ANIMALGS: 4D ANIMAL RECONSTRUCTION FROM MONOCULAR VIDEO WITH 3D GAUSSIAN SPLATTING

Anonymous authors
Paper under double-blind review

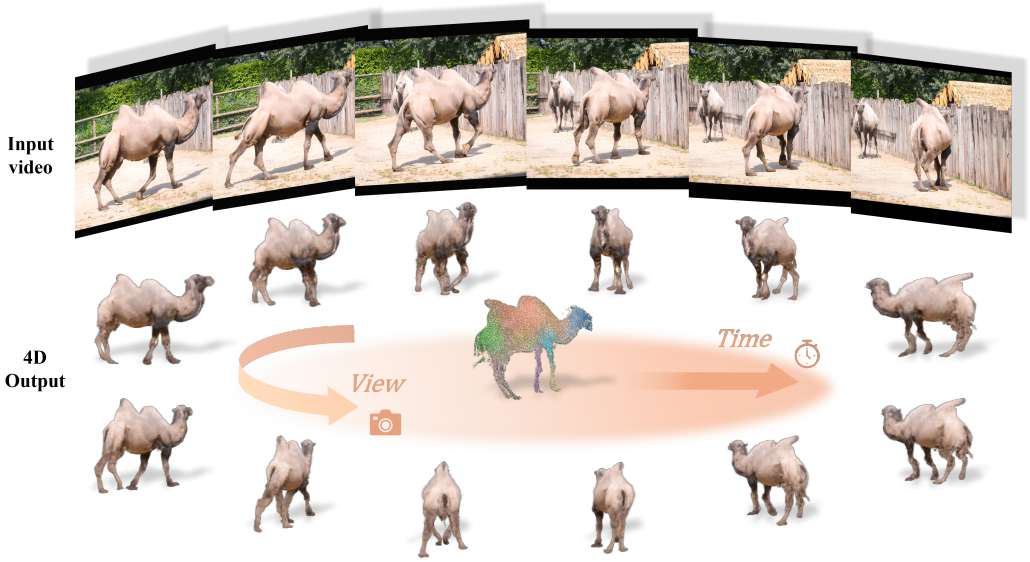

Figure 1: Given a monocular video of an animal (top), **AnimalGS** produces a high-fidelity, consistent 4D model (bottom), enabling free-viewpoint rendering across both time and viewing angles. Center: canonical 3D Gaussians colored by skinning weights.

ABSTRACT

Reconstructing 4D animals from monocular videos is challenging due to large inter-species variation, complex articulations, and the lack of reliable templates. We introduce **AnimalGS**, a test-time optimization framework built on a 3D Gaussian Splatting representation for high-fidelity 4D reconstructions from single videos. Grounded in the insight that robust reconstruction emerges from pose-guided optimization rather than strict shape priors, AnimalGS treats priors as coarse initializations and integrates joint-aware and symmetry-aware designs to progressively disentangle motion and appearance. This leads to empirically strong generalization across diverse species and robustness to mismatching with shape priors. Extensive experiments demonstrate the superior performance of our approach [1] in geometry, motion, and temporal consistency across a wide variety of animal species.

1 INTRODUCTION

Animals in the natural world display a stunning diversity of shapes and behaviors. Accurately reconstructing their 3D shape and motion from visual data is crucial for various applications ranging from wildlife monitoring, animal conservation and ethology research, to immersive media content creation. Despite the wide accessibility of monocular video, the task of creating realistic 4D animal models from monocular video presents a significant challenge in computer vision. This is primarily

[1] Our code and results are to be published upon paper acceptance.

due to the inherent complexity of animal morphology and behaviors, as well as the fact that their appearance and motion are only partly observable from a monocular video.

The task of animal reconstruction presents unique challenges compared to 3D human reconstruction. Human models, such as SMPL (Loper et al., 2023; Pavlakos et al., 2019), benefit from well-studied anatomical structures and abundant 3D motion capture datasets. In contrast, animals of diverse species, ranging from camels, elephants to birds, exhibit extreme shape and motion variations, yet very little animal motion capture benchmarks are available. The pioneer SMAL model (Zuffi et al., 2017) is a parametric SMPL-like animal model learned from a limited collection of toy figurines; it captures a rather limited category of species and lacks realistic details of shape and motion. The dilemma of both scarcely labeled, partially observable animal data, and extraordinarily diverse shape & motion variations across animal species, forces a trade-off where, category-specific methods (Badger et al., 2020; Wang et al., 2021; Wu et al., 2023; Rueegg et al., 2022; Rüegg et al., 2023) achieve higher reconstruction quality by training on annotated dataset but struggle with generalization, while category-agnostic approaches (Li et al., 2024; Aygun & Mac Aodha, 2024; Jakab et al., 2024) improve coverage at the cost of reconstruction fidelity. It motivates us to consider an alternative pathway of learning-based test-time optimization that does not require training from a labeled dataset, except for having a prior model to facilitate our initial shape reconstruction.

Extending to 4D reconstruction from monocular videos reveals a fundamental tension between representation flexibility and computational efficiency. Mesh-based methods (Yang et al., 2021a;b; Sabathier et al., 2024) are efficient but topologically constrained, while neural implicit methods (Yang et al., 2022; 2023a) offer flexibility at prohibitive computational costs. Recent 3D Gaussian Splatting approaches (Kerbl et al., 2023; Lei et al., 2024) provide a middle ground but still rely on parametric templates. Diffusion-driven methods (Ren et al., 2024a; Jiang et al., 2025) achieve impressive synthesis but sacrifice input fidelity. Existing methods either depend on rigid templates that limit generalization or generative priors that compromise reconstruction accuracy. We argue this trade-off stems from treating shape priors as strict constraints rather than flexible initializations.

Inspired by the above observations, we present **AnimalGS**, a pose-guided test-time optimization framework built on 3D Gaussian splatting. Our key insight is that robust 4D animal reconstruction is not dependent on highly accurate shape priors, which is contrary to common assumptions. Specifically, AnimalGS treats the animal shape prior as a coarse initialization and employs a hierarchical two-stage strategy: first, articulated motion is refined using joint-aware anchors together with a symmetry-aware temporal encoding that exploits bilateral cues to stabilize poses; second, non-rigid effects are captured via pose-guided deformation conditioned on global articulation context. This progressive disentanglement of motion, geometry, and appearance enables temporally coherent reconstructions across diverse species and behaviors.

In summary, our approach features the following key contributions:

- A novel test-time optimization framework is proposed, enabling 4D reconstruction of shapes and behaviors of a wide variety of animals from single monocular videos. This is achieved without access to well-annotated training dataset, or additional input requirement such as multi-view generative priors and category-specific shape templates.
- Our framework consists of two stages: a pose refinement stage followed by a pose-guided deformation stage (Figure 2). By introducing joint-aware anchors and symmetry-aware encoding, it progressively disentangles motion and appearance, enabling robust optimization even under inaccurate initialization.
- Extensive experiments demonstrate state-of-the-art performance across diverse animal species and behaviors. AnimalGS achieves 16.5% higher PSNR than prior work on APT-v2 (Table 1) and uniquely maintains quality on short sequences where existing methods fail.

2 Related Work

3D Animal Reconstruction. Reconstructing 3D animals is more challenging than reconstructing humans due to interspecific variation, complex articulations, and limited 3D data. Parametric models such as SMAL (Zuffi et al., 2017) provides the first skinned multi-animal template from toy figurines, following by various extensions and refinement (Zuffi et al., 2018; 2019; Rueegg et al., 2022; Rüegg et al., 2023). CSM-based methods (Kulkarni et al., 2019; 2020) predict dense image-to-surface

mappings but remain tied to predefined templates. Template-free methods (Yao et al., 2022; 2023; Liu et al., 2023a) discover parts and skeletons from sparse images through optimizing primitive part representation. Recent learning-based approaches scale to Internet data: UMR (Li et al., 2020), MagicPony (Wu et al., 2023), and Farm3D (Jakab et al., 2024) learn category-specific models, whereas FAUNA (Li et al., 2024) and SAOR (Aygun & Mac Aodha, 2024) aim for category-agnostic reconstruction. The evolution in 3D animal priors has grounded a natural basis for 4D reconstruction, yet they are often treated as fixed constraints, and are severely limited in generalizing to unseen animal species. Instead, shape prior is engaged in our approach as merely coarse initialization, which has been empirically demonstrated to notably contribute to flexible and faithful 4D recovery across species.

Dynamic Animal Reconstruction Extending static 3D models to capture temporal dynamics from monocular videos remains a central challenge in animal reconstruction. Deformation-based approaches (Yang et al., 2021a;b; Wu et al., 2022) represent objects by deforming an initial sphere mesh with fixed face connectivity, which struggle to recover fine surface details due to the limitations of template reliance. Yang et al. (2022; 2023a) adopt NeRF-based representations for greater topological flexibility but suffer from prohibitive computational costs and lack explicit surface geometry. Hybrid explicit approaches (Sabathier et al., 2024; Lei et al., 2024) have recently emerged, among which GART (Lei et al., 2024) is highlighted for leveraging BITE initialization (Rüegg et al., 2023) and combining it with 3D Gaussian Splatting (Kerbl et al., 2023), allowing for more flexible shape representation. In spite of substantial progress, existing methods remain constrained by relying on fixed mesh topologies or category-specific templates, and typically fail to provide a systematic framework for jointly refining pose and shape under severe prior mismatches. In contrast, our method employs a progressive, pose-guided optimization strategy that allows the coarse prior to evolve during reconstruction, enabling robust adaptation to diverse animal species.

Gaussian Splatting for Dynamic Scenes 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023) revolutionizes novel view synthesis by providing an explicit, flexible representation that achieves real-time performance with high fidelity. For dynamics, *time-augmented* 3DGS (e.g., 4DGS, Grid4D, Hybrid 3D–4D) encodes temporality in Gaussians but scales poorly with sequence length (Wu et al., 2024; Jiawei et al., 2024; Oh et al., 2025). *Deformation-based* variants keep a canonical 3DGS and learn warps (Deformable3DGS, SC-GS; spline extensions) for compact memory and smooth motion (Yang et al., 2024; Huang et al., 2024; Song et al., 2025). We adopt the deformation paradigm and drive warps with pose cues to improve temporal coherence for articulated animals.

Video-to-4D Generation A parallel line of work leverages generative priors for video-to-4D synthesis. Zero-1-to-3 (Liu et al., 2023b) pioneered this direction by leveraging diffusion models to hallucinate novel views from a single image. SV4D (Xie et al., b) and SV4D 2.0 (Yao et al., 2025) extend this concept to videos, enforcing multi-frame and multi-view consistency. However, these methods struggle with long video sequences, often losing geometric detail and requiring fixed-length inputs. Other methods that directly generate 4D models, such as Splat4D (Yin et al., 2025), L4GM (Ren et al., 2024a), and GVF-Diffusion (Jiang et al., 2025) usually demand large model ensembles and high memory usage. They will also produce results with significant inconsistencies in both appearance and shape compared to the input video. In contrast, we pursue reconstruction-only supervision from the input video, avoiding generative mismatch while retaining faithfulness.

3 Our Approach

Our goal is to recover time-varying 4D representations of animals from a monocular video sequence $\{I^t\}_{t=1}^T$ using a canonical-deformation formulation. A canonical 3D Gaussian Splatting model G_{can} undergoes hierarchical transformation: first through articulated pose refinement (G^t_{pose}), then pose-guided non-rigid deformation (G^t_{deform}). The following sections detail initialization (Sec. 3.1), pose refinement (Sec. 3.2), deformation modeling (Sec. 3.3), and self-supervised optimization (Sec. 3.4), with the complete pipeline illustrated in Figure 2 with a cow video as an example.

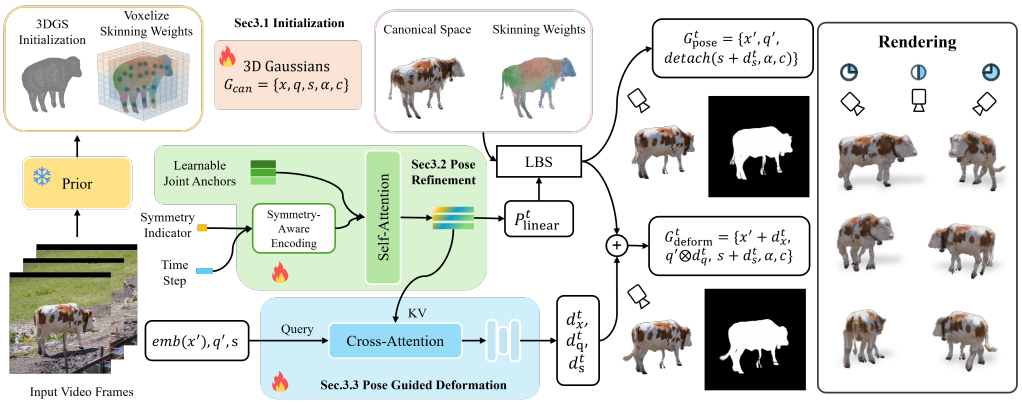

Figure 2: **AnimalGS** pipeline overview. From a monocular video input, an initial animal representation is constructed with the Fauna prior (Sec. 3.1), as 3DGS initialization, i.e. a canonical 3D Gaussian set G_{can}. This is then followed by two stages. The **Pose Refinement** stage integrates learnable joint anchors and time step encoded by a symmetry indicator to estimate articulated transformations, yielding intermediate representations G^t_{pose} (Sec. 3.2). The **Pose-Guided Deformation** stage then predicts non-rigid displacements to obtain the final time-specific representation G^t_{deform} (Sec. 3.3). **Right:** The resulting 4D representation can be rendered from arbitrary viewpoints and time steps.

3.1 Initialization from Prior

We initialize our canonical model using outputs from Fauna (Li et al., 2024), which provides a coarse category-agnostic estimate of animal shape and pose from a single image. Given an input frame I^t, Fauna predicts $(V^t, W^t, C^t, P^t) = F(I^t)$, where $V^t \in \mathbb{R}^{N_v \times 3}$ are mesh vertices in rest pose, $W^t \in \mathbb{R}^{N_v \times J}$ are skinning weights, $C^t \in \mathbb{R}^{4\times4}$ is the camera pose, and $P^t \in \mathbb{R}^{J\times4\times4}$ are joint transformations. Since vertices and skinning weights vary across frames, we use only the first-frame outputs (V^1, W^1) for initialization and omit superscripts for clarity. Note that Fauna also predicts per-frame deformed meshes, but we discard them and retain only the coarse rest-pose shape for canonical initialization, as subsequent deformations are explicitly modeled within our framework. Inspired by Jiang et al. (2022); Lei et al. (2024), we further embed W into a voxel grid to enable trilinear interpolation for dynamically created Gaussians during adaptive density control.

Following Kerbl et al. (2023), we represent canonical 3D Gaussian as $G_{\text{can}} = \{\mathbf{x}, \mathbf{q}, \mathbf{s}, \boldsymbol{\alpha}, \mathbf{c}\}$, where each Gaussian is parameterized by its center $\mathbf{x}$, orientation quaternion $\mathbf{q}$, scale $\mathbf{s}$, opacity α, and view-dependent spherical harmonic coefficients $\mathbf{c}$. We initialize the Gaussian centers $\mathbf{x}$ from the rest-pose mesh vertices V, while other attributes are randomly initialized.

3.2 Pose Refinement

Initial per-frame poses P^t predicted by the prior model are often unreliable due to limited views and articulation ambiguity in real-world videos. Our pose refinement module improves robustness by estimating per-joint transformations for linear blend skinning (LBS), mapping the canonical representation G_{can} to posed states G^t_{pose}.

Instead of relying directly on noisy joint detections, we introduce learnable joint anchors that provide a stable articulation-aware representation. To incorporate bilateral symmetry, we design a **symmetry-aware temporal encoding**:

$$\mathbf{e}^m_t = \text{emb}(t \cdot m) \oplus m, \quad m \in \{-1, 1\}, \tag{1}$$

where m is a symmetry indicator distinguishing original (m=1) from flipped (m=-1) views, and $\oplus$ denotes vector concatenation. This encoding ensures flipped frames are treated as mirror-symmetric counterparts rather than independent temporal observations, enabling consistent use of symmetry cues (see Sec. 3.4 for details).

Anchors combined with $\mathbf{e}^m_t$ are processed by a self-attention block to produce joint-specific temporal features $\mathbf{F_J}^t \in \mathbb{R}^{J\times K}$. These features are then projected to per-joint transformations $P^t_{\text{linear}} \in$

$\mathbb{R}^{J\times 7}$ (quaternion rotation and translation) and also forwarded to the subsequent deformation stage, providing a temporal and joint-aware context. As illustrated in Figure 2, applying LBS with P^t_{linear} updates Gaussian centers $\mathbf{x}$ and orientations $\mathbf{q}$, yielding the posed Gaussian $G^t_{\text{pose}} = \{\mathbf{x}', \mathbf{q}', \mathbf{s}, \boldsymbol{\alpha}, \mathbf{c}\}$.

3.3 Pose-Guided Deformation

While LBS-based pose refinement captures articulated motion, real animals exhibit complex non-rigid effects that cannot be modeled by skeletal transformations alone. Our pose-guided deformation module addresses this by predicting pose-conditioned, spatially varying offsets $(\mathbf{d}^t_x, \mathbf{d}^t_q, \mathbf{d}^t_s)$ to refine the Gaussian representation beyond articulation.

The key observation is that non-rigid deformations are tightly coupled with articulated pose, e.g. muscle bulging or skin motion varies with joint configuration. To capture this dependency, we adopt a cross-attention mechanism that conditions local Gaussian deformations on global joint-aware context. As shown in Figure 2, queries are constructed from Gaussian attributes—centers $\mathbf{x}'$ (with positional encoding), posed orientations $\mathbf{q}'$ and scales $\mathbf{s}$—while the joint-aware features $\mathbf{F}^t_J$ from the pose refinement module serve as keys and values. Using quaternion multiplication $\otimes$, the final deformed representation is $G^t_{\text{deform}} = \left\{\mathbf{x}' + \mathbf{d}^t_x,\ \mathbf{q}' \otimes \mathbf{d}^t_q,\ \mathbf{s} + \mathbf{d}^t_s,\ \boldsymbol{\alpha}, \mathbf{c}\right\}$. We deform only the geometric parameters while preserving appearance $(\boldsymbol{\alpha}, \mathbf{c})$, ensuring consistent texture and color across time.

3.4 Optimization

We optimize AnimalGS through *test-time optimization* on each input video. The overall objective is:

$$L_{\text{total}} = L_{\text{pose}} + L_{\text{deform}} + L_{\text{smooth}}, \tag{2}$$

where L_{pose} corresponds to the pose refinement stage, L_{deform} to the pose-guided deformation stage, and L_{smooth} is a regularization term. We use a differentiable Gaussian rasterizer (Kerbl et al., 2023) to render RGB images $\hat{I}$, silhouettes $\hat{S}$, and normal maps $\hat{N}$ from G with camera parameters C. Supervision combines photometric and silhouette objectives:

$$\mathcal{L}_{rgb}(\cdot) = (1-\lambda_{\text{ssim}})\,\mathcal{L}_1(\cdot) + \lambda_{\text{ssim}}\mathcal{L}_{ssim}(\cdot), \qquad \mathcal{L}_{sil}(\cdot) = \mathcal{L}_{bce}(\cdot) + \mathcal{L}_{dice}(\cdot). \tag{3}$$

with $\lambda_{\text{SSIM}} = 0.2$ in all experiments. All terms follow standard definitions.

Pose Refinement Loss The pose refinement stage uses silhouette-only supervision to isolate articulated motion from appearance, preventing texture artifacts from corrupting pose estimates while ensuring robust geometric alignment. Non-pose parameters are detached during rendering to block appearance-driven gradients, so that the posed Gaussian is represented as $\mathcal{G}^t_{\text{pose}} = \{\mathbf{x}', \mathbf{q}', \text{detach}(\mathbf{s} + \mathbf{d}^t_s, \boldsymbol{\alpha}, \mathbf{c})\}$, where adding the scale offset $\mathbf{d}^t_s$ can stabilizes downstream optimization without affecting pose gradients. The loss combines silhouette supervision with optional prior regularization:

$$L^t_{\text{pose}} = \lambda_{\text{pose}} \cdot \mathcal{L}_{\text{sil}}(\hat{S}^t_{\text{pose}}, S^t_{\text{SAM}}) + \lambda_{\text{prior}}(t) \cdot \|\hat{\mathbf{P}}^t_{\text{linear}} - \mathbf{P}^t\|_2, \tag{4}$$

where $\hat{S}^t_{\text{SAM}}$ are masks from Grounded-SAM (Ren et al., 2024b), $\lambda_{\text{pose}} = 0.2$, and $\lambda_{\text{prior}}(t) = \mathbf{1}_{[t\leq 4000]}$ provides early guidance.

Pose-Guided Deformation Loss Unlike pose refinement, which focuses only on articulated alignment, the deformation stage jointly optimizes all Gaussian parameters—including appearance—to capture fine-grained details beyond skeletal motion. Given the deformed representation G^t_{deform}, we supervise both silhouettes and RGB renderings:

$$L^t_{\text{deform}} = \mathcal{L}_{rgb}(\hat{I}^t_{\text{deform}}, I^t) + \mathcal{L}_{sil}(\hat{S}^t_{\text{deform}}, S^t_{\text{SAM}}), \tag{5}$$

where $\hat{S}^t_{\text{deform}}$ and $\hat{I}^t_{\text{deform}}$ denote the rendered silhouette and RGB image from G^t_{deform}, and I^t is the input frame. This stage ensures accurate geometry while recovering realistic textures and capturing non-rigid deformations.

Smoothness Regularization To improve geometric stability, we regularize surface normals under random views of the deformed 3D Gaussians G^t_{deform}. For each rotated view θ, we render a silhouette

Figure 3: Visual comparison with the SOTA methods of Fauna (Li et al., 2024), GART (Lei et al., 2024), D-3DGS (Yang et al., 2024), GVFDiffusion (Jiang et al., 2025), at two randomly chosen time steps, t and t′.

$\hat{S}^t_\theta$ and normal map $\hat{N}^t_\theta$, and penalize local angular variation using a total-variation style loss with absolute cosine similarity to avoid inward and outward facing ambiguities.

$$\mathcal{L}_{\text{smooth}} = \lambda_{\text{smooth}} \sum_{d\in\{x,y\}} \frac{\sum_{i,j} w^d_{i,j}\big(1 - |\hat{N}^t_\theta(i,j)\cdot\hat{N}^t_\theta(i+\delta_d, j+\delta'_d)|\big)}{\sum_{i,j} \hat{S}^t_\theta(i,j)\cdot\hat{S}^t_\theta(i+\delta_d, j+\delta'_d) + \epsilon}, \tag{6}$$

We use unit offsets $(\delta_x, \delta'_x) = (0,1)$ and $(\delta_y, \delta'_y) = (1,0)$, and $\lambda_{\text{smooth}}(t) = \mathbf{1}_{[7000<t<14000]}$ only activates at mid-training to avoid over-smoothing. This encourages view-invariant normal smoothness, reducing noise and flickering in novel-view renderings.

Table 1: **Input-view** quality on three datasets. Best in **bold**, second best underlined.

Method	DAVIS			Online			APTv2		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Fauna (Li et al., 2024)	18.281	0.761	0.280	16.669	0.774	0.279	19.561	0.760	0.267
D-3DGS (Yang et al., 2024)	23.569	**0.924**	0.109	22.726	**0.913**	0.116	19.922	**0.861**	0.167
GART (Lei et al., 2024)	19.486	0.810	0.201	21.006	0.841	0.181	18.564	0.807	0.168
GART (w/ tto) (Lei et al., 2024)	21.347	0.859	0.171	23.128	0.883	0.158	19.167	0.834	0.150
GVFDiffusion (Jiang et al., 2025)	16.419	0.836	0.174	16.835	0.857	0.167	14.820	0.778	0.256
Ours	**25.720**	0.911	**0.087**	**25.464**	0.912	**0.089**	**23.843**	0.860	**0.140**

Table 2: **Novel-view** quality on three datasets. Best in **bold**, second best underlined. A dash indicates metric not reported by the method.

Method	DAVIS			Online			APTv2		
	KID-16V↓	FVD-F↓	FVD-Diag↓	KID-16V↓	FVD-F↓	FVD-Diag↓	KID-16V↓	FVD-F↓	FVD-Diag↓
Fauna (Li et al., 2024)	0.279	—	—	0.334	—	—	0.247	—	—
D-3DGS (Yang et al., 2024)	0.211	1176.787	1042.756	0.213	1162.396	1265.113	0.326	1245.334	1182.464
GART (Lei et al., 2024)	0.216	1750.899	1675.388	0.233	1470.950	1547.561	0.230	1355.4969	992.195
GART (w/ tto) (Lei et al., 2024)	0.208	1680.705	1579.871	0.238	1473.927	1364.587	0.228	1134.661	948.274
GVFDiffusion (Jiang et al., 2025)	0.145	1872.192	1270.189	0.179	1673.419	1387.732	0.274	1715.471	1575.461
Ours	**0.144**	**897.713**	**810.408**	**0.156**	**1003.253**	**1162.509**	**0.166**	**992.239**	**887.060**

Bilateral Symmetry Augmentation Naively treating flipped frames as independent samples discards their inherent geometric relationship. Meanwhile, imperfect camera calibration prevents strict symmetry enforcement, leading to ambiguity and conflicting supervision. To leverage symmetry without introducing noise, we build on the symmetry-aware encoding in Eq. 1 and construct four augmented samples for each frame I^t under two geometric interpretations:

$$\mathcal{V}_{\text{orig}} = \{(I^t, C^t, P^t), (I^t_{\text{flip}}, C^t_{\text{sym}}, P^t_{\text{flip}})\} \quad (m = 1), \tag{7}$$

$$\mathcal{V}_{\text{flip}} = \{(I^t_{\text{flip}}, C^t_{\text{flip}}, P^t_{\text{flip}}), (I^t, C^t_{\text{flip,sym}}, P^t)\} \quad (m = -1), \tag{8}$$

with symmetric cameras $C_{\text{sym}} = MC$ computed via sagittal plane reflection.

This augmentation enriches supervision by exposing the model to both original and mirrored interpretations, while the symmetry indicator ensures consistent temporal encoding and suppresses calibration inconsistencies. As a result, bilateral symmetry is enforced effectively, leading to more stable geometry and motion reconstruction.

Stabilization Strategies. We employ a pose blending with annealing scheme to enable stable refinement. The predicted pose $\hat{\mathbf{P}}^t_{\text{linear}}$ is blended with the prior pose to produce the final transformation:

$$\mathbf{P}^t_{\text{linear}} = w(t) \cdot \mathbf{P}^t + (1 - w(t)) \cdot \hat{\mathbf{P}}^t_{\text{linear}}, \tag{9}$$

where $w(t)$ anneals from 1 to 0 over 7K iterations. This allows gradual refinement from the initialization while enabling joint optimization of all modules from the start, preventing overfitting to inaccurate priors. In addition, we adopt the annealing smooth training mechanism (Yang et al., 2024), which injects decaying Gaussian noise into the temporal coordinate t during early iterations. This improves temporal smoothness under pose inaccuracies without incurring extra overhead.

4 Experiment

4.1 Dataset

We collect 87 videos from three sources: online collection (11 videos), DAVIS (Perazzi et al., 2016) (8 videos), and APTv2 (Yang et al., 2023b) (68 videos). For APTv2, all videos contain 15 frames, except for two, which were manually composed by concatenating similar clips. We design a semi-automatic preprocessing pipeline: first, we extract animal masks using Grounded-SAM (Ren et al., 2024b) with category text prompts and compute smoothed bounding boxes, then estimate animal and camera parameters for both original and horizontally flipped sequences using Fauna (Li et al., 2024). Finally, we select temporally stable frames via DBSCAN clustering (Ester et al., 1996) on camera trajectories and sample every fifth frame for testing, resulting in a 4:1 train/test split.

Table 3: Ablation study on our online collections. Best in **bold**, second best underlined.

Variant	Input View			Novel View		
	PSNR↑	SSIM↑	LPIPS↓	KID-V↓	FVD-F↓	FVD-Diag↓
random Init	25.052	0.905	0.105	0.216	1061.154	1219.191
w/o Deform	24.394	0.896	0.098	0.176	1051.321	1305.3614
w/o Joint Anchors	25.462	**0.912**	**0.088**	0.163	965.899	**1130.662**
w/o Symmetry Encoding	25.244	0.909	0.096	**0.156**	**959.335**	1154.033
w/o L_{spin}	25.426	0.911	0.089	0.170	969.896	1217.878
full	**25.464**	**0.912**	0.089	0.162	1003.253	1162.509

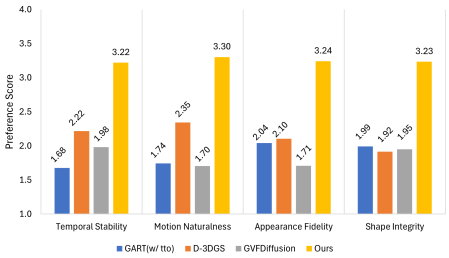

Figure 4: User study preference scores across four perceptual dimensions.

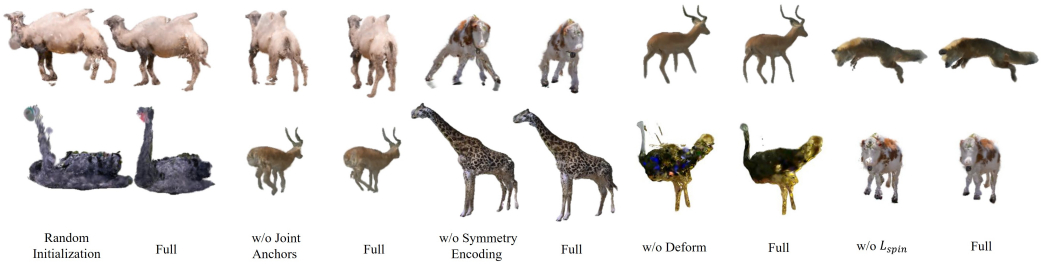

Figure 5: Novel-view synthesis under ablations, showing degraded reconstructions without certain components, while the full model remains stable and accurate.

4.2 Implementation Details

We implement our method in PyTorch. Optimization is run for 20K iterations on DAVIS and online videos, and 10K on APTv2. Each iteration trains on $\mathcal{V}_{\text{orig}}$ or $\mathcal{V}_{\text{flip}}$ group. The symmetric augmentation only guide the final pose-guided deformation stage. All Gaussian parameters follow the learning rate schedule of 3DGS, and other modules are optimized with a single Adam (Kingma & Ba, 2015) with an exponentially decaying learning rate from 8×10^{-4} to 1.6×10^{-4}. Rendering speed scales with the number of optimized Gaussians; on the online collection, we achieve an average of 109.9 FPS, and on DAVIS, 110.7 FPS, measured on a single NVIDIA A6000 GPU at resolution 512×512.

4.3 Results and Comparisons

Baselines We compare to 4 state-of-the-art methods representing different reconstruction paradigms: (1) **Fauna** (Li et al., 2024): Learning-based single-image 3D reconstruction, also serving as our initialization prior; (2) **GART** (Lei et al., 2024): Test-time optimization using SMPL/SMAL priors for articulated 3D reconstruction from monocular video. We report both GART (optimized on training frames only) and GART (w/ tto), where the model is further refined on test frames, following their original evaluation protocol; (3) **D-3DGS** (Yang et al., 2024): Dynamic 3DGS with learnable deformation fields; (4) **GVFDiffusion** (Jiang et al., 2025): 4D reconstruction using pretrained 3D diffusion models (Trellis (Xiang et al., 2025)). For fair comparison, we enhance GART and D-3DGS with flip augmentation, treating flipped frames as new timestamps to avoid shape aliasing.

Input-view reconstruction Table 1 reports the input-view reconstruction results on DAVIS, Online, and APTv2, evaluated on test-set frames. Our method achieves the best PSNR and LPIPS across all three datasets, significantly outperforming prior works. Although D-3DGS obtains slightly higher SSIM, our method remains highly competitive while consistently providing superior perceptual quality. These results indicate that our approach faithfully reconstructs the observed views with both high fidelity and realism.

Novel-view reconstruction Table 2 reports novel-view quality. For Fauna, we only report KID-16V since it is single-image based. We follow Xie et al. (a) and adopt FVD-F (temporal coherence at a fixed view) and FVD-Diag (spatio-temporal consistency). Since FVD-V requires equal frame/view

counts and biases results on long videos, we instead introduce KID-V (See Figure 7), which uniformly samples novel views and computes Kernel Inception Distance (Bińkowski et al., 2018). KID-V is unbiased and stable even with few images, making it well-suited to our limited-view-per-time setting. KID-V is measured on novel views from the test set, while FVD-V and FVD-Diag computed over the entire generated video. As shown in Table 2, our method achieves state-of-the-art performance across most metrics, with consistently strong results on all datasets. Our method remains top-ranked overall.

Qualitative Comparison Figure 3 shows reconstructions across four species. Fauna (Li et al., 2024) produces only coarse shapes with approximate color. GART (Lei et al., 2024) and D-3DGS (Yang et al., 2024) capture body motion but fail on limbs—notably missing the stork's legs entirely. GVFDiffusion (Jiang et al., 2025) struggles with short sequences, producing distorted artifacts on APTv2 clips and hallucinating extra legs on the stork despite reasonable cow reconstruction. In contrast, our method achieves faithful, temporally consistent results across all species, preserving fine details in both input and novel views.

Ablation Study Table 3 quantifies each component's contribution. Since no ground-truth novel views exist, current metrics are only approximate: FVD favors temporal smoothness, while KID and FVD cannot capture shape integrity or motion naturalness. This explains the gap between numerical scores and the visual quality in Figure 5, where our full model yields superior geometry despite comparable metrics. Qualitative results further reveal characteristic failure modes: missing joint anchors or random initialization lead to severe shape errors, removing symmetry encoding or deformation degrades pose and motion, and omitting L_{smooth} introduces surface artifacts. These observations validate the necessity of our complete framework, while highlighting the need for perceptually aligned metrics in future evaluation.

User study Since existing quantitative metrics may not fully capture the perceptual quality of novel-view synthesis, we conducted a user study with 56 participants evaluating 15 videos sampled from our three datasets. Participants were asked to rank 4 methods (Ours, GART, D-3DGS, GVFDiffusion) along 4 perceptual dimensions: temporal stability, motion naturalness, appearance fidelity, and shape integrity. Rankings were converted to scores (4=best, 1=worst) and averaged. As shown in Figure 4, our method consistently achieves the highest preference scores across all dimensions, indicating superior perceptual quality compared to all baselines. See A.3 for detailed protocols.

4.4 Discussion and Limitation.

Our results highlight that treating animal priors as coarse initialization, rather than strict constraints, enables robust 4D reconstruction across species. Unlike recent diffusion-based approaches that prioritize plausibility over fidelity and degrade with limited frames, our optimization-based framework emphasizes realistic reconstruction that remains closely aligned with the input video even with severe prior mismatches. This suggests a broader lesson for articulated non-human reconstruction: balancing prior knowledge with optimization flexibility is key to realism and generalization. Nevertheless, our approach still struggles under limited viewpoints, inaccurate camera priors, or strong occlusions, and subtle head motions remain challenging as silhouettes provide insufficient 3D cues. Future work could incorporate stronger geometric supervision (e.g., depth, keypoints) or learned deformation priors to address these limitations.

5 Conclusion

We presented AnimalGS, a test-time optimization framework for 4D animal reconstruction from monocular video. Our key insight that robust reconstruction arises from pose-guided optimization rather than accurate shape priors enables generalization without multi-view supervision or category-specific templates. By introducing joint-aware anchors and symmetry-aware encoding, our method disentangles motion from appearance and remains robust to prior mismatches. Extensive experiments across diverse species demonstrate clear improvements over state-of-the-art baselines in both reconstruction quality and temporal consistency, validated through quantitative metrics and user studies. While developed for animals, these principles may extend to other non-rigid objects, suggesting hybrid approaches that couple optimization precision with stronger geometric cues and multi-view synthesis.

REFERENCES

Mehmet Aygun and Oisin Mac Aodha. Saor: Single-view articulated object reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10382–10391, 2024.

Marc Badger, Yufu Wang, Adarsh Modh, Ammon Perkes, Nikos Kolotouros, Bernd G Pfrommer, Marc F Schmidt, and Kostas Daniilidis. 3d bird reconstruction: a dataset, model, and shape recovery from a single view. In *European conference on computer vision*, pp. 1–17. Springer, 2020.

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *International Conference on Learning Representations*, 2018.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pp. 226–231, 1996.

Yi-Hua Huang, Yang-Tian Sun, Ziyi Yang, Xiaoyang Lyu, Yan-Pei Cao, and Xiaojuan Qi. Sc-gs: Sparse-controlled gaussian splatting for editable dynamic scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4220–4230, 2024.

Tomas Jakab, Ruining Li, Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Farm3d: Learning articulated 3d animals by distilling 2d diffusion. In *2024 International Conference on 3D Vision (3DV)*, pp. 852–861. IEEE, 2024.

Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. Gaussian variation field diffusion for high-fidelity video-to-4d synthesis. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Xu Jiawei, Fan Zexin, Yang Jian, and Xie Jin. Grid4D: 4D decomposed hash encoding for high-fidelity dynamic gaussian splatting. *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL `http://arxiv.org/abs/1412.6980`.

Nilesh Kulkarni, Abhinav Gupta, and Shubham Tulsiani. Canonical surface mapping via geometric cycle consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2202–2211, 2019.

Nilesh Kulkarni, Abhinav Gupta, David F Fouhey, and Shubham Tulsiani. Articulation-aware canonical surface mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 452–461, 2020.

Jiahui Lei, Yufu Wang, Georgios Pavlakos, Lingjie Liu, and Kostas Daniilidis. Gart: Gaussian articulated template models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19876–19887, 2024.

Xueting Li, Sifei Liu, Kihwan Kim, Shalini De Mello, Varun Jampani, Ming-Hsuan Yang, and Jan Kautz. Self-supervised single-view 3d reconstruction via semantic consistency. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pp. 677–693. Springer, 2020.

Zizhang Li, Dor Litvak, Ruining Li, Yunzhi Zhang, Tomas Jakab, Christian Rupprecht, Shangzhe Wu, Andrea Vedaldi, and Jiajun Wu. Learning the 3d fauna of the web. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9752–9762, 2024.

Di Liu, Anastasis Stathopoulos, Qilong Zhangli, Yunhe Gao, and Dimitris Metaxas. Lepard: Learning explicit part discovery for 3d articulated shape reconstruction. *Advances in Neural Information Processing Systems*, 36:54187–54198, 2023a.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9298–9309, 2023b.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 851–866. 2023.

Seungjun Oh, Younggeun Lee, Hyejin Jeon, and Eunbyung Park. Hybrid 3d-4d gaussian splatting for fast dynamic scene representation, 2025. URL https://arxiv.org/abs/2505.13215.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Jiawei Ren, Kevin Xie, Ashkan Mirzaei, Hanxue Liang, Xiaohui Zeng, Karsten Kreis, Ziwei Liu, Antonio Torralba, Sanja Fidler, Seung Wook Kim, and Huan Ling. L4gm: Large 4d gaussian reconstruction model. In *Proceedings of Neural Information Processing Systems(NeurIPS)*, Dec 2024a.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024b.

Nadine Rueegg, Silvia Zuffi, Konrad Schindler, and Michael J Black. Barc: Learning to regress 3d dog shape from images by exploiting breed information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3876–3884, 2022.

Nadine Rüegg, Shashank Tripathi, Konrad Schindler, Michael J Black, and Silvia Zuffi. Bite: Beyond priors for improved three-d dog pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8867–8876, 2023.

Remy Sabathier, Niloy J Mitra, and David Novotny. Animal avatars: Reconstructing animatable 3d animals from casual videos. In *European Conference on Computer Vision*, pp. 270–287. Springer, 2024.

Mingyang Song, Yang Zhang, Marko Mihajlovic, Siyu Tang, Markus Gross, and Tunç Ozan Aydın. Spline deformation field. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–10, 2025.

Yufu Wang, Nikos Kolotouros, Kostas Daniilidis, and Marc Badger. Birds of a feather: Capturing avian shape models from images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14739–14749, 2021.

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.

Shangzhe Wu, Ruining Li, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. Magicpony: Learning articulated 3d animals in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8792–8802, 2023.

Yuefan Wu, Zeyuan Chen, Shaowei Liu, Zhongzheng Ren, and Shenlong Wang. Casa: Category-agnostic skeletal animal reconstruction. *Advances in Neural Information Processing Systems*, 35: 28559–28574, 2022.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21469–21480, 2025.

Yiming Xie, Chun-Han Yao, Vikram Voleti, Huaizu Jiang, and Varun Jampani. Sv4d: Dynamic 3d content generation with multi-frame and multi-view consistency. In *The Thirteenth International Conference on Learning Representations*, a.

Yiming Xie, Chun-Han Yao, Vikram Voleti, Huaizu Jiang, and Varun Jampani. Sv4d: Dynamic 3d content generation with multi-frame and multi-view consistency. In *The Thirteenth International Conference on Learning Representations*, b.

Gengshan Yang, Deqing Sun, Varun Jampani, Daniel Vlasic, Forrester Cole, Huiwen Chang, Deva Ramanan, William T Freeman, and Ce Liu. Lasr: Learning articulated shape reconstruction from a monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15980–15989, 2021a.

Gengshan Yang, Deqing Sun, Varun Jampani, Daniel Vlasic, Forrester Cole, Ce Liu, and Deva Ramanan. Viser: Video-specific surface embeddings for articulated 3d shape reconstruction. *Advances in Neural Information Processing Systems*, 34:19326–19338, 2021b.

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.

Gengshan Yang, Chaoyang Wang, N Dinesh Reddy, and Deva Ramanan. Reconstructing animatable categories from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16995–17005, 2023a.

Yuxiang Yang, Yingqi Deng, Yufei Xu, and Jing Zhang. Aptv2: Benchmarking animal pose estimation and tracking with a large-scale dataset and beyond, 2023b.

Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20331–20341, 2024.

Chun-Han Yao, Wei-Chih Hung, Yuanzhen Li, Michael Rubinstein, Ming-Hsuan Yang, and Varun Jampani. Lassie: Learning articulated shapes from sparse image ensemble via 3d part discovery. *Advances in Neural Information Processing Systems*, 35:15296–15308, 2022.

Chun-Han Yao, Wei-Chih Hung, Yuanzhen Li, Michael Rubinstein, Ming-Hsuan Yang, and Varun Jampani. Hi-lassie: High-fidelity articulated shape and skeleton discovery from sparse image ensemble. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4853–4862, 2023.

Chun-Han Yao, Yiming Xie, Vikram Voleti, Huaizu Jiang, and Varun Jampani. Sv4d 2.0: Enhancing spatio-temporal consistency in multi-view video diffusion for high-quality 4d generation. *arXiv preprint arXiv:2503.16396*, 2025.

Minghao Yin, Yukang Cao, Songyou Peng, and Kai Han. Splat4d: Diffusion-enhanced 4d gaussian splatting for temporally and spatially consistent content creation. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–10, 2025.

Silvia Zuffi, Angjoo Kanazawa, David W Jacobs, and Michael J Black. 3d menagerie: Modeling the 3d shape and pose of animals. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6365–6373, 2017.

Silvia Zuffi, Angjoo Kanazawa, and Michael J Black. Lions and tigers and bears: Capturing non-rigid, 3d, articulated shape from images. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 3955–3963, 2018.

Silvia Zuffi, Angjoo Kanazawa, Tanya Berger-Wolf, and Michael J. Black. Three-d safari: Learning to estimate zebra pose, shape, and texture from images "in the wild". In *International Conference on Computer Vision*, October 2019.

A APPENDIX

A.1 THE USE OF LARGE LANGUAGE MODELS

In the process of preparing this paper submission, Large Language Models (LLMs) were used only as a writing-assist tool. Specifically, they were used to polish the text for better clarity and fluency, as well as to correct minor grammatical errors. The LLMs didn't contribute to the research ideation, methodology, or substantive writing of the paper. The authors take full responsibility for all the submitted contents.

A.2 EFFECT OF INITIALIZATION STRATEGIES

Figure 6 compares different initialization strategies for our 3D Gaussian representation. With prior-based initialization, the canonical space already provides a coarse but structured shape, leading to stable canonical 3DGS and skinning weights. In contrast, random initialization starts from an unstructured point cloud that produces unstable intermediate results in the canonical stage. Nevertheless, optimization can still converge to a configuration compatible with the skinning weights, although the process is less stable and less reliable than with prior guidance.

A.3 EVALUATION PROTOCOL OF USER STUDY

We conducted a user study with 56 participants, evaluating reconstruction results of animal videos sampled from our three datasets. Among them, 8 participants had prior experience in 3D modeling, while the remaining participants had no such experience. Each participant was asked to rank four reconstruction methods (Ours, GART, D-3DGS, GVFDiffusion) across four perceptual dimensions.

A total of 56 participants evaluated the reconstructed animal videos using a comprehensive rating system. Four methods were ranked from best to worst across four evaluation dimensions: Q1: 3D Temporal Stability (consistency of reconstructed shape and texture over time), Q2: Animal Motion Naturalness (realism of reconstructed movements), Q3: Appearance Fidelity (visual realism including identity consistency and appearance quality), and Q4: Shape Integrity (correctness and completeness of 3D geometry including structural soundness, completeness, and proportional accuracy). Each participant assessed four different methods based on one 4D GIF visualization for questions Q1 and Q2, evaluating the complete reconstructed animal, and 3D visualizations rendered at three distinct temporal moments for questions Q3 and Q4. All evaluation criteria (Q1-Q4) were presented with detailed explanations and examples to ensure consistent and informed participant assessments.

45 participants completed all evaluation questions; the remaining participants provided partial responses. To ensure fair analysis, weighted scores were calculated for each question based on the actual number of votes received. The evaluation process converted participant rankings into numerical scores (ranging from 4 for best performance to 1 for worst), which were then averaged to determine final ratings.

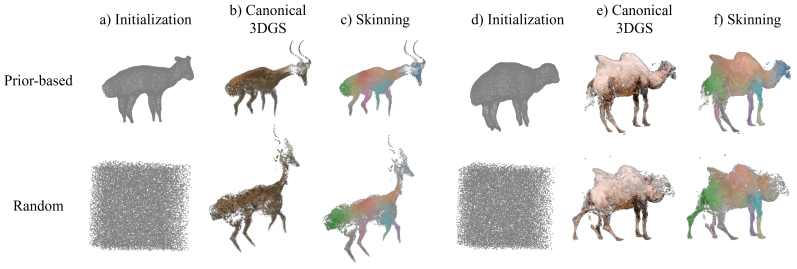

Figure 6: Effect of different initialization strategies.

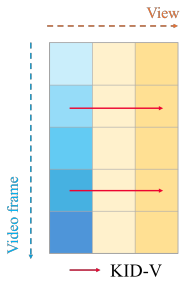

Figure 7: Illustration of KID-V metric.