NSFA: A Multimodal Intelligence System for Automated Wildlife Exploitation Detection on Social Media

"See it. Tag it. Understand it. Act on it."

In collaboration with Magnetic Core Nature Lab & Wild Around

Technical White Paper · September 2026

Keywords. NSFA, wildlife ethics, multimodal AI, computer vision, content moderation, conservation technology, risk scoring, large language models, platform governance.

Abstract

The proliferation of wildlife-related content on social media platforms has created an urgent need for scalable, accurate, and context-aware moderation systems. This paper introduces NSFA (Not Safe For Animals), a proposed content classification and moderation standard for animal exploitation on social media, and presents NSFA, its technical implementation: a multimodal intelligence framework that integrates computer vision, natural language processing, and large language model (LLM) reasoning to detect and classify potentially exploitative human–animal interactions. Inspired by widely recognised internet content labels such as NSFW (Not Safe For Work), NSFA proposes a universally understandable category for content that may be unsafe for animals, making animal exploitation visible as a distinct digital harm category within platform governance. Unlike unimodal classifiers, NSFA combines image-level analysis with textual context (captions, profile descriptions, location metadata) to produce a structured risk score and interpretable moderation label. This paper describes the NSFA standard, the system's architecture, multi-factor risk scoring methodology, ethical framework, and potential applications in conservation monitoring and platform governance. The system is designed so that contextual integration is expected to improve classification accuracy over vision-only baselines, particularly for ambiguous edge cases involving professional conservationists and accredited sanctuaries; a formal empirical evaluation to substantiate this design goal is planned as future work (see Section 7.5).

The NSFA Standard

0.1 What is NSFA?

NSFA (Not Safe For Animals) is a proposed content classification and moderation standard for social media platforms that identifies content depicting animal exploitation, harmful captivity, wildlife trafficking indicators, unethical wildlife tourism, and other forms of animal harm. Inspired by widely recognised internet content labels such as NSFW (Not Safe For Work), NSFA introduces a simple and universally understandable category for content that may be unsafe for animals. Its purpose is to make animal exploitation visible as a distinct digital harm category rather than allowing it to remain hidden within broader moderation systems or treated as isolated incidents.

NSFA is founded on the belief that wildlife exploitation is not merely a conservation issue but also a platform governance issue. Images and videos showing captive wildlife, direct human interaction with wild animals, exotic pet ownership, animal performances, or trafficking-related activity can normalize harmful behaviour, stimulate demand, and contribute to broader wildlife exploitation networks. Research has shown that social media content can influence public perception, increase the desirability of exotic species, and indirectly facilitate illegal wildlife trade [1, 2, 15].

0.2 The Gap NSFA Addresses

In an ideal world, social media platforms would already treat animal exploitation as a dedicated moderation category, supported by automated detection, human review, and reporting pathways to appropriate conservation and law-enforcement agencies. However, despite possessing the technical capability to do so, most platforms lack consistent frameworks, standards, or transparency around this type of content. NSFA seeks to address this gap by proposing a common language, classification framework, and moderation standard that can be adopted across platforms, organisations, and communities.

NSFA does not advocate for automatic censorship of all flagged content. Instead, it promotes a graduated approach combining awareness, moderation, human review, research, and intelligence generation. The goal is to ensure that content potentially harmful to animals is identified, contextualised, and handled responsibly rather than amplified without scrutiny.

0.3 Mission and Vision

MISSION

To establish NSFA (Not Safe For Animals) as a globally recognised content classification standard for animal exploitation, enabling social media platforms, researchers, conservation organisations, journalists, and investigators to identify, assess, and respond to harmful animal-related content in a consistent and transparent manner.

VISION

A digital ecosystem where content that contributes to animal suffering, wildlife trafficking, or exploitative human–animal interactions is no longer normalised, but recognised, documented, and addressed through responsible moderation, public awareness, and evidence-based intervention.

0.4 NSFA: The Technical Implementation

NSFA is the technical research initiative and white paper that demonstrates how the NSFA standard can be implemented using modern artificial intelligence. NSFA is a multimodal intelligence system that combines computer vision, natural language processing, metadata analysis, and large language model reasoning to detect indicators of animal exploitation across social media content. By analysing images, captions, hashtags, locations, and account behaviour together, the system is designed to produce contextual risk assessments that are intended to be more informative than image-only moderation approaches; comparative evaluation of this design goal is discussed in Section 7.5.

The purpose of NSFA is not to replace human decision-making, but to demonstrate that the technical barriers to large-scale detection of animal exploitation are already low. NSFA serves as both a proof of concept and a challenge to platform operators: if open and accessible technologies can identify animal exploitation at scale, then the continued absence of animal exploitation as a recognised moderation category is no longer a technical limitation: it is a policy choice.

Introduction

1.1 Background and Motivation

Illegal wildlife trade (IWT) is a multi-billion-dollar criminal enterprise ranking among the largest illegitimate businesses globally after narcotics, and the Internet, particularly social media, has become a major marketplace for wildlife products [1]. With an estimated 2.5 billion users, platforms such as Facebook, Instagram, Twitter, and Weibo offer wildlife traders cost-effective reach, vast outreach, and relative anonymity [1], with recent evidence

documenting platform-scale facilitation of criminal wildlife trade on Facebook specifically [23]. Research by World Animal Protection (2019) identified over 550,000 posts on a single platform depicting wildlife in tourist entertainment contexts, the majority originating from venues associated with animal cruelty [2]. Empirical studies have documented active trade across platforms: Xu et al. (2019) collected 138,357 tweets over 14 days and identified 53 posts from 38 unique users suspected of promoting ivory sales [3]; a parallel study of Facebook Chinese-language content collected 10,303 unique posts over 45 days and identified 639 signal posts from 268 unique users directly marketing wildlife products including elephant ivory, hawksbill shell, and rhino horn [4].

The massive volume of user-generated content makes manual monitoring of these platforms untenable. Human moderators exhibit significant variance in judgement, particularly when content superficially resembles ethical wildlife education but involves covert exploitation. Automated systems relying solely on image classification fail to account for critical contextual signals embedded in accompanying text, poster credibility, or geographical metadata. This has motivated the development of multimodal intelligence systems that combine computer vision, natural language processing, machine learning, and structured reasoning to automatically detect and classify wildlife exploitation at scale.

1.2 Problem Statement

Existing content moderation pipelines are predominantly unimodal, operating on either visual features or textual content in isolation. This architectural limitation produces high false-negative rates for exploitative content that presents visually benign, such as a photograph of a person standing near a chained tiger described as “wildlife volunteering”. Conversely, false positives arise when legitimate veterinary documentation is misclassified as exploitation due to superficial visual similarity to harmful content.

1.3 Contributions

This paper makes the following primary contributions:

1. A multimodal risk assessment framework integrating vision, language, and metadata signals.
2. A structured, interpretable risk scoring rubric informed by known ethical and unethical wildlife interaction patterns.
3. A facility knowledge base cross-referencing post location against catalogued sanctuaries and exploitative venues.
4. A human-in-the-loop moderation interface enabling auditable review workflows.

Related Work

2.1 Wildlife Monitoring with Computer Vision

Prior work in automated wildlife monitoring has primarily focused on species identification and population estimation in controlled ecological contexts. Norouzzadeh et al. (2019) trained a CNN (VGG-16) to classify 20 African wildlife species from camera-trap images with 87.5% accuracy using 111,467 images, employing Grad-CAM to reveal that the model learned species-specific visual features [5]. Follmann et al. (2021) demonstrated an end-to-end pipeline combining Mask R-CNN for animal detection and (2+1)D ResNet for action recognition in camera-trap videos, achieving 63.8% average precision for detection and up to 94.1% accuracy for action recognition [6]. These techniques are directly transferable to detecting wildlife products (e.g., pangolin scales, rhino horn, ivory) in social media images and videos. However, such models are not designed to assess ethical dimensions of human-animal interactions, nor do they process accompanying social context.

2.2 Multimodal Content Moderation and NLP

Di Minin et al. (2018) proposed a framework for investigating illegal wildlife trade on social media using machine learning, outlining how NLP can infer the meaning of verbal descriptions, detect species and locations, and classify sentiment toward illegal wildlife trade [1]. Xu et al. (2019) demonstrated an unsupervised biterm topic model (BTM) combined with keyword filtering to detect wildlife product promotion on Twitter, classifying posts into topic clusters and identifying signal messages based on trading language and e-commerce references [3]. A critical finding is that 47% of wildlife selling posts on Facebook used code words instead of species keywords, with some posts replacing text entirely with animal graphic emojis [4], underscoring the need for adaptive NLP beyond keyword matching.

Yang et al. (2017) developed a directly analogous multimodal approach for detecting illicit drug dealing on Instagram, using deep learning image classification with decision-level fusion of image and text classifiers, achieving over 88% accuracy on drug-related post recognition [7]. They also extracted account-level behavioral features (including percentage of illicit posts, temporal patterns, and follower/following ratios), achieving an F1-score of 0.51 for dealer account detection [7]. This methodology is directly transferable to wildlife exploitation detection. Hateful content detection has similarly benefited from joint visual-textual architectures such as VisualBERT [8] and CLIP [9], which outperform unimodal counterparts on nuanced classification tasks.

2.3 LLM Fine-Tuning for Wildlife Trafficking Detection

A recent and significant advance in the field is WildLLM (Antala, Mou, & Lee, 2026), a text-based framework that fine-tunes large language models using parameter-efficient methods (LoRA) to detect wildlife product trading (WLT) posts on social media, with a focus on ivory trafficking on Twitter/X [20]. The study demonstrates that out of 138,357 collected social media posts, only 53 were confirmed WLT posts (a WLT post rate of just 0.038%), highlighting the severe class imbalance that makes detection particularly challenging. WildLLM addresses this through: (i) LLM-based data augmentation to generate realistic synthetic WLT posts, alleviating class imbalance; (ii) in-context learning and prompt engineering to improve augmentation realism; and (iii) LoRA fine-tuning of LLaMA-3.1-8B and Qwen2.5-7B. The framework achieved 0.854 MCC and 0.927 Macro F1, a 21% MCC and 8.7% Macro F1 improvement over the best baseline [20]. Crucially, WildLLM also demonstrates that traffickers deliberately use euphemisms such as “white gold” for ivory, spelling variants, emojis, and code words, confirming that straightforward keyword filters produce both excessive false positives and false negatives.

The WildLLM findings directly motivate NSFA's multimodal approach: purely text-based models, even when fine-tuned, remain vulnerable to novel obfuscations, whereas NSFA's joint visual and textual reasoning is designed to provide a complementary signal that persists even when captions are deliberately obscured, a hypothesis that has not yet been empirically tested (see Section 7.5). The LoRA fine-tuning methodology described in WildLLM represents a promising future direction for NSFA's caption analysis module.

2.4 Network Analysis and LLM-Assisted Reasoning

Beyond individual post detection, van der Merwe et al. (2015) demonstrated how social network analysis (SNA) can identify key players in wildlife trafficking networks using centrality measures computed from communication data, generating actionable intelligence that specifies optimal arrest sequences and predicts network resilience [10]. Applied to a rhino horn trafficking case in Kruger National Park, their system identified key intermediaries whose removal would maximally disrupt the criminal network [10].

The emergence of instruction-tuned large language models with vision capabilities (GPT-4o, Gemini 1.5 Pro) has enabled a new class of moderation architectures in which structured reasoning is elicited via carefully designed prompts [11]. NSFA employs this paradigm, using chain-of-thought prompting to decompose the moderation task into discrete analytical sub-tasks, with the goal of improving both accuracy and interpretability; the accuracy contribution of this design choice remains to be empirically validated (Section 7.5).

System Architecture

NSFA is structured as a five-stage processing pipeline that transforms raw social media content into a moderation decision. Figure 1 illustrates the end-to-end data flow.

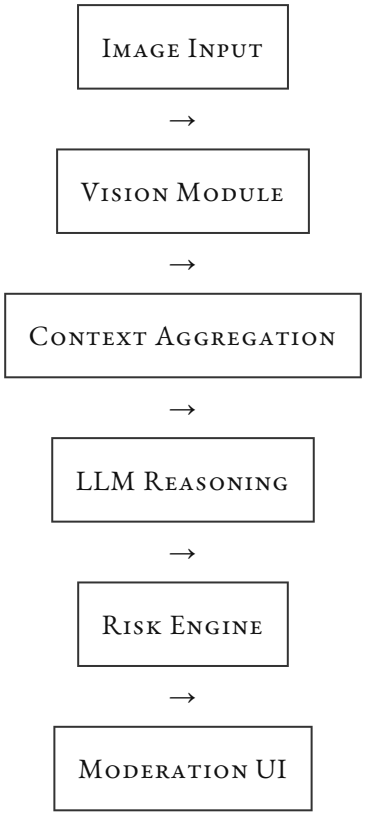


Figure 1. NSFA processing pipeline.

3.1 Vision Module

The Vision Module processes the input image using a multimodal vision-language model. It performs: (i) species identification and taxonomic classification; (ii) human presence detection and pose estimation; (iii) environment classification across seven categories (wild, zoo, domestic, tourism, urban, studio, unknown); and (iv) interaction type characterization (e.g., physical contact, confinement, riding, posing).

3.2 Context Aggregation Module

Prior to LLM inference, available textual and metadata signals are aggregated. These include: the post caption, hashtags, and any embedded text; the poster's profile description; and the tagged or inferred post location. These inputs are serialized into a structured prompt template, ensuring consistent representation across all inference requests.

3.3 Multimodal Reasoning Layer

The core reasoning component invokes a large language model with both the image and the aggregated context. A structured chain-of-thought prompt guides the model through five analytical sub-tasks: visual analysis, caption interpretation, poster credibility assessment, facility ethics evaluation, and overall risk synthesis. The model returns a structured JSON response conforming to a predefined schema, ensuring downstream compatibility.

3.4 Risk Scoring Engine

The Risk Scoring Engine applies an expert-designed additive rubric to the model's structured output, producing a continuous risk score in the range [0.0, 1.0]. The rubric is detailed in Section 4.

3.5 Moderation Interface

Analyses are persisted to a structured database and surfaced in a human review interface. Moderators may approve, restrict, or escalate content, appending notes that feed back into system refinement.

Risk Scoring Methodology

NSFA employs a transparent, additive risk rubric. Each detected factor contributes a fixed increment (or decrement) to a baseline score of 0.0, with the final value clamped to [0.0, 1.0].

Table 1. Risk Scoring Rubric

Risk Factor	***Δ Score***	***Rationale***
Wild animal in non-wild environment	**+0.20**	Indicates removal from natural habitat
Human physically interacting with wild animal	**+0.30**	Direct exploitation signal
Animal dressed in outfit or costume	**+0.30**	Anthropomorphisation and welfare concern
Tourism or staged setting detected	**+0.20**	Commercial exploitation indicator
Suspicious caption keywords	**+0.20**	Linguistic markers of trade or trivialisation
Known unethical facility	**+0.40**	High-confidence institutional risk factor
Verified veterinarian / conservationist	**−0.20**	Professional credential mitigates risk
Accredited ethical sanctuary	**−0.20**	Institutional legitimacy mitigates risk
Cartoon / animation detected	**→ 0.00**	Non-real content; exempt from scoring

4.1 Classification Thresholds

The continuous risk score is mapped to five ordinal classification labels:

ethical	*score* ∈ *[0.00, 0.20)*
likely_ethical	*score* ∈ *[0.20, 0.40)*
uncertain	*score* ∈ *[0.40, 0.55)*
likely_unethical	*score* ∈ *[0.55, 0.75)*
unethical	*score* ∈ *[0.75, 1.00]*

4.2 Extended Risk Scoring Framework

To capture a broader spectrum of exploitative behaviours, the NSFA risk model is extended to include additional high-confidence indicators beyond the core rubric.

Table 2. Extended Risk Scoring Rubric

Additional Risk Factor	***Δ Score***	***Rationale***
Wildlife-derived products (fur, ivory, skins)	**+0.40**	Indicator of commercial exploitation
Hunting trophies / deceased animals	**+0.40**	Evidence of lethal interaction
Non-accredited captivity	**+0.30**	High probability of welfare violations
Unregulated animal transport	**+0.30**	Potential trafficking signal
Exotic pet ownership signals	**+0.30**	Demand-generation indicator
Repeated exposure patterns	**+0.20**	Behavioural reinforcement risk
Trafficking-related language	**+0.40**	Strong linguistic exploitation signal

Facilities are further classified into three tiers:

- **Accredited (low risk):** AZA-accredited institutions, PAWS-certified rehabilitation centres, Born Free Foundation facilities.
- **Unverified (medium risk):** Facilities with no documented accreditation or regulatory oversight.
- **High-risk / Blacklisted:** Roadside zoos, entertainment venues, travelling circuses, trafficking-linked facilities.

The output schema has also been extended to capture a richer set of signals per analysis. The following fields are now persisted alongside the core risk score:

Table 3. Extended Output Schema Fields

Field	*Type*	*Description*
`wildlife_product_types`	string[]	Specific products identified (e.g. ivory, pangolin scales, rhino horn)
`transaction_intent_explicit`	boolean	True if an explicit offer to buy, sell, or ship wildlife is detected
`evasion_tactics_detected`	boolean	True if code words, emoji substitution, or private-channel redirects are found
`evasion_tactics_notes`	string	Human-readable description of the specific evasion technique identified
`caption_sentiment`	enum	Inferred sentiment of the caption: positive / neutral / negative / ambiguous

4.3 Updated Risk Function

The composite risk score is computed as a weighted linear combination of feature groups:

$$r = \alpha V + \beta C + \gamma M + \delta H$$

V: visual features (species, environment, interaction type)

C: contextual features (caption, location, facility)

M: multimodal reasoning output (LLM-synthesised score)

H: historical behaviour (repeated exposure patterns, account history)

Weights α, β, γ, δ are initially set by expert judgement and sum to 1.0, with empirical calibration against labelled data planned as future work (see Section 7.5). The final score is clamped to [0.0, 1.0].

Contextual Analysis Modules

5.1 Advanced Caption and Language Analysis

A critical limitation of earlier moderation systems is their reliance on naive keyword matching for textual analysis. Wildlife traffickers and exploitative content creators have rapidly adapted to platform moderation by developing an evolving lexicon of code words, alternative terminology, and emoji-based substitutions designed to evade automated detection.

NSFA addresses this through structured multimodal NLP prompting that instructs the reasoning model to:

- **Detect code words and alternative terminology:** Terms such as “baby”, “tame”, “hand-raised”, “available”, “rehoming”, “collection”, and “specimen” frequently appear in exploitation contexts and are now explicitly flagged.
- **Identify emoji-based species substitution:** Sellers routinely substitute species names with animal emojis to bypass text filters. The system treats these signals as potential transaction indicators when combined with other contextual cues.
- **Infer transaction intent:** The system distinguishes “sale offer” posts (containing pricing, availability, shipping details) from observational wildlife content, producing a structured `transaction_intent_explicit` boolean field.
- **Detect private channel redirection:** Phrases such as “DM for details”, “WhatsApp me”, “contact for pricing”, and “link in bio” are recognised as strong evasion signals, populating the `evasion_tactics_detected` field and `evasion_tactics_notes` with a human-readable explanation.
- **Sentiment analysis:** The overall sentiment of the caption towards wildlife trade is classified as positive, neutral, negative, or ambiguous, providing an additional signal for risk stratification.

5.2 Specific Wildlife Product Identification

Earlier versions of the system detected wildlife-derived products as a binary flag. The updated model now performs granular product identification, returning a structured list of specific items identified, such as ivory, rhino horn, pangolin scales, bushmeat, or exotic skins. This shift from boolean to typed enumeration enables downstream analysis of which specific product categories are most prevalent, supporting targeted conservation policy responses and more precise risk scoring.

5.3 Poster Credibility Assessment

The system evaluates the poster's profile description to assess professional legitimacy. Indicators of verified expertise, such as affiliation with recognised wildlife organisations (WWF, IUCN, AZA), veterinary credentials, or peer-reviewed research activity, are associated with risk reduction. Conversely, commercial signals (e.g., “exotic pet breeder”, “wildlife experience tours”) increase the assessed risk. Profiles providing no description or vague self-identification alongside wild animal imagery result in elevated uncertainty.

5.4 Facility and Location Intelligence

When a post location is provided, the system cross-references it against a curated knowledge base of known facilities. Facilities classified as unethical include travelling circuses, roadside zoos, venues with documented cruelty violations (e.g., Tiger Temple, Thailand; Surabaya Zoo, Indonesia), and organisations listed on international watchdog databases. Ethical reference facilities include AZA-accredited institutions, Born Free Foundation sanctuaries, and PAWS-certified wildlife rehabilitation centres. A match with a known unethical facility adds +0.40 to the risk score and triggers the `known_unethical_facility` tag.

5.5 Animated and Non-Photographic Content

To prevent false positives in artistic contexts, the Vision Module first determines whether the submitted image is photographic or illustrative. Cartoons, CGI renderings, and illustrated artworks are immediately classified as ethical with a risk score of 0.0, bypassing all subsequent moderation stages.

Account-Level Behavioural Analysis

Post-level classification, while necessary, is insufficient to disrupt well-organised trafficking networks. Sellers operating on social media routinely spread activity across multiple posts, use coordinated accounts, and adapt their language iteratively in response to platform enforcement. NSFA's architecture is designed to evolve toward account-level behavioural profiling as a complementary analytical layer.

6.1 Enhanced Profile Description Analysis

In the current implementation, the poster's profile description is analysed for behavioural signatures that signal commercial intent or systematic exploitation, even when not explicitly stated. Indicators include language advertising specific species, phrases encouraging private contact (“DM for availability”), commercial shipping references, or the promotion of “collections” and “breeding programmes” outside accredited contexts. These signals refine the `is_verified_professional` and `professional_assessment` fields and contribute to the contextual risk sub-score.

6.2 Historical Data Integration (Roadmap)

A full account-level risk profile requires aggregating a user's historical analysis records prior to each new inference. In practice, this means querying prior `Analysis` records associated with the same account and passing a structured summary to the LLM as additional context, for

example: “This account has submitted 12 prior analyses; 8 were flagged as likely_unethical; average risk score: 0.74; frequent species: pygmy slow loris, ball python.” This history-augmented context would substantially increase the accuracy of the `risk_score_llm` sub-score for repeat actors.

6.3 Proposed Account Behaviour Metrics

The following persistent metrics are proposed for a `UserBehaviourProfile` entity or extension of the built-in User entity, enabling longitudinal risk profiling:

Table 4. Proposed Account-Level Behavioural Metrics

Metric	***Description***
`total_posts_analyzed`	Total number of content items submitted or ingested for this account
`total_flagged_posts`	Count of items labelled likely_unethical or unethical
`average_risk_score`	Rolling average composite risk score across all analyses
`frequent_species_posted`	List of species appearing most frequently in the account's content
`evasion_pattern_detected`	Boolean: does the account repeatedly employ code words, emoji substitution, or private redirects?
`transaction_intent_rate`	Proportion of posts containing explicit transaction intent signals

6.4 From Post Detection to Network Disruption

The ultimate goal of account-level analysis is to shift the system's orientation from reactive post classification toward proactive network disruption. By identifying clusters of coordinated accounts exhibiting correlated risk patterns, NSFA can surface trafficking networks rather than isolated offenders, providing enforcement agencies and platform trust-and-safety teams with intelligence at the scale required to meaningfully disrupt supply chains.

Ethical Framework and Limitations

7.1 Transparency and Interpretability

Every moderation decision produced by NSFA is accompanied by an AI-generated reasoning statement, a structured tag set, and the complete risk score breakdown. This design ensures that moderators and content creators can understand and contest automated decisions, a prerequisite for ethically deployable AI systems [14].

7.2 Human-in-the-Loop Design

NSFA is explicitly designed as a decision-support tool, not an autonomous enforcement mechanism. All high-risk classifications are routed to human moderators for review. The system supports three moderation outcomes: Approved, Restricted, and Escalated, each associated with a reviewer-authored note for accountability.

7.3 Cultural and Contextual Bias

The risk rubric reflects predominantly Western conservation norms and may require regional calibration for contexts where human-animal relationships are governed by distinct cultural or subsistence frameworks. Future work will address this limitation through geographically stratified dataset curation and rubric localisation.

7.4 Model Limitations

As with all LLM-based reasoning systems, NSFA is susceptible to hallucination, particularly in facility identification and species taxonomy. The structured JSON output schema mitigates this partially by constraining the output space. Ongoing human review and feedback loops are essential for maintaining system reliability.

7.5 Limitations of Current Validation

Evidentiary status

NSFA is presented in this paper as a proof-of-concept and a documented methodology, not as a validated production system. To date, no formal empirical evaluation has been conducted against an annotated benchmark dataset: there is no ablation study comparing the full multimodal pipeline against a vision-only baseline, no inter-rater reliability assessment of the risk-scoring rubric, and no measurement of precision, recall, or false-positive rates on a held-out test set. The risk-scoring weights presented in Section 4 are informed by domain literature and expert intuition but have not yet been statistically calibrated against labelled data. Accordingly, all claims in this paper about the system's accuracy or its superiority over unimodal approaches should be read as design objectives rather than measured results. A rigorous evaluation plan (including dataset construction, annotation protocols, baseline comparisons, and rubric calibration) is the primary prerequisite for future development and is outlined in Section 9.

7.6 The Impact of Wildlife Exploitation Imagery on Global Trafficking Dynamics

The proliferation of wildlife-related imagery on social media platforms has measurable downstream effects on conservation outcomes, particularly through its influence on public perception, demand formation, and illegal wildlife trade dynamics.

7.6.1 Perceptual Normalization and Cognitive Bias

Empirical studies demonstrate that the depiction of wild animals in anthropogenic contexts, such as proximity to humans or within domestic-like environments, alters public perception of species status and welfare. Exposure to images of primates in human settings has been shown to increase the likelihood that viewers perceive such species as non-endangered and suitable as pets [15]. This phenomenon contributes to:

- misclassification of conservation status by the general public
- normalization of inappropriate human–animal interactions
- attenuation of perceived ethical concern

7.6.2 Demand Formation and Behavioral Signaling

Social media content functions as a behavioral signaling mechanism, shaping user aspirations and consumption patterns. Influencer-driven exposure to exotic species can transform wildlife into objects of desirability, thereby increasing demand for exotic pets, wildlife tourism experiences, and animal-derived luxury goods. This demand-side pressure is a recognised driver of wildlife exploitation and trade [1].

7.6.3 Indirect Facilitation of Illegal Wildlife Trade

Even in the absence of explicit commercial intent, wildlife-related content can act as implicit advertising for trafficking networks. Increased visibility and normalization reduce social stigma and facilitate market expansion, price inflation for rare species, and recruitment of new participants in illicit trade. The United Nations Office on Drugs and Crime (UNODC) identifies online platforms as key enablers in modern wildlife trafficking ecosystems [16].

7.6.4 Policy Implications

These findings underscore the necessity of moderation systems that address indirect and upstream drivers of harm, rather than focusing solely on explicit abuse. NSFA operationalises this approach through contextual risk assessment, multimodal interpretation, and early-stage intervention mechanisms.

Applications and Recommendations

NSFA is designed for deployment across the following use cases:

- **Social Media Platform Moderation:** Integration via API into platform review queues to pre-screen wildlife content before publication or during appeals.
- **Conservation NGO Monitoring:** Systematic surveillance of hashtags and accounts associated with suspected wildlife trafficking or tourism exploitation.
- **Regulatory Compliance:** Support for government agencies enforcing CITES provisions related to digital media [17].
- **Academic Research:** Provision of structured, labelled datasets for downstream research in animal welfare AI and multimodal ethics classification.
- **Public Awareness:** Educational tools that provide transparent explanations of why specific content is flagged, raising platform-user awareness of ethical wildlife engagement.

8.1 Recommendations for Social Media Platform Integration

To achieve meaningful impact, NSFA should be deployed within a hybrid human–AI moderation framework that integrates automated detection with user education and human oversight.

8.1.1 Hybrid Moderation Architecture

We propose a three-tiered system:

1. **Automated Pre-Screening:** All user-generated content is evaluated using NSFA at upload, producing a calibrated risk score that routes content to the appropriate tier.
2. **Awareness-Based User Intervention:** For intermediate-risk content, platforms should implement non-punitive interventions including contextual warning labels, educational overlays, and links to conservation resources. This aligns with established best practices in behavioural nudging and digital harm reduction [18].
3. **Human Moderation Escalation:** High-risk cases are routed to trained moderators for final decision-making, ensuring contextual interpretation, accountability, and contestability of outcomes.

8.1.2 Awareness-Driven Moderation

NSFA introduces a paradigm shift from punitive enforcement toward educational moderation, where the primary objective is to inform users, reshape behaviour, and increase awareness of ethical wildlife interactions. This approach is particularly relevant for ambiguous cases where intent is unclear but potential harm exists.

8.1.3 Integration into Platform Infrastructure

NSFA can be embedded into existing platform systems, including:

- content ranking algorithms (reducing visibility of exploitative content)
- account scoring systems (tracking repeated high-risk behaviour)
- recommendation engines (preventing amplification of harmful content)

8.1.4 Toward Industry Standardisation

Currently, wildlife exploitation is not systematically addressed in platform governance frameworks. We propose its inclusion as a standard moderation category, comparable to hate speech, adult content, and violent material. This would enable cross-platform consistency, shared benchmarks, and regulatory alignment.

8.2 Real-World Precedent: ECO-SOLVE Global Monitoring System

The viability of AI-powered wildlife trade monitoring at scale is demonstrated by the ECO-SOLVE initiative, an EU-funded programme led by the Global Initiative Against Transnational Organized Crime [21]. Its Global Monitoring System (GMS) connects data hubs in six key countries (Brazil, Colombia, Thailand, Indonesia, Nigeria, and South Africa) to track illicit wildlife markets across social media and e-commerce platforms. By 2025, GMS had detected 5,791 illicit postings featuring live animals, lion pelts, whale teeth, ivory, and more [21]. In a focused investigation of pirarucu (an Amazon fish), ECO-SOLVE identified 150 illicit trade advertisements from 74 unique sellers within the first three months alone, with 65% of advertisements originating from just 20 sellers, many of which were registered businesses [21].

ECO-SOLVE's operational experience directly reflects the challenges NSFA is designed to address.

> *"They constantly change the way they advertise, using new keywords, emojis or coded language. And they hop between different platforms or move conversations onto encrypted messaging apps like WhatsApp or Telegram to avoid detection."*
>
> Russell Gray, Head of ECO-SOLVE Data Component [21]

This observation reinforces the necessity of NSFA's evasion-detection capabilities and the critical importance of the `evasion_tactics_detected` field. Notably, ECO-SOLVE has announced future plans to incorporate AI image recognition, directly aligning with NSFA's current multimodal architecture.

8.3 Government and Regulatory Applications

NSFA provides a foundation for data-driven regulatory oversight and cross-sector collaboration in wildlife protection.

8.3.1 Regulatory Signaling and Early Warning Systems

At scale, NSFA enables the identification of emerging exploitation trends, high-risk species, and geographic hotspots. These insights can support organisations such as CITES [17], the INTERPOL Environmental Crime Programme, and national enforcement agencies.

8.3.2 Geospatial Mapping of Exploitation Networks

Aggregated platform data can be transformed into geospatial intelligence systems, enabling identification of exploitative facilities, mapping of tourism-related abuse clusters, and detection of trafficking corridors. Such maps can complement existing tools such as the TRAFFIC wildlife trade monitoring network [19] and IUCN conservation databases.

8.3.3 Platform–Government Data Collaboration

A privacy-preserving collaboration model should be implemented, where platforms share anonymized, aggregated risk data and trend-level analytics, while ensuring compliance with data protection regulations and prevention of surveillance misuse.

8.3.4 Enforcement and Legal Support

Structured outputs from NSFA can assist in identifying repeat offenders, documenting behavioural patterns, and linking digital content to physical locations. These capabilities enhance investigative workflows, evidentiary pipelines, and cross-border enforcement co-ordination.

8.3.5 Societal Impact

By integrating AI systems, social platforms, and regulatory bodies, NSFA contributes to a global digital infrastructure for wildlife protection, enabling proactive intervention, scalable monitoring, and improved conservation outcomes.

Challenges and Future Directions

Despite the demonstrated capabilities of multimodal wildlife exploitation detection reported in the literature, several significant challenges remain unresolved, both in NSFA's current implementation and in the broader field.

9.1 Evolving Evasion Tactics

Sellers continuously adapt their language to evade detection, using code words, emojis, and indirect language [4]. As documented by Xu et al. (2020), 47% of wildlife selling posts used code words instead of species keywords. NLP models must continuously adapt through active learning and model retraining to keep pace with these evolving strategies. The current NSFA implementation already detects known evasion patterns; however, generalization to novel obfuscation techniques requires ongoing human-in-the-loop feedback.

9.2 Data Access and Platform Restrictions

Social media platforms have progressively restricted API access following privacy scandals, making systematic data collection increasingly difficult [1, 12]. Much wildlife trade migrates to private groups or encrypted messaging applications (WhatsApp, WeChat), which are effectively opaque to external monitoring [12]. Posts are often self-deleted after a completed sale, requiring near real-time detection and archiving pipelines [3]. Stringham et al. (2021)

provide a six-step framework for internet-based wildlife trade monitoring that addresses these challenges through platform-specific API strategies and web scraping, approaches that would substantially expand NSFA's data ingestion capabilities in a future version [12].

9.3 Training Data and Model Generalizability

Deep learning models require large, consistently labeled training datasets. While general-purpose datasets provide initial training for species recognition, more specific labeled data is needed for wildlife products (pangolin scales, rhino horn, bushmeat) [1]. CNNs can also learn spurious correlations, such as associating background vegetation with particular species, that degrade generalizability [5]. Crowd-sourced initiatives involving domain experts are needed to create sufficiently specific and geographically diverse training datasets.

9.4 Network-Level Detection (Roadmap)

The current system operates at the post and account level. A comprehensive intelligence system must extend to network disruption, identifying clusters of coordinated accounts exhibiting correlated risk patterns. The social network analysis methodology demonstrated by van der Merwe et al. (2015), using eigenvector centrality, betweenness centrality, and brokerage scores to identify key trafficking intermediaries, represents the target architecture for a future NSFA network analysis module [10].

9.5 Multi-Language and Temporal Capabilities (Roadmap)

Current studies and the NSFA implementation are predominantly English-language, but wildlife trade on social media occurs globally across multiple languages [3, 4]. Future versions should integrate multi-language NLP capabilities. Additionally, Lamba et al. (2019) highlighted the potential of real-time monitoring systems using Internet-of-Things (IoT) technology and automated acoustic classification [13], directions that would complement NSFA's image-based detection with broader environmental sensing.

9.6 Verification of Illegality

Determining whether a detected transaction is actually illegal requires knowledge of jurisdiction-specific laws, CITES appendix listings, and permit systems [1, 3]. Not all wildlife trade is prohibited, and distinguishing legal from illegal transactions remains an open challenge requiring integration of regulatory databases and legal expertise alongside AI detection.

9.7 Intelligence-Led Enforcement: Field Trial Evidence

Beyond online detection, a 2026 Australian field trial by Meagher et al. (Taronga Conservation Society Australia / UNSW) demonstrated how machine learning tools can be integrated into physical enforcement operations, enabling intelligence-led interdiction of illegal wildlife shipments [22]. The Detect and Trace study used ML-based screening systems in combination with investigator intelligence to improve seizure rates, providing evidence that AI-assisted detection can translate into real-world enforcement outcomes when integrated with operational workflows [22]. This represents a relevant proof point for the enforcement pipeline that NSFA's structured output is designed to feed: from online content flagging → moderation escalation → law enforcement intelligence briefs → physical interdiction. Future versions of NSFA should incorporate structured export formats compatible with enforcement agency intelligence workflows.

9.8 Structural Barriers and the Deployment Paradox

A May 2026 GI-TOC report synthesising findings from a multi-stakeholder Bangkok workshop on AI and environmental crime enforcement identifies a critical "deployment paradox" directly relevant to NSFA's roadmap: AI is most valuable in under-resourced, data-poor enforcement contexts, yet these same conditions make responsible and effective deployment significantly more difficult [24]. The report's key findings illuminate several structural challenges that constrain any AI-based wildlife monitoring system:

- **Structural barriers outweigh technical barriers.** Weak data foundations, limited digitisation, fragmented governance, and legal uncertainty determine success or failure more decisively than the sophistication of the tools themselves [24]. For NSFA, this underscores the importance of investing in annotation infrastructure and regulatory alignment alongside model development.
- **Human-in-the-loop is non-negotiable.** The GI-TOC workshop confirmed that AI systems are most reliable and actionable when they augment rather than replace human judgement, with clear human review mechanisms at decision points [24]. NSFA's moderation interface is designed with this principle as a first-class constraint.
- **Transnational crime requires cross-border AI architectures.** Environmental crime operates across jurisdictions, but most AI systems are nationally designed, creating interoperability and data-sharing gaps [24]. A future NSFA deployment should prioritise API-level integration with cross-border enforcement intelligence platforms such as INTERPOL's I-24/7 and the GMS dashboard.

- **Bias toward visible, small-scale actors.** If not carefully designed, AI systems may disproportionately surface low-level content while overlooking coordinated criminal networks operating through obfuscation [24]. NSFA's account-level behavioural analysis module (Section 6) is specifically designed to counteract this bias by identifying network-level patterns rather than isolated posts.
- **AI as a public good.** The report argues that open, collaborative approaches, with shared data and infrastructure, are critical to preventing institutional dependency and ensuring equitable access [24]. NSFA's fully documented methodology and access-by-request model directly embody this principle.

These findings from practitioner experience in diverse enforcement contexts provide independent external context for NSFA's design choices and highlight priority areas for future development: data infrastructure investment, cross-jurisdictional interoperability, and rigorous human oversight protocols.

Conclusion

This paper has presented NSFA, a multimodal intelligence system for the automated detection of wildlife exploitation in social media content. By integrating computer vision, natural language processing, structured LLM reasoning, and facility intelligence, the system is designed to achieve a level of contextual understanding not attainable by unimodal approaches. The additive risk scoring rubric provides transparency and auditability, while the human-in-the-loop moderation interface ensures that automated decisions remain contestable and accountable.

NSFA extends beyond a technical moderation system to serve as a behavioural intervention and policy-support framework. By transforming social media content into structured, actionable intelligence, it aims to enable both platforms and regulatory bodies to address wildlife exploitation through proactive, context-aware, and scalable mechanisms.

This project exists as a functional, working proof-of-concept. Its implementation, built in a matter of days using commercially available multimodal AI and a modest engineering effort, suggests that the technical barrier to deploying wildlife exploitation detection at scale is low, though this claim, like the system's accuracy claims, has not yet been formally validated (Section 7.5). The methodology is documented in full in this paper, and the cost of the proof-of-concept is negligible relative to the resources commanded by the world's largest social media corporations.

The continued absence of wildlife exploitation as a recognised moderation category in major platform governance frameworks raises a question this paper cannot answer on its own: whether that absence reflects a genuine technical limitation or a policy choice. NSFA is

offered as a proof of concept and a starting point for that conversation, and the NSFA standard (Not Safe For Animals) is proposed as common language for an industry that has not yet created one for itself. See it. Tag it. Understand it. Act on it.

Future work will focus on: (i) empirical evaluation on large-scale annotated wildlife datasets; (ii) fine-tuning of species classification and facility recognition modules; (iii) multi-language caption analysis; (iv) integration with platform APIs for real-time ingestion pipelines; and (v) longitudinal study of the system's impact on downstream trafficking indicators. NSFA represents an early step toward technology-enabled global wildlife protection at scale, contingent on the validation work outlined in Section 7.5.

Access & Licensing

NSFA is released as a documented methodology and a hosted proof-of-concept, not as a self-hostable codebase. The complete system architecture, risk scoring rubric, prompt design, and structured output schema are published in full in this white paper, enabling any researcher or engineer to understand, critique, and independently reimplement the approach.

The running system itself is operated as an independent initiative and is made available to partner organisations (platforms, conservation NGOs, researchers, and enforcement agencies) on request. This model keeps the live instance maintained and consistent, avoids the fragmentation and drift inherent in public forks, and ensures that interested parties engage through a single, accountable channel rather than maintaining divergent copies.

Organisations interested in piloting NSFA, integrating its outputs, or collaborating on further development are invited to contact the authors. The methodology described herein is shared openly as a research contribution; the hosted service and its underlying implementation remain under the stewardship of the NSFA team.

References

[1] Di Minin, E., Fink, C., Hiippala, T., & Tenkanen, H. (2018). A framework for investigating illegal wildlife trade on social media with machine learning. *Conservation Biology*, 33(1), 210–213. doi:10.1111/cobi.13104

[2] World Animal Protection. (2019). *Wildlife Selfie Code: Social Media and Wildlife Tourism.* World Animal Protection International. worldanimalprotection.org

[3] Xu, Q., Li, J., Cai, M., & Mackey, T. K. (2019). Use of Machine Learning to Detect Wildlife Product Promotion and Sales on Twitter. *Frontiers in Big Data*, 2, 28. doi:10.3389/fdata.2019.00028

[4] Xu, Q., Cai, M., & Mackey, T. K. (2020). The illegal wildlife digital market: an analysis of Chinese wildlife marketing and sale on Facebook. *Environmental Conservation*, 48(1), 58–66. doi:10.1017/s0376892920000235

[5] Norouzzadeh, M. S., et al. (2019). Insights and approaches using deep learning to classify wildlife. *Scientific Reports*, 9, 8137. doi:10.1038/s41598-019-44565-w

[6] Follmann, A., et al. (2021). Identification of animals and recognition of their actions in wildlife videos using deep learning techniques. *Ecological Informatics*, 61, 101215. doi:10.1016/j.ecoinf.2021.101215

[7] Yang, Y., et al. (2017). Tracking Illicit Drug Dealing and Abuse on Instagram Using Multimodal Analysis. *ACM Transactions on Intelligent Systems and Technology*, 8(4), 58. doi:10.1145/3011871

[8] Li, L. H., et al. (2020). VisualBERT: A simple and performant baseline for vision and language. arXiv: 1908.03557.

[9] Radford, A., et al. (2021). Learning transferable visual models from natural language supervision. *ICML 2021*, 8748–8763.

[10] van der Merwe, J. S., et al. (2015). Federated databases and actionable intelligence: using social network analysis to disrupt transnational wildlife trafficking criminal networks. *Security Informatics*, 4, 9. doi: 10.1186/s13388-015-0018-8

[11] OpenAI. (2024). *GPT-4o System Card.* OpenAI Technical Report.

[12] Stringham, O. C., et al. (2021). A guide to using the internet to monitor and quantify the wildlife trade. *Conservation Biology*, 35(4), 1130–1139. doi:10.1111/cobi.13675

[13] Lamba, A., Cassey, P., Segaran, R. R., & Koh, L. P. (2019). Deep learning for environmental conservation. *Current Biology*, 29(19), R977–R982. doi:10.1016/j.cub.2019.08.016

[14] Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box. *Harvard JOLT*, 31(2), 841–887.

[15] Ross, S. R., et al. (2008). Chimpanzees in media: implications for public perception and conservation. *American Journal of Primatology*, 70(8), 797–803. doi:10.1002/ajp.20576

[16] UNODC. (2020). *World Wildlife Crime Report: Trafficking in Protected Species.* United Nations Office on Drugs and Crime. unodc.org

[17] CITES. (2023). *What is CITES?* Convention on International Trade in Endangered Species of Wild Fauna and Flora. cites.org

[18] Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving Decisions About Health, Wealth, and Happiness.* Yale University Press.

[19] TRAFFIC. (2024). *Wildlife Trade Monitoring Network.* traffic.org

[20] Antala, P., Mou, G., & Lee, K. (2026). WildLLM: Uncovering Hidden Wildlife Trafficking on Social Media with Augmented and Fine-Tuned LLMs. *Proceedings of ICWSM 2026.* Worcester Polytechnic Institute. wpi.edu

[21] Siggia, S. (2025). ECO-SOLVE: Using AI to Disrupt Global Wildlife Trafficking. *ACAMS Today.* April 21, 2025. acams.org

[22] Meagher, P., Cincotta, J., Phan, H. T. H., et al. (2026). Detect and Trace: An Australian Field Trial Using Machine-Learning Tools to Combat Illegal Wildlife Trade. *Animals*, 16(5), 731. doi:10.3390/ani16050731

[23] Gray, R. J., & Haysom, S. (2026). *Wildlife Has a Facebook Problem: Evidence of Platform-Scale Facilitation of Criminal Trade in Wildlife.* Global Initiative Against Transnational Organized Crime (GI-TOC). April 2026. globalinitiative.net

[24] ECO-SOLVE / Global Initiative Against Transnational Organized Crime. (2026). *Artificial Intelligence for Environmental Crime Enforcement: Insights from the Bangkok Workshop on AI and Environmental Crime.* GI-TOC. May 2026. globalinitiative.net