MOTION IS STRUCTURE: ZORA FOR ZERO-SHOT 3D RIGGING FROM IN-THE-WILD 4D DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Automating the creation of articulated 3D characters is a longstanding challenge in computer graphics and embodied AI, with applications that span gaming, visual effects, and physical simulation. Recent advances in monocular 4D reconstruction enable extraction of dynamic 3D surfaces from an input single-view video, yet these outputs remain *baked geometry*: temporally re-playable but not editable or animatable. Converting them into articulated assets with explicit skeletons and skinning weights is a critical unsolved step. Existing automatic rigging methods rely on rig-label supervision or category-specific priors, making them brittle under severe self-contact and out-of-distribution topologies such as multi-limbed creatures or non-humanoid figures. We introduce **ZORA** (Zero-shot Optimization for Rigging Assets), a training-free framework built on a single principle: *motion is structure*. Given a reconstructed 4D surface, ZORA extracts a topology-aware kinematic scaffold by routing geodesic paths through regions of high temporal motion variance, suppressing spurious shortcuts at self-contact without any learned prior. It then recovers an explicit linear blend skinning rig via joint optimization of skinning weights and per-frame rigid transformations under four complementary objectives: motion affinity, topological locality, temporal smoothness, and volumetric consistency. Crucially, no template skeleton or labeled training data is required, coherent 4D motion alone serves as a geometric prior, binding vertices with correlated dynamics to the same articulated part. ZORA produces editable, production-ready rigs across a diverse range of characters, including those with complex topologies where prior methods degrade.

1 INTRODUCTION

Recent advances in image-to-video generation (Hu et al., 2022; Ni et al., 2023; Ren et al., 2024; Xing et al., 2025; Li et al., 2025) and monocular 4D reconstruction (Gafni et al., 2021; Wang et al., 2025c; Chen et al., 2025; Wu et al., 2025) have made it increasingly feasible to recover dynamic 3D content from unconstrained visual input. However, such outputs remain *dynamic surfaces* rather than *articulated assets*: they can be rendered and replayed, but they do not provide the rig structure required for editing, control, simulation, or animation reuse. Meanwhile, existing automatic rigging methods mainly operate in the static mesh-to-rig regime (James & Twigg, 2005; Baran & Popović, 2007; Xu et al., 2020; Liu et al., 2025a), typically relying on rig-labeled datasets, category-specific priors, or learned shape statistics from pre-rigged assets. This reliance creates a critical bottleneck, as the vast majority of "in-the-wild" visual data, whether unconstrained 3D scans, or generated 4D assets, inherently lacks any form of manual structural annotation.

This static, data-driven paradigm has three fundamental limitations. First, because learned riggers inherit the distributional biases of existing rigged datasets, they can generalize poorly to out-of-distribution topologies such as fantasy creatures, multi-limbed objects, or abstract generated assets. Second, static geometry alone is often underdetermined: parts that are spatially proximate, self-contacting, or fused in a rest pose, such as crossed arms or wings tightly folded against a bird's torso, cannot be reliably separated based on shape alone. Third, ideal rigging should be related to the target animation, i.e., non-moving parts should not be involved in the rigging, and parts with complicated deformation require more detailed rigging.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

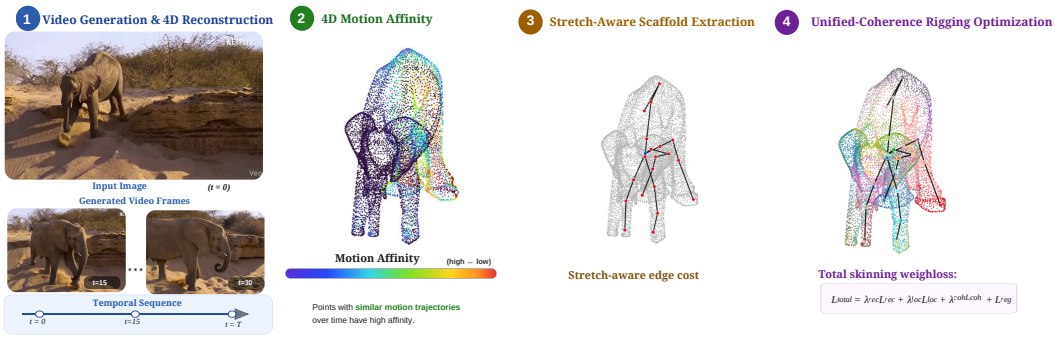


Figure 1: **Overview of the ZORA Pipeline.** ZORA extracts functional rigs by observing dynamic motion. (1) Given an input image or short clip, we synthesize and reconstruct a 4D sequence, (2) calculating motion affinity to group cohesive surface regions. (3) We then extract a stretch-aware kinematic scaffold, (4) and optimize the skinning weights through the skeleton.

We argue that *motion is structure*. Over time, vertices belonging to the same articulated part exhibit coherent trajectories, while joints, contacts, and part boundaries appear as discontinuities in motion affinity. Thus, given a monocular 4D reconstruction, rigging can be reformulated as the problem of recovering the latent articulated model that best explains the observed deformation. This raises a central question: *Can an editable articulated rig be recovered from monocular 4D motion without rig labels, category templates, or semantic part annotations?*

In this work, we answer this question by introducing **ZORA** (**Z**ero-shot **O**ptimization for **R**igging **A**ssets), a physics-constrained *4D-to-rigging* framework that extracts articulated structures directly from in-the-wild 4D dynamics. Bypassing dataset bias and manual labels entirely, ZORA utilizes kinematic motion affinity as an explicit geometric prior for both articulation discovery and skinning optimization. Starting from a monocular dynamic reconstruction, our method requires zero semantic templates or category-specific skeletons. Instead, ZORA first extracts a topology-aware kinematic scaffold by temporally severing spurious contacts via trajectory variance, estimating a geodesic root–extremity hierarchy, and strictly anchoring joint proposals within the volumetric interior of the shape. By explicitly enforcing topological locality, multi-scale coherence, and volumetric consistency, ZORA ensures that the deformed mesh behaves physically like solid biological tissue. Together, these unified components enable the robust recovery of editable, production-ready articulated rigs entirely from zero-shot 4D observations.

As a result, the proposed method converts unconstrained 4D-derived geometry into reusable articulated rigs without labeled rigs, manual annotation, or pre-defined skeletal models. Our main contributions are as follows:

- We introduce a fully zero-shot 4D-to-rigging formulation in which 4D motion affinity serves as the supervisory signal for articulation discovery, eliminating the need for labeled rigs or category-specific skeletal templates.
- We propose a topology-aware yet template-free kinematic initialization procedure that extracts joint proposals from a motion-refined 4D mesh graph via temporal cutting, geodesic routing, and adaptive branch consolidation.
- We develop a unified joint optimization objective that combines topological locality, affinity-guided consistency, multi-scale coherence, and temporal smoothness to recover stable skinning weights and articulated motion from monocular dynamic reconstructions.

2 Related Work

Automatic Rigging and Motion-Aware Articulation. Early geometry pipelines (James & Twigg, 2005; Baran & Popović, 2007) and classical trajectory clustering frameworks like RASR (Le & Deng, 2014) pioneered motion-driven extraction but remain highly sensitive to topological noise because they treat motion merely as a downstream analytical signal. Conversely, recent learning-based and generative methods advance template-free rigging by predicting structures directly from

3D shapes (Liu et al., 2019; Pan et al., 2021; Xu et al., 2020; Song et al., 2025b; Liu et al., 2025a; Song et al., 2025a; Zhang et al., 2025; Sun et al., 2025b;a), operating primarily in the static mesh-to-rig regime by inferring articulation from dataset priors rather than observed kinematics. Even modern motion-aware architectures (Yang et al., 2021; Xu et al., 2022) remain strictly supervised by ground-truth annotations. In contrast, our framework tackles a fundamentally different setting: zero-shot 4D-correspondence-to-rigging. By elevating motion affinity to an explicit structural constraint within the optimization loop, ZORA physically regularizes corrupted upstream 4D inputs from unconstrained video, successfully extracting robust articulated rigs without requiring rig-labeled datasets, category-specific templates, or prior training.

Motion as Geometric Structure. Visual geometry has advanced dynamic 3D recovery from classical non-rigid structure-from-motion (Garg et al., 2013; Dai et al., 2014) to neural models that predict continuous 4D representations directly from monocular video (Cai et al., 2022; Wang et al., 2023; 2025b;a; Zhou et al., 2025; Wu et al., 2024; Jiang et al., 2026). However, these approaches yield temporally replayable surfaces, leaving the extraction of continuous Linear Blend Skinning (LBS) rigs severely underdetermined. While prior works leverage motion to infer structure, they diverge fundamentally from our objective by targeting rigid object segmentation (Liu et al., 2023), learning latent pose representations via training phases (He et al., 2025; Yoo et al., 2024), or producing representation-specific assets like Gaussians and deformation graphs (Liu et al., 2025b; Bozic et al., 2021). In contrast, ZORA treats 4D reconstructions strictly as upstream evidence. We formulate a robust, zero-shot optimization explicitly designed to survive defective, in-the-wild 4D inputs, directly extracting industry-standard, editable kinematic skeletons and continuous soft-tissue skinning weights without requiring any prior training.

In summary, visual geometry and automatic rigging remain weakly connected because classical extraction shatters under 4D reconstruction noise. ZORA bridges this gap with a robust formulation explicitly designed to survive defective, corrupted, in-the-wild 4D inputs. By embedding 4D motion affinity directly into a self-supervised optimization loop, ZORA actively corrects topological noise and resting self-contacts, successfully extracting standard, editable skinned rigs from corrupted upstream evidence without requiring template skeletons or rig-labeled supervision.

3 Method: 4D-Constrained Optimization for Topology-Agnostic 3D Rigging

Instead of inferring articulation directly from raw video pixels or static shape templates, ZORA formulates zero-shot rigging as an inverse optimization problem driven entirely by upstream 4D geometric evidence. We jointly optimize the explicit kinematic skeleton ($\mathcal{S}$), the continuous semantic skinning weights ($\mathbf{W}$), and the per-frame rigid bone transformations by strictly constraining them to mutually satisfy the observed 4D surface trajectories. By embedding 4D motion affinity and a Stretch-Aware Scaffold as optimization constraints, our framework severs spurious contacts and regularizes the geometry to produce a valid, topology-agnostic Linear Blend Skinning rig.

3.1 4D Lifting and Motion Affinity

Video Generation and 4D Reconstruction. Let $I \in \mathbb{R}^{H\times W\times 3}$ be the input image. A video generation model $\mathcal{G}$ produces a multi-frame sequence $\mathcal{V} = \{F_t\}_{t=1}^{T} = \mathcal{G}(I)$, implicitly encoding the plausible kinematic articulation of the subject. Alternatively, our method can also take existing video as input. A 4D reconstruction module processes $\mathcal{V}$ to yield a temporally-consistent mesh $\mathcal{M}$ with N vertices, providing per-vertex spatiotemporal trajectories: $\mathbf{X} = \{\mathbf{x}_i^t\}_{i=1}^{N},\ t \in [1, T]$.

Temporal Motion Affinity. Motivated by Hu et al. (2025), which shows that VGGT attention layers implicitly encode dynamic cues, we extract motion-coherence signals from Zhou et al. (2025) without manual annotations or category priors. These cues are mined from PAGE-4D attention via Gram similarity, temporally aggregated, and refined with projection-gradient information, providing soft kinematic evidence for separating dynamically coherent regions. For each vertex pair (i, j), the temporal motion affinity A_{ij} is defined as the inverse of their displacement variance across all frames:

$$A_{ij} = \exp(-\frac{1}{\tau T}\sum_{t=1}^{T} \|\Delta\mathbf{x}_i^t - \Delta\mathbf{x}_j^t\|^2) \tag{1}$$

where $\Delta \mathbf{x}_i^t$ is the per-frame displacement and $\tau > 0$ is a temperature hyperparameter. To ensure topological consistency, $\mathbf{A}$ is regularized via a Dirichlet energy functional $\mathcal{L}_{\text{aff}} = \sum_{(i,j)\in\mathcal{E}}(A_{ij} - A_{ij}^{\text{geo}})^2$, where A_{ij}^{geo} decays exponentially with respect to the geodesic distance $d_{\text{geo}}(i,j)$ on the mesh manifold.

3.2 Stretch-Aware Kinematic Scaffold Extraction

Given the reconstructed 4D mesh sequence $\mathcal{X} = \{\mathbf{X}^t\}_{t=1}^T$, where $\mathbf{X}^t = \{\mathbf{x}_i^t\}_{i=1}^V$, we first extract an initial kinematic scaffold that serves as the structural initialization for rig optimization. Unlike purely static skeletonization methods, our scaffold is computed using temporal deformation cues. The key observation is that vertices connected by a valid articulated part should preserve their relative distance over time, whereas spurious contacts or shortcuts across independently moving parts exhibit large temporal stretch.

We construct a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where vertices $\mathcal{V}$ correspond to mesh vertices, and edges $\mathcal{E}$ are formed from mesh connectivity together with local proximity queries. Rather than removing potentially ambiguous proximity edges, we assign each edge a stretch-aware traversal cost. For an edge $e = (i,j)$, let $d_e^t = \|\mathbf{x}_i^t - \mathbf{x}_j^t\|_2$ be its length at frame t. We define the rest length and temporal stretch as

$$d_e^0 = \max(d_e^1, \epsilon), \qquad s_e = \max_t d_e^t - \min_t d_e^t, \tag{2}$$

where ϵ is a small number to avoid numerical instability. The edge cost is then defined as $c_e = d_e^0(1 + \alpha \frac{s_e}{d_e^0})$, with α controlling the strength of the stretch penalty. This cost preserves stable local connections while strongly penalizing edges that connect parts with inconsistent motion.

Using the resulting graph metric, we estimate a topological root by minimizing average shortest-path distance to sampled surface nodes:

$$r = \arg\min_{u\in\mathcal{V}_s} \frac{1}{|\mathcal{V}_s|} \sum_{v\in\mathcal{V}_s} d_{\mathcal{G}}(u,v), \tag{3}$$

where $\mathcal{V}_s$ is a sampled vertex set and $d_{\mathcal{G}}$ is the Dijkstra distance under costs c_e. We then select K extremities using graph farthest-point sampling and trace shortest paths from each extremity to the root. Joint candidates are sampled along these root-extremity paths and consolidated using geodesic non-maximum suppression. Finally, each joint proposal is projected toward the volumetric interior by moving it along the inward normal direction, ensuring that the initialized skeleton lies inside the animated surface rather than on the boundary.

3.3 Unified-Coherence Rigging Optimization

Given the extracted kinematic scaffold, we optimize a standard Linear Blend Skinning (LBS) rig. The unknowns are the skinning weights $\mathbf{W} = \{w_{ib}\}_{i=1,b=1}^{V,B}$, constrained by $w_{ib} \geq 0$ and $\sum_{b=1}^B w_{ib} = 1$, alongside the per-frame bone transformations, where V and B are the numbers of vertices and bones. Let $\mathbf{G}_b^t \in SE(3)$ denote the global transformation of bone b at frame t. The predicted vertex position is given by:

$$\hat{\mathbf{x}}_i^t = \pi\Big(\sum_{b=1}^B w_{ib}\mathbf{G}_b^t(\mathbf{G}_b^1)^{-1}\tilde{\mathbf{x}}_i^1\Big), \tag{4}$$

where $\tilde{\mathbf{x}}_i^1$ is the homogeneous canonical vertex coordinates and $\pi(\cdot)$ maps coordinates back to $\mathbb{R}^3$. We optimize the rig using a unified objective:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{loc}}\mathcal{L}_{\text{loc}} + \lambda_{\text{coh}}\mathcal{L}_{\text{coh}} + \mathcal{L}_{\text{reg}}. \tag{5}$$

Reconstruction & Locality: The reconstruction term $\mathcal{L}_{\text{rec}}$ measures the L_2 fidelity between the predicted vertices and the observed 4D sequence. To prevent standard weight-bleeding pathologies where distant bones incorrectly influence a vertex, we introduce a topological locality prior $\mathcal{L}_{\text{loc}}$. Let $d_{\mathcal{G}}(i,b)$ denote the stretch-aware geodesic distance from vertex i to bone b. We heavily penalize cross-space influence:

$$\mathcal{L}_{\text{loc}} = \frac{1}{V}\sum_{i=1}^V\sum_{b=1}^B w_{ib}\frac{d_{\mathcal{G}}(i,b)^2}{D^2}, \tag{6}$$

where D is the scale-normalizing bounding-box diagonal.

	$E_{LBS} \downarrow$	$E_{norm} \downarrow$	$E_{coh} \downarrow$	$E_{viol} \downarrow$	$E_{cross} \downarrow$
MoRig	0.5098 ± 0.4194	0.5557 ± 0.2019	0.2258 ± 0.3767	0.3334 ± 0.3651	0.9422 ± 0.1489
RASR	0.0111 ± 0.0053	0.2300 ± 0.1258	0.0241 ± 0.0078	0.0673 ± 0.0484	0.6749 ± 0.2777
Ours	0.0586 ± 0.0264	0.3708 ± 0.1821	0.0653 ± 0.0234	0.0107 ± 0.0106	0.2040 ± 0.2696

Table 1: **Quantitative comparison on the in-the-wild AnimalKingdom dataset.** Because this dataset lacks ground-truth skeletons, we evaluate performance using our proposed unsupervised metrics. The results quantitatively capture MoRig's structural collapse on out-of-distribution topologies, evidenced by much larger E_{viol} and E_{cross}. In contrast, our zero-shot method successfully enforces topological validity while simultaneously achieving an order-of-magnitude reduction in E_{LBS}.

Kinematic Affinity: The core of our zero-shot optimization is the coherence term $\mathcal{L}_{\text{aff}}$ in $\mathcal{L}_{\text{coh}}$, which forces the skinning weights to respect both local mesh geometry and global 4D dynamics. Specifically, we use the affinity matrix A obtained from upstream 4D motion cues. Vertices exhibiting high motion affinity are explicitly encouraged to receive identical skinning weights via a normalized Dirichlet energy:

$$\mathcal{L}_{\text{aff}} = \frac{1}{Z_A} \sum_{(i,j)\in\mathcal{P}_A} A_{ij} \|\mathbf{w}_i - \mathbf{w}_j\|_2^2, \tag{7}$$

where $\mathcal{P}_A$ denotes high-confidence affinity pairs. This ensures that vertices with coherent dynamic behavior are bound to the same bone. $\mathcal{L}_{\text{coh}}$ also includes mesh-level coherence $\mathcal{L}_{\text{mesh}}$ (see Appendix A). Finally, $\mathcal{L}_{\text{reg}}$ encompasses auxiliary regularizers for temporal smoothness and weight balance, the exact formulations of which are detailed in Appendix A.

4 Experiments

Implementation Details. We implement ZORA using PyTorch. All zero-shot optimizations are executed on a single NVIDIA RTX 4090 GPU. The computing resources used in our implementation are listed in Appendix C.

Datasets. To rigorously evaluate our method's topology-agnostic generalization, we first benchmark against *ModelsResource* (Xu et al., 2022), a dataset featuring a vast spectrum of morphologies and severe dynamic articulations. Furthermore, to assess performance on unconstrained, in-the-wild scenarios, we introduce a novel unlabeled dataset of 4D structures. This collection is curated using unannotated sequences from the *Animal Kingdom* dataset [1] (Ng et al., 2022), alongside highly diverse synthetic assets generated via Veo 3.1 (Wiedemer et al., 2025).

Baselines. We establish **MoRig** (Xu et al., 2022) and **RASR** (Le & Deng, 2014) as our primary baselines. As the current state of the art (SOTA) in motion-driven auto-rigging, MoRig serves as the most direct and competitive point of comparison for our framework. We compute two variants of this baseline, **MoRig-single**($\text{MoRig}_{\text{sgl}}$), which processes single-view point clouds rendered from each frame to simulate real-world occlusions, and **MoRig-standard** ($\text{MoRig}_{\text{std}}$), which utilizes full-surface point clouds densely sampled from the complete 4D sequence to establish a theoretical performance upper bound under perfect 3D visibility. To rigorously isolate the benefits of our algorithmic formulation in a zero-shot setting, we include **RASR**, a classical mesh-sequence rigging framework that consumes the exact same temporally consistent 4D vertex trajectories as our method. Furthermore, we qualitatively compare our zero-shot approach with SOTA supervised rigging models, specifically RigAnything (Liu et al., 2025a) and Puppeteer (Song et al., 2025a). Despite these baselines relying on massive, domain-specific 3D training corpora, comprising 9,388 (RigAnything) and over 74,000 (Puppeteer) annotated assets, respectively, our optimization-based formulation achieves highly competitive structural extraction without requiring a single training example.

4.1 Quantitative Evaluation of Unannotated Rigging

For generated or in-the-wild 4D assets, ground-truth skeletons and skinning weights are strictly unavailable. We therefore evaluate the recovered rigs using a custom suite of reconstruction-based, motion-based, and structural validity based metrics. Crucially, to ensure the reliability of our unsupervised metrics, we ground them in well-established geometric principles. Our motion metric builds on As-Rigid-As-Possible (ARAP) kinematics (Sorkine et al., 2007; Yang et al., 2022) to

[1] https://sutdcv.github.io/Animal-Kingdom/

preserve local structure during movement. Furthermore, boundary violation uses Signed Distance Fields (SDF) (Bogo et al., 2016); because SDFs explicitly quantify the spatial distance to the nearest surface, they provide a mathematically precise way to measure and penalize unnatural collisions in articulated models. Detailed mathematical formulations for all metrics are provided in Appendix D.

LBS Reconstruction (E_{LBS}) measures the scale-normalized mean per-vertex error between the observed 4D sequence and the sequence reconstructed via our extracted linear blend skinning rig. This serves as the standard kinematic fidelity metric in auto-rigging literature (Xu et al., 2020; Kavan & Žára, 2005; Kavan et al., 2008).

Motion Coherence (E_{coh}) evaluates local rigidity. Grounded in established As-Rigid-As-Possible (ARAP) principles (Sorkine et al., 2007; Yang et al., 2022), we transform vertex trajectories into the local coordinate frame of their assigned bones and compute the temporally averaged variance. Lower variance indicates that vertices bound to the same bone exhibit highly consistent kinematics.

Boundary Violation (E_{viol}) measures continuous volumetric validity. Following standard boundary-adherence practices in physics-based modeling (Bogo et al., 2016), we sample points along predicted bone segments and utilize the animated mesh's Signed Distance Field (SDF) to strictly penalize any bone portions that pierce outside the object volume.

Cross-Space Bones (E_{cross}) evaluates discrete structural failures. Grounded in classic curve-skeleton principles where valid kinematic chains must strictly reside within the object's interior (Baran & Popović, 2007; Au et al., 2008; Cornea et al., 2024), it measures the frequency of spurious skeleton edges that form unsupported shortcuts through empty space. This is defined as bones whose outside-volume ratio exceeds a strict distance tolerance γ and duration fraction ρ.

Surface Normal Consistency (E_{norm}) evaluates local geometric preservation. While positional metrics like E_{LBS} capture global kinematics, they are often blind to high-frequency artifacts such as unnatural twisting or local folding. We penalize these structural distortions by computing the average cosine distance between the original 4D mesh face normals and the LBS-reconstructed face normals.

Table 1 presents a quantitative comparison on the *AnimalKingdom* dataset. Since this dataset provides no ground-truth skeletons, these metrics are crucial for assessing true rig quality. The comparison reveals a stark contrast in performance: MoRig's reliance on out-of-domain training priors leads to severe topological degradation, quantified by extreme cross-space bone allocations ($E_{cross} = 0.9422$) and pervasive volume intersections ($E_{viol} = 0.3334$). In contrast, our zero-shot optimization successfully bridges surface tracking with physical realism. We achieve an 88% reduction in raw positional error ($E_{LBS} = 0.0586$) alongside vast improvements in surface normal accuracy (E_{norm}) and local rigidity ($E_{coh} = 0.0653$). Ultimately, these unlabeled metrics strongly demonstrate that our framework produces structurally sound and temporally coherent rigs across complex morphologies that fundamentally break supervised models.

4.2 Evaluation on Datasets with Ground-Truth

Relying solely on exact-match structural metrics for ground-truth datasets is inherently flawed for zero-shot methods. These datasets often reflect arbitrary human stylistic conventions rather than strict kinematic laws, giving supervised baselines an unbalanced advantage. To provide a fair evaluation, we focus on functional motion reproduction rather than strict structural imitation. We measure this functional animation quality using **Animation MPVE** (Mean Per-Vertex Error). By reconstructing the animated mesh via linear blend skinning using the predicted skeleton and weights, a lower Animation MPVE demonstrates that the rig faithfully explains the observed 4D motion, regardless of stylistic differences in bone placement. For more details, please check Appendix D.

Table 2 evaluates the functional and structural validity of the extracted rigs. When compared against the realistic, monocular baseline ($\text{MoRig}_{\text{sgl}}$), our zero-shot formulation demonstrates overwhelming superiority, reducing positional reconstruction error (E_{LBS}) by more than an order of magnitude (0.0172 vs. 0.2338). Furthermore, we compare our method against $\text{MoRig}_{\text{std}}$, a fully supervised baseline provided with perfect 3D visibility (acting as a theoretical upper bound). As expected, this data-driven upper bound leverages its internalized 3D training priors to achieve slightly tighter raw surface fitting (E_{LBS}, E_{norm}, MPVE) and surface boundary adherence (E_{viol}). Crucially, however, our strictly optimization-based approach surpasses even this supervised upper bound in internal structural integrity. By effectively enforcing physical and geometric constraints, our method

	$E_{LBS} \downarrow$	$E_{norm} \downarrow$	$E_{viol} \downarrow$	$E_{cross} \downarrow$	MPVE $\downarrow$
$\text{MoRig}_{\text{sgl}}$	0.2338 ± 0.0295	0.1213 ± 0.0463	0.0067 ± 0.0088	0.0760 ± 0.1196	0.0341 ± 0.0136
$\text{MoRig}_{\text{std}}$	0.0151 ± 0.0131	0.0824 ± 0.0404	0.0032 ± 0.0055	0.0432 ± 0.1216	0.0120 ± 0.0055
RASR	0.0290 ± 0.0129	0.1571 ± 0.0810	0.0042 ± 0.0041	0.0756 ± 0.1005	0.0290 ± 0.0129
Ours	0.0172 ± 0.0087	0.0966 ± 0.0405	0.0191 ± 0.0094	0.0274 ± 0.0583	0.0180 ± 0.0174

Table 2: **Comparison against supervised baselines on ModelsResource.** Results reported as mean $\pm$ standard deviation. Despite operating entirely zero-shot, our method reduces positional error (E_{LBS}) by over an order of magnitude compared to the monocular baseline ($\text{MoRig}_{\text{sgl}}$). Furthermore, by strictly enforcing physical and geometric priors, our approach prevents unsupported kinematic shortcuts, achieving superior internal skeletal topology (E_{cross}) even when compared to a fully supervised 3D upper bound ($\text{MoRig}_{\text{std}}$).

achieves the lowest cross-space bone error ($E_{cross} = 0.0274$ vs. 0.0432). This proves that our zero-shot formulation successfully prevents the unsupported kinematic shortcuts that frequently plague data-driven models, ultimately yielding a robust, animation-ready rig without requiring any domain-specific training.

Additional qualitative comparisons on *ModelsResource* assets against MoRig, RigAnything, and Puppeteer are provided in Appendix E.2. While all evaluated methods successfully extract valid rigs for standard humanoid subjects, the heavily supervised baselines (RigAnything and Puppeteer) suffer a severe drop in quality when applied to non-humanoid topologies, such as avian meshes. In contrast, our zero-shot approach, ZORA, consistently recovers highly accurate and functional skeletal structures across both humanoid and out-of-distribution animal morphologies.

4.3 Comparison on Unannotated Topologies

Objects in-the-wild. As illustrated in Figure 2, current state-of-the-art rigging models struggle significantly to generalize to in-the-wild 3D geometries. Baselines such as MoRig and RigAnything fail entirely on non-standard morphologies, particularly invertebrates (*e.g.*, worms and centipedes) and aquatic animals (*e.g.*, seals and squids). While Puppeteer (Song et al., 2025a) demonstrates stronger generalization, successfully recovering functional rigs for familiar dataset staples like sheep and elephants (see Appendix E.1), it remains fundamentally brittle when applied to unconstrained captures. Specifically, Puppeteer consistently fails to produce valid topological structures for aquatic life (*e.g.*, squids), and particularly invertebrates (e.g., worms and centipedes). This instability highlights the limitations of relying on static shape priors or biased training distributions. In contrast, by grounding the rigging process in 4D spatio-temporal dynamics, ZORA robustly extracts kinematically valid rigs across this diverse array of complex, out-of-distribution morphologies.

Rigging from Generative Videos. To evaluate out-of-distribution robustness on emerging synthetic media, we visualize rigging results on high-fidelity generated videos (Appendix E.3, Fig. 8). When applied to these assets, standard baselines such as MoRig and RigAnything fail entirely to extract viable kinematic structures. Furthermore, while heavily supervised models like Puppeteer succeed on standard humanoids, they exhibit severe topological failures on fantastical or non-standard compositions. For instance, when processing complex, out-of-distribution geometries (*e.g.*, a dinosaur), Puppeteer struggles to deduce a valid skeletal topology, resulting in notable structural inconsistencies. More critically, when presented with extreme morphological anomalies, such as multi-limbed entities (*e.g.*, the six-armed figure Nezha), Puppeteer's strong dataset bias overrides the visual evidence, forcing a canonical bipedal template and entirely ignoring the supplementary appendages. Conversely, our zero-shot formulation bypasses these biases by relying entirely on observed kinematics. Grounded strictly in actual motion, our method successfully extracts all articulating components, proving its unparalleled adaptability to the boundless morphological diversity of generative video.

Generative Motion Augmentation for In-the-Wild Videos. A fundamental challenge of motion-driven, zero-shot rigging is that short, in-the-wild video clips frequently capture a limited range of motion. If a subject does not articulate a specific limb during the recording, the optimizer receives no kinematic evidence to discover that joint. To overcome this limitation, we leverage state-of-the-art video generation models to artificially extend the original footage. By synthesizing novel animations that force the subject into new poses, we generate the missing kinematic signals required for complete skeletal extraction. As shown in Fig. 9 at App. E.3, augmenting the source video

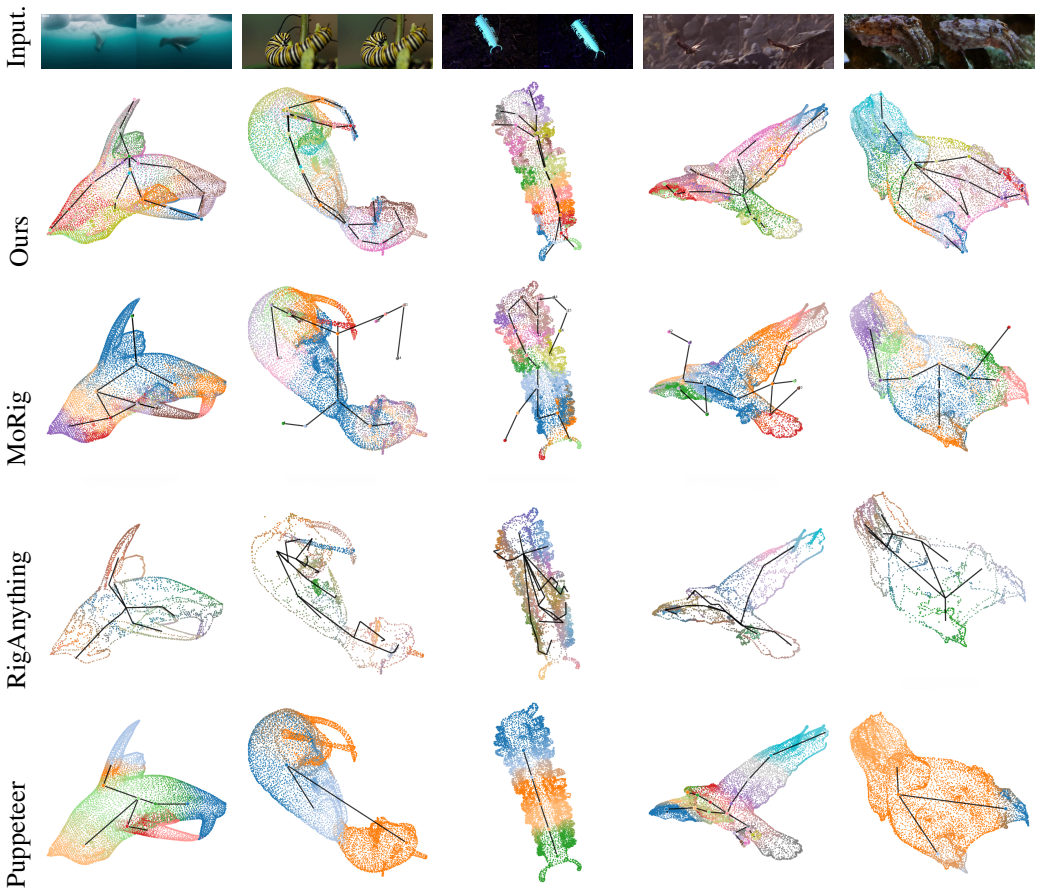


Figure 2: **Zero-shot extraction on in-the-wild AnimalKingdom data.** Each column highlights a distinct, complex animal topology. While traditional regression-based state-of-the-art baselines completely fail to capture the correct underlying skeletal structure, resulting in topologically broken and functionally unusable rigs, our method successfully adapts to these diverse, unseen shapes. This demonstrates the severe generalization limits of data-driven baselines and the robust, out-of-distribution adaptability of our zero-shot approach. For more rigging results, please check Appendix E.1.

with generative motion allows our framework to successfully discover previously dormant joints, significantly improving the topological completeness and functional quality of the final rig.

4.4 Ablation Study

To validate our physical and geometric constraints, we conduct an ablation study on *ModelsResource* dataset, utilizing both ground-truth structural metrics and our GT-free motion validity suite. Our baseline configuration uses mesh and affinity coherence with the skeleton-scale prior disabled. We systematically evaluate the following variants by isolating individual components: w/o $\mathcal{L}_{\text{loc}}$, and w/o $\mathcal{L}_{\text{mesh}}$. Furthermore, we ablate our auxiliary regularizers, w/o $\mathcal{L}_{\text{reg}}$, and Uniform Mesh (disabling part-aware weighting) and visualize the intermediate optimization stage (w/o Affinity and w/o Stretch Scaffold) at Sec. 3.2.

As shown in Table 3, the full ZORA formulation is essential for balancing positional tracking with structural realism. While removing the spatial regularizer (*w/o* $\mathcal{L}_{\text{mesh}}$) artificially minimizes continuous tracking errors (E_{LBS} dropping to 0.0135), this indicates severe overfitting; the optimizer fragments skinning weights to force exact vertex alignment, catastrophically degrading surface integrity (E_{norm} rises to 0.1293, E_{coh} to 0.0449). Furthermore, ablating the structural priors (*w/o* $\mathcal{L}_{\text{loc}}$, $\mathcal{L}_{\text{reg}}$) compromises internal kinematic validity, significantly increasing cross-space bones (E_{cross} to 0.1110). Finally, relying on naive *Uniform* weighting uniformly worsens both tracking precision and

	$E_{\text{LBS}} \downarrow$	$E_{\text{norm}} \downarrow$	$E_{\text{coh}} \downarrow$	$E_{\text{cross}} \downarrow$	MPVE $\downarrow$
Full (ours)	0.0291 ± 0.0109	0.0966 ± 0.0405	0.0273 ± 0.0095	0.1080 ± 0.1110	0.0293 ± 0.0118
w/o $\mathcal{L}_{\text{mesh}}$	0.0135 ± 0.0075	0.1293 ± 0.0516	0.0449 ± 0.0259	0.1027 ± 0.1472	0.0137 ± 0.0084
w/o $\mathcal{L}_{\text{loc}}$	0.0291 ± 0.0109	0.0967 ± 0.0406	0.0273 ± 0.0095	0.1090 ± 0.1126	0.0294 ± 0.0118
w/o $\mathcal{L}_{\text{reg}}$	0.0290 ± 0.0109	0.0964 ± 0.0405	0.0274 ± 0.0095	0.1110 ± 0.1151	0.0292 ± 0.0118
Uniform	0.0307 ± 0.0112	0.0977 ± 0.0400	0.0272 ± 0.0093	0.1225 ± 0.1239	0.0309 ± 0.0120

Table 3: Unsupervised ablation study on ModelsResource. While removing the primary spatial regularizer (*w/o* $\mathcal{L}_{\text{mesh}}$) artificially lowers E_{LBS} and MPVE, it causes severe structural and geometric collapse, evidenced by sharp increases in surface normal (E_{norm}) and coherence (E_{coh}) errors. Our full formulation successfully balances raw tracking with kinematic and topological validity. (*Uniform* denotes disabling part-aware weighting in favor of naive uniform weights.)

topology. By preventing degenerate solutions, such as fragmented skinning or unsupported skeletons, our complete objective ensures functional, production-ready rigs.

Qualitative Analysis of the Structural Prior. As shown in App. F (Fig. 10), removing the Stretch Scaffold causes the optimizer to prioritize naive surface fitting over anatomical logic. Without explicit topological constraints, predicted joints drift from true articulation points or collapse entirely, forcing the model to compensate by overfitting skinning weights. In contrast, our full formulation successfully anchors joints to valid kinematic hinges, extracting a structurally sound, production-ready rig.

To understand why removing the affinity prior artificially inflates continuous reconstruction metrics, we visualize the predicted skinning weights (Appendix F, Fig. 11). Without this prior, the optimizer overfits to observed poses, assigning fragmented, illogical weights to force exact positional alignment. Consequently, this unregularized structure suffers severe local collapse under novel articulations (Figs. 11d,e). Conversely, our affinity prior enforces smooth, continuous weight distribution across related surface regions. This guarantees cohesive movement and a robust, production-ready rig, fully justifying the minor trade-off in raw positional error.

Robustness to 4D Artifacts. We comprehensively stress-test ZORA against common 4D reconstruction corruptions, including temporal jitter, correspondence drift, missing geometry, and topological noise (*e.g.*resting self-contacts). While the classical RASR (Le & Deng, 2014) baseline remains highly competitive on clean, well-registered studio scans, its reliance on discrete downstream trajectory clustering causes catastrophic structural collapse under in-the-wild artifacts. Specifically, resting self-contacts trick classical methods into mathematically fusing distinct body parts. In contrast, ZORA's formulation actively embeds motion as an optimization constraint: it serves as a natural low-pass filter against high-frequency jitter, while its stretch-aware scaffold successfully severs spurious topological fusions and bridges missing geometry. Consequently, ZORA maintains structural fidelity and gracefully absorbs the exact unconstrained artifacts that shatter classical extraction pipelines. A full quantitative analysis mapping the failure boundaries of both methods is provided in Appendix G Table 5and 6.

5 Conclusion and Future Work

Guided by the principle that motion is structure, we present an end-to-end framework that extracts interpretable kinematic representations directly from 4D video. By treating spatio-temporal dynamics as a geometric prior, our approach overcomes the brittleness of static heuristics. We integrate geodesic weight masking, infinite log-barrier constraints, and dynamic topological surgery to formulate a robust optimization pipeline. Ultimately, our method generalizes across arbitrary topologies, reliably translating raw continuous motion into high-fidelity articulated structures. For limitation and future work, please check Appendix H.

AI use statement

In this work, we used generative AI tools for editing, refining, and polishing the academic writing of the manuscript after the initial manual drafting phase. We have not used generative AI tools for generating scientific ideas, creating experimental results, or formulating our core technical claims, and the automated generation of empirical data or quantitative metrics is not applicable to this work. We have reviewed all AI-assisted work. Specifically, the authors wrote the original manuscript text

and subsequently used AI solely to elevate the academic tone and conciseness; all AI-revised text was rigorously verified by the authors line-by-line to ensure it accurately preserved our original scientific intent, factual accuracy, and mathematical formulations.

Ethics statement

As a zero-shot, optimization-based framework, ZORA does not require training on massive, pre-annotated human datasets. This fundamentally mitigates common ethical concerns related to data privacy, consent, and representational bias that are often inherent in large-scale supervised modeling and generative pipelines. Our experiments utilize established, publicly available datasets for evaluation, and we do not foresee any direct negative societal impacts or dual-use concerns arising from our topological rigging extraction method.

Reproducibility statement

To facilitate the exact reproduction of our experiments and results, we have included the complete source code within the supplementary material. Detailed derivations of our optimization losses are documented in Appendix A, the evaluation metrics are defined in Appendix D, and all comprehensive hyperparameter settings are explicitly detailed in Appendix I. Upon acceptance of this paper, the entire codebase will be publicly released under an open-source license to support further research and adoption by the broader community.

References

Oscar Kin-Chung Au, Chiew-Lan Tai, Hung-Kuo Chu, Daniel Cohen-Or, and Tong-Yee Lee. Skeleton extraction by mesh contraction. *ACM transactions on graphics (TOG)*, 27(3):1–10, 2008.

Ilya Baran and Jovan Popović. Automatic rigging and animation of 3d characters. *ACM Transactions on graphics (TOG)*, 26(3):72–es, 2007.

Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European conference on computer vision*, pp. 561–578. Springer, 2016.

Aljaz Bozic, Pablo Palafox, Michael Zollhofer, Justus Thies, Angela Dai, and Matthias Nießner. Neural deformation graphs for globally-consistent non-rigid reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1450–1459, 2021.

Hongrui Cai, Wanquan Feng, Xuetao Feng, Yan Wang, and Juyong Zhang. Neural surface reconstruction of dynamic scenes with monocular rgb-d camera. *Advances in Neural Information Processing Systems*, 35:967–981, 2022.

Jianqi Chen, Biao Zhang, Xiangjun Tang, and Peter Wonka. V2m4: 4d mesh animation reconstruction from a single monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11643–11653, 2025.

Nicu D Cornea, Deborah Silver, and Patrick Min. Curve-skeleton properties, applications, and algorithms. *IEEE Transactions on visualization and computer graphics*, 13(3):530–548, 2024.

Yuchao Dai, Hongdong Li, and Mingyi He. A simple prior-free method for non-rigid structure-from-motion factorization. *International Journal of Computer Vision*, 107(2):101–122, 2014.

Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8649–8658, 2021.

Ravi Garg, Anastasios Roussos, and Lourdes Agapito. Dense variational reconstruction of non-rigid surfaces from monocular video. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pp. 1272–1279, 2013.

Guangzhao He, Chen Geng, Shangzhe Wu, and Jiajun Wu. Category-agnostic neural object rigging. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22078–22088. IEEE, 2025.

Yaosi Hu, Chong Luo, and Zhenzhong Chen. Make it move: controllable image-to-video generation with text descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18219–18228, 2022.

Yu Hu, Chong Cheng, Sicheng Yu, Xiaoyang Guo, and Hao Wang. Vggt4d: Mining motion cues in visual geometry transformers for 4d scene reconstruction. *arXiv preprint arXiv:2511.19971*, 2025.

Doug L James and Christopher D Twigg. Skinning mesh animations. *ACM Transactions on Graphics (TOG)*, 24(3):399–407, 2005.

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.

Ladislav Kavan and Jiří Žára. Spherical blend skinning: a real-time deformation of articulated models. In *Proceedings of the 2005 symposium on Interactive 3D graphics and games*, pp. 9–16, 2005.

Ladislav Kavan, Steven Collins, Jiří Žára, and Carol O'Sullivan. Geometric skinning with approximate dual quaternion blending. *ACM Transactions on Graphics (TOG)*, 27(4):1–23, 2008.

Binh Huy Le and Zhigang Deng. Robust and accurate skeletal rigging from mesh sequences. *ACM Transactions on Graphics (TOG)*, 33(4):1–10, 2014.

Teng Li, Guangcong Zheng, Rui Jiang, Shuigen Zhan, Tao Wu, Yehao Lu, Yining Lin, Chuanyun Deng, Yepan Xiong, Min Chen, et al. Realcam-i2v: Real-world image-to-video generation with interactive complex camera control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 28785–28796, 2025.

Isabella Liu, Zhan Xu, Yifan Wang, Hao Tan, Zexiang Xu, Xiaolong Wang, Hao Su, and Zifan Shi. Riganything: Template-free autoregressive rigging for diverse 3d assets. *ACM Transactions on Graphics (TOG)*, 44(4):1–12, 2025a.

Lijuan Liu, Youyi Zheng, Di Tang, Yi Yuan, Changjie Fan, and Kun Zhou. Neuroskinning: Automatic skin binding for production characters with deep graph networks. *ACM Transactions on Graphics (ToG)*, 38(4):1–12, 2019.

Shaowei Liu, Saurabh Gupta, and Shenlong Wang. Building rearticulable models for arbitrary 3d objects from 4d point clouds. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21138–21147. IEEE, 2023.

Yu Liu, Baoxiong Jia, Ruijie Lu, Chuyue Gan, Huayu Chen, Junfeng Ni, Song-Chun Zhu, and Siyuan Huang. Videoartgs: Building digital twins of articulated objects from monocular video. *arXiv preprint arXiv:2509.17647*, 2025b.

Xun Long Ng, Kian Eng Ong, Qichen Zheng, Yun Ni, Si Yong Yeo, and Jun Liu. Animal kingdom: A large and diverse dataset for animal behavior understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19023–19034, June 2022.

Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18444–18455, 2023.

Xiaoyu Pan, Jiancong Huang, Jiaming Mai, He Wang, Honglin Li, Tongkui Su, Wenjun Wang, and Xiaogang Jin. Heterskinnet: A heterogeneous network for skin weights prediction. In *Proceedings of the ACM on computer graphics and interactive techniques*, volume 4. Association for Computing Machinery, 2021.

Weiming Ren, Huan Yang, Ge Zhang, Cong Wei, Xinrun Du, Wenhao Huang, and Wenhu Chen. Consisti2v: Enhancing visual consistency for image-to-video generation. *arXiv preprint arXiv:2402.04324*, 2024.

Chaoyue Song, Xiu Li, Fan Yang, Zhongcong Xu, Jiacheng Wei, Fayao Liu, Jiashi Feng, Guosheng Lin, and Jianfeng Zhang. Puppeteer: Rig and animate your 3d models. *arXiv preprint arXiv:2508.10898*, 2025a.

Chaoyue Song, Jianfeng Zhang, Xiu Li, Fan Yang, Yiwen Chen, Zhongcong Xu, Jun Hao Liew, Xiaoyang Guo, Fayao Liu, Jiashi Feng, et al. Magicarticulate: Make your 3d models articulation-ready. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15998–16007, 2025b.

Olga Sorkine, Marc Alexa, et al. As-rigid-as-possible surface modeling. In *Symposium on Geometry processing*, volume 4, pp. 109–116, 2007.

Mingze Sun, Junhao Chen, Junting Dong, Yurun Chen, Xinyu Jiang, Shiwei Mao, Puhua Jiang, Jingbo Wang, Bo Dai, and Ruqi Huang. Drive: Diffusion-based rigging empowers generation of versatile and expressive characters. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21170–21180, 2025a.

Mingze Sun, Shiwei Mao, Keyi Chen, Yurun Chen, Shunlin Lu, Jingbo Wang, Junting Dong, and Ruqi Huang. Armo: Autoregressive rigging for multi-category objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7721–7730, 2025b.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5294–5306, 2025a.

Qianqian Wang, Vickie Ye, Hang Gao, Weijia Zeng, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. Shape of motion: 4d reconstruction from a single video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9660–9672, 2025b.

Yikai Wang, Yinpeng Dong, Fuchun Sun, and Xiao Yang. Root pose decomposition towards generic non-rigid 3d reconstruction with monocular videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13890–13900, 2023.

Zihan Wang, Jeff Tan, Tarasha Khurana, Neehar Peri, and Deva Ramanan. Monofusion: Sparse-view 4d reconstruction via monocular fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8252–8263, 2025c.

Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners. *arXiv preprint arXiv:2509.20328*, 2025.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, Yue Qian, Xiaohang Zhan, and Yueqi Duan. 4d-fly: Fast 4d reconstruction from a single monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16663–16673, 2025.

Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20310–20320, 2024.

Jinbo Xing, Long Mai, Cusuh Ham, Jiahui Huang, Aniruddha Mahapatra, Chi-Wing Fu, Tien-Tsin Wong, and Feng Liu. Motioncanvas: Cinematic shot design with controllable image-to-video generation. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–11, 2025.

Zhan Xu, Yang Zhou, Evangelos Kalogerakis, Chris Landreth, and Karan Singh. Rignet: Neural rigging for articulated characters. *arXiv preprint arXiv:2005.00559*, 2020.

Zhan Xu, Yang Zhou, Li Yi, and Evangelos Kalogerakis. Morig: Motion-aware rigging of character meshes from point clouds. In *SIGGRAPH Asia 2022 conference papers*, pp. 1–9, 2022.

Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. Banmo: Building animatable 3d neural models from many casual videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2863–2873, 2022.

Ze Yang, Shenlong Wang, Sivabalan Manivasagam, Zeng Huang, Wei-Chiu Ma, Xinchen Yan, Ersin Yumer, and Raquel Urtasun. S3: Neural shape, skeleton, and skinning fields for 3d human modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13284–13293, 2021.

Seungwoo Yoo, Juil Koo, Kyeongmin Yeo, and Minhyuk Sung. Neural pose representation learning for generating and transferring non-rigid object poses. *arXiv preprint arXiv:2406.09728*, 2024.

Jia-Peng Zhang, Cheng-Feng Pu, Meng-Hao Guo, Yan-Pei Cao, and Shi-Min Hu. One model to rig them all: Diverse skeleton rigging with unirig. *ACM Transactions on Graphics (TOG)*, 44(4):1–18, 2025.

Kaichen Zhou, Yuhan Wang, Grace Chen, Xinhai Chang, Gaspard Beaudouin, Fangneng Zhan, Paul Pu Liang, and Mengyu Wang. Page-4d: Disentangled pose and geometry estimation for vggt-4d perception. *arXiv preprint arXiv:2510.17568*, 2025.

A Detail for Loss Functions

In the main text, we introduced the unified optimization objective for ZORA. Here, we provide the complete mathematical formulations for the auxiliary coherence priors and regularizers grouped under $\mathcal{L}_{\text{reg}}$ and $\mathcal{L}_{\text{coh}}$. Extended Coherence Priors ($\mathcal{L}_{\text{coh}}$): Beyond the 4D motion affinity ($\mathcal{L}_{\text{aff}}$), our coherence objective incorporates mesh-scale structural priors:

$$\mathcal{L}_{\text{coh}} = \mathcal{L}_{\text{mesh}} + \mathcal{L}_{\text{aff}}.$$

At the mesh scale, neighboring vertices on the rest topology are encouraged to have compatible skinning weights to prevent local tearing:

$$\mathcal{L}_{\text{mesh}} = \frac{1}{|\mathcal{E}_m|} \sum_{(i,j)\in\mathcal{E}_m} \|\mathbf{w}_i - \mathbf{w}_j\|_2^2,$$

where $\mathcal{E}_m$ represents the set of mesh edges.

To suppress temporal jitter in the extracted motion, $\mathcal{L}_{\text{reg}}$ regularizes the second-order finite difference of the learned bone translations $\boldsymbol{\delta}_b^t$:

$$\mathcal{L}_{\text{reg}} = \frac{1}{(T-2)B} \sum_{t=2}^{T-1} \sum_{b=1}^{B} \|\boldsymbol{\delta}_b^{t+1} - 2\boldsymbol{\delta}_b^t + \boldsymbol{\delta}_b^{t-1}\|_2^2.$$

Finally, to prevent degenerate solutions where the network collapses all vertex assignments onto a trivially small subset of bones, we apply a weak bone-usage regularizer based on the mean weight $\bar{w}_b = \frac{1}{V}\sum_{i=1}^{V} w_{ib}$:

$$\mathcal{L}_{\text{bal}} = -\frac{1}{B} \sum_{b=1}^{B} \log(\bar{w}_b + \epsilon).$$

B Exporting an Editable GLB Rig

After optimization, we export the recovered skeleton, skinning weights, and animation into a standard GLB asset using the glTF linear blend skinning formulation. Since glTF supports at most four joint influences per vertex, we retain the top-4 weights for each vertex. Let $\mathcal{N}_4(i) = \text{TopK}_4(\{w_{ib}\}_{b=1}^{B})$ denote the four largest bone weights for vertex i. The exported weights are renormalized as $\hat{w}_{ib} = \frac{w_{ib}}{\sum_{m\in\mathcal{N}_4(i)} w_{im}}$ with $b \in \mathcal{N}_4(i)$, and $\hat{w}_{ib} = 0$ otherwise.

Let $\mathbf{J}_b$ be the rest position of bone b. The inverse bind matrix is

$$\mathbf{B}_b^{-1} = \begin{bmatrix} \mathbf{I} & -\mathbf{J}_b \\ \mathbf{0}^\top & 1 \end{bmatrix}. \tag{8}$$

Method	Wall Time	CPU Memory	GPU Memory	Training Time	Training GPU
Puppeteer	137.39s	9.13 GiB	4.46 GiB	92h	8×A100
RigAnything	40.40s	3.30 GiB	3.68 GiB	-	-
ZORA (ours)	24.99s	10.77 GiB	6.98 GiB	-	no training

Table 4: Computational resources required for rigging one sample on an RTX 4090 GPU and a 13th Gen Intel Core i7 CPU. We don't train the models, but instead collect the checkpoints from RigAnything [2] and Puppeteer [3].

For each frame t, the optimized bone transform is exported as

$$\mathbf{M}_b^t = \begin{bmatrix} \mathbf{R}_b^t & \mathbf{J}_b + \boldsymbol{\delta}_b^t \\ \mathbf{0}^\top & 1 \end{bmatrix}, \tag{9}$$

where $\mathbf{R}_b^t$ is converted to a unit quaternion for glTF keyframes, and $\boldsymbol{\delta}_b^t$ is the learned translation offset relative to the rest joint.

The final skinned vertex position is $\mathbf{v}_i'(t) = \pi(\sum_{b\in\mathcal{N}_4(i)} \hat{w}_{ib}\mathbf{M}_b^t\mathbf{B}_b^{-1}\tilde{\mathbf{v}}_i)$. Substituting the definitions of $\mathbf{M}_b^t$ and $\mathbf{B}_b^{-1}$, each bone contribution becomes $\mathbf{R}_b^t(\mathbf{v}_i - \mathbf{J}_b) + \mathbf{J}_b + \boldsymbol{\delta}_b^t$, which matches the continuous optimization objective. For tree-structured skeletons, the optimized global transforms are converted into parent-relative glTF node transforms before export. Thus, the final GLB contains an editable skeleton hierarchy, top-4 normalized skinning weights, and animation keyframes that reproduce the recovered 4D motion under standard LBS.

C Computing Resources

The computational analysis in Table 4 highlights that while ZORA (ours) requires competitive inference time and memory compared to state-of-the-art baselines, it completely eliminates the substantial overhead of model training, such as the 92-hour A100 training cycle required by Puppeteer. By operating as a training-free, zero-shot framework, our method offers a high-performance alternative that is immediately deployable without the need for extensive high-end GPU clusters or large-scale annotated datasets.

D Detailed Evaluation Metrics

In the main text, we introduced a custom suite of reconstruction-based, motion-based, and structural validity metrics for evaluating rigs extracted from in-the-wild 4D assets. Here, we provide their complete mathematical formulations.

Let $\mathbf{X} = \{\mathbf{X}^t\}_{t=1}^T$, where $\mathbf{X}^t = \{\mathbf{x}_i^t\}_{i=1}^V$, denote the observed 4D mesh sequence.

Let $\tilde{\mathbf{x}}_i^1$ be the canonical homogeneous vertex position. A recovered rig consists of a set of bones $\mathcal{E}$, skinning weights w_{ib} for vertex i and bone b, and per-frame bone transformations $\mathbf{G}_b^t \in SE(3)$. To ensure fair evaluation across arbitrary topologies, we normalize all distance-based metrics by D, the bounding-box diagonal of the canonical mesh.

LBS Reconstruction (E_{LBS}). To measure how well the rig captures the observed motion in a scale-invariant manner, we compute the normalized LBS reconstruction error. The reconstructed vertex is given by: $\hat{\mathbf{x}}_i^t = \pi\left(\sum_{b\in\mathcal{E}} w_{ib}\mathbf{G}_b^t(\mathbf{G}_b^1)^{-1}\tilde{\mathbf{x}}_i^1\right)$, where $\pi(\cdot)$ converts homogeneous coordinates back to $\mathbb{R}^3$. The normalized error is: $E_{\text{LBS}} = \frac{1}{DTV}\sum_{t=1}^T\sum_{i=1}^V \|\hat{\mathbf{x}}_i^t - \mathbf{x}_i^t\|_2$.

Motion Coherence (E_{coh}). A valid bone should explain the motion of its assigned vertices using an approximately rigid transformation. We transform each observed vertex trajectory into the local coordinate frame of bone b: $\mathbf{z}_{ib}^t = \pi\left((\mathbf{G}_b^t)^{-1}\tilde{\mathbf{x}}_i^t\right)$. If vertex i is well-explained by bone b, $\mathbf{z}_{ib}^t$ should remain nearly constant over time. Defining its temporal mean as $\bar{\mathbf{z}}_{ib} = \frac{1}{T}\sum_{t=1}^T \mathbf{z}_{ib}^t$, we compute the motion coherence error as the weighted local-frame variance: $E_{\text{coh}} = \frac{1}{D^2}\frac{\sum_{b\in\mathcal{E}}\sum_{i=1}^V w_{ib}\left(\frac{1}{T}\sum_{t=1}^T \|\mathbf{z}_{ib}^t - \bar{\mathbf{z}}_{ib}\|_2^2\right)}{\sum_{b\in\mathcal{E}}\sum_{i=1}^V w_{ib}}$.

Boundary Violation (E_{viol}). We measure volumetric validity using the signed distance field (SDF) of the animated mesh. Let $s^t(\mathbf{p})$ be the signed distance to the mesh at frame t, where $s^t(\mathbf{p}) > 0$ indicates that point $\mathbf{p}$ lies outside the mesh volume. For each predicted bone $b = (u, v)$, we sample K points along the bone segment at frame t: $\mathbf{p}_{b,k}^t = (1 - \lambda_k)\mathbf{j}_u^t + \lambda_k \mathbf{j}_v^t$, where $\lambda_k = \frac{k-1}{K-1}$. The boundary violation enforces that bones remain inside the mesh: $E_{\text{viol}} = \frac{1}{DT|\mathcal{E}|K} \sum_{t=1}^{T} \sum_{b \in \mathcal{E}} \sum_{k=1}^{K} \max(0, s^t(\mathbf{p}_{b,k}^t))$.

Cross-Space Bones (E_{cross}). We additionally penalize discrete structural failures. For each bone b, we compute its outside-volume ratio: $r_b = \frac{1}{TK} \sum_{t=1}^{T} \sum_{k=1}^{K} \mathbb{I}\left[s^t(\mathbf{p}_{b,k}^t) > \epsilon D\right]$, where ϵ is a small distance tolerance. A bone is classified as a "cross-space bone" if $r_b > \rho$, where ρ is the maximum allowed outside-volume fraction. The cross-space bone score is the average occurrence of these failures: $E_{\text{cross}} = \frac{1}{|-|} \sum_{b \in -} \mathbb{I}[r_b > \rho]$.

Surface Normal Consistency (E_{norm}). While E_{LBS} captures global positional fidelity, it does not explicitly penalize high-frequency surface artifacts such as local folding, crushing, or "candy-wrapper" twisting common in poor skinning assignments. To evaluate local surface preservation, we measure the normal consistency. Let F denote the set of all triangular faces in the mesh. Let $\mathbf{n}_f^t$ be the unit normal vector of face f on the observed 4D mesh at frame t, and $\hat{\mathbf{n}}_f^t$ be the unit normal of the corresponding face reconstructed via our extracted linear blend skinning rig. The normal consistency error is computed as the temporally averaged cosine distance:

$$E_{\text{norm}} = \frac{1}{T|F|} \sum_{t=1}^{T} \sum_{f \in F} \left(1 - \mathbf{n}_f^t \cdot \hat{\mathbf{n}}_f^t\right).$$

Lower E_{norm} indicates that the recovered rig successfully preserves the original geometric details and surface smoothness during dynamic articulation.

Animation MPVE. For datasets with ground-truth sequences, we evaluate the end-to-end animation quality of the recovered rig. Using the predicted skeleton, skinning weights, and per-frame joint transformations, we reconstruct the animated mesh via linear blend skinning and compute the Mean Per-Vertex Error (MPVE), MPVE $= \frac{1}{TV} \sum_{t=1}^{T} \sum_{i=1}^{V} \|\hat{\mathbf{x}}_i^t - \mathbf{x}_i^t\|_2$, where T is the total number of frames in the 4D sequence, and V is the total number of mesh vertices. A lower Animation MPVE indicates that the recovered rig faithfully explains the observed 4D motion.

E More Visualization Results

To further demonstrate the robustness and generalizability of our zero-shot rigging framework, this section provides extended qualitative visualizations across a highly diverse set of data modalities. First, we provide expanded evaluations on in-the-wild video sequences from the *AnimalKingdom* dataset, showcasing successful skeleton extraction across a wide variety of out-of-distribution animal topologies. Second, we present additional results on complex, artist-created 3D sequences from the *ModelsResource* dataset, highlighting our method's precision on stylized geometries. Finally, to demonstrate ZORA's immediate applicability to emerging synthetic media, we apply our framework to high-fidelity, AI-generated videos produced by Veo model. Together, these extended results confirm that our approach consistently extracts structurally sound and kinematically valid rigs regardless of the input source or topological complexity.

E.1 Results on Animal Kingdom

To further validate the robustness of our framework to diverse morphological structures, Figs. 3, 4, and 5 present extended qualitative comparisons on the in-the-wild *AnimalKingdom* dataset. As demonstrated across these extended samples, data-driven baselines exhibit varying degrees of structural failure when faced with out-of-distribution shapes. Specifically, MoRig consistently suffers from complete structural collapse, while RigAnything achieves only sporadic success on a narrow subset of topologies. Notably, even the highly supervised Puppeteer model, which produces high-quality

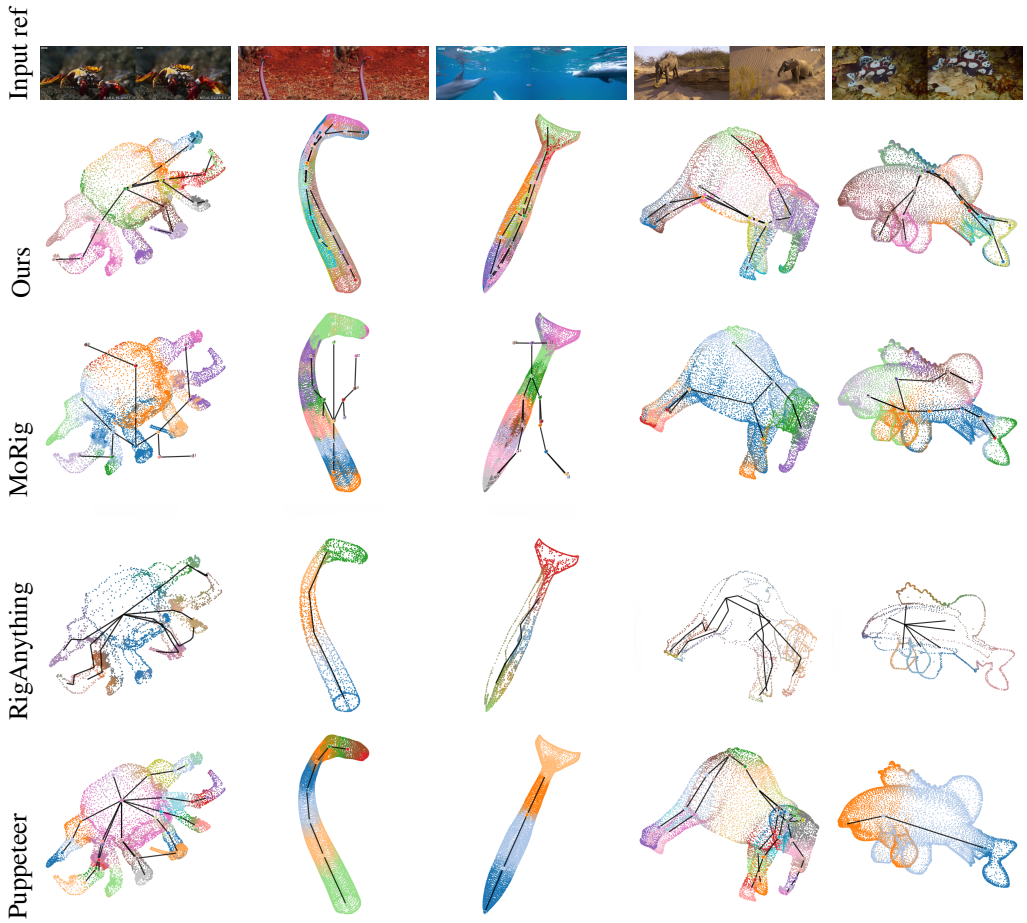


Figure 3: Qualitative basic comparison on selected AnimalKingdom samples. Columns show crab, cobra, dolphin, elephant, and fish.

rigs for common quadrupeds such as sheep, horses, and elephants, remains fundamentally brittle. It systematically fails on aquatic life, insects, and specific terrestrial mammals like foxes, indicating a severe lack of true topological robustness. In contrast, our zero-shot formulation reliably extracts topologically correct and kinematically cohesive rigs across this entire spectrum of species.

E.2 Example from ModelsResources

Figure 7 provides additional qualitative comparisons utilizing stylized, artist-created assets from the *ModelsResource* dataset. Our analysis reveals a clear domain bias in existing state-of-the-art baselines. While heavily supervised models such as Puppeteer and RigAnything successfully recover accurate skeletal structures for standard humanoid subjects, they exhibit severe topological degradation when applied to out-of-distribution morphologies, such as avian meshes. In contrast, our zero-shot formulation relies on universal physical and geometric priors, allowing it to consistently extract high-quality, functionally sound rigs across both humanoid and non-humanoid structures alike.

E.3 Results from Generative Videos

The rapid advancement of text-to-video generative models presents both a challenge and a unique opportunity for 3D character animation. In this section, we explore our framework's bidirectional relationship with generative media: first, its unparalleled ability to rig highly diverse, AI-generated assets, and second, its novel use of generative models to augment incomplete real-world kinematic data.

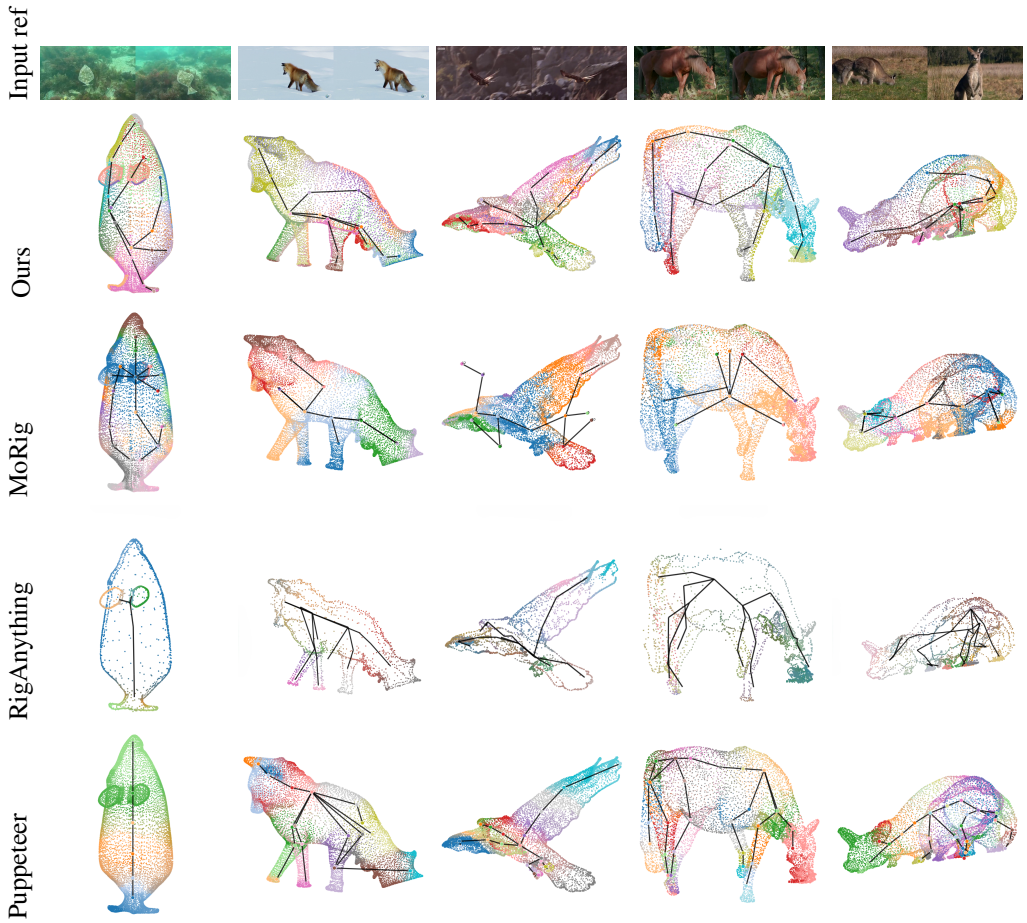


Figure 4: Qualitative basic comparison on selected AnimalKingdom samples. Columns show flatfish, fox, hawk, hourse, and kangaroo.

E.4 Out-of-Distribution Robustness on Synthetic Media

To evaluate out-of-distribution robustness on emerging synthetic media, we visualize rigging results on high-fidelity generated videos (Figure 8). When applied to these generative assets, baselines such as MoRig and RigAnything fail entirely, unable to extract any viable kinematic structure. Furthermore, while heavily supervised baselines like Puppeteer successfully rig standard humanoid subjects, they exhibit severe topological failures when faced with out-of-distribution or fantastical compositions. For instance, when attempting to rig a generated video of a dinosaur, Puppeteer struggles to deduce a valid skeletal topology, resulting in notable structural inconsistencies and an inaccurate rig. More critically, when presented with extreme morphological anomalies, such as the mythological figure Nezha featuring three heads and six arms, Puppeteer's strong dataset bias overrides the visual evidence. It forces a standard bipedal template onto the subject, entirely failing to acknowledge the extra appendages. Conversely, our zero-shot formulation bypasses dataset bias by relying entirely on observed kinematics. Grounded in actual motion, our method successfully extracts all articulating limbs, proving its unparalleled adaptability to the boundless morphological diversity of generative video.

E.5 Generative Kinematic Augmentation for In-the-Wild Captures

A fundamental challenge of motion-driven, zero-shot rigging is that short, in-the-wild video clips frequently capture a limited range of motion. If a subject does not articulate a specific limb during the recording, the optimizer receives no kinematic evidence to discover that joint. To overcome this limitation, we leverage state-of-the-art video generation models to artificially extend the original

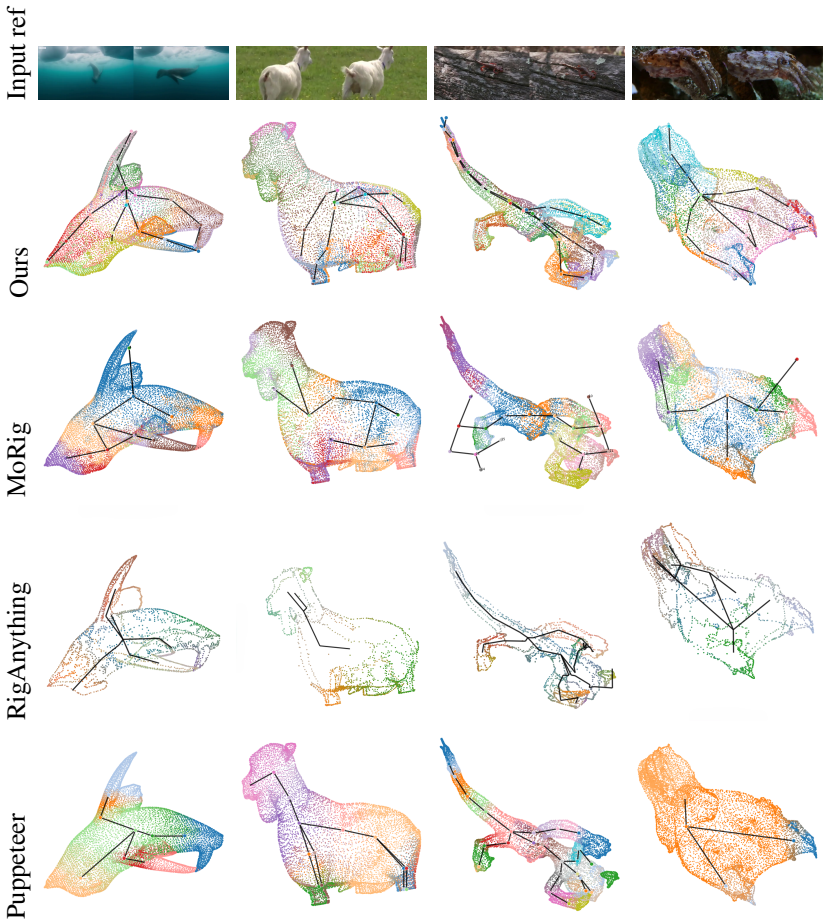


Figure 5: Qualitative basic comparison on selected AnimalKingdom samples. Columns show seal, goat, lizard, and cuttlefish.

footage. By synthesizing novel animations that force the subject into new poses, we generate the missing kinematic signals required for complete skeletal extraction. As shown in Figure 9, augmenting the source video with generative motion allows our framework to successfully discover previously dormant joints, significantly improving the topological completeness and functional quality of the final rig.

F Qualitative Ablations of Structural and Affinity Priors

To evaluate the impact of our intermediate structural prior, we visualize the extracted rigs with and without the Stretch Scaffold in Figure 10. Without the scaffold's explicit distance and topology constraints, the neural optimizer prioritizes naive surface fitting over anatomical logic. Consequently, predicted joints drift away from true articulation points (*e.g.*, elbows and knees), floating aimlessly inside the mesh volume or collapsing entirely. To compensate for this degraded kinematic structure, the optimizer inevitably overfits the skinning assignments. In contrast, our full method successfully anchors joints to kinematically logical hinges, yielding a functional, production-ready rig.

Furthermore, to explain why removing the affinity prior artificially improves certain continuous reconstruction metrics, we visualize the predicted skinning weights in Figure 11. Without affinity, the optimizer severely overfits to the observed poses by assigning fragmented, physically illogical skinning weights to individual vertices to force exact positional alignment. As explicitly highlighted in Figure 11(d) and (e), this lack of regularization causes the underlying skinning structure to give way under novel articulation, resulting in severe local collapse and unnatural deformation. Conversely, our

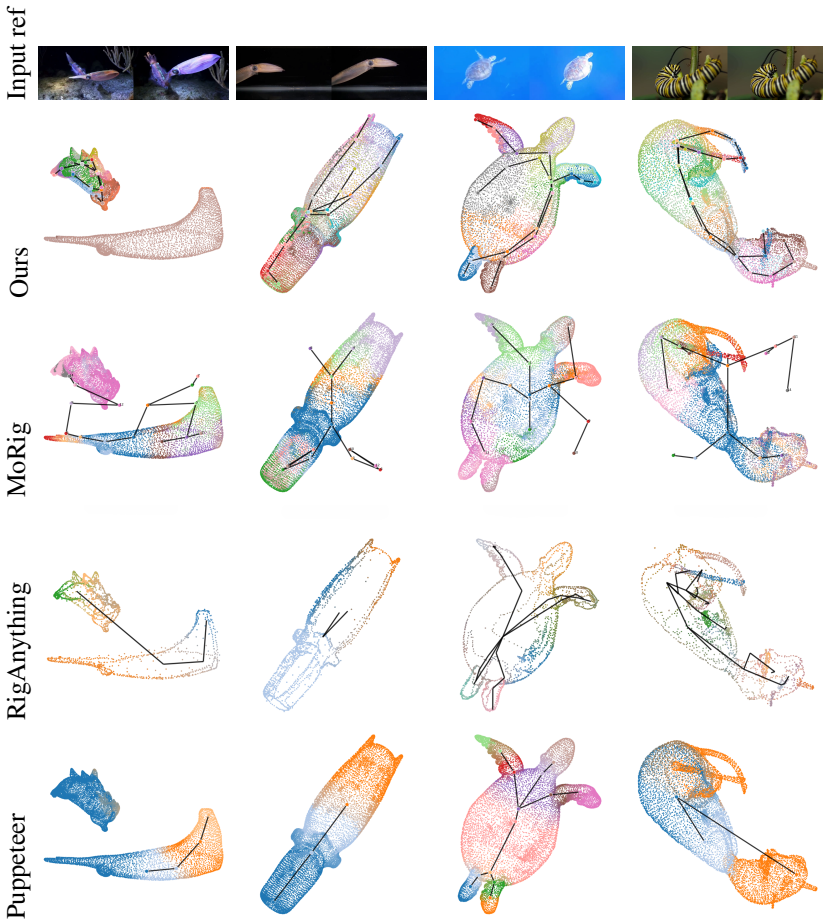


Figure 6: Qualitative basic comparison on selected AnimalKingdom samples. Columns show two squids, partial shown squid, turtle, and caterpillar.

affinity prior guarantees that visually and temporally related surface regions share smooth, continuous skin weights. This ensures that distinct body parts move cohesively, yielding a superior, structurally sound rig despite minor trade-offs in raw positional error.

G THE ROBUSTNESS FOR 4D ARTIFACTS.

Our sensitivity analysis at Table 5 and 6, stress-tests ZORA against the raised failure modes. Because motion-affinity optimization acts as a low-pass filter, ZORA stays stable under high-frequency jitter (even at 30%) and degrades only under deeper structural corruption: correspondence drift beyond 30% raises normal error (E_norm 0.096→0.474), and missing geometry beyond 10% deprives the scaffold of the manifold needed for geodesic routing, roughly doubling the cross-space penalty (E_cross 0.116→0.221). Baseline [a] tolerates jitter but breaks down under the two corruptions typical of in-the-wild reconstruction. Under self-contact (topology corruption), its Euclidean trajectory clustering fuses briefly co-moving parts and its joint solve places articulations through the contact, collapsing Joint F1 from 0.322 to 0.114 at 40%. Under missing geometry, reconstruction fidelity degrades steadily (E_LBS 0.029→0.041) as bones route through gaps. ZORA, which treats motion as a constraint inside the optimization rather than a downstream clustering stage, is built to survive exactly these cases. On self-contact the separation is decisive and confirmed by three independent labelled metrics at 40% corruption: Joint CD (0.062 vs. 0.094, 34% lower), Joint F1 (0.301 vs. 0.114, 2.6×), and Anim MPVE (0.030 vs. 0.046, 34% lower); its stretch-aware scaffold severs spurious contacts before geodesic routing. Under missing geometry, ZORA's reconstruction fidelity holds (E_LBS 0.023 at 40%, at or below clean) while RASR(Le & Deng, 2014) rises to 0.041. Notably, on clean,

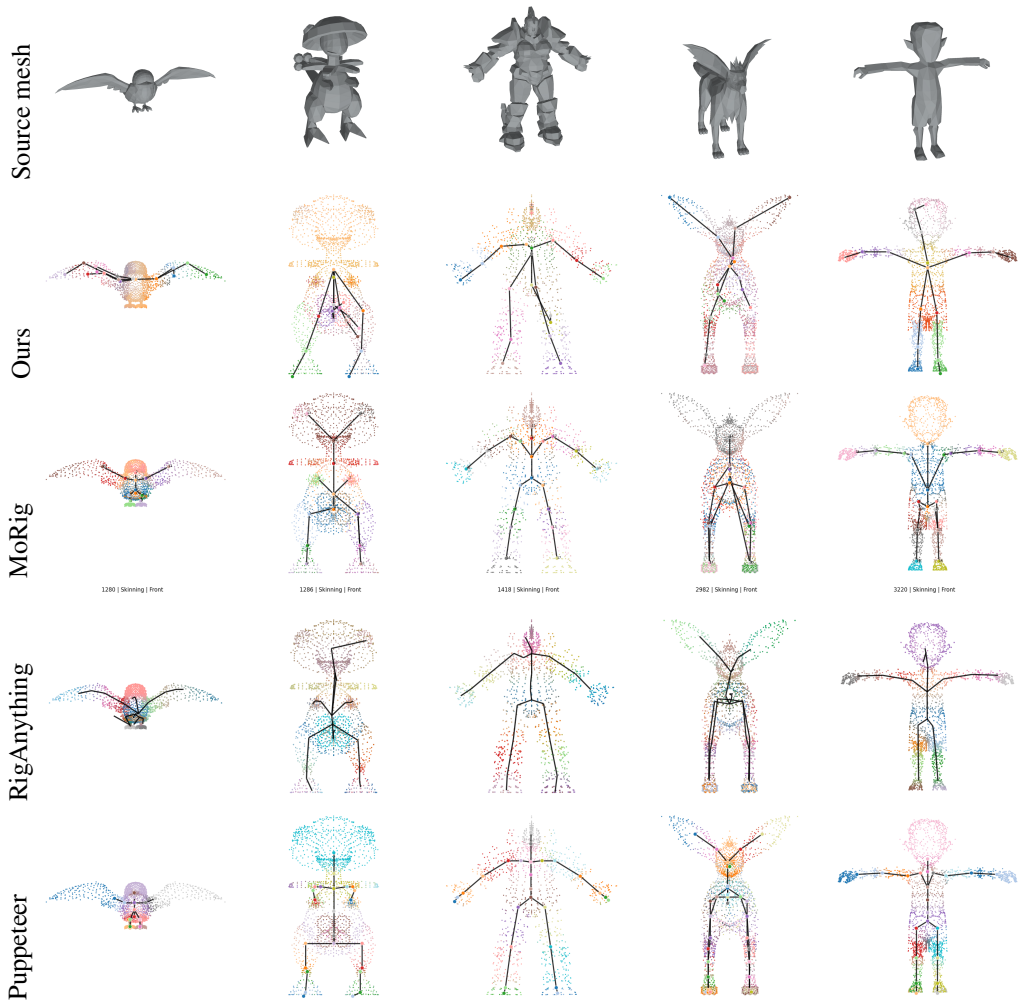


Figure 7: Qualitative basic comparison on selected ModelsResources samples.

well-tracked input [a] is competitive or better (Joint CD 0.050 vs. 0.060; Bone F1 0.372 vs. 0.323), consistent with its design for registered studio scans, the separation emerges only under in-the-wild corruption. The agreement of these ground-truth metrics with the unsupervised E-metrics further indicates the latter track genuine structural quality. In short, the perturbation that most damages [a], self-contact, is precisely the one ZORA is built to absorb.

H Limitation and Future Work

While ZORA effectively extracts kinematics from motion, it operates as a purely geometric downstream process, making it sensitive to the failure modes of the underlying 4D reconstruction. Missing geometry, temporal flicker, or topological noise can mislead the geodesic routing step, resulting in extraneous branches. A promising direction for future work is to jointly optimize the 4D reconstruction and the rig, allowing the articulated structure to regularize the surface reconstruction in return. Additionally, because ZORA lacks semantic awareness, it produces an undirected kinematic scaffold rather than a semantically meaningful joint hierarchy. It cannot infer joint types or recognize specific anatomy, meaning the resulting rigs may not perfectly align with human animation conventions. Incorporating foundation models, such as vision-language models, to guide joint type classification and structural hierarchy remains a natural and exciting next step.



Figure 8: Qualitative basic comparison on selected Generated samples.

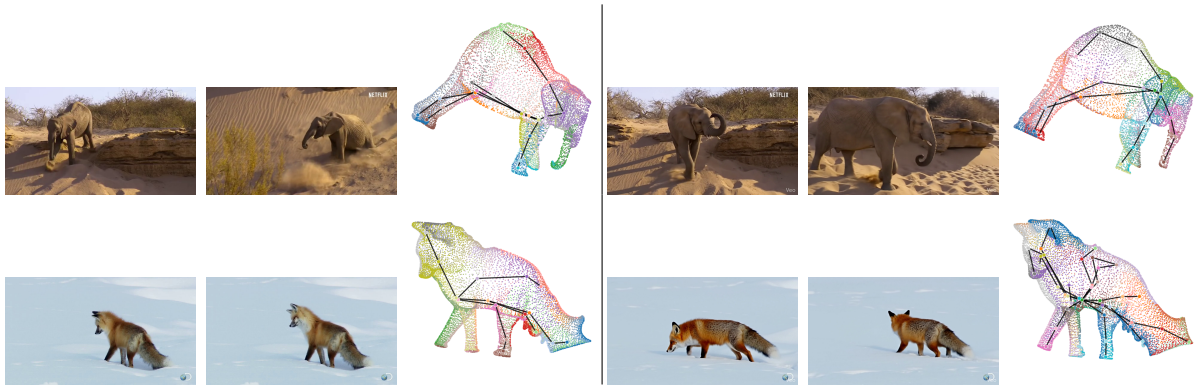
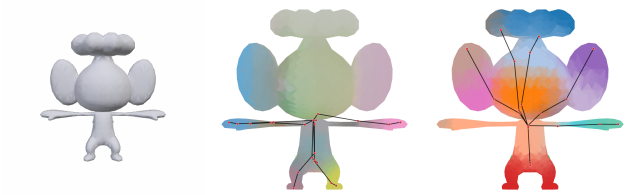


Figure 9: Given a single frame from an original video, video generation can synthesize alternative subsequent frames that lead to different rigging results. Left: It is the original video from Animal Kingdom to rig. Right: It is the generated video to rig

I HYPER-PARAMETER SETTING

The behavior of motion-driven rig recovery can vary with the scale, topology, and motion range of the input sequence. We therefore report the main hyper-parameters of our formulation and provide one reasonable reference setting that can be used as a starting point. These values should not be



Source mesh Full binding w/o Stretch binding

Figure 10: Effect of Stretch Scaffold on sample 14760. The columns show the source mesh and rest-pose skinning results for the full and ablated variants. Removing Stretch Scaffold reduces E_{LBS} from 0.02448 to 0.01953, but slightly increases E_{norm} from 0.07260 to 0.07309, evaluated over all six observed frames. Joint accuracy also deteriorates: Joint F1 drops from 0.5000 to 0.0870, and Joint CD increases from 0.0212 to 0.0766. This example shows that lower reconstruction error does not necessarily imply more accurate joint placement.



(a) Input mesh (b) Full (c) w/o affinity (d) Full (e) w/o affinity

Figure 11: **Impact of the affinity prior on skinning weight distribution.** (a) Input stylized mesh. (b) Our full formulation enforces physical cohesion, yielding smooth, contiguous global skinning weights. (c) Without the affinity prior, the optimizer severely overfits to the observed 4D dynamics by fragmenting the skinning assignments. To explicitly highlight this failure, we isolate the influence of the exact same bone in (d) and (e). While the full model (d) maintains logical surface boundaries, the unregularized model (e) fragments skinning weights to force an exact fit, a flaw that causes the mesh to collapse when animated.

interpreted as a unique optimum for every asset; changing them can alter the extracted scaffold, skinning smoothness, and reconstruction–regularization trade-off.

For the stretch-aware scaffold, K controls the number of extremities selected by graph farthest-point sampling, and thus influences the number of root–extremity branches. The path sampling step controls the density of joint candidates along each branch. The bridge radius determines the local proximity edges added to the mesh graph, while the stretch penalty α controls how strongly temporally inconsistent shortcuts are discouraged. The inward offset moves surface-sampled joint candidates toward the interior of the shape, and ϵ is used only as a numerical floor for stable stretch normalization.

For the unified rigging objective, the weights $\lambda_{\mathrm{rec}}, \lambda_{\mathrm{loc}}, \lambda_{\mathrm{coh}}, \lambda_{\mathrm{bal}}$ respectively balance LBS reconstruction fidelity, topological locality, mesh/motion-affinity coherence, and bone-usage regularization. The coherence blend weights $w_{\mathrm{mesh}}, w_{\mathrm{aff}}$ control the relative contribution of mesh-scale smoothness and motion-affinity consistency. The affinity top-k, affinity threshold, and affinity geodesic radius define the sparse high-confidence affinity graph used by the coherence term. The locality radius and cutoff specify the geodesic support of vertex–bone assignments. The initialization scale controls the spread of the initial skinning logits, and the softmax temperature controls the sharpness of the skinning distribution during optimization.

Table 7 lists a practical reference setting. The most sensitive parameters are typically K, α, λ_{loc}, λ_{coh}, and the affinity sparsification parameters, since these directly affect whether nearby but independently moving parts are separated or smoothed together. For glTF export, we follow the standard four-influence limit by retaining the top-4 skinning weights per vertex and renormalizing them before writing the final GLB.

Table 5: ZORA rigging under different noise conditions on ModelsResource. TJ: Temporal jitter; Corr: Correspondence; MP: Missing parts; TN: Topology noise; Occ: Occlusion

Noise	Joint CD ↓	Joint F1 ↑	Bone F1 ↑	Skin L1 ↓	IoU ↑	MPVE ↓	E_{LBS} ↓	E_{norm} ↓	E_{coh} ↓	E_{viol} ↓	E_{cross} ↓
Clean	0.0602 ± 0.0196	0.3338 ± 0.1343	0.3227 ± 0.1286	1.5155 ± 0.1733	0.1350 ± 0.0604	0.0299 ± 0.0103	0.0297 ± 0.0101	0.0966 ± 0.0396	0.0277 ± 0.0090	0.0054 ± 0.0033	0.1162 ± 0.1132
TJ 1%	0.0598 ± 0.0197	0.3271 ± 0.1253	0.3405 ± 0.1346	1.4878 ± 0.1726	0.1436 ± 0.0582	0.0295 ± 0.0108	0.0293 ± 0.0106	0.0971 ± 0.0413	0.0278 ± 0.0098	0.0061 ± 0.0046	0.1311 ± 0.1269
TJ 5%	0.0560 ± 0.0195	0.3256 ± 0.1346	0.3376 ± 0.1374	1.4341 ± 0.1959	0.1653 ± 0.0679	0.0275 ± 0.0109	0.0273 ± 0.0107	0.0990 ± 0.0418	0.0254 ± 0.0094	0.0057 ± 0.0038	0.1282 ± 0.1207
TJ 10%	0.0564 ± 0.0194	0.3136 ± 0.1200	0.3624 ± 0.1353	1.4122 ± 0.1975	0.1658 ± 0.0664	0.0269 ± 0.0101	0.0268 ± 0.0097	0.1097 ± 0.0414	0.0247 ± 0.0084	0.0055 ± 0.0037	0.1113 ± 0.1153
TJ 30%	0.0576 ± 0.0201	0.2993 ± 0.1219	0.3767 ± 0.1367	1.3866 ± 0.1865	0.1708 ± 0.0671	0.0263 ± 0.0099	0.0267 ± 0.0095	0.1927 ± 0.0476	0.0244 ± 0.0082	0.0055 ± 0.0041	0.1228 ± 0.1256
TJ 40%	0.0584 ± 0.0201	0.2886 ± 0.1234	0.3660 ± 0.1253	1.4140 ± 0.1700	0.1685 ± 0.0660	0.0265 ± 0.0104	0.0273 ± 0.0099	0.2444 ± 0.0505	0.0249 ± 0.0085	0.0052 ± 0.0035	0.1105 ± 0.1090
Corr 1%	0.0743 ± 0.0225	0.2702 ± 0.1103	0.2492 ± 0.1015	1.6190 ± 0.1695	0.0900 ± 0.0434	0.0323 ± 0.0108	0.0322 ± 0.0106	0.1152 ± 0.0394	0.0291 ± 0.0105	0.0044 ± 0.0042	0.0762 ± 0.1046
Corr 5%	0.0761 ± 0.0243	0.2489 ± 0.1065	0.2399 ± 0.1035	1.6003 ± 0.1809	0.0925 ± 0.0471	0.0320 ± 0.0112	0.0323 ± 0.0106	0.1751 ± 0.0361	0.0298 ± 0.0115	0.0048 ± 0.0050	0.0935 ± 0.1450
Corr 10%	0.0742 ± 0.0229	0.2577 ± 0.1063	0.2429 ± 0.1147	1.5876 ± 0.1884	0.0956 ± 0.0564	0.0319 ± 0.0109	0.0328 ± 0.0104	0.2447 ± 0.0326	0.0306 ± 0.0111	0.0048 ± 0.0052	0.0807 ± 0.1119
Corr 30%	0.0689 ± 0.0196	0.2626 ± 0.1249	0.2621 ± 0.1041	1.5302 ± 0.1880	0.1272 ± 0.0619	0.0304 ± 0.0114	0.0336 ± 0.0104	0.4749 ± 0.0253	0.0317 ± 0.0105	0.0054 ± 0.0055	0.0839 ± 0.1157
Corr 40%	0.0684 ± 0.0215	0.2622 ± 0.1246	0.2785 ± 0.1059	1.5292 ± 0.2014	0.1272 ± 0.0641	0.0295 ± 0.0111	0.0338 ± 0.0098	0.5621 ± 0.0218	0.0317 ± 0.0089	0.0053 ± 0.0045	0.0957 ± 0.1205
MP 1%	0.0656 ± 0.0211	0.2847 ± 0.1209	0.2931 ± 0.1181	1.5576 ± 0.1670	0.1135 ± 0.0571	0.0306 ± 0.0102	0.0303 ± 0.0098	0.1053 ± 0.0399	0.0285 ± 0.0093	0.0059 ± 0.0034	0.1276 ± 0.1114
MP 5%	0.0748 ± 0.0256	0.2543 ± 0.1300	0.2669 ± 0.1337	1.5894 ± 0.1780	0.0968 ± 0.0587	0.0313 ± 0.0109	0.0305 ± 0.0096	0.1366 ± 0.0814	0.0287 ± 0.0099	0.0058 ± 0.0035	0.1117 ± 0.1049
MP 10%	0.0812 ± 0.0290	0.2380 ± 0.1308	0.2758 ± 0.1109	1.6312 ± 0.1626	0.0864 ± 0.0549	0.0334 ± 0.0118	0.0315 ± 0.0103	0.1533 ± 0.1096	0.0297 ± 0.0108	0.0073 ± 0.0046	0.1600 ± 0.1535
MP 30%	0.0918 ± 0.0314	0.2122 ± 0.1167	0.2383 ± 0.1059	1.6932 ± 0.1610	0.0670 ± 0.0538	0.0372 ± 0.0142	0.0264 ± 0.0098	0.1783 ± 0.0974	0.0270 ± 0.0104	0.0092 ± 0.0069	0.2215 ± 0.2087
MP 40%	0.1005 ± 0.0514	0.1963 ± 0.1099	0.2478 ± 0.1142	1.6904 ± 0.1632	0.0682 ± 0.0609	0.0384 ± 0.0129	0.0234 ± 0.0109	0.1995 ± 0.1144	0.0240 ± 0.0123	0.0092 ± 0.0081	0.2222 ± 0.2370
TN 1%	0.0603 ± 0.0196	0.3331 ± 0.1384	0.3222 ± 0.1252	1.5159 ± 0.1712	0.1356 ± 0.0606	0.0299 ± 0.0103	0.0297 ± 0.0101	0.0964 ± 0.0395	0.0277 ± 0.0091	0.0055 ± 0.0034	0.1146 ± 0.1117
TN 5%	0.0603 ± 0.0198	0.3386 ± 0.1328	0.3217 ± 0.1253	1.5095 ± 0.1686	0.1348 ± 0.0578	0.0301 ± 0.0104	0.0298 ± 0.0102	0.0967 ± 0.0404	0.0276 ± 0.0092	0.0056 ± 0.0034	0.1148 ± 0.1153
TN 10%	0.0606 ± 0.0198	0.3356 ± 0.1403	0.3204 ± 0.1376	1.5005 ± 0.1786	0.1340 ± 0.0551	0.0301 ± 0.0105	0.0298 ± 0.0103	0.0964 ± 0.0395	0.0279 ± 0.0094	0.0059 ± 0.0040	0.1204 ± 0.1182
TN 30%	0.0626 ± 0.0195	0.2960 ± 0.1155	0.3162 ± 0.1280	1.5156 ± 0.1901	0.1300 ± 0.0571	0.0302 ± 0.0104	0.0300 ± 0.0102	0.0955 ± 0.0382	0.0275 ± 0.0092	0.0068 ± 0.0035	0.1230 ± 0.1240
TN 40%	0.0622 ± 0.0193	0.3005 ± 0.1210	0.3055 ± 0.0967	1.5338 ± 0.1614	0.1321 ± 0.0508	0.0303 ± 0.0105	0.0301 ± 0.0104	0.0951 ± 0.0389	0.0280 ± 0.0094	0.0076 ± 0.0055	0.1329 ± 0.1262
Occ 1%	0.0756 ± 0.0260	0.2267 ± 0.1251	0.2199 ± 0.0884	1.6229 ± 0.1554	0.0881 ± 0.0504	0.0313 ± 0.0101	0.0307 ± 0.0096	0.1024 ± 0.0379	0.0308 ± 0.0122	0.0094 ± 0.0074	0.1983 ± 0.1818
Occ 5%	0.0803 ± 0.0320	0.2318 ± 0.1270	0.2573 ± 0.1020	1.6017 ± 0.1632	0.0963 ± 0.0577	0.0326 ± 0.0113	0.0302 ± 0.0099	0.0993 ± 0.0381	0.0287 ± 0.0092	0.0058 ± 0.0047	0.1100 ± 0.1457
Occ 10%	0.0876 ± 0.0336	0.2163 ± 0.1250	0.2364 ± 0.1010	1.6350 ± 0.1534	0.0861 ± 0.0585	0.0340 ± 0.0110	0.0304 ± 0.0094	0.0982 ± 0.0385	0.0301 ± 0.0108	0.0053 ± 0.0066	0.0746 ± 0.1119
Occ 30%	0.0901 ± 0.0307	0.2117 ± 0.1159	0.2354 ± 0.0926	1.6105 ± 0.2023	0.0824 ± 0.0576	0.0360 ± 0.0111	0.0283 ± 0.0083	0.0926 ± 0.0359	0.0278 ± 0.0083	0.0038 ± 0.0030	0.0560 ± 0.0846
Occ 40%	0.0827 ± 0.0300	0.2233 ± 0.1033	0.2474 ± 0.0974	1.6098 ± 0.2013	0.0804 ± 0.0559	0.0375 ± 0.0117	0.0266 ± 0.0076	0.0860 ± 0.0359	0.0268 ± 0.0079	0.0039 ± 0.0028	0.0452 ± 0.0699

Table 6: RASR rigging under different noise conditions on ModelsResource. TJ: Temporal jitter; Corr: Correspondence; MP: Missing parts; TN: Topology noise; Occ: Occlusion

Noise	Joint CD ↓	Joint F1 ↑	Bone F1 ↑	Skin L1 ↓	IoU ↑	MPVE ↓	E_{LBS} ↓	E_{norm} ↓	E_{coh} ↓	E_{viol} ↓	E_{cross} ↓
Clean	0.0503 ± 0.0175	0.3222 ± 0.1485	0.3719 ± 0.1261	1.6554 ± 0.1755	0.1145 ± 0.0679	0.0290 ± 0.0129	0.0290 ± 0.0129	0.1571 ± 0.0810	0.0273 ± 0.0112	$0.0156 \pm 0.0099^{\dagger}$	$0.2067 \pm 0.1692^{\dagger}$
TJ 1%	0.0503 ± 0.0175	0.3222 ± 0.1485	0.3719 ± 0.1261	1.6459 ± 0.1729	0.1171 ± 0.0719	0.0276 ± 0.0119	0.0276 ± 0.0119	0.1500 ± 0.0707	0.0265 ± 0.0113	0.0040 ± 0.0037	0.0702 ± 0.0924
TJ 5%	0.0503 ± 0.0175	0.3222 ± 0.1485	0.3719 ± 0.1261	1.6443 ± 0.1798	0.1187 ± 0.0718	0.0284 ± 0.0125	0.0284 ± 0.0125	0.1630 ± 0.0814	0.0273 ± 0.0114	0.0042 ± 0.0039	0.0761 ± 0.0976
TJ 10%	0.0503 ± 0.0175	0.3222 ± 0.1485	0.3719 ± 0.1261	1.6496 ± 0.1581	0.1163 ± 0.0670	0.0279 ± 0.0114	0.0279 ± 0.0114	0.1776 ± 0.0799	0.0267 ± 0.0104	0.0042 ± 0.0038	0.0764 ± 0.1017
TJ 30%	0.0503 ± 0.0175	0.3222 ± 0.1485	0.3719 ± 0.1261	1.6538 ± 0.1522	0.1122 ± 0.0664	0.0285 ± 0.0123	0.0285 ± 0.0123	0.2166 ± 0.0994	0.0279 ± 0.0118	0.0042 ± 0.0041	0.0738 ± 0.0970
TJ 40%	0.0503 ± 0.0175	0.3222 ± 0.1485	0.3719 ± 0.1261	1.6505 ± 0.1412	0.1096 ± 0.0607	0.0279 ± 0.0119	0.0279 ± 0.0119	0.2277 ± 0.0986	0.0278 ± 0.0107	0.0041 ± 0.0042	0.0736 ± 0.0948
Corr 1%	0.0503 ± 0.0175	0.3222 ± 0.1485	0.3719 ± 0.1261	1.6478 ± 0.1719	0.1180 ± 0.0680	0.0274 ± 0.0116	0.0274 ± 0.0116	0.1678 ± 0.0748	0.0266 ± 0.0115	0.0042 ± 0.0041	0.0762 ± 0.1009
Corr 5%	0.0503 ± 0.0175	0.3222 ± 0.1485	0.3719 ± 0.1261	1.6561 ± 0.1538	0.1054 ± 0.0629	0.0264 ± 0.0115	0.0264 ± 0.0115	0.2115 ± 0.0812	0.0270 ± 0.0107	0.0039 ± 0.0033	0.0720 ± 0.0957
Corr 10%	0.0503 ± 0.0175	0.3222 ± 0.1485	0.3719 ± 0.1261	1.6576 ± 0.1464	0.1016 ± 0.0587	0.0268 ± 0.0109	0.0268 ± 0.0109	0.2548 ± 0.0862	0.0278 ± 0.0107	0.0040 ± 0.0038	0.0709 ± 0.0965
Corr 30%	0.0503 ± 0.0175	0.3222 ± 0.1485	0.3719 ± 0.1261	1.6584 ± 0.1323	0.0936 ± 0.0515	0.0268 ± 0.0122	0.0268 ± 0.0122	0.3319 ± 0.1054	0.0288 ± 0.0119	0.0040 ± 0.0041	0.0746 ± 0.1028
Corr 40%	0.0503 ± 0.0175	0.3222 ± 0.1485	0.3719 ± 0.1261	1.6516 ± 0.1282	0.0897 ± 0.0470	0.0273 ± 0.0121	0.0273 ± 0.0121	0.3491 ± 0.1090	0.0292 ± 0.0117	0.0042 ± 0.0042	0.0751 ± 0.1022
MP 1%	0.0511 ± 0.0168	0.3189 ± 0.1420	0.3675 ± 0.1170	1.6642 ± 0.1624	0.1123 ± 0.0674	0.0290 ± 0.0140	0.0290 ± 0.0140	0.1641 ± 0.0879	0.0274 ± 0.0120	0.0042 ± 0.0041	0.0728 ± 0.1059
MP 5%	0.0517 ± 0.0173	0.3048 ± 0.1408	0.3690 ± 0.1286	1.6570 ± 0.1682	0.1153 ± 0.0792	0.0302 ± 0.0124	0.0302 ± 0.0124	0.1854 ± 0.1017	0.0280 ± 0.0103	0.0042 ± 0.0036	0.0717 ± 0.0865
MP 10%	0.0571 ± 0.0196	0.2870 ± 0.1501	0.3541 ± 0.1260	1.6831 ± 0.1493	0.0997 ± 0.0598	0.0337 ± 0.0195	0.0337 ± 0.0195	0.2161 ± 0.1413	0.0302 ± 0.0178	0.0044 ± 0.0041	0.0772 ± 0.1060
MP 30%	0.0726 ± 0.0230	0.2003 ± 0.1252	0.3316 ± 0.1110	1.7106 ± 0.1976	0.0813 ± 0.0663	0.0386 ± 0.0228	0.0386 ± 0.0228	0.2290 ± 0.1158	0.0336 ± 0.0202	0.0055 ± 0.0045	0.1070 ± 0.1203
MP 40%	0.0753 ± 0.0259	0.2003 ± 0.1302	0.3516 ± 0.1152	1.7307 ± 0.1770	0.0749 ± 0.0603	0.0411 ± 0.0247	0.0411 ± 0.0247	0.2401 ± 0.1147	0.0358 ± 0.0208	0.0066 ± 0.0051	0.1377 ± 0.1381
TN 1%	0.0517 ± 0.0176	0.3240 ± 0.1399	0.3742 ± 0.1291	1.6552 ± 0.1710	0.1221 ± 0.0793	0.0291 ± 0.0145	0.0291 ± 0.0145	0.1580 ± 0.0908	0.0278 ± 0.0127	0.0041 ± 0.0037	0.0745 ± 0.1025
TN 5%	0.0531 ± 0.0172	0.2863 ± 0.1259	0.3284 ± 0.1269	1.6088 ± 0.2066	0.1253 ± 0.0715	0.0288 ± 0.0117	0.0288 ± 0.0117	0.1548 ± 0.0687	0.0274 ± 0.0102	0.0046 ± 0.0048	0.0732 ± 0.0952
TN 10%	0.0569 ± 0.0202	0.2434 ± 0.1185	0.3090 ± 0.1147	1.6193 ± 0.2083	0.1230 ± 0.0605	0.0294 ± 0.0126	0.0294 ± 0.0126	0.1612 ± 0.0813	0.0284 ± 0.0117	0.0055 ± 0.0057	0.0934 ± 0.1139
TN 30%	0.0847 ± 0.0290	0.1492 ± 0.0864	0.3443 ± 0.0924	1.7232 ± 0.1704	0.0870 ± 0.0565	0.0419 ± 0.0268	0.0419 ± 0.0268	0.2314 ± 0.1393	0.0405 ± 0.0247	0.0052 ± 0.0043	0.0863 ± 0.1115
TN 40%	0.0940 ± 0.0296	0.1140 ± 0.0632	0.3497 ± 0.1100	1.7323 ± 0.1517	0.0857 ± 0.0555	0.0459 ± 0.0294	0.0459 ± 0.0294	0.2576 ± 0.1427	0.0457 ± 0.0282	0.0050 ± 0.0040	0.0860 ± 0.1355
Occ 1%	0.0503 ± 0.0175	0.3222 ± 0.1485	0.3719 ± 0.1261	1.6654 ± 0.1595	0.1067 ± 0.0645	0.0299 ± 0.0133	0.0299 ± 0.0133	0.1675 ± 0.0767	0.0277 ± 0.0121	0.0043 ± 0.0041	0.0780 ± 0.1049
Occ 5%	0.0503 ± 0.0175	0.3222 ± 0.1485	0.3719 ± 0.1261	1.6722 ± 0.1510	0.1059 ± 0.0578	0.0298 ± 0.0141	0.0298 ± 0.0141	0.1657 ± 0.0819	0.0271 ± 0.0121	0.0041 ± 0.0037	0.0765 ± 0.1045
Occ 10%	0.0503 ± 0.0175	0.3222 ± 0.1485	0.3719 ± 0.1261	1.6657 ± 0.1585	0.1068 ± 0.0633	0.0287 ± 0.0108	0.0287 ± 0.0108	0.1667 ± 0.0723	0.0262 ± 0.0099	0.0041 ± 0.0040	0.0715 ± 0.1002
Occ 30%	0.0503 ± 0.0175	0.3222 ± 0.1485	0.3719 ± 0.1261	1.6675 ± 0.1434	0.0961 ± 0.0565	0.0298 ± 0.0126	0.0298 ± 0.0126	0.1811 ± 0.0789	0.0274 ± 0.0113	0.0042 ± 0.0040	0.0752 ± 0.1076
Occ 40%	0.0503 ± 0.0175	0.3222 ± 0.1485	0.3719 ± 0.1261	1.6703 ± 0.1455	0.0975 ± 0.0533	0.0290 ± 0.0118	0.0290 ± 0.0118	0.1812 ± 0.0767	0.0271 ± 0.0103	0.0043 ± 0.0041	0.0806 ± 0.1075

Hyper-parameter	Reference value	Meaning
K	5	Number of Graph-FPS extremities for scaffold extraction
Path sampling step	15	Joint sampling stride along root–extremity paths
Bridge radius	0.02	Local proximity radius, relative to max mesh extent
α	100	Stretch-aware edge penalty strength
Inward offset	0.03	Joint inward projection distance, relative to max mesh extent
ϵ	10^{-4}	Numerical floor for stretch normalization
Optimization steps	1000	Number of gradient-based rigging optimization iterations
Learning rate	10^{-2}	Adam learning rate
λ_{rec}	10	LBS reconstruction weight
λ_{loc}	50	Topological locality weight
λ_{coh}	80	Mesh/affinity coherence weight
λ_{bal}	1	Bone-usage balance weight
w_{mesh}	0.5	Mesh-scale coherence blend weight
w_{aff}	0.5	Motion-affinity coherence blend weight
Affinity top-k	32	Number of retained affinity neighbors per vertex
Affinity threshold	0.20	Minimum retained affinity value
Affinity geodesic radius	$0.35D$	Local geodesic support for affinity pairs
Locality radius	$0.25D$	Hard vertex–bone geodesic mask radius
Locality cutoff	0.30	Soft locality cutoff as a fraction of the mask radius
Initialization scale	0.05	Gaussian skinning-logit scale relative to mean vertex–joint distance
Softmax temperature	$2.0 \rightarrow 1.0$	Annealed skinning-weight temperature
Export top-k	4	Number of glTF joint influences retained per vertex

Table 7: Reference hyper-parameter setting for ZORA. D denotes the bounding-box diagonal of the canonical mesh. These values are intended as a practical starting point rather than a claim of a universally optimal configuration.