SAM 3D Animal: Promptable Animal 3D Reconstruction from Images in the Wild

Xuyi Hu[1*] **Jin Lyu**[2*] **Jiuming Liu**[1] **Yebin Liu**[3] **Silvia Zuffi**[4] **Liang An**[5†] **Stefan Goetz**[1]

[1]University of Cambridge [2]Southern University of Science and Technology [3]Tsinghua University
[4]IMATI-CNR, Milan, Italy [5]Nanjing University



Figure 1: A promptable view of multi-animal 3D reconstruction. We present **SAM 3D Animal**, a promptable framework that addresses the problem of joint 3D reconstruction of multiple animals from a single image. Here, we show the input image together with the overlay reconstruction results.

Abstract

3D animal reconstruction in the wild remains challenging due to large species variation, frequent occlusions, and the prevalence of multi-animal scenes, while existing methods predominantly focus on single-animal settings. We present SAM 3D Animal, the first promptable framework for multi-animal 3D reconstruction from a single image. Built on the SMAL+ parametric animal model, our method jointly reconstructs multiple instances and supports flexible prompts in the form of keypoints and masks which enable more reliable disambiguation in crowded and occluded scenes. To train such a model, we further introduce Herd3D, a multi-animal 3D dataset containing over 5K images, designed to increase diversity in species, interactions, and occlusion patterns. Experiments on the Animal3D, APTv2, and Animal Kingdom datasets show that our framework achieves state-of-the-art results over both existing model-based and model-free methods, demonstrating a scalable and effective solution for prompt-driven animal 3D reconstruction in the wild. Project page: `https://georgehux.com/SAM3D-Animal-project-page/`

1 Introduction

Animals are a fundamental part of the visual world, yet 3D reconstruction research remains heavily skewed toward humans. Human-centric methods have advanced pose and shape estimation dramatically [21, 61, 60, 18, 47, 6, 29]. In contrast, animal reconstruction still suffers from scarce datasets, broad species variation, and inconsistent anatomical definitions.

[*]Equal contribution.
[†]Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Parametric animal models such as SMAL [65] and SMAL+ [63] provide an effective basis for recovering 3D pose and shape from a single image [64, 53, 39, 36, 2]. These approaches typically focus on one animal at a time and often rely on pre-cropped inputs or strong object detections. However, many in-the-wild animal scenes contain multiple individuals with mutual occlusion, and complex interactions that invalidate single-animal assumptions.

Multi-animal 3D reconstruction raises unique challenges beyond those of the single-object case. First, instance association becomes ambiguous when animals overlap or occlude one another. Second, pose and shape estimation must be jointly consistent across multiple hypotheses, since mistakes on one individual can be amplified by false depth ordering or incorrect occlusion reasoning. Third, available datasets rarely provide dense multi-animal 3D annotations, which hinders supervised learning for crowded scenes.

To overcome these challenges, we draw inspiration from promptable reconstruction in human vision. Recent works such as SAM 3D Body [56] demonstrates that explicit prompts can guide a model to focus on a desired subject and resolve ambiguity in cluttered scenes. Prompts can take the form of keypoints or masks, each providing a different level of spatial and semantic guidance.

In this paper, we introduce **SAM 3D Animal**, the first promptable framework for multi-animal 3D reconstruction, see Fig. 1. Our model uses the SMAL+ template [63] and can ingest optional prompts in two modalities: keypoints for skeletal alignment and masks for precise silhouette discrimination. This promptable design allows SAM 3D Animal to recover multiple animals jointly from a single image. Different from SAM 3D Body, which reconstructs a single prompted subject per forward pass, our model adopts a set-prediction paradigm that recovers all animal instances in one shot via DETR-style [12] bipartite matching, eliminating the need for per-instance bounding-box cropping.

However, training such a multi-instance model with only 2D-annotated datasets is insufficient, as 2D keypoints and silhouettes alone cannot provide the per-instance 3D shape supervision needed to resolve inter-animal occlusions. To address this, we propose Herd3D, a multi-animal 3D dataset containing over 5K images with per-instance ground-truth meshes, designed to increase diversity in species, interactions, and occlusion patterns. The generation pipeline of Herd3D is adapted from GenZoo [39], where each animal is synthetically generated to produce an image with an image-aligned SMAL+ model.

To demonstrate the effectiveness of SAM 3D Animal, we compare it with state-of-the-art animal mesh recovery methods on publicly available datasets including Animal3D [53], APTv2 [57], and Animal Kingdom [38]. Even without any prompt, our model achieves competitive or superior results compared to existing methods. When prompts are provided, performance improves consistently across all benchmarks, with up to 54% AP gain and 80% mAP gain on the out-of-domain Animal Kingdom dataset over the strongest baseline, as well as 5.2 PA-MPJPE improvement on Animal3D. Ablation studies confirm that Herd3D brings consistent improvements, particularly on multi-animal benchmarks, and that keypoint prompts are the dominant contributor among prompt modalities, with performance scaling monotonically as the number of keypoints increases.

2 Related Work

2.1 Model-Free Reconstruction

Model-free animal reconstruction learns 3D structure directly from image or video collections without assuming a predefined template [17, 52]. Early methods model category-specific articulated animals from single-view image collections by separating a predefined skeleton prior from instance-specific deformations [59, 51, 50, 19]. Later approaches extend to a wider variety of species, either by learning a unified shape model [30] or by applying linear skinning to deform learned 3D object shapes [4]. However, these methods still struggle with extreme poses, heavy occlusions, and limited viewpoint coverage, often producing geometrically ambiguous reconstructions.

2.2 Model-Based Reconstruction

Model-based animal reconstruction typically relies on predefined quadruped templates such as SMAL [65] that encode the shape and articulation structure of specific animal families. These models either fit predefined 3D templates to animal images using 2D observations such as keypoints or

silhouettes [64, 8, 9], or directly reconstruct the 3D shape from image or video observations [13, 55, 59, 66, 35, 14]. This parametric formulation offers interpretable and controllable representations, which make the reconstructed animals readily animatable and editable. Recent works further extend SMAL-based reconstruction to broader quadruped species and training settings. AWOL [63] maps CLIP-style embeddings to the SMAL$^+$ parameter space for language- and image-guided animal shape generation. RAW [24] reconstructs animals jointly with their surrounding environment, including multi-animal scenes. However, it relies on rigid animal assets rather than articulated animal models, and therefore does not address fine-grained articulated animal reconstruction.

2.3 Animal Pose Estimation Datasets

In comparison to humans, the construction of large-scale animal datasets is significantly more challenging because animals are difficult to capture in controlled environments and exhibit substantial morphological diversity across species. Animal benchmarks, such as Stanford Extra [7], Animal Pose [10], and AwA-Pose [5], remain limited to 2D annotations. Existing 3D animal datasets, such as Animal3D [53] and GenZoo [39] predominantly focus on single-animal instances. Additionally, multi-view quadrupedal animal datasets focus mostly on single species such as dogs [41, 48, 22], horses [26], and some agriculture animals like pigs [3], goats [20] or cattle [40]. Meanwhile, real-captured dynamic datasets targeting wildlife animals such as WildDepth [1] and AiM [62], and sim-to-real datasets for diverse species [35, 32, 37, 34, 27], all focus on single instances. Previous large-scale social animal benchmarks like APT-36K [58] and Animal Kingdom [38] provide only 2D annotations. This limitation restricts the development of methods that can jointly model inter-animal occlusions, spatial relationships, and pose dependencies in multi-animal scenes.

2.4 Promptable Mesh Reconstruction

Promptable mesh reconstruction has recently emerged in human mesh recovery, where auxiliary cues guide 3D estimation under occlusion and crowding. PromptHMR [46] incorporates full-image context with spatial and semantic prompts for pose and shape estimation. SAM 3D Body [56] extends this idea to full-body recovery through a promptable encoder-decoder architecture supporting keypoint and mask prompts. SAM-Body4D [16] further leverages temporally consistent masklets to produce coherent mesh trajectories from videos. However, these methods are designed for humans, whereas our work targets the animal domain, where large morphological diversity, inter-animal occlusion, and multi-instance interactions must be jointly considered.

3 SAM 3D Animal Model

3.1 Preliminary

SMAL+. SMAL+ [63], denoted as $\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma})$, extends the original SMAL [65] model by incorporating training samples from D-SMAL [43] and hSMAL [25], alongside new species such as the giraffe, bear, mouse and rat. This results in a broader, 145-dimensional shape space learned from a total of 145 animals. The inputs to SMAL+ are the shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{145}$ and the pose parameters $\boldsymbol{\theta} \in \mathbb{R}^{35\times 3}$ (using an axis-angle representation) and the global translation $\boldsymbol{\gamma} \in \mathbb{R}^3$. By applying linear blendshapes and Linear Blend Skinning (LBS), SMAL+ outputs a posed mesh with vertices $\mathbf{V} \in \mathbb{R}^{3889\times 3}$ and faces $\mathbf{F} \in \mathbb{N}^{7774\times 3}$.

3.2 End-to-end Multi-Instance Network

Given an animal image, our model can reconstruct all animals in the image without requiring preprocessed bounding boxes, and support masks or keypoints as prompts.

Encoder. Starting from the image $\boldsymbol{x} \in \mathbb{R}^{H\times W\times 3}$, we utilize the ViT-Huge Encoder [15] to generate the feature tokens $\mathcal{F} \in \mathbb{R}^{(H_0\times W_0)\times C_0}$, where C_0, H_0, and W_0 are the channels, the height and width of the feature map, respectively. In our case, $W = H = 512$, $C_0 = 1280$, $W_0 = H_0 = 32$.

Decoder. Inspired by SAM 3D Body [56], the decoder is a SAM-style promptable Transformer (see Fig. 2). Specifically, it takes the feature tokens $\mathcal{F}$ and a set of query tokens as input, then performs cross-attention, and finally predicts the SMAL+ parameters, cameras, and bounding boxes. Note

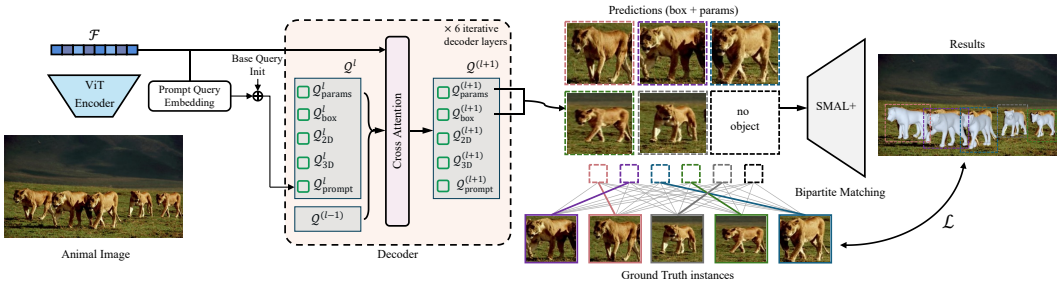


Figure 2: SAM 3D Animal model structure.

that, different from SAM 3D Body which predicts single person at a time, our model directly predict $P = 30$ possible instances at a time, eliminating the need for bounding box input. The query tokens consist of five distinct token groups at state index $l = 0, \ldots, L$, where $L = 6$ and $l = 0$ denotes the decoder input:

$$\mathcal{Q}^{(l)} = \text{Concat}(\mathcal{Q}^{(l)}_{\text{params}}, \mathcal{Q}^{(l)}_{\text{box}}, \mathcal{Q}^{(l)}_{\text{2D}}, \mathcal{Q}^{(l)}_{\text{3D}}, \mathcal{Q}^{(l)}_{\text{prompt}}) \in \mathbb{R}^{N \times D}, \tag{1}$$

where $\mathcal{Q}^{(l)}_{\text{params}}$, $\mathcal{Q}^{(l)}_{\text{box}}$, $\mathcal{Q}^{(l)}_{\text{2D}}$, and $\mathcal{Q}^{(l)}_{\text{3D}}$ represent the SMAL+ pose tokens, bounding box tokens, 2D keypoint tokens, and 3D keypoint tokens, respectively, while $\mathcal{Q}^{(l)}_{\text{prompt}}$ represents the interaction prompt tokens. The four base query groups, $\mathcal{Q}^{(0)}_{\text{params}}$, $\mathcal{Q}^{(0)}_{\text{box}}$, $\mathcal{Q}^{(0)}_{\text{2D}}$, and $\mathcal{Q}^{(0)}_{\text{3D}}$, are learnable embeddings. They are randomly initialized once before training and subsequently optimized jointly with the network. At inference time, their learned values are used without reinitialization. In contrast, $\mathcal{Q}^{(0)}_{\text{prompt}}$ is input-dependent and constructed from the keypoint prompts provided for the current image, rather than being a randomly initialized base query. A keypoint prompt $p_i = (u_i, v_i, c_i)$ comprises normalized coordinates $(u_i, v_i) \in [0, 1]^2$ and a label: an Animal3D keypoint ID 0–25, a negative point -1, or an invalid, missing, or padded point -2. Its token is

$$q^{(0)}_{\text{prompt},i} = W_p e_i + b_p, \qquad e_i = \begin{cases} \text{PE}(u_i, v_i) + E(c_i), & c_i \in \{0, \ldots, 25\}, \\ E_{\text{neg}}, & c_i = -1, \\ E_{\text{invalid}}, & c_i = -2, \end{cases}$$

where PE denotes positional encoding, $E(c_i), E_{\text{neg}}, E_{\text{invalid}}$ are learned label embeddings, and W_p, b_p are learned projection parameters. Prompts enter the first decoder layer through $\mathcal{Q}^{(0)} = \text{Concat}(\mathcal{Q}^{(0)}_{\text{base}}, \mathcal{Q}^{(0)}_{\text{prompt}})$, with $\mathcal{Q}^{(0)}_{\text{base}}$ collecting the four base query groups. The feature dimension is $D = 1024$. At inference, $N = 55P + N_{\text{prompt}} = 1650 + N_{\text{prompt}}$: each hypothesis has one parameter, one previous-estimate, one box, and 26 tokens each for 2D and 3D keypoints; N_{prompt} counts prompt tokens including padding.

Decoder block $l + 1$ maps $\mathcal{Q}^{(l)}$ to $\mathcal{Q}^{(l+1)}$ (Fig. 2):

$$\mathcal{Q}^{(l+1)} = \mathcal{D}_{l+1}(\mathcal{Q}^{(l)}, A^{(l)}, \mathcal{F}), \qquad l = 0, \ldots, L-1, \tag{2}$$

Here, $\mathcal{D}_{l+1}$ comprises self-attention, image cross-attention, and a feed-forward network; $A^{(l)}$ denotes query positional augmentation. Fig. 2 highlights cross-attention. Its lower $\mathcal{Q}^{(l-1)}$ label schematically denotes previous-estimate conditioning, encoded once before decoding (using the initial estimate if none is supplied); it does not index a decoder layer.

For $l = 0, \ldots, L-2$, SMAL+ and camera predictions yield projected 2D coordinates $\hat{x}^{(l+1)}_{\text{2D}}$ and pelvis-centered 3D coordinates $\bar{x}^{(l+1)}_{\text{3D}}$. Feedback updates the next input in place:

$$\begin{aligned} \mathcal{Q}^{(l+1)}_{\text{2D}} &\leftarrow \mathcal{Q}^{(l+1)}_{\text{2D}} + \phi_{\text{feat}}(\mathcal{F}(\hat{x}^{(l+1)}_{\text{2D}})), \\ A^{(l+1)}_{\text{2D}} &\leftarrow m^{(l+1)} \odot \phi_{\text{pos}}(\hat{x}^{(l+1)}_{\text{2D}}), \end{aligned} \tag{3}$$

$$A^{(l+1)}_{\text{3D}} \leftarrow \psi_{\text{pos}}(\bar{x}^{(l+1)}_{\text{3D}}). \tag{4}$$

Here, $m^{(l+1)}$ selects in-frame projections with valid positive depth, $\mathcal{F}(\hat{x}^{(l+1)}_{\text{2D}})$ denotes masked bilinear samples, and $\phi_{\text{feat}}, \phi_{\text{pos}}, \psi_{\text{pos}}$ are learned mappings. Eqs. (3)–(4) follow Eq. (2), producing

$(\mathcal{Q}^{(l+1)}, A^{(l+1)})$ for the next block; 3D token contents and unlisted components remain unchanged. Fig. 2 omits these updates. The sixth block skips feedback and generates final predictions from its parameter and box groups; "params" includes SMAL+ and camera parameters. Algorithm 1 in the Appendix details the complete inference forward pass.

Bipartite Matching for Multi-Animal Instances. To enable end-to-end training without heuristic post-processing such as Non-Maximum Suppression (NMS), we adopt a set prediction formulation following the DETR paradigm [12]. Specifically, we employ bipartite matching via the Hungarian algorithm [23] to find the optimal one-to-one assignment between the fixed-size set of P predicted animal hypotheses and the M ground-truth instances. The matching cost is a weighted combination of bounding box $\mathcal{L}_1$ distance, Generalized IoU [42], focal-style confidence penalty [44], and masked 2D keypoint distance. Once the optimal assignment is established, predicted outputs are reordered for loss computation. See Appendix for more details.

3.3 Loss Functions

After establishing the correspondence between predicted outputs and ground-truth labels via bipartite matching, we optimize the model using a multi-task loss function formulated as:

$$\mathcal{L} = \lambda_{\text{params}}\mathcal{L}_{\text{params}} + \lambda_{\text{2D}}\mathcal{L}_{\text{2D}} + \lambda_{\text{3D}}\mathcal{L}_{\text{3D}} + \lambda_{\text{box}}\mathcal{L}_{\text{box}}, \tag{5}$$

where $\lambda_{\{\cdot\}}$ denotes the weighting coefficients used to balance the respective loss contributions.

Parameter Loss ($\mathcal{L}_{\text{params}}$) computes the ℓ_2 distance between the predicted SMAL+ shape and pose parameters and their corresponding ground-truth values if provided.

Keypoint Losses ($\mathcal{L}_{\text{2D}}$, $\mathcal{L}_{\text{3D}}$) represent the ℓ_1 distance between the ground-truth 2D and 3D keypoint positions and the ones regressed from predicted SMAL+, respectively.

Bounding Box Loss ($\mathcal{L}_{\text{box}}$) supervises localization accuracy through a combination of coordinate regression, geometric alignment, confidence scoring and the denoising training strategy of DN-DETR [28]:

$$\mathcal{L}_{\text{box}} = \mathcal{L}_{\text{coord}} + \mathcal{L}_{\text{giou}} + \mathcal{L}_{\text{conf}} + \mathcal{L}_{\text{dn}}, \tag{6}$$

where $\mathcal{L}_{\text{coord}}$ is an ℓ_1 loss over normalized bounding box coordinates, and $\mathcal{L}_{\text{giou}}$ refers to the Generalized IoU (GIoU) loss [42]. To refine the objectness score, $\mathcal{L}_{\text{conf}}$ employs a Binary Cross-Entropy (BCE) loss, where the actual IoU between the matched predicted and ground-truth boxes serves as the soft target for the predicted confidence.

4 Herd3D Dataset



Figure 3: Example from the Herd3D dataset. This figure shows a generated scene with eight dogs, together with the corresponding canny map, depth map, RGB image, and 3D reconstruction overlay. The full Herd3D dataset contains multi-animal scenes with 2 to 8 animals per image.

To support multi-animal 3D reconstruction in real-world scenarios, we construct Herd3D, a large-scale dataset specifically designed for multi-animal scenes which contains over 5K images and 118 species (see Fig. 3). We believe that GenZoo [39] provides a strong and practical starting point for constructing large-scale animal datasets, because it couples a parametric animal model with controllable image synthesis, which enables scalable generation of paired images and geometry while maintaining pose and shape consistency. Building on GenZoo, we adapt the pipeline for multi-animal data generation. To construct group layouts, we sample up to 8 animals per image and place them on a shared ground plane. For each instance, we set $t_y = 0$ and sample translations with $t_x \in [-1.5, 1.5]$ using non-adjacent horizontal bins to limit excessive overlap, and $t_z \in [8, 50]$ from

predefined depth intervals while constraining the group depth span to at most 30; we add a small x and z jitter within ± 1.5 and apply a fixed ground alignment offset of 0.3. We further diversify global orientation by sampling pitch in $[-15^\circ, 15^\circ]$ and yaw in $[0^\circ, 360^\circ]$. To accommodate the increased complexity of multi-animal scenes-including frequent occlusions, higher ambiguity in instance-wise orientation and limited pose diversity, we adapt the GenZoo pipeline with several targeted modifications. We (i) impose scene layout constraints by placing all animals on a shared ground plane, (ii) expand the pose pool by integrating Animal3D poses [53] to increase pose diversity, (iii) replace the ControlNet backend with Qwen-Image-ControlNet-Union [49] to better preserve geometry and occlusion ordering, and (iv) resolve multi-animal orientation ambiguity via a two-stage Qwen3-VL-8B-Instruct [45] prompting scheme, which first predicts left-to-right per-animal facing directions and then composes a single coherent final prompt that integrates the species information, camera settings, and scene attributes. Each synthetic image has a resolution of 1024 × 1024 and includes annotations for SMAL+ parameters, 2D keypoints, 3D keypoints and bounding boxes.

5 Experiments

Datasets. We curate a comprehensive training corpus of 49.2K images containing both 2D and 3D annotations. Specifically, we aggregate the training splits of Animal Pose [10], APTv2 [57], AwA-Pose [5], Stanford Extra [7], Animal3D [53], and our newly introduced Herd3D. For evaluation, following the protocol in AniMer [36], we report results on two in-domain datasets (Animal3D and APTv2) alongside an out-of-domain (OOD) dataset Animal Kingdom.

Baselines. We benchmark our approach against three recent state-of-the-art (SOTA) methods to ensure a comprehensive evaluation across different architectural paradigms. For model-based techniques, we compare with AniMer [36], a transformer-based architecture utilizing the SMAL model, and GenZoo [39], which builds upon the SMAL+ variant. Additionally, we include 3D Fauna [30] as a representative SOTA model-free reconstruction approach.

Evaluation Metrics. We evaluate 3D accuracy using the Procrustes-Aligned Mean Per Joint Position Error (PA-MPJPE). For 2D accuracy, we report the Percentage of Correct Keypoints (PCK) and keypoint AP/mAP based on Object Keypoint Similarity (OKS) [31]. OKS is computed over visible keypoints, with errors normalized by the ground-truth visible-keypoint bounding-box area and $\sigma = 0.05$. We report AP^{kp}_{50} and AP^{kp}_{75} at OKS thresholds of 0.50 and 0.75, and $\text{mAP}^{kp}_{50:95}$ averaged over thresholds from 0.50 to 0.95 in steps of 0.05.

Implementation Details. Our network is optimized using AdamW [33] with an initial learning rate of 2×10^{-5}, incorporating a linear warmup over the first 15 epochs. Similar to AniMer [36], we employ a two-stage training strategy consisting of 250 epochs for the first stage and 250 epochs for the second. We apply prompt dropout for robustness: the mask prompt is dropped with 50% probability, the entire keypoint prompt is dropped with a probability of $p_{\text{drop}} = 0.2$, and otherwise each keypoint is independently masked out with a rate sampled uniformly from $[0,\ 0.7]$ per step, encouraging the model to handle partial or absent prompts at inference time. Training is distributed across four RTX 4090 GPUs with a gradient accumulation step of 16. To balance the objective function, our empirical loss weighting factors are set to $\lambda_{\text{params}} = 1$, $\lambda_{\text{2D}} = 5$, $\lambda_{\text{3D}} = 5$, and $\lambda_{\text{box}} = 1$.

5.1 Comparison

Comparison without prompts. We present the quantitative results in Table 1. Without any prompt, our method already achieves competitive or superior performance relative to existing approaches. On Animal3D, the prompt-free variant attains a PA-MPJPE of 80.7 mm, slightly outperforming AniMer (81.0 mm), while achieving a higher mAP (49.3 vs. 47.2). On APTv2, keypoint localization improves substantially, with PCK@0.1 reaching 87.9, far surpassing GenZoo (64.1) and AniMer (62.4), though AP remains lower (49.4 vs. 55.5 for GenZoo), indicating that the two paradigms exhibit complementary strengths under different metrics. On the OOD Animal Kingdom benchmark, our prompt-free results lead all metrics, demonstrating stronger generalization to unseen scenes.

Prompt-driven performance gains. A key advantage of our framework is its ability to leverage auxiliary prompts at inference time. When supplied with keypoints from an off-the-shelf ViTPose [54] estimator, performance improves consistently: on APTv2, AP rises from 49.4 to 55.5 and mAP from 23.5 to 27.9; on Animal Kingdom, AP increases from 45.0 to 50.5. This practical variant already

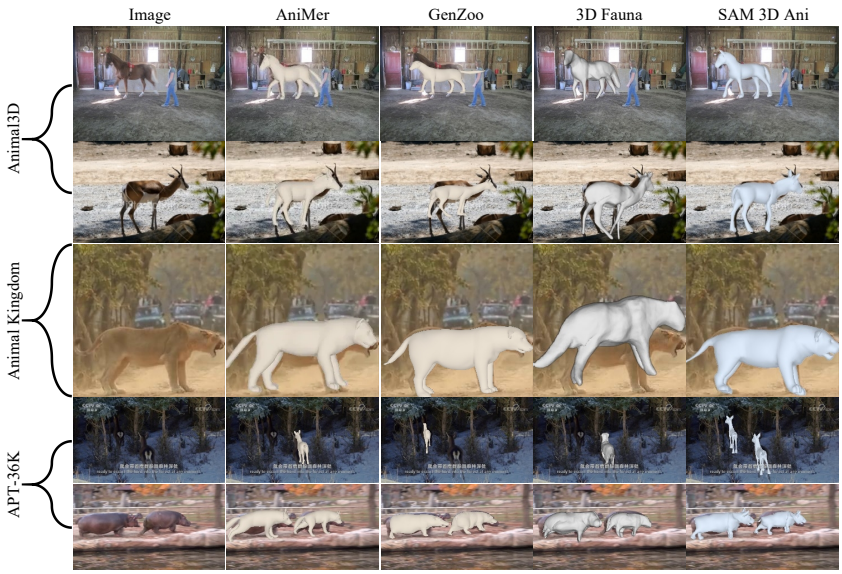


Figure 4: Qualitative comparisons on Animal3D, Animal Kingdom and APT-36K datasets. We compare our results with AniMer [36], GenZoo [39] and 3D Fauna [30]. AniMer, GenZoo, and 3D Fauna require pre-cropped single-animal images as input, whereas our method directly processes the full image and retrieves reconstructions whose confidence exceeds a predefined threshold, eliminating the need for cropping.

Table 1: **Quantitative comparisons on the Animal3D dataset, APTv2 dataset and Animal Kingdom dataset.**

Method	Animal3D			APTv2			Animal Kingdom		
	PA-MPJPE↓	AP↑	mAP↑	PCK@0.1↑	AP↑	mAP↑	PCK@0.1↑	AP↑	mAP↑
3D Fauna [30]	202.8	63.0	10.0	32.6	39.5	7.0	31.4	13.5	2.9
GenZoo [39]	123.2	85.3	34.0	64.1	55.5	19.2	42.1	23.6	4.7
AniMer [36]	81.0	93.4	47.2	62.4	51.8	20.1	58.2	39.7	10.4
Ours	80.7	92.8	49.3	87.9	49.4	23.5	59.2	45.0	12.6
Ours (w/ vitpose)	80.7	92.8	50.0	88.8	55.5	27.9	63.0	50.5	14.5
Ours (w/ GT prompt)	**75.8**	**93.8**	**53.4**	**89.0**	**57.4**	**29.7**	**66.2**	**60.1**	**17.7**

matches or surpasses the best baseline on most metrics. With ground-truth keypoint prompts, the gains become substantially larger: on APTv2, PCK@0.1 reaches 89.0 (vs. 62.4 for AniMer) and AP reaches 57.4 (vs. 55.5 for GenZoo); on Animal Kingdom, AP improves to 60.1 and mAP to 17.7, roughly doubling AniMer's 10.4. These results confirm that prompting provides a scalable mechanism for improving reconstruction quality, with performance increasing monotonically as prompt fidelity improves—a unique advantage that existing methods cannot replicate.

Qualitative comparison. Fig. 4 presents visual comparisons across the three benchmarks. 3D Fauna, as a model-free approach, produces coarse reconstructions that lack geometric detail. GenZoo and AniMer yield plausible shapes but exhibit less accurate alignment with the input image. Our method consistently produces reconstructions that are better aligned with the observed pose and viewpoint. Additional results spanning diverse species and challenging in-the-wild scenarios are shown in Fig. 5.

Performance under Varying Numbers of Animals. We further evaluate how reconstruction performance changes with the number of animals in an image. We stratify the APTv2 test set into images containing 1, 2, 3, 4, and 5+ annotated animals, corresponding to 3,502, 1,384, 557, 166, and 426 images, respectively (6,035 images in total). This grouping captures increasingly challenging conditions with stronger crowding, occlusion, and smaller animal scale. We report instance-averaged PCK@0.1, AP, and mAP for each group in Table 2. For direct comparison with existing methods, *Ours (w/ bbox)* follows the same protocol as the baselines: each animal is cropped using its ground-truth bounding box and reconstructed independently. Under this setting, SAM 3D Animal achieves the best performance across all metrics and animal-count groups. For example, it

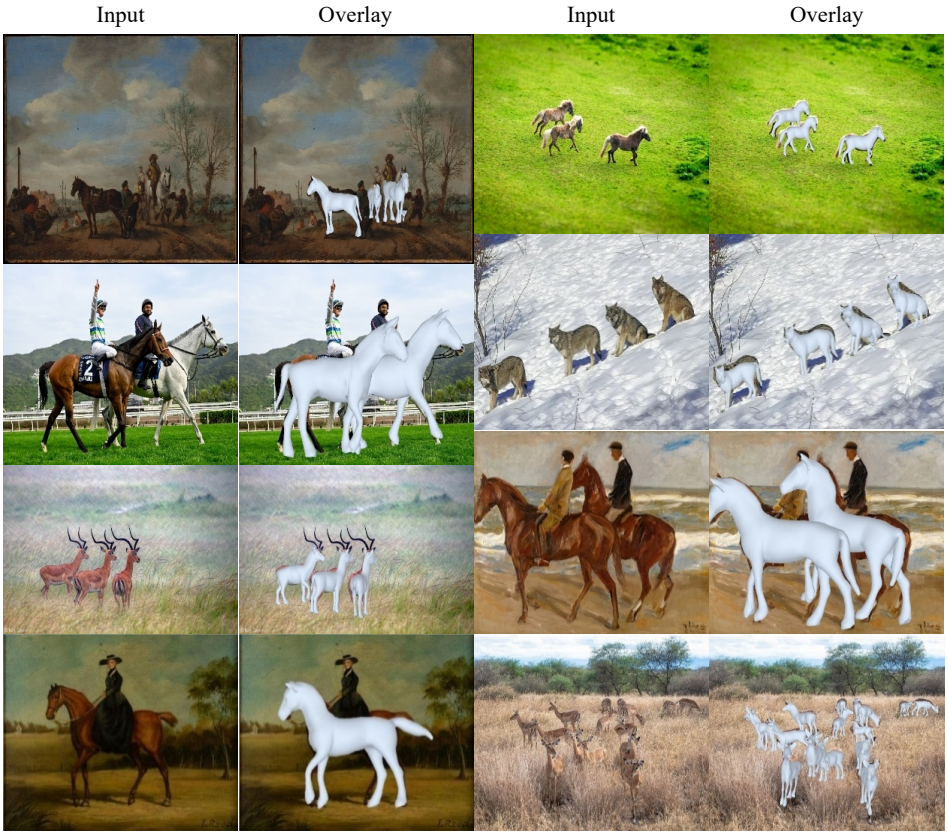


Figure 5: Qualitative evaluation of SAM 3D Animal. For each example, we show: (a) the input image, and (b) the 3D reconstruction overlay. To demonstrate the robustness of SAM 3D Animal, we visualize results across diverse animal species and challenging in-the-wild scenarios, including unusual poses, large viewpoint variations, and crowded scenes with several animals.

Table 2: **Quantitative comparisons on the APTv2 dataset stratified by the number of animals per image.**

Method	1 Animal			2 Animals			3 Animals			4 Animals			5+ Animals		
	PCK@0.1↑	AP↑	mAP↑	PCK@0.1↑	AP↑	mAP↑	PCK@0.1↑	AP↑	mAP↑	PCK@0.1↑	AP↑	mAP↑	PCK@0.1↑	AP↑	mAP↑
3D Fauna [30]	39.8	44.6	8.4	31.5	40.4	8.0	32.5	38.1	7.8	28.7	35.6	5.7	26.8	34.8	4.5
GenZoo [39]	68.6	57.5	19.8	58.7	39.3	12.0	59.7	40.4	12.7	57.5	45.6	15.2	48.9	24.1	6.6
AniMer [36]	78.4	69.0	29.8	64.2	53.1	19.6	61.8	55.9	20.9	52.0	50.2	19.2	46.7	31.5	10.4
Ours (w/o GT prompt)	83.4	81.1	42.9	64.9	50.4	21.8	62.1	48.0	21.2	42.7	31.9	12.7	28.9	13.3	4.0
Ours (w/ bbox)	**86.8**	**84.1**	**46.8**	**74.5**	**63.2**	**30.0**	**78.2**	**66.5**	**33.6**	**66.9**	**55.0**	**26.0**	**57.7**	**37.8**	**14.9**

obtains 86.8/84.1/46.8 PCK/AP/mAP on single-animal images and remains strongest for images with 5+ animals, reaching 57.7/37.8/14.9 and outperforming the best baseline by 11.0, 6.3, and 4.5 points, respectively. Without GT prompts, *Ours (w/o GT prompt)* processes the full image and jointly reconstructs all animals without bounding boxes, keypoints, or masks. It remains competitive for single-animal and moderately crowded scenes, but performance decreases more noticeably for 4 and 5+ animals, where jointly detecting and reconstructing all instances becomes substantially more difficult under severe occlusion and scale variation.

5.2 Ablation

We ablate three design axes to isolate their respective contributions: training data, prompt modality, and prompt density. Results are reported in Tables 3 and 4, with qualitative examples in Fig. 6.

Effect of Herd3D. Removing Herd3D from the training set leads to a consistent performance drop across all three benchmarks, with the largest degradation observed on APTv2 (Table 3). This is

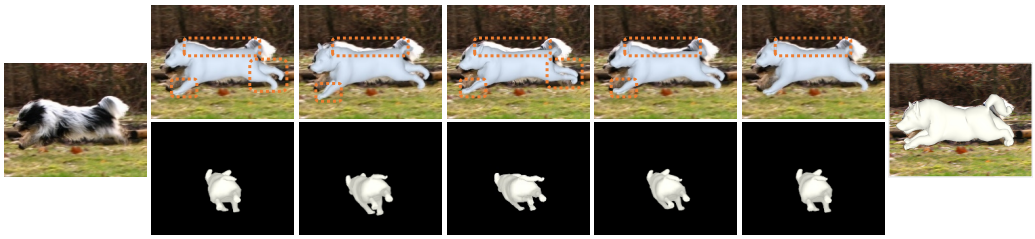


Figure 6: Ablation studies. Keypoint prompting, mask prompting, and training with our Herd3D dataset each lead to improved performance, as discussed in Sec. 5.2.

Table 3: **Ablation study on the Animal3D dataset, APTv2 dataset and Animal Kingdom dataset.**

Variant	Animal3D			APTv2			Animal Kingdom		
	PA-MPJPE↓	AP↑	mAP↑	PCK@0.1↑	AP↑	mAP↑	PCK@0.1↑	AP↑	mAP↑
w/o Herd3D	79.5	93.8	51.6	88.0	48.7	24.7	62.5	51.1	15.7
w/o prompt	80.7	92.8	49.3	87.9	49.4	23.5	59.2	45.0	12.6
w/o kp	80.8	93.1	49.8	87.9	49.7	23.5	59.3	45.4	12.8
w/o mask	**75.8**	**94.1**	**53.4**	88.9	57.3	29.6	66.0	**60.1**	17.5
Full	**75.8**	93.8	**53.4**	**89.0**	**57.4**	**29.7**	**66.2**	**60.1**	**17.7**

expected: Herd3D is the primary source of multi-animal scenes with per-instance annotations, and its absence disproportionately affects benchmarks that feature crowded or overlapping subjects. The result validates that our curated dataset fills a genuine gap in existing training resources rather than merely increasing data volume.

Keypoint vs. mask prompts. Comparing the prompt modality variants in Table 3 reveals a clear asymmetry: keypoint prompts are the dominant contributor, while mask prompts provide only marginal gains. Specifically, removing the mask prompt (*w/o mask*) leaves performance nearly unchanged from the full model, whereas removing keypoint prompts (*w/o kp*) reduces results to the level of the unprompted baseline. We attribute this to two factors. First, keypoints encode the articulated skeletal structure directly, which is precisely what the SMAL model needs to resolve pose ambiguity; masks, by contrast, convey only silhouette-level information that is largely redundant with the image features already extracted by the backbone. Second, the mask prompts are generated by SAM [11] at inference time, and segmentation errors, particularly on thin limbs or under occlusion, introducing noise that dilutes the prompt signal. Nonetheless, the full model still edges ahead of the keypoint-only variant on APTv2 and Animal Kingdom, suggesting that mask prompts offer a small but consistent complementary benefit when segmentation quality is adequate.

Number of keypoints. Table 4 displays the performance as a function of the number of keypoint prompts, ranging from 0 to 15. For each sample, we randomly select the specified number of keypoints from all available annotations as the prompt input. Performance improves monotonically, with the steepest gain occurring between 0 and 5 keypoints, even a sparse set of randomly chosen landmarks is sufficient to substantially disambiguate pose. Beyond 5, improvements continue at a diminishing rate, indicating that the initial keypoints resolve the most salient ambiguities while additional ones refine secondary joints. This graceful scaling is practically appealing: users can supply as few or as many keypoints as available, and the model extracts value from each additional annotation without requiring a fixed-size input.

5.3 Impact of Prompt Strategy Under Occlusion

To understand how different prompt strategies behave under varying levels of occlusion, we partition the test set by the number of visible keypoints into three groups (Low, Mid, High) and evaluate three prompt modes: GT Prompt (ground-truth 2D keypoints), ViTPose Prompt (automatically detected by ViTPose [54]), and No Prompt. Results are reported using mAP on both APTv2 and Animal Kingdom, as shown in Fig. 7. Note that for Animal Kingdom, the lowest visibility group starts at 8–11 keypoints because the dataset's annotations contain a minimum of 8 visible keypoints per instance.

Table 4: **Ablation study on the number of prompt keypoints.**

Number	Animal3D			APTv2			Animal Kingdom		
	PA-MPJPE↓	AP↑	mAP↑	PCK@0.1↑	AP↑	mAP↑	PCK@0.1↑	AP↑	mAP↑
0	80.8	93.1	49.8	87.9	49.7	23.5	59.3	45.4	12.8
5	79.5	**94.1**	51.7	88.9	54.6	26.7	63.9	56.9	16.3
10	78.6	**94.1**	51.7	**89.1**	56.8	28.7	65.3	58.6	17.1
15	**77.4**	93.8	**52.5**	89.0	**57.5**	**29.6**	**66.0**	**60.2**	**17.7**

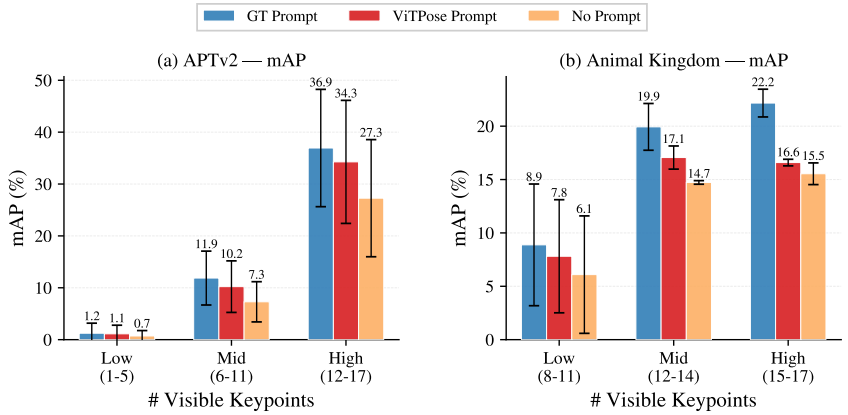


Figure 7: **Performance under different visibility levels.** We group test samples by the number of visible keypoints into Low, Mid, and High buckets. (a) mAP on APTv2. (b) mAP on Animal Kingdom. Error bars denote standard deviation across visibility counts within each group.

Prompt matters more under heavy occlusion. Across both datasets, the relative gain of prompting over the No Prompt baseline is largest in the Low visibility group and diminishes as more keypoints become visible. On APTv2, ViTPose Prompt improves mAP over No Prompt by 57% relative in the Low group, 40% in Mid, and 26% in High. On Animal Kingdom, the same trend holds: 28% relative gain in Low, 16% in Mid, and 7% in High. This confirms that the model relies more heavily on prompt-provided spatial priors when visual evidence is limited, and that prompting offers the greatest practical benefit precisely in the most challenging scenarios.

ViTPose Prompt is a practical alternative to GT. On APTv2, ViTPose Prompt consistently performs close to the GT upper bound across all visibility levels, demonstrating that an off-the-shelf keypoint detector can serve as an effective substitute for manual annotations at inference time.

6 Conclusion

We presented SAM 3D Animal, a promptable framework for multi-animal 3D reconstruction from a single image. Unlike prior animal reconstruction methods that predominantly focus on animal-centric images, our approach reconstructs multiple animals jointly through a set-prediction formulation and supports flexible keypoint and mask prompts to resolve ambiguity in crowded and occluded scenes.

Limitation. While SAM 3D Animal shows strong performance, it remains limited by the SMAL+ shape space and is therefore mainly applicable to quadruped-like animals. Moreover, relative depth ordering between animals is not explicitly constrained, which can cause inaccurate spatial arrangements under severe occlusion. Future work could explore more flexible animal representations, explicit depth-aware scene reasoning, and multi-view refinement by decoupling view-dependent and view-invariant parameters.

References

[1] Muhammad Aamir, Naoya Muramatsu, Sangyun Shin, Matthew Wijers, Jia-Xing Zhong, Xinyu Hou, Amir Patel, Andrew Loveridge, and Andrew Markham. Wilddepth: A multimodal dataset for 3d wildlife perception and depth estimation. In *European Conference on Computer Vision*, pages 40–58. Springer, 2026.

[2] Liang An, Jin Lyu, Li Lin, Pujin Cheng, Yebin Liu, and Xiaoying Tang. Animer+: Unified pose and shape estimation across mammalia and aves via family-aware transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):3233–3249, 2026.

[3] Liang An, Jilong Ren, Tao Yu, Tang Hai, Yichang Jia, and Yebin Liu. Three-dimensional surface motion capture of multiple freely moving pigs using mammal. *Nature Communications*, 14(1):7727, 2023.

[4] 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*, pages 10382–10391, 2024.

[5] Prianka Banik, Lin Li, and Xishuang Dong. A novel dataset for keypoint detection of quadruped animals from images. *arXiv preprint arXiv:2108.13958*, 2021.

[6] Fabien Baradel, Matthieu Armando, Salma Galaaoui, Romain Brégier, Philippe Weinzaepfel, Grégory Rogez, and Thomas Lucas. Multi-hmr: Multi-person whole-body human mesh recovery in a single shot. In *European Conference on Computer Vision*, pages 202–218. Springer, 2024.

[7] 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 *European Conference on Computer Vision*, pages 195–211. Springer, 2020.

[8] Benjamin Biggs, Thomas Roddick, Andrew Fitzgibbon, and Roberto Cipolla. Creatures great and smal: Recovering the shape and motion of animals from video. In *Asian Conference on Computer Vision*, pages 3–19. Springer, 2018.

[9] Piotr Borycki, Yizhe Zhu, Yongqiang Gao, PrzemysĹ Spurek, et al. Smal-pets: Smal based avatars of pets from single image. *arXiv preprint arXiv:2603.17131*, 2026.

[10] Jinkun Cao, Hongyang Tang, Hao-Shu Fang, Xiaoyong Shen, Cewu Lu, and Yu-Wing Tai. Cross-domain adaptation for animal pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9498–9507, 2019.

[11] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pages 138846–138923, 2026.

[12] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[13] Thomas J Cashman and Andrew W Fitzgibbon. What shape are dolphins? building 3d morphable models from 2d images. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):232–244, 2012.

[14] Gyeongsu Cho, Changwoo Kang, Donghyeon Soon, and Kyungdon Joo. Dogrecon: Canine prior-guided animatable 3d gaussian dog reconstruction from a single image: G. cho et al. *International Journal of Computer Vision*, 133(9):6332–6346, 2025.

[15] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[16] Mingqi Gao, Yunqi Miao, and Jungong Han. Sam-body4d: Training-free 4d human body mesh recovery from videos. *arXiv preprint arXiv:2512.08406*, 2025.

[17] Shubham Goel, Angjoo Kanazawa, and Jitendra Malik. Shape and viewpoint without keypoints. In *European Conference on Computer Vision*, pages 88–104. Springer, 2020.

[18] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14783–14794, 2023.

[19] 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)*, pages 852–861. IEEE, 2024.

[20] Bo Jin, Jin Lyu, Bin Zhang, Tao Yu, Liang An, Yebin Liu, Meili Wang, et al. Monocular mesh recovery and body measurement of female saanen goats. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 38670–38678, 2026.

[21] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018.

[22] Sinead Kearney, Wenbin Li, Martin Parsons, Kwang In Kim, and Darren Cosker. Rgbd-dog: Predicting canine pose from rgbd sensors. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8333–8342. IEEE, 2020.

[23] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[24] Peter Kulits, Michael J Black, and Silvia Zuffi. Reconstructing animals and the wild. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16565–16577, 2025.

[25] Ci Li, Nima Ghorbani, Sofia Broomé, Maheen Rashid, Michael J Black, Elin Hernlund, Hedvig Kjellström, and Silvia Zuffi. hsmal: Detailed horse shape and pose reconstruction for motion pattern recognition. *arXiv preprint arXiv:2106.10102*, 2021.

[26] Ci Li, Ylva Mellbin, Johanna Krogager, Senya Polikovsky, Martin Holmberg, Nima Ghorbani, Michael J Black, Hedvig Kjellström, Silvia Zuffi, and Elin Hernlund. The poses for equine research dataset (pferd). *Scientific Data*, 11(1):497, 2024.

[27] Ci Li, Yi Yang, Zehang Weng, Elin Hernlund, Silvia Zuffi, and Hedvig Kjellström. Dessie: Disentanglement for articulated 3d horse shape and pose estimation from images. In *Asian Conference on Computer Vision*, pages 268–288. Springer, 2024.

[28] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13619–13627, 2022.

[29] Hongjie Li, Heng Yu, Jiaman Li, Hong-Xing Yu, Ehsan Adeli, C Karen Liu, and Jiajun Wu. Anylift: Scaling motion reconstruction from internet videos via 2d diffusion. *arXiv preprint arXiv:2604.17818*, 2026.

[30] 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*, pages 9752–9762, 2024.

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[32] Dongxia Liu, Jie Ma, Xiaochen Yang, Jiancheng Zhang, Bin Xia, Zhehan Kan, Nisha Huang, Jun Liang, Wenming Yang, and Jin Li. Mozoo: Unleashing video diffusion power in animal fur and muscle simulation. *arXiv preprint arXiv:2605.13857*, 2026.

[33] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[34] Haimin Luo, Teng Xu, Yuheng Jiang, Chenglin Zhou, Qiwei Qiu, Yingliang Zhang, Wei Yang, Lan Xu, and Jingyi Yu. Artemis: Articulated neural pets with appearance and motion synthesis. *ACM Trans. Graph.*, 41(4), jul 2022.

[35] Jin Lyu, Liang An, Pujin Cheng, Yebin Liu, and Xiaoying Tang. 4dequine: Disentangling motion and appearance for 4d equine reconstruction from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 32496–32506, June 2026.

[36] 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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17486–17496, 2025.

[37] Lucas Maximilian Martini, Alex Lappe, Anna Bognár, Rufin Vogels, and Martin Giese. Bigmaq: a big macaque motion and animation dataset bridging image and 3d pose representations. In *International Conference on Learning Representations*, volume 2026, pages 117144–117175, 2026.

[38] 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*, pages 19023–19034, 2022.

[39] Tomasz Niewiadomski, Anastasios Yiannakidis, Hanz Cuevas-Velasquez, Soubhik Sanyal, Michael J Black, Silvia Zuffi, and Peter Kulits. Generative zoo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8492–8502, 2025.

[40] Ian Noronha, Heather Neave, and Upinder Kaur. Moocap: A multi-view benchmark for cow-object-human interaction and behavior dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27324–27333, 2026.

[41] Yichen Peng, Jyun-Ting Song, Chen-Chieh Liao, Kris Kitani, Hideki Koike, and Erwin Wu. Interpet4d: A multimodal 4d human-pet interaction dataset for pet motion generation. In *European Conference on Computer Vision*, pages 437–454. Springer, 2026.

[42] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019.

[43] 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*, pages 3876–3884, 2022.

[44] Chi Su, Xiaoxuan Ma, Jiajun Su, and Yizhou Wang. Sat-hmr: Real-time multi-person 3d mesh estimation via scale-adaptive tokens. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 16796–16806, June 2025.

[45] Qwen Team. Qwen3 technical report, 2025.

[46] Yufu Wang, Yu Sun, Priyanka Patel, Kostas Daniilidis, Michael J Black, and Muhammed Kocabas. Prompthmr: Promptable human mesh recovery. In *Proceedings of the computer vision and pattern recognition conference*, pages 1148–1159, 2025.

[47] Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. Tram: Global trajectory and motion of 3d humans from in-the-wild videos. In *European Conference on Computer Vision*, pages 467–487. Springer, 2024.

[48] Zan Wang, Siyu Chen, Luya Mo, Xinfeng Gao, Yuxin Shen, Lebin Ding, and Wei Liang. Dogmo: A large-scale multi-view rgb-d dataset for 4d canine motion recovery. *arXiv preprint arXiv:2510.24117*, 2025.

[49] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025.

[50] Shangzhe Wu, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. Dove: Learning deformable 3d objects by watching videos. *International Journal of Computer Vision*, 131(10):2623–2634, 2023.

[51] 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*, pages 8792–8802, 2023.

[52] Shangzhe Wu, Ameesh Makadia, Jiajun Wu, Noah Snavely, Richard Tucker, and Angjoo Kanazawa. De-rendering the world's revolutionary artefacts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6338–6347, 2021.

[53] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9099–9109, 2023.

[54] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in neural information processing systems*, 35:38571–38584, 2022.

[55] 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, 2021.

[56] Xitong Yang, Devansh Kukreja, Don Pinkus, Anushka Sagar, Taosha Fan, Jinhyung Park, Soyong Shin, Jinkun Cao, Jiawei Liu, Nicolas Ugrinovic, Matt Feiszli, Jitendra Malik, Piotr Dollar, and Kris Kitani. Sam 3d body: Robust full-body human mesh recovery, 2026.

[57] Yuxiang Yang, Yingqi Deng, Yufei Xu, and Jing Zhang. Aptv2: Benchmarking animal pose estimation and tracking with a large-scale dataset and beyond. *arXiv preprint arXiv:2312.15612*, 2023.

[58] Yuxiang Yang, Junjie Yang, Yufei Xu, Jing Zhang, Long Lan, and Dacheng Tao. Apt-36k: A large-scale benchmark for animal pose estimation and tracking. *Advances in Neural Information Processing Systems*, 35:17301–17313, 2022.

[59] 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.

[60] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12287–12303, 2023.

[61] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11446–11456, 2021.

[62] Brian Nlong Zhao, Jiajun Wu, and Shangzhe Wu. Web-scale collection of video data for 4d animal reconstruction. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

[63] Silvia Zuffi and Michael J Black. Awol: Analysis without synthesis using language. In *European Conference on Computer Vision*, pages 1–19. Springer, 2024.

[64] 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*, pages 3955–3963, 2018.

[65] 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*, pages 6365–6373, 2017.

[66] Silvia Zuffi, Ylva Mellbin, Ci Li, Markus Hoeschle, Hedvig Kjellström, Senya Polikovsky, Elin Hernlund, and Michael J Black. Varen: Very accurate and realistic equine network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5374–5383, 2024.

A Bipartite Matching Details

We formulate the assignment between the predicted entities and ground-truth targets as a set prediction problem. Let $\hat{\mathcal{Y}} = \{\hat{y}_i\}_{i=1}^{N}$ denote the set of predictions, where each prediction comprises bounding box coordinates, a confidence score, and 2D keypoint locations. Let $\mathcal{Y} = \{y_j\}_{j=1}^{M}$ denote the set of ground-truth targets ($M \leq N$). Following DETR [12], the assignment is determined by searching for an optimal injection $\hat{\sigma} : \{1, \ldots, M\} \to \{1, \ldots, N\}$ that minimizes the overall bipartite matching cost:

$$\hat{\sigma} = \arg\min_{\sigma} \sum_{i=1}^{M} \mathcal{C}_{\text{match}}(\hat{y}_{\sigma(i)}, y_i) \tag{7}$$

where $\sigma(i) \in \{1, \ldots, N\}$ is the prediction index assigned to the i-th ground-truth target, and $\mathcal{C}_{\text{match}}(\hat{y}_{\sigma(i)}, y_i)$ is the pair-wise assignment cost between prediction $\sigma(i)$ and ground-truth target i. To ensure that the assigned predictions maintain high localization precision and structural pose alignment, our matching cost is a weighted composition of four distinct terms:

$$\mathcal{C}_{\text{match}} = \lambda_{\text{conf}}\mathcal{C}_{\text{conf}} + \lambda_{\text{bbox}}\mathcal{C}_{\text{bbox}} + \lambda_{\text{giou}}\mathcal{C}_{\text{giou}} + \lambda_{\text{kpts}}\mathcal{C}_{\text{kpts}} \tag{8}$$

Specifically, the weighting factors in the matching cost (Eq. 8) are set as follows: $\lambda_{\text{conf}} = 1$, $\lambda_{\text{bbox}} = 1$, $\lambda_{\text{giou}} = 1$, and $\lambda_{\text{kpts}} = 10$. The terms are defined as follows:

- **Bounding Box Cost ($\mathcal{C}_{\text{bbox}}$):** The $\mathcal{L}_1$ distance between the normalized center coordinates, width, and height of the predicted and ground-truth bounding boxes.
- **GIoU Cost ($\mathcal{C}_{\text{giou}}$):** The Generalized Intersection over Union (GIoU) cost [42], which provides a scale-invariant geometric alignment constraint for the bounding boxes.
- **Confidence Cost ($\mathcal{C}_{\text{conf}}$):** A focal-style penalty applied to the predicted confidence score $\hat{c}_i \in [0, 1]$. It is formulated as $\alpha(1 - \hat{c}_i)^{\gamma}(-\log \hat{c}_i)$ with focusing parameters $\alpha = 0.25$ and $\gamma = 2.0$ following [44], encouraging the model to match ground truth with high-confidence predictions.
- **Keypoint Cost ($\mathcal{C}_{\text{kpts}}$):** A pose-aware geometric constraint computed as the mean $\mathcal{L}_1$ distance between the predicted 2D keypoints and the corresponding ground-truth keypoints, strictly masked to account only for visible joints.

The optimal permutation $\hat{\sigma}$ is computed efficiently using the Hungarian algorithm [23]. Once the optimal assignment is established, the predicted outputs of multiple animals are reordered to perfectly align with the ground-truth sequence, allowing the subsequent calculation of the standard supervisory loss functions.

B More Details about Herd3D

We provide additional qualitative and statistical details of Herd3D in this section. Fig. 8 visualizes representative multi-animal examples with varying group sizes, where the number of visible animals gradually increases from two to eight. These examples illustrate the diversity of species, poses, viewpoints, background environment, and occlusion patterns covered by Herd3D, as well as the corresponding structural cues used for annotation and visualization, including edge maps, depth maps, RGB images, and pose overlays. Table 5 further summarizes the taxonomic composition of the training set by listing the top-10 animal families and representative species within each family. Together, these results show that Herd3D covers both dense multi-animal scenes and broad cross-species diversity, making it suitable for studying scene-centric animal pose understanding.

C Failure Cases in Herd3D Construction

During the construction of Herd3D, we observed several failure cases when rendering multi-animal scenes with Qwen-ControlNet. These failures are particularly common when animal meshes have large overlapping regions or severe inter-instance occlusions. In such cases, the rendered image may not faithfully preserve the intended 3D geometry and pose. For example, an animal whose head is oriented away from the camera may be incorrectly rendered with a forward-facing face, leading to inconsistent orientation between the mesh and the generated image. Local semantic errors may

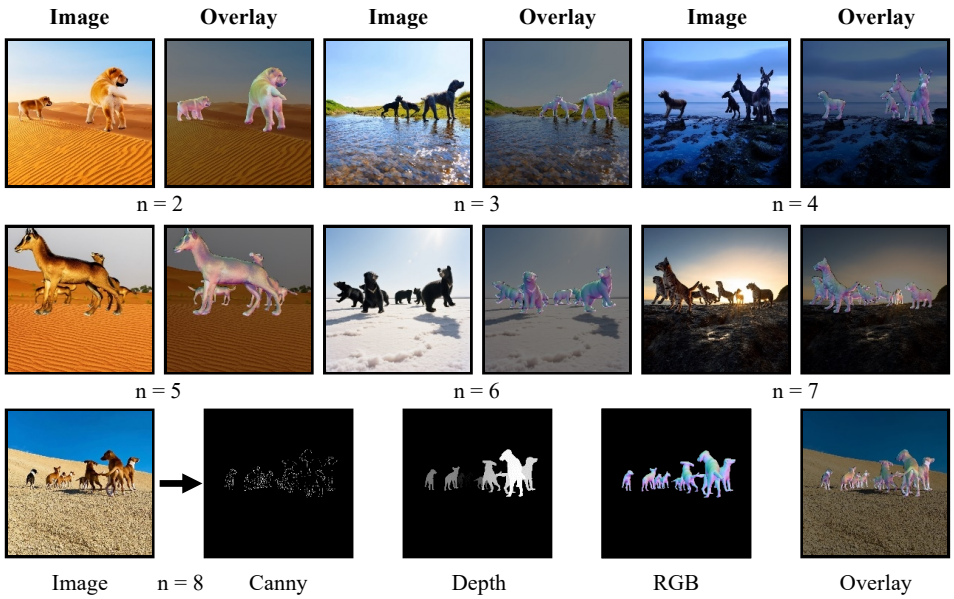


Figure 8: Herd3D multi-animal dataset. The images include dogs, horses, antelopes, bears, and cats, with the number of animals per image gradually increasing from 2 to 8 .



Figure 9: Failure cases of data generation.

also occur, where small body parts are misinterpreted, such as ears being rendered as noses or other facial structures. In addition, when two animals are spatially close, the renderer may blend their body regions, causing the torso or limbs of one animal to be partially rendered onto another. These artifacts indicate that dense multi-animal scenes remain challenging for image-conditioned generative rendering, especially under heavy occlusion and close physical interaction. We provide failure cases in Fig. 9.

Table 5: **Top-10 animal families in the Herd3D training image set.** Families are ranked by the number of labelled training images. For each family, we report representative species with their corresponding image counts.

Rank	Family	Images	Representative species
1	Canidae	3,238	Great Dane (105), Curly Coated Retriever (86), Irish Setter (85), Clumber Spaniel (84), Irish Wolfhound (84), Irish Water Spaniel (82), American Cocker Spaniel (73), American Eskimo Dog (71), Cairn Terrier (70), Alaskan Malamute (64)
2	Bovidae	324	Cow (106), Cattle (95), American Bison (21), Antelope (21), Goat (21), Bighorn Sheep (20), Buffalo (20), Impala (20)
3	Equidae	305	Horse (127), Feral Horse (115), Domestic Horse (21), Donkey (21), Common Zebra (21)
4	Mustelidae	187	Honey Badger (73), Eurasian Otter (54), African Clawless Otter (20), American Badger (20), American Mink (20)
5	Rhinocerotidae	157	Black Rhinoceros (157)
6	Ursidae	151	American Black Bear (90), Brown Bear (21), Da Xiong Mao (20), Giant Panda (20)
7	Felidae	143	Bay Lynx (22), European Wildcat (21), Bay Cat (20), Caracal (20), Cat (20), Cheetah (20), Cougar (20)
8	Giraffidae	63	Angolan Giraffe (21), Cape Giraffe (21), Giraffe (21)
9	Herpestidae	42	Gray Meerkat (42)
10	Viverridae	41	Bear Cat (21), African Civet (20)

Table 6: Forward-pass inference time under different numbers of animals on APTv2. The reported time is measured in milliseconds per image and excludes data loading, preprocessing, and metric computation.

Method	1 Animal	2 Animals	3 Animals	4 Animals	5+ Animals
3D Fauna	195.9 (195.9)	198.1 (396.2)	196.6 (589.9)	198.1 (792.6)	195.4 (1518.1)
GenZoo	7.8 (7.8)	7.8 (15.5)	8.0 (24.0)	8.0 (31.8)	7.8 (60.9)
AniMer	7.6 (7.6)	7.6 (15.1)	7.6 (22.7)	7.6 (30.3)	7.6 (59.5)
Ours	79.4	79.2	84.7	80.3	80.5

D More Ablation Experiments

Fig. 10 further demonstrates the effect on the number of prompt keypoints. With more keypoints used as prompt information, SAM 3D Animal produces results aligned better with input image.



Input Image | w/o kp | kp = 5 | kp = 10 | kp = 15

Figure 10: Ablation study on the number of prompt keypoints.

E Inference Efficiency

We additionally compare the inference efficiency of different methods on APTv2. We group test images by the number of annotated animals in the original image: 1, 2, 3, 4, and 5+ animals, containing 3,502, 1,384, 557, 166, and 426 images, respectively. Table 6 reports the average inference time for each group.

During measurement, we excluded data loading, preprocessing, and metric computation. All measurements were performed on a single NVIDIA RTX 4090 GPU. Note that, for 3D Fauna, GenZoo and AniMer, the input is cropped animal image resized to 256x256 resolution with padding. For SAM 3D Animal, the input is the whole uncropped image resized to 512x512 resolution with padding. Note that, although our method is currently slower than GenZoo and AniMer without any acceleration techniques, it maintains a stable time cost for different numbers of animals, thus having the potential to scale up for inferring a large number of animals efficiently.

F Robustness to Image Quality

We further study how input image quality affects reconstruction performance on Animal3D. We divide the test set into three approximately balanced groups using an object-centric image-quality score that equally combines percentile-normalized effective animal resolution and crop sharpness. The resulting groups are: *Low* ($[0, 0.3903]$, 107 images), *Medium* ($[0.3903, 0.6270]$, 106 images), and *High* ($[0.6301, 0.9483]$, 107 images). Table 7 reports PA-MPJPE, AP, and mAP for each group.

For direct comparison with the baselines, *Ours (w/ bbox)* follows the same evaluation protocol: each animal is cropped using its ground-truth bounding box and reconstructed independently. In contrast, *Ours (w/o GT prompt)* takes the full image as input and reconstructs the animal without ground-truth bounding boxes, keypoints, or masks.

Image quality primarily affects the 2D reconstruction metrics, while PA-MPJPE remains relatively stable. In the prompt-free setting, mAP increases from 35.9 on low-quality images to 59.7 on high-quality images. Bounding-box localization substantially improves robustness to degraded images: the low-to-high AP gap is reduced from 12.2 to 6.6 points, while the mAP gap decreases from 23.8 to 15.6 points. These results suggest that accurate localization can partially compensate for

Table 7: **Quantitative comparisons on the Animal3D dataset under different image-quality level.**

Method	Low			Medium			High		
	PA-MPJPE↓	AP↑	mAP↑	PA-MPJPE↓	AP↑	mAP↑	PA-MPJPE↓	AP↑	mAP↑
3D Fauna [30]	194.7	66.3	10.5	212.8	59.2	9.4	201.0	63.4	10.1
GenZoo [39]	124.8	80.4	29.8	118.6	88.7	35.5	124.3	85.0	37.3
AniMer [36]	**82.6**	86.3	40.8	78.1	98.5	48.8	82.3	95.5	47.2
Ours (w/o GT prompt)	84.3	85.0	35.9	79.5	96.2	52.5	78.4	97.2	59.7
Ours (w/ bbox)	83.2	**92.5**	**47.3**	**78.0**	**99.1**	**57.8**	**76.0**	**99.1**	**62.9**

Table 8: Effect of keypoint prompt locations on the 252-image common-visible subset of Animal3D. *Distal* uses the four fore- and hind-paw endpoints; *Intermediate* uses the four intermediate limb joints, approximately corresponding to the front elbows and rear knees/hocks; *Proximal* uses the four shoulder and hip keypoints; and *Random* samples four visible keypoints from all 26 Animal3D keypoints.

Prompt type	PA-MPJPE ↓	AP ↑	mAP ↑
No Prompt	77.6	95.6	51.4
Random	76.4	95.4	52.3
Distal	76.1	**96.4**	52.5
Intermediate	76.0	96.0	**54.0**
Proximal	**75.8**	95.6	53.4

limited visual detail, while the bbox-based setting maintains strong performance across all three image-quality levels.

G Effect of different types of keypoint prompts.

In Table 8, we conduct an additional controlled experiment on Animal3D to investigate whether the anatomical location of keypoint prompts affects reconstruction quality. To ensure a fair comparison, we select a common subset of 252 images in which all candidate keypoints are visible and provide exactly four keypoint prompts for each image. We compare Distal keypoints (the four fore- and hind-paw endpoints), Intermediate keypoints (the front elbows and rear knees/hocks), Proximal keypoints (the shoulder and hip keypoints), and Random keypoints (four visible keypoints sampled from all 26 Animal3D keypoints). The Random result is averaged over five seeds, and mask prompts are disabled to isolate the effect of keypoint location.

The results confirm that keypoint selection has a measurable effect. Anatomically selected prompts generally outperform random prompts. Proximal keypoints achieve the best PA-MPJPE of 75.8, suggesting that shoulder and hip landmarks provide stronger constraints for global 3D pose recovery. Intermediate keypoints achieve the best mAP of 54.0, indicating that articulated limb joints are more informative for fine-grained pose estimation. Distal keypoints achieve the best AP of 96.4. Therefore, distal keypoints are informative, as suggested by the reviewer, but they are not consistently optimal across all metrics. We include this controlled analysis and the corresponding discussion in the revised manuscript.

H Complete Inference Forward Pass

Algorithm 1 summarizes the prompt-conditioned forward pass using the indexing of Eqs. (1)–(4). Here, κ denotes image/camera metadata, $\hat{\Theta}$ collects SMAL+ parameters, and $\hat{C}$ denotes camera predictions.

Algorithm 1 SAM 3D Animal inference forward pass ($P = 30$, $L = 6$)

Require: Image $\boldsymbol{x}$, keypoint prompts $\mathcal{P}$, optional mask M, metadata κ, optional previous estimate η_{prev}
Ensure: P scored hypotheses with SMAL+ parameters, cameras, boxes, meshes, and keypoints

$\mathcal{F} \leftarrow \text{ImageFeatures}(\boldsymbol{x}, M, \kappa)$
$\eta_0 \leftarrow \text{LearnedInitialEstimate}(P)$
$\eta_{\text{prev}} \leftarrow \eta_0$ if no previous estimate is supplied
$(\mathcal{Q}^{(0)}_{\text{base}}, A^{(0)}_{\text{base}}) \leftarrow \text{InitBase}(\eta_0, \eta_{\text{prev}}, \kappa)$
$\mathcal{P} \leftarrow \text{PadMissingPrompts}(\mathcal{P}, c = -2)$
$\mathcal{Q}^{(0)}_{\text{prompt}} \leftarrow \text{Concat}_i\big(W_p\,\text{EncodePoint}(p_i) + b_p\big)$
$\mathcal{Q}^{(0)} \leftarrow \text{Concat}(\mathcal{Q}^{(0)}_{\text{base}}, \mathcal{Q}^{(0)}_{\text{prompt}})$ ▷ Inject prompts once
$A^{(0)} \leftarrow \text{Concat}(A^{(0)}_{\text{base}}, \mathcal{Q}^{(0)}_{\text{prompt}})$
for $l = 0, \ldots, L-1$ **do**
  $\mathcal{Q}^{(l+1)} \leftarrow \mathcal{D}_{l+1}(\mathcal{Q}^{(l)}, A^{(l)}, \mathcal{F})$ ▷ Eq. (2)
  $A^{(l+1)} \leftarrow A^{(l)}$
  $(\hat{\Theta}, \hat{C}, \hat{B}, \hat{s})^{(l+1)} \leftarrow \text{Predict}(\mathcal{Q}^{(l+1)}_{\text{params}}, \mathcal{Q}^{(l+1)}_{\text{box}}; \eta_0)$
  $(\mathbf{V}, \mathbf{X}_{\text{3D}})^{(l+1)} \leftarrow \mathcal{M}(\hat{\Theta}^{(l+1)})$
  $(\hat{x}_{\text{2D}}, d)^{(l+1)} \leftarrow \text{Project}(\mathbf{X}^{(l+1)}_{\text{3D}}, \hat{C}^{(l+1)}; \kappa)$
  if $l < L-1$ **then** ▷ Feedback only after the first five blocks
    $m^{(l+1)} \leftarrow \text{ValidProjection}(\hat{x}^{(l+1)}_{\text{2D}}, d^{(l+1)})$
    $\bar{x}^{(l+1)}_{\text{3D}} \leftarrow \text{CenterPelvis}(\mathbf{X}^{(l+1)}_{\text{3D}})$
    $f^{(l+1)} \leftarrow m^{(l+1)} \odot \text{BilinearSample}(\mathcal{F}, \hat{x}^{(l+1)}_{\text{2D}})$
    $\mathcal{Q}^{(l+1)}_{\text{2D}} \leftarrow \mathcal{Q}^{(l+1)}_{\text{2D}} + \phi_{\text{feat}}(f^{(l+1)})$
    $A^{(l+1)}_{\text{2D}} \leftarrow m^{(l+1)} \odot \phi_{\text{pos}}(\hat{x}^{(l+1)}_{\text{2D}})$ ▷ Eq. (3)
    $A^{(l+1)}_{\text{3D}} \leftarrow \psi_{\text{pos}}(\bar{x}^{(l+1)}_{\text{3D}})$ ▷ Eq. (4)
  end if
end for
return $\{\hat{\Theta}^{(L)}_j, \hat{C}^{(L)}_j, \hat{B}^{(L)}_j, \hat{s}^{(L)}_j, \mathbf{V}^{(L)}_j, \mathbf{X}^{(L)}_{\text{3D},j}, \hat{x}^{(L)}_{\text{2D},j}\}_{j=1}^{P}$

Initialization and prompting. ImageFeatures extracts ViT features with the configured mask and ray conditioning. InitBase uses learned parameters to construct the four base query groups and the previous-estimate conditioning tokens included in the parameter-related inputs. Their initial positional augmentations contain the previous-estimate embeddings, with zero augmentation for the box and keypoint groups. EncodePoint is the piecewise encoding defined in Section 3.2: valid keypoint IDs use position-plus-label embeddings, while labels -1 and -2 use the negative and invalid embeddings. Missing prompts are represented by invalid entries.

Predictions and feedback. Predict includes decoder-output normalization and the SMAL+, camera, box, and confidence heads. Project produces crop-image coordinates and depths; m selects in-frame projections with valid positive depth. The mask is applied before feature projection and after positional encoding, as in the implementation. Feedback uses the keypoint subsets associated with the 2D and 3D token groups, and preserves 3D token contents and all unlisted components. All hypotheses are returned with confidence scores; a configured confidence threshold can select output instances. During training, denoising queries, auxiliary supervision, and bipartite matching supplement this inference path as described in Section 3.3 and Appendix A.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the main contributions: SAM 3D Animal, a promptable multi-animal 3D reconstruction framework; Herd3D, a multi-animal 3D dataset with per-instance SMAL+ annotations; and empirical comparisons on Animal3D, APTv2, and Animal Kingdom. These claims are supported by the quantitative comparisons, ablation studies, and prompt analysis reported in the experiments.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: The conclusion discusses two main limitations: the method is constrained by the SMAL+ shape space and is therefore mainly applicable to quadruped-like animals, and relative depth ordering between animals is not explicitly constrained under severe occlusion.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not present theoretical results, theorems, or formal proofs. The mathematical formulations are used to describe the model architecture, bipartite matching, and loss functions.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper describes the model architecture, Herd3D generation process, training datasets, evaluation datasets, metrics, optimizer, learning rate, training schedule, prompt dropout strategy, loss weights, and compute setup. The supplementary material further provides details of the bipartite matching formulation and experiment results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The code will be made publicly available upon paper acceptance.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the training corpus, evaluation datasets, baselines, metrics, optimizer, initial learning rate, warmup schedule, training epochs, prompt dropout strategy, gradient accumulation, GPU setup, and loss weights. These details are sufficient to understand the reported experimental results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper reports error bars for the visibility-level analysis in Fig. 8, where the error bars denote standard deviation across visibility counts within each group. However, the main quantitative comparison and ablation tables do not report repeated-run error bars or statistical significance tests because running inference and evaluation is computationally expensive due to optimization.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper reports that training is distributed across four NVIDIA RTX 4090 GPUs with gradient accumulation in experiment section.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work focuses on 3D animal reconstruction from public and synthetic animal imagery. It does not involve human subjects, personally identifiable information, or safety-critical deployment.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Potential positive impacts include wildlife monitoring, animal behavior analysis, and digital content creation as mentioned in introduction, while potential negative impacts may include misuse for intrusive animal surveillance or inaccurate downstream analysis.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer:[N/A]

    Justification: The proposed model and dataset are not high-risk dual-use assets such as language models, biometric identification systems, or unsafe scraped datasets. The new dataset is generated through a controlled synthetic pipeline.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper cites the existing datasets, models, and methods used, including SMAL+, GenZoo, Animal3D, APTv2, Animal Kingdom, ViTPose, and Qwen-based components.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The paper introduces Herd3D and documents its generation process, scene layout constraints, animal placement strategy, pose sources, ControlNet backend, orientation prompting scheme, image resolution, and available annotations.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing experiments or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve human subjects, user studies, or crowdsourced human participant experiments, so IRB approval is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The paper describes the use of Qwen3-VL-8B-Instruct in the Herd3D generation pipeline to resolve multi-animal orientation ambiguity, and also describes the use of Qwen-Image-ControlNet-Union as the image generation backend. These uses are part of the dataset generation method rather than only writing or editing assistance.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.