Sculpt4D: Generating 4D Shapes via Sparse-Attention Diffusion Transformers

Supplementary Material

1. Model Details

The network architecture is instantiated as a 21-layer Diffusion Transformer block with a hidden dimension of 2,048 and 16 attention heads, resulting in a head dimension of 128. The model processes a spatiotemporal input sequence generated by the VAE encoder [6], where each frame consists of 4,096 spatial tokens derived from a 64-channel latent input. Conditioning is provided via cross-attention to visual context embeddings with a dimensionality of 1,370. Within the attention mechanisms, we employ RMSNorm for the normalization of query and key projections to enhance training stability, while standard LayerNorm is utilized for pre-norm blocks.

To facilitate effective information propagation across the network depth, the model incorporates symmetric long-range skip connections between the shallow and deep layers. Specifically, the feature maps from the first 10 layers are cached during the forward pass. The 11th layer acts as a central processing block without skip interactions. Subsequently, the final 10 layers (layers 11 through 20) retrieve the corresponding cached features in reverse order. In these layers, the input hidden states are concatenated along the channel dimension with the retrieved features, temporarily doubling the hidden dimension to 4,096. This concatenated representation is immediately projected back to the standard 2,048 dimensions via a linear transformation and normalized before entering the attention blocks. Within each block, the architecture follows a factorized design that sequentially applies spatial self-attention, cross-attention, and Block Sparse Attention. To scale model capacity, the feed-forward networks in the final six layers are replaced by Mixture-of-Experts (MoE) [4] modules, featuring eight experts per layer with a top-2 routing strategy.

Temporal positional information is encoded using Rotary Positional Embeddings (RoPE) [5] applied exclusively to the temporal attention layers. We pre-compute sinusoidal frequencies for the frame sequence and apply rotation to the query and key tensors. To match the architectural head dimension of 128, the sine and cosine components—initially computed for half the dimension—are interleaved and duplicated along the last axis. Crucially, to maintain numerical stability during mixed-precision training, this rotational transformation is explicitly cast to and executed in 32-bit floating-point precision before the tensors are reverted to the model's native data type for subsequent attention computation.

Table A1. **Ablation study.**

	Chamfer↓	IoU↑	F-Score↑	PFLOPs
A: w/o attention sink	0.0986	0.3442	0.3375	169.8
B: Fixed stride	0.1124	0.3298	0.3306	167.1
C: Aggressive decay	0.0991	0.3420	0.3365	145.0
D: Conservative decay	0.0968	0.3454	0.3388	233.6
E: Full attention	0.0958	0.3466	0.3402	425.7
F: **Ours**	0.0972	0.3451	0.3383	186.3

2. Ablation Study on Attention Mask

To validate our Block Sparse Attention, we conducted ablation studies on its core components: the First-Frame Anchor and the Time-Decaying Sparsity mask. Our design addresses the trade-off between structural integrity and efficiency in 4D generation. First, the "First-Frame Anchor" is introduced to mitigate structural degradation by providing a constant reference for global coherence. Second, recognizing that full attention is computationally prohibitive, we hypothesize that information density decays over time: immediate neighbors require dense attention for motion dynamics, whereas distant frames provide semantic context efficiently captured by sparse connections.

To test these hypotheses, we designed two sets of comparisons. Set 1: Effectiveness of First-Frame Anchor. To assess the necessity of the anchor for sustaining global generation quality, we trained a variant (Model A) where the global connection to the first frame is removed, relying solely on the relatively sparse mask. Set 2: Impact of Stride Strategies. To investigate the optimal trade-off between motion smoothness and efficiency, we compared our proposed stride schedule against three representative alternatives: Model B (Fixed stride) applies a constant sparsity (stride=4) across all temporal distances; Model C (Aggressive decay) uses an immediate exponential decay schedule $[1, 2, 4, 8, 16, 32]$; and Model D (Conservative decay) uses a conservative schedule $[1, 1, 2, 2, 4, 4]$. Our proposed Model E (Ours) utilizes a "Delayed Exponential" schedule $[1, 1, 2, 4, 8, 16]$. We quantitatively evaluated the generation quality using Chamfer Distance, Intersection over Union (IoU), and F-Score. The results, summarized in Tab. A1, confirm our hypotheses. Comparing Model A (w/o Anchor) with our full model (Model F), we observe that removing the anchor leads to a degradation in all metrics. This indicates that without the global reference, the model struggles to achieve high geometric fidelity across the sequence. Regarding stride strategies, Model B (Fixed stride) performs the worst, suggesting that uniform sparsity fails to capture essential local motion details. Model C (Aggressive decay)

improves efficiency but suffers in geometric accuracy due to the rapid loss of local information. Model D (Conservative decay) achieves competitive results but incurs significantly higher computational overhead. Our proposed Model F achieves the best balance, delivering high IoU and F-Score comparable to the dense Step Decay strategy while maintaining the efficiency benefits of exponential sparsity. This confirms that our hybrid strategy—maintaining dense local attention for generation quality while using exponential sparsity for long-range dependencies—is optimal for 4D generation.

3. Computational Analysis

Table A2. **Computational analysis.**

Frames	PFLOPs_{sparse}	PFLOPs_{full}	$\frac{Sparse}{Full}$	$\frac{Sparse\_attn}{Full\_attn}$
8	84.5	123.2	68.6%	58.1%
16	186.3	425.7	43.8%	35.2%
32	425.0	1584.9	26.8%	21.5%

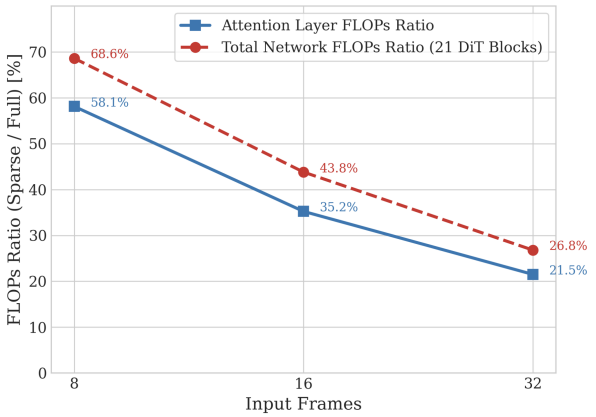

Figure A1. **Computational scaling analysis of the sparse temporal attention mechanism**. The lines show the FLOPs ratio (Sparse/Full) for the core temporal attention layer and the entire network across varying input frame counts.

As shown in Tab. A2, we quantitatively evaluate the computational benefits of our proposed sparse temporal attention mechanism compared to a standard full attention baseline across different input sequence lengths. When analyzing the core temporal attention layer in isolation, our method demonstrates a highly favorable scaling trend driven by the block-sparse connectivity: while processing a shorter sequence of 8 frames results in 58.11% of the full attention FLOPs, this ratio drops significantly to 35.23% for 16 frames and further decreases to a remarkable 21.50% when handling long 32-frame sequences. This efficiency gain translates directly to the total network computation across the 21 DiT blocks. The full attention network requires 425.7 PFLOPs for 16 frames, whereas our sparse implementation reduces the total cost to 186.3 PFLOPs, or 43.8% of the baseline. For the most challenging case of 32 frames, the full attention model requires a prohibitive 1584.9 PFLOPs; our sparse design dramatically cuts this down to only 425.0 PFLOPs, achieving a mere 26.8% of the total computational load. These results strongly validate the superior efficiency of our block-sparse approach, particularly as the temporal dimension increases.

Fig. A1 clearly illustrates the effectiveness of our sparse attention strategy as the temporal dimension scales. The two curves, representing the FLOPs ratio for the spatiotemporal attention layer and the total network, both exhibit a sharp non-linear decline as the input frame count increases from 8 to 32. This decreasing ratio confirms that the complexity of our sparse mechanism grows much slower than the $O(N^2)$ complexity of full attention. Crucially, the Total Network FLOPs ratio (the upper curve) is consistently higher than the layer-specific ratio (the lower curve). This difference arises because the total network cost includes computationally fixed components, such as Spatial Attention, Cross-Attention, and the Feed-Forward Network (FFN), which must be executed in both the sparse and full baselines. As the temporal attention becomes more sparse, these fixed costs occupy a larger percentage of the total budget, thus raising the overall Network FLOPs ratio, despite the massive savings achieved at the attention layer itself.

4. Additional Visual Quality Assessment

Table A3. **Results comparison.**

Method	LPIPS ↓	CLIP ↑	FVD ↓	Time ↓
Hunyuan3D	0.131	0.803	1276.2	24 min
DreamMesh4D	0.145	0.835	914.9	45 min
V2M4	0.152	0.827	952.0	45 min
Ours	0.098	0.916	483.1	**7 min**
Ours-full	**0.094**	**0.919**	**477.8**	16 min

In Tab. A3, we provide a comprehensive quantitative comparison of our method against several baselines (Hunyuan3D [2], DreamMesh4D [3], and V2M4 [1]), focusing on video-based quality assessment and overall inference time. To rigorously evaluate the generated 4D sequences, we employ LPIPS for perceptual distance, CLIP score for visual fidelity, and Fréchet Video Distance (FVD) for temporal coherence. As shown in the table, our method significantly outperforms all baseline approaches across all quality metrics. Most notably, our framework achieves a substantial reduction in FVD. Furthermore, we evaluate the computational efficiency of each approach. Our default configuration ("Ours") requires only 7 minutes to generate a complete sequence, establishing a new standard for efficiency compared to existing methods that take up to 45 minutes. For scenarios requiring maximum geometric and visual quality, our "Ours-full" configuration achieves the best

overall performance (LPIPS of 0.094, CLIP of 0.919, and FVD of 477.8) with a modest increase in inference time to 16 minutes. This demonstrates that our framework not only delivers state-of-the-art 4D generation quality but also provides a highly practical and flexible trade-off between speed and performance.

5. Generalization to Longer Sequences

Table A4. **Scalability analysis.**

Frames	Chamfer ↓	IoU ↑	F-Score ↑
8	0.099	0.338	0.315
16	0.102	0.339	0.315
32	0.106	0.334	0.314
64	0.114	0.326	0.310

To evaluate the temporal scalability of Sculpt4D, we investigate its ability to generate sequences longer than those seen during training. Specifically, while our model is trained exclusively on 16-frame sequences, we conduct inference on extended sequences of up to 64 frames without any additional fine-tuning. Tab. A4 presents the quantitative geometric evaluation—measuring Chamfer distance, Intersection over Union (IoU), and F-Score—across varying sequence lengths (8, 16, 32, and 64 frames). The results demonstrate remarkably robust performance. Notably, even when extrapolating to 4× the training length (64 frames), the metrics remain highly stable (e.g., the F-Score only experiences a marginal shift from 0.315 to 0.310). This confirms that our incorporated sparse attention mechanism effectively preserves temporal coherence and geometric fidelity, enabling strong zero-shot generalization to significantly longer temporal contexts.

6. More Visualization Results

Fig. A2 and Fig. A3 present additional visualizations of the mesh sequences. We select six time frames and show two views for each frame, with the small images on the left corresponding to the input views.

References

[1] Jianqi Chen, Biao Zhang, Xiangjun Tang, and Peter Wonka. V2m4: 4d mesh animation reconstruction from a single monocular video. *arXiv preprint arXiv:2503.09631*, 2025. 2

[2] 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. 2

[3] Zhiqi Li, Yiming Chen, and Peidong Liu. Dreammesh4d: Video-to-4d generation with sparse-controlled gaussian-mesh hybrid representation. *NeurIPS*, 2024. 2

[4] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *ICLR*, 2017. 1

[5] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 1

[6] Biao Zhang, Jiapeng Tang, Matthias Niessner, and Peter Wonka. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. *TOG*, 2023. 1

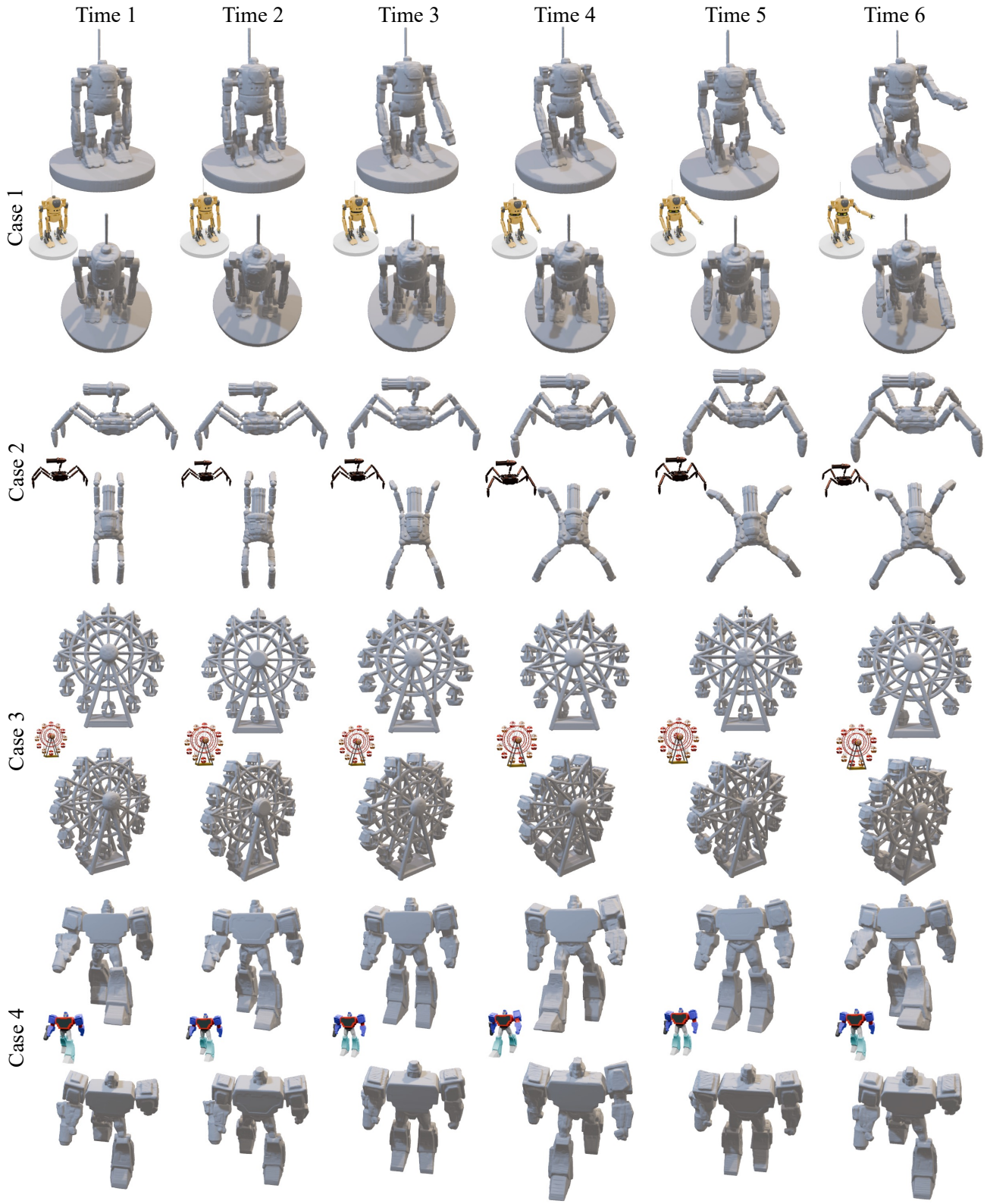

Figure A2. **More 4D mesh sequence results.**

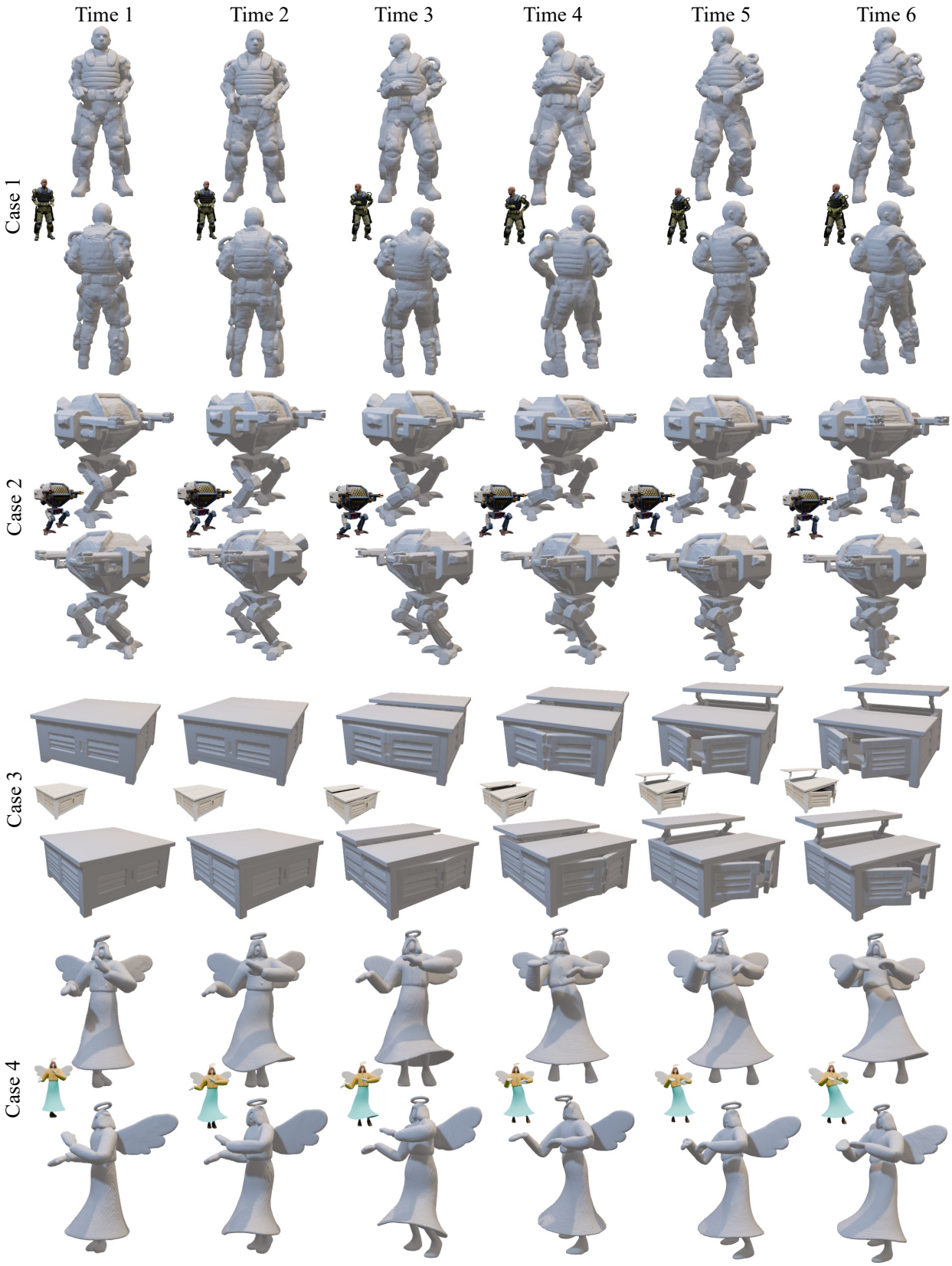

Figure A3. **More 4D mesh sequence results.**