Diffusion-based free-viewpoint synthesis for dataset expansion and wildlife classification

Jess Tam[1,2], **Hideo Saito**[1]
[1]Keio University
[2]UNSW Sydney
j.tam@unsw.edu.au, hs@keio.jp

Abstract

Processing wildlife imagery for conservation and management poses significant challenges, especially when limited data hinders the ability of classification models to extract sufficient features from each wildlife class. In this study, we propose the use of a 3D free-viewpoint image-to-video generative model to augment the dataset by synthesizing new images for fine-tuning classification models. Our results demonstrate a notable improvement in the F1-score, increasing from a baseline of 0.338 to 0.525. While the quality of the synthesized images can be further enhanced, particularly in terms of incorporating wildlife-specific semantics, the study highlights the potential of generative AI not only for media creation but also for advancing environmental monitoring applications, despite challenges such as high computational cost.

Code — https://github.com/jesstytam/dimensionX_for_wildlife_cls

Introduction

Creating automated wildlife monitoring systems is essential for effective conservation and management (Christin, Hervet, and Lecomte 2019; Nakagawa et al. 2023; Tuia et al. 2022; Fergus et al. 2024; Weinstein 2018). However, collecting large datasets for building such systems in the wild is challenging, especially for elusive species. To overcome the limited availability of wildlife imagery, modern computer vision approaches can offer novel solutions to synthesize new data that were previously not possible. These solutions include generative models, such as GANs (Goodfellow et al. 2014), VAEs (Yu et al. 2024), and diffusion models (Sohl-Dickstein et al. 2015).

GANs work by training a generator to generate synthetic samples to fool a discriminator from distinguishing them from real samples. However, GANs can suffer from mode collapse if a sub-optimal loss function is implemented, causing the discriminator to become stuck at a local minimum without converging during training. As a result, synthesized imagery may lack diversity, where the images appear similar or of the same class (Kossale, Airaj, and Darouichi 2022; Thanh-Tung and Tran 2020; Saatchi and Wilson 2017).

VAEs consist of an encoder that extracts features into a continuous latent space as a Gaussian prior, and a decoder then reconstructs samples by learning from the underlying probabilistic distribution from the latent space. However, as features are compressed into a low-dimensional latent space that follows a simple Gaussian distribution, fine details of the features are often lost, reducing the realism of the synthesized imagery (Yacoby, Pan, and Doshi-Velez 2020).

On the other hand, diffusion models learn through a process of gradually adding and then removing Gaussian noise from images. By iteratively denoising random noise back into realistic images, they can capture fine-grained visual details and complex data distributions, producing higher quality and more diverse samples without relying on compact latent representations that are used in VAEs. In addition, diffusion models are more stable to train than GANs, as they do not depend on an adversarial objective and are less susceptible to mode collapse, resulting in more reliable convergence and consistent image quality.

In this study, we demonstrate a novel pipeline that incorporates free-viewpoint video frames generated from images using the 3D free-viewpoint image-to-video model DimensionX (Sun et al. 2025). We show that we can increase the training dataset size using synthetic multi-view sequences, which enhances classification performance. This approach highlights the potential of generative models to address data limitations in ecological applications and support more robust automated wildlife monitoring. To summarise, our contributions are as follows:

- We expand our training dataset by generating multi-view imagery using a 3D view synthesis diffusion model, producing novel geometric perspectives of each individual from multiple angles to enhance viewpoint diversity.
- We apply and evaluate this generative approach within a wildlife classification pipeline.

Related work

Single-image to 3D scene generation

Reconstructing a 3D scene from a single view is a challenging problem, as it lacks multiple viewpoints to infer depth, geometry, or camera pose. Recent progress in 3D scene generation and reconstruction has mainly been driven

by diffusion-based approaches. Scene-level diffusion models go beyond reconstructing isolated objects and instead synthesize entire environments in which foreground subjects interact with the background and environment. These approaches can broadly be categorised into single-stage and two-stage pipelines. Single-stage methods (Gao* et al. 2024; Yu et al. 2025) train a unified diffusion model that directly maps a single image to a 4D representation, jointly learning spatial and temporal consistency. In contrast, two-stage methods (Sun et al. 2025; Sargent et al. 2024; Zhao et al. 2024) first generate intermediate 2D or 3D features before refining them into coherent spatio-temporal scenes.

Video-based diffusion frameworks offer a way to unify spatial and temporal reasoning. By learning in 3D or 4D latent spaces, they produce temporally smooth and spatially realistic view transitions without relying on explicit geometry. Treating view synthesis as a controllable video diffusion task enables the capturing of fine-grained features across frames, ensuring consistent appearance and structure across viewpoints (Ma et al. 2025; Wen et al. 2025; Li et al. 2024).

DimensionX (Sun et al. 2025) builds on this approach by framing single-image 3D scene generation as a controllable video diffusion process. Its core innovation is the ST-Directors, which decouples spatial and temporal control through two LoRA (Hu et al. 2022) modules: the *S-Director* for spatial viewpoint control and the *T-Director* for temporal motion control. The S-Director is trained on spatially varying datasets where camera parameters change while the 3D scene remains static ($S(t) = S_0$), enabling controlled camera motion and novel viewpoint generation. Conversely, the T-Director is trained on temporally varying datasets where camera parameters are fixed ($C(t) = C_0$), allowing realistic object motion control.

For our classification task, we employ only the S-Director to generate spatially variant video sequences, with the animal and background held static. This allows the apparent camera trajectory to vary while maintaining the underlying scene structure, producing realistic multi-view sequences that improve viewpoint diversity for training and evaluation.

Methods

Dataset description

The dataset was collected by NSW National Parks and Wildlife Services (NPWS) within the state of New South Wales (NSW) in Australia. The images were collected using motion-sensitive cameras placed within multiple national parks. The original dataset contained 41,022 images in total across 24 classes of wildlife, including both non-native and native Australian wildlife. Training and testing sets were created by stratifying the dataset temporally by 80:20 .

For this study, we sample a subset from the training dataset as video generation is computationally expensive. We select 14 classes of mammals with the most images available, and then randomly sample 300 images from each class to build the baseline training set.

Free-viewpoint video generation

To increase the size of our dataset with multi-view sequences, we employ the S-Director from DimensionX (Sun et al. 2025), a controllable extension of CogVideoX (Yang et al. 2024) that enables free-viewpoint video generation. This approach renders the animal(s) in each image from multiple viewpoints while preserving object structure and scene coherence.

Architecture overview (Fig. 1). Built on CogVideoX, a text encoder and a 3D causal VAE map the text prompt and image into latent spaces, represented by z_{text} and z_{vision}, respectively. The latents are then processed by the a transformer-based video diffusion model (DiT; (Peebles and Xie 2022)) backbone to learn 2D and diffusion priors, which denoises and decodes the latents into coherent video frames. To achieve spatial control, the S-Director LoRA module is injected into the attention layers of the DiT backbone.

Viewpoint projection. The S-Director adjusts the attention weights to guide camera trajectories C_t that vary spatial viewpoint while holding time constant. The rendering process can be described as the projection of a 4D scene $S(T)$ onto the 2D image plane:

$$I_t(u, v) = P_{C_t}\big(S(T)\big), \tag{1}$$

where P_{C_t} represents the projection operator under the current camera pose.

Video generation and training set augmentation. From a single-view image, a still multi-view video sequence is created. The videos generated using the S-Director orbit the camera leftwards around the centre of each image at a mean of 103 degrees, as estimated with VGGT (Wang et al. 2025) (see Appendix for details). For each input image, we generate a 4-second, 12 frames-per-second video, i.e., 49 total frames per video. See Appendix for examples of video sequences.

Each generated frame represents a novel camera viewpoint of the same static subject, effectively sampling local variations in appearance and geometry. This expands the diversity of training data available for fine-tuning without requiring new image collection in the field.

To examine if text conditioning can influence the quality of synthetic outputs, we designed a short and a long prompt. The short prompt provides a concise description of each image ("A camera trap image taken in southeastern Australia."), while the long prompt, with the addition of a negative prompt, adds explicit object and scene-level instructions ("A realistic 3D label, clearly visible and unobstructed…") to increase the number of conditioning tokens interacting with the visual latent through cross-attention layers (Hertz et al. 2022). Details of prompts are shown in Table 1.

Then, we run MegaDetector v6 (Beery, Morris, and Yang 2019) with a confidence threshold of 0.8 on all frames to draw bounding box. Frames without bounding box predictions are discarded to ensure that only frames containing a clearly visible animal are retained. We then use the box coordinates for segmentation with Segment Anything Model 2 (SAM2; (Ravi et al. 2025)) using the default thresholds (predicted IOU threshold of 0.88 and stability score threshold

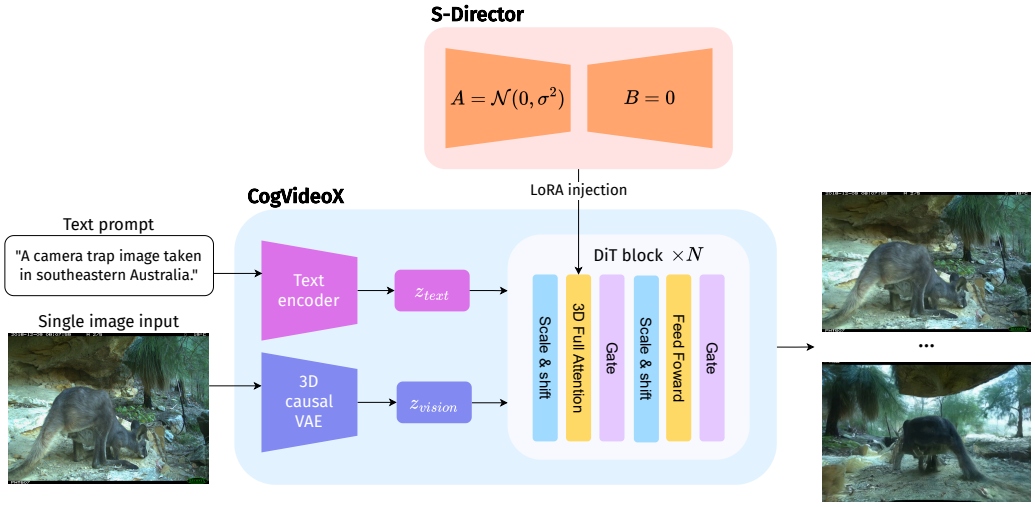

Figure 1: Overview of model architecture used in our pipeline, adapting CogVideoX with the DiT backbone and the S-Director LoRA module injection from DimensionX.

of 0.95) to retain high-confidence segmentations and create masked images to increase the size of the training datasets. As the synthetic frames lack ground truth, the thresholds we set serve as a quality control step to minimise error propagation, rather than a measure of absolute accuracy.

Experiments

We devise four experiments to generate new training data to test the effectiveness of using free-viewpoint data for classification. An overview of how our training datasets are created is shown in Fig. 2, while the prompts and configurations for video generation are shown in Table 1.

Baseline (Experiment 1) - no generated data. For our baseline experiment, no synthesized data was included. Directly using the raw images, we used MegaDetector to draw bounding boxes. Using the bounding boxes as prompts, we created segmented masks using SAM2. Finally, we fine-tuned the classifier using the masked images.

Experiment 2 - Generated from raw images + short prompt. Directly from the raw images, we generate one video per image using a short text prompt. In low-quality or low-contrast inputs, we observed that the generated wildlife subjects gradually blended into the background or became partially occluded during the latter half of the sequence (Appendix S4, S5). To minimise the inclusion of distorted frames and reduce redundancy, we retained only the first two seconds of each video, sampling one frame every 0.5 seconds (four frames per video). We then process the frames with MegaDetector and SAM2 to create masked images to increase the training set size.

Experiment 3 - Generated from SAM2 masked images + short prompt. We first apply SAM2 to the raw images to remove the background. The masked images are then used as input for video generation using a short prompt. From each video, all 49 frames are extracted and processed with MegaDetector to create bounding boxes to use in SAM2 to create final training data.

Experiment 4 - Generated from SAM2 masked images + long prompt. Same as that of Experiment 3, with the exception of using a long prompt and a negative prompt for video generation.

Model fine-tuning

We included identical geometric and photometric augmentations across all experiments to serve as our traditional data augmentation baseline. These transformations include horizontal and vertical flips, random rotations, colour jitter, Gaussian blur, and grayscale conversion. By applying the same augmentation pipeline to both the baseline and diffusion-augmented datasets, we isolate the additional contribution of the generative views from standard 2D image-space transformations.

All models are fine-tuned using ResNet-50 as the backbone and optimised with the Adam optimiser. Training is performed on a single NVIDIA RTX A6000 GPU for 20 epochs, with a batch size of 128 and 16 CPU workers. We use a learning rate of 0.0001 with a step scheduler (step size of 5 epochs, decay factor of 0.1), and employ cross-entropy loss for all experiments. We also perform 5-fold cross-validation to prevent overfitting.

Evaluation metrics

We evaluate model performance using three common classification metrics: precision, recall, and F1-score, for each class and the overall dataset.

Precision (p) measures the proportion of correctly predicted positive samples (true positives, TP) among all samples predicted as positive (TP and false positives, FP):

$$p = \frac{TP}{TP + FP} \tag{2}$$

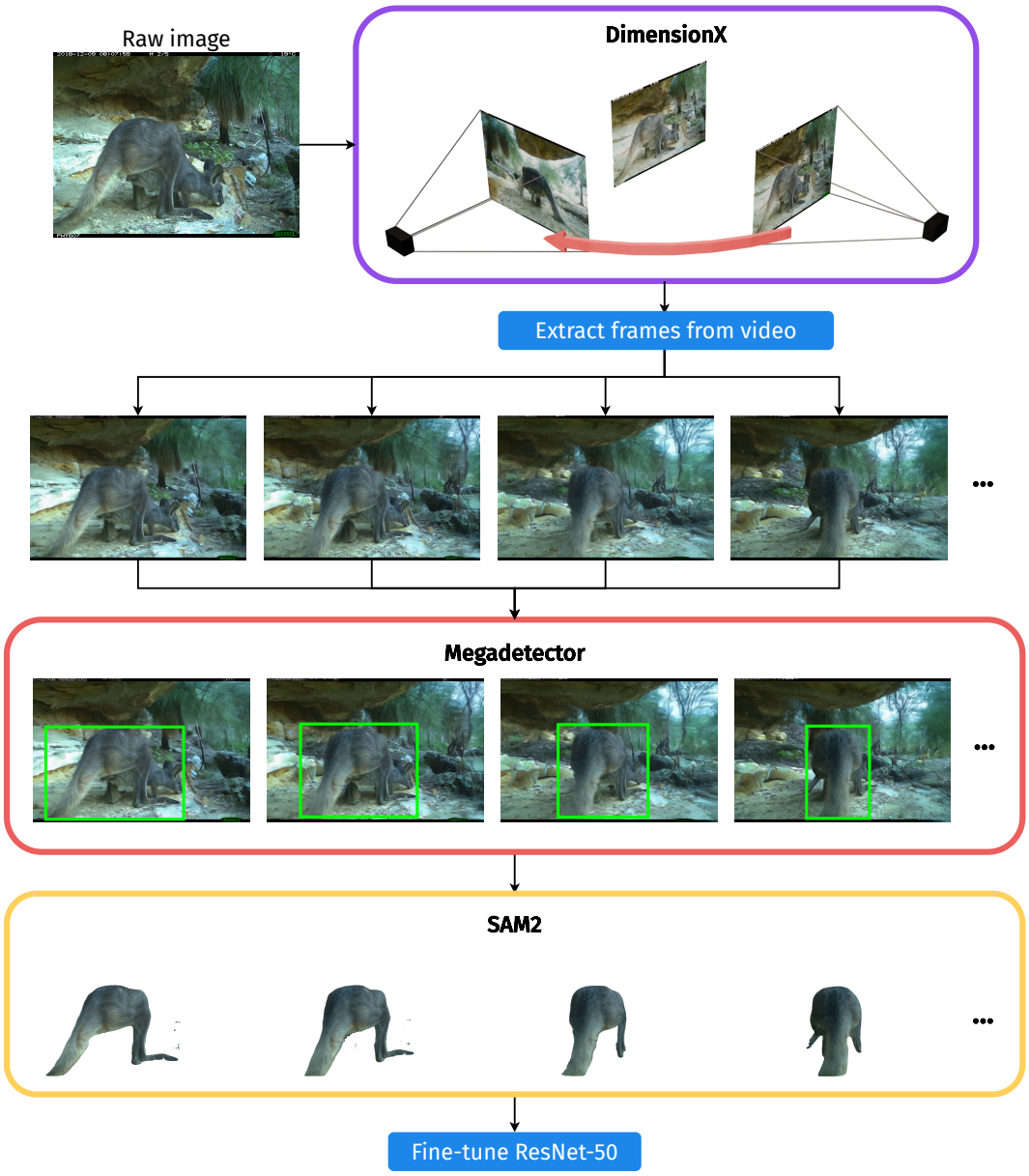

Figure 2: Pipeline illustrating how we created our training dataset.

	Baseline	Experiment 2	Experiment 3	Experiment 4
Generate videos	No	Yes	Yes	Yes
Videos generated from	/	Raw images	SAM2 masked images	SAM2 masked images
Prompt	/	"A camera trap image taken in southeastern Australia."	"A camera trap image taken in southeastern Australia."	"A realistic 3D {label}, clearly visible and unobstructed in the center of the frame. The camera smoothly orbits around the animal without anything blocking the view. The scene is a dry forest with a clean, open view." **Negative prompt:** "whiteout, overexposed, washed out, glowing, unrealistic lighting, blur, occlusion, other animals, trees in front, grass covering, deformities"
Frames extracted	/	4 frames (every 0.5 seconds of the first 2 seconds) (frames without detections were discarded)	All frames with bounding box predictions from MegaDetector (frames without detections were discarded)	All frames with bounding box predictions from MegaDetector (frames without detections were discarded)

Table 1: Summary of experimental configurations for video generation.

Recall (r) measures the proportion of correctly predicted positive samples among all actual positive samples (TP and false negatives, FN):

$$r = \frac{TP}{TP + FN} \tag{3}$$

The F1 score combines precision and recall into a single harmonic mean, ensuring that both metrics are balanced:

$$F1 = 2\frac{pr}{p + r} \tag{4}$$

Results

Evaluation metrics

Across all metrics - precision, recall, and F1-score, there was an overall improvement over the baseline results (Fig. 3), where the precision of baseline increased from 0.393 to 0.478, 0.535, and 0.540, recall increased from 0.367 to 0.449, 0.559, and 0.548, and F1 increased from 0.338 to 0.426, 0.525, and 0.515 respectively.

However, when comparing the results between the two models that also used video frames generated from SAM2 masked images, the model with a shorter prompt demonstrated slightly better performance than the model with a more detailed prompt. While precision increased from 0.535 to 0.540, recall dropped from 0.559 to 0.548, and F1 dropped from 0.525 to 0.515.

Confusion matrices

Echoing the metrics, normalised confusion matrices show a gradual improvement in the diagonals with the inclusion of video frames over the baseline (Fig. 4, 5). However, when comparing the matrices, we observed frequent misclassifications between the Brown Bandicoot and the Long-nosed Bandicoot, as seen in the results from Experiment 4. Example video frames from both species are provided in the Appendix to illustrate the similarity and distortions that may have contributed to these errors.

Conclusion and Discussion

We demonstrate that 3D view synthesis can effectively expand limited training datasets by generating novel viewpoints of wildlife for downstream tasks such as classification. While view synthesis has been explored as a data augmentation strategy in other domains (Zhou et al. 2023; Ma et al. 2024), it remains a developing research direction for environmental monitoring. Research in generative and diffusion models are rapidly expanding. Here, we demonstrate the use of one model for such pipeline. However, the comparison against other generative models remains outside the scope of this study.

Variables affecting classification performance

When comparing Experiments 3 and 4, while overall results appeared similar, several classes performed better in Experiment 3, where a shorter text prompt was used. This suggests

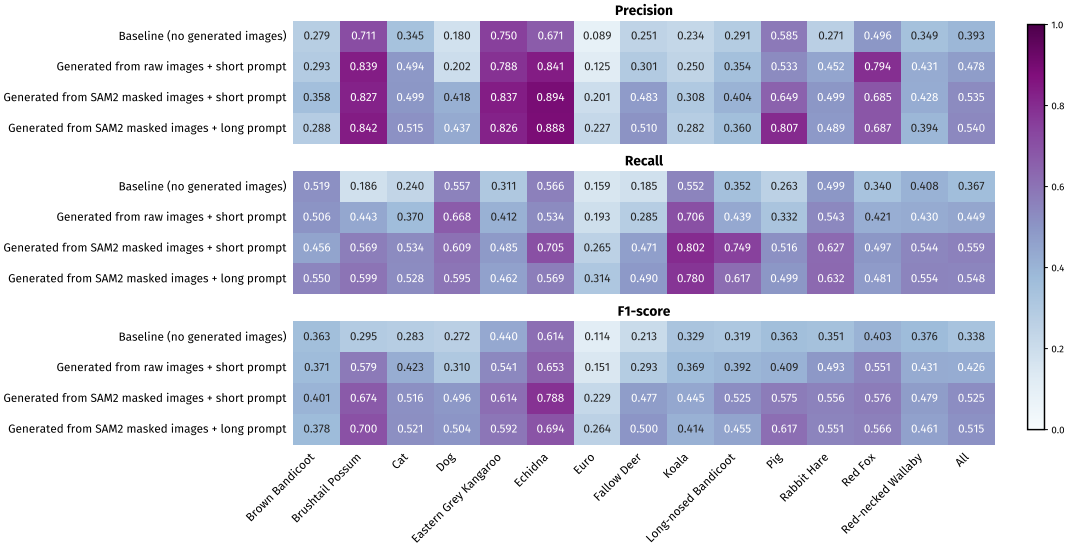

Figure 3: Precision, recall, and F1-score of all four experiments.

Figure 4: Confusion matrix of baseline (Experiment 1) results, showing normalised values

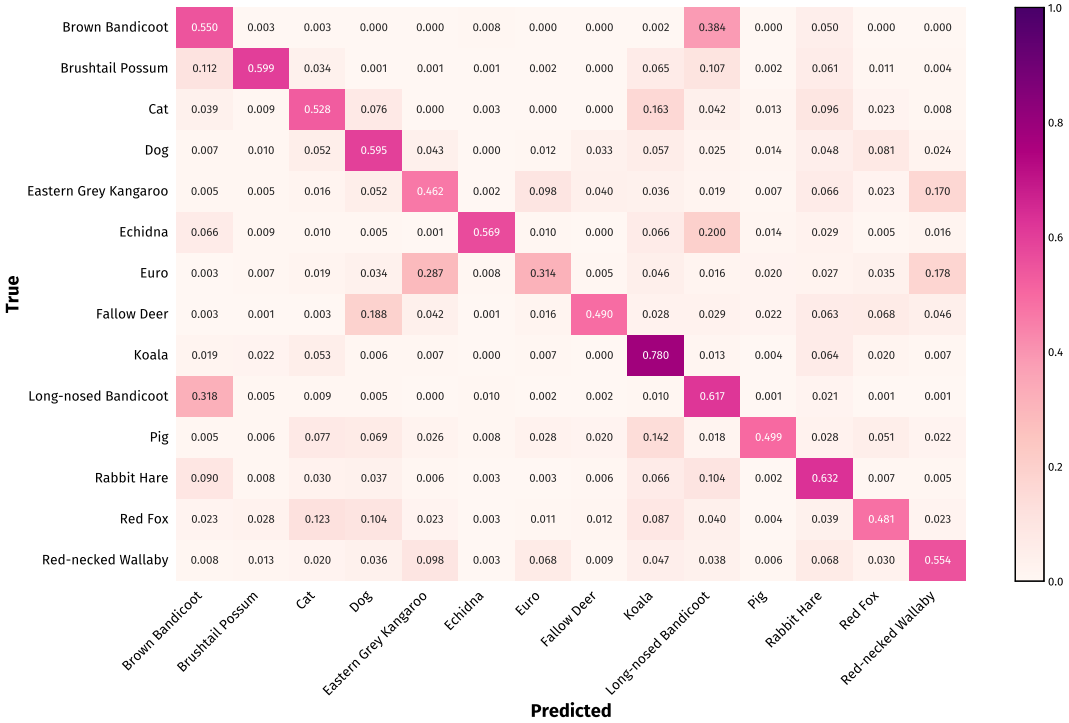

Figure 5: Confusion matrix of results after including free-viewpoint video frames generated from SAM2 images using a long prompt and a negative prompt (Experiment 4), showing normalised values

that longer prompts may have introduced irrelevant contextual information, leading to reduced accuracy, an effect also reported in large language and multi-modal models (Levy, Jacoby, and Goldberg 2024; Shi et al. 2023). In contrast, the comparison between the results in Experiments 2 and 3 show that using segmented inputs is more valuable in reducing background noise, and thus producing more consistent multi-view representations for better classification.

Morphological characteristics of each species also had a strong influence on classification performance. In particular, misclassifications were frequent between Brown and Long-nosed Bandicoots (Fig. 5). As they have similar physical characteristics, overlapping morphologies may create difficulties to distinguish, especially with their small body size, and where diagnostic features only cover small regions of the body that are easily occluded in camera trap images. Further, geometric distortions in the synthetic data could contribute to confusion between the two classes, due to the model backbone's limited semantic understanding of wildlife anatomy. Further explanations on such failure cases are outlined in the Appendix.

Semantics limitations affecting video quality

The lack of wildlife-specific semantic understanding could lead to limited quality in some synthesized data. This limitation manifests as unrealistic limb configurations and blurred fur textures. Nevertheless, several strategies could improve the semantic understanding of diffusion models for wildlife imagery, such as bettering visual understanding with VLMs (Choi et al. 2025) and incorporating 3D priors that encode body geometry (Chan et al. 2023), including using museum or field specimens.

Utilities of generative AI for wildlife monitoring

As diffusion-based models continue to innovate (Ma et al. 2025; Wen et al. 2025; Li et al. 2024; Ahsan et al. 2025), they present new opportunities for wildlife research beyond classification (Rafiq et al. 2025). In addition to increasing training set size as shown here and other domains (Zhou et al. 2023; Ma et al. 2024; Chen et al. 2022), they can enable the 3D reconstruction of animals in higher definition with more realistic simulation of coat patterns and lighting. Beyond classification, with improved wildlife-specific semantics, these methods could support other ecological applications, such as animal pose estimation and behaviour analysis.

Limitations and future directions

While diffusion-based generative models can produce higher quality imagery compared to those from GANs or VAEs, they remain computationally expensive and time-consuming to run. Future work can focus on improving the efficiency of such models, such as through model distillation (Qin et al. 2025) or implementing lightweight architectures (Shen et al. 2025), to ensure accessibility and efficiency. Another limitation is that synthetic data generated directly from raw data often do not have ground-truth labels. As such, detector and segmentation outputs may con-

tain uncertainties that can be challenging to be quantitatively evaluated. Incorporating confidence weighting by both human and model could improve robustness of models used for downstream tasks (Yánez et al. 2024). In addition, as a proof of concept, we only demonstrated the use of a single generative framework, rather than performing an exhaustive ablation across multiple generative models. Nevertheless, direct comparisons with other approaches, such as NeRF-based reconstruction (Mildenhall et al. 2021), can provide further insights into the benefits and trade-offs between realism and controllability of different approaches. Overall, our study demonstrates that diffusion-based view synthesis offers a scalable pathway for improving wildlife image analysis under data scarcity.

Acknowledgments

We thank the reviewers for their valuable feedback that helped to further enhance the quality of this paper. This research was enabled by the Research Training Program scholarship from the Australian Government and the Baxter Family scholarship.

References

Ahsan, M. M.; Raman, S.; Liu, Y.; and Siddique, Z. 2025. A comprehensive survey on diffusion models and their applications. *Applied Soft Computing*, 181(113470).

Beery, S.; Morris, D.; and Yang, S. 2019. Efficient Pipeline for Camera Trap Image Review. arXiv.

Chan, E. R.; Nagano, K.; Chan, M. A.; Bergman, A. W.; Park, J. J.; Levy, A.; Aittala, M.; De Mello, S.; Karras, T.; and Wetzstein, G. 2023. Generative novel view synthesis with 3d-aware diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4217–4229.

Chen, Y.; Yang, X.-H.; Wei, Z.; Heidari, A. A.; Zheng, N.; Li, Z.; Chen, H.; Hu, H.; Zhou, Q.; and Guan, Q. 2022. Generative adversarial networks in medical image augmentation: a review. *Computers in Biology and Medicine*, 144: 105382.

Choi, D.; Son, G.; Kim, S. Y.; Paik, G.; and Hong, S. 2025. Improving Fine-grained Visual Understanding in VLMs through Text-Only Training. In *Proceedings of the AAAI 2025 Workshop on Vision-Language Models*. Palo Alto, CA: Association for the Advancement of Artificial Intelligence.

Christin, S.; Hervet, É.; and Lecomte, N. 2019. Applications for deep learning in ecology. *Methods in Ecology and Evolution*, 10(10): 1632–1644.

Fergus, P.; Chalmers, C.; Longmore, S.; and Wich, S. 2024. Harnessing artificial intelligence for wildlife conservation. *Conservation*, 4(4): 685–702.

Gao*, R.; Holynski*, A.; Henzler, P.; Brussee, A.; Martin-Brualla, R.; Srinivasan, P. P.; Barron, J. T.; and Poole*, B. 2024. CAT3D: Create Anything in 3D with Multi-View Diffusion Models. In *Advances in Neural Information Processing Systems*.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-Prompt Image Editing with Cross Attention Control. *arXiv preprint arXiv:2208.01626*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Kossale, Y.; Airaj, M.; and Darouichi, A. 2022. Mode collapse in generative adversarial networks: An overview. In *2022 8th International Conference on Optimization and Applications (ICOA)*, 1–6. IEEE.

Levy, M.; Jacoby, A.; and Goldberg, Y. 2024. Same task, more tokens: the impact of input length on the reasoning performance of large language models. *arXiv preprint arXiv:2402.14848*.

Li, X.; Zhang, Q.; Kang, D.; Cheng, W.; Gao, Y.; Zhang, J.; Liang, Z.; Liao, J.; Cao, Y.-P.; and Shan, Y. 2024. Advances in 3d generation: A survey. *arXiv preprint arXiv:2401.17807*.

Ma, R.; Ma, T.; Guo, D.; and He, S. 2024. Novel view synthesis and dataset augmentation for hyperspectral data using NeRF. *IEEE Access*, 12: 45331–45341.

Ma, Y.; Feng, K.; Hu, Z.; Wang, X.; Wang, Y.; Zheng, M.; He, X.; Zhu, C.; Liu, H.; He, Y.; et al. 2025. Controllable video generation: A survey. *arXiv preprint arXiv:2507.16869*.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Nakagawa, S.; Lagisz, M.; Francis, R.; Tam, J.; Li, X.; Elphinstone, A.; Jordan, N. R.; O'Brien, J. K.; Pitcher, B. J.; Van Sluys, M.; et al. 2023. Rapid literature mapping on the recent use of machine learning for wildlife imagery. *Peer Community Journal*, 3.

Peebles, W. S.; and Xie, S. 2022. Scalable diffusion models with transformers. 2023 IEEE. In *CVF International Conference on Computer Vision (ICCV)*, volume 4172.

Qin, H.; Chen, L.; Kong, M.; Lu, M.; and Zhu, Q. 2025. Distilling multi-view diffusion models into 3d generators. *arXiv preprint arXiv:2504.00457*.

Rafiq, K.; Beery, S.; Palmer, M. S.; Harchaoui, Z.; and Abrahms, B. 2025. Generative AI as a tool to accelerate the field of ecology. *Nature Ecology & Evolution*, 9(3): 378–385.

Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; Mintun, E.; Pan, J.; Alwala, K. V.; Carion, N.; Wu, C.-Y.; Girshick, R.; Dollár, P.; and Feichtenhofer, C. 2025. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations (ICLR)*.

Saatchi, Y.; and Wilson, A. G. 2017. Bayesian GAN. In *Advances in Neural Information Processing Systems 30 (NIPS)*.

Sargent, K.; Li, Z.; Shah, T.; Herrmann, C.; Yu, H.-X.; Zhang, Y.; Chan, E. R.; Lagun, D.; Fei-Fei, L.; Sun, D.; and Wu, J. 2024. ZeroNVS: Zero-Shot 360-Degree View Synthesis from a Single Real Image. In *Computer Vision and Pattern Recognition (CVPR)*.

Shen, H.; Zhang, J.; Xiong, B.; Hu, R.; Chen, S.; Wan, Z.; Wang, X.; Zhang, Y.; Gong, Z.; Bao, G.; et al. 2025. Efficient Diffusion Models: A Survey. *Transactions on Machine Learning Research (TMLR)*.

Shi, F.; Chen, X.; Misra, K.; Scales, N.; Dohan, D.; Chi, E. H.; Schärli, N.; and Zhou, D. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, 31210–31227. PMLR.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In Bach, F.; and Blei, D., eds., *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, 2256–2265. Lille, France: PMLR.

Sun, W.; Chen, S.; Liu, F.; Chen, Z.; Duan, Y.; Zhang, J.; and Wang, Y. 2025. Dimensionx: Create any 3d and 4d scenes from a single image with controllable video diffusion. In *CVF International Conference on Computer Vision (ICCV)*.

Thanh-Tung, H.; and Tran, T. 2020. Catastrophic forgetting and mode collapse in GANs. In *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–10. IEEE.

Tuia, D.; Kellenberger, B.; Beery, S.; Costelloe, B. R.; Zuffi, S.; Risse, B.; Mathis, A.; Mathis, M. W.; Van Langevelde, F.; Burghardt, T.; et al. 2022. Perspectives in machine learning for wildlife conservation. *Nature communications*, 13(1): 792.

Wang, J.; Chen, M.; Karaev, N.; Vedaldi, A.; Rupprecht, C.; and Novotny, D. 2025. VGGT: Visual Geometry Grounded Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Weinstein, B. G. 2018. A computer vision for animal ecology. *Journal of Animal Ecology*, 87(3): 533–545.

Wen, B.; Xie, H.; Chen, Z.; Hong, F.; and Liu, Z. 2025. 3D Scene Generation: A Survey. *arXiv preprint arXiv:2505.05474*.

Yacoby, Y.; Pan, W.; and Doshi-Velez, F. 2020. Failure modes of variational autoencoders and their effects on downstream tasks. In *International Conference on Learning Representations*.

Yánez, F.; Luo, X.; Minero, O. V.; and Love, B. C. 2024. Confidence-weighted integration of human and machine judgments for superior decision-making. *arXiv preprint arXiv:2408.08083*.

Yang, Z.; Teng, J.; Zheng, W.; Ding, M.; Huang, S.; Xu, J.; Yang, Y.; Hong, W.; Zhang, X.; Feng, G.; et al. 2024. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. *arXiv preprint arXiv:2408.06072*.

Yu, L.; Lezama, J.; Gundavarapu, N. B.; Versari, L.; Sohn, K.; Minnen, D.; Cheng, Y.; Birodkar, V.; Gupta, A.; Gu, X.; Hauptmann, A. G.; Gong, B.; Yang, M.-H.; Essa, I.; Ross, D. A.; and Jiang, L. 2024. Language Model Beats Diffusion–Tokenizer is Key to Visual Generation. In *International Conference on Learning Representations*.

Yu, W.; Xing, J.; Yuan, L.; Hu, W.; Li, X.; Huang, Z.; Gao, X.; Wong, T.-T.; Shan, Y.; and Tian, Y. 2025. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhao, Y.; Lin, C.-C.; Lin, K.; Yan, Z.; Li, L.; Yang, Z.; Wang, J.; Lee, G. H.; and Wang, L. 2024. Genxd: Generating any 3d and 4d scenes. In *International Conference on Learning Representations*.

Zhou, A.; Kim, M. J.; Wang, L.; Florence, P.; and Finn, C. 2023. Nerf in the palm of your hand: Corrective augmentation for robotics via novel-view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17907–17917.

Appendix

Camera orbit angle calculation

We estimated the orbit angles of each synthetic video of Experiments 2 to 4 by estimating the camera poses with VGGT (Visual Geometry Grounded Transformer) (Wang et al. 2025). As we observed some outlying values, we report results both with and without these outliers. The standard deviations of all orbit angles decreased substantially after removing the outliers. Outliers were identified using the interquartile range (IQR) method, where the lower bound is defined as $Q1 - (1.5 \times IQR)$ and upper bound as $Q3 + (1.5 \times IQR)$. The distributions of results are illustrated in Fig. S1 and S2, with the mean values summarised in Tables S1 and S2, grouped by experiment and class.

Training data distribution

We conducted four experiments in this study, with each having a different set of images. The final size of each training set are illustrated in Fig. S3.

Video sequence examples

Fig. S4 - S10 show examples of still video sequences we generated with DimensionX for Experiments 2 to 4, showing the camera orbiting towards the left.

In these experiments, we observed several failure cases occurring especially common in the second half of the videos, such as incorrect animal geometry (fused or disappearing body parts), degradation of fur textures, or animals blending into the background or environment.

In particular, for the Brown and Long-nosed Bandicoots, (Fig. S4, S8, S9), due to their similar anatomy, in addition to their small size, they often appear smaller in camera trap images compared to other wildlife. This means that distortions were not uncommon, which may have lead to the frequent misclassifications between them, as outlined in the main text.

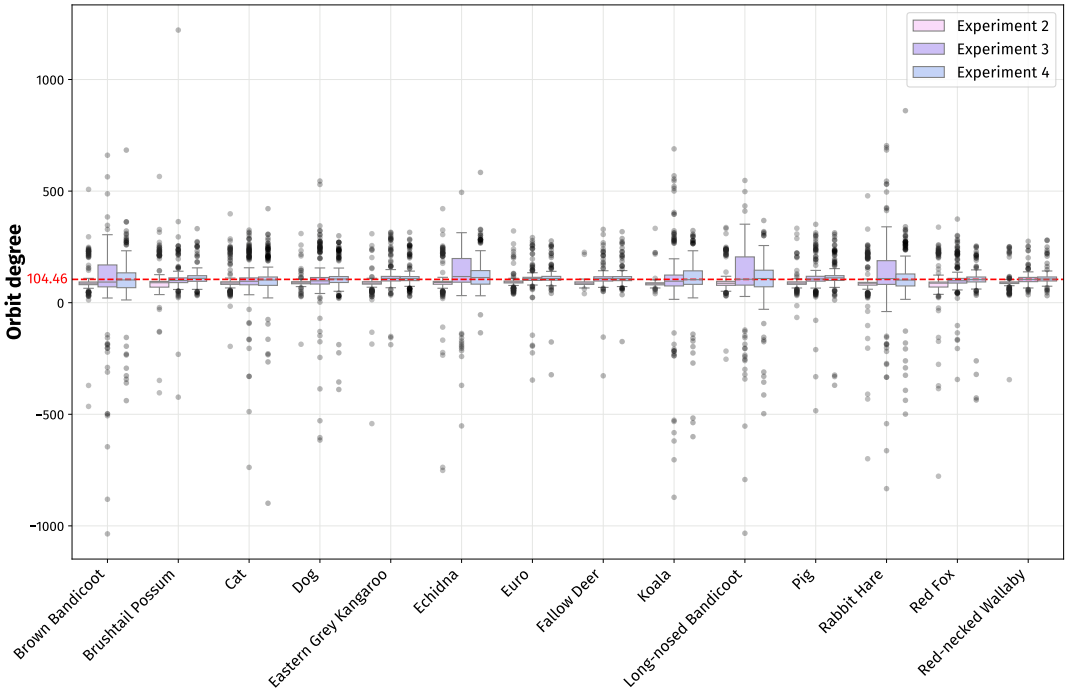

Figure S1: The distribution of orbit angles of all VGGT estimations. Outliers are illustrated by black circles. The mean orbit angle is 104.46 degrees.

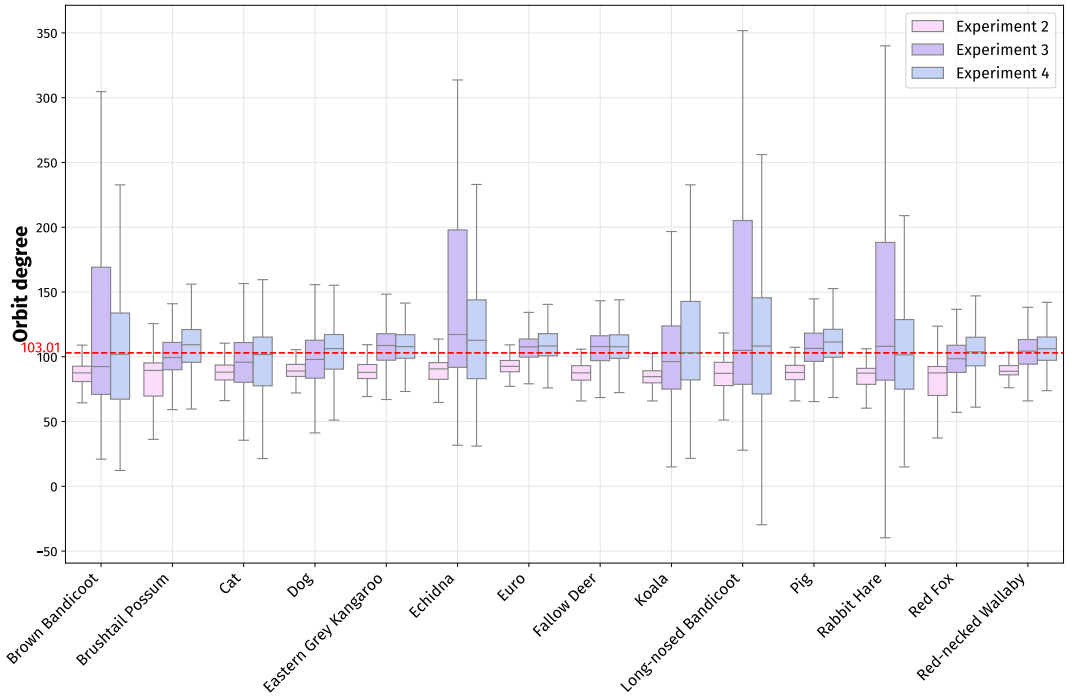

Figure S2: The distribution of orbit angles of VGGT estimations after removing outliers. The mean orbit angle is 103.01 degrees.

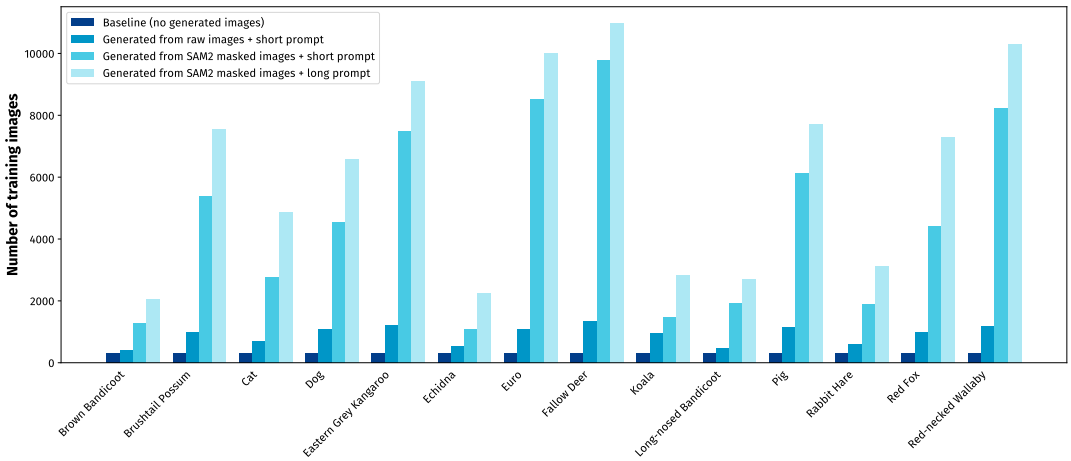

Figure S3: The distribution of the training datasets of our four experiments.

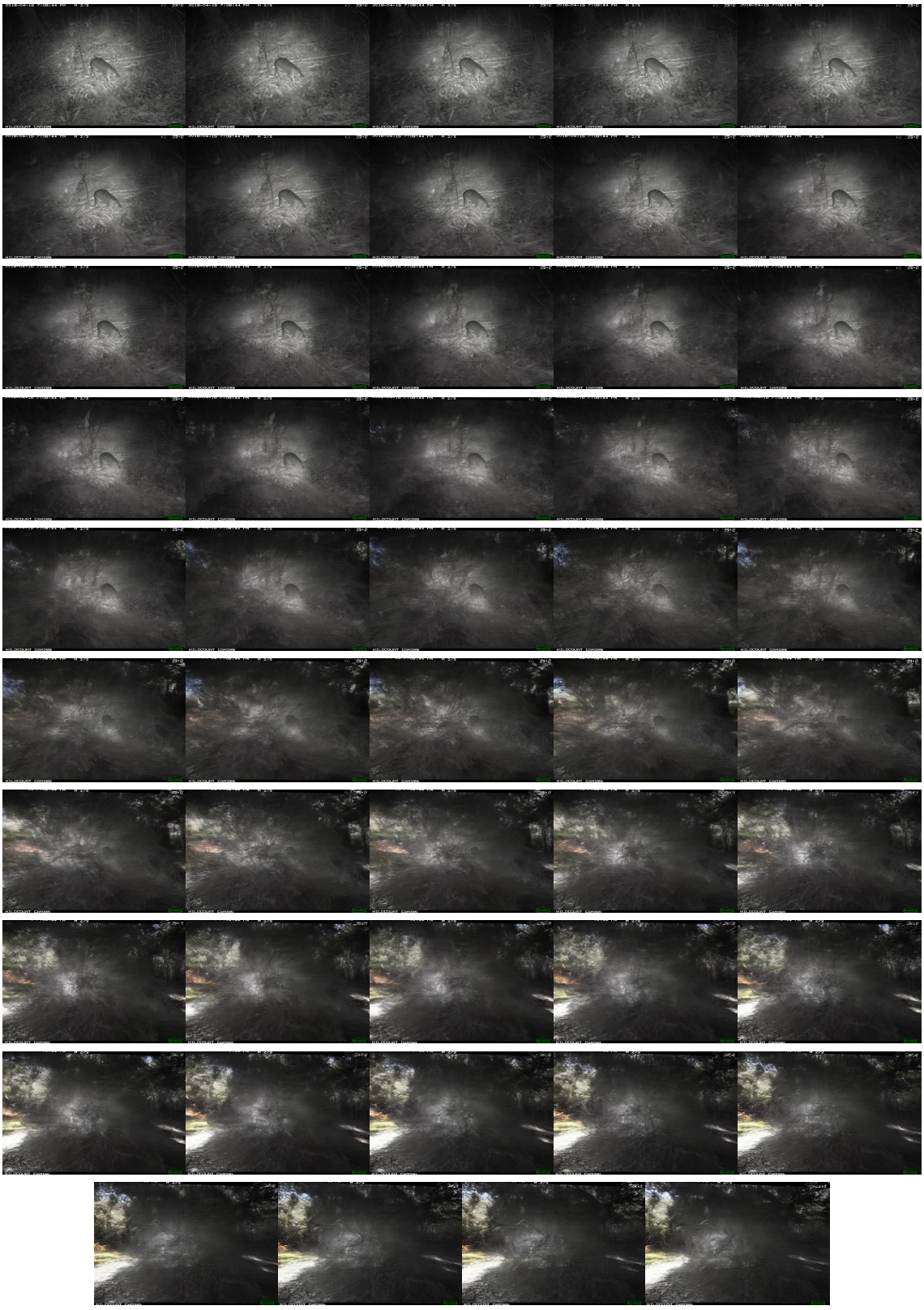

Figure S4: Video sequence generated for Experiment 2, where videos were generated directly from the raw images, showing the camera orbiting around a Brown Bandicoot, with the bandicoot gradually becoming occluded by the environmental elements.

Figure S5: Video sequence generated for Experiment 2, where videos were generated directly from the raw images, showing the camera orbiting around a Koala, with the Koala gradually disappearing into the background scene.

Figure S6: Video sequence generated for Experiment 2, where videos were generated directly from the raw images, showing the camera orbiting around a Euro (Common Wallaroo), where the limbs were not accurately synthesized in the second half of the video when moving from the right to left side of the body.

Figure S7: Video sequence generated for Experiment 3, where we first created segmented images using SAM2, and then generated videos using a short text prompt, showing the camera orbiting around a Euro (Common Wallaroo), where incorrect geometry was generated on the left side of the body.

Figure S8: Video sequence generated for Experiment 4, where we created segmented images with SAM2 first, as shown in Experiment 3, before generating videos using a long prompt and a negative prompt, showing the camera orbiting around a Brown Bandicoot. As the camera orbits around the bandicoot, the geometry between the snot and the left each became ambiguous and increasingly similar.

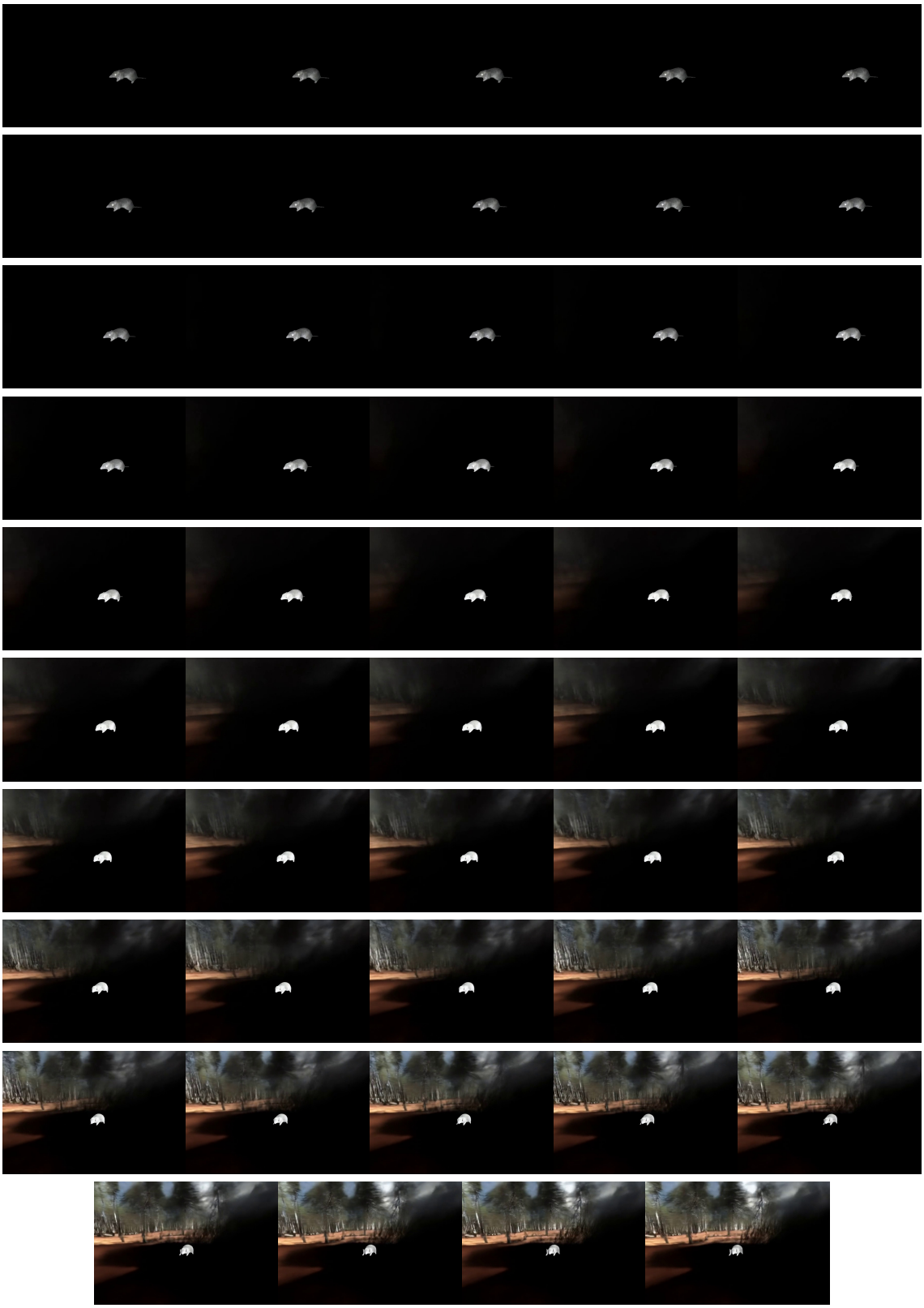

Figure S9: Video sequence generated for Experiment 4, where we created segmented images with SAM2 first, as shown in Experiment 3, before generating videos using a long prompt and a negative prompt, showing the camera orbiting around a Long-nosed Bandicoot. Although the synthesized geometry of the bandicoot remained accurate, there was a reduction in quality of fur texture in the second half of the sequence.

Figure S10: Video sequence generated for Experiment 4, where we created segmented images with SAM2 first, as shown in Experiment 3, before generating videos using a long prompt and a negative prompt, showing the camera orbiting around a Euro (Common Wallaroo). When orbiting from the right to left side of the wallaroo's body, the model synthesized incorrect limb and body geometry.

Table S1: Mean orbit angles of all videos, summarised across each experiment and class. Values include outliers and are rounded to three decimal places.

Experiment	Class	Degree
Experiment 2	Brown Bandicoot	94.053 (±65.768)
	Brushtail Possum	95.039 (±72.051)
	Cat	96.159 (±50.981)
	Dog	90.540 (±29.137)
	Eastern Grey Kangaroo	84.413 (±50.260)
	Echidna	90.328 (±91.664)
	Euro	93.671 (±28.259)
	Fallow Deer	88.033 (±13.738)
	Koala	86.378 (±21.421)
	Long-nosed Bandicoot	96.145 (±55.818)
	Pig	90.105 (±33.154)
	Rabbit Hare	90.160 (±88.537)
	Red Fox	85.863 (±86.028)
	Red-necked Wallaby	90.087 (±39.743)
Experiment 3	Brown Bandicoot	101.920 (±156.693)
	Brushtail Possum	109.273 (±84.053)
	Cat	102.955 (±96.402)
	Dog	106.082 (±107.043)
	Eastern Grey Kangaroo	115.765 (±52.943)
	Echidna	128.704 (±105.313)
	Euro	107.265 (±54.626)
	Fallow Deer	109.846 (±44.063)
	Koala	109.247 (±157.366)
	Long-nosed Bandicoot	110.876 (±147.822)
	Pig	109.423 (±68.135)
	Rabbit Hare	125.417 (±151.521)
	Red Fox	106.445 (±65.455)
	Red-necked Wallaby	106.965 (±26.082)
Experiment 4	Brown Bandicoot	109.182 (±103.732)
	Brushtail Possum	111.376 (±36.224)
	Cat	104.341 (±90.195)
	Dog	108.589 (±65.503)
	Eastern Grey Kangaroo	113.780 (±39.603)
	Echidna	131.713 (±72.970)
	Euro	111.521 (±41.790)
	Fallow Deer	112.180 (±37.147)
	Koala	113.827 (±104.487)
	Long-nosed Bandicoot	112.038 (±101.798)
	Pig	114.187 (±62.352)
	Rabbit Hare	111.521 (±109.145)
	Red Fox	102.280 (±62.490)
	Red-necked Wallaby	109.650 (±27.559)

Table S2: Mean orbit angles of videos excluding outlier values, summarised across each experiment and class, and rounded to three decimal places.

Experiment	Class	Degree
Experiment 2	Brown Bandicoot	86.150 (±10.095)
	Brushtail Possum	82.564 (±18.037)
	Cat	87.556 (±9.693)
	Dog	89.209 (±7.064)
	Eastern Grey Kangaroo	88.116 (±9.006)
	Echidna	87.730 (±13.963)
	Euro	92.730 (±6.501)
	Fallow Deer	87.383 (±7.796)
	Koala	84.679 (±6.912)
	Long-nosed Bandicoot	85.832 (±14.273)
	Pig	87.872 (±8.597)
	Rabbit Hare	82.766 (±14.175)
	Red Fox	80.833 (±18.557)
	Red-necked Wallaby	89.324 (±5.770)
Experiment 3	Brown Bandicoot	122.154 (±72.603)
	Brushtail Possum	101.243 (±17.258)
	Cat	95.236 (±27.849)
	Dog	97.526 (±23.109)
	Eastern Grey Kangaroo	107.847 (±15.972)
	Echidna	142.439 (±70.663)
	Euro	106.824 (±12.303)
	Fallow Deer	106.505 (±14.078)
	Koala	111.075 (±56.800)
	Long-nosed Bandicoot	132.105 (±78.099)
	Pig	106.144 (±18.284)
	Rabbit Hare	133.040 (±75.694)
	Red Fox	97.741 (±16.943)
	Red-necked Wallaby	103.986 (±13.449)
Experiment 4	Brown Bandicoot	110.583 (±57.713)
	Brushtail Possum	107.138 (±20.728)
	Cat	96.200 (±27.811)
	Dog	103.833 (±21.398)
	Eastern Grey Kangaroo	107.368 (±14.608)
	Echidna	122.693 (±55.171)
	Euro	109.226 (±13.449)
	Fallow Deer	107.800 (±14.987)
	Koala	119.687 (±58.927)
	Long-nosed Bandicoot	119.996 (±67.860)
	Pig	110.352 (±20.107)
	Rabbit Hare	104.995 (±45.872)
	Red Fox	104.498 (±17.756)
	Red-necked Wallaby	106.640 (±13.978)