RigMo: Unifying Rig and Motion Learning for Generative Animation

Supplementary Material

Overview of Supplementary Materials

Due to the strict page limit of the main manuscript, several important details, analyses, and visualizations could not be presented. This supplementary document provides a comprehensive extension of our method and experiments. Specifically, we include the following additions:

- **Comparison with prior work.**
- **RigMo-VAE.** We provide side-by-side comparisons of the 48-token and 128-token configurations to highlight their differences in reconstruction quality.
- **MotionDiT.** We include a detailed description of the experimental settings, training configuration, dataset preparation, and complete quantitative evaluations, accompanied by an in-depth analysis of the observed trends.
- **Video Demonstrations.** We further provide demo videos showcasing RigMo's generated results (1-frame $\rightarrow$ 9-frame prediction), along with side-by-side visual comparisons of reconstruction results against state-of-the-art baselines that could not fit in the main paper.

Together, these supplementary materials offer a more complete view of RigMo's effectiveness and support the claims made in the main manuscript.

Comparison with prior work.

Existing approaches related to RigMo fall broadly into two categories: dynamic 3D generation methods and rigging–structure prediction methods. Despite impressive progress, none of these paradigms jointly model rigging and motion in a unified, self-supervised manner, which fundamentally limits their applicability to general deformable objects.

Dynamic 3D generation. Recent 4D VAE models such as AnimateAnyMesh and GVFDiffusion learn trajectories for each vertex or Gaussian primitive. While effective for short-term sequence reconstruction, these models do not recover an explicit rig or kinematic structure, and typically require hundreds of latent tokens (often $512+$) to encode motion. In contrast, RigMo uses only 48 latent tokens yet produces a compact, interpretable rig together with temporally coherent motion parameters. Another line of work extends static 3D VAEs, including Hunyuan3D and Step1X3D, to per-frame mesh generation. These methods operate on independent frames and therefore suffer from slow inference, lack of temporal smoothness, inconsistent geometry across time, and, most critically, no vertex correspondence between the generated shapes. Finally, motion–generation models designed for humans or articulated categories (e.g., AnyTop, many human motion diffusion/transformer models) assume access to ground-truth rigs. Such assumptions break down for general objects where no consistent rig annotations exist, and rigging conventions vary widely across datasets and artists, resulting in poor cross-category generalization.

Method	Explicit Rig	Temporal Consist.	Rig Annotation
Dynamic 3D Generation			
AnimateAnyMesh / GVFDiff.	✗	✓	none
Hunyuan3D / Step1X3D	✗	✗	none
Human/AnyTop Motion Gen.	✓(GT)	✓	GT rigs
Rigging Prediction			
UniRig / Magic Articul.	✓	✗	GT rigs
RigMo (Ours)			
RigMo	✓	✓	none

Table 4. Comparison between RigMo and representative method families.

Rigging prediction. Methods such as UniRig and Magic Articulate aim to predict rig structures from static geometry. These methods depend heavily on large-scale human-annotated rig datasets; however, manual rigging contains inconsistencies, annotation noise, and large variations across artists and modeling standards. Moreover, available datasets are small, category-specific, and difficult to scale, which limits generalization to diverse objects and motion patterns. RigMo, in contrast, requires no human annotation: it infers the underlying kinematic structure directly from observed motion. This motion-driven formulation yields rigging that is physically meaningful, consistent across instances, and automatically aligned with the actual deformation behavior of the object.

By unifying rigging and motion into a single generative model, RigMo overcomes the core limitations of both categories. It provides interpretable rigs, consistent mesh deformations, temporally stable trajectories, and strong generalization without relying on manual labels—offering a principled and scalable solution for general-object animation.

MotionDiT Evaluation

To more comprehensively evaluate the motion generation capability of MotionDiT, we design a set of controlled experiments that minimize sampling randomness and isolate the contribution of the diffusion-based temporal prediction module. Each experiment is conditioned on two complementary signals, described below.

Method	1→1		1→9	
	Train (L1/L2)	Val (L1/L2)	Train (L1/L2)	Val (L1/L2)
w/o frame cond.	2.13±0.46 / 1.52±0.33	2.48±0.48 / 1.74±0.35	2.63±0.94 / 1.98±0.72	3.10±0.91 / 2.22±0.69
latent rig cond.	1.89±0.42 / 1.31±0.32	2.01±0.43 / 1.45±0.31	2.15±0.85 / 1.44±0.64	2.51±0.84 / 1.78±0.60
full MotionDiT	**1.43**±0.41 / **0.99**±0.30	**1.77**±0.41 / **1.23**±0.29	**1.61**±0.83 / **1.18**±0.61	**1.86**±0.82 / **1.37**±0.56

Table 5. Ablation study of MotionDiT ($\times 10^{-2}$) under two sparse-conditioning settings. We report Chamfer Distance (L1 / L2) for both training and validation sets. The three variants include: (1) w/o frame condition (no frame-mask guidance), (2) latent rig condition (using only the rig-branch latent feature), and (3) full MotionDiT (latent rig feature + decoded skinning weights + Gaussian bone centers + frame-mask guidance). Experiments are done on DT4D test dataset.

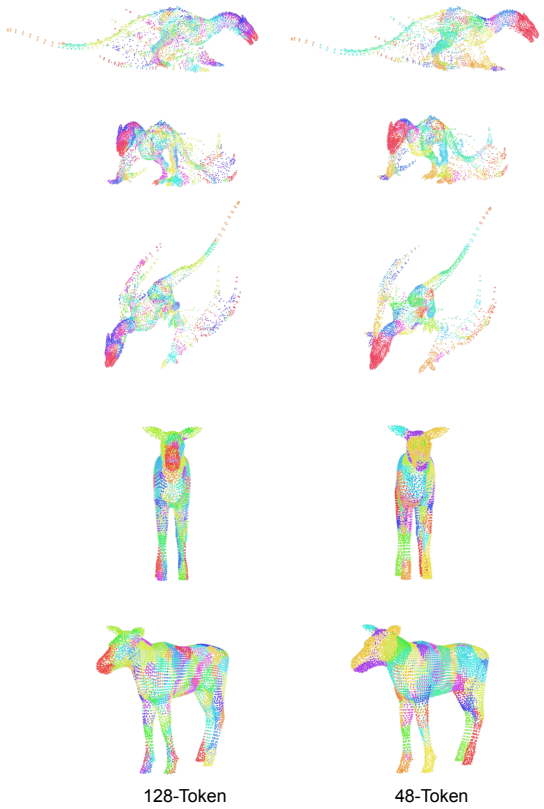

Figure 6. Side-by-side skinning weights comparisons of the 48-token and 128-token configurations.

- Rigging condition: MotionDiT receives rig-aware structural cues extracted from the RigMo-VAE rig branch. These include the latent rig feature as well as the decoded, physically meaningful rig attributes such as skinning weights and Gaussian bone centers, injected into the first cross-attention layer (Fig. 3). This conditioning provides the model with articulation structure and the expected ranges of deformation.
- Frame condition (mask pattern): A sparse set of observed frames is encoded into a frame-mask sequence and fed into the second cross-attention layer. This mask pattern controls which frames are visible to the model and enables controlled interpolation and long-horizon prediction.

We evaluate MotionDiT under two sparse-conditioning settings:

1. 1-frame → 1-frame prediction: Given a single input frame at time t, the model predicts the next frame at $t+1$, focusing on local temporal smoothness and short-term motion fidelity.
2. 1-frame → 9-frame prediction: Given one observed frame, the model predicts the following nine frames. This more challenging setting evaluates the ability of the model to produce temporally coherent and physically plausible long-horizon motion.

These sparse-conditioning setups avoid free-form motion sampling and instead provide controlled evaluation of the temporal prediction capability of MotionDiT. They also allow a clean examination of the effect of different rig-conditioning strategies.

We compare three variants:

- w/o frame condition: Removes the frame-mask signal entirely, forcing the model to infer temporal structure without any observed-frame guidance.
- Latent rig condition: Conditions the model only on the latent rig feature extracted from the RigMo-VAE rig encoder, without using decoded skinning weights or Gaussian bone centers.
- Full MotionDiT: Uses the complete rig-conditioning bundle, including the latent rig feature, decoded skinning weights, and Gaussian bone centers, providing explicit structural and physically interpretable priors for motion prediction.

Quantitative results on both the training and validation splits, measured using Chamfer Distance (CL_1 and CL_2), are summarized in Table 5. Across both prediction horizons, the full MotionDiT model achieves strong motion fidelity and temporal coherence.