPU, increasing the sequence length from 8 to 32 frames increases peak memory usage from 12.9 GB to 30.4 GB and inference time from 90 to 368 seconds per asset for the geometry generation stage.

A2 User Study Details

We conduct a user study to evaluate the perceived quality of our generated dynamic 3D reconstructions against prior methods. Participants compare pairwise rendered videos and select which result looks better overall, considering visual quality, temporal consistency, and preservation of details.

A2.1 Instructions Shown to Participants

Participants were shown the following task description:

> **4D Reconstruction.** In this study, you will compare pairs of rendered 4D reconstruction results. Each question shows a single video containing: an input video, Result A, and Result B. Your task is to decide which generated result, A or B, looks better overall.

Participants were instructed to consider the following criteria:

1. **Overall visual quality:** which result looks more realistic and visually appealing.
2. **Temporal consistency:** which result has smoother and more coherent motion over time.
3. **Preservation of details:** which result better preserves fine details.

Each question used the following prompt:

> Watch the input video and the two generated results, A and B. Which result looks better overall?

The available answer choices were:

- Result A is better.
- Result B is better.
- Tie / no clear difference.

Participants were told that there were 25 questions in total, that they could go back to change previous answers, and that the study would take approximately 10 minutes.

A2.2 Participants and Comparisons

Table A2: **User study summary.**

Statistic	Value
Total unique participant IDs	21
Participants completing all 25 questions	18
Partial participants	3
Total response rows collected	458
Questions per participant	25
RGB comparisons	15
Normal-map comparisons	10

The study contains 25 pairwise comparison questions per participant. Among these, 15 questions evaluate RGB rendered appearance, while 10 questions evaluate normal-map renderings that emphasize surface geometry. In the RGB category, we compare against Motion 3-to-4 [2], Mesh4D [7], and SS4D [13]. In the normal-map category, we compare against ActionMesh [22] and ShapeGen4D [36]. The question order was randomized independently for each participant. Responses were collected anonymously.

We compute the win rate as follows:

$$\text{WinRate} = \frac{W + 0.5T}{N} \tag{A8}$$

where W, L, N are the number of wins, losses, and ties for our method.

A3 Helix4DBench Construction

Our test set is constructed from the example images released in the official Trellis2 repository. We use each image as a static object input and generate a video using Wan2.2 [27] with the corresponding text prompt in Tab. A3. Each prompt shares the same prefix:

> *Fixed camera, pure white background, no camera movement or rotation.*

We therefore omit this shared prefix from the table for compactness and report only the motion-specific prompt suffix. For each image filename, we report only its first eight characters, i.e., `filename[:8]`; for the special filename `1.png`, we report the prefix as `1`.

Table A3: **Image IDs and their prompts.** The image ID column shows the filename prefix which is the first eight characters of the original image filename.

Image ID	Prompt suffix
`2bb09323`	A steampunk mechanical pig statue where the rock base stays fixed while the pig springs to life, galloping vigorously in place with large exaggerated leg strides, pistons pumping hard, steam bursting from joints, and glowing parts pulsing brightly.
`4dae7ef0`	A medieval wooden catapult where the base stays grounded while the throwing arm swings forcefully upward in a full arc, launching a projectile, with chains whipping and ropes snapping taut with dramatic motion.
`5c80e5e0`	A fierce bear rearing up on its hind legs, swiping both front claws aggressively through the air, roaring with its jaw wide open, and twisting its torso side to side with powerful exaggerated motion.
`5a6c81d3`	A detailed hiking backpack where zippers rapidly unzip, multiple compartments swing open wide, straps flap loosely, and the whole bag bounces as if being shaken out and unpacked with energetic motion.

Image ID	Prompt suffix
7b540da3	A robed woman turning her full torso dramatically from side to side, arms sweeping outward, robes billowing and flowing with wide expressive gestures and dynamic cloth movement.
8ce83f6a	A stylized mermaid swaying her torso and arms expressively while her long hair flows and whips around, as glowing jellyfish rise and drift around her with large undulating tentacle motion.
7d7659d5	A mechanical suit of armor raising both arms high, clenching and unclenching its fists, twisting its torso left and right, with every joint visibly bending and articulating with bold mechanical motion.
8e12cf09	A retro mechanical robot waving both arms up and down, bending at the waist, turning its head left and right, with gears spinning and joints clicking in large expressive movements. The viewpoint does not change.
7bd0521d	A grotesque goblin bust where the face contorts dramatically, jaw dropping wide open, tongue thrashing out, eyes bulging, and the head tilting and shaking with exaggerated snarling facial motion.
9c306c7b	A stylized traveler marching energetically in place with high knee lifts, arms swinging wide, backpack bouncing, and gear swaying with lively exaggerated walking motion.
50b70c5f	A Viking warrior swinging his axe in a wide arc overhead, the lantern swaying broadly on its chain, his torso twisting with the swing, cape and armor shifting with strong dynamic motion.
61fea9d0	A stylized archer pulling the bowstring all the way back with full arm extension, torso twisting into the draw, then releasing the arrow with a dramatic snap and follow-through motion.
80ad7988	A stylized armored figure raising the golden chalice high overhead with a full arm lift, then bringing it down and tilting it forward as if pouring, with the other arm gesturing outward expressively.
95db3c13	A quadruped battle robot shifting its weight between legs, turret swiveling rapidly back and forth, legs stomping and repositioning with heavy mechanical steps and visible hydraulic motion.
154c8867	A stylized potted plant where stems shoot upward rapidly, leaves unfurl and spread wide, colorful flowers burst into full bloom, and vines curl and extend outward with fast energetic growing motion.
454e7d8a	A stylized human lifting and turning the small robot in his hands, the robot flailing its arms and kicking its legs, while the human leans and shifts his body to keep hold of it with lively playful motion.
901d8de4	A glass bottle aquarium where fish dart and chase each other in fast circles, bubbles stream upward rapidly, and aquatic plants sway dramatically with strong swirling water motion inside the bottle.
3903b879	A humanoid robot marching in place with high exaggerated steps, swinging its arms wide, turning its torso and head dramatically from side to side with bold mechanical motion.
26717a7d	A large tropical tree where branches burst outward rapidly, leaves unfurl and flutter vigorously in the wind, vines whip and spiral dynamically, and the entire canopy sways with strong dramatic motion.
39488b45	A marble bust where the face dramatically comes to life, eyes widening, mouth opening and closing as if speaking passionately, head turning left and right with bold expressive facial motion.
52284bf4	A marble angel statue where the wings spread wide open in a dramatic full extension, arms lifting outward, and the entire upper body shifting into a new majestic pose with sweeping lifelike motion.
65433d02	A vintage typewriter where keys hammer down rapidly in quick succession, the carriage slides across with a snap, the return lever slams back, and paper feeds upward fast with energetic clacking motion.
3723615e	A stylized witch leaping high into the air, arms spread wide, hat flying off her head, hair whipping upward, hanging bottles and charms swinging wildly with exaggerated bouncy motion.

Image ID	Prompt suffix
a13d176c	An ornate steampunk pistol where gears spin rapidly, the hammer cocks back and snaps forward, the slide racks with a sharp motion, and small valves and pistons pump with fast visible mechanical action.
1	A steampunk mechanical device where gears spin rapidly at different speeds, pistons pump hard with visible force, glowing fluid rushes through transparent tubes, and steam bursts from valves with energetic continuous motion.
f94e2b76	A stylized turkey character with a jetpack performing a playful leap upward, legs tucking in, wings and feathers bouncing wildly, and the jetpack flaring with bright exhaust bursts and lively energetic motion.
f5332118	A stone king statue rising forcefully from a seated throne to a full standing position, gripping his sword tightly, with heavy weighty motion, stone cracking and shifting, and dust particles falling off.
e4d6b2f3	A stylized rider on a stationary motorcycle turning his torso toward the viewer and raising a hand high to wave enthusiastically with bold upper-body motion and natural arm swing.
e1046572	A stylized farmer dropping a basket from his hands, the basket tumbling downward, while his facial expression shifts dramatically from calm to wide-eyed surprise with exaggerated body recoil.
cdf996a6	A fierce goblin warrior swinging its weapon in wide aggressive arcs, torso twisting with each swing, feet shifting for balance, with powerful fluid slashing motion and visible follow-through.
c3d714bc	A monstrous treasure chest where the base stays still while the massive mouth slowly closes with sharp teeth interlocking, the tongue retracting inward, and drool dripping with smooth organic motion.
c2125d08	A wooden windmill where the structure remains still while the turbine blades spin smoothly and continuously with steady natural motion and subtle creaking sway.
6b6d89d4	A pair of rugged leather boots performing a rhythmic tango-style tap dance in place with precise energetic footwork, heels clicking, toes tapping, and bouncing with lively motion.
0e4984a9	A brick wall where leaves and tendrils rapidly grow and spread outward, curling and expanding to fully cover the surface with fast organic creeping motion and unfurling vines.
290af2dd	A golden armored helmet where the metal slowly melts and deforms downward with soft fluid motion, edges dripping and pooling naturally, gold surface warping and sagging dramatically.
ab3bb3e1	An hourglass where sand pours smoothly and continuously from the top chamber to the bottom, grains cascading and piling up with visible flowing motion.
c9340e74	A bat perched on a tomb lifting off and flapping its wings vigorously to fly upward with strong wing beats, body rising and legs releasing their grip with dramatic takeoff motion.
0f168a4b	A mounted turret firing rapid glowing energy shots forward, the upper gun assembly swiveling and tracking targets, muzzle flashing brightly, with the base remaining completely static.
25d412fe	Potion bottles shattering and exploding outward with colorful liquid splashing in all directions, glass shards flying, and vibrant fluid arcing through the air with dramatic explosive motion.
7baa867b	An ornate chair tipping over sideways and crashing to the ground with a heavy fall, legs bouncing on impact, and the whole frame rocking with dramatic toppling motion.
bb319089	A truck house where doors and windows swing open wide outward with strong creaking motion, shutters flapping, hinges stretching, and the whole structure revealing its interior dramatically.
a306e2ee	A ship violently breaking apart, hull splintering outward, masts snapping and falling, wood fragments and planks flying in all directions with dramatic destructive wrecking motion.

Image ID	Prompt suffix
ee8ecf65	A stone angel slowly spreading its wings wide open in a full dramatic extension, feathers fanning outward, arms lifting, and the pose shifting with majestic lifelike motion.
51b1b31d	A small castle suddenly igniting as fire breaks out and spreads rapidly across the structure, flames licking upward, smoke billowing, and embers flying with intense burning motion.
f351569d	A stone golem slowly bending its massive knees and lowering itself down to sit, stone grinding and cracking, dust falling from joints, with heavy deliberate motion.
a3d0c28c	A mech activating and firing its arm-mounted weapon with bright glowing energy blasts, recoil shaking the arm, muzzle flashing, and spent casings ejecting with powerful shooting motion.
f8a7eafe	A humanoid mech gripping its rifle firmly and firing rapid glowing shots forward, shoulders bracing against recoil, barrel flashing, with bold aggressive shooting motion.
dd4c51c1	A police robot performing energetic dance moves, shifting its weight side to side, swinging its arms rhythmically, head bobbing, and hips swaying with lively exaggerated dance motion.
be7deb26	A princess character twirling gracefully with her flowing dress fanning outward wide, butterflies fluttering around her, arms outstretched, hair swirling with elegant spinning dance motion.
b358d0eb	Colorful fish swimming actively around a bowl with energetic fin movements, darting and chasing each other, bubbles rising rapidly with lively aquatic motion.
cd3c309f	A dragon head on armor breathing a powerful burst of glowing fire forward, flames roaring and flickering intensely, heat shimmer distorting the air, jaws opening wide with fierce dramatic motion.
fdf979f5	A potted plant rapidly growing taller with stems shooting upward, leaves and colorful flowers sprouting and expanding quickly, branches spreading outward with fast energetic growing motion.

A4 TexVerse test set

We construct a fixed TexVerse test set of 32 held-out assets from the TexVerse-1K dataset. The assets are obtained from the official TexVerse Hugging Face repository: https://huggingface.co/datasets/YiboZhang2001/TexVerse. Table A4 lists the corresponding asset paths in the repository.

A5 Societal Impacts

Our work reduces the barrier to creating high-quality 4D generation, which can benefit creators in animation, gaming, and VR/AR applications.



Figure A1: **Qualitative comparison on TexVerse test set.** The left column shows input video frames. Each pair of rows shows two frames from the generated sequence, with normal renderings from ground truth and each method. Baselines (ActionMesh [22], Motion 3-to-4 [2], Mesh4D [7]) struggle with topology changes: the opening door fuses shut, character limbs and capes stick to the body, and motion fails to propagate to the second frame. Our method follows the ground-truth motion and preserves fine geometry across frames.

Table A4: **TexVerse test set.** List of the 32 held-out TexVerse-1K assets used in our experiments.

#	Asset path
1	glbs/glbs_1k/000-047/ef95775fcbae453a924071ff4bf3acb1_1024.glb
2	glbs/glbs_1k/000-047/ed43cd02f71346ec9313429638db6af7_1024.glb
3	glbs/glbs_1k/000-046/eab7714de6d9488cafe6bfba15a1349b_1024.glb
4	glbs/glbs_1k/000-074/d9d84b1f6bed433eb69b6ab6d16c3bc5_1024.glb
5	glbs/glbs_1k/000-043/dafeb1c2ec924298a2503a8d03150f3a_1024.glb
6	glbs/glbs_1k/000-040/c867839965ac44dc975e48f849c03159_1024.glb
7	glbs/glbs_1k/000-038/bddeaa8febea4481adb8fb449e372f1e_1024.glb
8	glbs/glbs_1k/000-043/d8e3f050f3e4453bb7adf5b0aae229f3_1024.glb
9	glbs/glbs_1k/000-085/b9d748ccc31344e2966d3f8176aed656_1024.glb
10	glbs/glbs_1k/000-037/bb66a60bf1884d2a930d2bdf2e342fbf_1024.glb
11	glbs/glbs_1k/000-030/99d607aef15340a3bfd253b99089d676_1024.glb
12	glbs/glbs_1k/000-029/91bf7ee81767422fae8ef8cbc843bf1c_1024.glb
13	glbs/glbs_1k/000-027/8a383eb03ef24d4f8df3b68aa121cf9d_1024.glb
14	glbs/glbs_1k/000-021/6a1516209ee8411da12acf7b1ec3e98f_1024.glb
15	glbs/glbs_1k/000-082/6d4835a9aedc4f369580cc45db413932_1024.glb
16	glbs/glbs_1k/000-015/4dd70b77276b4cb195bb01cf269b6c0b_1024.glb
17	glbs/glbs_1k/000-016/52f8934bbe1c4da8911c615399f6a8ea_1024.glb
18	glbs/glbs_1k/000-013/427bc35ebcf34795be1c1f776cb1307a_1024.glb
19	glbs/glbs_1k/000-055/2a4d2840fc904211bab65e2c4a9f52a0_1024.glb
20	glbs/glbs_1k/000-080/1eab696ff8934d44b5343747f8cf3eaf_1024.glb
21	glbs/glbs_1k/000-003/10a8bbaa7ad648f6a5990b9923e4fa06_1024.glb
22	glbs/glbs_1k/000-002/0b80aedfcc8b4f7f99fca567a768d8a5_1024.glb
23	glbs/glbs_1k/000-052/099f508befcd4ebe9930502b48978042_1024.glb
24	glbs/glbs_1k/000-017/57d76986e57a492a865875b54351ddd7_1024.glb
25	glbs/glbs_1k/000-018/5a5bbca8e18744ebacdce8320f014a1e_1024.glb
26	glbs/glbs_1k/000-066/8ec8f8a2022d491fa0d74279c4f07304_1024.glb
27	glbs/glbs_1k/000-030/9943ce328fff4c298ecdb92d03bdebb5_1024.glb
28	glbs/glbs_1k/000-084/92e28edf31074225ad28b652b426b4f4_1024.glb
29	glbs/glbs_1k/000-034/aab888ae4c4f4e91ba053c39221ffb05_1024.glb
30	glbs/glbs_1k/000-037/ba253b4d995140aca8a968f4a86b2a40_1024.glb
31	glbs/glbs_1k/000-042/d6c86c5eeb204bfd985f7bdfb270327d_1024.glb
32	glbs/glbs_1k/000-049/f5cf2511b7ce49588ba01565d9afc687_1024.glb

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We describe our main contributions in Sec. 4 and provide detailed ablations, comparisons, and analysis in Sec. 5. The claims in the abstract and introduction are supported by these experimental results and are scoped to the settings evaluated in the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We explain our limitations in Sec. A1.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: Our paper does not include formal theoretical results, theorems, or proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides the information needed to reproduce the main results. Section 5 (Implementation Details) specifies the data source, how to process, and training details to replicate our implementation.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We are unable to release the code or trained model checkpoints due to company policy. To support reproducibility, we provide detailed descriptions of the architecture, training procedure, hyperparameters, and data curation pipeline in Section 5 and the supplementary, our Trellis4DBench is constructed entirely from publicly available assets which is detailed in Sec. A3, and our TexVerse test set is detailed in Sec. A4.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: Section 5 specifies the data curation pipeline, train/test splits, and our configuration, with additional details provided in the supplementary.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Due to the high computational cost of training each stage (32 A100 GPUs for $\sim$2 days per stage), we report single-run results without error bars. To mitigate this, we evaluate on multiple benchmark to provide a robust assessment of the relative performance across methods.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 5 (Implementation Details) reports the compute resources used for training: each of the three stages is fine-tuned on 32 A100 GPUs for 20K iterations (approximately 2 days per stage), and Tab. 5 additionally reports normalized wall-clock time.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms with the NeurIPS Code of Ethics. We don't see harmful applications or fairness/privacy concerns specific to this 4D generation method.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We discuss both positive and negative societal impacts in the supplementary.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: We do not release model weights and training code.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All existing assets used in this work, including TexVerse-1k [44], Trellis2 [31], Wan2.2 [27], and the baseline methods (SS4D [13], ShapeGen4D [36], Mesh4D [7], Motion 3-to-4 [2], ActionMesh [22]) and ActionBench [22], are properly cited.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: We do not release any new assets. Model weights and code are not released due to company policy, and our 52-video benchmark is not redistributed as a standalone asset; instead, we describe its construction in the supplementary so it can be reproduced from publicly available data.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: We conducted a user study comparing our results against baselines in a pairwise A/B setting. The full instructions given to participants and details of the study setup are provided in the supplementary.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: Our user study involves only anonymous pairwise A/B comparisons of rendered 4D outputs with no collection of personal or sensitive information.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not used as a component of our core method.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.