GSC Advanced Research and Reviews

eISSN: 2582-4597 CODEN (USA): GARRC2

Cross Ref DOI: 10.30574/gscarr

Journal homepage: https://gsconlinepress.com/journals/gscarr/

(RESEARCH ARTICLE)

Advancements in diffusion models for high-resolution image and short form video generation

Ayodele R. Akinyele [1], Frederick Ogunseye [2], Adewale Asimolowo [3], Geoffrey Munyaneza [1], Oluwatosin Mudele [4] and Oluwole Olakunle Ajayi [5,*]

[1] *Kenan-Flagler Business School, University of North Carolina at Chapel Hill, North Carolina, USA.*
[2] *Fuqua School of Business, Duke University, Durham, North Carolina, USA.*
[3] *Ross School of Business, University of Michigan, USA.*
[4] *School of Computing, Engineering and Built Environment, Glasgow Caledonian University, Glasgow, United Kingdom.*
[5] *Community and Program Specialist, UHAI For Health Inc, Worcester, Massachusetts, USA.*

GSC Advanced Research and Reviews, 2024, 21(02), 508–520

Publication history: Received on 07 October 2024; revised on 17 November 2024; accepted on 19 November 2024

Article DOI: https://doi.org/10.30574/gscarr.2024.21.2.0441

Abstract

This paper offers an in-depth review of the most recent advancements in diffusion models, particularly highlighting their transformative role in high-resolution image generation and their emerging applications in short-form video generation. Diffusion models, a class of generative models rooted in probabilistic frameworks, have rapidly gained prominence due to their ability to produce photorealistic and detailed outputs by reversing a noise-infusion process. Their strength lies in their capacity to generate high-quality media that exceeds previous limitations of generative models like GANs, especially in terms of diversity and training stability. The study utilized five scientific databases and a systematic search strategy to identify research articles on PubMed, Google Scholar, Scopus, IEEE, and Science Direct relating to the topic. Furthermore, books, dissertations, master's theses, and conference proceedings were utilized in this study. This study encompassed all publications published until 2024. The review begins by delving into the fundamental principles underlying diffusion models, which revolve around the process of gradually adding and removing noise from an image or video over a series of time steps. This section emphasizes the mathematical foundation of diffusion processes, particularly the forward process of noise addition and the reverse process of denoising, which enables these models to generate media with fine detail. A significant portion of this review is dedicated to the impact of diffusion models on high-resolution image and short-form video generation as well as success metrics for evaluating short-form video generation, curation, and summarization, areas where they have been especially transformative. Conclusively, this paper provides a comprehensive exploration of how diffusion models have reshaped the landscape of media generation. From their foundational principles and technical evolution to their applications in high-resolution media and short-form video, the paper highlights both the profound potential of these models and the ongoing challenges that must be addressed for their responsible and scalable use.

Keywords: Diffusion Models; Video Creation; High Resolution Image; Short-form Video

1. Introduction

Digital marketing, social media, and entertainment require HD photos and short films. Digital media development has changed thanks to AI, ML, and deep learning (Hasan et al., 2024). Artificial intelligence has gained adoption for multiple applications in recent years (Akinyele et al., 2024, Mudele et al., 2019, Mudele et al., 2021a, Mudele et al., 2021b). During the previous decade, GANs and Variational Autoencoders pushed this change. Diffusion models can create high-fidelity material with appealing visuals and reliable timing. Since diffusion models from statistical mechanics can characterize

* Corresponding author: Oluwole Olakunle Ajayi

complex data distributions, they have returned to machine learning. These generative models reproduce high-resolution images and films from random noise via progressive noise addition and removal, unlike earlier methods (Chauhan et al., 2024). For detailed and coherent data that matches real-world photos and videos, models may iteratively reverse a diffusion process.

Social media, advertising, and VR require high-resolution images and videos. Premium content boosts TikTok, Instagram, and YouTube engagement. High-resolution photographs and videos attract people to these sites. Developers and content providers must properly create and curate it (Bushey, 2015). Though commonly utilized, GANs are unstable during training and struggle to produce fine features in high-resolution outputs. Due to probabilistic approximations, VAEs are stable but create lower-quality images (Sajjadi, 2024). Diffusion models provide more stable, accurate, and realistic media with better texture, lighting, and realism (Moser et al., 2024).

Instagram Reels, TikTok, and YouTube Shorts have made short videos popular. These services encourage brevity and distinctiveness. Interestingly, short videos are replacing long ones. Short form videomaking is tricky. Unlike photos, video temporal coherence requires error-free frame transitions (Engström et al., 2010). With so much video material published daily, platforms require algorithms to curate and summarize it to engage customers, AI, especially diffusion models, is important (Po et al., 2024). Diffusion models are increasingly adapting to video from photos. Recent video diffusion models create high-quality, temporal-aligned short-form videos. Real-time video creation, compilation, and summary also require this innovation (Chen et al., 2024). Watch duration, engagement rates, and sentiment analysis on social media platforms boost user engagement and retention, making them important short-form video curation, summary, and compilation success measures (Nguyen and Veer, 2024). The current review examines how diffusion model advances affect high-resolution image and short-form video generation, thereby connecting sentiment and engagement on social with high-resolution and short video generation.

2. Methods

The research employed five scientific databases and a systematic search methodology to locate publications on improvements in Diffusion Models for High-Resolution Image and short-form Video Generation (PubMed, Google Scholar, Scopus, IEEE, and Science Direct) (Zhao et al., 2020). Additionally, there existed books, dissertations, master's theses, and conference proceedings. The search terms "Short form Video" and the keywords "Video Creation," "High Resolution Image," and "Diffusion model" were entered into the search engine. A comprehensive list of abstracts was obtained and analyzed for the current investigation; any articles that satisfied the inclusion criteria were meticulously examined. The review included all papers released until 2024.

3. Results

3.1. Overview of Diffusion Models

Diffusion models, a deep generative modelling paradigm change, capture many with their potential. These models create incredibly realistic graphics, expanding digital content and innovation (Wang et al., 2024). Controllable and precise diffusion models meet user-defined generation needs better than Generative Adversarial Networks (GANs) (Cao et al., 2024). Stable diffusion model outputs demonstrate how diffusion models may govern image synthesis using text inputs to produce high-fidelity visuals that match textual descriptions (Jadhav et al., 2024). Recent years have seen many groundbreaking diffusion modelling theories and studies. Due to this research rush, newcomers struggle to establish themselves in this huge and continuously changing market (Markides, 2013).

Diffusion models' generative power makes them essential in vision-centric applications such image editing, inpainting, semantic segmentation, and anomaly detection (Wang et al., 2024). Diffusion probabilistic models, based on diffusion modelling, have acquired popularity. Research enthusiasm generates new models and scholarship constantly. Text-to-image generators like DALL-E, Imagen, and Stable Diffusion have raised the bar for image generation from text, sparking popular and academic interest in diffusion models. Text-to-video creation has expanded, showcasing advanced movies and boosting diffusion model interest (Bengesi et al., 2024). Statistic and temporal analysis indicate diffusion models' growing popularity in vision. This image highlights their growing prominence in generative modelling, indicating a mentality shift (Xing et al., 2023).

3.2. Working Principle and Framework of Diffusion models

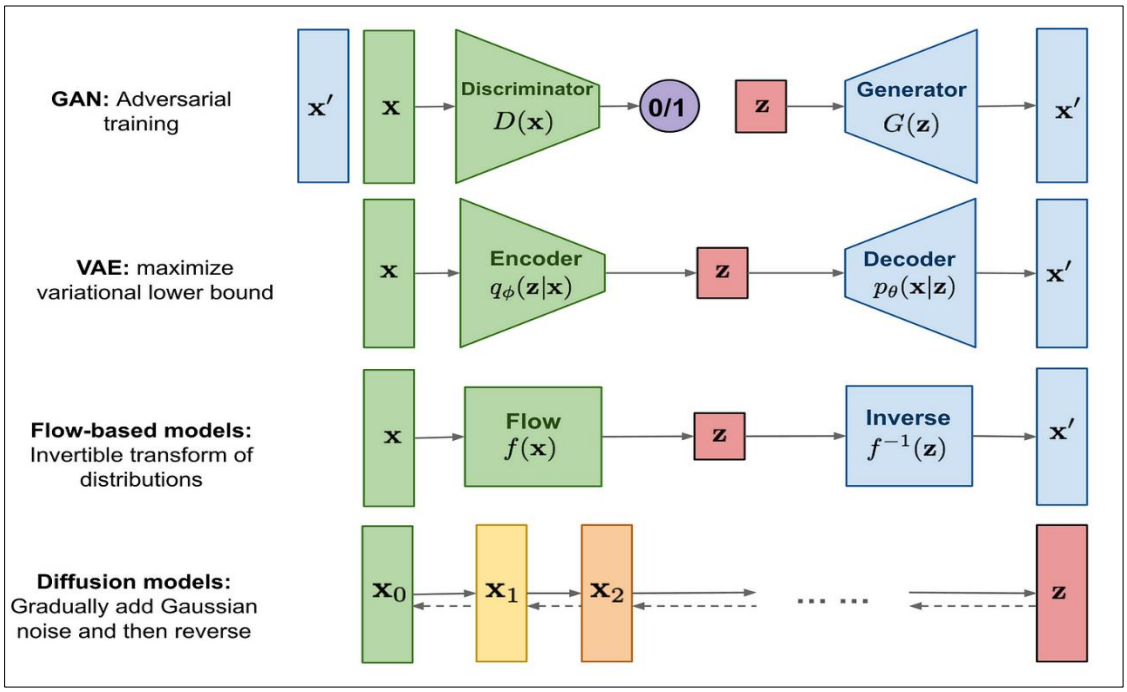

Figure 1 Diagrammatic Representation of Diffusion models framework (Singh, 2023)

3.2.1. Denoising Diffusion Probabilistic Models (DDPMs)

Forward Process

Reflecting numerous facets of the video curating and summarizing difficulty, the forward process in DDPMs where a clear image is progressively masked by noise mirrors several elements of This process of information deterioration provides insightful analysis of how we evaluate success in the production of short-form videos. The original image loses clarity when noise is introduced in the forward process, much as longer movies feature instances of varied importance (Chen et al., 2020). The difficulty for curators is determining which events include the most important material. Like maintaining important elements of an image amid rising noise, the success of a well-chosen video often depends on its capacity to sustain high information density packing maximum relevance into least time.

It gets harder to find the signal, which is the original picture, as noise increases in the forward process (Schofield et al., 2020). In the same way, less important material (noise) in long movies can hide important messages or events (the signal). What makes curated videos unique is how well they improve this ratio, or how well they pull out the most obvious "signal" from the original material (Khurana and Deshpande, 2021). The forward process is slow, which is related to the video summarizing metric of story coherence. The forward process of the DDPM keeps a smooth change from picture to noise (Wang et al., 2024) so a well-summarized video should keep the story's main arc even as it cuts down on the amount of information it contains. One way to tell if a recap is good is by how it moves from one important idea to the next without losing sight of the main point of the original story.

Reverse Process

In reverse process in DDPMs a structured picture forms from noise and have even stronger connections to video curation and compilation measures. Engagement rate, an important indicator of short-form videos, shows how well the DDPM can separate useful content from irrelevant noise (Huang et al., 2023). A well-made video should keep people interested by quickly creating clear themes or storylines, just like people become more interested in the output of the DDPM when they start to see shapes and patterns in it.

The sequential increase of the reverse process matches the measure of the retention rate. At each stage of the DDPM, the picture quality gets better, which keeps people's attention (Wang et al., 2022). In the same way, a good short-form video keeps people watching by always showing something interesting or useful; each part builds on the last (Rugrien and Funk, 2022). Relevance of the material is another important factor, which can be seen in how the DDPM reconstructs

different types of images based on how it was trained. Khan et al. (2024) say that a well-curated video should exactly match the interests of its viewers, just like a DDPM creates images that match how it distributes learned data. The different outputs that can be made from different noise sources in DDPMs is also linked to the content diversity metric in compilations. Like DDPMs can produce, successful compilations provide a spectrum of views or content kinds while keeping thematic consistency (Yu et al., 2024).

Shareability, which is a big part of short films, is like the "wow factor" of a fully realized DDPM output. Like how people are driven to share great photos made by AI (Guo et al., 202), a well-curated movie should be enough to get people to share it. In other words, the summarizing accuracy measure is based on how well the final DDPM output matches the original images. Like a DDPM tries to produce images indistinguishable from actual ones, a good summary or compilation should properly reflect the core of the original material (Wang et al., 2024).

3.2.2. score-Based Generative Models (SGMs)

SGMs are a lot like the process of selecting videos for viewing because they stress understanding the variation of the data distribution. As Bandi et al. (2023) say, the "score" in SGMs tells you how to change a sample so that it more closely matches the real data distribution. This idea works for figuring out how relevant material is in video curation. Niu et al. (2024) say that curators must narrow down huge amounts of information to the most important and relevant parts. This is like how SGMs constantly turn noise into useful data. The repeating nature of SGMs is also like the retention rate measure seen in short-form movies. Like every section of a well-curated video should preserve and increase viewer interest, every stage in an SGM gets the created content closer to the goal distribution (Olán, 2023). Just like SGMs get better at what they do over time, a video's success may be judged by how well it keeps people interested in the whole thing.

3.2.3. Stochastic Differential Equations

SDEs (Stochastic Differential Equations) helps to understand how video collections flow and make sense. Wang et al. (2024) say that the story coherence score in summarizing videos is like the smooth paths that SDEs show when turning noise into data. Even if the plot is shortened, a well-summarized movie should still have a clear progression from one important event to the next (He et al., 2024). Updating short videos with information relies on how well the drift and diffusion factors in SDEs work together (Bounoua et al., 2024). The drift term focuses the process on more likely data points, like how the curator's job is to draw attention to important information. To keep people interested, material needs to be varied, which is what the diffusion phrase means. Curation that works well finds a balance between fixed and stochastic factors, just like SDEs do (Cooper, 2024).

3.2.4. Relation to Other Generative Models

Variational Autoencoders (VAEs)

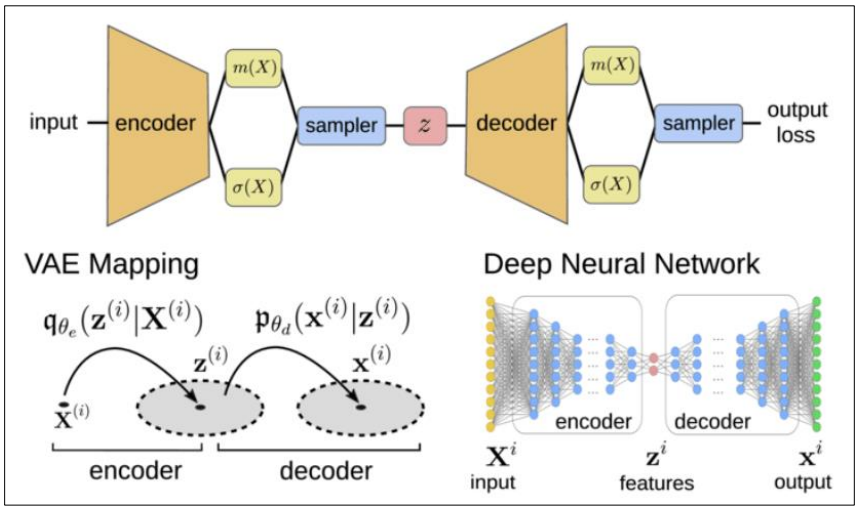

Figure 2 Variational Autoencoders mode of action (Asperti et al., 2021)

VAEs are a great way to think about video summary because they store data in a latent space and then decode it. VAEs' encoding process, which shrinks input data into a small latent representation, is like the goal of summarization, which

is to get to the heart of longer material. One of the major ways to measure success of summarization accuracy is to know how well the latent space of the summary capture the most important parts of the original content, this is like the content diversity measure in video compilations (Alhabeeb and Al-Shargabi, 2024). New samples are made by taking samples from the latent space and decoding them. Successful compilations offer a variety of content while staying true to the main theme, just like VAEs can make different outputs by picking places in the latent space that are different from each other. Being able to easily move through this empty space is also related to the engagement rate measure because it helps to make different but related content that keeps people interested (Jansen et al., 2023).

Generative Adversarial Networks

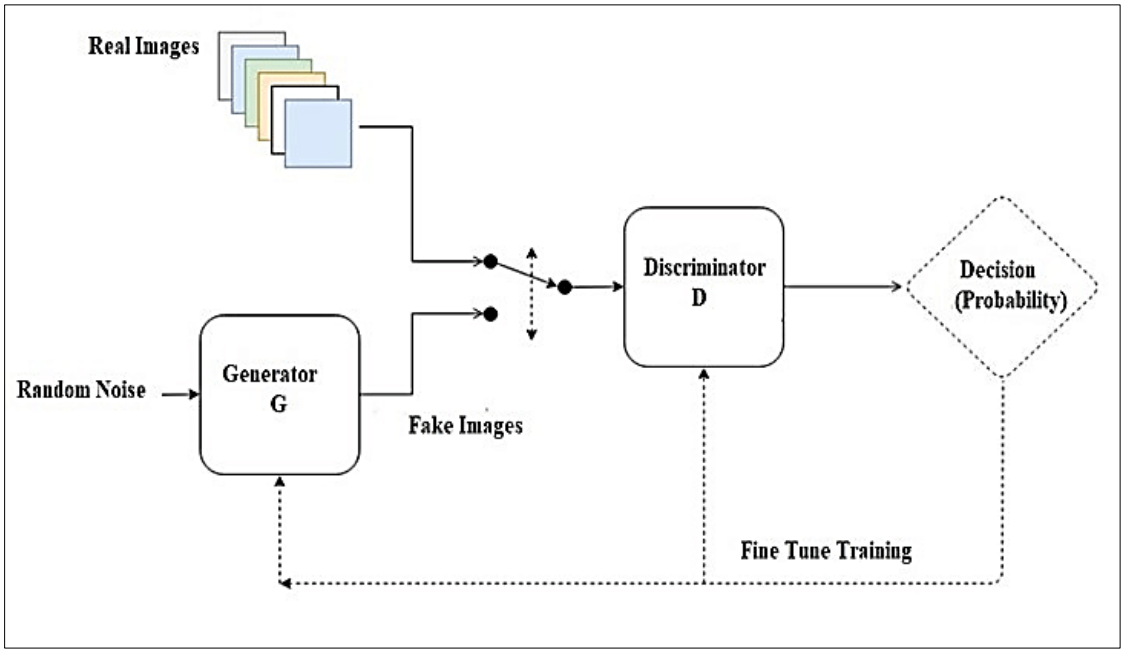

Figure 3 General Block Diagram of Generative Adversarial Networks (Remya Revi et al., 2021)

With their adversarial training system between a generator and a discriminator, GANs provide possibly the most direct analogy to the difficulties of short-form video production (Aslanidou-AEM, 2021). The generator aims to produce short-form films with content that, given the shareability measure, can mislead the discriminator alignment. Video curators want to build collections so appealing that viewers feel obliged to share them, much as a GAN's generator seeks to generate content indistinguishable from real data (Garon, 2023.). Tasked with separating generated from real data, the discriminator in GANs reflects the critical eye of the viewer in evaluating video quality. This has bearing on the engagement rate statistic for a successful short-form video (Kalateh et al., 2024). An effectively produced GAN should grab and hold viewer attention right away. Furthermore, offering insight into the balance between novelty and familiarity in content curation is GAN's adversarial character (Lan et al., 2020). Like curators who must discover fresh approaches on known subjects to keep audiences interested, the generator continuously develops to create original content that can evade the discriminator. These dynamic challenges curators to remain current with audience tastes and trending issues, therefore speaking to the content relevancy metric.

3.3. Applications of Diffusion Models

Diffusion models excel in art, design, and medical imaging. Their capacity to generate high-resolution photographs and films and accept human inputs makes them useful for artistic and technological concerns. Below are some key areas where diffusion models are making a significant impact:

3.3.1. Video Generation and Editing

Diffusion models produce realistic, high-resolution videos. Diffusion models automate and improve TikTok, Instagram Reels, and YouTube Shorts content. Video creators can use diffusion-based video generation models to create entire clips from text descriptions or low-resolution video drafts (Chen et al., 2023).

3.3.2. Text-to-Image Generation

Descriptions become images with text-to-image. Diffusion-based text-to-image generation is popular. The diffusion model can create high-quality samples from objects, textures, and shapes. Textual signals became images via Stable Diffusion (Shuai et al., 2024). GLIDE's region-based photo editing relies on integration. Natural language training allows GLIDE edit complex images. CLIP's language–image embeddings merge written and visual clues to enhance language's descriptive capacity. Picture quality is good with DDPM (Nichol et al., 2021). These two models provide unified and contextually relevant image alterations, making GLIDE excellent for accurate and creative photo editing. In classifier-free GLIDE, pictures convert text to images (Koh et al., 2024).

3.3.3. Text-to-Audio Generation

Grad-TTS and its predecessors upgraded text-to-audio generation. Grad-TTS, a popular text-to-speech model, uses Monotonic Alignment Search to convert encoder noise into voice like text input utilizing a score-based decoder and diffusion models (Wang et al., 2024). After that, GradTTS2 adapts the model to improve speech synthesis flexibility and quality. Diffsound anticipates and refines every mel-spectrogram token using a discrete diffusion model-based non-autoregressive decoder (Liu et al., 2024). EdiTTS increases generating control and precision by supplementing a coarsely modified mel-spectrogram with a score-based text-to-speech model (Zhang et al., 2023).

3.3.4. Text-to-3D Generation

Text-to-3D generation is a unique computer graphics and machine learning technology that creates 3D scenes from text. Using Imagen, Dream Fusion improves three-dimensional model synthesis by creating images from text (Liu et al., 2023). Directional text impacts Dream Fusion image creation. Direction of text influences visual perspectives. Dream Fusion renders 3D items with groundbreaking Mip-NeRF. Dream Fusion uses a pre-trained 2D text-to-image diffusion model for text-to-3D synthesis (Park et al., 2023). Randomly begun 3D models are optimized as NeRFs via probability density distillation loss. This loss function optimizes a parametric picture generator using the 2D diffusion model to create detailed and contextually relevant 3D representations from text descriptions (Chen et al., 2024). Dream Fusion shows the power of text-to-image models in 3D synthesis and the need to merge them with Mip-NeRF.

3.3.5. Molecule Generation and Drug Design

Biological and pharmaceutical molecule design uses diffusion models. These models enhance fragment-based drug design, a fundamental 3D molecular discovery tool. The E (3)-equivariant 3D-conditional diffusion model from DiffLinker (Igashov et al., 2024; Davies et al., 2024) produces molecular linkers well. To assemble molecules, a graph neural network calculates linker size and an equivariant diffusion model generates it. It advances the field by building linkers for many fragments and establishing atom count and attachment places (Zhang et al., 2023). This approach can synthesize 29-atom tiny molecules from nine heavy atoms. Equivariant graph neural networks and diffusion processes simplify training and improve performance and scalability by mimicking geometrically symmetric molecule structures. Several energy functions increase chemical synthesis and 3D point cloud uniformity in this model (Atz et al., 2024).

3.3.6. Medical Imaging

In the healthcare field, diffusion models have proven to be valuable for enhancing the quality and clarity of medical images, such as Magnetic Resonance Imaging (MRI) and Computed Tomography (CT) scans. These high-resolution images are critical for accurate diagnosis and treatment planning, especially in detecting small anomalies or early signs of diseases (Obuchowicz et al., 2024; Ajayi et al., 2024).

3.4. Current Challenges in Diffusion Models

Despite the impressive advancements in diffusion models for high-resolution image and short-form video generation, several challenges remain. These challenges are significant roadblocks to further optimization, widespread adoption, and ethical use of the technology. From the computational demands required to run these models to ethical considerations around their potential misuse, the following sections delve into the key challenges that need to be addressed.

3.4.1. Computational Complexity

Diffusion models, especially those designed for high-resolution image and video generation, are notoriously computationally expensive. The core mechanism of these models iteratively adding and removing noise through multiple steps demands substantial processing power and memory, even on high-end hardware. This computational intensity can limit the scalability and accessibility of diffusion models for everyday use (Ulhaq et al., 2022).

3.4.2. Temporal Consistency in Video Generation

While diffusion models have shown great promise in generating high-quality static images, extending this success to video generation introduces an additional layer of complexity inform of temporal consistency. In video generation, it is crucial that frames remain coherent and fluid across time to avoid visual artifacts that can detract from the overall viewing experience (Islam et al., 2023).

3.5. Ethical Concerns

Distribution models and generative AI struggle with hyper-realistic picture and film ethics. Media realism affects privacy, security, and misinformation. Use of diffusion models to create deepfakes of individuals doing or saying things they never did is unethical (Mustak et al., 2023). It could be used for political deception, fraud, and social narrative manipulation. This generated media looks like real footage, making deepfake detection and prevention difficult, threatening individuals and society.

Exploitation and creativity can overlap. AD, entertainment, and art are revolutionized by AI-generated content. Unauthorized AI model training on copyrighted information poses plagiarism and appropriation concerns (Sarkar, 2023). AI may be replacing human creativity and labour, sparking digital age ownership and remuneration issues among artists. Realism-based AI portraits bring privacy problems. Without consent, diffusion models can make photos and videos of people using their likenesses. This could compromise privacy by inappropriately using Ids (Oksanen et al., 2023). Ethical issues are being addressed by better deepfake detection, AI transparency, and regulatory frameworks. Developers are watermarking and limiting AI-generated content to prevent misuse.

4. Short-Form Video Generation

Short-form video has transformed social media enjoyment and information consumption. The rapid expansion of platforms like TikTok, Instagram Reels, and YouTube Shorts has made short-form video one of the most influential forms of digital media (Rugrien and Funk, 2022). These 15–60-second movies grab user attention in quick, snackable chunks. In the fast-paced digital world, their quick-consumption format is tempting. Short-form video is growing because to its simplicity of production, rapid distribution, and social media algorithms' viral potential. Short-form videos are easy to make and require little equipment or skill (Jennings, 2022). Creators, marketers, and businesses may now reach worldwide audiences through accessibility. Recent improvements in diffusion models can automatically generate engaging and visually appealing short-form video material (Hasan et al., 2024). The rise of short-form video is due to platforms like TikTok changing media consumption habits. TikTok's technology creates an addictive loop of content discovery by targeting user activity with personalized video feeds. Other platforms like Instagram Reels and YouTube Shorts use personalized content to drive engagement through a continual stream of bite-sized videos. Short-form video has made it easier for content makers to release videos without the production costs or time commitments of traditional forms (Li et al., 2022). Smartphones can film, edit, and upload short videos in minutes. This ease of production allows real-time viral trend creation and response, boosting audience growth and influence. small-form videos satisfy modern attention spans with small bursts of content. Short-form content is digestible, so consumers are more inclined to watch, share, and participate. High-quality short-form video material is difficult to make, despite its popularity. Short-form content is challenging to create a captivating narrative or message in a short period (Ahmed, 2024). Content makers must communicate clearly and engage the audience in the opening few seconds of the video. Marketers and companies face fierce competition in short-form video. Since millions of videos are uploaded daily, content and production quality are crucial to their success (Dwivedi et al., 2021). High-resolution images, clever storytelling, and seamless editing are now necessary to attract viewers in the saturated short-form video market. Viral short-form content means trends can shift quickly, so artists must generate videos quickly to stay relevant (Treske, 2015). Fast video production, editing, and publishing are essential, but quality typically suffers. Diffusion models help create short-form videos that fit modern content platforms' creative and technical needs (Hasan et al., 2024).

4.1. Advancements in Short-Form Video Generation

Short-form video creation is now possible with application diffusion models. Diffusion models automate video content generation, enabling high-quality, audience-targeted videos (Kreuter et al., 2013). These models learn from massive video datasets and generate new video content using patterns and structures. High-resolution videos with temporal consistency are a major improvement in diffusion models for short-form video creation (Shen et al., 2023). Temporal consistency ensures that video frames transition smoothly and coherently, preventing jerky movement and actions. In short-form video, where rapid action and quick cuts are common, this uniformity is crucial for visual appeal and professionalism.

Another innovation in short-form video is text-to-video generation. The diffusion model generates a video based on a simple textual description provided by authors using an AI-based method (Nixon et al., 2024). In marketing and advertising, firms can quickly create video content with cues like "a fast-paced video showing a product in use" or "a short tutorial explaining a new service." Content automation is possible with the model's ability to generate video sequences that match input text.

4.2. Success Metrics for Short-Form Video Curation, Summarization, and Compilation

In short-form video, where material is viewed quickly and trends change, judging success involves multiple variables. Learning how successfully a video curation, summarization, and compilation resonates with its audience and achieves its aims requires success measures. These metrics measure user engagement and how content affects retention, conversions, and emotional response. To optimize content for performance across platforms, these metrics help refine AI-generated videos, especially diffusion model videos.

4.2.1. Engagement Metrics

Engagement metrics are the main markers of short-form video viewer engagement. A video with high interactivity encourages viewers to like, share, and discuss. These metrics indicate how well AI-generated or curated videos engage and interact. AI models can optimize video output for high interaction rates by assessing engagement data and adapting material to audience preferences and behaviours (Hammar and Johansson, 2024).

4.2.2. Watch Time and Retention

Video watch time and retention rates are key indicators of audience engagement. Short-form videos must keep viewers' attention to succeed. Platform algorithms like TikTok and Instagram Reels prioritize high-retention videos in user feeds, so watch time is crucial. Retention numbers reveal content efficacy. AI systems may discover retention-boosting patterns like pacing, transitions, and visual effects and apply them to subsequent video generations (Dahl, 2024).

4.2.3. Conversion and Call-to-Action (CTA) Rates

Videos that promote businesses, products, or services use conversion metrics to measure success. CTAs, such as encouraging viewers to visit a website, buy, or read more, usually lead to conversion. This means that a short-form video's success depends on how well it motivates viewers. These movies are improved by AI models that determine which language, graphics, and timing drive conversions. This data-driven video generating method produces compelling, business-impacting videos (Sutherland, 2024).

4.2.4. Video Quality and Relevance

Video quality and relevance affect its success. Poorly produced videos perform worse than those with crisp images, smooth transitions, and expert editing. Relevance to audience interests and trends is also crucial, especially in social media's fast-changing world (Vernallis, 2013). AI models ensure that their content stands out in crowded social media feeds by focusing on quality and relevance, improving engagement and reach.

4.2.5. Sentiment Analysis

Understanding how people feel about a video requires sentiment analysis. Likes and shares reveal engagement, but sentiment analysis digs deeper into viewer emotions via comments, debates, and feedback (Munaro et al., 2021). This lets authors and companies know if their work is popular or controversial. AI-driven video generating models use sentiment analysis to match viewer emotions and expectations, improving content effectiveness and success.

4.3. Future Directions

Rapid advances in image and video diffusion models have benefited artificial intelligence and creative media. However, future advancements in this field are even more promising (Zhang et al., 2023). As academics and developers' perfect diffusion models, the next generation is projected to address present limits and enable real-time applications, hybrid modelling, and personalized content. The following sections discuss diffusion models' potential for more intelligent, efficient, and personalized media generation (Chen et al., 2024)

5. Conclusion

In conclusion, HD images and diffusion model videos revolutionized creative AI. Art, design, marketing, and customization benefit from realistic, temporally consistent media. Diffusion models differ because reverse diffusion

coherentized chaotic inputs. Unlike prior generative models, this method produces detailed images and films. Diffusion model revolutions affect many fields. Through diffusion, Instagram Reels, TikTok, and YouTube Shorts generate fast, high-quality films. Models adapt to interests, habits, and real-time interactions. Photorealism and creative abstraction are combined in diffusion models for digital storytelling by artists, designers, and filmmakers.

Combining diffusion, GANs, and reinforcement learning may improve diffusion-based media speed, quality, and diversity. The diffusion model innovation driven by advertising, education, and social media need for personalized information will create user-centric media. AI, media, and digital engagement will adopt new methods. They can revolutionize creativity and generate great results. Digital diffusion models will develop AI-driven content, revolutionizing media creation and consumption.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Ahmed, M. N. (2024). The Impact of Short-Form Video Content on Fan Engagement for Streamers. Ajayi, O. O., Wright-Ajayi, B., Mosaku, L. A., Davies, G. K., Moneke, K. C., and Adeleke, O. R. Enhancing Infectious Disease Management in Nigeria: The Role of Artificial Intelligence in Diagnosis and Treatment. *Clin Case Rep Int. 2024; 8*, *1670*.

[2] Ajayi, O. O., Wright-Ajayi, B., Mosaku, L. A., Davies, G. K., Moneke, K. C., Adeleke, O. R., ... and Mudele, O. (2024). Application of satellite imagery for vector-borne disease monitoring in sub-Saharan Africa: An overview. *GSC Advanced Research and Reviews*, *18*(3), 400-411.

[3] Akinyele, A.R., Ajayi, O.O., Munyaneza, G., Ibecheozor, U.H. and Gopakumar, N., (2024). Leveraging Generative Artificial Intelligence (AI) for cybersecurity: Analyzing diffusion models in detecting and mitigating cyber threats. GSC Advanced Research and Reviews, 21(2), pp.001-014.

[4] Alhabeeb, S. K., and Al-Shargabi, A. A. (2024). Text-to-Image Synthesis With Generative Models: Methods, Datasets, Performance Metrics, Challenges, and Future Direction. *IEEE Access*.

[5] Aslanidou-AEM, C. (2021). *Unsupervised Adversarial Video Summarization with Diverse Captions* (Doctoral dissertation, Aristotle University of Thessaloniki).

[6] Asperti, A., Evangelista, D., and Loli Piccolomini, E. (2021). A survey on variational autoencoders from a green AI perspective. *SN Computer Science*, *2*(4), 301.

[7] Atz, K., Cotos, L., Isert, C., Håkansson, M., Focht, D., Hilleke, M., ... and Schneider, G. (2024). Prospective de novo drug design with deep interactome learning. *Nature Communications*, *15*(1), 3408.

[8] Bandi, A., Adapa, P. V. S. R., and Kuchi, Y. E. V. P. K. (2023). The power of generative ai: A review of requirements, models, input–output formats, evaluation metrics, and challenges. *Future Internet*, *15*(8), 260.

[9] Bengesi, S., El-Sayed, H., Sarker, M. K., Houkpati, Y., Irungu, J., and Oladunni, T. (2024). Advancements in Generative AI: A Comprehensive Review of GANs, GPT, Autoencoders, Diffusion Model, and Transformers. *IEEE Access*.

[10] Bommasani, R., and Cardie, C. (2020, November). Intrinsic evaluation of summarization datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 8075-8096).

[11] Bounoua, M., Franzese, G., and Michiardi, P. (2024). Multi-modal latent diffusion. *Entropy*, *26*(4), 320.

[12] Bushey, J. (2015, January). Trustworthy citizen-generated images and video on social media platforms. In *2015 48th Hawaii International Conference on System Sciences* (pp. 1553-1564). IEEE.

[13] Cao, H., Tan, C., Gao, Z., Xu, Y., Chen, G., Heng, P. A., and Li, S. Z. (2024). A survey on generative diffusion models. *IEEE Transactions on Knowledge and Data Engineering*.

[14] Chakraborty, T., KS, U. R., Naik, S. M., Panja, M., and Manvitha, B. (2024). Ten years of generative adversarial nets (GANs): a survey of the state-of-the-art. *Machine Learning: Science and Technology*, *5*(1), 011001.

[15] Chauhan, A., Rabbani, S., Agarwal, D., Akhtar, N., and Perwej, Y. (2024). Diffusion Dynamics Applied with Novel Methodologies. *International Journal of Innovative Research in Computer Science and Technology*, *12*(4), 52-58.

[16] Chen, F., Yang, Z., Zhuang, B., and Wu, Q. (2024). Streaming Video Diffusion: Online Video Editing with Diffusion Models. *arXiv preprint arXiv:2405.19726*.

[17] Chen, H., Loy, C. C., and Pan, X. (2024). MVIP-NeRF: Multi-view 3D Inpainting on NeRF Scenes via Diffusion Prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5344-5353).

[18] Chen, H., Xia, M., He, Y., Zhang, Y., Cun, X., Yang, S., ... and Shan, Y. (2023). Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*.

[19] Chen, J., Shao, Z., Zheng, X., Zhang, K., and Yin, J. (2024). Integrating aesthetics and efficiency: AI-driven diffusion models for visually pleasing interior design generation. *Scientific Reports*, *14*(1), 3496.

[20] Chen, P., Zhang, Y., Tan, M., Xiao, H., Huang, D., and Gan, C. (2020). Generating visually aligned sound from videos. *IEEE Transactions on Image Processing*, *29*, 8292-8302.

[21] Cooper, A. F. (2024). *Between Randomness and Arbitrariness: Some Lessons for Reliable Machine Learning at Scale* (Doctoral dissertation, Cornell University).

[22] Dahl, H. (2024). How to Create Value in Social Media: a Postphenomenological Perspective: The technological mediation of social media between users and business-the effect on value of human factors engineering and customer engagement seen from a postphenomenological view.

[23] Davies, G. K., Ajayi, O. O., Wright-Ajayi, B., Mosaku, L. A., Moneke, K. C., Adeleke, O. R., ... and Mudele, O. (2024). Unravelling the complexity of environmental exposures and health: A novel exposome-centered framework for occupational and environmental epidemiology. *GSC Advanced Research and Reviews*, *19*(1), 026-032.

[24] Dwivedi, Y. K., Ismagilova, E., Hughes, D. L., Carlson, J., Filieri, R., Jacobson, J., ... and Wang, Y. (2021). Setting the future of digital and social media marketing research: Perspectives and research propositions. *International journal of information management*, *59*, 102168.

[25] Engström, A., Juhlin, O., Perry, M., and Broth, M. (2010, April). Temporal hybridity: Mixing live video footage with instant replay in real time. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1495-1504).

[26] Garon, J. (2023). A practical introduction to generative AI, synthetic media, and the messages found in the latest medium. *Synthetic Media, and the Messages Found in the Latest Medium (March 14, 2023)*.

[27] Guo, Y., Liu, Q., Chen, J., Xue, W., Fu, J., Jensen, H., ... and Xu, J. (2022). Pathway to future symbiotic creativity. *arXiv preprint arXiv:2209.02388*.

[28] Hammar, I. N., and Johansson, S. (2024). Exploring the Dynamics of Short-Format Videos on Social.

[29] Hasan, M., Athrey, K. S., Khalid, A., Xie, D., Younessian, E., and Braskich, T. (2024). Applications of Computer Vision in Entertainment and Media Industry. *Computer Vision: Challenges, Trends, and Opportunities*, 205.

[30] Hazra, T., and Anjaria, K. (2022). Applications of game theory in deep learning: a survey. *Multimedia Tools and Applications*, *81*(6), 8963-8994.

[31] He, C., Shen, Y., Fang, C., Xiao, F., Tang, L., Zhang, Y., ... and Li, X. (2024). Diffusion Models in Low-Level Vision: A Survey. *arXiv preprint arXiv:2406.11138*.

[32] He, J., Wang, X., Liu, S., Wu, G., Silva, C., and Qu, H. (2024). POEM: Interactive Prompt Optimization for Enhancing Multimodal Reasoning of Large Language Models. *arXiv preprint arXiv:2406.03843*.

[33] Huang, G. (2022). Towards meaningful and data-efficient learning: exploring GAN losses, improving few-shot benchmarks, and multimodal video captioning.

[34] Huang, Q., Park, D. S., Wang, T., Denk, T. I., Ly, A., Chen, N., ... and Han, W. (2023). Noise2music: Text-conditioned music generation with diffusion models. *arXiv preprint arXiv:2302.03917*.

[35] Igashov, I., Stärk, H., Vignac, C., Schneuing, A., Satorras, V. G., Frossard, P., ... and Correia, B. (2024). Equivariant 3D-conditional diffusion model for molecular linker design. *Nature Machine Intelligence*, 1-11.

[36] Islam, T., Miron, A., Liu, X., and Li, Y. (2023). FashionFlow: Leveraging diffusion models for dynamic fashion video synthesis from static imagery. *arXiv preprint arXiv:2310.00106*.

[37] Jabbar, A., Li, X., and Omar, B. (2021). A survey on generative adversarial networks: Variants, applications, and training. *ACM Computing Surveys (CSUR)*, *54*(8), 1-49.

[38] Jadhav, B., Jain, M., Jajoo, A., Kadam, D., Kadam, H., and Kakkad, T. (2024, July). Imagination Made Real: Stable Diffusion for High-Fidelity Text-to-Image Tasks. In *2024 2nd International Conference on Sustainable Computing and Smart Systems (ICSCSS)* (pp. 773-779). IEEE.

[39] Jansen, B. J., Aldous, K. K., Salminen, J., Almerekhi, H., and Jung, S. G. (2023). *Understanding Audiences, Customers, and Users Via Analytics: An Introduction to the Employment of Web, Social, and Other Types of Digital People Data*. Springer Nature.

[40] Jennings, E. M. P. (2022). *Short video marketing: A good strategy for small businesses on TikTok?* (Doctoral dissertation).

[41] Kalateh, S., Estrada-Jimenez, L. A., Hojjati, S. N., and Barata, J. (2024). A Systematic Review on Multimodal Emotion Recognition: Building Blocks, Current State, Applications, and Challenges. *IEEE Access*.

[42] Khan, W., Leem, S., See, K. B., Wong, J. K., Zhang, S., and Fang, R. (2024). A Comprehensive Survey of Foundation Models in Medicine. *arXiv preprint arXiv:2406.10729*.

[43] Khurana, K., and Deshpande, U. (2021). Video question-answering techniques, benchmark datasets and evaluation metrics leveraging video captioning: a comprehensive survey. *IEEE Access*, *9*, 43799-43823.

[44] Koh, J. Y., Fried, D., and Salakhutdinov, R. R. (2024). Generating images with multimodal language models. *Advances in Neural Information Processing Systems*, *36*.

[45] Kreuter, M. W., Farrell, D. W., Olevitch, L. R., and Brennan, L. K. (2013). *Tailoring health messages: Customizing communication with computer technology*. Routledge.

[46] Lan, L., You, L., Zhang, Z., Fan, Z., Zhao, W., Zeng, N., ... and Zhou, X. (2020). Generative adversarial networks and its applications in biomedical informatics. *Frontiers in public health*, *8*, 164.

[47] Lavda, F. (2024). *Improving the capabilities of Variational Autoencoder Models by exploring their latent space* (Doctoral dissertation, Université de Genève).

[48] Li, G., Wei, Y., Chi, Y., and Chen, Y. (2024). A Sharp Convergence Theory for The Probability Flow ODEs of Diffusion Models. *arXiv preprint arXiv:2408.02320*.

[49] Li, Y. Q., Kim, H. J., and Lee, H. G. (2022). Why Do Users Participate in Hashtag Challenges in a Short-form Video Platform? The Role of Para-Social Interaction. *정보화정책*, *29*(3), 82-104.

[50] Liu, Z., Dai, P., Li, R., Qi, X., and Fu, C. W. (2023). DreamStone: Image as a Stepping Stone for Text-Guided 3D Shape Generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[51] Lyu, Z., Kong, Z., Xu, X., Pan, L., and Lin, D. (2021). A conditional point diffusion-refinement paradigm for 3d point cloud completion. *arXiv preprint arXiv:2112.03530*.

[52] Markides, C. C. (2013). *Game-changing strategies: How to create new market space in established industries by breaking the rules*. John Wiley and Sons.

[53] Moser, B. B., Shanbhag, A. S., Raue, F., Frolov, S., Palacio, S., and Dengel, A. (2024). Diffusion models, image super-resolution and everything: A survey. *arXiv preprint arXiv:2401.00736*.

[54] Mudele, O. and Gamba, P., (2019), May. Mapping vegetation in urban areas using Sentinel-2. In 2019 Joint Urban Remote Sensing Event (JURSE) (pp. 1-4). IEEE.

[55] Mudele, O., Frery, A. C., Zanandrez, L. F. R., Eiras, A. E., and Gamba, P. (2021a). Dengue vector population forecasting using multisource earth observation products and recurrent neural networks. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14, 4390-4404.

[56] Mudele, O., Frery, A. C., Zanandrez, L. F. R., Eiras, A. E., and Gamba, P. (2021b). Modeling dengue vector population with earth observation data and a generalized linear model. Acta Tropica, 215, 105809.

[57] Munaro, A. C., Hübner Barcelos, R., Francisco Maffezzolli, E. C., Santos Rodrigues, J. P., and Cabrera Paraiso, E. (2021). To engage or not engage? The features of video content on YouTube affecting digital consumer engagement. *Journal of consumer behaviour*, *20*(5), 1336-1352.

[58] Mustak, M., Salminen, J., Mäntymäki, M., Rahman, A., and Dwivedi, Y. K. (2023). Deepfakes: Deceptions, mitigations, and opportunities. *Journal of Business Research*, *154*, 113368.

[59] Remya Revi, K., Vidya, K. R., and Wilscy, M. (2021). Detection of deepfake images created using generative adversarial networks: A review. In *Second International Conference on Networks and Advances in Computational Technologies: NetACT 19* (pp. 25-35). Springer International Publishing.

[60] Singh, V. (2023). An In-Depth Guide to Denoising Diffusion Probabilistic Models–From Theory to Implementation. *Learn Open CV*, *6*.

[61] Nguyen, T. T., and Veer, E. (2024). Why people watch user-generated videos? A systematic review and meta-analysis. *International Journal of Human-Computer Studies*, *181*, 103144.

[62] Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., ... and Chen, M. (2021). Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

[63] Niu, K., Liu, W., Sharif, N., and Zhu, D. (2024). Conditional Video Generation Guided by Multimodal Inputs: A Comprehensive Survey.

[64] Nixon, L., Apostolidis, K., Apostolidis, E., Galanopoulos, D., Mezaris, V., Philipp, B., and Bocyte, R. (2024). AI and data-driven media analysis of TV content for optimised digital content marketing. *Multimedia Systems*, *30*(1), 25.

[65] Obuchowicz, R., Strzelecki, M., and Piórkowski, A. (2024). Clinical Applications of Artificial Intelligence in Medical Imaging and Image Processing—A Review. *Cancers*, *16*(10), 1870.

[66] Oksanen, A., Cvetkovic, A., Akin, N., Latikka, R., Bergdahl, J., Chen, Y., and Savela, N. (2023). Artificial intelligence in fine arts: A systematic review of empirical research. *Computers in Human Behavior: Artificial Humans*, 100004.

[67] Olán, T. (2023). Generation and generalized detection of fully synthetic photorealistic images. *Image*.

[68] Park, J., Kwon, G., and Ye, J. C. (2023). ED-NeRF: Efficient Text-Guided Editing of 3D Scene using Latent Space NeRF. *arXiv preprint arXiv:2310.02712*.

[69] Po, R., Yifan, W., Golyanik, V., Aberman, K., Barron, J. T., Bermano, A., ... and Wetzstein, G. (2024, May). State of the art on diffusion models for visual computing. In *Computer Graphics Forum* (Vol. 43, No. 2, p. e15063).

[70] Rugrien, M. P., and Funk, S. (2022). *Social media Trend 2023: Short-form VS. Long-form Video* (No. 308246). Thammasat University. Faculty of Journalism and Mass Communication.

[71] Rugrien, M. P., and Funk, S. (2022). *Social media Trend 2023: Short-form VS. Long-form Video* (No. 308246). Thammasat University. Faculty of Journalism and Mass Communication.

[72] Sajjadi, S. M. M. (2024). *Enhancement and Evaluation of Deep Generative Networks with Applications in Super-Resolution and Image Generation* (Doctoral dissertation, Universität Tübingen).

[73] Sarkar, A. (2023, June). Exploring perspectives on the impact of Artificial Intelligence on the Creativity of Knowledge Work: Beyond Mechanised Plagiarism and Stochastic parrots. In *Proceedings of the 2nd Annual Meeting of the Symposium on Human-Computer Interaction for Work* (pp. 1-17).

[74] Schofield, R., King, L., Tayal, U., Castellano, I., Stirrup, J., Pontana, F., ... and Nicol, E. (2020). Image reconstruction: Part 1–understanding filtered back projection, noise and image acquisition. *Journal of cardiovascular computed tomography*, *14*(3), 219-225.

[75] Shen, L., Li, X., Sun, H., Peng, J., Xian, K., Cao, Z., and Lin, G. (2023, October). Make-it-4d: Synthesizing a consistent long-term dynamic scene video from a single image. In *Proceedings of the 31st ACM International Conference on Multimedia* (pp. 8167-8175).

[76] Shuai, X., Ding, H., Ma, X., Tu, R., Jiang, Y. G., and Tao, D. (2024). A Survey of Multimodal-Guided Image Editing with Text-to-Image Diffusion Models. *arXiv preprint arXiv:2406.14555*.

[77] Sutherland, K. E. (2024). Producing Videos that Pop. In *Strategic Social Media Management: Theory and Practice* (pp. 503-562). Singapore: Springer Nature Singapore.

[78] Talafha, S. (2022). *Generative Models in Natural Language Processing and Computer Vision*. Southern Illinois University at Carbondale.

[79] Treske, A. (2015). *Video theory*. transcript Verlag.

[80] Trinh, L. T., and Hamagami, T. (2024). Latent Denoising Diffusion GAN: Faster sampling, Higher image quality. *IEEE Access*.

[81] Ulhaq, A., Akhtar, N., and Pogrebna, G. (2022). Efficient diffusion models for vision: A survey. *arXiv preprint arXiv:2210.09292*.

[82] Vernallis, C. (2013). *Unruly media: YouTube, music video, and the new digital cinema*. Oxford University Press.

[83] Wang, W., Bao, J., Zhou, W., Chen, D., Chen, D., Yuan, L., and Li, H. (2022). Semantic image synthesis via diffusion models. *arXiv preprint arXiv:2207.00050*.

[84] Wang, X., He, Z., and Peng, X. (2024). Artificial-Intelligence-Generated Content with Diffusion Models: A Literature Review. *Mathematics*, *12*(7), 977.

[85] Williams, R. J., and Peng, J. (1990). An efficient gradient-based algorithm for on-line training of recurrent network trajectories. *Neural computation*, *2*(4), 490-501.

[86] Xing, Z., Feng, Q., Chen, H., Dai, Q., Hu, H., Xu, H., ... and Jiang, Y. G. (2023). A survey on video diffusion models. *ACM Computing Surveys*.

[87] Yu, X., Wang, Y., Chen, Y., Tao, Z., Xi, D., Song, S., and Niu, S. (2024). Fake Artificial Intelligence Generated Contents (FAIGC): A Survey of Theories, Detection Methods, and Opportunities. *arXiv preprint arXiv:2405.00711*.

[88] Zhang, C., Zhang, C., Zhang, M., and Kweon, I. S. (2023). Text-to-image diffusion models in generative ai: A survey. *arXiv preprint arXiv:2303.07909*.

[89] Zhang, C., Zhang, C., Zheng, S., Qiao, Y., Li, C., Zhang, M., ... and Hong, C. S. (2023). A complete survey on generative ai (aigc): Is chatgpt from gpt-4 to gpt-5 all you need? *arXiv preprint arXiv:2303.11717*.

[90] Zhang, Z., Yan, J., Liu, Q., Chen, E., and Zitnik, M. (2023). A systematic survey in geometric deep learning for structure-based drug design. *arXiv preprint arXiv:2306.11768*.

[91] Zhao, Y., Liu, Z., and Wu, J. (2020). Grassland ecosystem services: a systematic review of research advances and future directions. *Landscape Ecology*, *35*, 793-814.