WildAni4D: Towards 4D Animal Mesh Reconstruction

Supplementary Material

Overview

In this supplementary material, we provide additional details and extensive results to support the main paper. The document is organized as follows:

- Sec. A compares the proposed WildAni4D dataset with existing animal datasets and describes how the test set is constructed to ensure rigorous evaluation.
- Sec. B describes the test-time optimization procedure for animatable animal reconstruction.
- Sec. C justifies our rendering-based data generation pipeline and presents additional ablation studies.
- Sec. D presents additional results, including dataset samples, world-coordinate animal motion reconstruction, and animatable animal reconstruction results.

Please refer to the supplementary video for animated results that cannot be shown in this document.

A. Additional Dataset Details

A.1. Comparison with Existing Animal Datasets

The proposed WildAni4D dataset provides large-scale animal video sequences with complete annotations for 3D pose, shape, and global trajectory. In contrast, previous datasets are restricted to images or lack full 3D labels, as summarized in Table 1. Such 3D annotations are common in the human domain but remain largely unavailable for animals. This is because animals are difficult to capture in controlled setups and rarely come with accurate 3D labels. The lack of 3D supervision is therefore a central bottleneck for animal mesh recovery. WildAni4D addresses this bottleneck by providing long animal video sequences with temporally consistent 3D pose, shape, and global trajectory annotations.

A.2. WildAni4D Test Set Configuration

To evaluate the generalization capability of our model on unseen data distributions, we rigorously design the WildAni4D test set to be disjoint from the training set in terms of both animal species and environmental contexts. While the main paper outlines the synthetic generation pipeline involving dynamic textured animals and diverse 3D scenes, this section describes the specific split strategy used to minimize distributional overlap.

Unlike standard random splitting strategies that may include the same animal species in both training and testing, we intentionally hold out specific animal families from the training set. Specifically, we exclude the fox family and other morphologically distinct quadrupeds from the training data. These species are reserved exclusively for the test set. To further simulate the domain gap inherent in in-the-wild videos, we ensure that the background scenes used in the test set are never seen during training.

B. Details on Test-Time Optimization (TTO)

In the main paper, for the animatable animal reconstruction experiments, we also report results for a variant called *Ours-tto*. This variant uses test-time optimization to refine only the per-frame pose and global translation, while keeping the shape predicted by our model fixed. While our proposed AVT predicts a temporally consistent shape and smooth global trajectories, direct regression may sometimes lack pixel-perfect alignment with 2D evidence due to the smoothing effect of the temporal transformer. To demonstrate that our predictions provide a robust initialization for optimization-based refinement in the spirit of SMPLify [2], we run a simple test-time optimization that updates pose and translation *only* using a 2D keypoint reprojection loss, without any additional silhouette or photometric terms. This comparison is important because it shows that our model's disentangled shape and pose predictions provide a geometrically consistent prior that can be further refined to better match the image evidence, validating the robustness of our predictions.

C. Additional Experiments

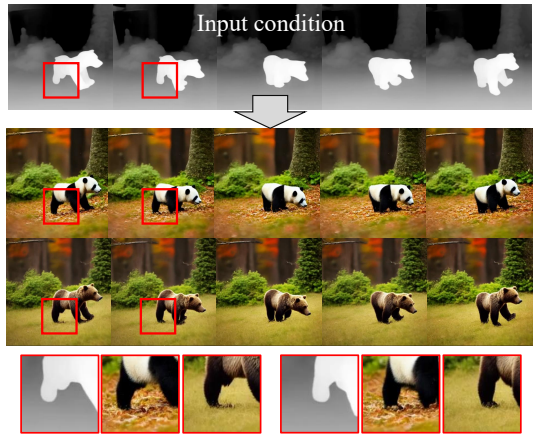
C.1. Why Rendering for 3D Video Data Generation?

When we synthesize a 3D animal video dataset without capturing new videos, two natural options arise: *generative video editing* and *3D rendering*. This motivates us to implement state-of-the-art methods for each option and compare them directly. In our implementation, we find that existing video generation models easily break pose and shape consistency, whereas rendering keeps them stable. This observation justifies our decision to adopt a rendering-based pipeline.

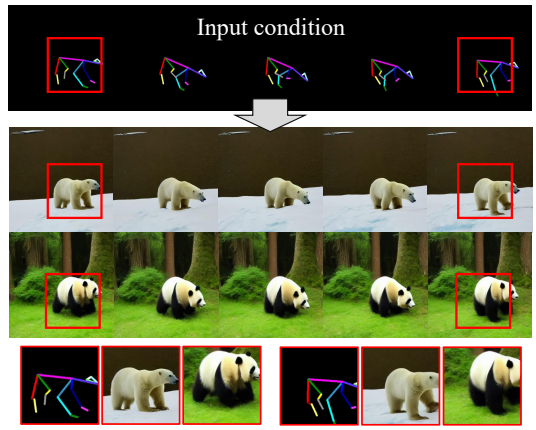
Generative video editing. VideoGrain [9] is a zero-shot video editing framework that takes an input video and applies fine-grained, text-driven modifications to specific objects, regions, or parts. VideoGrain outputs an edited video in which spatial and temporal details are consistently updated while preserving the original motion and structure. We synthesize animal videos using diffusion-based video editing methods such as VideoGrain. These videos serve

Table 1. **Comparison with existing animal datasets.** We compare our dataset with prior animal datasets in terms of available supervision and data scale. Our dataset is the only one that provides segmented and 3D-annotated *video* sequences (30K videos across 16 scenes), whereas existing datasets consist of image collections without video sequences.

	Ours	GenZoo [6]	AniMer [5]	Animal3D [8]	Stanford Ext. [1]	AP-10K [10]
Segmentation	✓	✓	✓	✓	✓	×
3D Annotation	✓	✓	✓	✓	×	×
#Images	> 1M	1M	9.7K	3.4K	8.1K	10K
Data type	Video	Image	Image	Image	Image	Image
#Videos	30K	×	×	×	×	×
#Scenes	16	×	×	×	×	×



(a) Depth-conditioned video generation



(b) Pose-conditioned video generation

Figure 1. **Results of conditioned generative video editing [9].** We explore generating diverse datasets by conditioning a generative model on mock-up videos, enabling flexible control over motion, appearance, and scene variation. (a) Depth-conditioned video generation. When using depth maps as guidance, the model occasionally produces local artifacts or incomplete structures, particularly around object boundaries. (b) Pose-conditioned video generation. While pose cues provide strong structural constraints, they can still lead to shape inconsistencies across frames, especially when the target motion is complex.

as training data for downstream tasks such as animal mesh recovery, so we generate them by conditioning the editing process on depth or 2D skeletons. Although the generated results roughly follow these conditions, they still exhibit two critical issues. The first issue is semantic inconsistency. When parts of the animal are occluded or blurred in the conditioning input, the generated structure can easily collapse. If only part of the paw is visible, the generated video often removes the paw region (see Fig. 1(a)). The second issue is shape drift. Even when we apply an animal pose condition to keep the pose fixed, the body size, thickness, and proportions still change noticeably from frame to frame (see Fig. 1(b)). This behavior breaks the basic assumption that a single sequence should represent one consistent individual.

3D rendering-based pipeline. To avoid these issues, we adopt the 3D rendering pipeline described in the main paper. We generate 3D animal meshes using SMAL [11], apply textures with SyncMVD [4], and then render the models in diverse scenes. This pipeline provides two key properties. The first is shape consistency. We keep the mesh topology and the shape parameters β fixed for each sequence, so the animal's shape does not change from frame to frame. The second is accurate 3D ground truth. The pipeline directly outputs numeric 3D quantities such as pose, shape, and global motion, which current generative video editing methods cannot reliably provide.

Taken together, these observations indicate that, at present, a 3D rendering pipeline is the most reliable way to generate training data for 4D animal reconstruction.

C.2. Ablation Studies

As described in the main paper, the proposed Animal Video Transformer (AVT) reconstructs world-grounded 4D animal motion from monocular video, including 3D pose, shape, and root trajectory in the recovered world frame. To achieve this, AVT utilizes features extracted from the pretrained GenZoo [6] backbone, which serves as an animal-specialized encoder. Furthermore, the model predicts a single sequence-level shape alongside per-frame pose and root motion to maintain temporal consistency. We analyze the effect of the model design using four variants of AVT. We

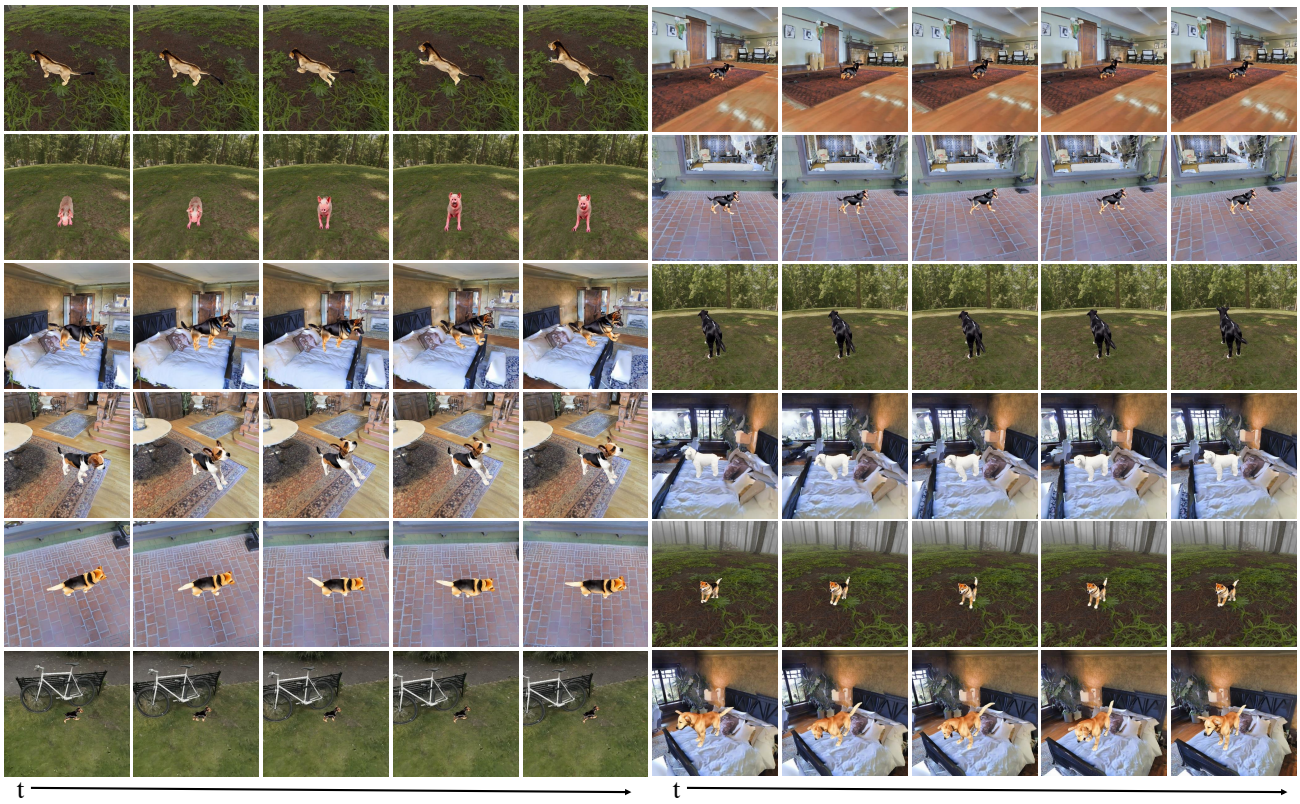


Figure 2. **Additional qualitative results of the proposed dataset.** By combining diverse factors such as animal motion, textures conditioned on text prompts, shape variation, scene context, and camera motion, we can generate a wide range of realistic video samples.

define each variant as follows:

- **AVT-c**: Simplest variant; an HMR2.0-like per-frame regressor with a pretrained AniMer [5] backbone.
- **AVT-d**: Temporally aware variant; AVT-c + temporal transformer with a pretrained AniMer backbone.
- **AVT-e**: Shape-consistent variant; AVT-d + sequence-level shape regressor with a pretrained AniMer backbone.
- **AVT** (*Ours*): Full AVT model; AVT-e with a pretrained GenZoo backbone.

Quantitative results for all variants are shown in Table 2. In the following, we analyze the ablation results with respect to the model architecture and the animal-specific backbone.

Model Architecture. As shown in Table 2, the temporal transformer plays a key role in smoothing motion over time. When we compare the per-frame AVT-c with the temporally aware AVT-d, the acceleration error decreases from 17.74 to 12.25, indicating that the predicted trajectories become less jittery and more stable the sequence. The sequence-level shape regressor then further improves stability by enforcing a single shape over the whole sequence. Comparing AVT-d with AVT-e, we observe the largest gain in the stability metrics: the acceleration error drops to 6.92 and the shape consistency error drops to 0.0681. This shows that explicit sequence-level shape modeling is crucial for producing coherent 4D motion rather than merely smoothing frame-wise predictions.

Table 2. **Ablation study**. We analyze the contribution of each architectural component. The sequence-level shape regressor substantially enhances temporal stability and pose accuracy.

	MPJPE↓	S-MPJPE↓	PA-MPJPE↓	Accel↓	Shape Consistency↓
AVT-c	123.21	81.89	54.40	17.74	0.6680
AVT-d	111.05	75.91	52.58	12.25	0.3340
AVT-e	108.58	74.54	52.58	**6.92**	**0.0681**
Ours	**91.29**	**62.44**	**40.13**	7.35	0.0703

These improvements in temporal smoothness and shape consistency also lead to better pose accuracy. Across AVT-c, AVT-d, and AVT-e, both MPJPE and PA-MPJPE steadily decrease. This trend shows that the temporal transformer and the sequence-level shape regressor not only stabilize motion over time but also regularize frame-wise predictions. These results suggest that they reduce ambiguity and shape drift across frames, which in turn leads to more accurate 3D joint estimates.

Feature Backbones. As shown in Table 2, the choice of feature backbone affects pose estimation accuracy. This becomes clear when we compare AVT-e with an AniMer backbone to our full AVT model with a GenZoo backbone under

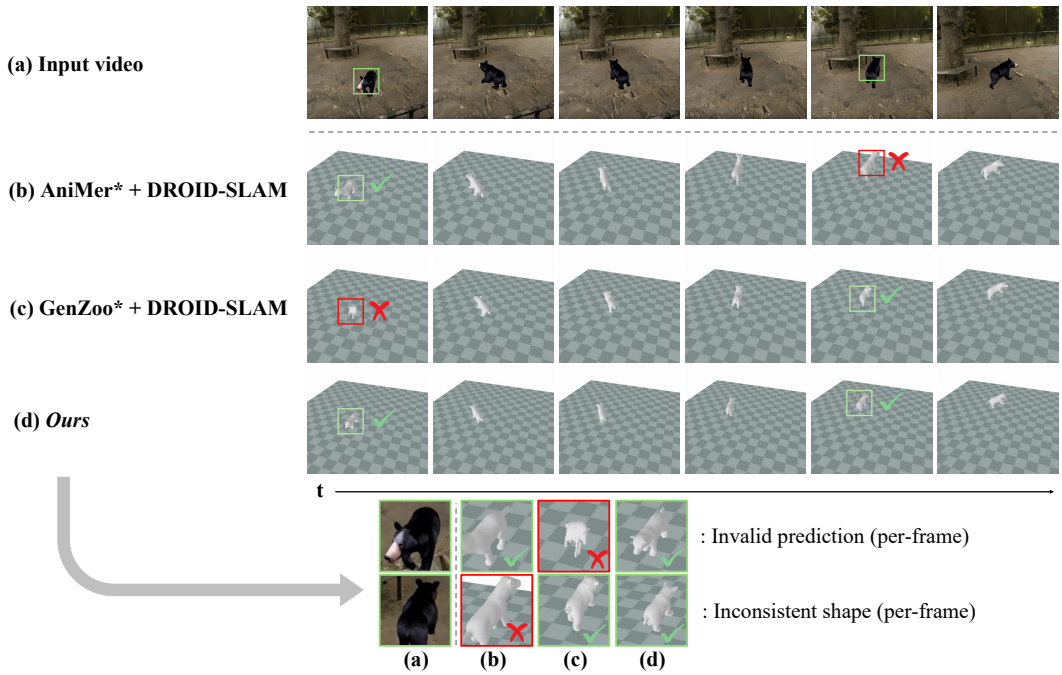


Figure 3. **Qualitative comparison of world-coordinate reconstruction.** We visualize the global trajectory and 3D motion reconstructed from a monocular video. We combine AniMer* and GenZoo* with DROID-SLAM [7] to lift per-frame predictions into world coordinates. (b) AniMer* exhibits inconsistent shape estimation across frames, failing to maintain the animal's identity. (c) GenZoo* suffers from invalid pose predictions, leading to physically implausible distortions. (d) In contrast, Ours maintains shape consistency and recovers a smooth, plausible trajectory throughout the sequence.

the same architecture: switching from AniMer to GenZoo reduces MPJPE from 108.58 to 91.29 and PA-MPJPE from 52.58 to 40.13, while keeping Accel and Shape Consistency at similar levels. These results suggest that the GenZoo encoder provides stronger animal-specific features.

D. Additional Results

D.1. Dataset Samples

We provide additional dataset samples in Fig. 2. Please refer to the supplementary video for animated visualizations of our results.

D.2. World-Coordinate Animal Motion Reconstruction

We visualize global trajectory reconstruction results in Fig. 3. We compare our method against AniMer* and GenZoo*. We integrate these per-frame baselines with DROID-SLAM [7] to estimate global camera motion. AniMer* and GenZoo* predict the pose independently at each frame. This per-frame design produces drift and jitter when we accumulate their predictions over time. Although we estimate camera motion with DROID-SLAM, AniMer* and GenZoo* still treat pose prediction as a purely per-frame problem, without jointly reasoning about camera and animal motion. As a result, their world trajectories exhibit drift and jitter over time.

Our AVT disentangles camera motion from animal motion. This separation allows the model to predict smooth poses that remain aligned across frames. As a result, the global trajectory stays consistent and matches the actual path in the video. This result shows that our world-coordinate reconstruction captures both local articulated motion and long-range movement in a coherent manner. Please refer to the supplementary video for animated visualizations of the world-coordinate animal motion reconstruction results.

D.3. More Results on Animatable Animal Reconstruction

In the main paper, to evaluate the quality and temporal consistency of the motion predicted by our model, we use GART [3] as a downstream application. GART represents the animal geometry using 3D Gaussians defined in a canonical space, which are articulated using the template prior. Here, we provide additional GART results for tiger and dog sequences in Fig. 4 and Table 3. As shown in the quantitative results, Ours-tto + GART achieves the best PSNR and SSIM, while also obtaining competitive LPIPS compared with AniMer* + GART. Visually, our model re-

covers sharp and plausible geometries, particularly preserving distinct shape details that are often blurred or distorted in other methods. Please refer to the supplementary video for animated visualizations of our application results for animatable animal reconstruction.

Table 3. **Quantitative results on in-the-wild videos.** We report both photometric and perceptual metrics on the rendered target views.

Data	Method	PSNR ↑	SSIM ↑	LPIPS ↓
Tiger	AniMer* + GART	14.58	0.5865	**0.3334**
	GenZoo* + GART	13.04	0.5456	0.4304
	Ours + GART	13.00	0.5681	0.4273
	Ours-tto + GART	**15.79**	**0.6265**	0.3399
Dog	AniMer* + GART	20.08	0.8054	0.2488
	GenZoo* + GART	17.76	0.7619	0.2909
	Ours + GART	19.83	0.8182	0.1771
	Ours-tto + GART	**24.13**	**0.8817**	**0.1354**

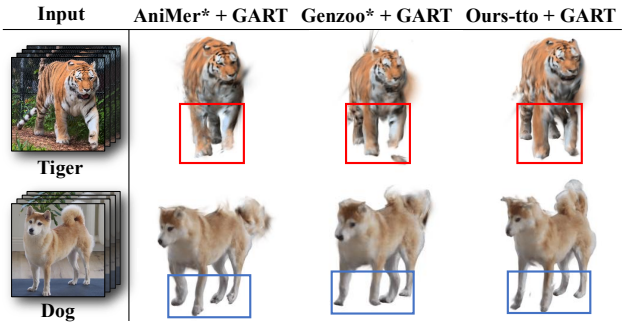


Figure 4. **Qualitative results on in-the-wild videos.** We render all methods from the same canonical pose for a fair comparison. AniMer* + GART and GenZoo* + GART show blurrier limbs and less stable parts than our method. Our method produces sharper textures and more consistent limb geometry across species, showing that our temporally coherent motion provides a stable input for animatable reconstruction.

References

[1] Benjamin Biggs, Oliver Boyne, James Charles, Andrew Fitzgibbon, and Roberto Cipolla. Who left the dogs out? 3d animal reconstruction with expectation maximization in the loop. In *ECCV*, 2020. 2

[2] 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 *ECCV*, 2016. 1

[3] Jiahui Lei, Yufu Wang, Georgios Pavlakos, Lingjie Liu, and Kostas Daniilidis. Gart: Gaussian articulated template models. In *CVPR*, 2024. 4

[4] Yuxin Liu, Minshan Xie, Hanyuan Liu, and Tien-Tsin Wong. Text-guided texturing by synchronized multi-view diffusion. In *SIGGRAPH Asia*, 2024. 2

[5] Jin Lyu, Tianyi Zhu, Yi Gu, Li Lin, Pujin Cheng, Yebin Liu, Xiaoying Tang, and Liang An. Animer: Animal pose and shape estimation using family aware transformer. In *CVPR*, 2025. 2, 3

[6] Tomasz Niewiadomski, Anastasios Yiannakidis, Hanz Cuevas-Velasquez, Soubhik Sanyal, Michael J Black, Silvia Zuffi, and Peter Kulits. Generative zoo. In *ICCV*, 2025. 2

[7] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. In *NeurIPS*, 2021. 4

[8] Jiacong Xu, Yi Zhang, Jiawei Peng, Wufei Ma, Artur Jesslen, Pengliang Ji, Qixin Hu, Jiehua Zhang, Qihao Liu, Jiahao Wang, et al. Animal3d: A comprehensive dataset of 3d animal pose and shape. In *ICCV*, 2023. 2

[9] Xiangpeng Yang, Linchao Zhu, Hehe Fan, and Yi Yang. Videograin: Modulating space-time attention for multi-grained video editing. In *ICLR*, 2025. 1, 2

[10] Hang Yu, Yufei Xu, Jing Zhang, Wei Zhao, Ziyu Guan, and Dacheng Tao. Ap-10k: A benchmark for animal pose estimation in the wild. *arXiv preprint arXiv:2108.12617*, 2021. 2

[11] Silvia Zuffi, Angjoo Kanazawa, David W Jacobs, and Michael J Black. 3d menagerie: Modeling the 3d shape and pose of animals. In *CVPR*, 2017. 2