Imperial College London

Department of Earth Science and Engineering

MSc in Applied Computational Science and Engineering

Independent Research Project
Project Plan

Dreambirds in Motion: AI-Driven Temporal Consistency in Surreal Video

by

Marceline Cheng

Email: xinyu.cheng24@imperial.ac.uk
GitHub username: esemsc-xc924
Repository: https://github.com/ese-msc-2023/irp-xc924

Supervisors:

Prof. Christopher Pain
Prof. James Coupe (Royal College of Arts)

June 2025

Abstract

This project tackles temporal consistency and feature preservation in video generation from surreal bird imagery using diffusion models. Existing techniques like AnimateDiff generate good image-to-video results. However, they struggle to preserve distinctive traits across frames. These traits include oversized anatomy, multiple heads, and abstract distortions. Our approach combines AnimateDiff with ControlNet conditioning, LoRA fine-tuning, and saliency mechanisms including CLIP-guided loss. Tests start with normal inputs before moving to stylized content. Assessment relies on Fréchet Video Distance (FVD) and CLIP feature analysis alongside expert review for frame consistency and image quality. Results will include an adaptable video creation pipeline, collected surreal bird videos, and findings on maintaining unique features in diffusion-based video production.d

1 Introduction

1.1 Motivation and Objectives

This project inspires from *Birds of the British Empire*, a collaboration that uses surreal avian imagery to explore identity, imperial histories, and speculative nature through generated images and videos. However, past studies such as TokenFlow [1] and RIFE [2] have highlighted two key challenges in generating surreal birds with diffusion-based models: maintaining temporal consistency and preserving outlier features, which are often lost during denoising due to their deviation from standard training distributions.

This project, *Dreambirds in Motion*, builds on the exsiting diffusion model AnimateDiff [3], integrating ControlNet [4] and LoRA [5] to address these challenges. By enhancing feature preservation and temporal alignment in surreal video synthesis, the project contributes to both creative AI applications and generative art practices, with implications for synthetic biology, interactive media, and speculative storytelling.

1.2 Literature Review

Video synthesis from still images relies on modular diffusion model design. AnimateDiff [3] adds motion modules to pretrained Stable Diffusion pipelines, enabling video generation without full retraining while maintaining image quality. ControlNet [4]provides conditional guidance through edge maps, depth, or sketches, crucial for unconventional content like surreal birds. LoRA [5] enables efficient fine-tuning by adapting specific model components, allowing customization for unusual visual features without extensive computational costs.

In order to generate surreal images, models need to accept and support creative deviations from normal shapes or structures. DreamBooth [6] fine-tunes text-to-image models to preserve specific subject identities and key visual features. Expanding on this, Text2LIVE [7] adds guided editing, making it useful for controlled and stylistic transformations.

Video synthesis still facing challenges in maintaining consistent motion and identity across a video sequence. Tune-A-Video [8] uses temporal attention and identity loss to improve coherence. DynamiCrafter [9] learns motion paths through latent trajectory learning, enhancing temporal continuity. Temporal Diffusion Models [10] extend the denoising process into the temporal domain to model inter-frame dependencies. VideoComposer [11] offers controllable motion guidance. While these methods significantly improve frame-to-frame consistency, they often struggle to preserve abstract or symbolic features over time.

Another significant challenge is to keep the important visual features thoughout the whole video, especially for the surreal or fantastical elements. PYoCo [12] addresses this by introducing saliency-aware conditioning and noise control, helping the model to retain key regions. Saliency-aware attention mechanism [13] was developed to teach the model to focus on key areas across frames. CLIP-guided loss functions [14] [15] are used to match generated images with both semantic and visual mainings, hence improving consistency. Additionally, StyleCLIP [15] offers fine-grained control over feature manipulation, allows users to edit the features while keeping the main identity unchanged.

To address this, newer metrics are proposed to focus on perceptual quality. Metrics such as Fréchet Video Distance (FVD) [16], VideoFID [17], and CLIP-based evaluation [14] evaluate videos based on perceptual quality and semantic consistency rather than pixel differences. These metrics suit artistic videos better, as they measure thematic consistency, identity preservation, and visual narrative coherence.

2 Methodology

2.1 Baseline Framework

The project uses the AnimateDiff pipeline with RealisticVision V5.1 as the base model for the still images to video creation. This framework adds motion capabilities to existing image models through modular components that can be easily integrated. We evaluate its performance by making short videos from artistic bird images, to see how well the style and details stay consistent across frames. This serves as our starting point.

2.2 Dataset Selection

We use a combination of still image and video bird datasets. These datasets include: *CUB-200-2011* [18] for fine-grained attribute reference, *VB100* [19] for pose-aware motion benchmarking, and *FBD-SV-2024* [20] for evaluating motion consistency under real-world environmental variations.

For the surreal image component, we construct a custom dataset containing artificially generated bird images with exaggerated, fantastical, dream-like features. These images are generated through prompt-based image generation using large-scale diffusion models (e.g., via OpenAI API or Stable Diffusion WebUI).

2.3 Feature Preservation Strategy

To preserve the dream-like unusual features in surreal bird imagery and video, we employ multiple techniques that complement each other across functional and temporal dimensions. We apply ControlNet for structural control, leveraging edge, depth, or sketch guidance maps to constrain the generation process. We Also use LoRA-based fine-tuning to achieve style adaptation. This enables pretrained models to adapt to surreal distributions. We experiment with two types of LoRA: motion-focused LoRA for consistent deformation across frames, and attention LoRA that re-weights internal attention to emphasize stylized or anomalous regions. To guide perceptual importance, we incorporate PYoco-inspired saliency-aware attention mechanisms that dynamically adjust focus on key features. CLIP saliency loss maintains semantic consistency by penalizing cross-frame semantic drift, preventing the loss of identity and anomalous features.

2.4 Temporal Consistency Strategy

Even for realism video generation using diffusion-based model, maintaining temporal coherence remains a critical challenge. We conduct a comparative analysis of different sampling schedulers, particularly DDIM and DDPM, then evaluate their impact on motion realism and inter-frame smoothness. Additionally, we incorporate latent-space consistency losses to reduce noise and flicker effects. By adapting Tune-A-Video's temporal attention module, we aim to preserve structured motion trajectories. This ensures smooth transitions despite artistic distortions or symbolic transformations.

2.5 Experimental Progression Plan

We test our approach progressively, starting with simple motion generation before advancing to complex stylistic animation. First, AnimateDiff processes standard bird images to set baseline performance. Next, guided synthesis integrates ControlNet to maintain structural coherence with stylized inputs. We then fine-tune models using LoRA on surreal bird datasets, capturing motion patterns for unconventional features. Finally, dynamic transformation trials evaluate visual continuity during mid-sequence style transitions.

2.6 Evaluation Protocol

We evaluate temporal and perceptual performance using both metrics and expert review. Quantitatively, we measure perceptual quality with Fréchet Video Distance (FVD) and track semantic consistency via CLIP-based similarity drift. Qualitatively, experts assess saliency retention, identity preservation, and motion smoothness. Selected outputs undergo artistic review to showcase technical-artistic impact.

2.7 Project Workflow

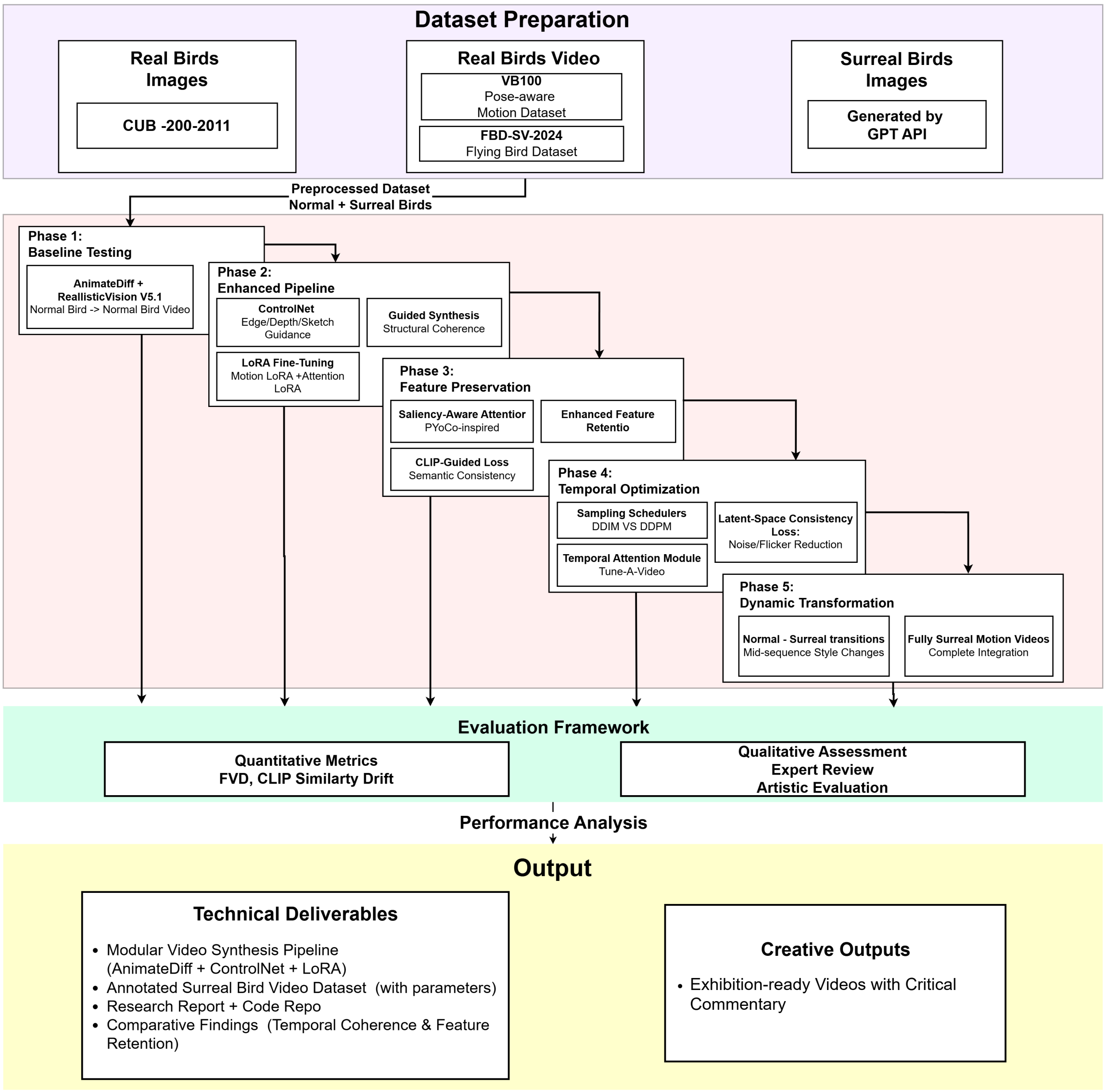


Figure 1: Total workflow for ”Dreambirds in Motion” project

3 Preliminary Results

To evaluate motion quality and structural preservation, we began by generating a base image in Stable Diffusion WebUI using the prompt: ”*A surreal bird with 4 wings flying in a purple sky*”. This initial frame was then processed through our AnimateDiff pipeline running the RealisticVision V5.1 model on the lab’s HPC cluster. Recognizing standard schedulers often caused temporal flickering, we built a custom DDIM scheduler specifically to smooth frame transitions and enhance motion coherence throughout the generated sequence. We tested two configurations:

1. Figure 2 shows the baseline AnimateDiff pipeline output. While exhibiting creative motion, the video suffers from structural deformation and object drift across frames.
2. Figure 3 demonstrates our ControlNet-enhanced approach. We extracted Canny edge maps from the base image using a custom Python script and used them to guide ControlNet. This approach significantly improving structural consistency throughout the video sequence.

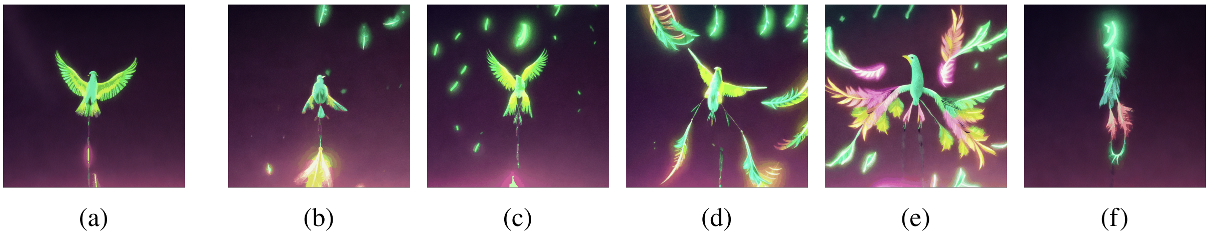


Figure 2: Video frame generated using the base AnimateDiff pipeline.
(a) Initial image generated with Stable Diffusion, used as the input of video generation.
(b-f) Frames sampled from a 24-frame video generation.

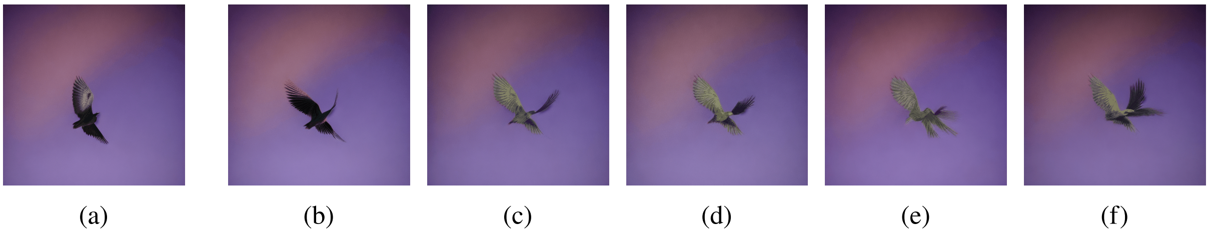


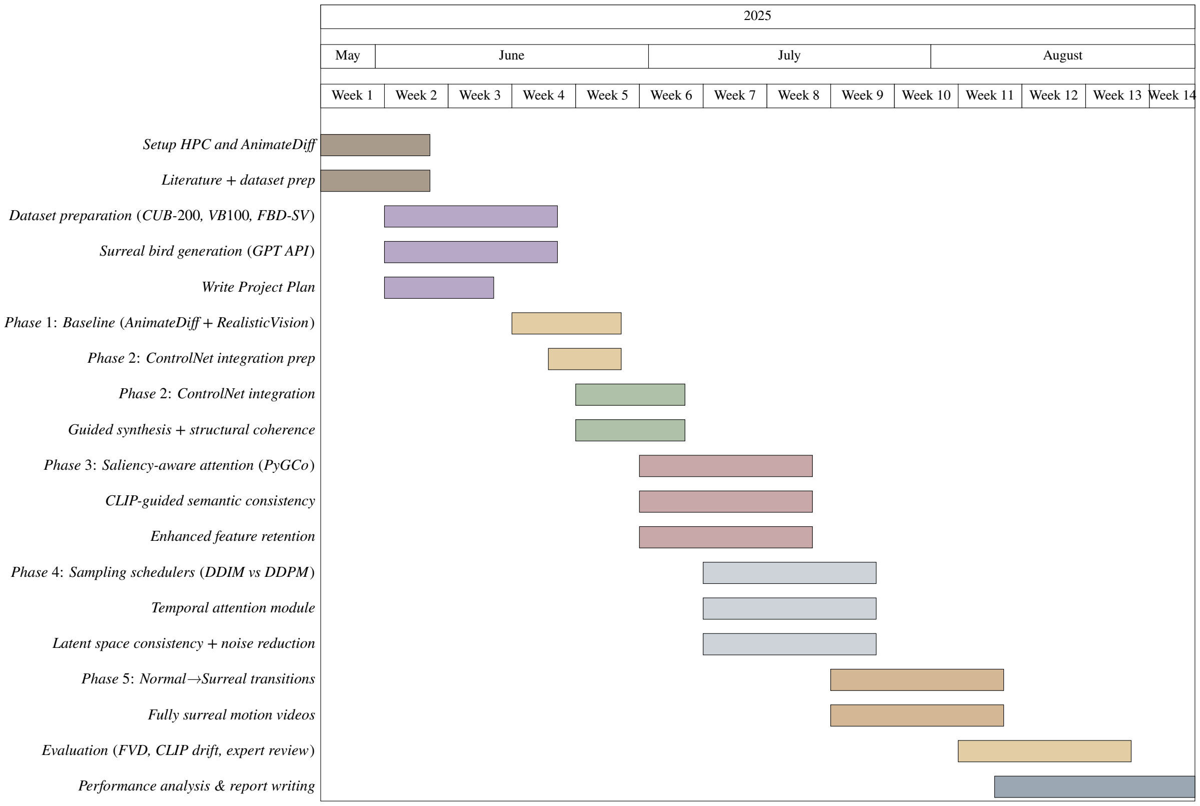
Figure 3: Video frame comparison using AnimateDiff with ControlNet (Canny edge)
(a) Initial image generated with Stable Diffusion, used as the input of video generation.
(b-f) Frames sampled from a 24-frame video generation.

4 Expected Outcomes and Deliverables

The project will produce a combination of technical, experimental, and optional creative outputs:

- A modular video synthesis pipeline for surreal bird imagery, built on AnimateDiff with ControlNet and LoRA integration, ensuring temporal consistency in generated sequences.
- An annotated dataset of surreal bird videos documenting experimental parameters (conditioning methods, schedulers, attention mechanisms) for research reuse.
- Comparative experimental findings on temporal coherence, feature retention, and conditioning efficacy, combining quantitative metrics with qualitative assessment.
- Complete MSc research documentation with technical report and well-documented code repository.
- Series of exhibition-ready visual outputs with critical commentary.

5 Future plan



AI Acknowledgement Statement

- **Tool Name and Version**: ChatGPT (4o)
- **Publisher/Provider**: OpenAI
- **URL**: https://chatgpt.com/?model=gpt-4o
- **Usage Description**: In the initial research phase, ChatGPT was used to analyze project requirements and background research, providing a clearer understanding of the project. It assisted in understanding specific AI frameworks (e.g., AnimateDiff, ControlNet) by clarifying documentation and explaining components in simpler terms. Additionally, it provided guidance on setting up environment dependencies and resolving module-related issues during HPC deployment.
- **Declaration**: All submitted work is my own. AI tools were used solely to support the development and understanding of the project, not to generate final content.

References

[1] Y. Ge, Y. Zhang, R. Du, *et al.*, "Tokenflow: Consistent diffusion features for consistent video editing," *arXiv preprint arXiv:2307.10373*, 2023. https://arxiv.org/abs/2307.10373.

[2] Z. Huang, T. Zhang, and P.-A. Heng, "Real-time intermediate flow estimation for video frame interpolation (rife)," *arXiv preprint arXiv:2011.06294*, 2022.

[3] S. Gu *et al.*, "Animatediff: Animate your personalized text-to-image diffusion models without specific tuning," *arXiv preprint arXiv:2307.04725*, 2023. https://arxiv.org/abs/2307.04725.

[4] Y. Zhang and M. Agrawala, "Adding conditional control to text-to-image diffusion models," *arXiv preprint arXiv:2302.05543*, 2023. https://arxiv.org/abs/2302.05543.

[5] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, W. Wang, and Z. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[6] N. Ruiz, Y. Li, V. Jampani, *et al.*, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," *arXiv preprint arXiv:2208.12242*, 2023. https://arxiv.org/abs/2208.12242.

[7] I. Bar-Tal, Y. Alaluf, O. Shapira, *et al.*, "Text2live: Text-driven layered image and video editing," *European Conference on Computer Vision (ECCV)*, 2022.

[8] J. Wu, Y. Ge, Y. Zhang, *et al.*, "Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation," *arXiv preprint arXiv:2212.11565*, 2022. https://arxiv.org/abs/2212.11565.

[9] Y. Hong, Z. Li, Y. Wang, *et al.*, "Dynamicrafter: Animating personalized text-to-image diffusion models," *arXiv preprint arXiv:2310.12190*, 2023. https://arxiv.org/abs/2310.12190.

[10] Z. Yang, J. Chen, Z. He, *et al.*, "Diffusion models for video generation," *NeurIPS*, 2022.

[11] Y. Tian, Z. Yu, C. Qian, *et al.*, "Videocomposer: Compositional video synthesis with motion controllability," *arXiv preprint arXiv:2311.11591*, 2023. https://arxiv.org/abs/2311.11591.

[12] D. Xu, Y. Zhang, J. Duan, *et al.*, "Pyoco: Pyramid-coordinated progressive noise conditioning for consistent video synthesis," *arXiv preprint arXiv:2308.14440*, 2023. https://arxiv.org/abs/2308.14440.

[13] J. Tang, Y. Zhang, and X. Wang, "Saliency-aware attention for temporal feature retention," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. Forthcoming.

[14] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning transferable visual models from natural language supervision," *International Conference on Machine Learning (ICML)*, 2021.

[15] O. Patashnik, Z. Wu, E. Shechtman, *et al.*, “Styleclip: Text-driven manipulation of stylegan imagery,” *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[16] T. Unterthiner, B. Nessler, G. Heigold, *et al.*, “Towards accurate generative models of video: A new metric & challenges,” *arXiv preprint arXiv:1812.01717*, 2018. https://arxiv.org/abs/1812.01717.

[17] Y. Wang, X. Qi, J. Zhang, and Z. Wang, “Improved video generation for multi-functional applications,” *arXiv preprint arXiv:2006.10738*, 2020.

[18] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, “The caltech-ucsd birds-200-2011 dataset,” in *Technical Report CNS-TR-2011-001*, 2011.

[19] Z. Cheng, W. Li, Y. Lu, and G. Huang, “Videobirds: A dataset for fine-grained bird classification in videos,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[20] X. Li, K. Zhou, M. Tang, and H. Chen, “Fbd-sv: A large-scale flying bird detection surveillance video dataset,” in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2024.