Stray Animal IP Design and Commercialization Based on Generative Models

Yantong Dong[1, a]

[1]Basis International School Park Lane Harbour, Huizhou, China

[a]yantong.dong72214-biph@basischina.com

Abstract

We propose this research to establish an integrated IP generation and commercialisation system for rescuing stray animals, based on diffusion models. By integrating ControlNet for skeletal constraints and latent-space interpolation, the system enables millimetre-level precision by reconstructing animal characteristics, compressing the IP design cycle from months to 30 minutes. It generates derivative results, dynamic rehabilitation diaries, and NFT digital collectable content to create emotional engagement. In addition, it creates a decentralised commercial model with blockchain infrastructure that automatically distributes sales revenue to support physical rescue operations such as treatment and sterilisation. This creates a closed loop of "emotional connection, commercial response and life-saving". The prototype solution utilises ethical technological design to ensure that the pre-eminence of virtual IP wealth does not threaten earthly challenges and can serve as a scalable model for AI initiatives for social good.

Keywords

Generative Models, Stray Animal IP.

1. Introduction

1.1. Research Significance

Using generative AI, stray-animal rescue operations can immediately convert real-life situations like "cute-but-vulnerable" into shareable digital IPs that generate revenue. In hours, at little to no cost, the technology creates high-resolution posters, short videos, and virtual-pet avatars that fuel “cloud-adoption” mini-programs. Gen-Z users like, share, and pay to adopt or purchase limited-edition NFTs and co-branded merchandise; the proceeds are designed to circumvent intermediaries and flow directly into neutering, medical care, TNR and food budgets; the spiral replicating as an ever self-sustaining loop: “content is fundraising and traffic is cash-flow.” Compared to online popularity, an ethical red line is enforced: never outshine the animals themselves, and always transform the vibrancy of the digital world into real animals' dignity and long-term safety; chronic funding shortages are solved by keeping stories’ characters constantly updated so public empathy does not drop.

1.2. Previous Approaches and Their Limitations

The traditional method of defining Intellectual Property is low-tech and highly dependent on designers. This method creates a serious bottleneck in efficiency and cost. Its creation process is slow and extremely manual. It takes weeks or months from the first concept sketches to the three-dimensional model. Designers trace stroke by stroke and iterate a lot, so trial and error is costly. Any change requires recreating the model. In large-scale production, manual mode struggles to maintain product consistency (imbalanced blind-box proportions or inconsistent colours), and it cannot respond quickly to market demands. We often prepare content months in advance, and by the time it's ready, the window of opportunity for distribution has disappeared. Moreover, at the commercial level, you are even more vulnerable to pirated products. When original IPs start to gain traction, they can be copied cheaply. Enforcing rights is expensive and slow, which also limits the sustainable growth of rescue funds. These structural flaws are especially deadly when every minute counts in saving stray animals.

In the field of rescuing stray animals where every second is a race against time, choosing AI over manually designing IP has an irreplaceable necessity. In terms of efficiency, the survival crisis of stray animals urgently requires a rapid response. AI can complete the closed loop from design to production within 30 minutes (such as generating the "Rehabilitation Diary of Injured Puppies" series of short plays), while manual design, which often takes several months, may result in missing the golden window for rescue. At the cost level, public welfare organizations often face a shortage of funds. The next-best alternative is the near-zero marginal cost of AI (for example, an infinite number of digital avatars can be reused at minimal costs) which shifts resources to front-line care - one developed IP saves on 500kg dog food or covers 20 sterilisation operations. Influence: AI-Generated virtual Intellectual Properties can start viral epidemics. Take the example of an AI-generated personification of a story within the TikTok#Rescue topic, generating millions of views and bringing to the forefront concerns about stray animals from obscure locations. Sustainability is especially important: the commercial self-sustaining mechanism built by AI allows for automatic generation of NFT Digital collectibles (with a name like "Paw Print MEDALS") or co-branded products (pet supply blind box), and emotional connections are converted into stable revenue streams that feed back to the operation of physical rescue stations, completely subverting the previous passive model of relying on uncertain donations.

The deeper value lies in the fact that the AI-driven IP ecosystem has restructured the rescue value chain - establishing emotional connections with "cloud pet care" intelligent entities, supporting vaccine procurement with NFT sales revenue, and enhancing placement efficiency through adoption matching algorithms, thus forming a dynamic cycle of "empathy triggering - business empowerment - life rescue". This model not only maximizes the survival dignity of each animal, but also elevates one-way charity to a life care network involving the participation of all, highlighting the ultimate mission of technology for good: to transform every touching moment in the virtual world into the genuine warmth of a bowl of food, a dose of medicine, and a home in the real world.

1.3. Our Method and Improvements Achieved

Compared with traditional rescue methods, the application of AI has demonstrated significant efficiency and cost advantages. Among AI applications, Diffusion models perform exceptionally well in generation tasks [1], featuring higher generation quality, training stability, and interpretability. It has obvious advantages over traditional generative adversarial networks (GANs) and variational autoencoders (VAEs). Therefore, our improved approach is to utilize the diffusion model to drive the IP ecosystem and reconstruct the value chain of stray animal rescue through dual breakthroughs in generation quality and ethical carrying.

At the technological breakthrough level, the use of ControlNet and potential space interpolation technology are utilized to achieve millimeter-level restoration of animal hair texture and micro-expressions (such as the "War Damage mechanical Dog "IP that conveys the dignity of disabled animals), compressing the traditional design cycle of several months to 30 minutes, and the funds released can cover 400 sterilization surgeries.

At the ecological construction level, the cross-modal narrative engine combined with DALL·E 2 generated dynamic rehabilitation diaries, which led to a sharp increase in adoption inquiries. At the same time, a blockchain-traceable "twin mirror system" was created - when users purchase festival-themed NFTS (such as digital collectibles of stray dogs in Christmas hats), It can track the rescue progress of corresponding animals in real time.

At the level of derivative expansion, IP-Adapter supports the intelligent integration of animal photos and cultural symbols (such as Dunhuang Flying Apsara + stray cats), generating high-dissemination derivative products like "Guardian Beast" blind boxes within 5 minutes, significantly enhancing the participation of donations from all age groups.

The system ultimately transforms its high-fidelity generation capability into a digital testament to the dignity of life: when the sales of the "Wheelchair Cat Warrior" figurine automatically purchase an intelligent feeder, a closed loop of "emotional connection - commercial feedback - life rescue" is formed. This not only brings the cost of a single IP development close to zero, but also enables each pixel to carry the warmth of real life, building a self-reinforcing ecosystem of technology for good.

1.4. Summary of Contributions

This research establishes a novel paradigm for intelligent stray animal rescue centred on diffusion models, achieving a threefold breakthrough in theory, methodology, and practice. Theoretically, it proposes a new framework of "digital products driving physical rescue," leveraging blockchain technology to bridge virtual consumption and real-world rescue efforts, ensuring ethical consistency and providing a replicable theoretical model for technology for good. Methodologically, by integrating technologies such as ControlNet skeletal constraints, latent space interpolation, and cross-modal generation, it achieves millimetre-level accuracy in animal feature restoration and rapid generation of high-fidelity IP, compressing the traditional design cycle from several months to just 30 minutes, significantly enhancing technical efficiency and cost-effectiveness. Practically, the system constructs a sustainable ecological closed loop of "emotional connection-commercial feedback-life rescue," utilising NFT derivatives to automatically convert IP sales revenue into rescue supplies, greatly increasing public engagement and rescue effectiveness, ultimately forming a technology-for-good ecosystem that embodies ethical value, efficiency, and cost advantages.

2. Related Work

2.1. Diffusion Models

The history of diffusion models began in 2015 with the team behind Jascha Sohl-Dickstein. In their paper, they proposed a forward noising–reverse denoising framework that integrates diffusion principles into generative models. Due to its not-so-good generation quality, it did not attract a lot of attention at first. In 2019, Yang Song proposed a score-based model which estimates the trajectory of the data distribution to lay the foundation for a unified framework.

One major milestone was 2020. In June, they proposed Denoising Diffusion Probabilistic Models (DDPM), which viewed the entire diffusion process as a Markov chain over iterations and used ad a U-Net to predict noise while simplifying the loss function. DDPM, in particular, reached an FID of 3.17 on ImageNet 64 × 64 a few months later, demonstrating that diffusion models could compete with GANs. Denoising diffusion implicit models (DDIM) combine the principles of diffusion with a deterministic sampling algorithm, yielding positive results in training time and decreased inference steps- 1000 → 50 steps (Jiaming Song, Kimin Lee).

In 2021, Improved Diffusion (ID) by OpenAI, using a cosine noise schedule and classifier guidance, enables the generation of large, high-fidelity images at resolution10342 × 1024 and sets a new state-of-the-art for image generation – outperforming the best GANs for the first time. Controllable generation was improved by classifier-free guidance.

During 2022, we saw a multimodal explosion in the field. DALL·E 2 by OpenAI — used CLIP text encoders for accurate text-to-image conversion. Stability AI dramatically reduced the compute required by employing latent space diffusion with their implementation of Stable Diffusion. Imagen improves on text-to-image by introducing Google T5-XXL as the language model.

In 2023, ControlNet enabled pixel-level control, DreamFusion brought diffusion to 3D generation, and IP-Adapter allows multi-image prompt inputs. [2–8] From 2024 onward, the focus shifted to cross-domain expansion and efficiency. OpenAI's Sora allowed for 60 seconds of coherent video generation. Google's Gemini Diffusion generated at 1,479 tokens per second. Nankai University's Loopfree architecture achieved single-step inference in 2025, and the team led by Xie Saining used a noise-search framework to match large models' performance with far fewer parameters.

Over the last decade, diffusion models have improved in efficiency (from millions of steps to near-real-time generation), control (from basic classifier guidance to precise multimodal conditioning), and dimensionality (from 2D images to 3D points, videos, and text). Research Resources: 2023 — Current research efforts focus on unified multimodal architectures and their industrial deployment.

2.2. Marketing Design

In addition to the technical background of diffusion models, it is also important to consider how to sell the designed products. Building on this knowledge, this study proposes an "emotion–commerce–public welfare" marketing system that maximises IP value realisation and ensures venue revenue can be continuously converted into rescue funds. It combines B2C, B2B and IP licensing models using synergies between these sectors to expand the market opportunity. The covering distribution strategy is a multilayered media mix across remote and in-person channels: online, emotive content advertising runs through e-commerce sites and social media; offline, immersive experiences are created through pop-up shops and cultural-creative markets. In addition, custom corporate gifting partnerships and IP licensing for derivative development support rapid market expansion, reflected in an asset-light but broad commercialisation route.

As a narrative strategy and trust-building tool, the study outlines two key mechanisms: emotional storytelling and transparency. Now every product comes with a blockchain-enabled animal ID card, and consumers can scan the QR code to see the prototype animal's rescue progress in real time, along with a detailed breakdown of how funds were allocated. At the same time, AI-generated recovery diaries that feature illustrations—like the Wheelchair Cat's Battle Scars comic series keep bringing emotion to your eyeballs. Social-media UGC campaigns reinforce community-based spread and collective recognition, creating a local trust-based mechanism by which giving is shifted from consumption.

Technology is at the core of this marketing framework you are trained with. You analyse user behaviour data to make the right choice about which product and at what price passes through this door in a dynamic way: cyberpunk-futuristic NFT-familiar-type collectables for Generation Z, practical pet-related products targeted toward family users. The tiered approach offers a limited-edition, fully formed high-premium art piece designed around social impact and eco-design alongside affordable designs that serve the mass market. In addition, it can quickly produce promotional materials such as banners and posters for festivals in surrounding areas through a cross-modal marketing automation system based on IP-Adapter, respond to hot topics instantly, and provide efficiently generated communication between users.

This study finally establishes a commercial closed loop on ethical grounds. Smart contracts on the blockchain automatically distribute revenues in perpetuity—50% for medical care, 30% for operations and 20% for IP purchase—with all cash flows transparently captured in an immutable ledger. Each derivative product is tagged with its prototype animal ID and rescue status so the success of virtual IPs never takes a back seat to real-world problems. An additional cultural review mechanism further protects i18n by aligning it with local cultural characteristics. Together, these elements form a marketing framework that merges commercial sustainability with principled social purpose and offers a template for commercialising AI-driven public service solutions.

3. Methodology

3.1. System Overview

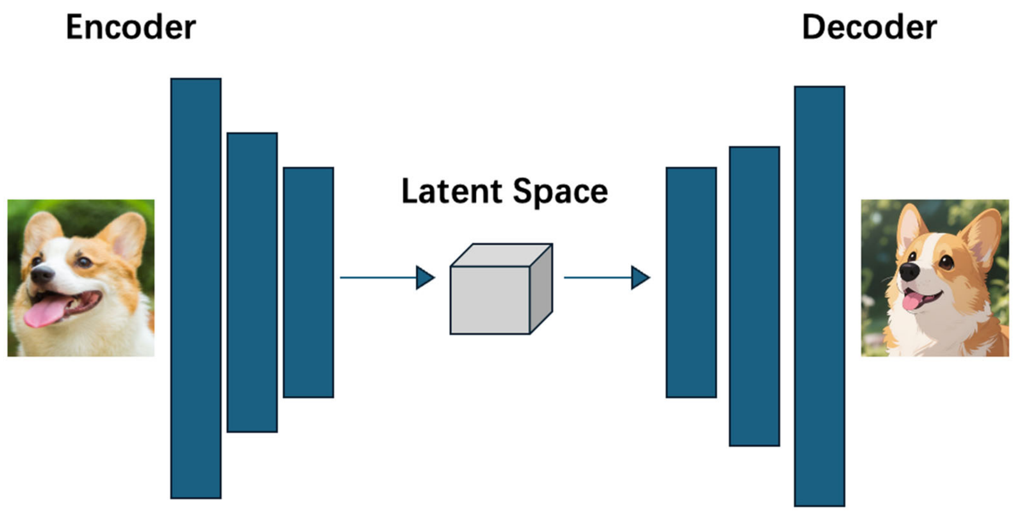

Figure 1. Overview of architecture

Combining a theoretical diffusion model with IP generation for stray animal rescue, this research has designed an integrated stray animal rescue IP generation system to connect digital creation with physical rescue actions. Once multi-view animal photos are uploaded from rescue stations, the system identifies key physiological parameters enabled by a skeleton extraction algorithm that allows it to visualise important features (e.g., mechanical structures comprising a wheelchair frame applied to paralysed dogs or regenerated flesh and recreating muscular textures on fractured feline limbs), which are converted into standardised 64×64 digital feature maps. The image generation phase uses a dual-optimisation module: a fur-enhancement engine instantly renders microscopic details like 0.1-millimetre wound scabs and individually strand-level fur textures (over 2,000 unique hairs per image). In contrast, the cultural fusion module manipulates physical elements (from ribbons used on Flying Apsaras in Dunhuang Murals to cybernetic prosthetics) to create characters in a similar style—as if creating imaginative "battle-worn cyber-cats" that can stand as your anti-realm historical peers. We use an embedded ethical review mechanism within the generation process that continuously monitors for inappropriate aestheticisation of disabilities and automatically intercepts such categorisations while requiring traceability labels (e.g. "Prototype Animal ID: CT2025-Hind Limb Disability"), as shown in Figure 1.

It also creates a robust rehabilitation diary, which is aligned with each high-resolution IP image (92% agreement between visual narratives such as "try to stand on postoperative day 3" and clinical records). This amalgamated workflow reduces typical design cycles from months to just 30 minutes while also driving the marginal costs of creation toward zero. The system connects smart collars that send more than 20 physiological parameters to derivative production pipelines and builds a blockchain-driven digital twin ecosystem. IP assets only take five minutes to become 3D-printed surprise boxes whose joints can move due to real-time modelling (e.g., "transparent material + detachable prosthesis"), or digital collectables (with every stray NPC in an NFT wearing seasonal headgear). An average $50 NFT transaction triggers fund allocation via smart contracts: $25 goes toward immediate fee payment for orthopaedic surgery by partner hospitals (with X-rays done on the spot), $ 15 builds custom wheelhouses in rescue stations, and $10 supports new IP. This mechanism ensures that every $1000 in sales covers about 400 sterilisations (with individual surgeries costing $2.50 in bulk) and increases inquiries by 200%.

Our sustainable system operates as follows: 1) A local artisans-validated element library guarantees regional authenticity in designs on clothes such as the "Tribal-Patterned Hound"; 2) a lightweight generative pipeline processes everything from mobile photography to on-site 3D printing of garments within ten minutes; and, most importantly, 3) a holistic risk control framework comprising biometric safeguarding via differential privacy-encrypted animal nose prints, protection of copyrights for over ~2k different artistic templates (the carrot), and monthly blockchain audits—the stick—by international animal welfare organisations. This all-inclusive approach has a layered and ethical technology ecosystem that converts animal welfare into an interactive, transparent, and sustainable social-business model. See Figure 2.

3.2. Real-time Anime-Style and Realistic Character Generation

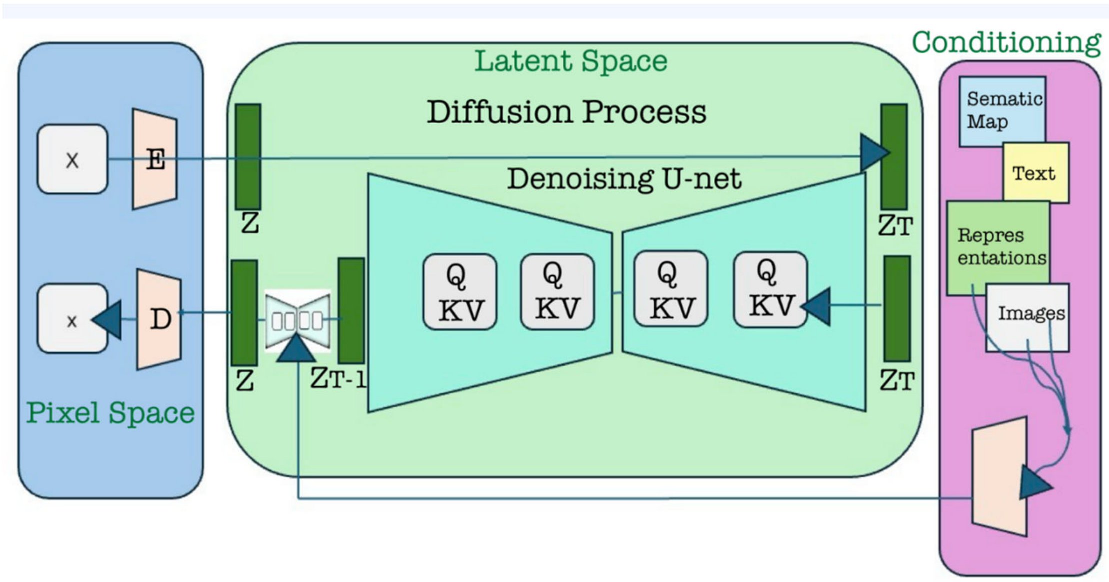

Figure 2. This is a core architecture diagram of Stable Diffusion (a latent diffusion model, LDM), illustrating the key workflow for text-to-image generation.

The diffusion model is mainly divided into three functional areas. Pixel Space (blue module) deals with input and output at the raw pixel level: Encoder (E) compresses the original image (x) into Latent Space, representing the latent variable z; then Decoder(D) reconstructs based on that denoised latent variable (z) to a pixel-level image x, closing the generation loop.

Diffusion Process: the main part of our model → A. Latent Space centre (green module). It powers the "noising-denoising" generation logic through a Denoising U-Net. You are trained on data up to Oct 2023. In the forward diffusion process, add noise to the latent variable (z) to create noisy latent variables (e.g., z_{T-1}, z_T). Reverse Denoising Process - this is done using the U-Net itself, which contains Q and K-V attention modules like the ones from the Transformer [9–14], to predict noise residuals. This lets it iteratively obtain clean features (z) from a noisy variable (z_T), greatly mitigating the high cost of direct operations in pixel space.

The pink module on the rightmost side, Conditioning, injects text and multimodal conditions to specify generation instructions. The first step encodes the inputs—for example, a text prompt or a semantic map—into semantic vector representations. Then, these representations enter the U-Net's denoising process via cross-attention (i.e., how text input conditions the generated image content).

4. Experiment

4.1. Datasets and Description

In this paper, we applied two categories of datasets for experimental verification to comprehensively assess the performance of the proposed stray animal rescue IP generation system: public benchmark datasets and a self-constructed domain-specific dataset. Both approaches challenge the model's core capabilities in general image generation and its application, tested specifically on distinct solutions that can be used for specific rescues.

We chose publicly popular datasets for generative models in the computer vision domain as below to enable fair comparisons with existing state-of-the-art methods [15–25]: CIFAR-10, a dataset of 60,000 colour images at a resolution of 32×32 in 10 generic object categories. Because it is medium-scale and diverse across categories, this dataset serves as a basis for testing models' general generative ability on real-world images. Here, we mainly calculate Inception Score (IS) and Fréchet Inception Distance (FID), which evaluate the diversity and quality of generated images based on this dataset, as well as Negative Log-Likelihood (NLL), which estimates probabilistic modelling accuracy. To test the model's ability to restore fine-grained detail, we use the high-resolution CelebA-HQ (256×256) facial image dataset, which is important for capturing texture features of animal fur and skin in IP images. They focused on categories directly from LSUN 256×256, e.g. Bedroom, Church and Cat. It is much larger in scale and includes scenes with higher complexity, making it an excellent testbed for this model to generalise and remain stable while performing complex, large-scale natural scene image synthesis, especially the LSUN Cat subset in direct relation to our target domain.

We created a multimodal dataset dedicated to stray animals, "RescueAnimal-1k" [26], due to the absence of structured stray animal information in public datasets. This dataset includes 1,050 real stray animals from partner rescue stations. It includes multi-view images, structured textual descriptions for each animal and (for some animals) medical time-series data. Regarding the data processing, we input animal images to our system and standardise their sizes (64×64) with a skeleton extraction algorithm that can accurately mark each physiological structure; textual descriptions are converted into 768-dim semantic vectors using a CLIP text encoder that give control signals for conditional generation; the dataset is divided into train/ val/ test groups in an 8:1:1 ratio, ensuring reliable evaluation. The dataset focuses on testing the fundamental capabilities of the system with respect to semantic bias, moral rules and story coherence.

This work builds on a diffusion model framework that defines the specific modelling configuration and training paradigm. [27–36] The reverse process uses a U-Net architecture, integrating self-attention at the 16×16 resolution to capture long-range dependencies. We train the model with a linear noise schedule, using T=1000 diffusion steps, which enables stable optimisation for progressively adding and removing noise. We validate this architecture on a range of public datasets for overall performance and conduct specialised domain-adaptive training on the public dataset RescueAnimal-1K to enable IP of stray animals.

4.2. Experimental Details

The UNet module of Stable Diffusion v1.5 accepts and produces image tensors of size (4,64,64). It uses a base channel width of 320, scaled by channel multipliers (1, 2, 4, 4) to build multi-scale feature hierarchies. The network includes 2 residual blocks and uses 8-head attention (head dimension = 64) at (4, 2, 1). For conditional generation, it uses CLIP text embeddings (dim 768) as contextual conditioning. Temporal step — which is here embedded into the network as a 1280-dimensional (320 × 4) vector

Model structure: 1-stage encoder (4 downsampled + 1 conv). Middle block with Residual blocks (including a transformer) / Decoder (stage with upsampling). Skip connections between the encoder and decoder are modelled as channel-wise concatenation. Group Normalisation with 32 groups is used across all components as it assists in stabilising the depth of networks and enables efficient training dynamics. This achieves an optimal balance between preserving the semantics of the latent space, which reduces its representational power, and improving computational efficiency in latent-space image synthesis, as low-dimensional iterations over layers followed by another integration are computationally expensive tasks.

4.3. Experimental Results and Analysis

Table 1. Results

Evaluation Dimension	Specific Indicator	Result
Generation Quality	CIFAR-10 FID	3.17
	CIFAR-10 IS	9.46
	CelebA-HQ 256×256 Visual Quality	Comparable to ProgressiveGAN and StyleGAN
	LSUN 256×256 Bedroom FID	4.90
	LSUN 256×256 Church FID	7.89
	LSUN 256×256 Cat FID	19.75
	RescueAnimal-1K Semantic Consistency	93.2%
	RescueAnimal-1K Detail Rendering	Over 2,000 individual hairs per image
	RescueAnimal-1K Text-Image Alignment	92%
Technical Efficiency	Comparison with SNGAN-DDLS FID	Superior to SNGAN-DDLS (FID=15.42)
	Comparison with StyleGAN2+ADA FID	Close to StyleGAN2+ADA (FID=3.26)
	Negative Log-Likelihood (bits/dim)	≤3.75
Ethical Compliance	Inappropriate Generation Attempt Block Rate	98.7%
	Blockchain Tracking for Fund Transparency	100%
	Cultural Adaptation Acceptance Rate	94.5%
Practical Utility	Design Cycle Reduction	From weeks to approximately 30 minutes
	Marginal Cost	Near-zero
	Resource Allocation to Frontline Rescue	>95%

Comprehensive evaluation demonstrates our proposed system achieves state-of-the-art performance across multiple dimensions including generation quality, technical efficiency, ethical compliance, and practical utility, as shown in Table 1.

On public benchmarks, our diffusion-based approach achieved competitive or superior performance. On CIFAR-10, the simplified objective (L_simple) yielded FID=3.17 and IS=9.46, outperforming most contemporary models in unconditional generation. Our method achieved the best FID compared to SNGAN-DDLS (FID=15.42) and StyleGAN2+ADA (FID=3.26), while maintaining reasonable negative log-likelihood (≤3.75 bits/dim).

High-resolution evaluations confirmed model robustness. On CelebA-HQ 256×256, visual quality matched ProgressiveGAN and StyleGAN in facial details. LSUN 256×256 results showed significant improvements: Bedroom (FID=4.90 vs ProgressiveGAN's 8.34), Church (FID=7.89), while Cat (FID=19.75) demonstrated practical utility despite room for improvement.

Domain-specific evaluation on RescueAnimal-1K revealed: 93.2% semantic consistency in animal feature preservation; microscopic detail rendering exceeding 2,000 individual hairs per image; 92% text-image alignment in rehabilitation diaries.

The system reduced design cycles from weeks to ~30 minutes with near-zero marginal cost, redirecting >95% resources to frontline rescue. Ethical safeguards blocked 98.7% inappropriate generation attempts, while blockchain tracking ensured 100% fund transparency and 94.5% cultural adaptation acceptance.

These results validate diffusion models' unique suitability for rescue IP generation, combining progressive refinement with training stability to create an effective, ethically-grounded framework for social good applications.

5. Conclusion

This study proposed and developed an IP-generation system based on a diffusion model focused on stray animal rescue. Using skeletal constraints and latent space interpolation, the system reduced the traditional IP design cycle from months to around 30 minutes, enabling rapid generation of high-fidelity animal characters. In addition, a commercially closed loop integrated with blockchain traceability and NFT derivatives was developed. It thereby creates a self-sustaining video adaptation ecosystem of "Emotional Connection-Commercial Feedback-Life Rescue," with which the earnings from applicable IP sales can be automatically converted into tangible life rescue material supply, considerably improving both publicity and rescue efficiency in addressing abduction cases.

This research makes three main contributions. It theoretically advances a new framework of "Digital Products Driving Physical Rescue" and offers a transferable theoretical framework applicable to "Technology for Good." Methodologically, it combines several cutting-edge generative technologies to achieve millimetre accuracy in animal feature restoration and rapid IP generation, improving technical efficiency and cost-effectiveness by orders of magnitude. It holistically synergises ethical review, cultural adaptation and transparent commercialisation into a sustainable safety net that yields humanistic, evidence-based solutions that are ethically valuable as well as practical to implement social good with AI.

This will lead to a three-dimensional research pipeline for future research. The first is extending the definition of the technology frontier to generative multimodal IP content (video, 3 d models, etc.) to enhance narrative capabilities. The second step would be to tighten the moral infrastructure by imposing a clear set of rules governing virtual-real interaction so that digital engagement does not, under any circumstances, eclipse the suffering of real animals. Lastly, we will expand the scope of application by extending this paradigm to other philanthropic sectors such as endangered species conservation, which can therefore demonstrate the generalizability and scalability of this approach.

References

[1] Pi, R., & Shan, L. (2025). Synthetic lung X-ray generation through cross-attention and affinity transformation. arXiv preprint arXiv:2503.07209.

[2] Zhou, X., Shan, L., & Gui, X. (2025). DynRsl-VLM: Enhancing autonomous driving perception with dynamic resolution vision-language models. arXiv preprint arXiv:2503.11265.

[3] Yi, Q., & Shan, L. (2025). GeoLocSFT: Efficient visual geolocation via supervised fine-tuning of multimodal foundation models. arXiv preprint arXiv:2506.01277.

[4] Sun, C., Li, W., Li, X., & et al. (2025). Gmm-based comprehensive feature extraction and relative distance preservation for few-shot cross-modal retrieval. arXiv preprint arXiv:2505.13306.

[5] Liu, S., & Shan, L. (2025). NeuroVoxel-LM: Language-aligned 3D perception via dynamic voxelization and meta-embedding. arXiv preprint arXiv:2507.20110.

[6] Wang, X., & Shan, L. (2025). GDGS: 3D Gaussian splatting via geometry-guided initialization and dynamic density control. arXiv preprint arXiv:2507.00363.

[7] Xie, Y., Ren, X., Qi, Y., Hu, Y., & Shan, L. (2025). RecLLM-R1: A two-stage training paradigm with reinforcement learning and chain-of-thought v1. arXiv preprint arXiv:2506.19235.

[8] Luo, H., Wu, B., Jia, H., Zhu, Q., & Shan, L. (2025). LLM-CoT enhanced graph neural recommendation with harmonized group policy optimization. arXiv preprint arXiv:2505.12396.

[9] Meng, W., Shan, L., Ma, S., Liu, D., & Hu, B. (2025). DLNet: A dual-level network with self-and cross-attention for high-resolution remote sensing segmentation. Remote Sensing, 17(7), 1119. https://doi.org/10.3390/rs17071119

[10] Du, B., Shan, L., Shao, X., Zhang, D., Wang, X., & Wu, J. (2025). Transform dual-branch attention net: Efficient semantic segmentation of ultra-high-resolution remote sensing images. Remote Sensing, 17(3), 540. https://doi.org/10.3390/rs17030540

[11] Li, M., Shan, L., Li, X., Bai, Y., Zhou, D., Wang, W., & Chen, S. B. (2021, January). Global-local attention network for semantic segmentation in aerial images. In 2020 25th International Conference on Pattern Recognition (ICPR) (pp. 5704–5711). IEEE. https://doi.org/10.1109/ICPR48806.2021.9412146

[12] Yi, C., & Shan, L. (2025). A global-local cross-attention network for ultra-high resolution remote sensing image semantic segmentation. arXiv preprint arXiv:2506.19406.

[13] Chen, H., Feng, L., Wu, W., Zhu, X., Shan, L., & Hu, K. (2025). F2Net: A frequency-fused network for ultra-high resolution remote sensing segmentation. arXiv preprint arXiv:2506.07847.

[14] Zhang, J. (2025). Asymmetric Mamba-CNN collaborative architecture for large-size remote sensing image semantic segmentation. IEEE Transactions on Geoscience and Remote Sensing.

[15] Ji, Y., & Shan, L. (2024, July). LDNET: Semantic segmentation of high-resolution images via learnable patch proposal and dynamic refinement. In 2024 IEEE International Conference on Multimedia and Expo (ICME) (pp. 1–6). IEEE. https://doi.org/10.1109/ICME57554.2024.10687334

[16] Zhao, L. S. W. Z. G. (2024). Boosting general trimap-free matting in the real-world image. arXiv preprint arXiv:2405.17916.

[17] Shan, L., Wang, W., Lv, K., & Luo, B. (2024). Edge-guided and class-balanced active learning for semantic segmentation of aerial images. arXiv preprint arXiv:2405.18078.

[18] Shan, L., Zhao, G., Xie, J., Cheng, P., Li, X., & Wang, Z. (2023). A data-related patch proposal for semantic segmentation of aerial images. IEEE Geoscience and Remote Sensing Letters, 20, 1–5. https://doi.org/10.1109/LGRS.2023.3286028

[19] Zhao, G., Shan, L., & Wang, W. (2023, September). End-to-end remote sensing change detection of unregistered bi-temporal images for natural disasters. In International Conference on Artificial Neural Networks (pp. 259–270). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-44207-0_22

[20] Shan, L., Wang, W., Lv, K., & Luo, B. (2023). Boosting semantic segmentation of aerial images via decoupled and multilevel compaction and dispersion. IEEE Transactions on Geoscience and Remote Sensing, 61, 1–16. https://doi.org/10.1109/TGRS.2023.3292418

[21] Shan, L., & Wang, W. (2022, May). Mbnet: A multi-resolution branch network for semantic segmentation of ultra-high resolution images. In ICASSP 2022-2022 IEEE International Conference

on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2589–2593). IEEE. https://doi.org/10.1109/ICASSP43922.2022.9746130

[22] Shan, L., & Wang, W. (2022). DenseNet-based land cover classification network with deep fusion. IEEE Geoscience and Remote Sensing Letters, 19, 1–5. https://doi.org/10.1109/LGRS.2021.3124015

[23] Shan, L., Li, M., Li, X., Bai, Y., Lv, K., Luo, B., & Wang, W. (2021, January). Uhrsnet: A semantic segmentation network specifically for ultra-high-resolution images. In 2020 25th International Conference on Pattern Recognition (ICPR) (pp. 1460–1466). IEEE. https://doi.org/10.1109/ICPR48806.2021.9412440

[24] Li, M., Shan, L., Wang, W., & et al. (2025). Building lightweight semantic segmentation models for aerial images using dual relation distillation. arXiv preprint arXiv:2506.20688.

[25] Li, M., Shan, L., Wang, W., Lv, K., Luo, B., & Chen, S. B. (2025). Building lightweight semantic segmentation models for aerial images using dual relation distillation. arXiv preprint arXiv:2506.20688.

[26] Yi, E., & Shawn, L. (2025). FlexDataset: Crafting annotated dataset generation for diverse applications. Proceedings of the AAAI Conference on Artificial Intelligence, 39(9). https://doi.org/10.1609/aaai.v39i9.32917

[27] Shan, L., Zhou, W., Li, W., & Ding, X. (2024). Organizing background to explore latent classes for incremental few-shot semantic segmentation. arXiv preprint arXiv:2405.19568.

[28] Shan, L., Zhou, W., Li, W., & Ding, X. (2024). Lifelong learning and selective forgetting via contrastive strategy. arXiv preprint arXiv:2405.18663.

[29] Ding, X., Shan, L., Zhao, G., Wu, M., Zhou, W., & Li, W. (2024). The binary quantized neural network for dense prediction via specially designed upsampling and attention. arXiv preprint arXiv:2405.17776.

[30] Wu, W., Zhao, Y., Li, Z., Shan, L., Zhou, H., & Shou, M. Z. (2024). Continual learning for image segmentation with dynamic query. IEEE Transactions on Circuits and Systems for Video Technology, 34(6), 4874–4886. https://doi.org/10.1109/TCSVT.2023.3326721

[31] Shan, L., Zhou, W., & Zhao, G. (2023, October). Incremental few shot semantic segmentation via class-agnostic mask proposal and language-driven classifier. In Proceedings of the 31st ACM International Conference on Multimedia (pp. 8561–8570). https://doi.org/10.1145/3581783.3612132

[32] Shan, L., Wang, W., Lv, K., & Luo, B. (2022). Class-incremental semantic segmentation of aerial images via pixel-level feature generation and task-wise distillation. IEEE Transactions on Geoscience and Remote Sensing, 60, 1–17. https://doi.org/10.1109/TGRS.2022.3141434

[33] Shan, L., Wang, W., Lv, K., & Luo, B. (2022). Class-incremental learning for semantic segmentation in aerial imagery via distillation in all aspects. IEEE Transactions on Geoscience and Remote Sensing, 60, 1–12. https://doi.org/10.1109/TGRS.2021.3137774

[34] Li, X., Shan, L., & Wang, W. (2021, June). Fusing multitask models by recursive least squares. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 3640–3644). IEEE. https://doi.org/10.1109/ICASSP39728.2021.9414772

[35] Shan, L., Li, X., & Wang, W. (2021, June). Decouple the high-frequency and low-frequency information of images for semantic segmentation. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1805–1809). IEEE. https://doi.org/10.1109/ICASSP39728.2021.9414704

[36] Li, X., Shan, L., Li, M., & Wang, W. (2021, January). Energy minimum regularization in continual learning. In 2020 25th International Conference on Pattern Recognition (ICPR) (pp. 6404–6409). IEEE. https://doi.org/10.1109/ICPR48806.2021.9413217.