Generating the Wild: Individual-Consistent Image-to-Video Generation for Wildlife

Yuzhuo Li **Di Zhao** **Xinyu Zhang** **Daniel Wilson** **Yun Sing Koh***
School of Computer Science, University of Auckland
Auckland, New Zealand
yil708@aucklanduni.ac.nz
{di.zhao,xinyu.zhang,daniel.wilson,y.koh}@auckland.ac.nz

Abstract

Individual-level wildlife identification often suffers from data scarcity, as varying observations of the same animal under diverse poses, viewpoints, and motions are rarely available. Image-to-video (I2V) generation offers a promising way to mitigate this limitation by synthesizing additional observations from a single reference image. However, existing I2V models mainly emphasize global layout, semantics, and motion, and therefore often fail to preserve fine-grained local appearance cues that distinguish one wildlife individual from another, such as fur texture, stripe boundaries, spot configurations, and contour transitions. We observe that these identity-critical cues are closely related to high-frequency information, which is essential for distinguishing individuals. To address this challenge, we propose **WildIcon**, a high-frequency-guided I2V framework for wildlife individual consistency. Specifically, WildIcon introduces a frequency-aware identity encoding branch that extracts individual-specific high-frequency cues from the reference image. Combined with isolated foreground information, the resulting identity tokens are then injected into cross-attention blocks as identity conditioning. Building on a frozen backbone with lightweight identity adaptation, WildIcon preserves fine-grained identity cues visible in the reference image while retaining the motion controllability and semantic fidelity of the base I2V model. In addition, to support the training and evaluation of wildlife individual-consistent I2V, we construct **WildlifeVid**, a wildlife-centric video dataset with high-quality, temporally coherent clips and individual-level identity labels. Experiments on I2V generation and downstream animal re-identification (ReID) show that WildIcon achieves stronger individual consistency than existing baselines, and that its filtered outputs can serve as useful candidate training augmentations for downstream ReID. Project page: https://ml-4-socialgood.github.io/WildIcon/.

1 Introduction

Wildlife conservation is essential for maintaining ecological balance, protecting endangered habitats, and sustaining essential ecosystem services [1; 2]. A key requirement in these efforts is individual-level identification for monitoring population dynamics, movement patterns, and individual condition in a scalable and non-invasive way [3; 4]. However, reliable wildlife identification remains fundamentally constrained by data scarcity, since repeated observations of the same individual under diverse poses, viewpoints, and environmental conditions are often unavailable [5; 6; 7].

To mitigate this limitation, prior work has explored data enrichment strategies, including data augmentation techniques (*e.g.,* cropping, flipping, and color jittering) [8; 9] and synthetic image

*Corresponding author.

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

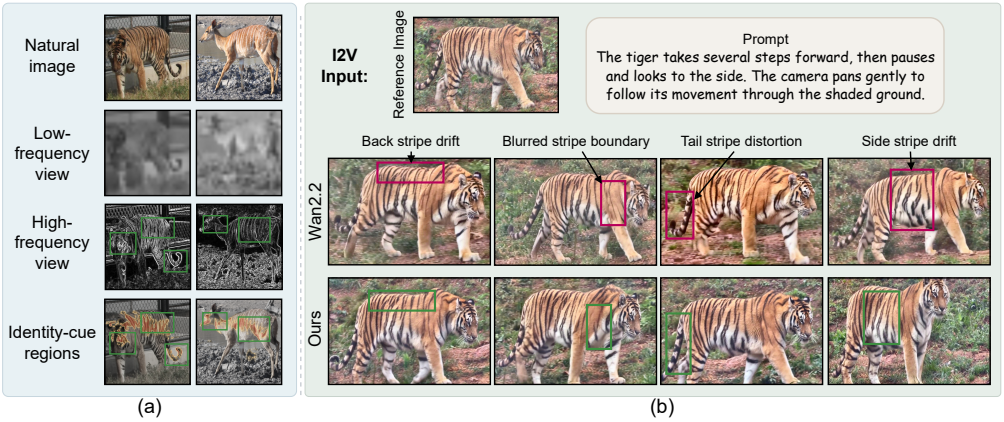


Figure 1: **Frequency-aware identity preservation in wildlife I2V generation.** (a) Low-frequency views preserve coarse body structure but suppress fine-grained identity cues, whereas high-frequency views reveal local identity-discriminative details. (b) Given the same reference image and prompt, Wan2.2 blurs or alters identity-critical stripe patterns across frames, while WildIcon better preserves these fine-grained cues for same-individual consistency.

generation [10; 11]. While conventional data augmentations can improve model robustness against moderate changes in scale, illumination and color, they remain limited to transformations of already observed samples and cannot synthesize genuinely new poses or complex out-of-plane viewpoint changes [12; 13; 14]. Image generation models can produce more diverse appearances and poses, but they operate on isolated frames without explicit temporal or geometric constraints, and therefore often fail to preserve subject identity under large viewpoint changes [15; 16; 17]. Image-to-video (I2V) generation is therefore a more promising direction: given a reference image, it can synthesize additional observations of the same individual under new motions, poses, and views [18; 19; 20; 21].

Despite this promise, existing I2V models are not designed for individual consistency in wildlife. Although they can generate realistic motion and preserve coarse subject appearance, maintaining consistency with the reference animal across frames requires preserving subtle local appearance cues that remain discriminative under changes in pose, viewpoint, and background. In wildlife, these cues often take the form of fine-grained visual patterns, such as fur or scale texture, stripe boundaries, spot configurations, and contour transitions [6; 22]. To better understand which visual information is most important for distinguishing individuals, we analyze the role of frequency components in wildlife identity. As shown in Figure 1, these fine-grained patterns are more closely related to high-frequency image information, which captures rapid spatial changes and fine visual structures [23; 24]. In contrast, low-frequency views mainly retain global content while suppressing many identity-critical details. This observation reveals a mismatch between the requirements of wildlife identity preservation and the modeling priorities of modern I2V models: while they are effective at modeling global layout, semantic content and temporal motion, they do not explicitly preserve the sparse local high-frequency details critical for wildlife identity [25]. As a result, they may generate visually plausible videos while still drifting in identity-critical appearance from one individual to another.

Motivated by this observation, we propose **WildIcon** (Sec. 4), a high-frequency-guided I2V framework tailored to wildlife individual consistency. WildIcon introduces a frequency-aware identity encoding branch that explicitly extracts foreground-filtered high-frequency cues from the reference image and encodes them together with foreground appearance information into identity tokens for identity-aware conditioning. These identity tokens are injected into selected cross-attention layers, enabling I2V models to preserve fine-grained individual identity, while the remaining blocks are kept frozen to retain the inherent motion controllability and semantic fidelity of the pretrained backbone. To enable training and evaluation, we further construct **WildlifeVid**, a wildlife-centric video dataset curated through a scalable pipeline (Sec. 3) that provides high-quality, temporally coherent clips and individual identity labels under realistic variation in viewpoint, motion, illumination and background. Together, WildIcon and WildlifeVid provide a modeling and data foundation for wildlife individual-consistent I2V generation.

We validate this framework on both wildlife I2V generation and downstream animal ReID. Experiments show that WildIcon outperforms existing baselines in individual consistency and downstream animal ReID performance by producing reliable, identity-aware synthetic videos. In summary, our main contributions are: (1) We construct WildlifeVid, a wildlife-centric video dataset comprising single-subject wildlife clips with identity labels and standardized prompts for training and evaluating individual-consistent I2V models. (2) We propose WildIcon, a high-frequency-guided I2V framework that makes sparse wildlife identity patterns explicit conditioning signals, enabling I2V models to preserve individual-specific local cues while maintaining motion controllability and semantic fidelity. (3) We demonstrate that wildlife individual-consistent I2V benefits downstream animal ReID: by converting limited real observations into additional individual-consistent training samples, WildIcon helps mitigate the data scarcity that limits individual-level wildlife identification.

2 Related Work

Image-to-video generation. Image-to-video (I2V) generation extends video foundation models by conditioning generation on a reference image and, optionally, a text prompt. Recent methods have explored a variety of conditioning strategies to inject visual guidance into pretrained video generators. For example, DynamiCrafter [26] incorporates image embeddings through cross-attention, while Stable Video Diffusion [27] and CogVideoX [28] concatenate conditional latents with noisy video latents. Wan [29] further adopts mask-guided conditioning and decoupled cross-attention to support high-quality I2V generation. While these methods improve controllability over text-only generation, maintaining fine-grained individual consistency remains challenging. For wildlife, identity cues are often sparse and closely tied to high-frequency structures, coarse reference-image conditioning is insufficient for individual consistency. Although DiT-based I2V backbones model global structure and motion well, they lack explicit mechanisms for preserving local identity-critical details, leading to appearance drift in generated wildlife videos.

Identity-preserving video generation. Beyond general I2V generation, recent work has studied subject-consistent and identity-preserving video generation, where the goal is to maintain a reference subject across generated frames while following text instructions. Subject-driven methods improve consistency through reference-image injection or cross-modal text-image-video alignment [30; 31; 32; 33], while human-centered identity-preserving methods use facial representations, landmarks, or frequency-decomposed facial cues to preserve person identity [25; 34]. Their conditioning mechanisms are mainly designed for general subject appearance or human facial identity, and therefore do not directly transfer to wildlife, where identity is often determined by sparse high-frequency patterns such as texture, stripes, spots, and contour transitions. WildIcon instead conditions generation on foreground-filtered high-frequency cues to better maintain fine-grained identity-relevant patterns visible in the reference.

Wildlife individual identity modeling. High-frequency information captures rapid spatial changes and fine visual structures, which in wildlife often correspond to individual-discriminative cues such as fur or scale texture, stripe boundaries, spot configurations, and contour details. AdaFreq [22] supports this view by showing that high-frequency cues provide an effective cross-species representation for wildlife ReID. While some recent generative efforts (*e.g.*, Syn-Tiger-360 [16]) further show that preserving cross-view biometric topologies improves downstream ReID, generalizable identity-preserving I2V generation for diverse wildlife remains underexplored. WildIcon extends wildlife identity modeling to this generative setting by using foreground-filtered high-frequency cues as conditioning signals for maintaining reference-supported individual appearance over time.

3 WildlifeVid Dataset

Developing wildlife individual-consistent I2V models requires wildlife videos with individual-level identity annotations, because these labels indicate which clips correspond to the same wildlife individual and support learning the visual cues that distinguish it from others of the same species. However, existing wildlife video datasets are primarily annotated for species, action, or behavior recognition, providing limited support for individual identity modeling. To address this gap, we present WildlifeVid, a wildlife-centric video dataset that provides individual identity labels, curated via a scalable pipeline for wildlife-consistent I2V generation. WildlifeVid is constructed by curating and re-annotating clips from four publicly accessible wildlife video sources: Animal-in-Motion

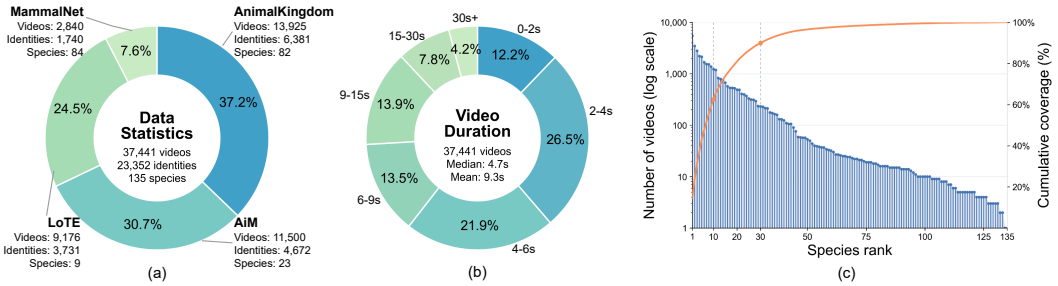


Figure 2: **Overview statistics of WildlifeVid.** (a) Composition of WildlifeVid across the four public data sources, showing the number of videos, identities, and species in each source. (b) Distribution of video durations. (c) Species distribution ranked by the number of videos, with cumulative coverage.

(AiM) [35], AnimalKingdom [36], MammalNet [37], and LoTE-Animal [38]. Figure 2 summarizes the dataset statistics.

3.1 Data Curation Pipeline

Since identity labels are assigned at the clip level, each retained clip should contain one dominant wildlife individual without identity changes over time. We therefore curate raw videos into single-animal temporal clips, retaining complete videos when possible and splitting or filtering videos only when identity changes. Each retained clip is paired with a standardized prompt and an individual identity label for individual-consistent I2V training.

Animal-centric clip extraction. Raw wildlife videos often contain shot changes, frames without a visible animal, or multiple animals in the same view, making them unsuitable for identity annotation. We therefore first split each source video into candidate clips based on shot changes using PySceneDetect [39], which detects large weighted average pixel changes across adjacent frames. We then remove clips that do not clearly show the target species using CLIP-based species filtering [40; 41]. Finally, we enforce a strict single-subject constraint on the remaining clips by retaining only clips with exactly one detected animal throughout the clip and discarding clips with zero or multiple detected animals.

Visibility and temporal stability filtering. After obtaining candidate single-subject clips, we further filter them to ensure that each clip can be assigned a reliable individual identity label. We sample frames from each candidate clip and examine the detected wildlife individual over time. We remove clips in which the animal is missing, the bounding box shows large changes in location or scale across adjacent frames, or the visible individual may change within the clip. Finally, we discard clips in which the wildlife's body is only partially visible within the frame, since incomplete bodies provide unreliable shape and appearance cues for identity learning.

Prompt generation. Text prompts provide controllable conditions for wildlife I2V training, guiding the model to follow the desired motion, viewpoint, and scene context. For each retained clip, we sample representative frames and use a multimodal captioning model to generate an initial caption. We then convert the raw caption into a standardized prompt by removing generic template text (*e.g.*, framing or style phrases that do not describe the visible content), unsupported inferences (*e.g.*, guessed behavior, intent, or attributes not clearly visible), repeated phrases, and content unrelated to the visible animal or motion. The final prompt (*e.g.*, "a tiger walks slowly through grass in a side view") is a concise sentence describing the visible content of the clip.

Identity annotation. We assign identity labels at the clip level by treating each retained clip as a candidate identity unit. For each clip, we use the temporal midpoint frame as its representative frame and extract a visual embedding using DINOv3 [42]. Candidate clips are grouped within the same species based on embedding similarity, which reduces cross-species mismatches and makes manual verification more tractable. High-similarity candidates are then manually verified, and clips confirmed to contain the same animal are assigned the same identity label.

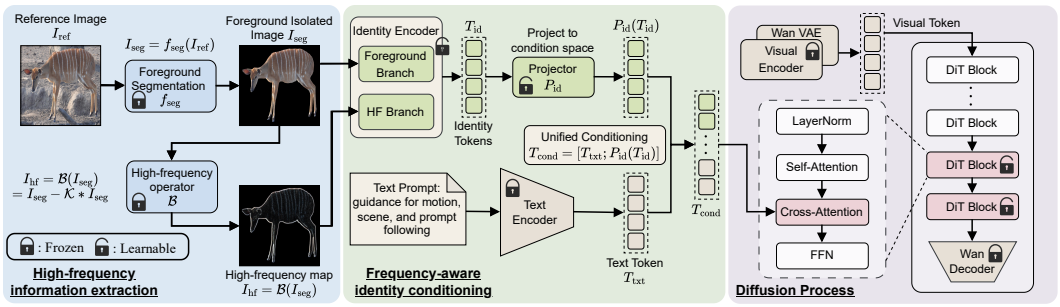


Figure 3: WildIcon augments a frozen Wan I2V backbone with a foreground-filtered high-frequency identity branch. Given a reference image I_{ref}, the model first obtains a foreground-isolated image I_{seg} and its high-frequency map $I_{\text{hf}} = \mathcal{B}(I_{\text{seg}})$. These complementary cues are encoded into identity tokens T_{id}, projected and fused with text tokens to form a unified condition sequence, and injected into selected middle and late DiT blocks through native cross-attention.

4 Methodology

In this section, we propose **WildIcon**, a high-frequency-guided I2V framework for individual-consistent wildlife generation. Given a reference image I_{ref} of a known wildlife individual and a text prompt y describing the desired motion or scene context, our goal is to generate a video $V = \{I_t\}_{t=1}^{T}$ consisting of T frames that remains temporally coherent, follows the prompt, and preserves the individual's fine-grained identity throughout the sequence. As illustrated in Figure 3, WildIcon consists of a frequency-aware identity encoder, an identity-conditioning module, and a DiT-based video backbone. Built on Wan2.2, WildIcon strengthens identity preservation by extracting foreground-filtered high-frequency cues from the reference image and injecting the resulting identity tokens into selected DiT blocks.

4.1 Preliminary

Our method is built on Wan2.2-I2V-A14B [29], a pretrained image-to-video generation framework, consisting of a variational autoencoder (VAE) [43], a transformer-based backbone [44] and a text encoder (*i.e.*, umt5-xxl [45]). Given a training video V, the VAE encoder maps it into a temporally and spatially compressed latent representation $\mathbf{z}_0$, which is used as the clean video latent in diffusion training. In Wan2.2, the transformer-based velocity prediction network Ψ_θ is trained with Flow Matching (FM) [46]. Specifically, given $\mathbf{z}_0$ and Gaussian noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, a noisy latent $\mathbf{z}_\sigma$ at noise level $\sigma \in [0, 1]$ is constructed as $\mathbf{z}_\sigma = (1 - \sigma)\mathbf{z}_0 + \sigma\epsilon$. The model is optimized to predict the target velocity with the condition c, which is formulated as:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{\mathbf{z}_0, \epsilon, \sigma, c}\left[\|\Psi_\theta(\mathbf{z}_\sigma, \sigma, c) - (\epsilon - \mathbf{z}_0)\|_2^2\right], \tag{1}$$

where c represents the conditioning features derived from the given text prompt y and the reference image I_{ref}.

4.2 Frequency-Aware Identity Encoding

Wildlife identity is often determined by fine-grained local appearance cues, such as fur texture, stripe boundaries, spot configurations, and sharp contour transitions. However, directly extracting high-frequency responses from the raw reference image is unreliable in natural environments, where background regions such as grass, leaves, shadows, and clutter can also produce strong high-frequency signals. To focus the identity representation on subject-centric appearance, we first isolate the wildlife foreground and then compute its corresponding high-frequency residual map. Specifically, given a reference image I_{ref}, we apply a foreground extraction operator $f_{\text{seg}}(\cdot)$ on I_{ref}, followed by a spatial high-pass operator $\mathcal{B}(\cdot)$ with a Gaussian smoothing kernel $\mathcal{K}$, to obtain the high-frequency residual map I_{hf}. The formulation is as follows:

$$I_{\text{seg}} = f_{\text{seg}}(I_{\text{ref}}), \quad I_{\text{hf}} = \mathcal{B}(I_{\text{seg}}) = I_{\text{seg}} - \mathcal{K} * I_{\text{seg}}. \tag{2}$$

Using I_{hf} alone for identity encoding is insufficient, because the residual map discards much of the foreground context and provides limited spatial support for interpreting the extracted high-frequency responses. We therefore use I_{seg} and I_{hf} as complementary identity cues: I_{seg} preserves foreground semantics, body structure, and spatial context, while I_{hf} emphasizes local texture changes and fine residual patterns. Since I_{seg} remains close to the natural image distribution while I_{hf} follows residual image statistics, we adopt a dual-branch identity encoder:

$$T_{\mathrm{id}} = \mathcal{G}_{\mathrm{id}}(I_{\mathrm{seg}}, I_{\mathrm{hf}}) = \mathcal{F}(\mathcal{E}_{\mathrm{vis}}(I_{\mathrm{seg}}) \oplus \mathcal{E}_{\mathrm{hf}}(I_{\mathrm{hf}})), \tag{3}$$

where $\mathcal{E}_{\mathrm{vis}}$ is a frozen DINO encoder [42] for foreground visual tokens, $\mathcal{E}_{\mathrm{hf}}$ is a lightweight trainable CNN encoder for high-frequency texture tokens. $\oplus$ denotes feature-wise concatenation along the channel dimension after token alignment, and $\mathcal{F}(\cdot)$ denotes a two-layer MLP-based fusion head with normalization.

4.3 Identity Conditioning

After obtaining the identity tokens T_{id}, we inject them into the frozen Wan I2V backbone to strengthen individual identity preservation during generation. Our design follows a simple principle: the native Wan model should remain responsible for coarse appearance, structural layout, motion dynamics, and prompt following, while the additional identity branch should provide complementary fine-grained cues that are otherwise difficult to preserve. We therefore retain the original image-conditioned generation path of Wan and augment the conditioning space of selected DiT blocks with identity-aware tokens derived from the reference image.

Given a text prompt y, the frozen text encoder produces text tokens T_{txt} that provide motion, scene, and semantic guidance, while the identity tokens T_{id} summarize fine-grained wildlife-specific appearance cues. We then project the identity tokens into the Wan text-conditioning feature space and concatenate them with the text tokens, as in Eq. (4). Here, $P_{\mathrm{id}}(\cdot)$ is a token-wise two-layer MLP projector, applied to the token feature dimension, and $[\cdot;\cdot]$ denotes token-wise concatenation.

$$T_{\mathrm{txt}} = \mathrm{TextEncoder}(y), \quad T_{\mathrm{cond}} = [\, T_{\mathrm{txt}};\, P_{\mathrm{id}}(T_{\mathrm{id}})\,] \tag{4}$$

Here, we inject T_{cond} only into a selected set $\mathcal{S}$ of middle and late DiT blocks, since early blocks mainly establish coarse layout and temporal structure, whereas middle and late blocks are better suited for refining appearance details and identity-sensitive textures. For each selected block $l \in \mathcal{S}$, let $X^{(l)}$ denote the incoming visual tokens. We update them through the native cross-attention layer:

$$\tilde{X}^{(l)} = X^{(l)} + \mathrm{Attn}^{(l)}\Big(X^{(l)}, T_{\mathrm{cond}}\Big), \quad l \in \mathcal{S}, \tag{5}$$

where $\mathrm{Attn}^{(l)}(\cdot,\cdot)$ denotes the cross-attention operation of the l-th DiT block, with visual tokens as queries and T_{cond} as keys and values. For blocks $l \notin \mathcal{S}$, the original Wan conditioning unchanged.

4.4 Optimization

WildIcon is trained with the FM objective in Eq. (1). To further encourage individual-level appearance preservation, we introduce an additional identity loss. Specifically, given the predicted velocity, we estimate the clean latent as $\hat{\mathbf{z}}_0 = \mathbf{z}_\sigma - \sigma\Psi_\theta(\mathbf{z}_\sigma, \sigma, c)$ and decode it into M sampled predicted frames $\{\hat{I}_m\}_{m=1}^{M}$. Using a frozen ReID encoder $\Phi(\cdot)$, the identity loss is defined as:

$$\mathcal{L}_{\mathrm{id}} = \frac{1}{M}\sum_{m=1}^{M}\Big[1 - \mathrm{sim}\Big(\Phi(\hat{I}_m), \Phi(I_{\mathrm{ref}})\Big)\Big], \tag{6}$$

where $\mathrm{sim}(u,v) = \frac{u^\top v}{\|u\|\,\|v\|}$ denotes the cosine similarity in the ReID feature space. $\mathcal{L}_{\mathrm{id}}$ encourages the decoded frames to remain close to the reference image in an identity-discriminative feature space. The final training objective is:

$$\mathcal{L} = \mathcal{L}_{\mathrm{FM}} + \lambda_{\mathrm{id}}\mathcal{L}_{\mathrm{id}}, \tag{7}$$

where λ_{id} controls the trade-off between video generation quality and identity preservation.

Table 1: Quantitative comparison on wildlife individual-consistent image-to-video generation. We compare WildIcon with representative I2V generation models.

Method	I2V Subject ↑	I2V Background ↑	Subject Consistency ↑	Motion Smoothness ↑	Text Relevance ↑	FVD ↓
CogVideoX-5B [28]	82.18	86.74	89.92	95.62	25.63	184
Wan2.2-5B [47]	86.02	88.01	90.16	95.94	26.57	175
Wan2.2-A14B [47]	86.23	88.27	90.79	96.43	**26.58**	152
ConsistI2V [48]	83.71	86.18	88.58	95.08	25.21	188
ConsisID [25]	84.06	86.29	89.72	95.56	25.98	168
Phantom [33]	85.64	87.53	91.02	96.11	26.25	172
WildIcon (Ours)	**87.40**	**89.36**	**92.94**	**96.85**	26.38	**139**

5 Experiments

In this section, we conduct qualitative and quantitative evaluations of popular I2V generators and our proposed WildIcon to assess their ability to maintain individual-level wildlife consistency.

5.1 Experimental Settings

Implementation Details. We build our model on Wan2.2-I2V-A14B [47], which generates 81-frame video clips at 832×480 resolution. For training, we employ the Adam optimizer with a learning rate of 10^{-4}. We use a per-GPU batch size of 1 with gradient accumulation over 4 steps (effective batch size 4). All experiments are conducted on NVIDIA RTX PRO 6000 Blackwell (96GB) GPUs.

Baselines. We compare WildIcon with two groups of baselines. First, we include representative general I2V and visual-consistency baselines, including Wan2.2 variants (I2V-A14B and TI2V-5B) [47], CogVideoX1.5-5B [28], and ConsistI2V [48]. Second, we compare with representative subject- or identity-consistent video generation methods, including Phantom [33] and ConsisID [25].

Evaluation metrics. An ideal I2V generator should not only align with the text prompt but also preserve the reference subject throughout the generated video. Accurately quantifying appearance drift is important for evaluating fine-grained video generation. Therefore, we evaluate our method using VBench-based metrics [49; 50] together with complementary video-generation metrics. (1) I2V Subject: Cosine similarity between DINO [42] features of the input image and the generated frames, measuring how well the reference subject is preserved in the generated video. (2) I2V Background: DreamSim [51] feature similarity between the input image and generated frames, assessing the visual consistency of the scene/background. (3) Subject Consistency: Average cosine similarity of DINO features across consecutive frames, evaluating subject appearance consistency throughout the video. (4) Text Relevance: We use CLIPScore [41] to measure the similarity between the generated videos and the input prompts. (5) Motion Smoothness: Motion prior score derived from a video frame interpolation model [52], assessing whether the generated motion remains smooth. (6) FVD: Fréchet Video Distance [53] measures the distributional distance between generated and real videos in the feature space of a pretrained video recognition network. Lower FVD indicates better overall spatio-temporal realism.

Benchmark. To evaluate whether a generator can preserve the identity of the same wildlife individual while following motion and scene instructions, we construct a fixed wildlife generation benchmark from WildlifeVid. The benchmark covers 10 species, with 10 identities per species. For each identity, we select 1 reference image and pair it with 2 text prompts, resulting in 200 generation cases in total. All benchmark identities and their associated clips are excluded from WildlifeVid fine-tuning. All methods are evaluated on this fixed benchmark under identical output resolution and video duration.

5.2 Experimental Results

Quantitative Results. As shown in Table 1, WildIcon achieves the best overall performance on our fixed wildlife generation benchmark. Compared with representative general-purpose I2V baselines and prior subject- or identity-consistent generation methods, our model delivers the clearest gains on identity preservation, improving Subject Consistency from 91.02% to 92.94%. These results suggest that WildIcon better preserves identity-critical wildlife appearance cues throughout the generated



Figure 4: **Examples of individual-consistent video generation by our WildIcon.** Given a reference image and a motion prompt, WildIcon generates videos that follow the specified motion while maintaining fine-grained appearance cues visible in the reference image across frames.

Table 2: **Ablation studies of WildIcon.** Subj., Back., Cons., and Mot. denote I2V Subject, I2V Background, Subject Consistency, and Motion Smoothness. FG denotes foreground-only input, HF denotes high-frequency-only input, LoRA-FT denotes LoRA fine-tuning of Wan2.2 on WildlifeVid, and Full denotes the default WildIcon setting. Default configurations are marked in gray.

(a) Identity branch input.

Setting	Subj.	Back.	Cons.	Mot.
FG-only	86.81	88.39	91.33	96.53
HF-only	86.69	88.34	91.48	96.28
FG+HF	**87.40**	**89.36**	**92.94**	**96.85**

(b) Identity injection strategy.

Setting	Subj.	Back.	Cons.	Mot.
Early	86.74	88.86	91.06	96.24
All	86.82	88.74	91.29	96.04
Mid&late	**87.40**	**89.36**	**92.94**	**96.85**

(c) Training/data effect.

Setting	Subj.	Back.	Cons.	Mot.
Frozen	86.23	88.27	90.79	96.43
LoRA-FT	86.95	88.69	91.63	96.45
Full	**87.40**	**89.36**	**92.94**	**96.85**

video while maintaining stable scene coherence and smooth temporal dynamics. Notably, these gains are achieved without sacrificing text relevance or motion quality, showing that stronger reference-conditioned appearance consistency remains compatible with the general generation capability of the pretrained I2V backbone.

Qualitative Results. Figure 4 presents representative examples of wildlife individual-consistent video generation. The generated videos exhibit consistent fine-grained identity cues across frames, such as fur texture, stripe boundaries, and local contour details, while maintaining realistic motion and scene coherence.

Challenging cases. Figure 5 further illustrates several challenging cases under incomplete or ambiguous reference evidence. For the nyala, its tail overlaps with another individual's head in the reference image, causing the overlap to be partially carried into the generated sequence. For the stoat and cat, vegetation obscures parts of the face; when the generated motion reveals regions that are not clearly visible in the reference, local appearance may drift, and unsupported details can emerge, such as the collar-like artifact in the cat example. For the tiger, thin fence wires crossing the animal are initially retained in the foreground, likely because these high-frequency structures are spatially entangled with the animal in the reference image. As the animal moves away from their original image locations, these artifacts gradually weaken. Overall, these examples show that generation is most reliable when the reference provides clear identity evidence and the requested motion does not require substantial extrapolation of unseen identity-bearing regions.



Figure 5: **Representative challenging cases under incomplete or ambiguous reference evidence.** Red boxes highlight local regions where overlap, occlusion, or scene structures in the reference image lead to unsupported or drifting appearance details during generation.

5.3 Ablation studies

Effect of Key Components. Table 2(a) studies the input design of the identity branch. Results show that combining foreground and high-frequency inputs outperforms either cue alone, confirming that foreground semantic context and high-frequency residual details provide complementary identity cues. Table 2(b) evaluates where identity conditioning should be injected. The default middle-to-late injection strategy outperforms early-block and all-block conditioning, suggesting that identity cues are most effective when introduced during feature refinement.

Effect of WildlifeVid Training. Table 2(c) examines the effect of training on WildlifeVid. LoRA fine-tuning of Wan2.2 improves over the frozen baseline, indicating that WildlifeVid provides useful supervision for wildlife I2V generation. However, the full WildIcon setting achieves stronger results than LoRA fine-tuning alone. This underscores that our architectural design is the key factor behind the performance boost.

6 WildIcon for Animal ReID

To further validate WildIcon, we evaluate on animal ReID, a fine-grained retrieval task for matching the same individual across different images. This experiment tests whether videos synthesized by different generators can provide useful training samples for downstream individual-level recognition. We use mean Average Precision (mAP) [56] and Cumulative Matching Characteristics (CMC) [57] as evaluation metrics. mAP measures overall retrieval performance by averaging precision across all query samples in the query set, while CMC measures the proportion of queries for which the correct identity appears within the top-k retrieved results. We report mAP and CMC at Rank-1 ($\%$), averaged over 10 runs, and provide the corresponding 95% confidence intervals in the Appendix.

Our evaluation follows a fixed synthetic-data construction pipeline. Starting from a curated real training subset, we use different I2V generators to synthesize videos under the same reference images and prompts, and then sample frames from the generated videos as synthetic training candidates. Each candidate is compared with its source reference image using a fixed DINO-based similarity filter, and only high-similarity samples are retained. We further apply a per-identity balanced selection rule to prevent synthetic samples from amplifying identity imbalance. This design supports two controlled comparisons. (1) To examine whether performance gains arise merely from increased data volume or from the additional supervision provided by reference-consistent generated samples, we include a classical data augmentation baseline with the same additional-sample budget. (2) To

Table 3: **Downstream animal ReID results on five species.** All settings use the same real training split, synthetic-data construction pipeline, and unchanged real-only query/gallery splits. We report mAP and Rank-1 (R1) (%) on each species. Best results within each backbone block are shown in bold. Results are averaged over ten runs; 95% confidence intervals are reported in the Appendix.

Model	Setting	Tiger		Nyala		Cow		Panda		Stoat	
		mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1
SigLIP-FT [54]	Real only	55.13	90.21	25.84	41.83	62.57	91.53	28.97	45.43	61.16	92.06
	+ Data Aug.	55.41	90.08	26.11	42.06	62.74	91.47	29.22	45.31	61.38	92.14
	+ Wan synth.	57.24	91.36	29.08	43.02	64.08	92.09	31.14	47.82	62.94	92.57
	+ Ours synth.	**59.72**	**93.41**	**32.01**	**46.36**	**65.84**	**92.11**	**32.26**	**49.36**	**65.18**	**94.71**
DINO-FT [42]	Real only	54.72	92.61	22.94	35.87	66.31	95.08	29.34	44.56	63.07	93.02
	+ Data Aug.	54.89	92.40	23.06	36.02	66.47	95.03	29.58	44.61	63.01	92.94
	+ Wan synth.	56.38	93.44	25.41	38.09	67.73	95.96	30.86	46.18	64.47	94.01
	+ Ours synth.	**58.21**	**94.27**	**27.86**	**40.72**	**69.12**	**96.54**	**31.94**	**47.63**	**66.13**	**94.87**
CLIP-ReID [55]	Real only	54.10	94.22	18.49	28.96	68.43	97.14	29.71	42.10	62.83	91.59
	+ Data Aug.	54.02	94.10	18.88	29.11	68.31	97.06	29.95	42.28	63.01	91.54
	+ Wan synth.	56.41	95.03	21.72	31.64	69.57	97.61	31.38	44.08	64.11	92.42
	+ Ours synth.	**58.36**	**95.96**	**24.58**	**34.29**	**71.04**	**97.85**	**32.47**	**45.96**	**65.72**	**93.36**

assess whether the generated data provides generally useful supervision, we evaluate all training sets with multiple ReID backbones, including SigLIP-FT [54], DINO-FT [42], and CLIP-ReID [55]. All remaining factors are held constant across generators, including the reference-image subset, prompt set, frame-extraction schedule, similarity-filtering rule, synthetic-data budget, ReID training setup, and the unchanged real-only query/gallery split.

As shown in Table 3, adding synthetic data consistently improves ReID performance over the real-only setting across all backbones. In contrast, classical data augmentation provides only marginal and less stable gains, indicating that the improvement is not simply caused by increasing the number of training images. Under the same synthetic-data budget, samples generated by WildIcon outperform those generated by Wan2.2 across the evaluated ReID models, indicating that downstream improvements are primarily driven by the quality of identity preservation in generated samples. Moreover, consistent improvements across different ReID architectures show that WildIcon provides reusable supervisory signals, supporting individual-consistent I2V generation as a practical data enrichment strategy for animal ReID. A generated-to-real retrieval evaluation further shows that WildIcon outperforms both Wan2.2 and WildlifeVid LoRA fine-tuning across all five datasets, while replacing DINOv3 with SigLIP for synthetic-sample selection preserves the downstream advantage (Appendix B.4).

7 Conclusion

In this paper, we explored individual-consistent I2V generation for wildlife as a practical direction for mitigating the data scarcity that limits wildlife identification. To support this task, we proposed WildlifeVid, a wildlife-centric video dataset with individual-level identity labels for training and evaluating individual-consistent I2V models. We further proposed WildIcon, a frequency-aware I2V framework that uses foreground-filtered high-frequency cues to better maintain identity-relevant appearance supported by the reference image. Experiments showed that WildIcon achieves stronger identity consistency than existing baselines, and that its generated samples consistently improve downstream animal ReID across five species. These results suggest that individual-consistent wildlife I2V generation can serve as a practical tool for enriching training data for wildlife identification and supporting broader biodiversity monitoring efforts when real observations are limited.

Acknowledgments and Disclosure of Funding

Funding. This research was supported by the Ministry of Business, Innovation and Employment (MBIE) Smart Ideas project "Stoat Re-Identification Powered by Artificial Intelligence for Effective Pest Management in Aotearoa" (UOA2505).

Competing Interests. The authors declare no competing interests.

References

[1] Lin Huang, Wei Cao, Xinliang Xu, Jiangwen Fan, and Junbang Wang. Linking the benefits of ecosystem services to sustainable spatial planning of ecological conservation strategies. *Journal of Environmental Management*, 222:385–395, 2018.

[2] Yihao Wu, Di Zhao, Yuzhuo Li, Matthew Alajas, Alistair S Glen, Jingfeng Zhang, Gillian Dobbie, Daniel Wilson, and Yun Sing Koh. Overcoming fine-grained visual challenges in animal re-identification via semantic feature alignment. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 371–381. IEEE, 2026.

[3] Pengfei Xu, Yuanyuan Zhang, Minghao Ji, Songtao Guo, Zhanyong Tang, Xiang Wang, Jing Guo, Junjie Zhang, and Ziyu Guan. Advanced intelligent monitoring technologies for animals: A survey. *Neurocomputing*, 585:127640, 2024.

[4] Yuzhuo Li, Di Zhao, Tingrui Qiao, Yihao Wu, Bo Pang, and Yun Sing Koh. Advancing wildlife conservation through multimodal animal re-identification with environmental metadata. *arXiv preprint arXiv:2609.34094*, 2026.

[5] Stefan Schneider, Graham W Taylor, Stefan Linquist, and Stefan C Kremer. Past, present and future approaches using computer vision for animal re-identification from camera trap data. *Methods in Ecology and Evolution*, 10(4):461–470, 2019.

[6] Yihao Wu, Di Zhao, Wayne Marcus Getz, Lingqiao Liu, Gillian Dobbie, Daniel Wilson, and Yun Sing Koh. Region-aware multimodal interleaving for animal re-identification. In *European Conference on Computer Vision*, pages 76–94. Springer, 2026.

[7] Yuzhuo Li, Di Zhao, Tingrui Qiao, Yihao Wu, Bo Pang, and Yun Sing Koh. Metawild: A multimodal dataset for animal re-identification with environmental metadata. In *Proceedings of the 33rd ACM international conference on multimedia*, pages 13009–13015, 2025.

[8] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48, 2019.

[9] Yaping Chai, Haoran Xie, and Joe S Qin. Text data augmentation for large language models: A comprehensive survey of methods, challenges, and opportunities. *Artificial Intelligence Review*, 59(1):35, 2026.

[10] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[11] Xin Xie and Dong Gong. Dymo: Training-free diffusion model alignment with dynamic multi-objective scheduling. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13220–13230. IEEE, 2025.

[12] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 13(3):1195–1215, 2020.

[13] Di Zhao, Jingfeng Zhang, Hongsheng Hu, Philippe Fournier-Viger, Gillian Dobbie, and Yun Sing Koh. Balancing invariant and specific knowledge for domain generalization with online knowledge distillation. In *IJCAI*, pages 2440–2448, 2025.

[14] Tingrui Qiao, Di Zhao, Caroline Walker, Chris Cunningham, and Yun Sing Koh. Zero-shot domain generalisation via prompt-driven feature refinement. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6184–6193. IEEE, 2026.

[15] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks through conditional image generation. *Advances in Neural Information Processing Systems*, 31, 2018.

[16] Qi Song, Weijia Xu, Bin Zhu, Xi Guo, Lei Bao, and Jianping Ge. Boosting tiger re-id with 360 video generation: A generative synthesis approach. *Neurocomputing*, page 133768, 2026.

[17] Xinyu Zhang, Dong Gong, Zicheng Duan, Anton van den Hengel, and Lingqiao Liu. Let your video listen to your music!–beat-aligned, content-preserving video editing with arbitrary music. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 12140–12149, 2025.

[18] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

[19] Jintao Rong, Xin Xie, Xinyi Yu, Linlin Ou, Xinyu Zhang, Chunhua Shen, and Dong Gong. When distillation breaks motion control: Restoring generative trajectories for fast video generators. In *European Conference on Computer Vision*, pages 247–265. Springer, 2026.

[20] Yuze Li, Dong Gong, Xiao Cao, Junchao Yuan, Dongsheng Li, Lei Zhou, Yun Sing Koh, Cheng Yan, and Xinyu Zhang. Let your image move with your motion!–implicit multi-object multi-motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11207–11217, 2026.

[21] Binglei Li, Mengping Yang, Zhiyu Tan, Xiaomeng Yang, Zhizhong Huang, Junping Zhang, and Hao Li. Diversedit++: Quantifying, analyzing, and promoting representation diversity in diffusion transformers, 2026. URL `https://arxiv.org/abs/2608.03082`.

[22] Chenyue Li, Shuoyi Chen, and Mang Ye. Adaptive high-frequency transformer for diverse wildlife re-identification. In *European Conference on Computer Vision*, pages 296–313. Springer, 2024.

[23] Shiqi Lin, Zhizheng Zhang, Zhipeng Huang, Yan Lu, Cuiling Lan, Peng Chu, Quanzeng You, Jiang Wang, Zicheng Liu, Amey Parulkar, et al. Deep frequency filtering for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11797–11807, 2023.

[24] Seunghoon Lee, Seongjae Kang, Inhyuk Park, Gitaek Kwon, Jihyeon Baek, and Doohyun Park. Frequency-balanced retinal representation learning with mutual information regularization. In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=K5tcKEQaUr`.

[25] Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyang Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and Li Yuan. Identity-preserving text-to-video generation by frequency decomposition. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12978–12988, 2025.

[26] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pages 399–417. Springer, 2024.

[27] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[28] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan.Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *The Thirteenth International Conference on Learning Representations*, 2025.

[29] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[30] Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. Dreamvideo: Composing your dream videos with customized subject and motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6537–6549, 2024.

[31] Yuming Jiang, Tianxing Wu, Shuai Yang, Chenyang Si, Dahua Lin, Yu Qiao, Chen Change Loy, and Ziwei Liu. Videobooth: Diffusion-based video generation with image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6689–6700, 2024.

[32] Tao Wu, Yong Zhang, Xintao Wang, Xianpan Zhou, Guangcong Zheng, Zhongang Qi, Ying Shan, and Xi Li. Customcrafter: Customized video generation with preserving motion and concept composition abilities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 8469–8477, 2025.

[33] Lijie Liu, Tianxiang Ma, Bingchuan Li, Zhuowei Chen, Jiawei Liu, Gen Li, Siyu Zhou, Qian He, and Xinglong Wu. Phantom: Subject-consistent video generation via cross-modal alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14951–14961, 2025.

[34] Yuechen Zhang, Yaoyang Liu, Bin Xia, Bohao Peng, Zexin Yan, Eric Lo, and Jiaya Jia. Magicmirror: Id-preserved video generation in video diffusion transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14464–14474, 2025.

[35] Brian Nlong Zhao, Jiajun Wu, and Shangzhe Wu. Web-scale collection of video data for 4d animal reconstruction. *arXiv preprint arXiv:2511.01169*, 2025.

[36] Xun Long Ng, Kian Eng Ong, Qichen Zheng, Yun Ni, Si Yong Yeo, and Jun Liu. Animal kingdom: A large and diverse dataset for animal behavior understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19023–19034, June 2022.

[37] Jun Chen, Ming Hu, Darren J Coker, Michael L Berumen, Blair Costelloe, Sara Beery, Anna Rohrbach, and Mohamed Elhoseiny. Mammalnet: A large-scale video benchmark for mammal recognition and behavior understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13052–13061, 2023.

[38] Dan Liu, Jin Hou, Shaoli Huang, Jing Liu, Yuxin He, Bochuan Zheng, Jifeng Ning, and Jingdong Zhang. Lote-animal: A long time-span dataset for endangered animal behavior understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20064–20075, October 2023.

[39] Brandon Castellano. Pyscenedetect. `https://www.scenedetect.com/`, 2026. Accessed: 2026-04-01.

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PmLR, 2021.

[41] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, 2021.

[42] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

[43] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[45] Hyung Won Chung, Xavier Garcia, Adam Roberts, Yi Tay, Orhan Firat, Sharan Narang, and Noah Constant. Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. In *The Eleventh International Conference on Learning Representations*, 2023.

[46] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

[47] Wan Team. Wan2.2: More powerful, more beautiful. `https://wan.video/blog/wan2.2`, 2025.

[48] Weiming Ren, Huan Yang, Ge Zhang, Cong Wei, Xinrun Du, Wenhao Huang, and Wenhu Chen. Consisti2v: Enhancing visual consistency for image-to-video generation. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

[49] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024.

[50] Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, et al. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[51] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: learning new dimensions of human visual similarity using synthetic data. In *Advances in Neural Information Processing Systems*, volume 36, pages 50742–50768, 2023.

[52] Zhen Li, Zuo-Liang Zhu, Ling-Hao Han, Qibin Hou, Chun-Le Guo, and Ming-Ming Cheng. Amt: All-pairs multi-field transforms for efficient frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9801–9810, 2023.

[53] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

[54] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[55] Siyuan Li, Li Sun, and Qingli Li. Clip-reid: exploiting vision-language model for image re-identification without concrete text labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1405–1413, 2023.

[56] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[57] Hyeonjoon Moon and P Jonathon Phillips. Computational and performance aspects of pca-based face-recognition algorithms. *Perception*, 2001.

[58] Shuyuan Li, Jianguo Li, Hanlin Tang, Rui Qian, and Weiyao Lin. Atrw: A benchmark for amur tiger re-identification in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2590–2598, 2020.

[59] Nkosikhona Dlamini and Terence L Van Zyl. Automated identification of individuals in wildlife population using siamese neural networks. In *International Conference on Soft Computing & Machine Intelligence*, pages 224–228. IEEE, 2020.

[60] Lili Fu, Shijun Li, Shuolin Kong, Ruiwen Ni, Haohong Pang, Yu Sun, Tianli Hu, Ye Mu, Ying Guo, and He Gong. Lightweight individual cow identification based on ghost combined with attention mechanism. *Plos one*, 17(10):e0275435, 2022.

[61] Le Wang, Rizhi Ding, Yuanhao Zhai, Qilin Zhang, Wei Tang, Nanning Zheng, and Gang Hua. Giant panda identification. *IEEE Transactions on Image Processing*, 30:2837–2849, 2021.

[62] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[63] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1010–1019, 2016.

[64] Najmeh Sadoughi, Yang Liu, and Carlos Busso. Msp-avatar corpus: Motion capture recordings to study the role of discourse functions in the design of intelligent virtual agents. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, volume 7, pages 1–6. IEEE, 2015.

[65] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in Neural Information Processing Systems*, 32, 2019.

[66] Wei Xiong, Wenhan Luo, Lin Ma, Wei Liu, and Jiebo Luo. Learning to generate time-lapse videos using multi-stage dynamic generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2364–2373, 2018.

[67] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1–11, 2019.

[68] Di Chang, Yichun Shi, Quankai Gao, Hongyi Xu, Jessica Fu, Guoxian Song, Qing Yan, Yizhe Zhu, Xiao Yang, and Mohammad Soleymani. Magicpose: realistic human poses and facial expressions retargeting with identity-aware diffusion. In *Proceedings of the 41st International Conference on Machine Learning*, pages 6263–6285, 2024.

[69] Mingyang Wu, Ashirbad Mishra, Soumik Dey, Shuo Xing, Naveen Ravipati, Hansi Wu, Binbin Li, and Zhengzhong Tu. Consid-gen: View-consistent and identity-preserving image-to-video generation. *arXiv preprint arXiv:2602.10113*, 2026.

[70] Di Zhao, Jingfeng Zhang, Hongsheng Hu, Philippe Fournier-Viger, Gillian Dobbie, and Yun Sing Koh. Unlearning during training: Domain-specific gradient ascent for domain generalization. In *The Fourteenth International Conference on Learning Representations*, 2026.

[71] Bo Pang, Tingrui Qiao, Caroline Walker, Chris Cunningham, and Yun Sing Koh. Cabin: Debiasing vision-language models using backdoor adjustments. In *IJCAI*, pages 484–492, 2025.

[72] Tingrui Qiao, Di Zhao, Yuzhuo Li, Bo Pang, Caroline Walker, Chris Cunningham, and Yun Sing Koh. Multiple images distract large multimodal models via attention fragmentation. In *European Conference on Computer Vision*, pages 361–379. Springer, 2026.

[73] Dawei Wang, Di Zhao, Xinyuan Liu, Marci Chi Ma, Xiaoyang Liu, Chengming Zhou, Gary Ushaw, and Richard Davison. Mars-ra: Rank aggregation for credit assignment via multimodal comparisons in embodied multi-agent cooperation. In *Proceedings of the Association for Computational Linguistics*, pages 29270–29289, 2026.

[74] Dawei Wang, Chengming Zhou, Di Zhao, Xinyuan Liu, Marci Chi Ma, Gary Ushaw, and Richard Davison. Towermind: A tower defence game learning environment and benchmark for llm as agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 26151–26159, 2026.

[75] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13320–13331, 2024.

[76] Zhedong Lin, Zhongsheng Wang, Qian Liu, Xinyu Zhang, and Jiamou Liu. Narratology meets text-to-image: a survey of consistency in ai generated storybook illustrations. *Artificial Intelligence Review*, 59(2):83, 2026.

[77] Zhongsheng Wang, Zhedong Lin, Qian Liu, Xinyu Zhang, and Jiamou Liu. Sdo: Subspace deconflicting operator for multi-adapter composition. *arXiv preprint arXiv:2608.13820*, 2026.

[78] OpenAI. GPT-5 Model Documentation. `https://platform.openai.com/docs/models/gpt-5`, 2025. Accessed: 2026-04-07.

A Implementation and Evaluation Details

A.1 Reproducibility

We open-source the implementation of WildIcon at `https://ml-4-socialgood.github.io/WildIcon/`. The repository contains the code needed to reproduce our individual-consistent wildlife I2V framework, including the frequency-aware identity encoder, foreground-isolated high-frequency extraction, identity-token projection, selective identity conditioning, training and inference pipelines, and the evaluation scripts used in our experiments.

Code. The released code is organized around the main components of WildIcon. It includes: (1) a foreground extraction and high-frequency preprocessing module that constructs I_{seg} and I_{hf} from a reference image; (2) a dual-branch identity encoder that encodes foreground appearance and high-frequency residual cues into identity tokens; (3) an identity-conditioning module that projects the identity tokens into the conditioning space and injects them into selected middle-to-late DiT blocks; and (4) training and inference scripts for generating wildlife videos from a reference image and a standardized text prompt. The repository also provides environment requirements, example commands, and configuration files for reproducing the main experimental settings.

Dataset and benchmark curation. To support reproducibility of WildlifeVid and the evaluation benchmark, the repository includes the metadata construction and curation utilities used in our pipeline. These scripts cover source-level metadata processing, clip-level filtering, standardized prompt organization, and identity-merging utilities.

Evaluation. The repository contains the evaluation scripts used to compute the video-generation metrics reported in the paper, including subject preservation, background consistency, subject consistency, motion smoothness, text relevance, and FVD evaluation.

A.2 Data Accessibility and Release

WildlifeVid is constructed by curating and re-annotating clips from existing wildlife video resources, including videos hosted on public platforms where applicable. We acknowledge that such source videos may be subject to copyright protection, platform terms of service, and the usage conditions of the original datasets. To respect these terms and mitigate copyright, redistribution, and ecological-sensitivity concerns, we do not redistribute the original videos or raw video frames. Instead, we publicly release only the annotations, which do not contain any raw video content. All derived data is non-identifiable and used solely for research purposes. The WildlifeVid full annotation files are distributed separately as supplementary material. Our newly created WildlifeVid annotations, prompts, scripts, and statistics are released under CC BY 4.0. Source-derived identifiers, labels, metadata, and media remain subject to the licenses and terms of their respective sources.

A.3 Animal ReID Datasets

We evaluate the downstream utility of WildIcon on five animal ReID datasets covering four public datasets, Tiger [58], Nyala [59], Cow [60], Panda [61], and our proprietary camera-trap Stoat dataset. For each dataset, we organize the data into a processed ReID split consisting of training, query, and gallery sets. The resulting dataset statistics are summarized in Table 4.

Animal-centric preprocessing. All real training, query, gallery, and synthetic-filtering inputs are processed as animal-centric crops using the same detection and preprocessing pipeline. The similarity used for synthetic-sample filtering is computed between the animal crop of each generated frame and the corresponding animal crop of its real reference image, rather than between full-scene images. This reduces the influence of background and scene context in both synthetic-sample selection and downstream ReID evaluation.

A.4 ReID Evaluation with Synthetic Data

To evaluate the downstream utility of generated videos, we construct synthetic training images from videos generated by each I2V model under a fixed pipeline. For fairness, all compared methods use the same reference-image pool, text prompts, video-generation settings, frame-sampling rule, similarity-filtering threshold, and final synthetic-data budget. The query and gallery sets are kept

Table 4: **Animal ReID datasets used for downstream evaluation.** We report the processed ReID splits used in our experiments. These statistics correspond to our curated splits and may differ from the raw dataset statistics reported in the original papers.

Animal	Dataset	Train		Query		Gallery		Total	
		Imgs	IDs	Imgs	IDs	Imgs	IDs	Imgs	IDs
Tiger	ATRW [58]	3,678	148	424	33	521	33	4,623	181
Nyala	Nyala [59]	1,213	179	354	58	375	58	1,942	237
Cow	CowDataset [60]	919	8	210	5	356	5	1,485	13
Panda	IPanda50 [61]	4,114	31	1,022	19	1,738	19	6,874	50
Stoat	Stoat (Proprietary)	1,390	37	403	46	780	46	2,573	83

fixed across all experiments and contain only real images; synthetic samples are used only to augment the training set.

Reference image selection. To prevent test-set leakage, reference images are selected from only the real training split of each ReID dataset. Let $\mathcal{D}_i^{\text{train}}$ denote the real training images of identity i. We construct a reference pool $\mathcal{R}_i \subseteq \mathcal{D}_i^{\text{train}}$ by excluding images that are unsuitable as I2V references, such as severely blurred, low-visibility nighttime, or infrared images with weak visible identity cues. This filtering is applied only to the reference pool for video generation; the original real, query, and gallery splits remain unchanged. The same reference pool is used for all compared generation methods.

Video generation protocol. For each selected reference image $r \in \mathcal{R}_i$, we generate two videos. The prompt is designed to encourage mild motion and viewpoint variation starting from the reference pose, such as slight turning, small body movement, or modest camera-view changes, while preserving the identity of the target animal as much as possible. All generated videos are processed using the default generation setting of 81 frames at 16 FPS.

Reference-evidence admission guardrail. For downstream ReID augmentation, we restrict generation to motion and viewpoint changes that remain supported by the visual evidence in the reference image. Prompts requiring a full 360° rotation, explicit exposure of the opposite side, or a camera orbit around the animal are excluded because they may reveal identity-bearing regions that are not observed in the reference. Generated candidates are further checked for whether sufficient reference-visible appearance remains comparable to the source image. Candidates that expose unsupported identity-bearing regions or otherwise become visually uncertain are discarded before the similarity-filtering stage and are not assigned the source identity label.

Frame sampling. To convert generated videos into synthetic training images, we sample frames at fixed temporal intervals. We discard the first 5 frames and the last 5 frames of each generated video to avoid frames that are overly close to the static input image and frames that may contain end-of-video artifacts, yielding 8 candidate synthetic frames per video. For an 81-frame video indexed from 0 to 80, we use

$$\mathcal{F}_{\text{samp}} = \{5, 15, 25, 35, 45, 55, 65, 75\},$$

DINO similarity filtering. After frame sampling, we filter synthetic candidates based on their appearance consistency with the corresponding source reference image. Let $\Phi(\cdot)$ denote the frozen DINOv3 image encoder. For a sampled frame x generated from the reference image r, we compute

$$\bar{\Phi}(x) = \frac{\Phi(x)}{\|\Phi(x)\|_2}, \quad s(x, r) = \bar{\Phi}(x)^\top \bar{\Phi}(r).$$

A sampled frame is retained only if $s(x, r) \geq \tau$, where $\tau = 0.90$ is fixed for all datasets and generation methods. This filtering step removes candidates with large appearance deviations from the source reference and retains samples that remain visually close to the reference under the fixed DINO representation.

Synthetic budget control. After similarity filtering, directly adding all retained frames could make the synthetic data overly dominant and amplify identity imbalance, because identities with more valid generated frames would receive disproportionately more synthetic samples. We therefore apply a second-stage budget control procedure to construct a balanced augmented training set.



Figure 6: **Qualitative comparison with representative video generators.** Given the same reference image and text prompt, we compare WildIcon with representative general I2V and identity-preserving video generation baselines.

For each dataset, we first set a target synthetic fraction ρ for the final training set. For an identity i with n_i^{real} real training images, the per-identity synthetic budget is computed as

$$b_i = \text{round}\left(\frac{n_i^{\text{real}} \cdot \rho}{1-\rho}\right),$$

so that the selected synthetic samples form approximately a fraction ρ of the final real-plus-synthetic training set. In our experiments, we use $\rho = 0.2$, corresponding to a target synthetic fraction of approximately 20%. We also limit each source reference image to a maximum of 2 selected synthetic frames to prevent a single reference image from dominating the augmented set. Within each identity, candidate synthetic frames are ranked by DINO similarity to the corresponding source reference image, and the highest-scoring frames are selected subject to the per-identity budget and per-source cap. If an identity has fewer valid candidates than its allocated budget, all valid candidates are retained, and the remaining budget is not filled with lower-similarity samples. Let $\mathcal{A}_i$ denote the selected synthetic samples for identity i, and let

$$\mathcal{A}_{\text{syn}} = \bigcup_i \mathcal{A}_i$$

denote the final selected synthetic subset across all identities. The augmented training set is then

$$\mathcal{D}_{\text{aug}}^{\text{train}} = \mathcal{D}_{\text{real}}^{\text{train}} \cup \mathcal{A}_{\text{syn}}.$$

The query and gallery sets remain unchanged throughout all ReID experiments. They contain only real images and never include synthetic samples. Therefore, synthetic data are used only to augment the real training set, and all reported ReID results are evaluated on the same fixed real-only query/gallery splits.

B Additional Results

B.1 Qualitative comparison with video generation baselines

Figure 6 complements the quantitative comparison in Table 1. Under the same reference images and prompts, existing video generators often produce plausible motion but exhibit identity-cue drift,

Table 5: **Quantitative ablation studies of WildIcon.** We study three factors: the input to the identity branch, the strategy for injecting identity conditioning, and the effect of the WildlifeVid dataset. The final row reports the full WildIcon model as the common reference for all ablations.

Method	I2V Subject ↑	I2V Background ↑	Subject Consistency ↑	Motion Smoothness ↑	Text Relevance ↑	FVD ↓
Wan2.2-A14B	86.23	88.27	90.79	96.43	**26.58**	152
Identity branch input						
Full image	86.57	88.41	91.07	96.45	26.56	159
Foreground only	86.81	88.39	91.33	96.53	26.49	148
High-frequency only	86.69	88.34	91.48	96.28	26.43	150
Injection strategy						
Early blocks	86.74	88.86	91.06	96.24	26.47	147
All blocks	86.82	88.74	91.29	96.04	26.24	144
Dataset effectiveness						
Wan2.2-A14B-FT	86.95	88.69	91.63	96.45	26.55	146
WildIcon (Ours)	**87.40**	**89.36**	**92.94**	**96.85**	26.38	**139**

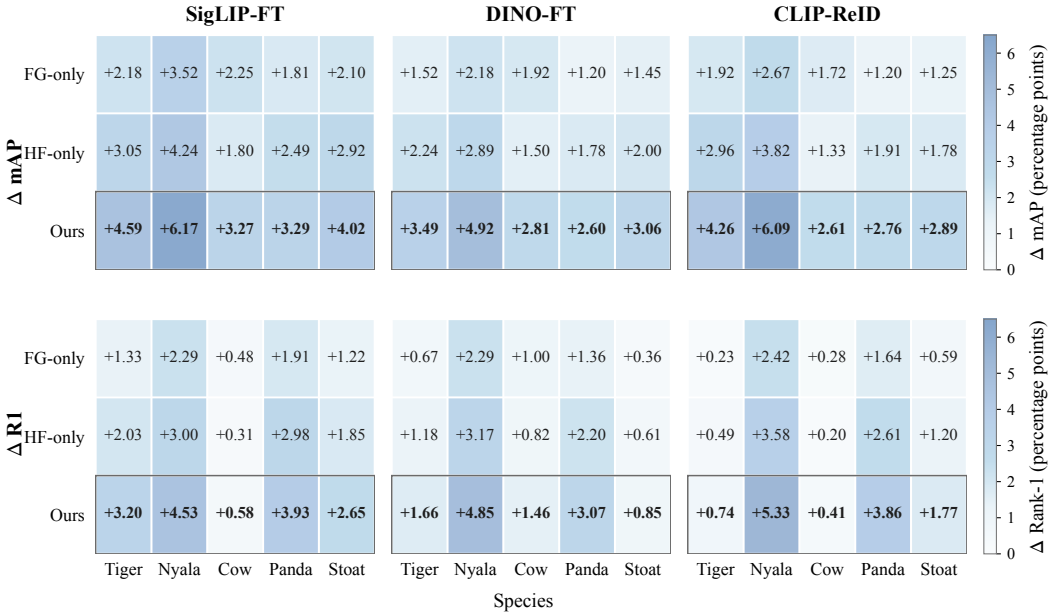


Figure 7: **Downstream ReID ablation of WildIcon identity-branch inputs.** We visualize the absolute performance gains over the real-only baseline for different identity-branch variants, including foreground-only conditioning (FG-only), high-frequency-only conditioning (HF-only), and the full WildIcon identity branch (Ours). Darker colors indicate larger downstream ReID gains.

including blurred stripe or spot boundaries, shifted fur texture, altered local markings, and inconsistent body contours. These failures indicate that generic image-conditioned or identity-preserving video generators do not explicitly preserve the fine-grained local cues required for individual consistency in wildlife. WildIcon better maintains these identity-critical patterns across frames while preserving coherent motion and prompt alignment.

B.2 Additional Ablation Results

Quantitative ablation studies of WildIcon. Table 5 provides a more complete ablation analysis of WildIcon, covering three factors: the identity-branch input, the identity-conditioning injection strategy, and the effect of WildlifeVid training.

First, the identity-branch input ablation shows that directly using the full reference image provides only limited gains over Wan2.2-A14B. This suggests that raw reference-image conditioning may

introduce background clutter and non-identity visual details that are not useful for preserving wildlife identity. In contrast, foreground-only conditioning improves both identity preservation and video realism by suppressing background interference. High-frequency-only conditioning yields slightly stronger subject consistency than foreground-only conditioning, indicating that local residual cues are useful for maintaining identity-sensitive texture and boundary details. However, it performs worse than the full WildIcon setting, because high-frequency residuals alone discard semantic foreground context and spatial support. The full WildIcon identity branch achieves the best overall results by combining foreground appearance with foreground-filtered high-frequency cues, confirming that these two sources provide complementary identity evidence.

Second, the injection-strategy ablation shows that injecting identity conditioning into early blocks or all blocks is less effective than the default middle-to-late strategy. Early-block injection improves background consistency but provides weaker subject consistency, suggesting that introducing identity cues too early may interfere with coarse layout and motion formation. All-block injection improves FVD compared with early-only injection, but it reduces motion smoothness and text relevance, indicating that overly persistent identity conditioning can constrain prompt-driven generation. The middle-to-late injection strategy gives the best balance, as these blocks are better suited for refining appearance details and identity-sensitive textures after global structure and motion have been established.

Third, the dataset-effectiveness ablation shows that LoRA fine-tuning Wan2.2-A14B on WildlifeVid improves over the frozen Wan2.2-A14B baseline, showing that WildlifeVid provides useful supervision for wildlife I2V generation. However, fine-tuning alone remains below the full WildIcon model across the main identity and realism metrics. This indicates that the gains are not solely due to domain-specific fine-tuning data, but also depend on the frequency-aware identity branch and selective identity conditioning.

Qualitative ablation studies of WildIcon. Figure 8 provides qualitative evidence for the ablation results in Table 5. The red boxes highlight representative local failures caused by removing or misplacing key components of WildIcon. With foreground-only conditioning, the model preserves the coarse body appearance but still produces localized identity drift, such as tail deformation in later frames. With high-frequency-only conditioning, local residual cues are available, but the lack of foreground semantic context leads to unstable appearance, including abnormal whitening and texture drift on the back. Injecting identity tokens only into early blocks disrupts temporal formation, leading to less coherent motion and more noticeable changes in head appearance across frames. Injecting identity conditioning into all blocks can over-constrain the generation process and introduce structural artifacts, such as distorted legs and feet. In contrast, the full WildIcon model preserves the nyala's stripe patterns, body contour, head appearance, and limb structure more consistently across the generated frames.

Downstream effect of identity-branch inputs. Figure 7 reports ReID gains over the real-only baseline for different identity-branch inputs. FG-only and HF-only synthetic data both improve downstream ReID, but neither consistently matches the full WildIcon setting. Combining foreground-isolated appearance with foreground-filtered high-frequency residuals yields the strongest gains across species and ReID backbones, indicating that the two cues are complementary for producing useful filtered synthetic training candidates.

B.3 ReID results with confidence interval

Figure 9 visualizes the 95% confidence intervals for the downstream ReID gains reported in Table 3. We plot the change in mAP and Rank-1 relative to the corresponding real-only baseline for each species and backbone. The dashed vertical line denotes zero gain, while points to the right indicate improved ReID performance after adding extra training samples.

Across the three backbones, classical data augmentation stays close to the zero-gain line, suggesting that simply increasing the training set through standard transformations provides limited benefit. In contrast, both synthetic-data settings generally improve mAP and Rank-1, with filtered WildIcon-generated samples producing the largest gains in most species-backbone combinations. The improvements are especially clear for mAP, where our method consistently shifts further to the right than Wan's synthetic data, indicating that the larger gains observed for WildIcon are consistent

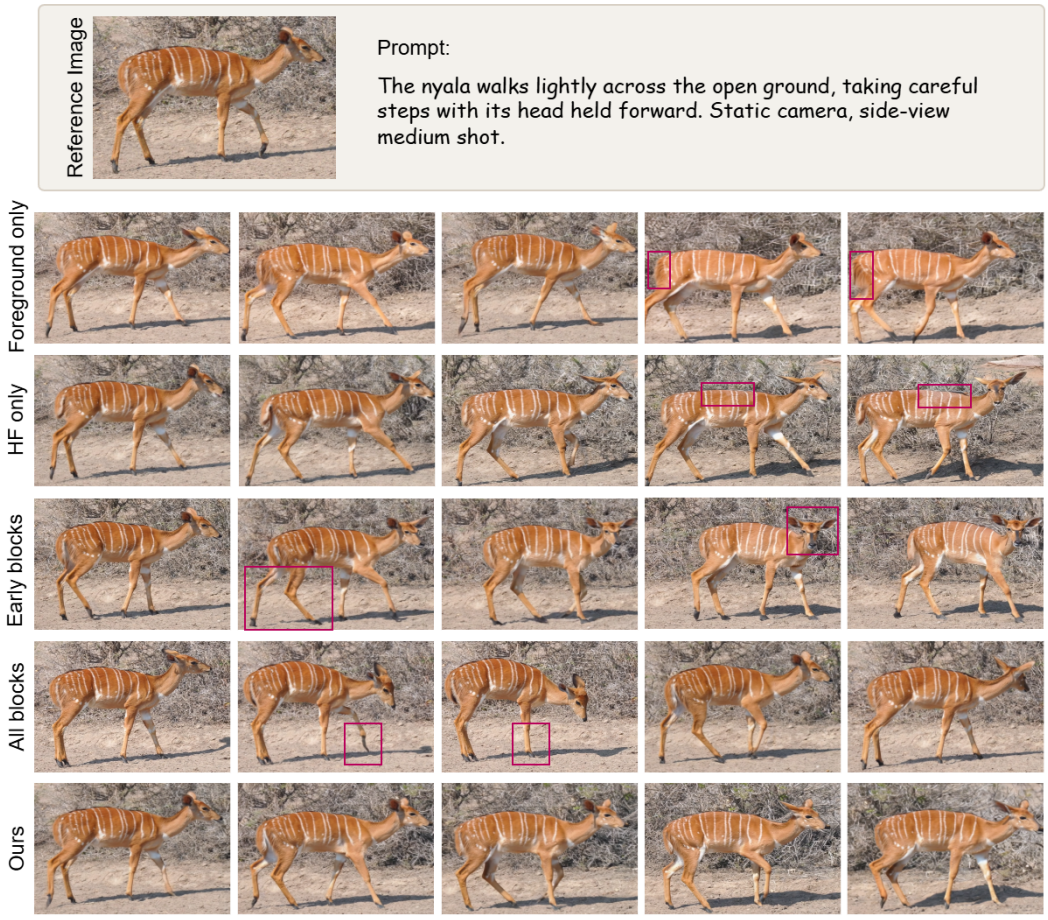


Figure 8: **Qualitative ablation studies of WildIcon.** We compare WildIcon variants under the same reference image and prompt. Red boxes mark representative artifacts, including tail deformation, back texture drift, head appearance changes, motion inconsistency, and limb distortion. The full WildIcon model better preserves local identity cues and structural consistency across frames.

with improved reference-conditioned appearance preservation providing more useful downstream training supervision.

B.4 Additional ReID Evaluations

Direct generated-to-real retrieval. To provide a more direct evaluation of individual correspondence, we test whether generated frames can retrieve held-out real images of the source individual among same-species negatives. For each eligible test identity, one image from the real query set is used solely as the I2V reference and excluded from the retrieval gallery. The gallery contains other held-out real images of the same identity as positives and real gallery images of all other test identities from the same species as negatives. The CLIP-ReID evaluator is trained only on the disjoint real training split and does not observe any test identity during training.

Using the same reference image, prompt, random seed, and generation settings, we generate one video per identity with frozen Wan2.2-A14B, Wan2.2 LoRA-fine-tuned on WildlifeVid (LoRA-FT), and WildIcon. To reduce dependence on frames that remain very close to the reference image, we exclude frames before index 15 and evaluate the same seven predetermined positions for every method: $\{15, 25, 35, 45, 55, 65, 75\}$. No DINO-based filtering, similarity thresholding, or manual frame selection is used in this experiment. Generated queries and real gallery images use the same animal-centric preprocessing pipeline.

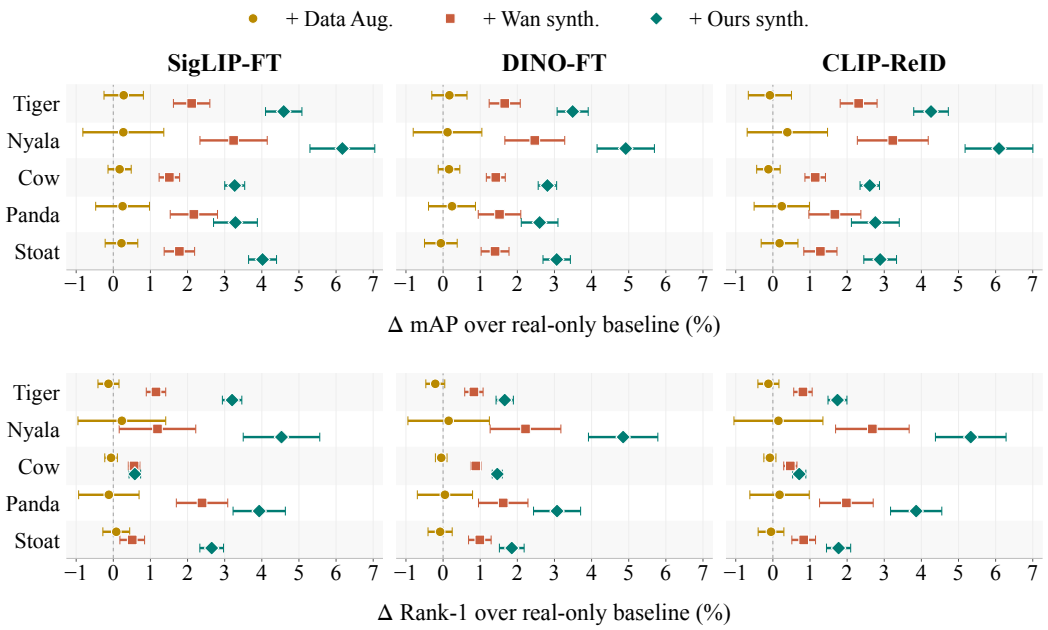


Figure 9: **Confidence intervals for downstream animal ReID gains.** For each backbone and species, we report the change in mAP and Rank-1 relative to the corresponding real-only baseline. Points denote mean gains over ten runs, and horizontal bars denote 95% confidence intervals. The dashed vertical line marks zero gain.

Table 6: **Direct generated-to-real retrieval.** Each entry reports frame-level mAP / Rank-1 (%). Generated frames are used as queries against a gallery containing held-out real images of the source individual and same-species negatives. No DINO-based filtering or manual frame selection is applied.

Query source	Tiger	Nyala	Cow	Panda	Stoat
Real reference	64.38 / 96.97	12.87 / 11.11	71.11 / 100.00	30.60 / 68.42	64.49 / 82.61
Wan2.2-A14B	42.73 / 78.79	4.18 / 3.97	54.62 / 85.71	21.87 / 47.37	42.36 / 60.25
LoRA-FT	43.91 / 82.68	4.95 / 5.29	54.74 / 88.57	21.68 / 49.62	46.83 / 67.08
WildIcon	**53.72 / 89.61**	**10.84 / 10.32**	**64.27 / 94.29**	**26.14 / 58.65**	**54.61 / 75.78**

As shown in Table 6, WildIcon achieves the highest mAP and Rank-1 among the generated-query methods across all five datasets. The real-reference queries remain stronger, as expected, while WildIcon narrows the gap more consistently than either frozen Wan2.2 or WildlifeVid LoRA fine-tuning. This experiment provides complementary evidence that WildIcon-generated frames retain source-specific cues that support retrieval against held-out real observations.

Sensitivity to the synthetic-sample selection representation. The generated-to-real experiment above is independent of DINO during both generated-frame selection and retrieval evaluation: frames are sampled at predetermined temporal positions and evaluated using CLIP-ReID. We additionally test whether the downstream advantage depends on using DINOv3 for synthetic-sample filtering.

We replace the DINOv3 selection representation with a frozen SigLIP encoder. Both representations operate on the same generated candidate pool and animal-centric crops. Because their similarity scales are not directly comparable, candidates are ranked within each identity, while the same per-identity synthetic budget and per-reference cap are retained. The ReID training protocol and real-only query/gallery sets remain unchanged, and CLIP-ReID is used as the downstream evaluator.

Table 7 reports the absolute mAP / Rank-1 difference between WildIcon and Wan2.2. WildIcon retains positive gains across all five datasets under both DINOv3- and SigLIP-based selection, indicating that the downstream improvement is not specific to a single synthetic-sample selection representation.

Table 7: **Effect of the synthetic-sample selection representation.** Each entry reports the absolute mAP / Rank-1 difference (%) between WildIcon and Wan2.2 using CLIP-ReID as the downstream evaluator.

Selection filter	Tiger	Nyala	Cow	Panda	Stoat
DINOv3	+1.95 / +0.93	+2.86 / +2.65	+1.47 / +0.24	+1.09 / +1.88	+1.61 / +0.94
SigLIP	+1.81 / +0.68	+2.94 / +2.11	+1.12 / +0.08	+1.57 / +2.03	+1.44 / +1.06

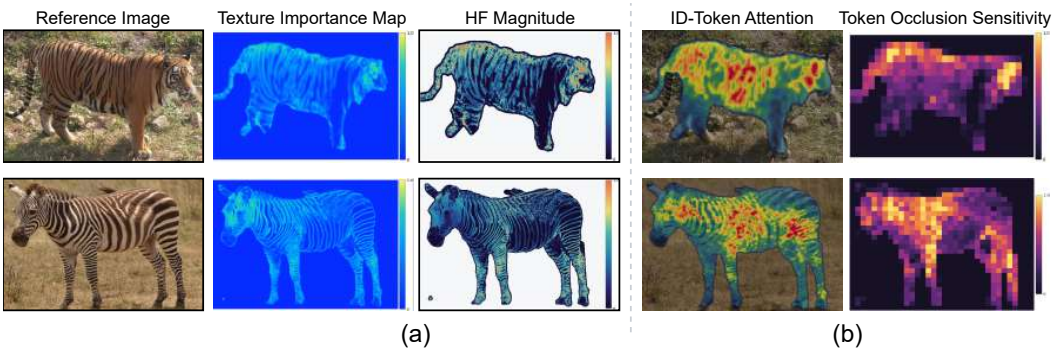


Figure 10: **Frequency-cue utilization analysis.** (a) Texture importance and high-frequency (HF) magnitude maps localize identity-related structures, such as stripe boundaries, tail patterns, and contour transitions. (b) ID-token attention and token occlusion sensitivity maps show that WildIcon attends to and depends on these high-frequency identity regions during generation.

B.5 Frequency-Cue Utilization Analysis

To further examine whether WildIcon actively uses high-frequency identity cues during generation, we provide a frequency-cue utilization analysis in Fig. 10. This analysis has two parts: cue localization and model utilization. The first part examines whether the high-frequency residuals spatially overlap with visible identity-discriminative regions, and the second part analyzes whether the injected identity tokens influence the denoising process.

Cue localization. Fig. 10(a) visualizes the relationship between identity-cue regions and high-frequency information. For each reference image, we identify identity-discriminative regions, including body stripes, facial markings, tail patterns, and contour transitions. These regions indicate the local visual structures that are likely to distinguish one individual from another within the same species.

We visualize the foreground-filtered high-frequency residual I_{hf} used by WildIcon by converting it into a scalar magnitude map. Since I_{hf} is a multi-channel residual image, we average the absolute residual magnitude over color channels:

$$H_{\mathrm{hf}}(u,v) = \frac{1}{C}\sum_{c=1}^{C}\left|I_{\mathrm{hf}}^{(c)}(u,v)\right|,$$

where (u,v) denotes the spatial location and C is the number of image channels. Let Ω_{fg} denote the foreground region obtained from I_{seg}. For visualization, we normalize the response within Ω_{fg}:

$$\widetilde{H}_{\mathrm{hf}}(u,v) = \frac{H_{\mathrm{hf}}(u,v) - \min_{(u',v')\in\Omega_{\mathrm{fg}}} H_{\mathrm{hf}}(u',v')}{\max_{(u',v')\in\Omega_{\mathrm{fg}}} H_{\mathrm{hf}}(u',v') - \min_{(u',v')\in\Omega_{\mathrm{fg}}} H_{\mathrm{hf}}(u',v') + \varepsilon}.$$

Pixels outside the foreground region are masked out in the visualization. Higher values in $\widetilde{H}_{\mathrm{hf}}$ indicate sharper local structures, such as stripe boundaries, texture transitions, and contour changes. The spatial overlap between marked identity-cue regions and the normalized high-frequency response shows that the foreground-filtered high-frequency residual captures local structures relevant to individual-level discrimination, rather than high-frequency clutter from the background.

Model utilization. Fig. 10(b) further analyzes whether the generation backbone uses the high-frequency cues. We consider two complementary visualizations: identity-token cross-attention and HF-guided identity-token occlusion contribution.

For the cross-attention analysis, we record the attention weights from video visual tokens to the unified condition sequence in the selected DiT blocks. Let $X_\sigma^{(l)}$ denote the video visual tokens at noise level σ and selected DiT block $l \in \mathcal{S}$, where $\mathcal{S}$ is the set of middle-to-late blocks used for identity conditioning. The projected identity tokens are denoted by $\bar{T}_{\mathrm{id}} = P_{\mathrm{id}}(T_{\mathrm{id}})$, and the unified condition sequence is $T_{\mathrm{cond}} = [T_{\mathrm{txt}}; \bar{T}_{\mathrm{id}}]$, where $[\cdot;\cdot]$ denotes token-wise concatenation. For attention head h, the query and key projections are

$$Q_{\mathrm{vid},\sigma}^{(l,h)} = X_\sigma^{(l)} W_Q^{(l,h)}, \quad K_{\mathrm{cond}}^{(l,h)} = T_{\mathrm{cond}} W_K^{(l,h)}.$$

The cross-attention matrix is computed as

$$A_{\mathrm{cond},\sigma}^{(l,h)} = \mathrm{softmax}\left(\frac{Q_{\mathrm{vid},\sigma}^{(l,h)} \left(K_{\mathrm{cond}}^{(l,h)}\right)^\top}{\sqrt{d_h}} \right),$$

where d_h is the feature dimension of each attention head, and the softmax is applied over the condition-token dimension. We retain the columns corresponding to $\bar{T}_{\mathrm{id}}$, yielding the identity-attention matrix $A_{\mathrm{id},\sigma}^{(l,h)}$. Each entry $A_{\mathrm{id},\sigma}^{(l,h)}(q,p)$ measures how strongly generated visual token q attends to identity token p.

We aggregate the identity-attention response by averaging over selected layers, attention heads, sampled noise levels, and generated visual tokens:

$$\alpha_p = \underset{l\in\mathcal{S},\, h\in\mathcal{H},\, \sigma\in\Sigma,\, q\in\mathcal{Q}}{\mathrm{Mean}} A_{\mathrm{id},\sigma}^{(l,h)}(q,p),$$

where α_p denotes the aggregated attention score of identity token p. The token-level scores $\{\alpha_p\}$ are then reshaped to the reference-image token grid and upsampled to the image resolution. High-response regions indicate reference identity-token locations that receive stronger attention from generated visual tokens. The resulting maps are concentrated around high-frequency identity regions, suggesting that WildIcon uses the injected identity tokens to refine local appearance cues during generation.

To provide a stronger diagnostic signal than attention alone, we also compute an HF-guided identity-token occlusion contribution map. Let $\mathcal{G}$ denote the set of spatial identity-token groups, where each group $g \in \mathcal{G}$ corresponds to a local region in the reference image. For each token group g, we construct an occluded identity-token sequence $T_{\mathrm{id}}^{\mathrm{occ}(g)} = \mathrm{Occlude}(T_{\mathrm{id}}; g)$ by replacing the corresponding local identity tokens with the mean identity-token feature. Here, $\mathrm{Occlude}(\cdot)$ keeps all identity tokens unchanged except those in group g, which are replaced by the mean feature computed over T_{id}. Let $c(T_{\mathrm{id}})$ denote the identity-augmented condition used in the selected cross-attention blocks, constructed from the text tokens and the projected identity tokens, i.e., $c(T_{\mathrm{id}}) = [T_{\mathrm{txt}}; P_{\mathrm{id}}(T_{\mathrm{id}})]$. Using the same noisy latent, text prompt, and noise level, we measure the change in the predicted velocity:

$$\Delta_g(\sigma) = \left\| \Psi_\theta(\mathbf{z}_\sigma, \sigma, c(T_{\mathrm{id}})) - \Psi_\theta\left(\mathbf{z}_\sigma, \sigma, c(T_{\mathrm{id}}^{\mathrm{occ}(g)})\right) \right\|_2,$$

where Ψ_θ is the velocity prediction network. The contribution score for token group g is obtained by averaging over sampled noise levels:

$$S_g = \mathbb{E}_{\sigma\in\Sigma}\left[\Delta_g(\sigma)\right],$$

where Σ denotes the set of sampled noise levels used for the analysis. For visualization, we normalize the contribution scores across all token groups:

$$\widetilde{S}_g = \frac{S_g - \min_{g'\in\mathcal{G}} S_{g'}}{\max_{g'\in\mathcal{G}} S_{g'} - \min_{g'\in\mathcal{G}} S_{g'} + \varepsilon}.$$

The normalized scores $\{\widetilde{S}_g\}$ are mapped back to the corresponding spatial token grid and visualized as a contribution heatmap. Regions with higher contributions indicate that occluding the corresponding local identity-token group causes a larger change in the denoising prediction. Compared with cross-attention, this perturbation-based visualization provides a more direct diagnostic of which HF-guided identity regions influence the generation dynamics.

Table 8: **Efficiency and memory profile.** We report 81-frame training and inference costs for Wan2.2-I2V-A14B and WildIcon at 832×480 resolution. Training time is reported as hours per 1K training samples. Memory is reported as peak per-GPU memory in GiB.

Method	GPUs	Train h/1K samples	Infer. s/video	Train mem.	Infer. mem.
Wan2.2	1 GPU	–	693.3	–	51.36
	4 GPUs	–	300.5	–	45.03
WildIcon	1 GPU	5.2	701.5	51.67	55.73
	4 GPUs	1.3	315.8	54.48	46.47

B.6 Efficiency and Cost Analysis

Table 8 reports the computational cost and peak GPU memory of WildIcon. All inference results are for 81-frame videos at 832×480 resolution with 50 sampling steps. Compared with the frozen Wan2.2-A14B backbone, WildIcon introduces only a small inference overhead. On a single GPU, inference time increases from 693.3s to 701.5s per video, and peak inference memory increases from 51.36 GiB to 55.73 GiB. Under the 4-GPU inference configuration, inference time increases from 300.5s to 315.8s per video, with peak per-GPU memory increasing from 45.03 GiB to 46.47 GiB.

For training, WildIcon requires 5.2 hours per 1K training samples on a single GPU and 1.3 hours per 1K training samples under the 4-GPU configuration. The peak training memory remains below 55 GiB per GPU in both settings, fitting comfortably within the 96 GiB memory budget of NVIDIA RTX PRO 6000 Blackwell GPUs. These results indicate that WildIcon adds a manageable training cost and modest inference overhead while preserving the practical efficiency of the frozen Wan2.2-A14B backbone.

C WildlifeVid: Additional Details

Table 9: Comparison of existing domain-specific video generation datasets and our WildlifeVid.

Dataset	Year	Domain	# Videos	Avg. Len (s)	Duration (h)	Resolution
UCF-101 [62]	2012	Human (Action)	13.3K	7.2	26.7	240p
NTU RGB+D [63]	2014	Human (Action)	114K	–	37	1080p
MSP-Avatar [64]	2015	Human (Action)	74	–	3	1080p
Taichi-HD [65]	2019	Human (Action)	3K	–	–	256p
SkyTimelapse [66]	2018	Sky	35K	–	–	360p
FaceForensics++ [67]	2019	Human (Face)	1K	–	–	Diverse
TikTok-v4 [68]	2023	Human (Dance)	350	–	1	–
ConsIDVid[69]	2025	Rigid Object	44.3K	8.4	104	Diverse
WildlifeVid (Ours)	2026	Wildlife	37K	9.3	96.4	Diverse

C.1 Comparison with Existing Video Datasets

Recent studies have explored a wide range of machine learning tasks and application domains [70; 71; 72; 73; 74]. In video generation, existing efforts primarily focus on collecting large, general-purpose video datasets to train video generative models [75; 76; 77]. However, domain-specific video datasets remain limited in both scale and diversity. As shown in Table 9, many existing domain-focused resources are mostly human-centric (*e.g.*, UCF-101 [62], Taichi-HD [65], FaceForensics++ [67]), making them inadequate for capturing fine-grained wildlife identity.

C.2 Video Collection

WildlifeVid is constructed by curating and re-annotating clips from four publicly accessible wildlife video sources, including Animal-in-Motion (AiM) [35], AnimalKingdom [36], MammalNet [37],

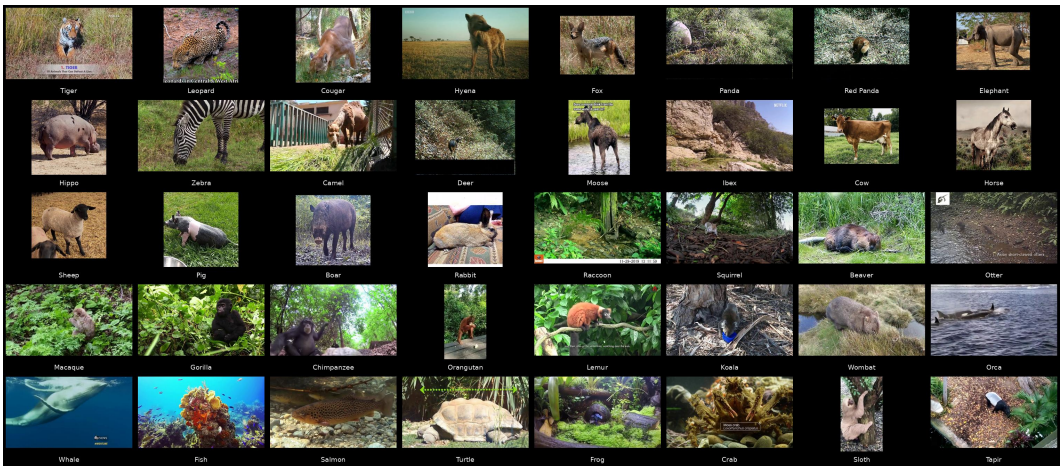


Figure 11: Random examples from WildlifeVid.

and LoTE-Animal [38]. These sources provide broad coverage of wildlife appearance, behavior, viewpoint, and environmental variation, but their original annotations are not sufficient for individual-consistent I2V training and evaluation. In particular, they are designed for tasks such as animal reconstruction, species recognition, action understanding, behavior analysis, and pose estimation, rather than individual-identity modeling. We therefore use these datasets as candidate video sources and apply a unified curation and re-annotation pipeline to obtain training-ready WildlifeVid clips.

AiM is a web-scale resource for 4D animal reconstruction, converting videos into object-centric clips with auxiliary annotations, such as masks, keypoints, and depth. The AiM benchmark contains 30K videos and 2M frames. AnimalKingdom is a large and diverse benchmark spanning 850 species across six major animal classes, with annotations for video grounding, fine-grained action recognition, and pose estimation. MammalNet contains over 18K videos totaling 539 hours, covering 173 mammal categories and 12 high-level behaviors; and LoTE-Animal focuses on understanding the behavior of endangered animals, with 10K annotated video sequences and 28K images collected over 12 years from 11 endangered species.

Each retained WildlifeVid sample is represented as a tuple:

$$(\mathcal{C}, y_{\mathcal{C}}, a_{\mathcal{C}}, \mathrm{id}_{\mathcal{C}}),$$

where $\mathcal{C}$ denotes a retained video clip, $y_{\mathcal{C}}$ is its standardized text prompt, $a_{\mathcal{C}}$ denotes the species label, and $\mathrm{id}_{\mathcal{C}}$ denotes the clip-level individual identity label. Figure 11 presents some random examples from WildlifeVid.

C.3 Licenses

WildlifeVid is constructed by curating and re-annotating clips from four existing wildlife video resources: Animal-in-Motion (AiM), AnimalKingdom, MammalNet, and LoTE-Animal. We cite the original creators of all source datasets and follow their corresponding licenses, access conditions, and terms of use. The upstream dataset licenses and access terms are summarized below.

- **Animal-in-Motion (AiM)** [35]: The official repository is released under the Apache-2.0 License and is available at the official repository.
- **MammalNet** [37]: The dataset annotations are distributed under CC BY, and the official code repository is released under the MIT License. The official repository is available at the MammalNet repository.
- **AnimalKingdom** [36]: The dataset is available through the official project page. We submitted the official download request form and obtained permission from the authors.
- **LoTE-Animal** [38]: The dataset is released under CC BY-SA 4.0 and is available through the official project page.

C.4 Data Curation Pipeline: Implementation Details

Shot splitting. Raw wildlife videos may contain shot changes, abrupt camera transitions, or multiple temporal segments with different visible individuals. We first split each source video into candidate clips with PySceneDetect [39], which detects large values of weighted average pixel change across frames. To avoid over-segmentation caused by brief visual fluctuations, we merge very short adjacent segments during shot splitting. This setting is used only to stabilize shot-boundary detection and does not impose a fixed duration threshold on the final retained clips. After this step, each candidate clip is further checked for species presence, single-subject visibility, and temporal identity consistency.

CLIP-based species filtering. We remove candidate clips that do not clearly contain the target species. For a candidate clip $\mathcal{C}$ with $L_{\mathcal{C}}$ decoded frames, we uniformly sample $n_{\mathcal{C}}$ frames over its duration and compare them with a species-level text prompt using CLIP [40; 41]. Let J_i denote the i-th sampled frame, and let $\mathrm{CLIP}_{\mathrm{img}}(\cdot)$ and $\mathrm{CLIP}_{\mathrm{text}}(\cdot)$ denote the CLIP image and text encoders, respectively. Given the source species label $a_{\mathcal{C}}$, we compute

$$s_{\mathrm{sp}}(\mathcal{C}) = \frac{1}{n_{\mathcal{C}}} \sum_{i=1}^{n_{\mathcal{C}}} \mathrm{sim}(\mathrm{CLIP}_{\mathrm{img}}(J_i), \mathrm{CLIP}_{\mathrm{text}}(\text{“a photo of } a_{\mathcal{C}}\text{”})) \,.$$

Candidate clips with $s_{\mathrm{sp}}(\mathcal{C}) < \rho_{\mathrm{sp}}$ are discarded. In our implementation, we set $\rho_{\mathrm{sp}} = 0.25$.

Single-subject enforcement. Since WildlifeVid is designed for individual-consistent I2V generation, each retained clip should contain one dominant wildlife individual without identity changes. To enforce this constraint, we use a Grounded-SAM2 tracking pipeline, where GroundingDINO-Tiny initializes text-grounded wildlife detections and SAM2 Hiera Large predicts and propagates instance masks over time. Bounding boxes are derived from the predicted instance masks. In our implementation, videos are resampled to 10 fps, new text-grounded detections are introduced every 50 frames, and the detection thresholds are set to box_threshold=0.6 and text_threshold=0.3. Very small detections are removed based on the mask-derived bounding-box area to suppress unreliable distant or spurious subjects. A candidate clip is rejected if any checked frame contains no valid wildlife instance or more than one valid wildlife instance.

Visibility and temporal stability filtering. After single-subject enforcement, we further refine the remaining clips to ensure reliable identity supervision. To remove unstable tracks, we compute the Intersection over Union (IoU) between bounding boxes in adjacent checked frames. If the IoU of any adjacent checked-frame pair falls below 0.5, the whole track is removed, since such cases usually indicate missed detections, tracking failure, abrupt shot changes, or possible identity switches. We also remove tracks in which the animal is heavily truncated by the image boundary. Specifically, if the mask-derived bounding box lies within 3 pixels of any image boundary in any checked frame, the corresponding track is discarded.

Prompt generation and cleaning. Text prompts provide controllable conditions for WildlifeVid clips, including the visible species, motion, viewpoint, and scene context. For each retained clip, we sample representative frames over its duration and query GPT-5 [78] with an instruction to describe only visible content in the sampled frames. The instruction template used for this step is shown in Figure 12. Since raw captions may contain generic template phrases, unsupported behavioral interpretations, repeated descriptions, or details unrelated to the target animal, we further clean them into standardized training prompts. Specifically, we remove non-visual or speculative content, such as inferred intent, emotion, causal explanations, or attributes that are not clearly visible in the sampled frames. We also remove irrelevant background descriptions when they do not affect the target animal or the requested motion. The final prompt is a concise factual sentence describing the visible animal, its motion, viewpoint, and scene context, *e.g.*, “a tiger walks slowly through grass in a side view”.

Identity Annotation Details WildlifeVid assigns identity labels at the clip level. After curation, each retained track is first treated as a candidate identity unit. To associate clips that contain the same individual, we perform within-species candidate grouping using visual embeddings. For each retained clip, we use the temporal midpoint frame as the representative image and extract an appearance embedding with a frozen DINOv3 ViT-H/16+ model [42]. All embeddings are ℓ_2-normalized before similarity computation. Within each species, we perform average-linkage agglomerative clustering in cosine space with a distance threshold of 0.15, corresponding approximately to candidate cross-clip merges with cosine similarity at least 0.85.

Instruction Template for Initial Wildlife Prompt Generation

System Prompt:
You are an expert wildlife video captioner. Your task is to write a short prompt-style description of a wildlife video clip using only directly visible content.

Rules:

- Write exactly 1–2 sentences under 40 words.
- Start with the animal species if identifiable.
- Describe visible motion, pose, or viewpoint.
- Optionally include brief scene context if clearly visible.
- Use concise prompt-style language rather than full narrative language.
- Do not mention camera movement, video quality, or artistic style.
- Do not speculate about identity, intent, emotion, or invisible attributes.
- Output only the prompt text.

User Prompt:
Write one short prompt-style sentence describing the wildlife clip using only visible content. Include the species, visible motion or pose, viewpoint if clear, and brief scene context if relevant. Do not mention camera behavior or make any unsupported inference.

Figure 12: Instruction template for initial wildlife caption generation.

High-similarity candidate merges are then manually reviewed. Annotators compare stable individual-specific appearance cues, including stripe topology, spot configuration, facial patches, horn shape, tail markings, contour details, and other persistent coat-pattern cues when applicable. A merge is accepted only when all the annotators agree that the clips contain the same individual; otherwise, the candidate is rejected. Accepted clips are assigned the same identity label, while rejected candidates retain their own identities.

C.5 Benchmark Protocol and Data Splits

We evaluate all models on a fixed wildlife I2V benchmark constructed from WildlifeVid. Benchmark identities are drawn from all four source datasets used to build WildlifeVid [35; 36; 37; 38]. The benchmark contains 10 species and 100 identities in total, with 10 identities per species. For each identity, we select one reference image and pair it with two prompts, resulting in 200 generation cases overall. Thus, each generation case is defined as a tuple of *(species, identity, reference image, prompt)*.

Prompt design. Each identity is paired with two prompts to evaluate identity preservation under different levels of motion and appearance variation. Prompt A contains mild motion and limited viewpoint change, such as slow forward movement, gentle body motion, or a slight head turn. Prompt B is more challenging and introduces stronger pose or viewpoint variation, such as turning around, body rotation, or larger camera movement.

Data splits and fixed evaluation protocol. The benchmark is kept fixed throughout all model comparisons and ablation studies. To avoid identity leakage between training and evaluation, all benchmark identities are excluded from the WildlifeVid training split, and the model is trained only on the remaining identities. All compared methods are evaluated using the same reference images, prompts, output resolution, and video duration.

D Limitations

This work focuses on a controlled, individual-consistent I2V setting in which the reference image contains a known wildlife individual, and the generated video is expected to preserve that individual's appearance across new poses, viewpoints, and motions. Consistent with this formulation, WildlifeVid

prioritizes single-subject, temporally coherent clips so that identity labels, prompts, and evaluation targets are well defined. This design makes the benchmark suitable for studying individual-level identity preservation. In contrast, broader full-scene wildlife generation involving multiple interacting animals, dense occlusion, or ambiguous identity association remains a complementary direction.

WildIcon operates in a single-reference setting and therefore cannot create new ground-truth identity evidence beyond what is supported by the conditioning image. When generation requires substantial viewpoint changes, particularly exposure of identity-bearing regions that are not visible in the reference, the model may extrapolate or hallucinate local markings, textures, or contours. Accordingly, individual consistency in this work refers to maintaining identity-relevant appearance supported by the reference image, rather than recovering verified appearance on unseen surfaces. Generated videos should therefore be treated as synthetic training candidates, not as genuine wildlife observations, ecological records, retrieval-database entries, or substitutes for real identity evidence.

WildIcon also depends on the visual evidence available in the reference image. The method is applicable when individual-discriminative cues, such as fur texture, stripe boundaries, spot configurations, or contour transitions, are visible in the input. When the reference image is extremely low-resolution, strongly blurred, heavily occluded, captured under very poor lighting, or depicts species with weak external individual markings, the available identity-conditioning signal is inherently limited.

E Broader Societal Impacts

Societal Impacts. WildIcon and WildlifeVid are developed to support data-scarce wildlife identification. The potential positive impact of WildIcon lies in using filtered generated samples as candidate training augmentations when real observations are limited. These samples can expose ReID models to additional appearance and motion variation during training, but they do not constitute additional field observations or new ground-truth identity evidence. Improved ReID models may in turn support wildlife monitoring workflows, while ecological inference and validation should continue to rely on real observations.

At the same time, the proposed technology may have unintended negative impacts. First, synthetic wildlife videos could be misused to fabricate evidence of animal sightings, population presence, or conservation outcomes. Such misuse may be especially harmful when generated media are presented without clear disclosure. Second, because the generated videos are synthetic observations rather than field-captured evidence, they should not be used directly to support ecological claims, such as confirming the presence, absence, movement, or population size of a species. Their intended role is to support model training and controlled evaluation, not to replace real observations or expert ecological assessment.

Safeguards and Responsible Release. We take several safeguards for responsible release. The dataset, WildlifeVid, is constructed from publicly accessible wildlife video sources and further curated into single-subject, temporally stable clips with prompts and individual-level identity labels. We do not redistribute the original videos or raw video frames. Access to source media remains governed by the respective data providers and their terms of use. We avoid releasing sensitive ecological metadata, such as precise camera locations, deployment identifiers, or fine-grained timestamps, when such information could increase the risk of unauthorized tracking or disturbance of wildlife.

For generated content, all synthetic videos released with the paper are labeled as synthetic and are intended for research, model evaluation, and downstream ReID training rather than as ecological evidence. We do not encourage the use of generated videos to support claims about real animal sightings, population presence, or conservation outcomes. Any released model weights, code, or generation examples will be accompanied by usage guidelines that prohibit presenting synthetic wildlife videos as real camera-trap footage or using the model to fabricate ecological evidence.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Limitations are discussed in Appendix D.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: Please refer to Section 5 and Appendix A.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: Please refer to the code and Appendix A for instructions of reproducing the results. The release provides code, configuration files, evaluation scripts, derived annotations, and instructions for reconstructing the dataset from permitted source assets. Raw source videos are not redistributed when restricted by source terms.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Please refer to Section 5 and Appendix A and C.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: Please refer to Appendix B. The ReID experiments are averaged over ten runs and reported with 95% confidence intervals. For I2V generation experiments, we report fixed-benchmark averages over the same 200 generation cases for all methods, since the goal is to compare methods under a controlled protocol rather than estimate sampling variance across random seeds.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Please refer to Appendix B.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: Please refer to Appendix E.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [Yes]

    Justification: Please refer to Appendix E.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: The creators or original owners of assets used in the paper are properly credited and the license and terms of use are explicitly mentioned and properly respected. Please refer to Appendix C.3.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: The paper introduces new assets, including annotations for the WildlifeVid dataset and code for the WildIcon framework, which are released with documentation and licensing information.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [Yes]

    Justification: We use LLM as a captioning tool in the dataset curation pipeline to generate initial video captions, please refer to Appendix C.4.

    Guidelines:

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