

Text-Driven Data Augmentation Tool for Synthetic Bird Behavioural Generation

David Mulero-Pérez¹[0000-0002-1712-7265], David Ortiz-Perez¹[0009-0008-4890-8217], Manuel Benavent-Lledo¹[0000-0001-8809-8476], Jose Garcia-Rodriguez¹[0000-0002-7798-3055], and Jorge Azorin-Lopez¹[0000-0003-4762-6927]

Department of Computer Technology, University of Alicante, Alicante, Spain
{dmulero,dortiz,mbenavent,jgarcia,jazorin}@dtic.ua.es

Abstract. Environmental conservation and biodiversity monitoring efforts have been greatly enhanced by the use of deep learning and computer vision technologies, particularly in protected areas such as parks and wildlife reserves. However, the development of accurate species recognition and behaviour understanding models is limited by the lack of detailed action annotations and reliance on static images, despite the availability of several animal datasets. This paper explores the use of generative models to tackle the challenge of creating comprehensive and diverse synthetic video data for identifying bird species and their behaviour. By utilising and fine-tuning different generative models, this study assesses the feasibility of synthetic video data generation to overcome these limitations. Our work focuses on generating realistic and varied video sequences that can improve machine learning algorithms for bird action detection and species recognition, thereby contributing to the protection and management of natural habitats. Throughout this investigation, we aim to provide new methodologies and tools for wildlife conservation technology, enhancing the monitoring and safeguarding of bird populations in their natural environments.

Keywords: Synthetic data · Data augmentation · Video generation · Bird behaviour

1 Introduction

The development of deep learning and computer vision technologies has ushered in a new era of environmental conservation and biodiversity monitoring, particularly in protected natural habitats such as parks and wildlife reserves [14, 16]. An essential aspect of these efforts is the precise identification of animal species and the comprehension of their behaviours through video analysis [20]. Birds present a unique challenge and opportunity due to their diverse species and behaviours. Developing sophisticated models to predict bird species and their actions from video data holds immense potential for enhancing the management and protection of natural environments. However, this development is heavily

contingent upon the availability of comprehensive and diverse video datasets. In this context, explainability becomes important as it ensures that the decisions made by these models are transparent and understandable [7], which is crucial for building trust among conservationists, researchers, and the general public.

Although there are numerous datasets available that catalogue animals, such as AnimalKingdom [13] and VB100 [5], many of these datasets are limited in their focus on static images. Additionally, they often lack annotations for the specific actions performed by the animals, which is a fundamental aspect for understanding their behaviour and interactions within their ecosystems. The lack of specificity and uniformity in the data format limits the ability to create models that can accurately recognise actions and identify species from video footage.

Generative models have emerged as a promising solution to augment existing datasets in response to these challenges. In particular, the use of generative models to create synthetic video data of animals, especially birds, performing various actions offers a novel approach to overcome the limitations of current datasets. The models have the potential to generate realistic and varied video sequences that can significantly enhance the training and performance of machine learning algorithms for action detection and species recognition in birds.

In this paper, we have focused on using and fine-tuning different generative models to assess the viability of these techniques for data augmentation purposes, as shown in Figure 1. The aim of this study is to bridge the gap in available datasets by generating synthetic video data that closely mimics real-world observations. Our exploration into the use of generative models for data augmentation is detailed in this paper, evaluating their effectiveness in creating useful datasets for applications such as bird action detection. This investigation aims to contribute to the wider field of wildlife conservation technology by providing new tools and methodologies for safeguarding and monitoring bird populations in their natural habitats.

In summary, our contributions are the following:

- We have developed a comprehensive pipeline for fine-tuning and utilising generative video models specifically tailored for the field of bird action recognition.
- Additionally, we have conducted a thorough evaluation and comparison of various generative techniques’ performance in the context of bird action recognition.

The remainder of the paper is structured as follows: Section 2 discusses related works and relevant datasets. Section 3 proposes a pipeline that utilises generative video models. The results are presented in Section 4. Finally, in Section 5, we present the conclusions of our work.

2 Related Works

The proliferation of deep learning and computer vision applications requires an expansive and detailed dataset for training purposes. The development of robust

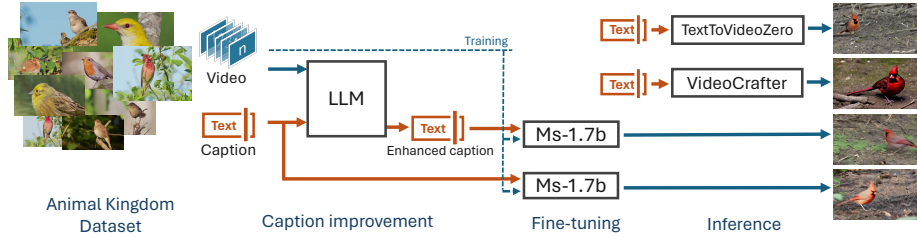


Fig. 1. End-to-End Pipeline for Synthetic Video Generation. The Text2VideoZero and VideoCrafter models are used in conjunction with the Video2Video models to produce higher quality videos. The text-to-video-ms-1.7b model has been retrained using the AnimalKingdom dataset and improved captions.

classification models [21] for bird species detection and action recognition is particularly challenging due to the scarcity of labelled data [17]. However, the lack of data variability limits the models’ ability to accurately understand the nuanced behaviours and appearances of diverse bird species. To address this issue, it is necessary to increase the amount of labelled data available for training. Synthetic video generation is a promising way to fill this gap and overcome the limitations imposed by the inadequacies of current datasets [12, 4]. By simulating a broad spectrum of scenarios and conditions, synthetic data offers the vital variability and scope often absent in traditional datasets.

2.1 Birds datasets

In the broader scope of animal datasets, the primary focus has been on cataloguing various animal species, encompassing details such as kind, species, and the actions they are performing, often annotated with bounding boxes [2, 23]. A standout in this category is the Animal Kingdom dataset [13], a comprehensive collection designed to deepen our understanding of animal behaviours across a spectrum of environmental conditions and viewpoints. This dataset is notable due to its breadth and depth, comprising 50 hours of annotated videos for video grounding tasks, 30,000 video sequences for fine-grained multi-label action recognition. Covering 850 species across 6 major animal classes, Animal Kingdom offers an unparalleled resource for the study of animal behaviour, including a significant subset dedicated to birds. With data on thousands of bird species performing hundreds of distinct actions, this dataset presents a goldmine for fine-tuning generative models aimed at bird video generation, thereby enhancing model performance in this specialised domain.

VB100 dataset [5] also offers multimodal data, including both video and audio recordings of birds engaging in various actions. It focuses on 100 bird species native to North America. This geographical specificity might limit its applicability for detecting species in Europe, where local species diversity and behaviours could differ significantly. In addition to video datasets, there are

also several image-based datasets such as Birds525 [6], CUB-200-2011 [18], and NABirds [19]. These datasets provide valuable insights for models to learn species identification through visual features. However, they fall short in supporting action detection due to the lack of temporal information.

The challenge remains in capturing and analysing the full range of bird actions, although each dataset has its strengths, particularly in species identification and understanding static visual features. This gap highlights the ongoing need for datasets that offer not only a broad range of species but also a depth of behavioural data, including temporal dynamics and multimodal inputs, to advance the field of bird action recognition significantly.

Name	Modalities	Labeled data	Samples	Bird species	Actions
VB100 [5]	Video, Audio	Species, actions	1416	100	–
Animal Kingdom [13]	Video, Audio	Species, actions	50h	1000	140
Birds525 [6]	Image	Species	84635	525	–
CUB-200-2011 [18]	Image	Species	5,994	200	–
NABirds [19]	Image	Species	48000	400	–

Table 1. Datasets used for bird species recognition.

2.2 Generative models

The advent of generative models for synthetic videos has led to significant advances but faces challenges in temporal coherence and realism. The exploration of generative image models based on textual descriptions has opened new avenues for generating photo-realistic images of animals, enriching the field with detailed and diverse visual data. This approach has seen remarkable applications in creating images that capture the essence and diversity of animal species with unprecedented fidelity.

In the realm of leveraging commonsense knowledge for image generation, the work of CD-GAN [22] presents a pioneering approach. It proposes a novel methodology to generate photo-realistic images grounded in entity-related commonsense knowledge. It incorporates generates high-resolution images guided by various commonsense knowledge in multiple stages to maintain text-image consistency. It has been demonstrated on the widely used CUB-birds dataset and achieves competitive results, showcasing its efficacy in generating realistic images consistent with textual descriptions.

However, the use of text-to-image models, while groundbreaking for static image generation, falls short when the goal extends to simulating behaviours or actions, necessitating a shift towards video generation models. For example,

Text2Video-Zero [9] adapts text-to-image methods for video generation, and despite its enhanced scene consistency, it struggles with temporal stability. On the other hand, VideoFusion [11] employs a decomposed diffusion process for improved video quality but grapples with realism in dynamic scenes. Similarly, Gen-2 Video [3] and VideoCrafter [1] offer detailed control over video editing through textual and visual descriptions, yet faces challenges in balancing structure and content for realistic portrayal. Collectively, these methods demonstrate progress but also highlight the ongoing need for refinement in realistic and temporally coherent video generation.

3 Synthetic video generation

In the pursuit of enriching the dataset for bird species detection and action recognition, our efforts have been directed towards the utilisation and fine-tuning of generative video models. This approach significantly augments the diversity and volume of bird video data available, facilitating the creation of synthetic videos depicting birds engaging in a variety of actions. The cornerstone of this endeavour is the AnimalKingdom dataset [13], renowned for its extensive collection of videos and accompanying textual descriptions of each video, including the actions performed.

3.1 Enhancing captions

Due to the fact that the textual descriptions provided by the AnimalKingdom dataset were found to be somewhat lacking in detail, we devised a pipeline to enhance these descriptions using Vision Language Models (VLMs) and Large Language Models (LLMs). The initial phase involved employing the *Blip2* model [10] for video captioning, guided by prompts designed to elicit detailed descriptions of the bird’s surroundings and its behaviour. Specifically, we asked, “Is the bird at any particular location like a river or forest? Describe the landscape.” and “Is the bird sitting, standing, or engaging in any other observable behaviour? Describe it.” The responses generated from these prompts were then amalgamated with the original dataset descriptions. A subsequent inference was performed on the LLM model *Mistral-7B-v0.1* [8] using a templated prompt that aimed to condense the information into cohesive sentences. This process yielded more detailed video descriptions, capturing nuances such as background elements and the bird’s actions more effectively.

3.2 Generative video models

These enhanced descriptions are used to fine-tune the *text-to-video-ms-1.7b* model based on the VideoFusion architecture. The model was trained using both the original and the newly generated descriptions, resulting in two fine-tuned versions for comparative evaluation against the original model. In addition to refining the descriptions, we explored the potential of two new generative models,

VideoCrafter [1] and TextToVideoZero [9]. These models, building upon the StableDiffusion2 architecture [15], have demonstrated the capability to produce high-quality videos. However, a notable limitation was their treatment of temporal dynamics, often resulting in videos with static backgrounds. To address this issue, we incorporated a video2video model, *zeroscope_v2_576w*, designed to enhance video realism by introducing motion and dynamism into the scenes. The integration of this model has markedly improved the realism and dynamic quality of the generated videos, making them more closely resemble real-life footage.

The development of these synthetic processing and generation pipelines has been instrumental in producing highly realistic videos. This has enabled us to generate a small dataset to validate our proposal for different bird species, such as: Shoebill, Great Reed Warbler, Cardinal, Tawny Owl or Sparrowhawk bird. Figure 2 shows some of the frames generated for various species. The qualitative evaluation of these videos will further attest to the efficacy of our approach in generating synthetic data that closely mimics real-world scenarios, thereby enriching the available dataset for bird species detection and action recognition research.



Fig. 2. Examples of synthetic videos generated for various bird species.

4 Results

This section presents the results obtained from generating synthetic bird videos using various text prompts across different video generation models. The experimentation involve the use of pre-trained models such as Text2VideoZero and VideoCrafter, followed by the application of a video-to-video (V2V) model to enhance the overall quality of the generated content. The aim of the subsequent V2V processing is to introduce more temporally coherent details, significantly improving the fluidity of the animations and minimising common artefacts such as flickering. The initial results show that the bird movements and actions are convincingly realistic, with commendable temporal coherence. However, the variability in the generated videos is limited, and the backgrounds tend to be overly simplistic and lacked intricacy.

To overcome these limitations, we fine-tune the *text-to-video-Ms-1.7b* model, using both the original video captions from the AnimalKingdom dataset and

the extended captions derived from the responses previously generated. The aim of this approach is to enhance the video content by adding more detailed and dynamic backgrounds, which could potentially increase the realism and diversity of the generated videos. The videos produced by these fine-tuned models show a significant improvement in terms of environmental richness and detail, featuring more dynamic and varied backgrounds as well as movements from the birds. However, this enhancement come at a cost. The fidelity of the videos noticeably decreased compared to those generated by the pre-trained models. The fine-tuned models successfully introduced a greater depth of scene complexity, but slightly compromised on the visual clarity and sharpness that characterised the original generative models’ outputs. Figure 3 shows an example of an original video from the Animal Kingdom and VB100 dataset, as well as synthetic videos generated by the models. To generate the synthetic example videos, the following text prompt was used “A cardinal bird eats from the ground”.

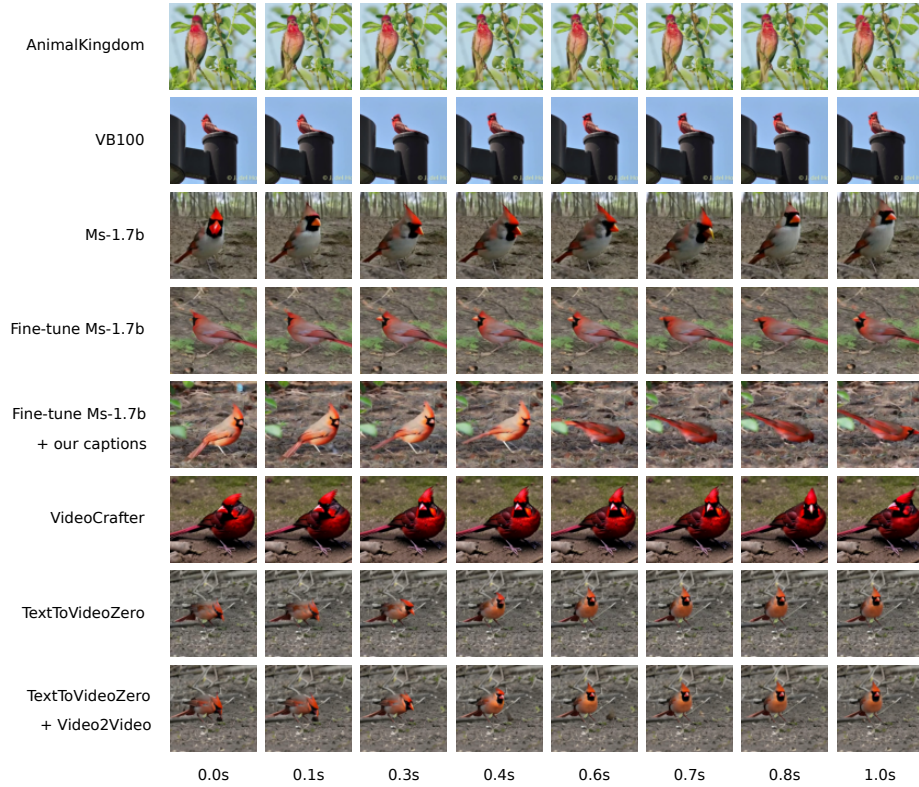


Fig. 3. Comparative analysis of video sequences of Cardinal bird (*Cardinalis cardinalis*) from different datasets, including AnimalKingdom and VB100, alongside sequences generated by AI generative models such as Text2Video-Zero with Stable Diffusion 2, VideoCrafter, and text-to-video-ms-1.7b.

5 Conclusions

In this study, we have explored the capabilities of generative models in augmenting video datasets, specifically focusing on the identification and behaviour analysis of bird species. Our research leverage and fine-tune various generative video models to create synthetic data that address the significant limitations posed by existing animal datasets, mainly their focus on static images and lack of detailed behavioural annotations.

The core of our work involved developing a sophisticated pipeline for generating synthetic videos that not only depict birds in various actions but also incorporate enhanced descriptions and backgrounds. This process involved improving video descriptions with the aid of LLMs and fine-tuning models using both original and enhanced descriptions. Additionally, we explored the use of cutting-edge models such as VideoCrafter and TextToVideoZero, further refining their generated outputs through video2video techniques, to achieve more realistic and temporally coherent video sequences.

Our findings suggest that this pipeline significantly contributes to the enhancement of bird video datasets, providing a robust foundation for training more accurate species and behaviour detection models. The variability and the realistic nature of the generated videos underscore the potential of generative models in overcoming the challenges of traditional datasets. Moreover, the methodology adopted in this research holds promise for application across different animal categories, opening avenues for comprehensive biodiversity monitoring and conservation efforts.

As a future work, we propose refining this pipeline to enable the generation of domain-specific synthetic datasets. This improvement will likely focus on enhancing the temporal coherence and realism of the generated videos, thereby broadening the scope of deep learning applications in environmental conservation and wildlife monitoring. Our work lays a foundational stone for future research in this area, suggesting a scalable and versatile approach to dataset augmentation that could revolutionise the field of wildlife conservation technology.

Acknowledgments We would like to thank “A way of making Europe” European Regional Development Fund (ERDF) and MCIN/AEI/10.13039/501100011033 for supporting this work under the “CHAN-TWIN” project (grant TED2021-130890B-C21). HORIZON-MSCA-2021-SE-0 action number: 101086387, REMARKABLE, Rural Environmental Monitoring via ultra wide-ARea networKs And distriButed federated Learning. This work has also been supported by a Spanish national and two regional grants for PhD studies, FPU21/00414, CIACIF/2021/430 and CIACIF/2022/175. Finally, we would like to thank the University Institute for Computer Research at the UA for their support.

References

1. Chen, H., Zhang, Y., Cun, X., Xia, M., Wang, X., Weng, C., Shan, Y.: Videocrafter2: Overcoming data limitations for high-quality video diffusion models. arXiv preprint arXiv:2401.09047 (2024)
2. Chen, K., Song, H., Change Loy, C., Lin, D.: Discover and learn new objects from documentaries. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3087–3096 (2017)
3. Esser, P., Chiu, J., Atighehchian, P., Granskog, J., Germanidis, A.: Structure and content-guided video synthesis with diffusion models. arXiv e-prints pp. arXiv–2302 (2023)
4. Garcia-Garcia, A., Martinez-Gonzalez, P., Oprea, S., Castro-Vargas, J.A., Orts-Escolano, S., Garcia-Rodriguez, J., Jover-Alvarez, A.: The robotrix: An extremely photorealistic and very-large-scale indoor dataset of sequences with robot trajectories and interactions. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 6790–6797 (2018). <https://doi.org/10.1109/IROS.2018.8594495>
5. Ge, Z., McCool, C., Sanderson, C., Wang, P., Liu, L., Reid, I., Corke, P.: Exploiting temporal information for dcnn-based fine-grained object classification. In: 2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA). pp. 1–6. IEEE (2016)
6. Gerry: BIRDS 525 SPECIES- IMAGE CLASSIFICATION. <https://www.kaggle.com/datasets/gpiosenka/100-bird-species> (2023), accessed: 2024-02-26
7. Górriz, J., Álvarez Illán, I., et al.: Computational approaches to explainable artificial intelligence: Advances in theory, applications and trends. *Information Fusion* **100**, 101945 (2023). <https://doi.org/https://doi.org/10.1016/j.inffus.2023.101945>, <https://www.sciencedirect.com/science/article/pii/S1566253523002610>
8. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.d.l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al.: Mistral 7b. arXiv preprint arXiv:2310.06825 (2023)
9. Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., Shi, H.: Text2video-zero: Text-to-image diffusion models are zero-shot video generators. arXiv:2303.13439 (2023)
10. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models (2023)
11. Luo, Z., Chen, D., Zhang, Y., Huang, Y., Wang, L., Shen, Y., Zhao, D., Zhou, J., Tan, T.: Videofusion: Decomposed diffusion models for high-quality video generation. In: IEEE/CVF CVPR (June 2023)
12. Martinez-Gonzalez, P., Oprea, S., Garcia-Garcia, A., Jover-Alvarez, A., Orts-Escolano, S., Garcia-Rodriguez, J.: Unrealrox: an extremely photorealistic virtual reality environment for robotics simulations and synthetic data generation. *Virtual Reality* **24**, 271–288 (2020)
13. Ng, X.L., Ong, K.E., Zheng, Q., Ni, Y., Yeo, S.Y., Liu, J.: Animal kingdom: A large and diverse dataset for animal behavior understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19023–19034 (June 2022)
14. Pino, J., Rodà, F., Ribas, J., Pons, X.: Landscape structure and bird species richness: implications for conservation in rural areas between natural parks. *Landscape and urban planning* **49**(1-2), 35–48 (2000)

15. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
16. Ruiz-Ponce, P., Ortiz-Perez, D., Garcia-Rodriguez, J., Kiefer, B.: Poseidon: A data augmentation tool for small object detection datasets in maritime environments. *Sensors* **23**(7) (2023), <https://www.mdpi.com/1424-8220/23/7/3691>
17. Song, Q., Guan, Y., Guo, X., Guo, X., Chen, Y., Wang, H., Ge, J., Wang, T., Bao, L.: Benchmarking wild bird detection in complex forest scenes. *Ecological Informatics* **80**, 102466 (2024)
18. Van Horn, G., Branson, S., Farrell, R., Haber, S., Barry, J., Ipeirotis, P., Perona, P., Belongie, S.: Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 595–604 (2015). <https://doi.org/10.1109/CVPR.2015.7298658>
19. Van Horn, G., Branson, S., Farrell, R., Haber, S., Barry, J., Ipeirotis, P., Perona, P., Belongie, S.: Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 595–604 (2015)
20. Vélez, J., McShea, W., Shamon, H., Castiblanco-Camacho, P.J., Tabak, M.A., Chalmers, C., Fergus, P., Fieberg, J.: An evaluation of platforms for processing camera-trap data using artificial intelligence. *Methods in Ecology and Evolution* **14**(2), 459–477 (2023)
21. Yang, W., Liu, T., Jiang, P., Qi, A., Deng, L., Liu, Z., He, Y.: A forest wildlife detection algorithm based on improved yolov5s. *Animals* **13**(19), 3134 (2023)
22. Zhang, G., Xu, N., Yan, C., Zheng, B., Duan, Y., Lv, B., Liu, A.A.: Cd-gan: Commonsense-driven generative adversarial network with hierarchical refinement for text-to-image synthesis. *Intelligent Computing* **2**, 0017 (2023)
23. Zhang, L., Gao, J., Xiao, Z., Fan, H.: Animaltrack: A benchmark for multi-animal tracking in the wild. *International Journal of Computer Vision* **131**(2), 496–513 (2023)