Imperial College London

Department of Earth Science and Engineering

MSc in Applied Computational Science and Engineering

Independent Research Project
Final Report

Dreambirds in Motion: AI-driven Surreal Video Generation with Pose-Guided Temporal Consistency

by

Xinyu (Marceline) Cheng

Email: xinyu.cheng24@imperial.ac.uk
GitHub username: esemsc-xc924
Repository: https://github.com/ese-msc-2024/irp-xc924

Supervisors:

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

September 2025

Contents

Abstract

Producing bird motion videos with AI remains challenging due to high costs, technical barriers, and the scarcity of temporally annotated skeletal datasets. This study presents an end-to-end framework for generating surreal bird motion videos from static images through a three-stage pipeline. First, HRNet detects avian keypoints in the CUB-15 skeletal representation. Second, a Motion Diffusion Model (MDM) generates temporally coherent pose sequences conditioned on initial frames and action labels. Finally, AnimateDiff with multi-branch ControlNet renders skeleton-guided videos while preserving surreal stylistic traits. To support evaluation, we developed a motion analysis framework and a WebUI-based MOS rating system. Experiments demonstrate improved motion fidelity, temporal consistency, and feature preservation, highlighting a promising pathway for controllable, AI-driven wildlife video generation with artistic flexibility.

1 Introduction

1.1 Background and Research Gap

Generative AI has emerged as a powerful tool not only for scientific analysis but also for creative expression, offering new ways to represent animals and nature. Inspired by the *Birds of the British Empire* project, which reimagined avian forms through surreal aesthetics, this study extends the exploration from static images to dynamic videos. While still imagery can capture artistic traits, it cannot convey the temporal dimension of flight and behaviour. Our work therefore situates itself at the intersection of computational science and artistic practice, aiming to bring temporal coherence into surreal bird generation.

Bird motion videos hold increasing importance across applications such as animation, education, and game design, where dynamic behaviours like flying, gliding, and landing need to be realistically portrayed. Beyond these practical uses, surreal bird videos—characterised by eccentric forms, unusual colours, and dreamlike stylisation—have unique creative potential. They broaden the expressive range of digital media and highlight the value of AI as a medium for cross-disciplinary innovation.

Despite progress in diffusion-based models, generating high-quality bird videos remains challenging. Available datasets often lack either temporal continuity or explicit skeletal labels, limiting the training of motion-aware generative models. Methodologically, direct image-to-video or text-to-video generation frequently produces unstable outlines, uneven wing motions, and a loss of surreal features across frames. Together, these gaps hinder the simultaneous preservation of motion fidelity and stylistic consistency.

Skeleton-driven approaches offer a promising direction, as they disentangle motion from appearance and allow more precise structural control. However, existing skeleton priors such as OpenPose are designed for humans and transfer poorly to avian morphology. This mismatch highlights a broader limitation of current generative frameworks: the lack of domain-specific structural representations for non-human species. Addressing this challenge is essential for advancing temporally coherent and artistically consistent bird video generation, with implications for both computational research and creative applications.

1.2 Literature Review

Avian Datasets and Pose Estimation

Research on avian motion has been supported primarily by static image datasets. The CUB-200-2011 dataset provides 200 bird species, 11,788 images, and 15 annotated keypoints, forming the basis for fine-grained recognition and pose estimation [1]. VB100 contributes over one thousand bird flight videos that are valuable for behavioural studies, though they lack skeletal annotations [2, 3]. Broader multi-species resources such as AP-10K [4] and Animal-Pose [5] expand pose estimation to diverse taxa, while frameworks like DeepLabCut [6] and SLEAP [7] have been successfully applied in animal behaviour analysis. Early 3D bird pose datasets such as 3D-POP [8] and 3D-MuPPET [9] demonstrate the feasibility of motion capture in controlled environments. However, the absence of large-scale, temporally annotated

skeletal data for birds remains a major limitation. Moreover, adapting human-oriented skeletons like OpenPose [10] to avian morphology introduces distribution gaps, motivating the design of bird-specific skeletons such as CUB-15, later compared with reduced variants OP-5 and OP-9.

Surreal Image Generation and Style Control

Diffusion-based models, most notably Stable Diffusion [11], have enabled high-quality and controllable image synthesis. Structural guidance can be provided by ControlNet, which introduces edge, depth, or pose conditions [12]. Personalisation methods (e.g. DreamBooth [13], Textual Inversion [14], IP-Adapter [15]) support subject-specific fidelity and stylistic consistency. These approaches have proven effective for generating surreal bird images with unusual colours or exaggerated forms. However, they primarily target static imagery, and applying them frame by frame often results in flickering and inconsistent stylisation. Motion-aware extensions remain limited, leaving temporal coherence an open challenge. Recent multimodal systems such as BLIP-2 [16], LLaVA [17], and Flamingo [18] offer automated prompt generation and cross-modal consistency, suggesting promising directions for ensuring stylistic coherence across video frames.

Motion and Video Generation Frameworks

Video diffusion has advanced rapidly. Stable Video Diffusion [19] and AnimateDiff [20] extend text-to-image diffusion into video, while systems such as VideoComposer [21] and DynamiCrafter [22] improve compositional control and temporal consistency. In motion modelling, frameworks such as the Motion Diffusion Model (MDM) [23] and MotionCLIP [24] achieve strong results in human motion synthesis by learning structured dynamics from data. However, directly transferring these methods to birds is non-trivial due to distinct skeletal topology, wing kinematics, and aerodynamic constraints. Temporal consistency techniques—such as transformer-based temporal attention [25], recurrent regularisation, or optical-flow–based smoothing—are widely adopted in human video generation, but remain underexplored in the avian domain.

Skeleton/Condition-guided Video Synthesis

Pose-guided approaches have shown that disentangling motion from appearance leads to controllable, high-quality video synthesis. OpenPose-driven pipelines [10], DreamPose [26], and MusePose [27] demonstrate impressive temporal smoothness and appearance consistency in humans. Their success suggests that skeleton-conditioned generation may also be effective for birds, though adaptation is needed to account for flexible wings, rapid mid-air manoeuvres, and inter-species variation. In addition, multi-ControlNet strategies, combining edge, depth, and skeleton conditions [12], can further enhance structural fidelity and motion control, yet remain unexplored for avian motion.

Evaluation of Motion and Video Quality

Evaluating motion and video quality requires both objective and subjective measures. In pose estimation, Percentage of Correct Keypoints (PCK) [28] and Mean Per-Joint Position Error (MPJPE) [29] are widely used. For perceptual similarity, FID [30] and LPIPS [31] are standard, while benchmark suites such as VBench [32] provide comprehensive evaluation protocols across motion fidelity, identity preservation, and temporal stability. Subjective human evaluation, typically via Mean Opinion Score (MOS), remains essential for capturing perceptual qualities such as surreal trait preservation and overall video appeal.

In summary, existing resources lack temporally annotated skeletal data and current generative models struggle with avian morphology and dynamics. Few studies have addressed the joint challenge of controllable, temporally coherent, and stylistically consistent bird motion, motivating skeleton-driven approaches for surreal video synthesis.

1.3 Objective

The objective of this study is to develop and evaluate a skeleton-driven generative framework for surreal bird motion video synthesis that balances temporal coherence with stylistic consistency. The proposed pipeline consists of three stages: (i) HRNet-based detection of avian keypoints in the CUB-15 representation, (ii) a Motion Diffusion Model (MDM) for generating temporally plausible skeleton sequences, and (iii) AnimateDiff with multi-branch ControlNet conditioning (pose, edge, depth) for rendering skeleton-guided videos while preserving surreal traits.

The central research question is: *How effectively can skeleton-driven generative models simulate and extend avian motion while maintaining temporal coherence and surreal identity?* This is addressed through three sub-questions: (1) Can HRNet reliably detect avian keypoints, and how does skeleton design (CUB-15 vs. OpenPose variants) affect downstream motion modelling? (2) Can MDM generate realistic and flexible motion sequences that capture both natural dynamics and surreal extensions beyond biology? (3) Can AnimateDiff with multi-conditional ControlNet render skeleton-guided videos with smooth transitions and consistent surreal appearance?

The study hypothesises that: (a) skeleton-driven approaches will provide greater controllability and temporal coherence than direct image-to-video diffusion, (b) explicit conditioning via pose, edge, and depth is essential for preserving surreal features, and (c) bird-specific skeletons (CUB-15) will outperform reduced OpenPose proxies (OP-9) in aligning motion with rendering fidelity.

Evaluation combines objective motion metrics (PCK, MPJPE) with subjective perceptual ratings collected through a custom WebUI-based MOS system, enabling a systematic analysis of controllability, coherence, and surreal preservation in avian video synthesis.

2 Methodology

2.1 Workflow

We propose a three-stage framework that disentangles motion from appearance to generate controllable, temporally coherent, and visually consistent bird motion videos.

Stage 1: Pose Detection. An HRNet-based keypoint detector is trained on the CUB-200 dataset and reformatted into a 15-point avian skeletal representation (CUB-15). This stage converts static bird images into structured skeletons that capture anatomical landmarks, serving as conditioning signals for subsequent motion generation.

Stage 2: Motion Generation. A Motion Diffusion Model (MDM), trained on over 4,000 procedurally generated sequences with biomechanical constraints, produces species-aware skeletal dynamics. The model outputs temporally consistent trajectories in the format $[T, 15, 3]$, representing frame length, keypoint indices, and spatial coordinates (with visibility). Controllability is achieved via action labels (e.g., takeoff, gliding, hovering, soaring, diving, landing) and initial-pose conditioning to anchor the first frame.

Stage 3: Video Rendering. AnimateDiff with multi-branch ControlNet converts skeleton-guided trajectories into high-quality videos while preserving surreal traits. Three complementary conditioning signals are used: (i) Canny edges for global contours, (ii) depth maps for volumetric cues, and (iii) pose skeletons for frame-level kinematics. Pose conditioning is evaluated with the bird-specific CUB-15 skeleton and reduced OpenPose-style variants (OP-9). This comparison reveals how skeleton design impacts temporal fidelity, structural consistency, and controllability. All conditioning inputs are resolution-aligned; pose is provided per frame, whereas edge and depth are reused as static priors across frames.

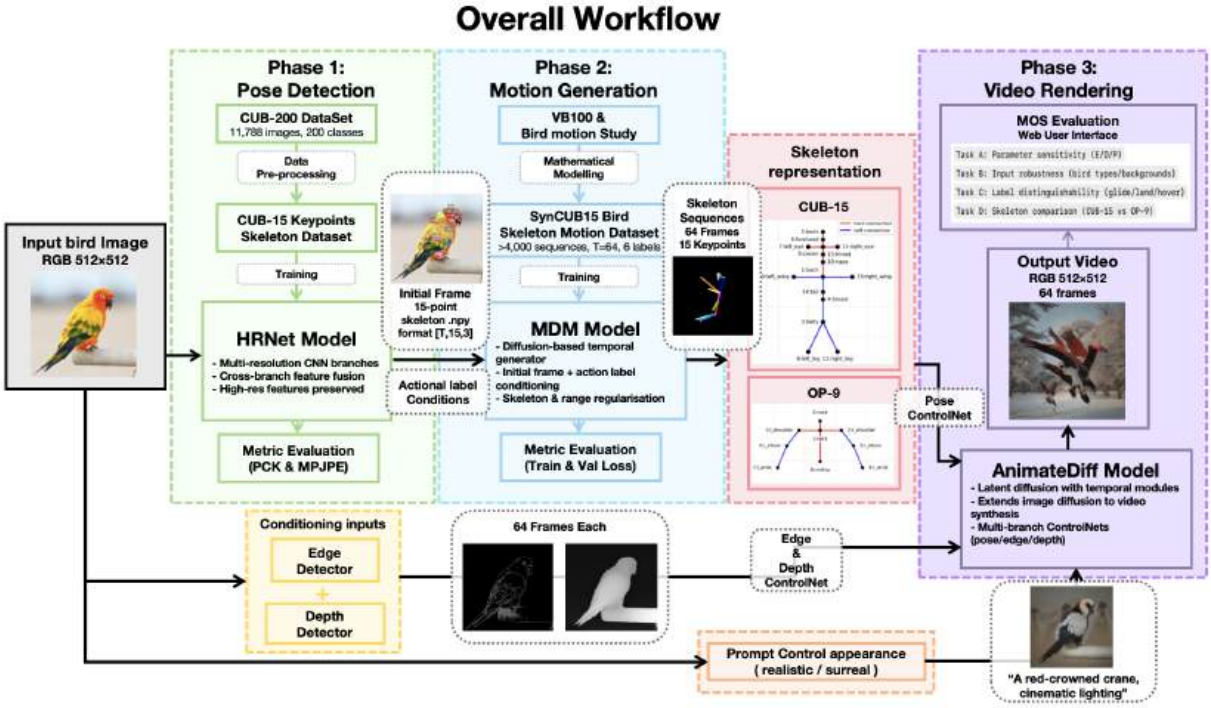


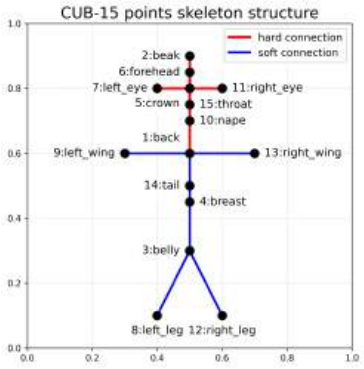
Figure 1: Overall workflow of the proposed Dreambirds pipeline.

2.2 Dataset

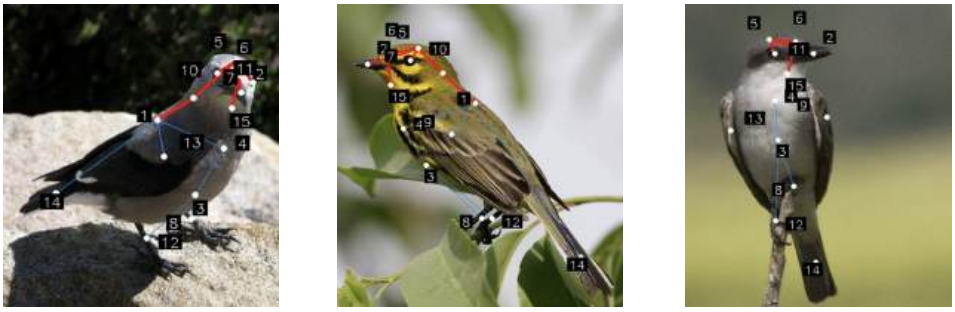
2.2.1 CUB dataset

The Caltech-UCSD Birds-200-2011 (CUB-200-2011) dataset contains 11,788 images across 200 bird species with 15 manually annotated part locations (beak, crown, wings, legs, tail). However, these keypoints lack structural connectivity, making them unsuitable for motion modelling. We therefore designed a custom 15-point skeleton schema (Fig. 2a) that links rigid anatomical parts (e.g., beak–crown–nape–back spine) and flexible appendages (e.g., back–wings, back–tail, belly–legs) into a coherent graph encoded in `skeleton.yaml`.

The annotations were converted into COCO-style format through a preprocessing pipeline, creating a structured dataset compatible with standard pose estimation frameworks. This restructured skeleton not only enabled HRNet training, but also provided a quantitative basis for evaluation using PCK and MPJPE metrics. Figure 2b illustrates the custom skeleton schema and example annotated images, which serve as input for HRNet training.



(a)



(b)

Figure 2: (a) Visualisation of the custom CUB-15 skeleton structure, with rigid (red) and flexible (blue) connections. (b) Pre-processed CUB-200 dataset images with annotated keypoints and skeletal connections, used as input for HRNet training.

2.2.2 SynCUB15 Motion Dataset

To address the scarcity of annotated bird motion data, we procedurally generated a synthetic dataset of pose sequences in the format $[T, 15, 3]$, where 15 anatomical keypoints are represented with $(x, y, \text{visibility})$. Biomechanical constraints such as fixed bone-length ratios and wing-folding limits were imposed to ensure anatomical plausibility. Each sequence thus encodes temporally consistent skeletal dynamics, serving as a surrogate for real-world motion capture and as training input for the Motion Diffusion Model (MDM).

Each sequence spans 64 frames, capturing the temporal evolution of the head, torso, wings, tail, and legs.

To enable controllable motion generation, we defined six canonical action categories—takeoff, gliding, hovering, soaring, diving, and landing—derived from a broader behavioural taxonomy. Each sequence was assigned a label from this taxonomy, later used as conditioning input for MDM training. Inference stability was further supported by anchoring the first frame to a detected skeleton. Biomechanical constraints and augmentation strategies were used to ensure realism and diversity, enhancing dataset variability and mitigating overfitting.

In total, the corpus contains over 4,000 labelled sequences, stratified into training and validation subsets, enabling MDM to learn species-aware, biomechanically grounded transitions. The VB100 dataset was additionally used as a qualitative reference for evaluating motion plausibility and informing the taxonomy, ensuring the dataset reflects common flight actions and natural variation. Overall, SynCUB15 provides a scalable surrogate for real-world motion capture, supporting MDM training for controllable bird motion generation. Figure 3 shows sample frames of SynCUB15 skeleton sequences.

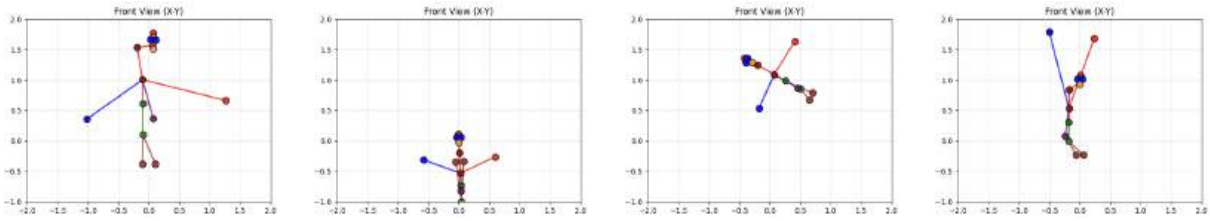


Figure 3: Sample frames front-view (XY) projections of CUB-15 skeleton sequences used for Motion Diffusion Model training.

2.3 Keypoint Detection using HRnet

Building on the restructured CUB-15 dataset (Section 2.2.1), the first stage requires a reliable pose detector for extracting structured skeletons from bird images. Since downstream motion generation and video synthesis depend on temporally consistent sequences, accurate localisation of both rigid anatomical landmarks (beak, crown, back, tail) and flexible appendages (wings, legs) is essential.

We adopt the High-Resolution Network (HRNet-W32) for its ability to maintain high-resolution representations throughout the network, enabling precise localisation of fine-scale details. Unlike conventional CNNs that progressively downsample, HRNet preserves high-resolution feature maps, which is advantageous for detecting small anatomical structures such as eyes and beak. Implementation is based on the MMPose framework, with input images resized to 256×256 and output heatmaps of 64×64 for 15 keypoints.

The trained HRNet is evaluated on the CUB-15 validation set. Results are reported both per keypoint and grouped into rigid vs. flexible categories, consistent with the skeleton definition in Section 2.2.1.

2.4 Pose Sequence Generation using MDM

The motion module transforms static bird skeletons into temporally coherent pose sequences. While the CUB dataset provides only still-image annotations, downstream video synthesis requires dynamic trajectories with species-aware movements and biomechanically plausible transitions.

We employ the Motion Diffusion Model (MDM), a generative framework that captures motion variability while preserving temporal smoothness. MDM combines the generative capacity of diffusion models with the sequential modelling power of Transformers, enabling the synthesis of smooth and diverse motion trajectories. The model learns to denoise skeleton sequences from Gaussian noise through iterative refinement, which makes it particularly effective for capturing long-range temporal dependencies compared to recurrent or adversarial approaches.

The model is trained on 64-frame sequences with six canonical action labels (takeoff, gliding, hovering, soaring, diving, landing) serving as conditioning signals. The objective follows the standard diffusion denoising formulation, supplemented by additional constraints to ensure anatomical plausibility and temporal stability. These include fixed bone lengths, spatial validity, and alignment with the ground-truth initial frame.

At inference, the model is conditioned on an initial skeleton frame and an action label. Generation proceeds by iterative denoising with classifier-free guidance, while softly anchoring the first frames to the user-provided pose. The final output is a 64-frame skeleton sequence that adheres to the chosen action label and evolves smoothly from the initial posture, enabling controllable bird motion generation.

The resulting MDM outputs serve as motion blueprints for the rendering stage, providing temporally consistent trajectories suitable for skeleton-conditioned video synthesis.

2.5 Video Generation using AnimateDiffusion+ ControlNet

The objective of this stage is to transform skeleton sequences into temporally coherent videos that preserve surreal traits while maintaining structural fidelity. To this end, we employ AnimateDiff for temporal modelling, combined with multiple ControlNets that provide complementary structural conditions.

AnimateDiff extends Stable Diffusion with a Motion Adapter, introducing temporal consistency across frames. ControlNet augments generation by conditioning on structural maps. Three signals are used: (i) Canny edges, preserving global silhouettes and preventing drift; (ii) depth maps, providing coarse volumetric cues; and (iii) OpenPose-style skeletons (CUB-15/OP-9), enforcing frame-level motion consistency. Together, these modules ensure visual appearance is controlled by static priors, while motion is governed by skeleton inputs.

The workflow proceeds as follows. Skeleton sequences from the Motion Diffusion Model are first smoothed with a Savitzky–Golay filter. The processed `.npy` files are then converted into OpenPose-style PNGs for ControlNet pose conditioning. Reference images are transformed into multi-frame Canny edge and depth sequences (64 frames), serving as consistent priors. Finally, these three branches (pose, edge, depth) are jointly fed into AnimateDiff, with branch scales tuned to balance motion adherence and appearance fidelity.

Appearance is specified either via textual prompts or directly from the reference image. Prompts introduce surreal traits such as unusual colours or textures, while image conditioning preserves realistic features like plumage and body markings. This dual mechanism allows flexible control between imaginative aesthetics and fidelity to the source bird.

All conditioning inputs are resolution-aligned and normalised. AnimateDiff inference generates frame-by-frame outputs, assembled into GIF or MP4 videos with logged seeds and configurations for reproducibility. The resulting videos achieve controllable motion guided by skeletons, while preserving surreal appearance through prompt conditioning and static edge/depth priors.

2.6 Evaluation

The evaluation strategy covers all three stages of the pipeline—pose detection, motion generation, and video rendering—ensuring that accuracy, temporal quality, and perceptual fidelity are assessed with complementary criteria. Since the final goal is to produce temporally coherent and stylistically consistent surreal bird motion videos, we employ both objective metrics and human-centred evaluation.

2.6.1 Pose Detection

For the HRNet-based detector, evaluation follows standard conventions in pose estimation. We report:

- **PCK (Percentage of Correct Keypoints):** The proportion of detected joints within a normalised error threshold τ of ground truth. We evaluate at $\tau \in \{0.05, 0.10, 0.15, 0.20\}$, and report results per keypoint as well as grouped into rigid vs. flexible categories.
- **MPJPE (Mean Per-Joint Position Error):** The mean Euclidean distance between predicted and ground-truth keypoints (pixels), complementing PCK by capturing localisation error.

2.6.2 Motion Generation

For the Motion Diffusion Model (MDM), evaluation focuses on the temporal quality and controllability of generated skeleton sequences:

- **Training and validation losses:** We monitor denoising loss, skeletal consistency, coordinate regularisation, and first-frame anchoring loss.
- **Qualitative error analysis:** Visualisation of skeleton sequences highlights characteristic errors such as wing asymmetry or left–right swaps.

2.6.3 Video Rendering

To complement objective evaluation metrics, we employed a **Mean Opinion Score (MOS)** procedure to assess the perceptual quality of the generated bird videos. We developed a lightweight web-based interface to facilitate this process: volunteer evaluators could watch randomized clips side-by-side with their corresponding skeleton sequences under unified resolution, frame rate, and duration, and then submit ratings online.

Each video was evaluated along four perceptual dimensions using a 1–5 Likert scale:

1. **Motion fidelity to skeleton**: 1 = clearly off-skeleton, 5 = strictly follows
2. **Temporal smoothness**: 1 = stutter or flicker, 5 = smooth
3. **Appearance consistency**: 1 = drifting/broken, 5 = stable
4. **Surreal trait preservation**: 1 = often lost, 5 = well preserved

An optional **overall quality** score (1 = poor, 5 = excellent) was also collected in some cases. For each clip, the ratings were averaged across evaluators to obtain the MOS for each dimension, which then served as the basis for subsequent analysis.

This MOS framework was applied to four experimental settings:

- **Task A Parameter sensitivity**: examining the effect of different parameter values on video quality
- **Task B Input robustness**: evaluating consistency across different bird species, backgrounds, and initial skeletons
- **Task C Label distinguishability**: measuring whether generated clips convey perceivable action labels, quantified primarily by recognition accuracy
- **Task D Skeleton comparison**: comparing CUB-15 versus OP-9 skeletal representations under identical parameters

By aggregating ratings across multiple evaluators, the MOS system provides a simple but effective perceptual metric, enabling us to identify optimal parameter ranges, evaluate robustness, verify label controllability, and quantify the influence of skeleton definitions on the generated outputs.

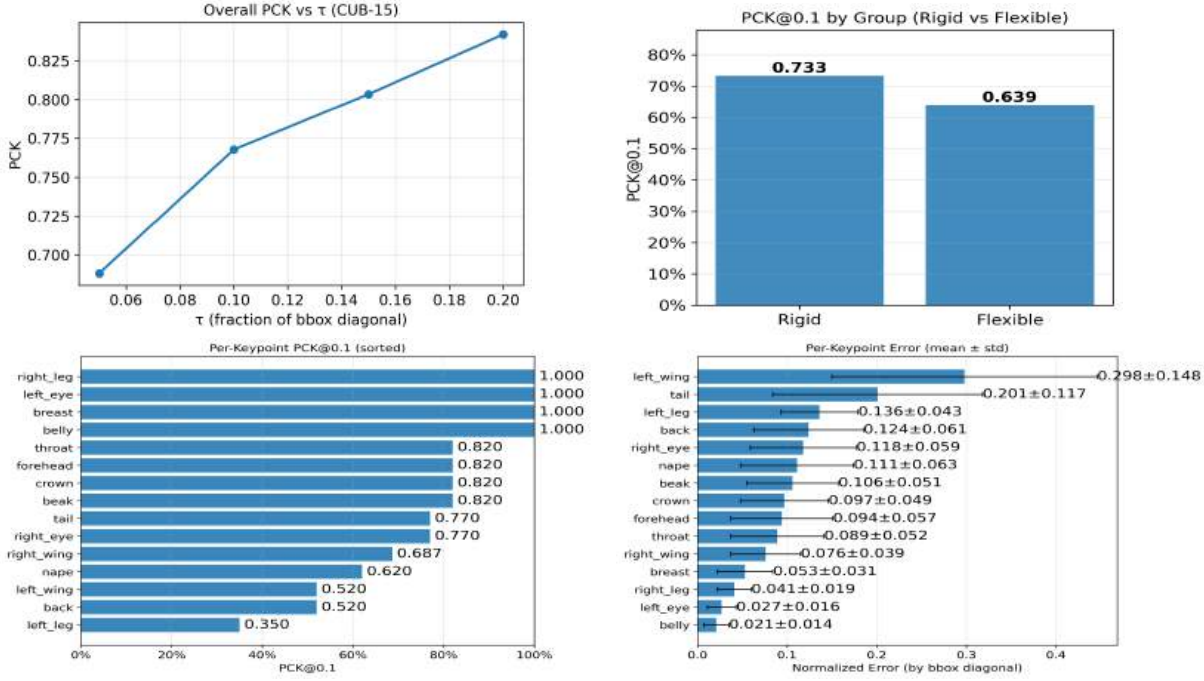
Further details, including the explicit mathematical definitions of all metrics and loss functions, and the screen shot of the MOS WebUI are provided in Appendix A and Appendix B.

3 Results

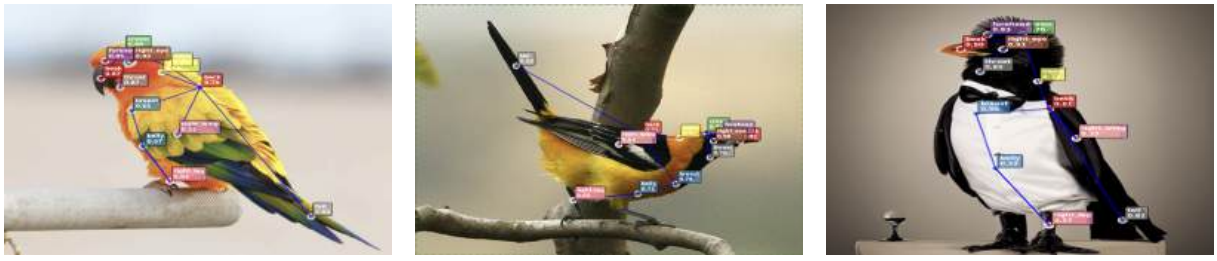
3.1 Keypoint Detection using HRnet

Fig. 4a shows the evaluation metrics of HRNet on the CUB-15 validation set. The PCK–threshold curve increases from PCK@0.05 = 0.689 to PCK@0.20 = 0.840, with PCK@0.1 = 0.733 at the operating point. Per-keypoint PCK@0.1 reveals variation across landmarks: *right leg*, *left eye*, *breast*, and *belly* achieve near-perfect accuracy (≈ 1.0), while *left wing*, *back*, and especially *left leg* perform lowest (≤ 0.5). Mid-range values are observed for *throat*, *forehead*, *crown*, and *beak* (≈ 0.82), as well as *tail* and *right eye* (≈ 0.77). Grouped PCK@0.1 yields 0.733 for rigid landmarks and 0.639 for flexible appendages. The mean ± standard deviation of per-keypoint errors further show the largest dispersions for *left wing* (0.298 ± 0.148) and *tail* (0.201 ± 0.117), whereas *belly*, *left eye*, and *right leg* remain below 0.05.

Fig. 4 presents examples of detected skeletons across bird classes.



(a) HRNet evaluation metrics: (top left) overall PCK curve, (top right) grouped PCK@0.1 (rigid vs flexible), (bottom left) per-keypoint PCK@0.1 ranking, and (bottom right) mean and standard deviation of keypoint errors.



(b) Example frames with detected 15-point skeletons from the generated sequence.

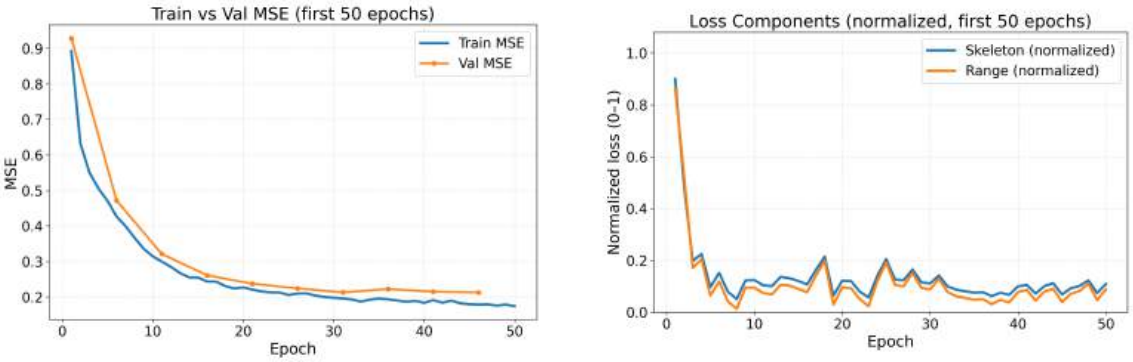
Figure 4: HRNet evaluation and output examples.

3.2 Pose Sequence Generation using MDM

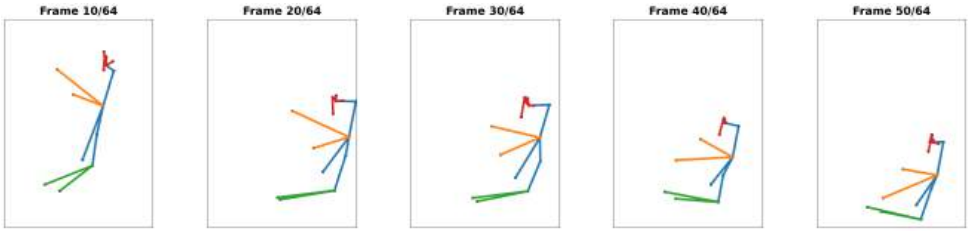
The Motion Diffusion Model (MDM) was trained on 64-frame sequences with six action labels. As shown in Fig. 5a(left), both training and validation MSE decreased steadily over the first 50 epochs. Fig. 5a(right) shows the normalised Skeleton and Range losses, which dropped rapidly before stabilising at low values.

Skeleton loss was computed as the averaged bone-length error, and Range loss as normalised coordinate deviations, both scaled to a comparable 0–1 range.

Representative samples from a generated 64-frame landing sequence are shown in Fig. 5b. The torso pitched upwards while descending, wings flared symmetrically into a braking posture, the tail aligned with the torso to stabilise attitude, and the legs extended forward in preparation for contact. The sequence remained free of bone crossings or frame-level jitter, and individual poses appeared anatomically consistent across time.



(a) Training curves (first 50 epochs): (left) Train vs Val MSE; (right) Skeleton and Range losses.



(b) Sample frames from the generated sequence.

Figure 5: Overall caption combining training curves and sample frames.

3.3 Video Generation using AnimateDiffusion + ControlNet

3.3.1 Qualitative results with CUB-15 skeletons

Generated sequences based on CUB-15 skeletons exhibit temporally coherent motion while preserving both realistic and surreal traits. Across test cases, outputs closely follow the provided skeleton trajectories, with wing flapping, body orientation, and overall flight dynamics remaining consistent over time. This stability is observed not only in naturalistic birds but also in surreal prompts, indicating robustness across different appearance domains.

For Bird 1 (realistic example, Fig 6b), the generated motion adheres tightly to the input skeleton, producing smooth wing beats and consistent body orientation throughout the sequence. For Bird 2 (surreal case, Fig 6c), unnatural colouration and morphology are preserved without disrupting temporal coherence, confirming that the method can maintain both motion fidelity and visual traits under diverse conditions.

The impact of ControlNet conditioning is also visible in parameter variations. Edge inputs preserve global silhouettes, depth maps reinforce volumetric perception, and skeleton conditioning enforces frame-to-frame alignment. Adjusting relative weights reveals clear trade-offs: stronger pose weights increase adherence to the skeleton but reduce fine details such as feather textures; stronger edge weights sharpen contours but may reduce motion flexibility; stronger depth weights enhance body volume but risk suppressing subtle dynamics. Background integration remains largely stable, though high edge weighting occasionally

introduces minor artefacts.

3.3.2 Qualitative results with OpenPose-09 skeletons

Since OpenPose was originally trained on human 18-keypoint skeletons, a clear distribution gap arises when directly applying it to avian motion. To investigate whether closer alignment with human-trained priors could improve controllability, we derived reduced skeletons from CUB-15. OP-5 retained only head and spine landmarks with wing roots, while OP-9 further added shoulder–elbow–wrist segments to mimic the human arm hierarchy (Fig 7a). Leg points and fine head landmarks were excluded due to frequent occlusion and lack of human correspondence. In the following experiments, OP-9 was compared directly with CUB-15.

Overall, OP-9 aligns more stably with OpenPose priors but provides less avian-specific guidance. To reach comparable fidelity with CUB-15, it required stronger edge, depth, and pose weighting. However, increasing edge and depth scales sharpened silhouettes at the expense of motion flexibility, while higher pose scales strengthened skeleton adherence but occasionally introduced artefacts reminiscent of human limbs. By contrast, the CUB-15 skeleton maintained a better balance between appearance fidelity and motion adherence under moderate parameters.

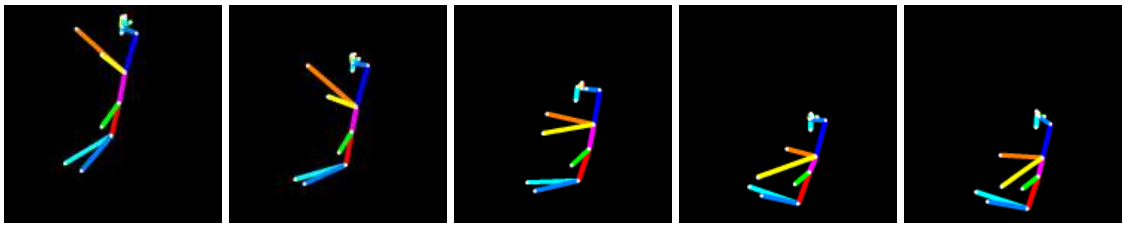
A qualitative comparison on Bird 3 (hover action, Fig 7b) further illustrates these differences. With CUB-15 and moderate ControlNet scales, the generated motion showed natural wing bending, tail spreading, and smooth temporal dynamics (Fig 7c). Using OP-9 with higher scales improved structural alignment but reduced flexibility, blurred wing and tail details, and in some frames introduced anthropomorphic artefacts (Fig 7d). These observations confirm that CUB-15 provides superior guidance for faithful and perceptually convincing video synthesis.

3.3.3 Quantitative evaluation with MOS ratings

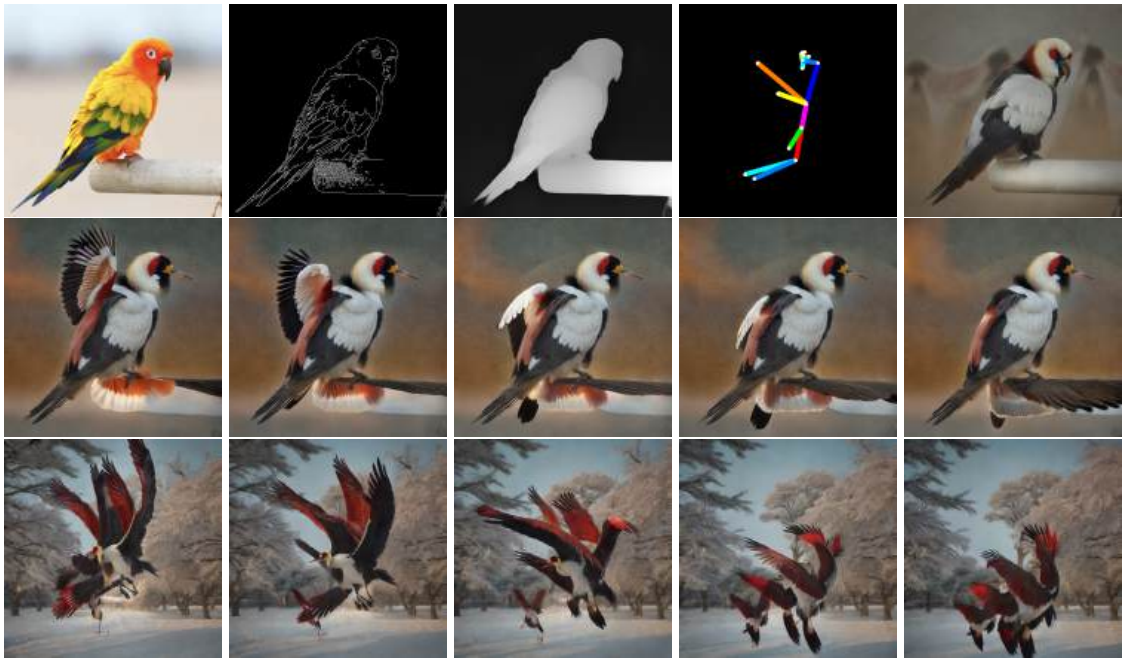
The MOS evaluation involved 18 raters across four tasks (Table 1). Task A arameter sensitivity: The best configuration was found at Edge=0.1, Depth=0.2, Pose=1.2, achieving the highest average rating. Task B Input robustness: Scores were consistent across different birds and contexts, with no single input causing systematic degradation. Task C Label distinguishability: Classification accuracy exceeded random chance, with gliding, landing, and hovering actions recognised by most raters. Task D Skeleton comparison: CUB-15 consistently outperformed OP-9 in subjective preference. The detailed raw scores of the MOS evaluation are provided in Appendix C. The web-based user interface and demonstration videos are available in the GitHub repository.

Task	**Condition**	**MOS**	**Notes**
A: Parameter sensitivity	E=0.1;D=0.2;P=1.2	4.11	Highest quality
	E=0.2;D=0.2;P=1.0	3.11	Lower quality
	E=0.2;D=0.4;P=1.4	4.06	Competitive
	E=0.2;D=0.0;P=1.3	2.61	Weakest
B: Input robustness	Bird1-glide	3.50	Moderate
	Bird2-land	4.17	Best performance
	Bird3-hove	3.94	Stable
C: Label distinguishability	Gliding	77.8%	Above random (16.7%)
	Landing	66.7%	Above random (16.7%)
	Hovering	72.2%	Above random (16.7%)
D: Skeleton comparison	Glide	CUB-15: 3.94 / OP-9: 3.06	CUB better
	Land	CUB-15: 3.72 / OP-9: 3.33	CUB better
	Hove	CUB-15: 4.00 / OP-9: 3.06	CUB better

Table 1: Mean Opinion Score (MOS) results across four evaluation tasks. Scores are averaged over 18 raters. For Task A Condition, E = *Edge scale,* D = *Depth scale,* P = *Pose scale*.



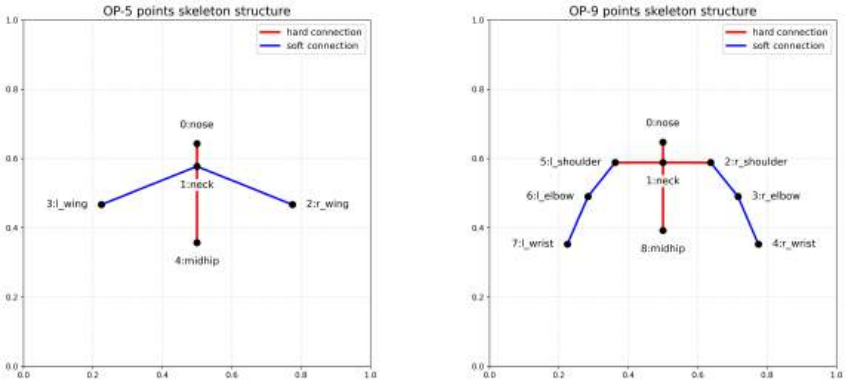
(a) Test skeleton sequence (selected frames) used as motion input.



(b) Results for Bird 1 under the prompt *"A red-crowned crane flapping wings, cinematic lighting"*; the first row shows conditioning inputs (reference image, Canny edges, depth map, OpenPose-style skeleton, generated look using prompt); subsequent columns correspond to different ControlNet weight settings — column 2: edge = 0.2, depth = 0.5, pose = 1.0; column 3: edge = 0.1, depth = 0.2, pose = 1.2.

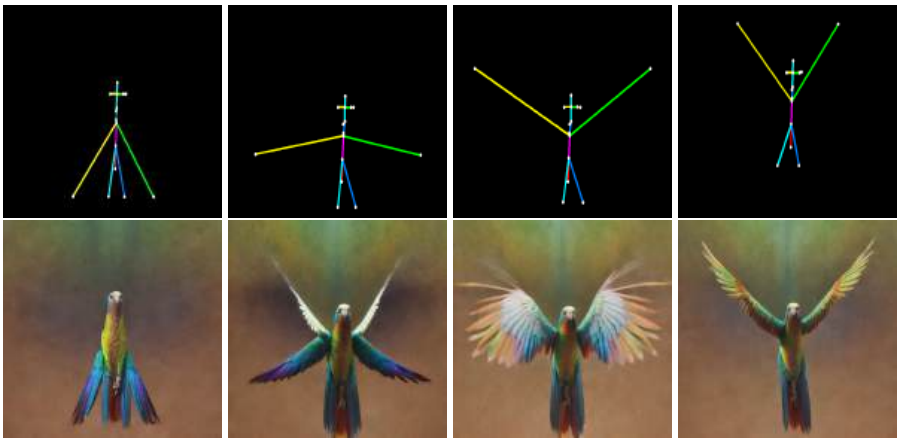
(c) Results for Bird 2 under the prompt *"A surreal purple bird in mid-air; lilac belly, violet back, wings with lavender edges"*; the first row shows conditioning inputs, while the second column illustrates generation with parameters edge = 0.1, depth = 0.2, pose = 1.2.

Figure 6: Representative qualitative results of AnimateDiffusion with CUB15 OpenPose ControlNet.

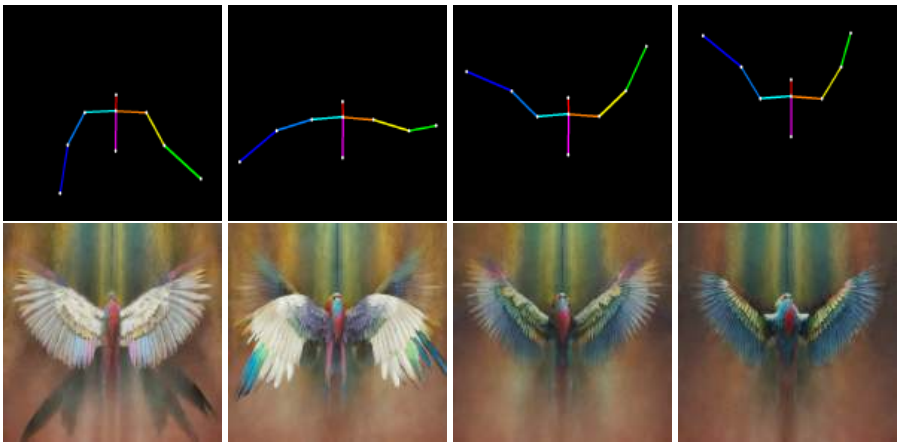


(a) Comparison of OP-5 and OP-9 skeleton structures.

(b) Conditioning inputs for bird 3: reference image, Canny edge map, depth map, and appearance generated by prompt: *"Front-facing parrot mid-air; rainbow plumage, flapping wings, cinematic lighting."*



(c) Results using CUB-15 skeleton with ControlNet weights edge = 0.1, depth = 0.1, pose = 1.0.



(d) Results using OP-9 skeleton with ControlNet weights edge = 0.2, depth = 0.2, pose = 1.5.

Figure 7: Representative qualitative results of AnimateDiffusion with ControlNet.

4 Discussion

4.1 HRNet-based keypoint detection

The HRNet-based detector showed clear variation in accuracy across anatomical landmarks. Rigid parts such as the right leg, breast, and belly achieved high precision, reflecting their consistent visibility and relatively low articulation. By contrast, weaker performance on the left wing, back, and left leg was linked to occlusions and pose variability, a consequence of the CUB-200 dataset bias toward side-standing birds with folded wings. Grouped PCK further confirmed that rigid structures are easier to localise than flexible appendages. These results indicate that wings and legs remain the main sources of error, suggesting the need for targeted augmentation strategies or biomechanical priors. While the detector provides a reliable basis for downstream stages, its weaknesses in localising flexible landmarks constrain the fidelity of subsequent motion generation and rendering.

4.2 MDM-based motion generation

The lower validation loss compared to training can be explained by augmentation and noise injection during training, which increase difficulty relative to the clean validation sequences. This indicates that the model generalises well despite stochastic perturbations. The stable convergence of Skeleton and Range losses further suggests that the MDM successfully enforced anatomical consistency while keeping joint coordinates within valid ranges. The landing motion visualisation also demonstrated that the model captured coordinated dynamics across body parts, such as torso pitching, bilateral wing flaring, and leg extension. Together, these results confirm that the MDM is capable of generating smooth and temporally coherent skeleton trajectories, providing a reliable motion prior for downstream rendering.

4.3 AnimateDiff-based video synthesis

4.3.1 CUB-15 skeletons

CUB-15 skeletons provide bird-specific structural cues that enable strong controllability and a balanced compromise between motion fidelity and appearance quality. The complementary roles of ControlNet branches illustrate that balanced integration is crucial for achieving both visual fidelity and temporal coherence. At the same time, parameter sensitivity highlights clear trade-offs: higher edge and depth weights sharpen silhouettes but suppress dynamic flexibility, while higher pose weights strengthen skeleton adherence but reduce fine details such as feather textures. Although the framework supports both realistic and surreal synthesis, the reliance on manual tuning points to the need for adaptive or learned calibration strategies.

4.3.2 OpenPose-09 skeletons

OP-9 aligns more stably with OpenPose priors but omits lower-body and detailed head landmarks, reducing the amount of avian-specific guidance. To reach comparable fidelity with CUB-15, OP-9 required stronger edge, depth, and pose signals, which introduced undesirable side effects. Stronger edge and depth weights restricted natural motion flexibility, while high pose weights sometimes amplified the human bias inherent in OpenPose training, leading to anthropomorphic artefacts. By contrast, CUB-15 maintained a better balance between appearance and motion under moderate settings. This comparison highlights that while human-trained priors can improve stability, domain-specific skeletons remain more effective, motivating hybrid designs or bird-specific ControlNets in future work.

4.3.3 MOS evaluation

The MOS study confirmed several aspects of controllability. First, parameter sensitivity showed that ControlNet scales directly influenced perceived motion fidelity, with excessive pose weighting leading to human-like bias. Second, robustness across inputs indicated that the pipeline generalised to different

birds and scenes, though minor drifts likely arose from skeleton jitter or background inconsistencies. Third, label distinguishability demonstrated that action labels acted as meaningful control signals, even if visually similar actions (e.g., gliding vs hovering) were sometimes confused. Finally, skeleton comparison underscored the decisive role of skeleton design: domain-specific representations such as CUB-15 provided stronger controllability and higher perceptual quality than reduced OpenPose mappings.

4.4 Overall analysis and future directions

Beyond individual components, several broader implications emerge.

First, dataset limitations restrict overall performance: CUB-200 provides limited coverage of extended wing poses and lacks temporally annotated skeleton sequences, which constrains both training and evaluation. Second, the 2D skeleton representation introduces inherent ambiguities in head–body orientation and depth, highlighting the need for coarse 3D lifting or orientation vectors. Third, while manual parameter tuning demonstrated controllability, it also revealed a lack of automated strategies for balancing ControlNet weights. Fourth, the action label study validated the dual controllability of skeletons and labels, but suggested that additional motion cues such as joint velocity or acceleration may enhance separability.

Overall, the proposed pipeline demonstrates clear strengths: controllable motion generation, temporally coherent skeleton sequences, and the ability to preserve surreal traits. However, limitations remain, including the loss of high-frequency textures, imperfect background integration, and reliance on 2D skeletons. Future work may address these issues by building bird-specific skeleton motion datasets, exploring 3D skeleton lifting, training ControlNets directly on avian data, automating parameter tuning, and integrating temporal transformers for improved long-term coherence.

5 Conclusion

This project investigated whether skeleton-driven generative methods can produce temporally coherent and stylistically consistent bird motion videos. We developed a three-stage pipeline combining HRNet for keypoint detection, a Motion Diffusion Model (MDM) for pose sequence generation, and AnimateDiff with multi-branch ControlNet conditioning for video synthesis. Experiments show that the pipeline generates sequences that follow skeleton trajectories while preserving both realistic and surreal traits, demonstrating controllability and coherence across diverse appearances.

The contributions are threefold. First, we introduced the CUB-15 skeleton representation, which provides bird-specific structural cues and outperforms reduced OpenPose-style variants such as OP-9. Second, we implemented an MDM with action-label conditioning and frame anchoring to generate smooth, anatomically plausible motion sequences. Third, we integrated AnimateDiff with edge, depth, and pose ControlNets, analysing parameter sensitivity and showing that balanced conditioning yields controllable and perceptually convincing results. Objective metrics (PCK, loss curves) and subjective MOS evaluations were combined for a comprehensive assessment.

The pipeline demonstrates controllable motion generation, temporal coherence, and preservation of surreal traits, but limitations remain. The CUB-200 dataset restricts coverage of extended wing poses and continuous sequences. The 2D skeleton introduces depth and orientation ambiguities, and ControlNet weights require manual tuning. High-frequency textures and background integration were also imperfectly preserved.

Future work may expand skeleton motion datasets with annotated sequences, incorporate 3D lifting or temporal transformer architectures, develop bird-specific ControlNets to mitigate human priors, and introduce automated strategies for weight balancing. These improvements would enhance robustness and usability. Beyond technical advances, the framework holds promise for animation, education, and creative media, where controllable and stylised bird motion can enrich both scientific study and artistic expression.

Acknowledgement

Firstly, I would like to thank my supervisors Prof. Christopher Pain and Prof. James Coupe for providing this fascinating project topic and their invaluable support throughout this period. I am also deeply grateful to Dr. Claire Heaney for hosting our weekly meetings and patiently answering all questions—it was always a pleasure to see her in our sessions. Special thanks to Dr. Aniket Joshi and Yueyan Li for all their technical support and guidance.

I would also like to express my heartfelt gratitude to my friends Rachel, Alisa, Ivy, and Shilan for their unwavering support in many different ways, helping me stay mentally grounded during these three months.

Most particularly, a big thank you to Ziqi Yue, another member of the Surreal Bird project. Although we worked on independent projects with different focuses, we spent countless hours meeting and discussing our work together, which made this journey truly enjoyable. I deeply appreciate his collaboration and friendship throughout this project.

AI Acknowledgment

- **Tool Name and Version**: ChatGPT (4o/5); Claude Sonnet 4
- **Publisher/Provider**: OpenAI; Anthropic
- **URL**: https://chatgpt.com/?model=gpt-4o; https://chatgpt.com/?model=gpt-5; https://claude.ai
- **Usage Description**: Generative AI tools were used in the initial research phase to analyze project requirements, understand AI frameworks by clarifying documentation and simplifying explanations, provide guidance on environment setup and resolving module issues during HPC deployment, and assist with debugging code throughout the development process.
- **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] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," Tech. Rep. CNS-TR-2011-001, Caltech, 2011.

[2] Z. Ge, C. McCool, C. Sanderson, P. Wang, L. Liu, I. Reid, and P. Corke, "Exploiting temporal information for dcnn-based fine-grained object classification," in *Proceedings of the International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pp. 1–6, IEEE, 2016.

[3] Z. Ge, C. McCool, C. Sanderson, P. Wang, L. Liu, I. Reid, and P. Corke, "Vb100 bird dataset (video and audio)," 2016.

[4] T. Yu, Y. Xu, X. Peng, Y. Wu, and L. Chen, "Ap-10k: A benchmark for animal pose estimation in the wild," *arXiv preprint arXiv:2108.12617*, 2021.

[5] J. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Cross-domain adaptation for animal pose estimation," in *ICCV*, 2019.

[6] A. Mathis, P. Mamidanna, K. M. Cury, T. Abe, V. N. Murthy, M. W. Mathis, and M. Bethge, "Deeplabcut: markerless pose estimation of user-defined body parts with deep learning," *Nature Neuroscience*, vol. 21, no. 9, pp. 1281–1289, 2018.

[7] T. D. Pereira *et al.*, "Sleap: A deep learning system for multi-animal pose tracking," *Nature Methods*, vol. 17, pp. 1–9, 2020.

[8] H. Naik, A. H. H. Chan, J. Yang, M. Delacoux, I. D. Couzin, F. Kano, and M. Nagy, "3d-pop: An automated annotation approach to facilitate markerless 2d-3d tracking of freely moving birds with marker-based motion capture," in *CVPR*, 2023.

[9] U. Waldmann, A. H. H. Chan, H. Naik, M. Nagy, I. D. Couzin, O. Deussen, B. Goldluecke, and F. Kano, "3d-muppet: 3d multi-pigeon pose estimation and tracking," *International Journal of Computer Vision*, 2024.

[10] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017.

[11] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022.

[12] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *ICCV*, 2023.

[13] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine-tuning text-to-image diffusion models for subject-driven generation," in *CVPR*, 2023.

[14] R. Gal *et al.*, "An image is worth one word: Personalizing text-to-image generation using textual inversion," *arXiv preprint arXiv:2208.01618*, 2022.

[15] H. Ye *et al.*, "Ip-adapter: Text-compatible image prompt adapter for text-to-image diffusion models," *arXiv preprint arXiv:2308.06721*, 2023.

[16] J. Li *et al.*, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597*, 2023.

[17] H. Liu *et al.*, "Visual instruction tuning," *NeurIPS*, 2023.

[18] J.-B. Alayrac *et al.*, "Flamingo: a visual language model for few-shot learning," *arXiv preprint arXiv:2204.14198*, 2022.

[19] A. Blattmann *et al.*, "Stable video diffusion: Scaling latent video diffusion models to large datasets," *arXiv preprint arXiv:2311.15127*, 2023.

[20] Y. Guo *et al.*, "Animatediff: Animate your personalized text-to-image diffusion models without specific tuning," *arXiv preprint arXiv:2307.04725*, 2023.

[21] X. Wang, H. Yuan, S. Zhang, D. Chen, J. Wang, Y. Zhang, Y. Shen, D. Zhao, and J. Zhou, "Videocomposer: Compositional video synthesis with motion controllability," *arXiv preprint arXiv:2306.02018*, 2023. Version 2, last revised 6 Jun 2023.

[22] J. Xing *et al.*, "Dynamicrafter: Animating open-domain images with video diffusion priors," in *ECCV*, 2024.

[23] G. Tevet, S. Raab, Y. Lipman, A. H. Bermano, and D. Cohen-Or, "Human motion diffusion model," *arXiv preprint arXiv:2209.14916*, 2022.

[24] G. Tevet *et al.*, "Motionclip: Exposing human motion generation to clip space," *arXiv preprint arXiv:2203.08063*, 2022.

[25] W. Yan *et al.*, "Temporally consistent transformers for video generation," in *ICML*, 2023.

[26] T. Karras *et al.*, "Dreampose: Fashion image-to-video synthesis via stable diffusion," in *ICCV*, 2023.

[27] M. Team, "Musepose: A pose-driven image-to-video framework for virtual human," 2024.

[28] M. Andriluka *et al.*, "2d human pose estimation: New benchmark and state of the art analysis," in *CVPR*, 2014.

[29] C. Ionescu *et al.*, "Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *TPAMI*, 2014.

[30] M. Heusel *et al.*, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *NeurIPS*, 2017.

[31] R. Zhang *et al.*, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018.

[32] Z. Huang *et al.*, "Vbench: Comprehensive benchmark suite for video generative models," in *CVPR*, 2024.

A Evaluation Metric Definitions

This appendix provides the formal definitions of all metrics and loss functions referenced in Section 2.6.

A.1 Pose Detection

Normalised point error. For keypoint i with ground-truth bounding box (w, h),

$$d_i = \frac{\|\hat{\mathbf{p}}_i - \mathbf{p}_i\|_2}{\sqrt{w^2 + h^2}}.$$

PCK (Percentage of Correct Keypoints). At threshold τ,

$$\text{PCK@}\tau = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[d_i \le \tau]\,.$$

MPJPE (Mean Per-Joint Position Error).

$$\text{MPJPE} = \frac{1}{N}\sum_{i=1}^{N} \|\hat{\mathbf{p}}_i - \mathbf{p}_i\|_2.$$

Left–right confusion. For symmetric pair (ℓ, r),

$$\text{Confusion} = \mathbb{1}\Big[\|\hat{\mathbf{p}}_\ell - \mathbf{p}_r\|_2 < \kappa\sqrt{w^2+h^2} \;\wedge\; \|\hat{\mathbf{p}}_r - \mathbf{p}_\ell\|_2 < \kappa\sqrt{w^2+h^2}\Big]\,,$$

with $\kappa = 0.1$.

A.2 Motion Generation

Noise-prediction MSE (diffusion objective).

$$\mathcal{L}_{\text{mse}} = \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\varepsilon}}\big[\,\|\hat{\boldsymbol{\varepsilon}}_\theta(\mathbf{x}_t, t) - \boldsymbol{\varepsilon}\|_2^2\,\big].$$

Skeleton-length consistency.

$$\mathcal{L}_{\text{skel}} = \frac{1}{|\mathcal{E}|}\sum_{(i,j)\in\mathcal{E}} \text{MSE}\Big(\big\|\hat{\mathbf{x}}_{0,i} - \hat{\mathbf{x}}_{0,j}\big\|_2, \big\|\mathbf{x}_i - \mathbf{x}_j\big\|_2\Big).$$

Coordinate-range regulariser.

$$\mathcal{L}_{\text{range}} = \mathbb{E}\Big[\,\big\|\max(\mathbf{0},\, \mathbf{x} - M)\big\|_2^2 + \big\|\max(\mathbf{0},\, m - \mathbf{x})\big\|_2^2\Big]\,.$$

First-frame anchoring.

$$\mathcal{L}_{\text{first}} = \frac{1}{15}\sum_{i=1}^{15} \big\|\hat{\mathbf{x}}_{0,i}^{(1)} - \mathbf{x}_{\text{init},i}\big\|_2^2.$$

Total training loss.

$$\mathcal{L} = \mathcal{L}_{\text{mse}} + \lambda_{\text{skel}}\mathcal{L}_{\text{skel}} + \lambda_{\text{range}}\mathcal{L}_{\text{range}} + \lambda_{\text{first}}\mathcal{L}_{\text{first}}.$$

A.3 Video Rendering

Mean Opinion Score (MOS). For criterion k with R raters and C clips,

$$\mathrm{MOS}_k = \frac{1}{RC} \sum_{c=1}^{C} \sum_{r=1}^{R} s_{r,c}^{(k)}.$$

B MOS Evaluation Web-based User Interface

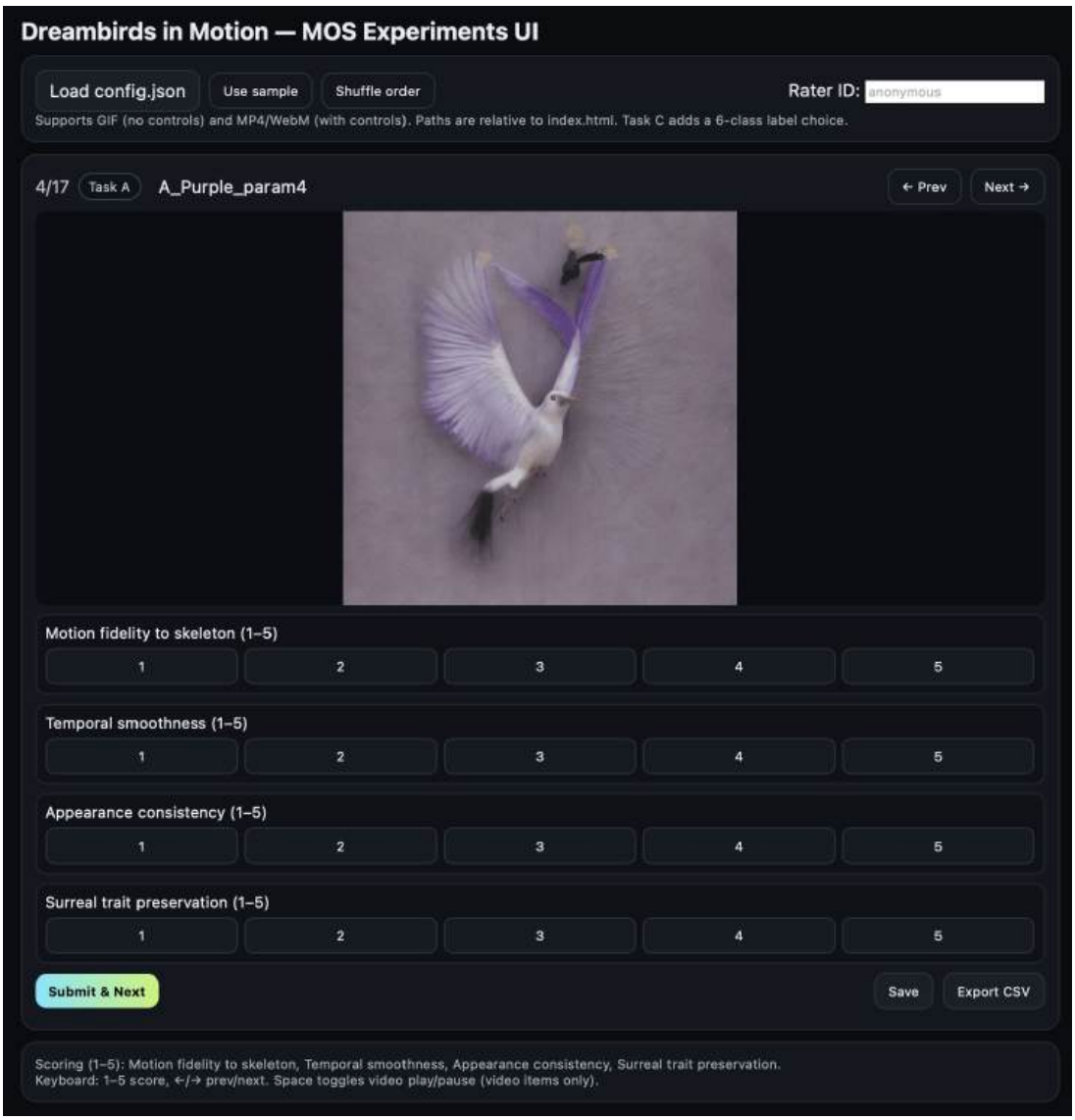


Figure 8: Screen shot of the MOS WebUI. The interface collects human ratings on motion fidelity, temporal smoothness, appearance consistency, and surreal trait preservation for generated bird videos.

C MOS Evaluation Results

Task A: Parameter sensitivity					Task B: Input robustness				
Rater	Param1	Param2	Param3	Param4	Rater	B_Black	B_Purple	B_Rainbow	B_Red
rater01	5	3	4	2	rater01	4	5	2	3
rater02	4	5	4	3	rater02	3	3	4	2
rater03	5	2	2	4	rater03	3	4	5	3
rater04	4	3	5	2	rater04	3	5	4	4
rater05	5	3	4	3	rater05	4	4	4	2
rater06	3	4	4	2	rater06	3	5	4	3
rater07	5	2	4	3	rater07	5	4	5	2
rater08	4	4	5	2	rater08	3	4	4	3
rater09	5	3	4	5	rater09	2	5	3	3
rater10	2	2	4	2	rater10	3	4	5	3
rater11	3	3	3	3	rater11	4	5	4	3
rater12	4	3	4	2	rater12	3	4	3	2
rater13	5	2	4	3	rater13	4	5	4	3
rater14	4	4	5	1	rater14	3	4	4	2
rater15	3	3	4	2	rater15	5	2	5	5
rater16	4	2	4	3	rater16	4	3	4	2
rater17	5	5	5	2	rater17	4	5	3	3
rater18	4	3	4	3	rater18	3	4	4	2
Mean	4.11	3.11	4.06	2.61	Mean	3.50	4.17	3.94	2.78

Table 2: MOS raw scores for Task A (parameter sensitivity) and Task B (input robustness).

Task C: Label distinguishability				Task D: Skeleton comparison						
Rater	C_Gliding	C_Landing	C_hovering	Rater	Black_CUB15	Black_OP9	Purple_CUB15	Purple_OP9	Rainbow_CUB15	Rainbow_OP9
rater01	gliding	landing	hovering	rater01	4	3	4	3	4	4
rater02	gliding	landing	hovering	rater02	3	1	2	3	4	3
rater03	gliding	landing	hovering	rater03	4	3	4	5	3	3
rater04	gliding	gliding	gliding	rater04	5	2	5	3	5	4
rater05	gliding	landing	hovering	rater05	2	4	4	4	4	3
rater06	gliding	gliding	hovering	rater06	4	3	5	1	4	2
rater07	gliding	landing	hovering	rater07	5	4	3	4	5	4
rater08	hovering	landing	hovering	rater08	4	3	4	4	4	1
rater09	gliding	landing	hovering	rater09	2	3	4	1	4	1
rater10	soaring	landing	hovering	rater10	4	1	3	4	4	3
rater11	gliding	takeoff	gliding	rater11	5	4	5	5	5	4
rater12	gliding	landing	hovering	rater12	4	3	2	4	3	4
rater13	landing	landing	hovering	rater13	4	4	4	3	5	3
rater14	gliding	hovering	soaring	rater14	3	3	4	4	4	4
rater15	gliding	landing	soaring	rater15	5	4	4	4	3	3
rater16	gliding	takeoff	hovering	rater16	4	4	1	2	4	2
rater17	hovering	gliding	gliding	rater17	5	2	5	3	4	4
rater18	gliding	landing	hovering	rater18	4	4	4	3	3	3
Correctness	77.8%	66.7%	72.2%	Mean	3.94	3.06	3.72	3.33	4.00	3.06

Table 3: MOS raw scores for Task C (label distinguishability) and Task D (skeleton comparison).