A CASE STUDY OF AI IN VISION AND SOUND GENERATION: THE EVOLUTION, APPLICATIONS, AND ETHICAL CONSIDERATIONS

by

Jingheng Huan

Signature Work Product, in partial fulfillment of the
Duke Kunshan University Undergraduate Degree Program

Mar 7, 2024

Signature Work Program
Duke Kunshan University

APPROVALS

Mentor: Peng Sun, Division of Natural and Applied Sciences

Marcia B. France, Dean of Undergraduate Studies

CONTENTS

Abstract iii

Acknowledgements iv

List of Figures v

1 Introduction 1
- 1.1 Artificial Intelligence (AI) . . . 1
- 1.2 Artificial Intelligence Generated Content (AIGC) . . . 2
- 1.3 Current Development Trends . . . 3
- 1.4 Case Study: Using AI tools to predict the future of soccer . . . 5

2 Methods 7
- 2.1 Workflow . . . 7
- 2.2 AIGC in Text (ChatGPT) . . . 8
- 2.3 AIGC in Image and Video (MidJourney Runway Topaz) . . . 11
- 2.4 AIGC in Sound (ElevenLabs) . . . 15

3 Results 23
- 3.1 Overview of Project Outcome . . . 23
- 3.2 Deliverable . . . 24

4 Discussion 31
- 4.1 Foundational Models . . . 31
- 4.2 Specialized Architectures . . . 31
- 4.3 Ethical and Societal Considerations . . . 32
- 4.4 Efficiency and Scalability . . . 32
- 4.5 Extension to Sound Generation . . . 33
- 4.6 Future Directions . . . 33
- 4.7 Mitigating Biases in Generative Systems . . . 34
- 4.8 Computational Efficiency and Resolution Enhancement in Visual Processing . . . 35

5 Conclusions 36

References 38

A Additional Material 42

ABSTRACT

This study embarks on an in-depth exploration of the transformative impact of Artificial Intelligence (AI) on the realms of image, video, and sound generation. It presents a comparative analysis of cutting-edge AI technologies, including Generative Adversarial Networks (GANs), Diffusion Models, and Neural Cellular Automata, highlighting their development, theoretical underpinnings, and practical applications. The research delves into the significant advancements these technologies have made possible, showcasing their ability to produce highly realistic and creative outputs. Moreover, it addresses the crucial ethical considerations surrounding the use of AI in content creation, emphasizing the need for robust guidelines to ensure the responsible deployment of these technologies. The findings of this study underline the potential of AI to revolutionize digital media production, pushing the boundaries of creativity and opening up new possibilities for innovation across various sectors. This abstract provides a concise overview of the research's scope, methodologies, and key contributions, setting the stage for a deeper discussion on the evolving role of AI in shaping the future of digital media.

本研究深入探讨了人工智能（AI）在图像、视频和声音生成领域的变革性影响。通过对最先进的 AI 技术进行比较分析，包括生成对抗网络（GANs）、扩散模型和神经细胞自动机，突出展示了这些技术的发展、理论基础和实际应用。研究深入探讨了这些技术实现的重大进步，展示了它们产生高度逼真和富有创造性输出的能力。此外，它还讨论了 AI 在内容创作中使用的关键伦理考虑，强调需要制定严格的指导方针，以确保这些技术的负责任部署。本研究的发现强调了 AI 革新数字媒体生产的潜力，推动创造力的边界，并在各个领域开辟了创新的新可能性。这个摘要提供了对研究范围、方法论和主要贡献的简洁概述，为深入讨论 AI 在塑造数字媒体未来中的不断发展作用奠定了基础。

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my SW mentor, Prof. Peng Sun, for his invaluable guidance, patience, and insight throughout the duration of this project. His expertise and encouragement were instrumental in navigating the challenges encountered during this journey. I also extend my gratitude to Xiaoqi Su in the SW team, whose support and advice enriched my research experience significantly.

Special thanks to the SW office for their support and SW grants, amounting to ¥1800, which was crucial for the experiential learning component of my project. This financial support was essential in realizing the objectives of my research, and I am deeply appreciative of their generosity.

Additionally, I wish to express my thanks to my family and friends for their encouragement and belief in my abilities. Their constant motivation and the confidence they placed in me were the foundation upon which I built my perseverance and dedication to this project.

Lastly, I acknowledge the contributions of the broader academic and research community whose works have provided the inspiration for this study. Their efforts in the field of AI have paved the way for emerging researchers to explore new frontiers and contribute to the ongoing dialogue in this dynamic and ever-evolving domain.

LIST OF FIGURES

1.1 A comprehensive overview of key milestones and product launches in the generative AI video domain throughout the year 2023, with a notable concentration of developments occurring in the fourth quarter. [35] 3
1.2 Comparative analysis of various AI video product offerings as of December 2023, detailing generation types, maximum length of generated clips, and additional features such as camera and motion control, showcasing the diversity and capabilities within the current market landscape.[35] 4
1.3 Progression of generative modeling techniques from 2016 to 2022, showcasing the evolution and scaling of model parameters in GAN, Transformer, and Diffusion methods, each color-coded by method type.[40] 5

2.1 Workflow of the video-creating project 7
2.2 Relationships in a Multimodal System 8
2.3 Using ChatGPT to generate vivid description prompt and prepare for creating the image stills. 10
2.4 Using ChatGPT's prompts to generate good quality image stills using MidJourney, and then upload them as the key frames to Runway Gen-2 for video clip generation. 12
2.5 Using the "Image + Text" mode in Runway Gen-2 to generate the video, with the help of motion brush function to control how the characters move 13
2.6 Using Topaz AI to upscale and improve the video quality, like changing the frame rate from 60 FPS to 120 FPS, removing noise, and sharpen the subject. 15
2.7 Using Professinal Voice Cloning from ElevenLabs to apply text to speech and add the voice over to the video 16
2.8 Using Professinal Voice Cloning from ElevenLabs to apply text to speech and add the voice over to the video 17
2.9 Autoregressive transformer model architecture used for audio processing tasks like transcription and voice synthesis. The model utilizes a stack of encoder and decoder blocks, with sinusoidal positional encodings and learned positional encodings for sequence representation. [54] 21
2.10 Using my clone voice to generate the voice over for the video, I recorded the training set by reading one chapter from the book *Elon Musk*, the audio file is 14 mins. 22

3.1 An in-depth analysis of an MP3 file showcasing its spectral dynamics, waveform, and amplitude in a colorful sound engineering software interface. 24
3.2 A pitch-black screen gradually reveals a soccer ball radiating a soft, ethereal glow. The light pulses gently, casting shadows that mimic fans gathered in a silent stadium, anticipation hanging in the air. 25
3.3 A swift zoom through time brings us to the present, landing on the electric atmosphere of the 2026 World Cup stadium. The architecture is a marvel of modern design, with sweeping curves and glowing lights, packed to the brim with a sea of fans from across the globe. 25
3.4 Close-up shots dive into the crowd's heart, capturing faces painted in national colors, eyes sparkling with excitement. Fans in synchronized motion, creating a living tapestry of support that breathes life into the stadium's structure. . . . 26
3.5 Transition to the crystal-clear, high-definition 2010s, where the speed and emotion of the game captivate. A player's face is frozen in a moment of triumph, surrounded by a blur of teammates and confetti, encapsulating modern soccer's glory. 26
3.6 A midfielder weaves through the opposition with balletic grace, the ball seemingly tethered to their feet. Each touch is a testament to years of dedication, the culmination of a lifetime's passion for the game. 27
3.7 The climactic moment as the ball arcs beautifully into the net, the scorer's joy uncontainable as they sprint towards the roaring fans, arms spread wide, embodying the euphoria of triumph. Teammates converge in a jubilant huddle, a symbol of unity and success. 27
3.8 The stadium, an architectural marvel, seamlessly blends organic designs with advanced technology. Bioluminescent structures illuminate the arena, while floating screens display live stats and player information. The atmosphere is electric, buzzing with anticipation for the historic final. 28
3.9 Inside the stadium, two teams of robots engage in a pre-game warm-up, showcasing incredible precision and agility. Their movements are fluid yet unmistakably mechanical, with LED lights tracing patterns across their metallic frames, each robot adorned with the colors of their respective teams. 28
3.10 Focus on a uniquely designed soccer ball at the center circle, glowing with internal lights and equipped with smart technology that allows it to record and respond to the game's flow. A player approaches and gives it a gentle tap, setting it aglow, ready for kickoff. 29
3.11 The kickoff is a spectacle of technology, with a robotic player skillfully propelling the ball into play using a swift, calculated motion. The ball itself is equipped with smart tech, glowing softly as it moves across the pitch, data streaming off it in visual waves that are visible to the audience. 29

3.12 Human players are shown warming up on the field, executing stretches and drills with the backdrop of a stadium that blends organic and technological aesthetics. The turf is a perfect green, with embedded lighting that responds to the players' movements, illuminating their paths across the pitch. 30
3.13 The climax of the segment is a goal that feels both alien and familiar. A robot striker launches the ball with pinpoint accuracy from outside the penalty box, the ball tracing a perfect arc into the top corner of the goal. The net lights up, and digital fireworks explode overhead, celebrating the score in a dazzling display of light and color. 30

Chapter 1

INTRODUCTION

1.1 Artificial Intelligence (AI)

Artificial Intelligence (AI) is an expansive field that integrates principles from computer science, mathematics, and neuroscience to set up systems capable of simulating human cognitive functions and executing tasks traditionally requiring human intellect [1]. This interdisciplinary approach transforms AI from a mere conceptual framework into an indispensable tool across diverse sectors such as technology, finance, and entertainment. The core of AI's functionality is its ability to learn from data and make decisions autonomously, without hard coding the explicit instructions. One of the foundational models in AI's evolution is machine learning [2], which includes techniques for classification, regression, and clustering [3]. These methods allow computers to learn patterns and make predictions from data, forming the basis for many early AI applications. As the field matured, researchers developed more sophisticated models, including neural networks [4], which imitate the structure and function of the human brain to perform complex pattern recognition tasks [5].

The evolution of AI is marked by significant milestones, particularly with the advent of deep learning [6] techniques, which have revolutionized areas like computer vision [7] and natural language processing (NLP) [8]. Deep learning, a subset of machine learning, employs complex neural networks with multiple layers of processing units, enabling the extraction of high-level features from raw input data. It improved the performance of AI systems in tasks ranging from image and speech recognition to language translation [9]. Recent years some specialized architectures have further pushed the boundaries of what AI can achieve. Variational Autoencoders (VAE), [10, 11], Generative Adversarial Networks (GANs) [12–16], and diffusion models [17–20] represent the forefront of AI research in data generation. GANs consist of two neural networks—the generator and the discriminator—competing against each other to generate new, synthetic instances of data that are indistinguishable from real data. This has proven especially powerful in the fields of vision and sound generation, enabling the creation of photorealistic images, videos, and lifelike synthetic audio [21].

The advancements in AI have not only expanded the horizons of what is technically feasible but have also paved the way for innovations in digital content creation that were previously unimaginable. The development and application of StyleGAN [22] and its successors have been pivotal in pushing the limits of image generation technologies. These models have the remarkable ability to produce images of astonishing realism, manipulate facial expressions with precision, and even venture into the realm of digital art creation [23]. Similarly, the advent of diffusion models marks a significant leap in the quality of generated content, setting unprecedented standards for both image and sound generation [17]. These models operate by iteratively refining data inputs through a process that progressively reduces noise, allowing for the creation of outputs that are not only highly detailed but also deeply resonant with the nuances of real-world sensory experiences. Such advancements underscore the multifaceted impact of AI in the digital domain, where the synergy between complex neural networks and deep learning techniques is not just enhancing the visual and auditory quality of synthetic media but is also broadening the scope of creative expression and immersion in virtual environments. This ongoing evolution in AI-driven content generation reflects a convergence of insights from across computer science, mathematics, and neuroscience, highlighting the interdisciplinary effort that underpins AI's capability to reinterpret and reimagine the way machines understand and interact with the world around us [24].

1.2 Artificial Intelligence Generated Content (AIGC)

In the field of Artificial Intelligence Generated Content (AIGC), platforms like ChatGPT [25] [26], MidJourney [27], Runway [28], and ElevenLabs [29] are redefining the boundaries of creative expression. ChatGPT, with its sophisticated large language model, excels in generating text that mocks human-like understanding and creativity, making it a helpful tool for crafting compelling narratives, dialogues, and even complex literary works. Its ability to process and produce coherent, contextually relevant content on a vast array of topics has democratized content creation, enabling writers and creators to generate high-quality text-based content efficiently. Similarly, MidJourney revolutionizes the visual aspect of content creation by transforming textual descriptions into detailed, high-resolution images, and also style transformation. This AI-driven approach to art and design allows creators to visualize abstract concepts and bring their most imaginative ideas to life, significantly simplifying the visual image storytelling process. Runway offers a toolkit of video generation and editing powered by AI, enabling creators to automate labor-intensive tasks such as object detection, background removal, and even generating video clips from textual prompts. This not only streamlines the video production process but also opens up new avenues for creativity and experimentation in visual content. ElevenLabs, with its state-of-the-art voice synthesis technology, provides solutions for text to speech and speech to text, changing the language of a speech, and even cloning the voice. Whether for dubbing, voiceovers, or virtual assistants, ElevenLabs offers unparalleled customization, allowing for the creation of unique vocal identities that can speak in multiple languages and tones. The combination of these AI models—ChatGPT, MidJourney, Runway, and ElevenLabs prelude a new age in digital content creation, where the fusion of text, image, video, and voice generated through AI not only enriches the content landscape but also offers creators the tools to craft experiences that were once unimaginable.

However, AIGC technology has also sparked ethical debates concerning the creation and use of AI-generated content. Deepfakes [30], which involve altering videos and images to the point of being indistinguishable from authentic media, have the potential to distort reality and spread misinformation. The manipulation of media content raises concerns about the impact on individuals' reputations and the broader implications for societal and political discourse. As artificial intelligence capabilities grow, the need for ethical guidelines that ensure transparency and prevent biases becomes more urgent. Coupled with the ethical concerns surrounding deepfakes is the issue of copyright in AI-generated images and videos. [31] The legal system is currently unprepared to tackle questions of ownership and violation when content is created by an AI. This prompts a discussion about the originality of AI-generated works and the ethical use of AI in creative processes. As AI becomes more pervasive, it will challenge traditional copyright laws, necessitating continuous research and the development of legal measures that keep pace with AI evolution.

1.3 Current Development Trends

The evolution of generative AI has unfolded with unexpected speed. The transition from large, centralized systems to more accessible and powerful technologies has seen generative AI quickly reach a phase where enthusiasts and professionals can experiment and innovate with considerable autonomy. The proliferation of smaller, more efficient foundation models, such as Meta's LLaMa family [32], StableLM [33], and Mistral [34], indicate a move towards greater performance with reduced resource requirements. These applications, alongside the increasing availability of open-source models, are democratizing AI by enabling a wider array of users to access state-of-the-art capabilities, fostering a more inclusive AI ecosystem.

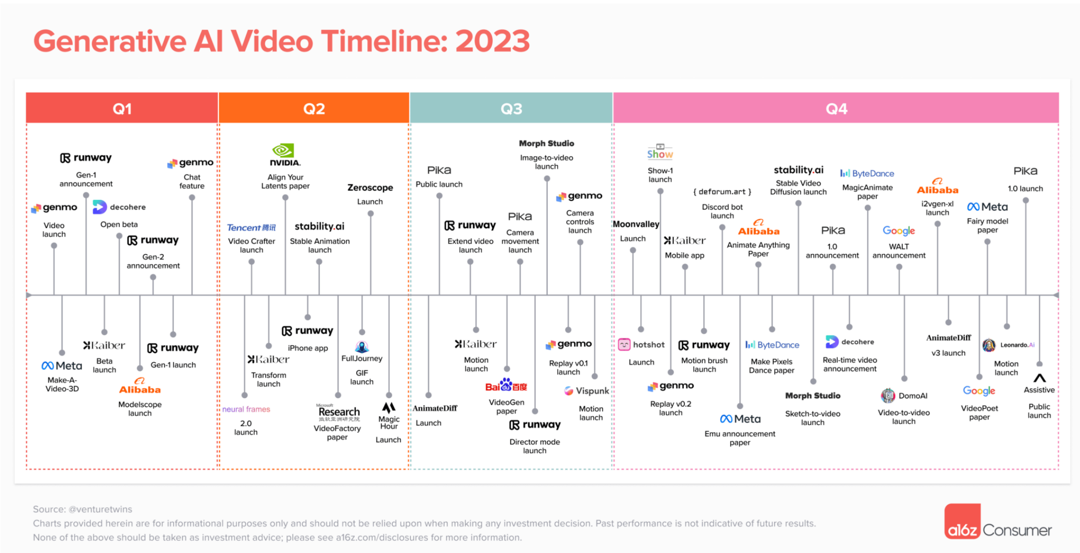


Figure 1.1: A comprehensive overview of key milestones and product launches in the generative AI video domain throughout the year 2023, with a notable concentration of developments occurring in the fourth quarter. [35]

As the demand for AI capabilities continues to surge, the AI industry faces practical challenges such as GPU shortages and rising cloud costs. This situation is pushing developers to innovate with hardware-efficient models and explore new deployment environments. The ability to run smaller models locally on devices is particularly significant, offering privacy advantages and opening up possibilities for edge computing and IoT applications. The ability to customize and fine-tune AI models is becoming increasingly important for businesses looking to differentiate their services. Open-source models offer the flexibility to develop tailored AI solutions, trained on proprietary data and optimized for specific industry needs. This approach is particularly relevant in sectors like healthcare, legal, and finance, which require specialized knowledge and expertise to generate high-quality content.

AI Video Products as of December 2023

Company	Generation Type	Max Length	Extend?	Camera Controls? (zoom, pan)	Motion Control? (amount)	Other Features	Format
Runway	Text-to-video, image-to-video, video-to-video	4 sec	Yes	Yes	Yes	Motion brush, upscale	Website
Pika	Text-to-video, image-to-video	3 sec	Yes	Yes	Yes	Modify region, expand canvas, upscale	Website
Genmo	Text-to-video, image-to-video	6 sec	No	Yes	Yes	FX presets	Website
Kaiber	Text-to-video, image-to-video, video-to-video	16 sec	No	No	No	Sync to music	Website
Stability	Image-to-video	4 sec	No	No	Yes		Local model, SDK
Zeroscope	Text-to-video	3 sec	No	No	No		Local model
ModelScope	Text-to-video	3 sec	No	No	No		Local model
AnimateDiff	Text-to-video, image-to-video, video-to-video	3 sec	No	No	No		Local model
Morph	Text-to-video	3 sec	No	No	No		Discord bot
Hotshot	Text-to-video	2 sec	No	No	No		Website
Moonvalley	Text-to-video, image-to-video	3 sec	No	No	No		Discord bot
Deforum	Text-to-video	14 sec	No	Yes	No	FX presets	Discord bot
Leonardo	Image-to-video	4 sec	No	No	Yes		Website
Assistive	Text-to-video, image-to-video	4 sec	No	No	Yes		Website
Neural Frames	Text-to-video, image-to-video, video-to-video	Unlimited	No	No	No	Sync to music	Website
Magic Hour	Text-to-video, image-to-video, video-to-video	Unlimited	No	No	No	Face swap, sync to music	Website
Vispunk	Text-to-video	3 sec	No	Yes	No		Website
Decohere	Text-to-video, image-to-video	4 sec	No	No	Yes		Website
Domo AI	Image-to-video, video-to-video	3 sec	No	No	Yes		Discord bot
FullJourney	Text-to-video, image-to-video	8 sec	No	Yes	No	Lipsyncing, face swap	Discord bot

Source: @venturetwins
Charts provided herein are for informational purposes only and should not be relied upon when making any investment decision. Past performance is not indicative of future results.
None of the above should be taken as investment advice; please see a16z.com/disclosures for more information.

Figure 1.2: Comparative analysis of various AI video product offerings as of December 2023, detailing generation types, maximum length of generated clips, and additional features such as camera and motion control, showcasing the diversity and capabilities within the current market landscape.[35]

The regulatory environment for AI is rapidly evolving, with significant developments in the EU, China, and the US. The EU's provisional agreement on the AI Act [36], China's measures to regulate AI usage [37], and the US's executive order on AI governance [38] reflect a global movement towards establishing legal frameworks to ensure responsible AI development and deployment. The outcome of ongoing legal battles, such as the New York Times' lawsuit against OpenAI [39], may have far-reaching implications for the future of AI regulation and its impact on innovation and deployment. As we look towards the future, the most impactful developments in AI may well be centered around governance, middleware, and data pipelines that make generative AI more trustworthy and accessible. With a more refined understanding of

AI capabilities, businesses are now focusing on integrating AI tools into existing services to enhance rather than revolutionize established processes. The challenge lies in striking a balance between leveraging the unique opportunities presented by AI while managing realistic expectations about its role in reshaping business practices.

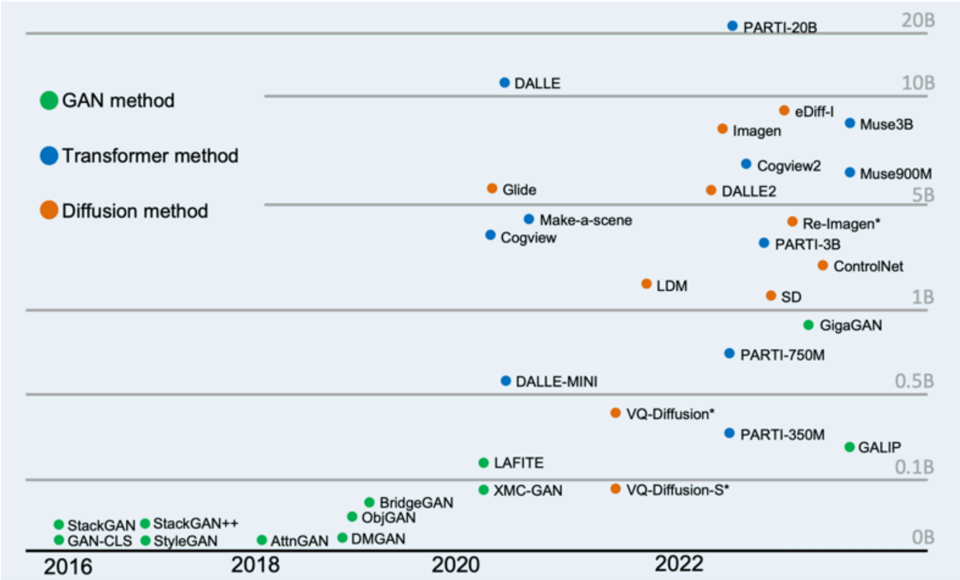


Figure 1.3: Progression of generative modeling techniques from 2016 to 2022, showcasing the evolution and scaling of model parameters in GAN, Transformer, and Diffusion methods, each color-coded by method type.[40]

In the field of image generation, models like MidJourney, DALL-E 3, and Stable Diffusion 3 have been at the forefront. Video generation has also seen substantial growth, with platforms such as Runway Gen-2, Pika, and Sora pushing the boundaries of what can be achieved in AI-generated cinematography. In the audio domain, AI's capability to replicate and generate human-like speech has made significant strides with models such as ElevenLabs, Speechify, and Whisper. These tools have not only democratized voice-over and audio content production but have also raised the bar for what can be expected from synthetic speech in terms of clarity, emotion, and expressiveness. The geographical landscape of AI development is also noteworthy, with the United States and China emerging as the primary incubators of AI innovation. Leading internet companies in these countries, including tech giants like Google, Meta, Alibaba, and ByteDance, have been instrumental in driving forward the development of AI technologies. Their investments have resulted in state-of-the-art models that, while most of them are not open-source.

1.4 Case Study: Using AI tools to predict the future of soccer

In an effort to visualize the potential of AI in the content creating, I create a project to design a one-minute video showcasing what the soccer game might look like in the future through the lens of AI. This journey was not just about prediction but also indicate how AI can revolutionize content creation and sports entertainment. This endeavor was made possible by several AI models, including ChatGPT for scripting, MidJourney for visual images generation, Runway for video editing and synthesis, and Topaz AI for enhancing video quality. This project not only demonstrates the capabilities of these AI tools but also offers a glimpse into the future of entertainment. The video has been shared on social platforms like YouTube, inviting viewers to engage with and reflect on the evolving intersection of technology and sports. This case study emphasized that user should treat AI as a colleague during the process, giving the clear and detailed prompts to ensure the result is close to the expectations. Even though AI tools can inspire the user during the brainstorming, but human still need to have a basic framework before starting the project. In the end, this project highlights how human creativity combined with AI's efficiency can unlock the opportunities for innovation and engagement in productivity.

Chapter 2

METHODS

2.1 Workflow

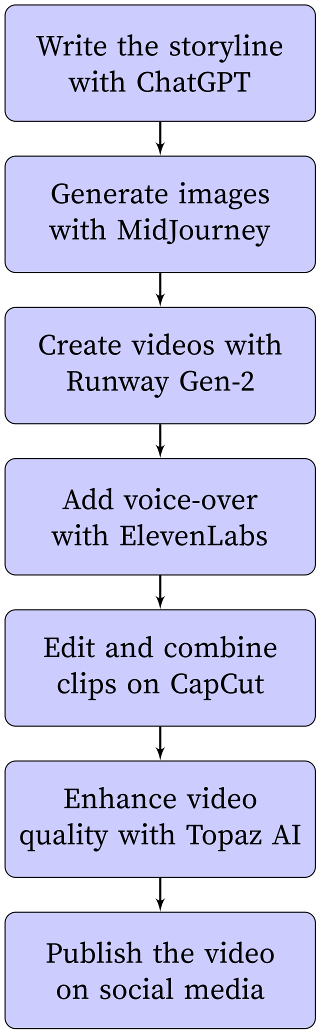


Figure 2.1: Workflow of the video-creating project

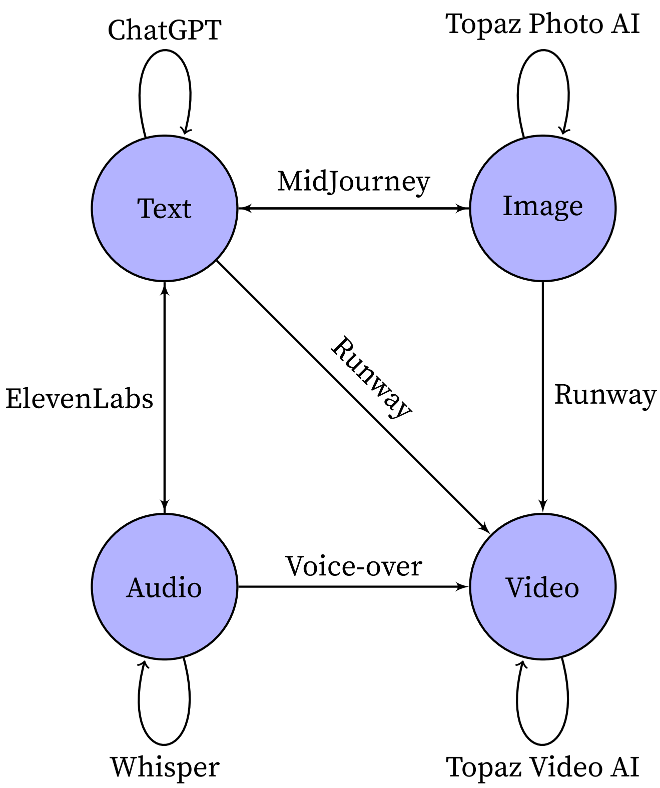


Figure 2.2: Relationships in a Multimodal System

2.2 AIGC in Text (ChatGPT)

2.2.1 Reinforcement Learning from Human Feedback (RLHF)

Reinforcement Learning from Human Feedback (RLHF) [41] can align the model's outputs with human preferences.

The objective of RLHF is to optimize the policy π to maximize the expected reward, which is indicative of the alignment between generated outputs and human preferences:

$$\max_{\pi} \mathbb{E}_{\tau \sim \pi}[R(\tau)], \tag{2.1}$$

where $\tau = (s_0, a_0, ..., s_{T-1}, a_{T-1}, s_T)$ represents a trajectory of states and actions, and $R(\tau)$ is the cumulative reward for trajectory τ.

Human feedback is used to train a reward model $R_\theta(s, a)$, parameterized by θ, which estimates the reward of taking action a in state s:

$$R_\theta(s, a) = \text{HumanFeedback}(s, a). \tag{2.2}$$

The policy π_ϕ, parameterized by ϕ, is optimized using the reward model as a surrogate for the true human feedback:

$$\phi^* = \arg\max_{\phi} \mathbb{E}_{\tau\sim\pi_\phi}[R_\theta(s,a)]. \tag{2.3}$$

Typically, Proximal Policy Optimization (PPO) or a similar algorithm is employed for policy optimization, ensuring stable and efficient learning:

$$L(\phi) = \hat{\mathbb{E}}_t\left[\min(r_t(\phi)\hat{A}_t, \text{clip}(r_t(\phi), 1-\epsilon, 1+\epsilon)\hat{A}_t)\right], \tag{2.4}$$

where $r_t(\phi) = \frac{\pi_\phi(a_t|s_t)}{\pi_{\phi_{old}}(a_t|s_t)}$ is the probability ratio, $\hat{A}_t$ is the advantage estimate at time t, and ϵ is a hyperparameter that controls clipping to reduce variance.

All the processes above indicate how RLHF is mathematically conceptualized in the context of improving GPT models, focusing on optimizing the model to generate text that aligns with human preferences based on feedback.

2.2.2 Transfer Learning

Given a pretrained model M with parameters θ, transfer learning [42] adjusts θ to new parameters θ' to better perform on a target task T with a smaller dataset D_T. The objective is to minimize the loss function $\mathscr{L}_T$ on D_T, leveraging knowledge from the source task S.

The optimization problem can be formalized as:

$$\theta' = \arg\min_{\theta} \mathscr{L}_T(f_\theta(D_T)),$$

where f_θ denotes the model's prediction function parameterized by θ.

The loss function $\mathscr{L}_T$ often takes the form of cross-entropy in text-based tasks:

$$\mathscr{L}_T = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{ic}\log(p_{ic}),$$

where N is the number of samples in D_T, C is the number of classes, y_{ic} is a binary indicator of whether class c is the correct classification for observation i, and p_{ic} is the predicted probability of observation i being in class c.

For fine-tuning, we update θ with a learning rate α specific to the target task:

$$\theta' = \theta - \alpha\nabla_\theta\mathscr{L}_T(f_\theta(D_T)).$$

To prevent catastrophic forgetting, a regularization term $\mathscr{R}(\theta)$ may be added, balancing the new task learning and retention of previously learned knowledge:

$$\theta' = \arg\min_{\theta} \mathscr{L}_T(f_\theta(D_T)) + \lambda\mathscr{R}(\theta),$$

where λ controls the importance of the regularization term, and $\mathscr{R}(\theta)$ can be defined as the difference between the new parameters θ' and the original pretrained parameters θ, such as in elastic weight consolidation (EWC):

$$\mathscr{R}(\theta) = \sum_i \frac{1}{2}\Omega_i(\theta_i - \theta_i^*)^2,$$

with Ω_i representing the importance of parameter i for the tasks learned during pretraining and θ^* denoting the parameters after pretraining.

Thus, transfer learning efficiently leverages pre-existing knowledge of GPT-4, enabling it to adapt to new tasks with remarkable efficiency and minimal data, bridging the gap between general and specific task performances.

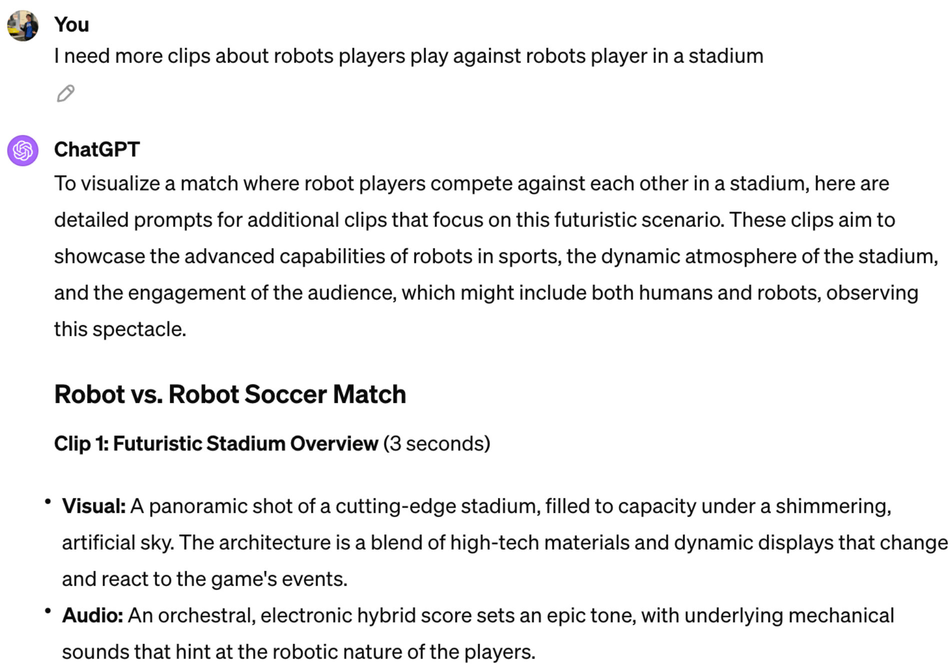


Figure 2.3: Using ChatGPT to generate vivid description prompt and prepare for creating the image stills.

ChatGPT utilizes Reinforcement Learning from Human Feedback (RLHF) and Transfer Learning to create a highly adaptive and user-aligned conversational model. Through RLHF, ChatGPT continuously refines its responses based on broad human feedback, optimizing its conversational results to align more closely with human preferences and expectations. This process ensures that the model's outputs are not only relevant but also resonate with users on a more personal level. Meanwhile, by employing Transfer Learning, ChatGPT can efficiently adapt to a wide range of conversational contexts and domains with relatively small amounts of task-specific data. This dual approach allows ChatGPT to maintain a broad understanding of language and context, gained from its initial training on diverse datasets, while also being able to specialize and improve in specific areas or topics as directed by user interactions and feedback. Thus, the integration of RLHF and Transfer Learning enables ChatGPT to offer conversational experiences that are both deeply informed and highly customized, making it a versatile tool for a variety of applications.

2.3 AIGC in Image and Video (MidJourney Runway Topaz)

2.3.1 Variational Autoencoders (VAE)

Variational Autoencoders (VAEs) are a subclass of autoencoders that are foundational for generative models. They are composed of an encoder and a decoder [43]. The encoder transforms input data into a latent space, whereas the decoder reconstructs the input data from this space. The latent space is typically modeled as a Gaussian distribution, with the encoder providing the parameters—mean μ and variance σ^2—for this distribution.

The encoder part of a VAE is responsible for compressing the input data into a latent representation. For a given data element x_i, the encoder produces a mean μ_i and a variance σ_i^2 that define a Gaussian distribution in the latent space. The encoding process can be represented by the following equation:

$$\log q_\phi \left(z^{(i)} | x^{(i)}\right) = \log \mathcal{N} \left(z^{(i)}; \mu^{(i)}, \sigma^{2(i)} I\right). \tag{2.5}$$

To allow for gradient descent methods to work, VAEs employ a reparameterization trick. This trick involves sampling a noise vector ϵ from a standard Gaussian distribution and then constructing the latent vector z_i by scaling the noise with the standard deviation and shifting by the mean:

$$z_i = \mu_i + \epsilon \times \sigma_i, \tag{2.6}$$

where $\epsilon \sim \mathcal{N}(0, I)$.

The decoder's goal is to reconstruct the input data from the latent representation. It tries to generate data that is as close as possible to the original input by maximizing the likelihood of the data given the latent space representation.

The training of VAEs involves maximizing the evidence lower bound (ELBO) on the marginal likelihood of the observed data. The ELBO can be represented as follows:

$$\mathcal{L}_b = -D_{KL}(q(z|x) \| p(z)) + \frac{1}{L} \sum_{i=1}^{L} (\log P(x|z)). \tag{2.7}$$

Furthermore, the ELBO can be decomposed into two terms: the first being the Kullback-Leibler divergence between the encoder's distribution and the prior distribution of the latent variables, and the second term being the expected log-likelihood of the data given the latent variables. For a latent space with dimensionality d, and assuming a standard Gaussian prior, the loss function simplifies to:

$$\mathcal{L}(\theta, x^i) = -\frac{1}{2} \sum_{i=1}^{d} \left(\mu_{(i)}^2 + \sigma_{(i)}^2 - \log \sigma_{(i)}^2 - 1\right) + \frac{1}{L} \sum_{i=1}^{L} \left(\log P(x^i|z)\right). \tag{2.8}$$

Despite their advantages for generative tasks, VAEs have limitations, particularly for image generation. Directly sampling from the Gaussian distribution often results in blurred images. Moreover, there is an information loss when projecting the data to a lower-dimensional latent space, which affects the quality of the reconstruction. However, VAEs find extensive applications in feature extraction and dimensionality reduction, and they can be particularly effective when combined with other generative models such as Generative Adversarial Networks (GANs) or diffusion models.

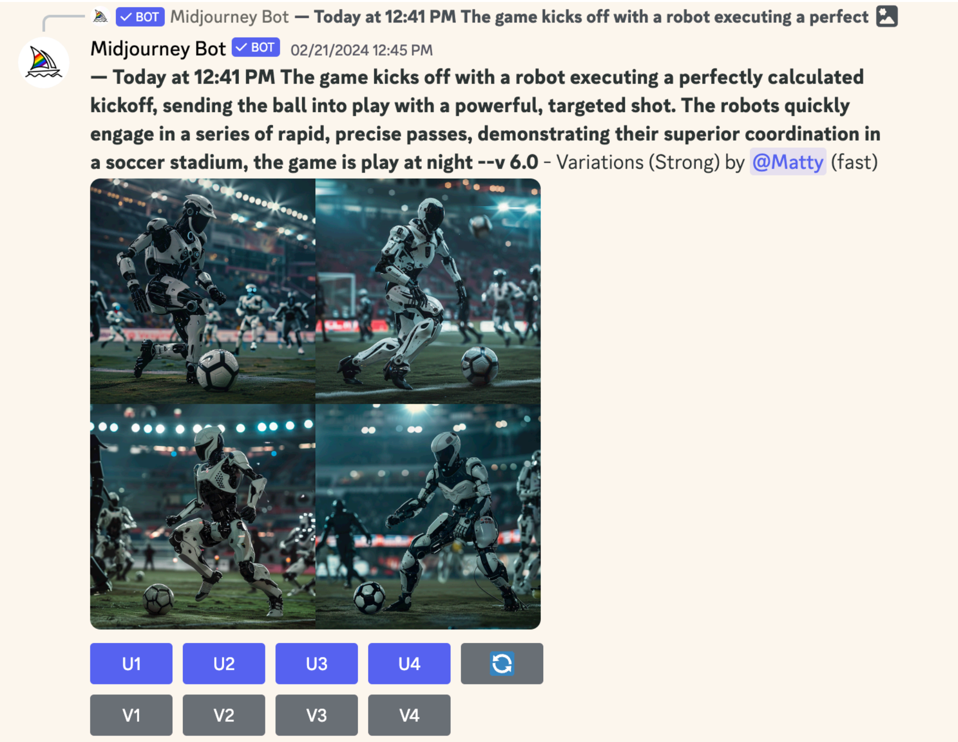


Figure 2.4: Using ChatGPT's prompts to generate good quality image stills using MidJourney, and then upload them as the key frames to Runway Gen-2 for video clip generation.

2.3.2 Generative Adversarial Networks (GAN)

Generative Adversarial Networks (GANs) represent a class of generative models that has significantly impacted the field of synthetic image generation. A GAN is composed of two distinct neural network models: a *generator* (G) and a *discriminator* (D). The generator attempts to produce data that is indistinguishable from genuine data, while the discriminator evaluates the authenticity of the data, distinguishing between actual and generated samples.

The objective function of a GAN is formulated as a min-max game between G and D:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))], \tag{2.9}$$

where p_{data} is the data distribution, x represents real data, z is a noise sample from distribution p_z, and $G(z)$ is the generated data.

During training, D is optimized to maximize the probability of correctly classifying a sample as real or fake. Simultaneously, G is optimized to minimize $\log(1 - D(G(z)))$. In practice, this is often achieved by alternating between the following gradient ascent step for D:

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[\log D(x^{(i)}) + \log(1 - D(G(z^{(i)}))) \right], \tag{2.10}$$

and the gradient descent step for G:

$$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log(1 - D(G(z^{(i)}))), \tag{2.11}$$

for a minibatch of m samples.

GANs have been utilized to generate high-fidelity images, outperforming other generative models in visual quality. However, they often struggle with mode collapse – a phenomenon where the generator produces a limited variety of outputs. Despite these challenges, GANs have shown impressive results in tasks beyond image generation, such as image super-resolution [44]and object detection [12].

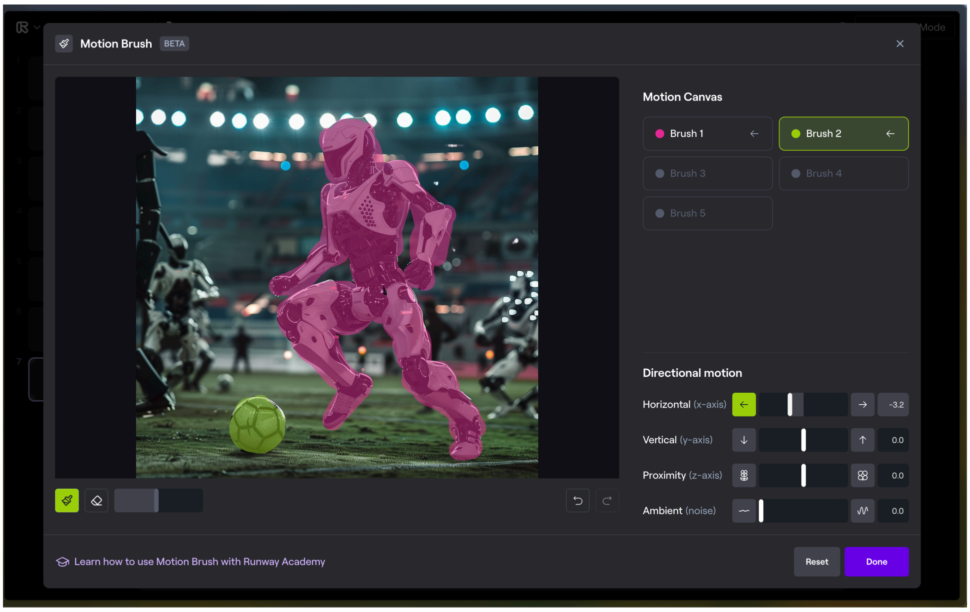


Figure 2.5: Using the "Image + Text" mode in Runway Gen-2 to generate the video, with the help of motion brush function to control how the characters move

2.3.3 Diffusion Models

Diffusion models [45, 46] are a class of generative models famous for their ability to generate high-quality images through a process that iteratively adds and removes noise. These models are characterized by their use of a Markov chain to gradually convert simple noise distributions into complex data distributions.

A diffusion model typically consists of two main components: the forward process (described in Algorithm 1) and the reverse process (described in Algorithm 2). The forward process models the gradual addition of noise to the data, while the reverse process models the generation of data from noise. Mathematically, the forward process can be described by:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \tag{2.12}$$

where β_t is a variance schedule over discrete timesteps $t = 1, 2, ..., T$.

Algorithm 1 Diffusion model training [40]

1: for every training iteration do
2: Sample t from discrete timestep $1, 2, \ldots, T$, or from continuous timestep $t \sim [0, 1]$.
3: Sample random noise $\epsilon \sim \mathcal{N}(0, 1)$.
4: Calculate x_t based on DDPM forward or SDE forward.
5: Update the model with noise prediction $\epsilon(x_t, t)$ or score function $s(x_t, t)$.
6: end for

The reverse process involves learning a model to reverse this noise addition:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)), \tag{2.13}$$

where $\mu_\theta(\mathbf{x}_t, t)$ and $\Sigma_\theta(\mathbf{x}_t, t)$ are functions parameterized by a neural network.

Algorithm 2 Diffusion model Inference [40]

1: Sample x_t from normal Gaussian distribution $x_t \sim \mathcal{N}(0, \mathbf{I})$.
2: Sample discrete timesteps from $1, 2, \ldots, T$ or continuous timestep from $[0, 1]$.
3: for t in Reverse(timesteps) do
4: Calculate the noise distribution $\epsilon(x_t, t)$ or score function $s(x_t, t)$ with the corresponding diffusion model.
5: Approximate x_{t-1} or $x_{t-\Delta t}$ based on the reverse function.
6: end for

The training of diffusion models, as outlined in Algorithm 1, involves adjusting the parameters of the neural network to maximize the likelihood of the data under the reverse process. Similarly, the inference procedure, as outlined in Algorithm 2, utilizes the trained model to generate new data samples by reversing the noise process.

Diffusion models have been applied to a variety of tasks beyond image generation, such as super-resolution and inpainting, demonstrating their versatility and potential in numerous domains of generative modeling.

MidJourney, Runway, and Topaz, while not explicitly disclosing which models they employ, potentially make use of advanced generative models like Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and Diffusion Models to offer cutting-edge image and video generation, enhancement, and editing capabilities. VAEs could be used to understand and encode the vast complexity of visual data into a manageable latent space, allowing these platforms to efficiently generate or modify content with a deep understanding of visual structures. GANs, known for their ability to produce high-quality synthetic images, might be instrumental in enabling these tools to create realistic images and videos that are difficult to distinguish from real ones, enhancing the authenticity and visual appeal of generated content.

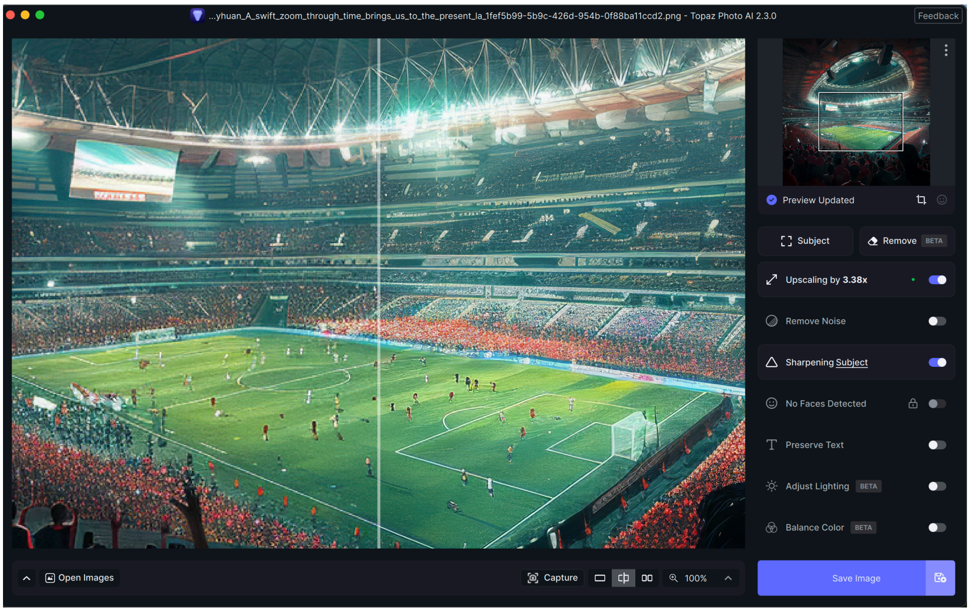


Figure 2.6: Using Topaz AI to upscale and improve the video quality, like changing the frame rate from 60 FPS to 120 FPS, removing noise, and sharpen the subject.

Diffusion Models can be implemented by generating high-fidelity images through iterative refinement and producing coherent visual outputs, particularly useful in tasks like image super-resolution, denoising, and text-to-image (TTI) synthesis. The theoretical integration of these models would enable MidJourney, Runway, and Topaz to harness the strengths of each, offering users a versatile tools for creating, enhancing digital media in ways that were previously unimaginable, thus pushing the boundaries of creative and technical possibilities in digital content creation.

2.4 AIGC in Sound (ElevenLabs)

2.4.1 Recurrent Neural Networks (RNN)

Recurrent Neural Networks (RNNs) [47–49] have shown remarkable capabilities in modeling sequential data, making them particularly suited for applications in sound synthesis. Unlike traditional feedforward neural networks, RNNs incorporate a feedback loop, allowing them to maintain a form of memory over input sequences. This characteristic is leveraged in sound synthesis to model the temporal dependencies of audio signals, enabling the generation of coherent and dynamic sound sequences.

The fundamental equation governing the operation of a basic RNN unit for sound synthesis can be described as follows:

$$h_t = \sigma(W_{ih}x_t + W_{hh}h_{t-1} + b_h), \tag{2.14}$$

where h_t represents the hidden state at time t, x_t is the input vector at time t, W_{ih} and W_{hh} are the weights of the input-to-hidden and hidden-to-hidden connections, respectively, b_h is the bias, and σ denotes the activation function, often a non-linear function such as the tanh or ReLU.

The output of the RNN, which corresponds to the synthesized sound at each timestep, is computed as:

$$y_t = W_{ho}h_t + b_o, \tag{2.15}$$

where y_t is the output vector at time t, W_{ho} represents the weights of the hidden-to-output connections, and b_o is the output bias.

To enhance the model's capacity to handle long-term dependencies, which are prevalent in complex sound sequences, Long Short-Term Memory (LSTM) [50] units or Gated Recurrent Units (GRUs) [51] can be incorporated. These advanced variants of RNNs introduce mechanisms such as forget gates (in LSTMs) and update gates (in GRUs) that allow the network to selectively remember or forget information. This capability significantly improves the network's ability to model sequences with long-range temporal dependencies, making it highly effective for tasks like sound synthesis where the coherence over time is crucial.

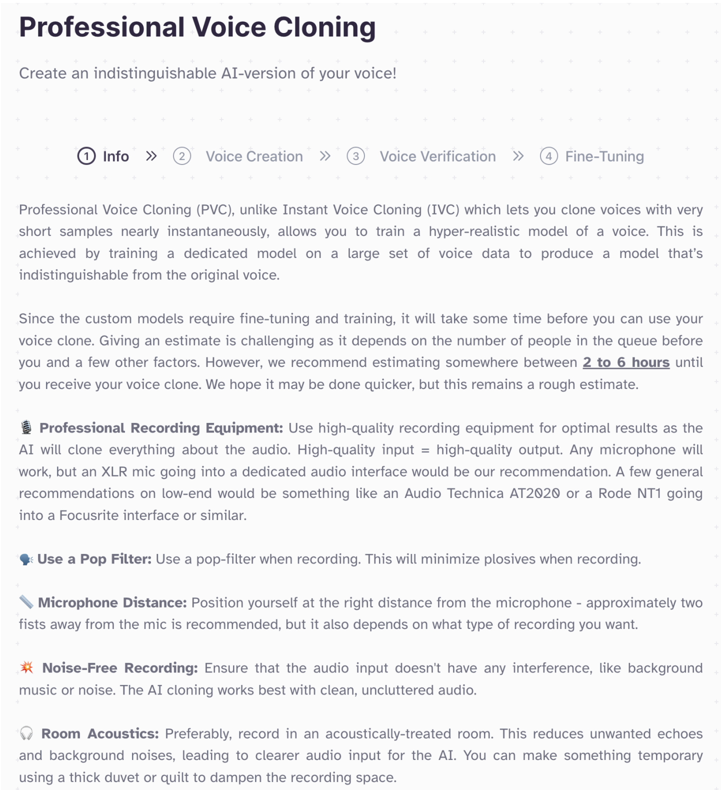
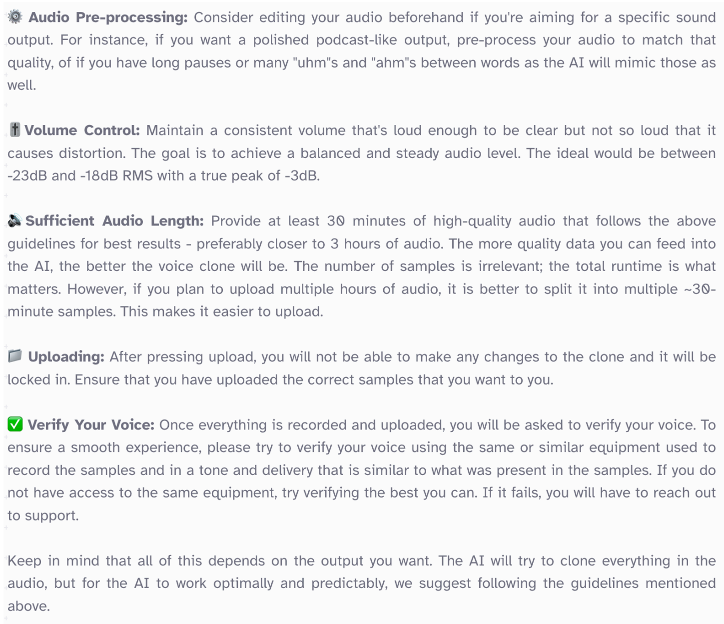


Figure 2.7: Using Professinal Voice Cloning from ElevenLabs to apply text to speech and add the voice over to the video

2.4.2 Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) units are a specific type of Recurrent Neural Network (RNN) architecture designed to overcome the limitations of traditional RNNs, particularly in handling long-term dependencies. LSTMs are adept at modeling the temporal relationships inherent in complex sound sequences, making them highly effective for sound synthesis tasks. The LSTM's architecture includes several key components—forget gate, input gate, cell state, and output gate—that work together to regulate the flow of information through the unit. These


Audio Pre-processing: Consider editing your audio beforehand if you're aiming for a specific sound output. For instance, if you want a polished podcast-like output, pre-process your audio to match that quality, of if you have long pauses or many "uhm"s and "ahm"s between words as the AI will mimic those as well.

Volume Control: Maintain a consistent volume that's loud enough to be clear but not so loud that it causes distortion. The goal is to achieve a balanced and steady audio level. The ideal would be between -23dB and -18dB RMS with a true peak of -3dB.

Sufficient Audio Length: Provide at least 30 minutes of high-quality audio that follows the above guidelines for best results - preferably closer to 3 hours of audio. The more quality data you can feed into the AI, the better the voice clone will be. The number of samples is irrelevant; the total runtime is what matters. However, if you plan to upload multiple hours of audio, it is better to split it into multiple ~30-minute samples. This makes it easier to upload.

Uploading: After pressing upload, you will not be able to make any changes to the clone and it will be locked in. Ensure that you have uploaded the correct samples that you want to you.

✅ **Verify Your Voice:** Once everything is recorded and uploaded, you will be asked to verify your voice. To ensure a smooth experience, please try to verify your voice using the same or similar equipment used to record the samples and in a tone and delivery that is similar to what was present in the samples. If you do not have access to the same equipment, try verifying the best you can. If it fails, you will have to reach out to support.

Keep in mind that all of this depends on the output you want. The AI will try to clone everything in the audio, but for the AI to work optimally and predictably, we suggest following the guidelines mentioned above.

Figure 2.8: Using Professinal Voice Cloning from ElevenLabs to apply text to speech and add the voice over to the video

components allow the LSTM to retain or discard information over long sequences, enabling the generation of coherent and dynamic audio content. The following equations provide a detailed mathematical representation of the LSTM's operation:

Forget Gate:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f). \tag{2.16}$$

It decides which information from the cell state is to be discarded. f_t represents the forget gate's activation at time t. σ is the sigmoid function, W_f is the weight matrix for the forget gate, h_{t-1} is the previous hidden state, x_t is the current input, and b_f is the bias.

Input Gate:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i), \tag{2.17}$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C). \tag{2.18}$$

It controls the extent to which new information is stored in the cell state. i_t is the input gate's activation, and $\tilde{C}_t$ is the candidate value for addition to the cell state. W_i and W_C are the weight matrices, and b_i and b_C are biases.

Cell State Update:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t. \tag{2.19}$$

It is updated by removing information deemed unnecessary by the forget gate and adding new candidate values scaled by the input gate's activation.

Output Gate:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o), \tag{2.20}$$

$$h_t = o_t * \tanh(C_t). \tag{2.21}$$

It determines the next hidden state h_t, which contains information about the current input and the previous state, to be passed to the network. The output gate's activation o_t scales the tanh of the updated cell state, deciding the part of the cell state to output.

In sound synthesis applications, LSTM models are trained on large datasets of audio samples, allowing them to learn and generate new sound sequences that closely mimic the characteristics of the training data. The ability to model long-term dependencies makes LSTMs particularly suited for generating complex and temporally coherent sound sequences, such as musical compositions or natural speech patterns. By optimizing the LSTM parameters (weights and biases) through backpropagation, the model can be fine-tuned to produce high-quality audio outputs that capture the nuances of human-generated sounds.

2.4.3 Autoregressive Transformers

Autoregressive Transformers [52, 53] have improved sound synthesis by leveraging self-attention mechanisms to model complex sequential data. Unlike RNNs, Transformers process sequences in parallel, allowing for more efficient training and better handling of long-term dependencies. In sound synthesis, autoregressive Transformers predict subsequent audio samples based on a sequence of past samples, capturing the temporal dynamics of sound. The core concept of the Transformer architecture in sound synthesis can be described by the following equations about the self-attention mechanism and the generation process:

Input Embedding and Positional Encoding:

$$X' = X + PE, \tag{2.22}$$

where X is the input sequence of sound samples or features, and PE is the positional encoding added to X to retain positional information.

Self-Attention Mechanism:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{2.23}$$

where Q, K, and V represent the queries, keys, and values matrices obtained from X', respectively, and d_k is the dimension of the key vectors. This equation calculates the attention weights and applies them to the values to produce an output that highlights important features of the input sequence.

Multi-Head Attention:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O, \tag{2.24}$$

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V). \tag{2.25}$$

Multi-head attention allows the model to focus on different positions, capturing various aspects of the sound sequence. W_i^Q, W_i^K, W_i^V, and W^O are parameter matrices for each head i and the output projection, respectively.

Position-wise Feed-Forward Networks:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2. \tag{2.26}$$

Each layer in the Transformer includes a feed-forward network applied to each position separately and identically. This consists of two linear transformations with a ReLU activation in between.

Output Generation:

$$P(y_t|y_{<t}) = \text{softmax}(y_{t-1}W + b). \tag{2.27}$$

The output at each timestep t, y_t, is predicted based on the previous outputs $y_{<t}$, where W and b are the weights and bias in the final linear layer that projects the decoder output to the space of possible audio samples. $P(y_t|y_{<t})$ represents the probability distribution over possible next samples given the previous samples.

Training and Fine-tuning:

During pre-training, the model learns to predict the next word in a sentence given the previous words, using a masked language modeling objective:

$$L_{\text{MLM}} = -\sum_{i \in M} \log P(w_i | w_{\setminus i}; \Theta), \tag{2.28}$$

where M denotes the set of masked positions, w_i is the word at position i, $w_{\setminus i}$ represents the sequence without the word at position i, and Θ are the model parameters.

Fine-tuning adjusts the pre-trained model to specific tasks or datasets. This phase enhances the model's performance on tasks such as conversational responses, sentiment analysis, and text summarization:

$$L_{\text{task}} = \sum_{(x,y) \in D_{\text{task}}} \log P(y|x; \Theta'), \tag{2.29}$$

where D_{task} is the task-specific dataset, (x, y) are the input-output pairs, and Θ' are the fine-tuned model parameters.

By training on sequences of audio samples, autoregressive Transformers learn to generate new sound sequences sample by sample, offering a powerful framework for high-quality sound synthesis, including music composition and speech generation, with the ability to capture the long-range dependencies characteristic of audio signals.

2.4.4 Voice Cloning & Text-to-Speech (TTS)

The applications of AI-driven voice synthesis, particularly through voice cloning and text-to-speech (TTS) technologies like ElevenLabs [29] and Whisper [54], has apparently closed the gap between synthetic and human speech. These systems utilize deep learning frameworks to analyze and replicate the nuances of human speech, including tonality and emotion. Voice cloning involves creating a digital replica of a target voice from a limited set of audio samples. The process can be described by the following stages:

Feature Extraction:

$$F = \mathrm{MFCC}(S), \tag{2.30}$$

where F represents the feature matrix extracted from the input speech signal S, and MFCC denotes the Mel-Frequency Cepstral Coefficients, a common feature used in voice synthesis to capture the timbral aspects of the speech.

Acoustic Modeling:

$$H = \mathrm{Encoder}(F; \theta_e), \tag{2.31}$$

$$\tilde{F} = \mathrm{Decoder}(H; \theta_d). \tag{2.32}$$

where H is the encoded representation of the speech features, θ_e and θ_d are the parameters of the encoder and decoder networks, respectively, and $\tilde{F}$ is the reconstructed feature matrix. The encoder-decoder framework is often implemented using deep neural networks, where the encoder learns a compressed representation of the speech features, and the decoder reconstructs the features, potentially in the target voice's style.

Voice Conversion:

$$V_t = \mathrm{Conversion}(H_t; \theta_c), \tag{2.33}$$

where V_t represents the target voice features, H_t is the encoded representation of the target speech, and θ_c are the parameters of the conversion model that transforms the source voice into the target voice.

Waveform Generation:

$$\hat{S} = \mathrm{WaveNet}(V_t; \theta_w), \tag{2.34}$$

where $\hat{S}$ is the synthesized speech waveform, and θ_w are the parameters of a WaveNet [55] model trained to convert the feature representation V_t back into the time-domain signal. WaveNet, a deep generative model of raw audio waveforms, is particularly effective in generating high-fidelity speech with natural intonations and expressions.

The process of voice cloning and modification enables the generation of synthetic speech that closely resembles a target human voice. By capturing the subtle nuances of speech, AI-driven voice synthesis technologies have opened new avenues for creating engaging and emotionally resonant digital communication experiences.

2.4.5 Audio Modification

There are several methods to tackle audio modification challenges such as noise reduction, sound separation, and audio restoration [56]. The processes are shown below:

Noise reduction aims to eliminate unwanted background noise from audio recordings, enhanc-

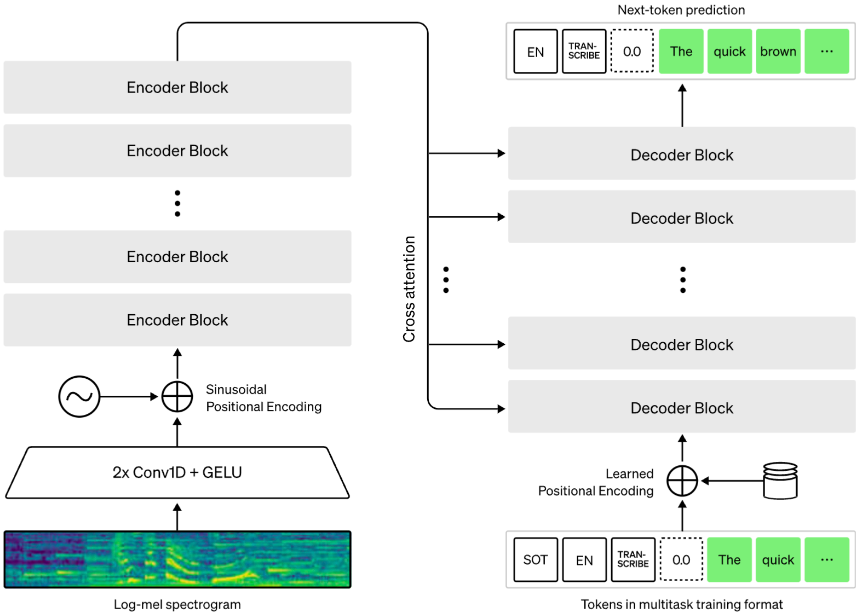


Figure 2.9: Autoregressive transformer model architecture used for audio processing tasks like transcription and voice synthesis. The model utilizes a stack of encoder and decoder blocks, with sinusoidal positional encodings and learned positional encodings for sequence representation. [54]

ing the clarity of the sound. This can be mathematically represented by:

$$S_{clean} = S_{noisy} - N, \tag{2.35}$$

where S_{clean} is the clean audio signal, S_{noisy} is the original noisy signal, and N represents the noise component. Advanced AI models, such as Deep Neural Networks (DNNs), are trained to estimate N accurately from S_{noisy}, allowing for effective noise removal.

Sound separation involves isolating individual sound sources from a mixed audio signal. This can be expressed as:

$$S_i = F(S_{mix}; \theta_i), \tag{2.36}$$

where S_i is the isolated sound of source i, S_{mix} is the mixed audio signal, and F is a function modeled by the AI system with parameters θ_i tailored to extract the i^{th} sound source.

Audio restoration focuses on recovering the original quality of degraded audio recordings. The process can be conceptualized as:

$$S_{restored} = R(S_{degraded}; \theta_r), \tag{2.37}$$

where $S_{restored}$ is the restored audio signal, $S_{degraded}$ is the degraded audio signal, and R represents the restoration function implemented by the AI with parameters θ_r. This function aims to reconstruct lost or corrupted signal components, effectively restoring the audio to its original state.

The final step often involves enhancing the waveform directly to improve the overall audio quality, which can be mathematically described by:

$$\hat{S} = G(S_{processed}; \theta_g), \tag{2.38}$$

where $\hat{S}$ is the enhanced audio waveform, $S_{processed}$ is the audio signal after noise reduction, sound separation, and restoration, and G is the enhancement function driven by AI with parameters θ_g. Models like WaveNet are examples of generative networks that can be used for this purpose, fine-tuning the audio quality by adjusting the waveform directly.

These core processes involved in audio modification, demonstrating how AI algorithms can significantly improve the clarity, fidelity, and overall quality of sound recordings.

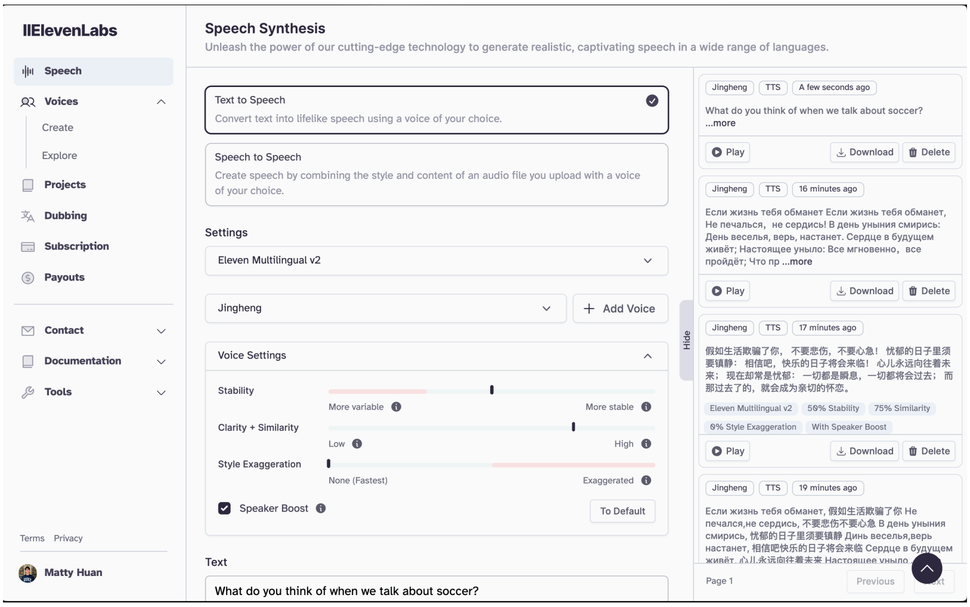


Figure 2.10: Using my clone voice to generate the voice over for the video, I recorded the training set by reading one chapter from the book *Elon Musk*, the audio file is 14 mins.

ElevenLabs, positioned at the forefront of AI-driven audio synthesis, particularly in voice cloning and text-to-speech (TTS) technologies, likely utilizes Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) units, and Autoregressive Transformers. These models excel in understanding and generating the temporal sequences crucial in replicating human speech's nuanced dynamics. While ElevenLabs has not publicly disclosed the specific models underpinning its technology, the theoretical basis of these models provides a clue to the sophisticated mechanism that enables ElevenLabs to produce highly realistic and expressive voice outputs. RNNs and LSTMs could be instrumental in modeling the sequential nature of speech, capturing the temporal dependencies necessary for coherent speech generation. Autoregressive Transformers, with their powerful self-attention mechanisms, offer a means to process long sequences in parallel, potentially enhancing the system's ability to manage the structure of speech at scale. Together, these models could form the basis of ElevenLabs' ability to analyze and replicate the unique attributes of individual voices, offering voice synthesis capabilities that find applications in entertainment, education, and beyond.

Chapter 3

RESULTS

3.1 Overview of Project Outcome

This video project is a mix of the latest AIGC tools and creative ideas, presents a narrative that is both visually stunning and thematically profound. It tells a story in one minute with series of crafted scenes, each contributing to a cohesive storyline that explores the symbiotic relationship between humanity and AI-driven robots. It is set in a futuristic landscape, the narrative delves into themes of the potential futures shaped by the integration of advanced AI entities within human society. This thematic exploration is brought to life through a variety of scenes, ranging from peaceful, digitally-rendered natural view to bustling, urban technology alive with the vibrancy of cyber-punk life. The selection and progression of scenes are carefully arranged to not only captivate the audience with visual magnificence but also to provoke thought and reflection on the evolving role of AI in our world.

The project's success in delivering a narrative is largely attributable to the use of AI technologies at each stage of the creative process. From the generation of high-fidelity image stills with MidJourney to the video synthesis facilitated by Runway Gen-2, the tools represent the potential of AI to the traditional boundaries of content creation. The narrative is further enriched with a voiceover, rendered through ElevenLabs' voice cloning technology, adding a layer of narrative depth that guides the audience through the unfolding story. This integration of AI-driven voice synthesis not only enhances the immersive quality of the video but also demonstrates the potential of AI in crafting engaging and emotionally resonant storytelling. The video's aesthetic and thematic achievements are underscored by Topaz AI, which elevates the visual quality through enhancements such as frame rate optimization and noise reduction, ensuring that the final product has a good overall quality.

This project, through its combination of AI capabilities and creative narrative design, opens a dialogue on the future of storytelling in the age of artificial intelligence. It stands as an example of how AI support to create content that is not just visually captivating, but also rich in thematic depth, challenging us to envision the limitless possibilities of creative expression in the digital era.

3.2 Deliverable

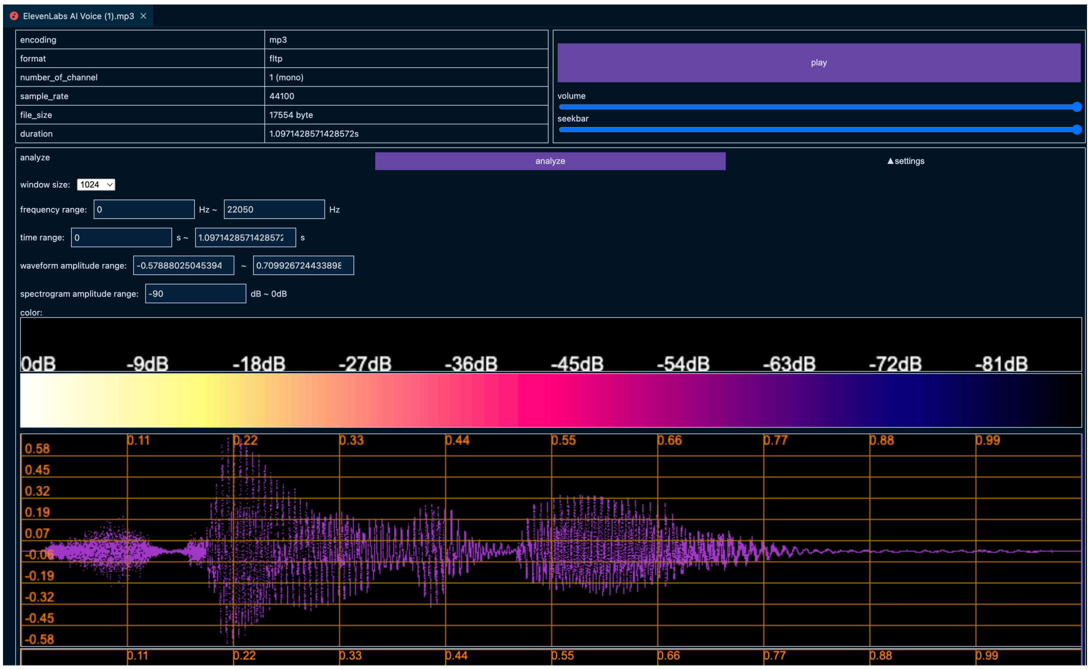


Figure 3.1: An in-depth analysis of an MP3 file showcasing its spectral dynamics, waveform, and amplitude in a colorful sound engineering software interface.

Figure 3.2: A pitch-black screen gradually reveals a soccer ball radiating a soft, ethereal glow. The light pulses gently, casting shadows that mimic fans gathered in a silent stadium, anticipation hanging in the air.

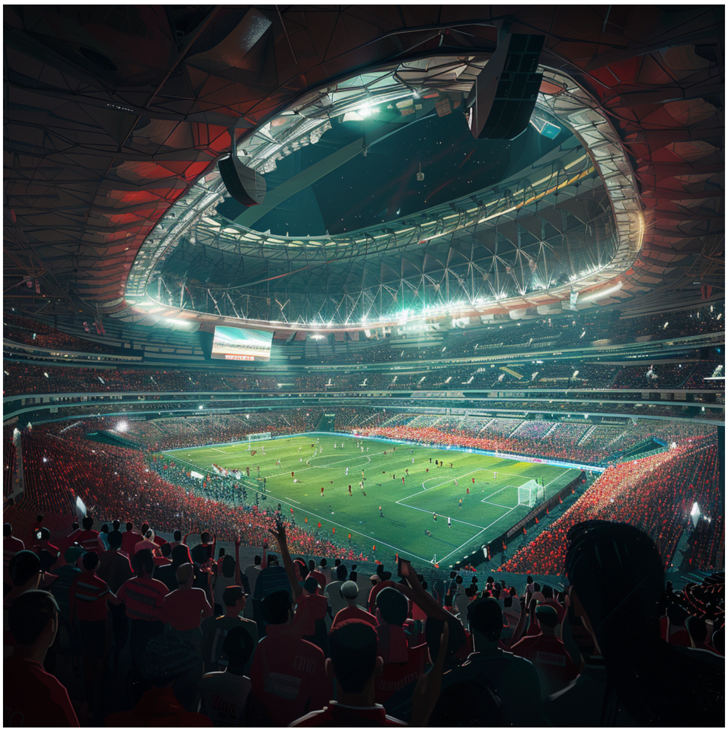


Figure 3.3: A swift zoom through time brings us to the present, landing on the electric atmosphere of the 2026 World Cup stadium. The architecture is a marvel of modern design, with sweeping curves and glowing lights, packed to the brim with a sea of fans from across the globe.

Figure 3.4: Close-up shots dive into the crowd's heart, capturing faces painted in national colors, eyes sparkling with excitement. Fans in synchronized motion, creating a living tapestry of support that breathes life into the stadium's structure.

Figure 3.5: Transition to the crystal-clear, high-definition 2010s, where the speed and emotion of the game captivate. A player's face is frozen in a moment of triumph, surrounded by a blur of teammates and confetti, encapsulating modern soccer's glory.

Figure 3.6: A midfielder weaves through the opposition with balletic grace, the ball seemingly tethered to their feet. Each touch is a testament to years of dedication, the culmination of a lifetime's passion for the game.

Figure 3.7: The climactic moment as the ball arcs beautifully into the net, the scorer's joy uncontainable as they sprint towards the roaring fans, arms spread wide, embodying the euphoria of triumph. Teammates converge in a jubilant huddle, a symbol of unity and success.

Figure 3.8: The stadium, an architectural marvel, seamlessly blends organic designs with advanced technology. Bioluminescent structures illuminate the arena, while floating screens display live stats and player information. The atmosphere is electric, buzzing with anticipation for the historic final.

Figure 3.9: Inside the stadium, two teams of robots engage in a pre-game warm-up, showcasing incredible precision and agility. Their movements are fluid yet unmistakably mechanical, with LED lights tracing patterns across their metallic frames, each robot adorned with the colors of their respective teams.

Figure 3.10: Focus on a uniquely designed soccer ball at the center circle, glowing with internal lights and equipped with smart technology that allows it to record and respond to the game's flow. A player approaches and gives it a gentle tap, setting it aglow, ready for kickoff.

Figure 3.11: The kickoff is a spectacle of technology, with a robotic player skillfully propelling the ball into play using a swift, calculated motion. The ball itself is equipped with smart tech, glowing softly as it moves across the pitch, data streaming off it in visual waves that are visible to the audience.

Figure 3.12: Human players are shown warming up on the field, executing stretches and drills with the backdrop of a stadium that blends organic and technological aesthetics. The turf is a perfect green, with embedded lighting that responds to the players' movements, illuminating their paths across the pitch.

Figure 3.13: The climax of the segment is a goal that feels both alien and familiar. A robot striker launches the ball with pinpoint accuracy from outside the penalty box, the ball tracing a perfect arc into the top corner of the goal. The net lights up, and digital fireworks explode overhead, celebrating the score in a dazzling display of light and color.

Chapter 4

DISCUSSION

4.1 Foundational Models

The journey into the realms of Artificial Intelligence (AI) commenced with the development of basic machine learning algorithms and the advent of neural networks, marking a seminal phase in the evolution of computational intelligence. These initial models established a fundamental framework for the analysis and interpretation of data patterns, heralding the dawn of computational learning. By enabling machines to learn from and make predictions based on data, these foundational technologies set the stage for a radical transformation in the way that information is processed and utilized. Neural networks, in particular, with their ability to mimic the neural structures of the human brain, introduced a novel paradigm for machine learning, facilitating the development of systems capable of performing complex tasks such as image recognition, natural language processing, and decision making with unprecedented accuracy. This era of AI research and development, characterized by the exploration and refinement of machine learning algorithms and neural network architectures, laid the crucial groundwork for subsequent advancements in AI, propelling the field towards the sophisticated and versatile AI systems we see today [3]. The significance of these early models cannot be overstated, as they not only provided the initial impetus for AI research but also continue to underpin the ongoing exploration and expansion of AI capabilities across various domains.

4.2 Specialized Architectures

The introduction of specialized architectures, notably Generative Adversarial Networks (GANs) and Diffusion Models, represents a watershed moment in the advancement of Artificial Intelligence (AI), heralding a new era of sophistication in AI capabilities. GANs, conceptualized as a system of two neural networks contesting with each other in a generative adversarial process, have fundamentally transformed the landscape of synthetic media creation. This architecture's unique capability to produce high-fidelity images and videos has significantly elevated the standard of realism and detail attainable in digital content, marking a departure from previous limitations. Similarly, Diffusion Models, which iteratively refine random noise into structured images through a reverse Markov process, have further extended the possibilities for generating intricate and lifelike synthetic visuals. These developments have not only expanded the horizons of what is achievable with AI in the realm of digital media but have also laid the foundation for novel applications across diverse fields such as entertainment, healthcare, and education. The impact of these specialized architectures on the AI domain has been profound, catalyzing a shift towards the creation of more realistic, dynamic, and nuanced digital experiences, and setting new benchmarks for the quality of AI--generated content [21]. The advent of GANs and Diffusion Models thus marks a significant milestone in the evolution of AI, underscoring the technology's growing ability to mimic, and in some cases surpass, the intricacies of human perception and creativity.

4.3 Ethical and Societal Considerations

As Artificial Intelligence (AI) technologies have evolved and matured, there has been a discernible shift towards addressing the broader ethical and societal implications associated with AI-generated content. This phase of AI development is characterized by a heightened awareness of the potential for biases embedded within AI systems, leading to concerted efforts aimed at identifying, understanding, and mitigating these biases to ensure fairness and equity in AI applications. Additionally, the capacity of AI to generate highly realistic synthetic media, such as deepfakes, has raised significant concerns about the potential for misuse in spreading misinformation and manipulating public perception. In response, researchers and developers are increasingly prioritizing the development of AI technologies that incorporate ethical considerations from the outset, emphasizing the creation of systems that are not only technologically advanced but also socially responsible. This involves the implementation of robust mechanisms for the detection and prevention of AI-generated content that could be used for malicious purposes, alongside initiatives to promote transparency and accountability in AI development processes. The focus on ethical and societal considerations marks a critical evolution in the field of AI, reflecting a growing recognition of the technology's profound impact on society and the imperative to guide its development in a direction that maximizes benefits while minimizing harms. This shift towards ethically conscious AI development is fundamental to ensuring that the advancements in AI technology contribute positively to society, fostering an environment where innovation is balanced with responsibility [57].

4.4 Efficiency and Scalability

The current focus within the field of Artificial Intelligence (AI) on enhancing efficiency and scalability underscores a pivotal shift towards developing algorithms that are both robust and adaptable to the exigencies of real-time and large-scale applications. This trend is driven by the increasing demand for AI technologies that can process and analyze vast amounts of data swiftly and accurately, facilitating their integration into various aspects of digital media production, from content creation to distribution. Innovations aimed at optimizing processing speeds are critical in this context, as they enable AI systems to perform complex computations more rapidly, thereby reducing latency and improving the user experience. Similarly, efforts to augment the scalability of AI systems are essential for accommodating the exponential growth in data volumes and the complexity of tasks that AI is expected to handle. These advancements are not merely technical achievements but also reflect a broader imperative to ensure that AI technologies remain relevant and effective in a rapidly evolving digital landscape. By prioritizing efficiency and scalability, researchers and developers are working to create AI systems that are not only theoretically advanced but also practically applicable, capable of supporting the dynamic and diverse needs of modern digital media production [58]. This alignment of AI development with the practical demands of real-world applications is crucial for the continued integration and expansion of AI across different sectors, promising to unlock new possibilities for innovation and creativity in the digital age.

4.5 Extension to Sound Generation

In tandem with the significant strides made in image and video synthesis, the exploration of Artificial Intelligence (AI) within the realm of sound generation has marked a pivotal expansion of AI's capabilities, catalyzing a transformative shift in the creation and manipulation of audio content. Advancements in AI-driven music composition, voice synthesis, and sound effects generation are at the forefront of this evolution, employing sophisticated algorithms to produce audio that is not only more natural and lifelike but also dynamic and sensitive to context. These developments in sound generation are predicated on deep learning models that analyze and learn from vast datasets of existing audio, enabling the generation of sound that can adapt to varying narrative and environmental cues, thereby enriching the auditory experience across a wide range of applications, from interactive media to immersive virtual environments. The integration of AI in sound generation thus represents a significant milestone, highlighting the technology's capacity to extend its influence beyond the visual domain and into the broader spectrum of multimedia content creation. This holistic approach to AI application in multimedia not only enhances the realism and engagement of digital experiences but also underscores the technology's role in shaping the future of content creation, where AI's contribution is envisaged to permeate all aspects of digital media, offering unprecedented opportunities for innovation and creativity.

4.6 Future Directions

Looking ahead, the trajectory of Artificial Intelligence (AI) development is increasingly oriented towards the integration and unification of diverse AI methodologies to forge comprehensive and holistic solutions, a direction that promises to significantly broaden the scope and impact of AI applications. This forward-thinking approach aims not only to enhance existing capabilities in image and video synthesis but also to extend AI's prowess into the realms of 3D object generation and the simulation of functional machines, thereby push-

ing the boundaries of digital creation into new dimensions. Moreover, the exploration of advanced applications in sound generation by leveraging AI techniques is set to further dissolve the distinctions between AI-generated content and the tangible reality, offering experiences that are more immersive and indistinguishable from the natural world. Such unified AI systems, which amalgamate various techniques ranging from Generative Adversarial Networks (GANs) and Diffusion Models to Neural Cellular Automata, are poised to transform a wide array of fields including entertainment, education, healthcare, and manufacturing. By enabling the creation of dynamic, interactive, and highly realistic digital environments and objects, this integrative approach underscores the potential of AI to not just augment but fundamentally redefine the landscape of digital content creation. The envisioned future of AI, as a confluence of disparate techniques yielding seamlessly integrated solutions, heralds a new era of innovation where the lines between digital and physical realities are increasingly blurred, fostering unprecedented levels of creativity and interaction [59].

4.7 Mitigating Biases in Generative Systems

The development of generative text-to-image models has brought forth the challenge of mitigating social biases that these models may perpetuate. Notably, biases towards certain genders or skin tones have been observed in the outcomes of these image generation models. To offset this, a novel approach has been proposed [57], which involves fine-tuning text-to-image models on synthetic data constructed from diverse text prompts. These prompts include a variety of perceived skin tones, genders, ethnicities, professions, and age groups, resulting in a more inclusive synthetic dataset. Fine-tuning models with this diverse data has shown to substantially improve group fairness metrics, reducing biases by significant margins. The process begins with the construction of text prompts through multiplicative combinations of various social qualifiers, which are then used to synthesize images that showcase a broad spectrum of human diversity. This method has proven to be effective, with diversity-finetuned models not only generating content that is more representative of darker skin tones and female genders but also improving overall fairness in model outcomes.

Furthermore, the evaluation of biases in text-to-image models is formulated within a group fairness framework, which assesses whether model outputs unfairly favor certain subgroups over others. By adopting this framework, the research moves beyond merely measuring bias to actively promoting fairness in AI-generated content. This is complemented by user studies that validate the absence of undesirable visual artifacts in finetuned models' outputs. This approach not only showcases the capability to generate more inclusive content but also highlights the potential for AI to serve as a tool for social change. By actively mitigating biases, AI models can foster digital environments that are reflective of a wide spectrum of human experiences and perspectives, thereby promoting inclusivity and diversity. Future work need to address additional forms of biases and explore the applicability of mitigative strategies in video models, thereby extending the principles of fairness and inclusivity to broader multimedia contexts.

4.8 Computational Efficiency and Resolution Enhancement in Visual Processing

In the rapidly evolving domain of image and video processing, implementations such as Skip-Convolutions [58] have significantly improved computational efficiency, enabling high-quality visual outputs without compromising on speed or fidelity. These innovative methods bypass traditional linear processing pathways, allowing for selective data transmission across layers, thereby reducing the computational load while preserving or even enhancing the quality of video content. This technique proves especially beneficial for tasks that demand real-time processing, including live streaming, augmented reality (AR) applications, and instant video communication platforms, by accelerating video editing and enhancement operations. Furthermore, enhancing the resolution of images and videos has become increasingly crucial, with technologies like AI-based upscaling playing a pivotal role. AI upscaling techniques intelligently increase the resolution of digital media, ensuring that even content produced at lower resolutions can meet the current standards expected by consumers, without the need for extensive computational resources. [60] By integrating such approaches, Skip-Convolutions, along with AI upscaling, address the growing demand for premium video content that can be swiftly produced, edited, and shared, marking a shift towards more efficient content creation paradigms. Consequently, these advancements not only elevate the technical capabilities of video processing software but also substantially contribute to the broader field of digital media, where speed, efficiency, and enhanced resolution are extremely important. This progress highlights the ongoing necessity for innovation in computational methodologies to navigate the challenges presented by the increasing demand for multiple video content across various digital platforms.

Chapter 5

CONCLUSIONS

This comprehensive study has elucidated the transformative role of Artificial Intelligence (AI) in the domains of image, video, and sound generation, underscoring the profound implications these technologies hold for the future of digital media production. Through a meticulous exploration of Generative Adversarial Networks (GANs), Diffusion Models, and Neural Cellular Automata, the research has unveiled the significant advancements in AI-driven generative technologies, demonstrating their capacity to produce outputs that are not only photorealistic but also creatively unparalleled. The application of these AI models across various sectors, from healthcare and entertainment to education and digital manufacturing, signifies a paradigm shift in how content is conceived, created, and consumed. The integration of AI in these processes not only enhances efficiency and scalability but also opens up new vistas for innovation, allowing for the creation of content that transcends traditional boundaries of creativity and realism. Furthermore, the critical assessment of the ethical landscape surrounding AI-generated content has highlighted the importance of developing robust ethical guidelines and frameworks to govern the use of these powerful technologies, ensuring that they contribute positively to society and do not perpetuate existing biases or facilitate misinformation.

The project's successful utilization of AI technologies like ChatGPT, MidJourney, Runway Gen-2, ElevenLabs, and Topaz Video AI in creating a compelling video narrative serves as a tangible testament to the potential of AI to revolutionize content creation. The synergy between AI-generated text, images, videos, and voiceovers has not only achieved a high degree of narrative cohesion and visual sophistication but also showcased the practical feasibility of employing AI in streamlining complex production processes without sacrificing artistic integrity. This harmonious integration of AI with human creativity has paved the way for a new era of storytelling, where digital narratives are more engaging, expressive, and immersive. The project's outcomes emphasize the importance of a collaborative approach to AI-driven content creation, where human oversight and creative input guide the AI's capabilities towards achieving a shared vision, thus maximizing the potential of these technologies to enhance storytelling and visual communication.

Looking forward, the trajectory of AI in digital media production is set towards further integration and unification of diverse AI methodologies, promising to expand the scope and impact of AI applications across all facets of content creation. The continuous evolution of AI technologies is poised to redefine the landscape of digital content creation, blurring the lines between digital and physical realities and fostering new levels of creativity and interaction. As AI systems become more autonomous and creative, the imperative for ethical considerations and guidelines becomes increasingly paramount, ensuring that the advancements in AI not only push the boundaries of what is possible in content creation but also align with societal values and ethical principles. The future of AI in digital media production is bright, heralding a new chapter in creative expression and storytelling that is bound only by the limits of our imagination. This study not only contributes to the academic discourse on AI and generative media but also provides a roadmap for future explorations in this dynamic and evolving field.

REFERENCES

[1] Stuart J Russell and Peter Norvig. *Artificial intelligence a modern approach*. London, 2010.

[2] Michael I Jordan and Tom M Mitchell. "Machine learning: Trends, perspectives, and prospects". In: *Science* 349.6245 (2015), pp. 255–260.

[3] Jiaxin Huang et al. "Large language models can self-improve". In: *arXiv preprint arXiv:2210.11610* (2022).

[4] Oludare Isaac Abiodun et al. "State-of-the-art in artificial neural network applications: A survey". In: *Heliyon* 4.11 (2018).

[5] Jürgen Schmidhuber. "Deep learning in neural networks: An overview". In: *Neural networks* 61 (2015), pp. 85–117.

[6] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *nature* 521.7553 (2015), pp. 436–444.

[7] Athanasios Voulodimos et al. "Deep learning for computer vision: A brief review". In: *Computational intelligence and neuroscience* 2018 (2018).

[8] KR1442 Chowdhary and KR Chowdhary. "Natural language processing". In: *Fundamentals of artificial intelligence* (2020), pp. 603–649.

[9] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[10] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. "Stochastic backpropagation and approximate inference in deep generative models". In: *International conference on machine learning*. PMLR. 2014, pp. 1278–1286.

[11] Diederik P Kingma and Max Welling. "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114* (2013).

[12] Ian Goodfellow et al. "Generative adversarial nets". In: *Advances in neural information processing systems* 27 (2014).

[13] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. "Generating videos with scene dynamics". In: *Advances in neural information processing systems* 29 (2016).

[14] Sergey Tulyakov et al. "Mocogan: Decomposing motion and content for video generation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 1526–1535.

[15] Aidan Clark, Jeff Donahue, and Karen Simonyan. "Adversarial video generation on complex datasets". In: *arXiv preprint arXiv:1907.06571* (2019).

[16] Tim Brooks et al. "Generating long videos of dynamic scenes". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 31769–31781.

[17] Robin Rombach et al. “High-resolution image synthesis with latent diffusion models”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 10684–10695.

[18] Jonathan Ho et al. “Imagen video: High definition video generation with diffusion models”. In: *arXiv preprint arXiv:2210.02303* (2022).

[19] Andreas Blattmann et al. “Align your latents: High-resolution video synthesis with latent diffusion models”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 22563–22575.

[20] Agrim Gupta et al. “Photorealistic video generation with diffusion models”. In: *arXiv preprint arXiv:2312.06662* (2023).

[21] Niv Granot et al. “Drop the gan: In defense of patches nearest neighbors as single image generative models”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 13460–13469.

[22] Tero Karras, Samuli Laine, and Timo Aila. “A style-based generator architecture for generative adversarial networks”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 4401–4410.

[23] Or Patashnik et al. “Styleclip: Text-driven manipulation of stylegan imagery”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 2085–2094.

[24] David Silver et al. “Mastering the game of Go with deep neural networks and tree search”. In: *nature* 529.7587 (2016), pp. 484–489.

[25] Christoph Leiter et al. “Chatgpt: A meta-analysis after 2.5 months”. In: *arXiv preprint arXiv:2302.13795* (2023).

[26] R OpenAI. “Gpt-4 technical report. arxiv 2303.08774”. In: *View in Article* 2 (2023), p. 13.

[27] *Midjourney Explore*. https://www.midjourney.com/explore. Accessed: Feb 26 2024.

[28] RunwayML Inc. *RunwayML*. https://runwayml.com/. Accessed: Feb 26 2024.

[29] Eleven Labs. *Eleven Labs*. https://elevenlabs.io/. Accessed: Feb 26 2024.

[30] Mika Westerlund. “The emergence of deepfake technology: A review”. In: *Technology innovation management review* 9.11 (2019).

[31] Kalin Hristov. “Artificial intelligence and the copyright dilemma”. In: *Idea* 57 (2016), p. 431.

[32] Hugo Touvron et al. “Llama: Open and efficient foundation language models”. In: *arXiv preprint arXiv:2302.13971* (2023).

[33] Stability AI. *StableLM: A project by Stability AI*. GitHub repository. Oct 5, 2023. URL: https://github.com/Stability-AI/StableLM.

[34] Albert Q Jiang et al. “Mistral 7B”. In: *arXiv preprint arXiv:2310.06825* (2023).

[35] Justine Moore. *Why 2023 Was AI Video’s Breakout Year and What to Expect in 2024*. https://a16z.com/why-2023-was-ai-videos-breakout-year-and-what-to-expect-in-2024/. Accessed: 2024-02-17. 2023.

[36] European Parliament. *EU AI Act: First Regulation on Artificial Intelligence*. https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence. Accessed: 2024-02-28. 2023.

[37] The State Council of the People's Republic of China. *Interim Measures for the Management of Generative Artificial Intelligence Services [生成式人工智能服务管理暂行办法]*. https://www.gov.cn/zhengce/zhengceku/202307/content_6891752.htm. Accessed: 2024-02-28. July 2023.

[38] The White House. *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/. Accessed: 2024-02-28. Oct. 2023.

[39] Michael M. Grynbaum and Ryan Mac. *The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work*. https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html. Accessed: 2024-02-28. Dec. 2023.

[40] Fengxiang Bie et al. "RenAIssance: A Survey into AI Text-to-Image Generation in the Era of Large Model". In: *arXiv preprint arXiv:2309.00810* (2023).

[41] Paul F Christiano et al. "Deep reinforcement learning from human preferences". In: *Advances in neural information processing systems* 30 (2017).

[42] Lisa Torrey and Jude Shavlik. "Transfer learning". In: *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*. IGI global, 2010, pp. 242–264.

[43] Kyunghyun Cho et al. "On the properties of neural machine translation: Encoder-decoder approaches". In: *arXiv preprint arXiv:1409.1259* (2014).

[44] Emily L Denton, Soumith Chintala, Rob Fergus, et al. "Deep generative image models using a laplacian pyramid of adversarial networks". In: *Advances in neural information processing systems* 28 (2015).

[45] Jascha Sohl-Dickstein et al. "Deep unsupervised learning using nonequilibrium thermodynamics". In: *International conference on machine learning*. PMLR. 2015, pp. 2256–2265.

[46] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models". In: *Advances in neural information processing systems* 33 (2020), pp. 6840–6851.

[47] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. "Unsupervised learning of video representations using lstms". In: *International conference on machine learning*. PMLR. 2015, pp. 843–852.

[48] Silvia Chiappa et al. "Recurrent environment simulators". In: *arXiv preprint arXiv:1704.02254* (2017).

[49] David Ha and Jürgen Schmidhuber. "World models". In: *arXiv preprint arXiv:1803.10122* (2018).

[50] Yong Yu et al. "A review of recurrent neural networks: LSTM cells and network architectures". In: *Neural computation* 31.7 (2019), pp. 1235–1270.

[51] Rahul Dey and Fathi M Salem. "Gate-variants of gated recurrent unit (GRU) neural networks". In: *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*. IEEE. 2017, pp. 1597–1600.

[52] Wilson Yan et al. "Videogpt: Video generation using vq-vae and transformers". In: *arXiv preprint arXiv:2104.10157* (2021).

[53] Chenfei Wu et al. “Nüwa: Visual synthesis pre-training for neural visual world creation”. In: *European conference on computer vision*. Springer. 2022, pp. 720–736.

[54] OpenAI. *Whisper*. https://openai.com/research/whisper. Accessed: 2024-02-27. 2024.

[55] Aaron van den Oord et al. “Wavenet: A generative model for raw audio”. In: *arXiv preprint arXiv:1609.03499* (2016).

[56] Simon Godsill, Peter Rayner, and Olivier Cappé. *Digital audio restoration*. Springer, 2002.

[57] Piero Esposito et al. “Mitigating stereotypical biases in text to image generative systems”. In: *arXiv preprint arXiv:2310.06904* (2023).

[58] Amirhossein Habibian et al. “Skip-convolutions for efficient video processing”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 2695–2704.

[59] Patrick Esser, Peter Michael, and Soumyadip Sengupta. “Towards Unified Keyframe Propagation Models”. In: *arXiv preprint arXiv:2205.09731* (2022).

[60] Andrey Ignatov et al. “Real-time video super-resolution on smartphones with deep learning, mobile ai 2021 challenge: Report”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 2535–2544.

Appendix A

ADDITIONAL MATERIAL

The link of the video is here: https://youtube.com/shorts/C_Ng1JBiA-c?feature=shared