

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 AIGC in Image and Video	3
1.4 AIGC in Sound	7
1.5 Current Development Trends	12
2 Material and Methods	16
2.1 Workflow	16
2.2 ChatGPT	16
2.3 MidJourney	18
2.4 Runway Gen-2	18
2.5 ElevenLabs	19
2.6 Topaz AI	21
3 Results	22
3.1 Overview of Project Outcome	22
3.2 ChatGPT-Assisted Storyline and Description Generation	23
3.3 Image Stills Generation with MidJourney	24
3.4 Video Clips Generation Using Runway Gen-2	24
3.5 Voice Over Creation with ElevenLabs	25
3.6 Video Quality Enhancement with Topaz AI	26
4 Discussion	28
4.1 Foundational Models	28
4.2 Specialized Architectures	29
4.3 Ethical and Societal Considerations	29
4.4 Efficiency and Scalability	30

4.5	Extension to Sound Generation	30
4.6	Future Directions	31
5	Conclusions	32
	References	34
A	Additional Material	38

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 throughout this research. Also, I am grateful to the other members of the SW team for their support and guidance. In the end, I would like to thank the SW office for their support and guidance, and also extend my thanks to the generous funding provided by SW grants, ¥1800, which made the experiential learning part possible. Finally, I would like to thank my family and friends for their support and encouragement.

LIST OF FIGURES

1.1	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. [47]	11
1.2	The development of generative AI Video Timeline in 2023, the most updates were launched in the 4th quarter. [50]	13
1.3	The comparison of 21 public AI Video Products as of December 2023 in terms of Generation Type, Max Length of clips, Scalability, Controllability, and Usage. [50]	13
1.4	Evolutionary Timeline of TTI Models: Distinguished by parameter size and colors. An asterisk next to a model indicates its parameter count excludes text encoders. [36]	14
2.1	Using ChatGPT to generate vivid description prompt and prepare for creating the image stills.	17
2.2	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.	18
2.3	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	19
2.4	Using Professional Voice Cloning from ElevenLabs to apply text to speech and add the voice over to the video	20
2.5	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.	20
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.	21

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 cre-

ation, 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 AIGC in Image and Video

1.3.1 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 [32]. 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:

$$\log q_\phi(z^{(i)}|x^{(i)}) = \log \mathcal{N}(z^{(i)}; \mu^{(i)}, \sigma^{2(i)}I). \quad (1.1)$$

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:

$$z_i = \mu_i + \epsilon \times \sigma_i, \quad (1.2)$$

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:

$$\mathcal{L}_b = -D_{KL}(q(z|x)\|p(z)) + \frac{1}{L} \sum_{i=1}^L (\log P(x|z)). \quad (1.3)$$

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:

$$\mathcal{L}(\theta, x^i) = -\frac{1}{2} \sum_{i=1}^d (\mu_{(i)}^2 + \sigma_{(i)}^2 - \log \sigma_{(i)}^2 - 1) + \frac{1}{L} \sum_{i=1}^L (\log P(x^i|z)). \quad (1.4)$$

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.

1.3.2 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 :

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{data}(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G(z)))] , \quad (1.5)$$

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 :

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^m \left[\log D(x^{(i)}) + \log(1 - D(G(z^{(i)}))) \right], \quad (1.6)$$

and the gradient descent step for G :

$$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^m \log(1 - D(G(z^{(i)}))), \quad (1.7)$$

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 [33] and object detection [12].

1.3.3 Diffusion Models

Diffusion models [34, 35] 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:

$$q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I}), \quad (1.8)$$

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

Algorithm 1 Diffusion model training [36]

- 1: for every training iteration do
 - 2: Sample t from discrete timestep $1, 2, \dots, T$, 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:

$$p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)), \quad (1.9)$$

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 [36]

-
- 1: Sample x_t from normal Gaussian distribution $x_t \sim \mathcal{N}(0, I)$.
 - 2: Sample discrete timesteps from $1, 2, \dots, T$ 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.

1.3.4 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 [37], 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 multi-

media contexts.

1.3.5 Computational Efficiency and Resolution Enhancement in Visual Processing

In the rapidly evolving domain of image and video processing, implementations such as Skip-Convolutions [38] 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. [39] 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.

1.4 AIGC in Sound

1.4.1 Recurrent Neural Networks (RNN)

Recurrent Neural Networks (RNNs) [40–42] 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:

$$h_t = \sigma(W_{ih}x_t + W_{hh}h_{t-1} + b_h) \quad (1.10)$$

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:

$$y_t = W_{ho}h_t + b_o \quad (1.11)$$

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) [43] units or Gated Recurrent Units (GRUs) [44] 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.

1.4.2 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 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:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (1.12)$$

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:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (1.13)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (1.14)$$

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:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (1.15)$$

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:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (1.16)$$

$$h_t = o_t * \tanh(C_t) \quad (1.17)$$

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.

1.4.3 Autoregressive Transformers

Autoregressive Transformers [45, 46] 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:

$$X' = X + PE \quad (1.18)$$

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:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1.19)$$

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:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O \quad (1.20)$$

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (1.21)$$

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:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \quad (1.22)$$

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:

$$P(y_t|y_{<t}) = \text{softmax}(y_{t-1}W + b) \quad (1.23)$$

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.

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.

1.4.4 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 [47], 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:

$$F = \text{MFCC}(S) \quad (1.24)$$

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:

$$H = \text{Encoder}(F; \theta_e) \quad (1.25)$$

$$\tilde{F} = \text{Decoder}(H; \theta_d) \quad (1.26)$$

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:

$$V_t = \text{Conversion}(H_t; \theta_c) \quad (1.27)$$

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:

$$\hat{S} = \text{WaveNet}(V_t; \theta_w) \quad (1.28)$$

where \hat{S} is the synthesized speech waveform, and θ_w are the parameters of a WaveNet [48] 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.

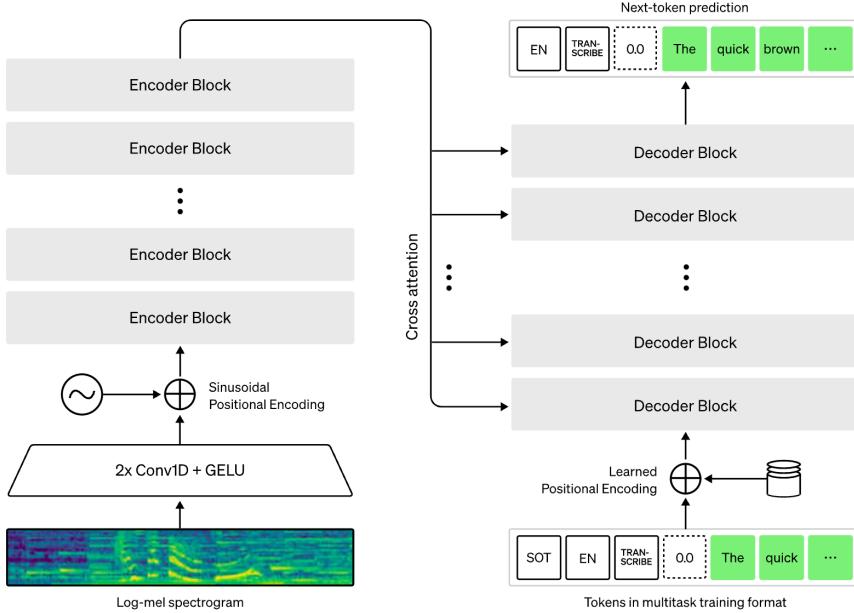


Figure 1.1: Autoregressive transformer model architecture used for audio processing tasks like transcription and voice synthesis. The model utilizes a stack of encoder and decoder blocks, with sinusoidal positional encodings and learned positional encodings for sequence representation. [47]

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.

1.4.5 Audio Modification

There are several methods to tackle audio modification challenges such as noise reduction, sound separation, and audio restoration [49]. The processes are shown below:

Noise reduction aims to eliminate unwanted background noise from audio recordings, enhancing the clarity of the sound. This can be mathematically represented by:

$$S_{clean} = S_{noisy} - N \quad (1.29)$$

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:

$$S_i = F(S_{mix}; \theta_i) \quad (1.30)$$

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:

$$S_{restored} = R(S_{degraded}; \theta_r) \quad (1.31)$$

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:

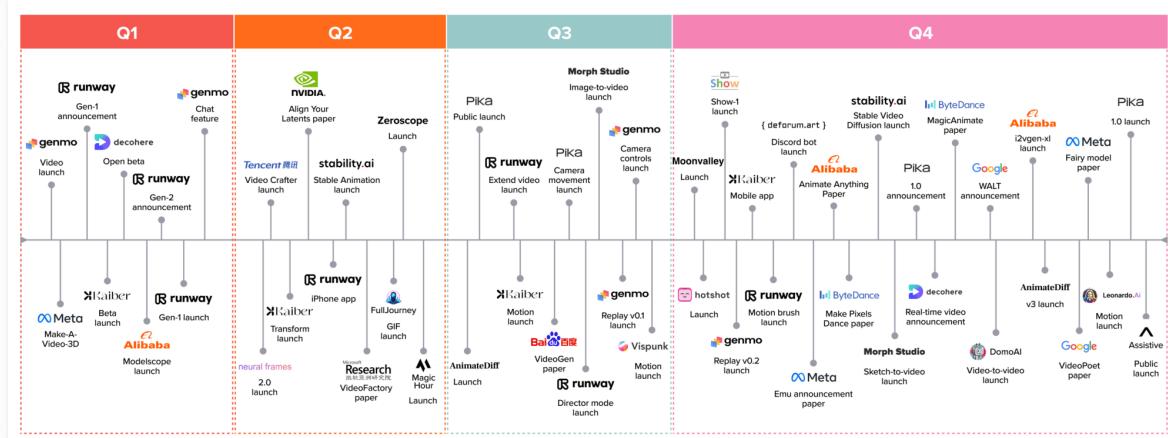
$$\hat{S} = G(S_{processed}; \theta_g) \quad (1.32)$$

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.

1.5 Current Development Trends

Generative AI Video Timeline: 2023



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 af6z.com/disclosures for more information.



Figure 1.2: The development of generative AI Video Timeline in 2023, the most updates were launched in the 4th quarter. [50]

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 af6z.com/disclosures for more information.



Figure 1.3: The comparison of 21 public AI Video Products as of December 2023 in terms of Generation Type, Max Length of clips, Scalability, Controllability, and Usage. [50]

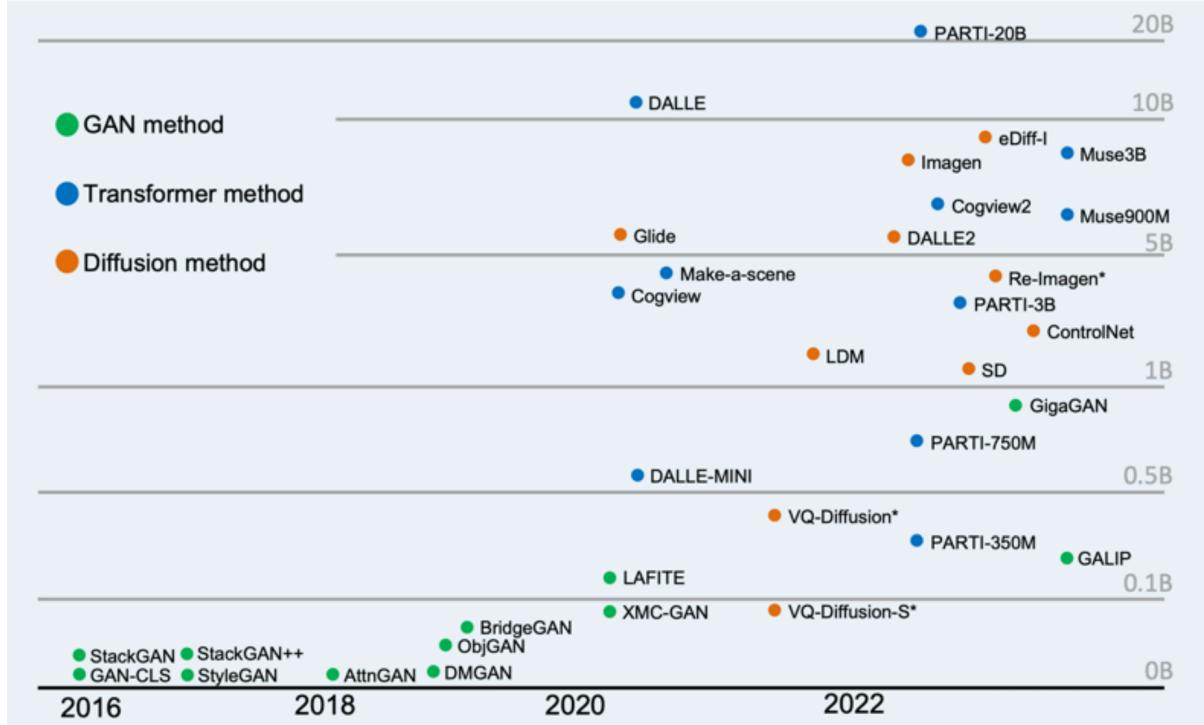


Figure 1.4: Evolutionary Timeline of TTI Models: Distinguished by parameter size and colors. An asterisk next to a model indicates its parameter count excludes text encoders. [36]

The current landscape of AI-driven advancements in image, video, and now sound generation is witnessing unprecedented growth, characterized by both the deepening sophistication of technologies and the broadening scope of applications. In the realm of image and video generation, the pursuit of high-fidelity and high-resolution synthesis has led to the development of models like Latent Diffusion Models, which excel in generating images with remarkable detail and clarity [17]. The exploration of AI's capabilities has expanded into 3D space, with Neural Cellular Automata models facilitating the creation of complex three-dimensional objects and functional machines, marking a significant leap from traditional 2D image manipulation [sudhakaran2021growing]. Efforts to enhance real-time processing and efficiency have brought about innovations such as Skip-Convolutions, optimizing video processing tasks to achieve faster outputs without compromising quality [38].

Ethical considerations have also come to the forefront, with initiatives aimed at mitigating biases in text-to-image generation and developing safeguards against the misuse of AI-generated deepfakes [37]. The push towards photorealism sees advanced algorithms augmenting computer-generated imagery to levels indistinguishable from actual photographs, enhancing the authenticity of digital creations [51]. Moreover, the trend towards unification of different AI techniques for seamless and integrated content creation solutions is evident, promising a more cohesive approach to digital media production [52].

Expanding into the auditory domain, AI's influence on sound generation mirrors these trends, with significant advancements in creating and modifying sounds with high precision. AI-

driven technologies are now capable of synthesizing music, voice, and sound effects with unprecedented realism, catering to diverse applications from virtual assistants to immersive game environments. In music composition, AI algorithms harness vast datasets to produce compositions across genres, democratizing music creation [53]. Voice synthesis technologies like WaveNet and Tacotron are refining speech generation to include emotional inflections and accents, broadening the horizon for interactive media [48]. Moreover, in sound effects generation, AI models trained on sound libraries are crafting unique auditory experiences, enhancing the realism of virtual spaces [54].

These developments signify not merely a technological evolution but a comprehensive transformation of the digital content landscape. The integration of sophisticated AI in image, video, and sound generation is not only pushing the boundaries of creativity and realism but also addressing efficiency, ethical, and accessibility concerns. As these trends continue to unfold, they herald a future where AI's role in digital media production becomes increasingly central, shaping the way content is created, consumed, and perceived across various platforms.

Chapter 2

MATERIAL AND METHODS

2.1 Workflow

The workflow of this video-creating project is shown below:

- 1: Write the storyline of the video, including the description prompt of each scene and the narrative flow. I used ChatGPT to polish them to make them more concise and coherent.
- 2: Generate and edit the image stills by using MidJourney.
- 3: Use Runway Gen-2 to generate the video that based on the footage as the key frames.
- 4: Edit and combine the video clips on CapCut.
- 5: Use ElevenLabs to clone my voice and train the text to speech model.
- 5: Add voice over, music and sound effects.
- 6: Use Topaz Video AI to improve the video quality like (change the frame rate from 60 FPS to 120 FPS)
- 6: Publish the video on social media. (like YouTube and Instagram)

2.2 ChatGPT

ChatGPT, a state-of-the-art language model developed by OpenAI, operates on the transformer architecture, a breakthrough in deep learning that has revolutionized natural language processing (NLP). It is pre-trained on a diverse corpus of text from the internet, enabling it to grasp a wide array of language nuances, styles, and topics. This pre-training involves unsupervised learning on vast datasets, followed by fine-tuning for specific tasks, which allows ChatGPT to generate coherent, contextually relevant text across many domains without task-specific data. The model's architecture, consisting of multiple layers of transformer blocks, employs self-attention mechanisms to process input text in parallel, significantly improving efficiency and understanding of complex language patterns. Its large parameter count, in the order of billions, captures deep linguistic structures, making it adept at a range of tasks from simple text generation to complex question-answering and summarization without needing explicit programming for each task. The transformer model [55] and subsequent innovations by OpenAI, including GPT-3's methodology [56], underscore the

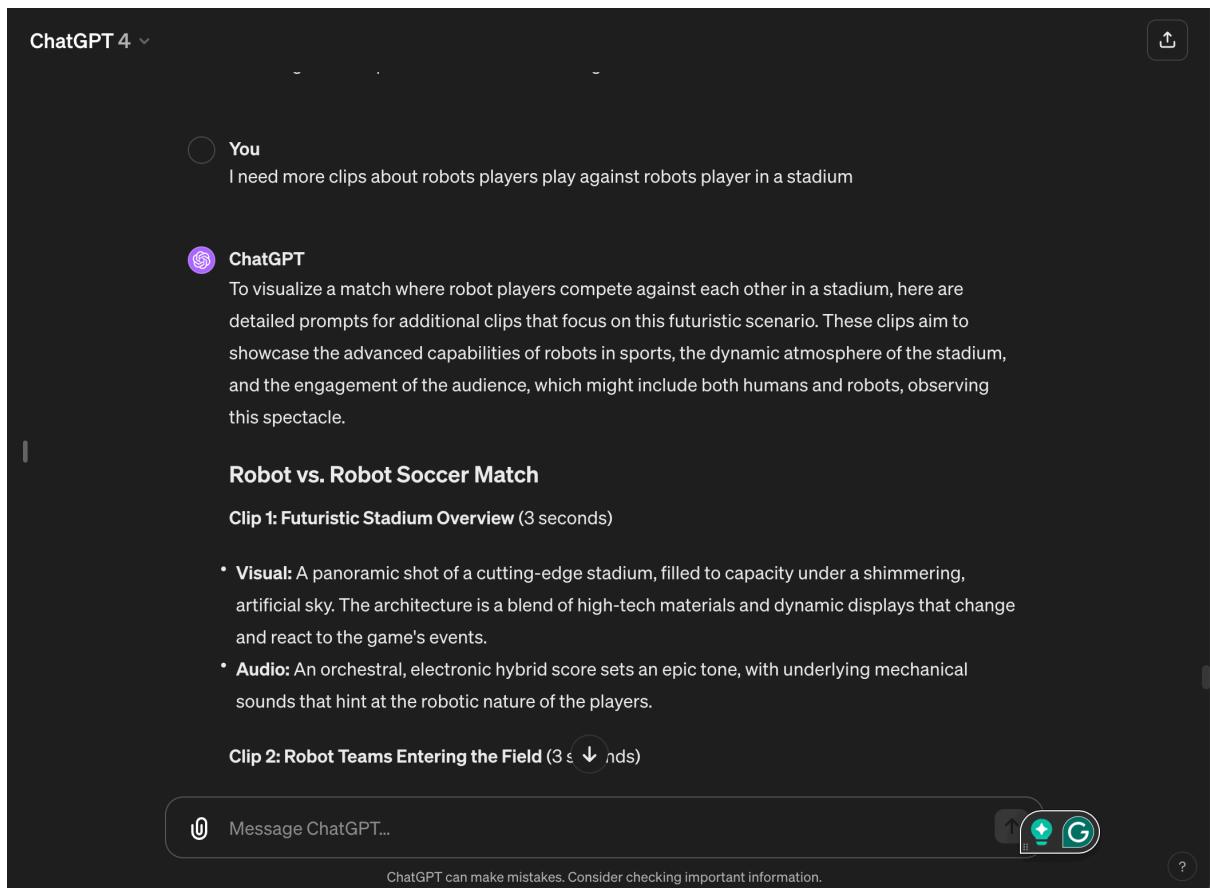


Figure 2.1: Using ChatGPT to generate vivid description prompt and prepare for creating the image stills.

technical foundation and advanced capabilities of ChatGPT, making it a quintessential tool in advancing AI's role in understanding and generating human language.

2.3 MidJourney

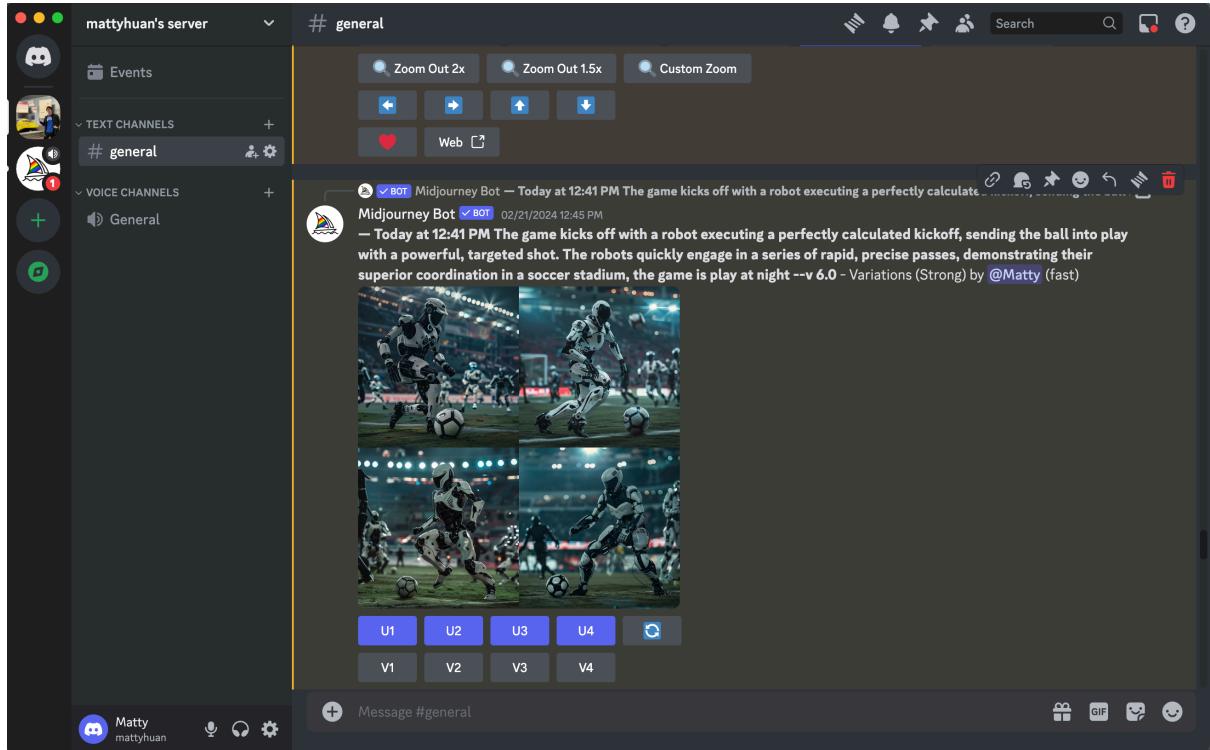


Figure 2.2: Using ChatGPT’s prompts to generate good quality image stills using MidJourney, and then upload them as the key frames to Runway Gen-2 for video clip generation.

MidJourney is an innovative AI model that specializes in generating highly detailed and creative visual content, operating on a complex framework that leverages deep learning techniques to interpret and visualize textual descriptions into compelling images. It utilizes a variant of generative adversarial networks (GANs), where two neural networks work in tandem; one generates images based on textual input, while the other evaluates these images against real-world examples to enhance fidelity and creativity. This iterative process, rooted in the adversarial training methodology, allows MidJourney to produce images that are not only visually stunning but also contextually aligned with the input prompts. The model’s architecture is designed to understand and interpret a wide range of descriptive languages, enabling users to guide the creative process through detailed prompts. Advances in GANs and their application in image generation [12] [22], underpin MidJourney’s capabilities. These foundational studies demonstrate the technical and theoretical framework that allows MidJourney to push the boundaries of AI-driven creativity, showcasing its potential to transform visual content generation across various domains.

2.4 Runway Gen-2

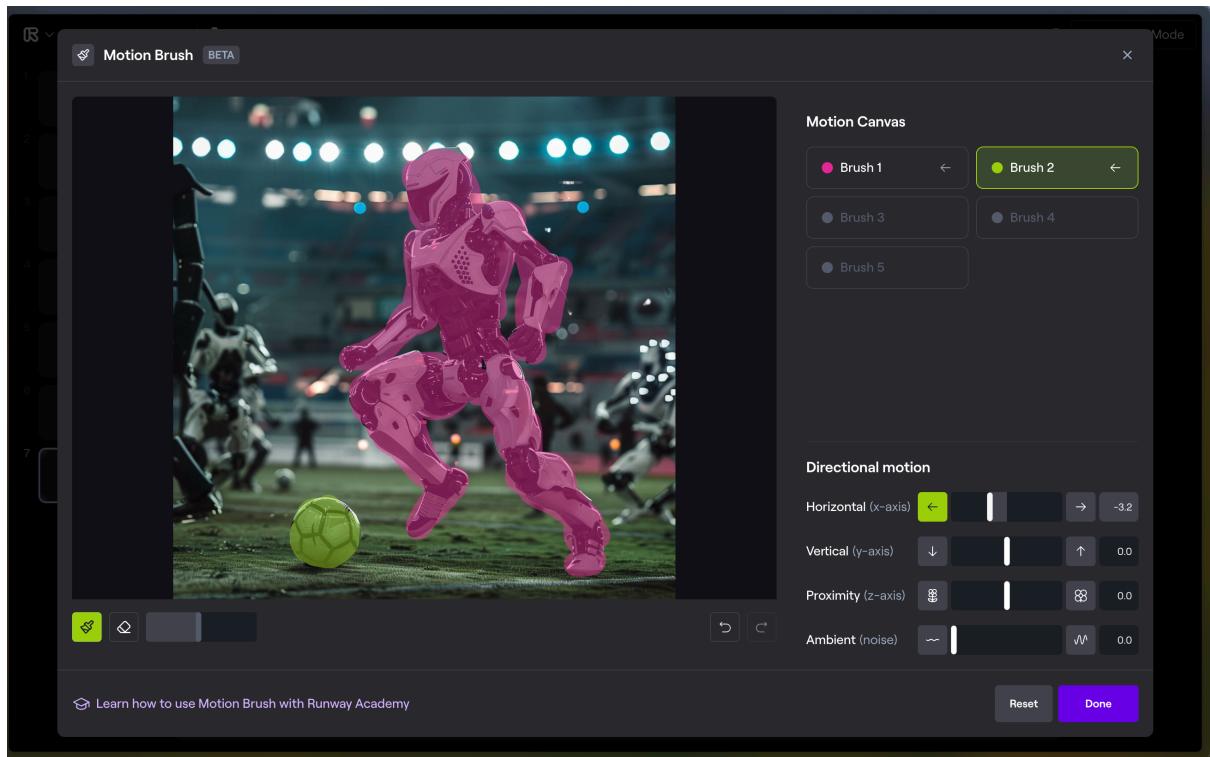


Figure 2.3: Using the "Image + Text" mode in Runway Gen-2 to generate the video, with the help of motion brush function to control how the characters move

Runway Gen-2 is an advanced machine learning platform designed for creative and generative tasks, leveraging cutting-edge models and frameworks to facilitate the creation of digital art, design, and multimedia content. At its core, Runway Gen-2 incorporates a variety of pre-trained models, including those based on GANs (Generative Adversarial Networks) and VAEs (Variational Autoencoders) [11], allowing users to generate highly detailed and diverse outputs. The platform is built on a modular architecture that supports seamless integration of new models and customizations, making it adaptable to evolving creative needs. It employs a user-friendly interface that democratizes access to complex AI technologies, enabling artists, designers, and creators to experiment with AI without deep technical knowledge. Significant advancements in Runway Gen-2's underlying technology include the optimization of model training and inference processes, which are crucial for real-time performance and high-quality generation capabilities. This optimization is supported by efficient data processing and model deployment strategies, ensuring that users can work with large datasets and complex models effectively.

2.5 ElevenLabs

ElevenLabs leverages advanced deep learning techniques to create highly realistic and customizable synthetic voices. At its core, the technology utilizes state-of-the-art voice synthesis models, which are trained on vast datasets of human speech. These models, often based on variants of the transformer neural network architecture, enable the generation of speech that closely mimics human intonation, emotion, and nuances. The process involves two

Figure 2.4: Using Professional Voice Cloning from ElevenLabs to apply text to speech and add the voice over to the video

Figure 2.5: 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.

main stages: voice cloning [57], where a target voice is replicated from a small sample, and text-to-speech (TTS) synthesis, where the cloned voice is used to convert written text into spoken words. This sophisticated approach ensures a high degree of vocal fidelity and naturalness, making it suitable for applications in audiobooks, virtual assistants, and more. The methodology behind ElevenLabs' technology draws upon recent advancements in machine learning and NLP, emphasizing the importance of ethical considerations and user consent in voice cloning.

2.6 Topaz AI

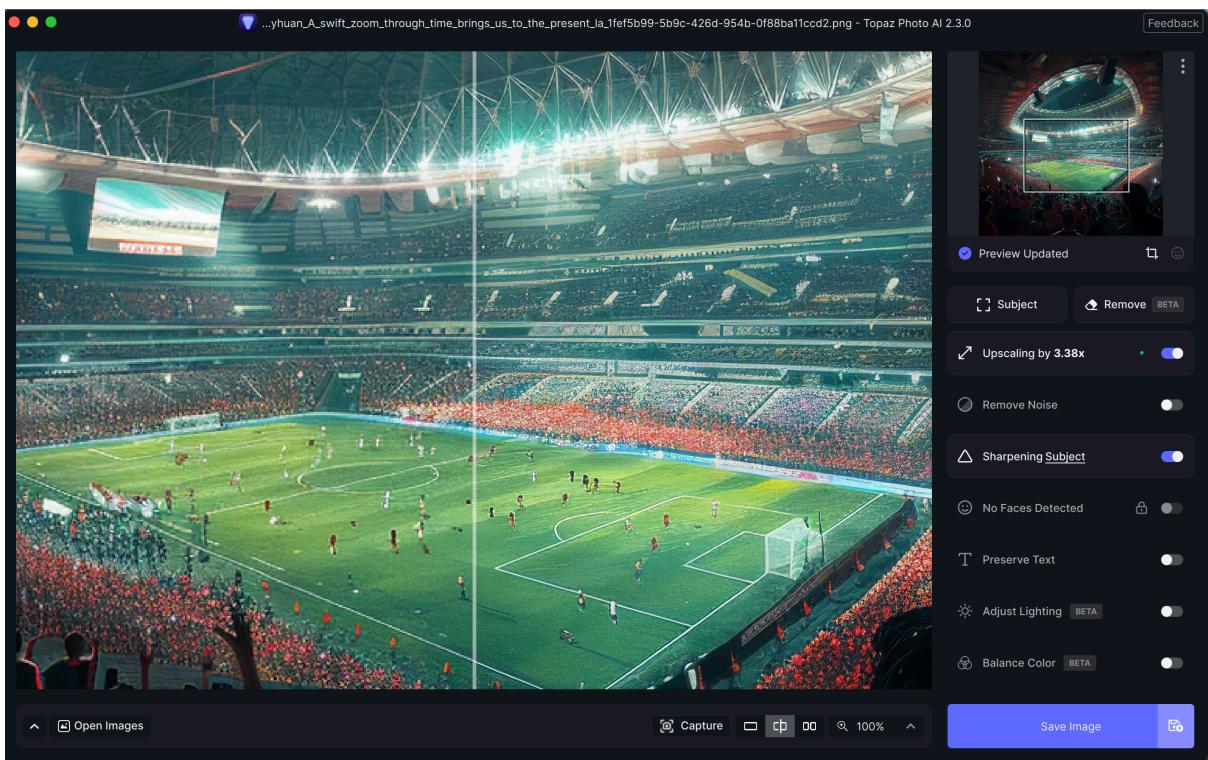


Figure 2.6: Using Topaz AI to upscale and improve the video quality, like changing the frame rate from 60 FPS to 120 FPS, removing noise, and sharpen the subject.

Topaz AI represents a suite of cutting-edge tools designed to enhance image and video quality through deep residual learning [58]. At its core, Topaz AI utilizes deep learning algorithms to perform tasks such as upscaling, denoising, and sharpening images and videos. The technology is based on convolutional neural networks (CNNs), a type of deep neural network particularly effective for processing visual data. These networks are trained on vast datasets of images, allowing the models to learn how to identify and correct imperfections in new images with remarkable accuracy.

Chapter 3

RESULTS

3.1 Overview of Project Outcome

The video project, a culmination of cutting-edge AI technology and creative vision, presents a narrative that is both visually stunning and thematically profound. Spanning a five-minute duration, it intricately weaves together a series of meticulously crafted scenes, each contributing to a cohesive storyline that explores the symbiotic relationship between humanity and artificial intelligence. Set against the backdrop of futuristic landscapes, the narrative delves into themes of exploration, discovery, and 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 tranquil, digitally-rendered natural vistas to bustling, urban technoscapes alive with the vibrancy of futuristic life. The selection and progression of scenes are carefully orchestrated to not only captivate the audience with visual splendor but also to provoke thought and reflection on the evolving role of AI in our world.

The project's success in delivering a compelling narrative is largely attributable to the sophisticated use of AI technologies at each stage of the creative process. From the generation of high-fidelity image stills with MidJourney to the seamless video synthesis facilitated by Runway Gen-2, the tools employed exemplify the potential of AI to transcend traditional boundaries of content creation. The narrative is further enriched with a voiceover, rendered with remarkable realism 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 nuanced potential of AI in crafting engaging and emotionally resonant storytelling. The video's aesthetic and thematic achievements are underscored by the technical prowess of Topaz AI, which elevates the visual quality through enhancements such as frame rate optimization and noise reduction, ensuring that the final product stands as a testament to the harmony between technological innovation and creative expression.

This project, through its successful fusion of AI capabilities and creative narrative design, not only showcases the practical applications of these technologies in producing high-quality digital content but also opens a dialogue on the future of storytelling in the age of artificial intelligence. It stands as a compelling example of how AI can be leveraged to create content that is not just visually captivating, but also rich in thematic depth and narrative complexity, challenging us to envision the limitless possibilities of creative expression in the digital era.

3.2 ChatGPT-Assisted Storyline and Description Generation

In the process of transforming the initial concepts into a coherent and engaging narrative, ChatGPT played an indispensable role, serving as both a creative partner and a critical editor. The tool was pivotal in refining the storyline, ensuring that the flow between scenes was seamless and that the narrative arc was compelling from start to finish. By inputting rough outlines and thematic elements into ChatGPT, the project leveraged the AI's ability to suggest enhancements, fill narrative gaps, and enrich the storyline with deeper context. This collaboration resulted in a narrative that was not only structurally sound but also imbued with a level of depth and nuance that might have been challenging to achieve otherwise. The AI's contributions were particularly valuable in ensuring consistency in tone and pacing, which are crucial for maintaining viewer engagement in a narrative that traverses a range of futuristic landscapes and complex human-AI interactions.

Beyond the overarching narrative structure, ChatGPT's role extended to the crafting of detailed scene descriptions, which were critical for the subsequent visual content creation phase. For each scene, the AI was provided with a brief, outlining the desired visual elements, mood, and thematic intentions. In response, ChatGPT generated descriptions that vividly captured the essence of these scenes, providing a rich tapestry of details that guided the visual realization process. These descriptions not only included visual cues but also offered suggestions on mood, lighting, and potential dynamics within the scene, thereby ensuring that the visual content was not only aesthetically pleasing but also thematically coherent. The iterative refinement of these descriptions, through a feedback loop between the creators and ChatGPT, exemplified the synergistic potential of combining human creativity with AI's generative capabilities, leading to scene descriptions that were both imaginative and aligned with the project's vision.

The integration of ChatGPT in this early stage of the project underscores the transformative potential of AI in the realm of creative storytelling and content creation. By harnessing ChatGPT's capabilities, the project not only achieved a higher level of narrative cohesion and detail but also highlighted the broader implications of AI as a tool for creative augmentation. The success of this collaboration between human creativity and artificial intelligence sets a precedent for future projects, suggesting a paradigm where AI's role extends beyond mere assistance to becoming a central component in the creative process. This approach not only enhances the efficiency and depth of content creation but also opens new avenues for exploring complex themes and narratives in digital storytelling.

3.3 Image Stills Generation with MidJourney

The creation of image stills, a critical component of the video project, was significantly enhanced by the utilization of MidJourney, an AI tool known for its ability to generate visually stunning and highly detailed images. Leveraging the refined scene descriptions developed with the assistance of ChatGPT, MidJourney was tasked with translating these textual narratives into compelling visual art. The process involved feeding the AI with descriptive prompts, meticulously crafted to encapsulate the essence and intricacies of each scene. MidJourney responded with a series of image stills that not only matched the creators' vision in terms of aesthetic appeal but also surpassed expectations by adding layers of complexity and depth to the visual narrative. The images generated by MidJourney stood out for their vivid representation of futuristic concepts, capturing the imagination with their realism and attention to detail. This phase of the project highlighted the AI's ability to interpret and visualize creative prompts, transforming abstract ideas into tangible visual expressions that significantly contributed to the storytelling process.

The iterative collaboration between the creative team and MidJourney was instrumental in refining the visual output to align closely with the project's thematic goals. Through a process of selection and refinement, the team worked with the AI to fine-tune the images, adjusting elements such as color schemes, composition, and thematic details to ensure coherence with the overall narrative flow. This back-and-forth not only optimized the visual quality of the stills but also ensured that each image contributed effectively to the unfolding story, enhancing the emotional and thematic resonance of the scenes. The flexibility and responsiveness of MidJourney to creative feedback underscored the potential of AI as a dynamic tool in the content creation process, capable of accommodating nuanced artistic direction and contributing creatively to the project's vision.

The successful integration of AI-generated image stills into the video project underscored the transformative impact of artificial intelligence on the field of digital media production. By leveraging MidJourney's capabilities, the project not only achieved a high degree of visual sophistication but also demonstrated the feasibility of utilizing AI to streamline the content creation process without compromising on creativity or artistic integrity. This approach to visual storytelling, where AI plays a central role in bringing narrative concepts to life, represents a significant advancement in the capabilities of content creators, offering new possibilities for exploring and expressing complex themes through digital media. The collaboration between human creativity and AI in this phase of the project serves as a compelling example of the synergies that can be achieved, paving the way for future endeavors that harness the power of AI to enhance and expand the boundaries of visual art and storytelling.

3.4 Video Clips Generation Using Runway Gen-2

The transition from static image stills to dynamic video clips marked a pivotal phase in the video project, facilitated by the advanced capabilities of Runway Gen-2. This AI-driven plat-

form, known for its robust video processing and generation features, was tasked with breathing life into the meticulously crafted scenes, transforming the still imagery into seamless and engaging video sequences. Utilizing the "Image + Text" mode, the project team input the detailed scene descriptions and corresponding stills generated by MidJourney into Runway Gen-2. The AI responded by crafting video clips that not only preserved the visual fidelity of the original stills but also introduced motion and continuity, creating a sense of progression and narrative flow. The application of Runway Gen-2's motion brush tool was particularly noteworthy, as it allowed for precise control over the animation of characters and elements within the scenes, ensuring that each movement contributed to the storytelling in a meaningful way. This phase highlighted the AI's ability to interpret creative direction and apply complex video synthesis techniques, thereby enhancing the narrative with dynamic visual storytelling.

The collaborative process between the creative team and Runway Gen-2 was characterized by an iterative approach, where generated video clips were reviewed and refined to meet the project's exacting standards. Adjustments were made to the pacing, transitions, and animations within the clips, ensuring that each sequence flowed smoothly into the next and that the overall narrative cohesion was maintained. This fine-tuning process underscored the importance of human oversight in achieving the desired outcome, with the creative team leveraging Runway Gen-2's flexibility to enact changes that aligned with the project's vision. The AI's capacity to adapt to creative feedback and generate revised clips efficiently was instrumental in realizing a video that was both visually captivating and narratively compelling. This phase of the project demonstrated the synergistic potential of combining human creativity with AI's technical prowess, achieving a balance between artistic expression and technological innovation.

The successful integration of AI-generated video clips into the final project underscored the transformative potential of artificial intelligence in the realm of video production. By utilizing Runway Gen-2, the project not only achieved a high level of visual and narrative dynamism but also showcased the feasibility of employing AI to streamline complex video creation processes. This innovative approach to video storytelling, where AI plays a central role in enhancing the visual narrative, represents a forward leap in the digital content creation landscape. It offers content creators new avenues for artistic exploration and expression, enabling the production of rich, immersive narratives that might have been unattainable through traditional methods alone. The collaborative achievements of this phase serve as a testament to the evolving relationship between human creativity and artificial intelligence, highlighting a future where AI's role in creative industries continues to expand, opening up unprecedented possibilities for storytelling and visual communication.

3.5 Voice Over Creation with ElevenLabs

The addition of a voiceover to the video project represented a critical juncture in enhancing its narrative depth and emotional resonance, achieved through the innovative use of Eleven-

Labs' voice cloning technology. This phase involved the creation of a synthetic voice that could narrate the video's storyline with a level of expressiveness and nuance akin to a human speaker. By submitting a sample of the desired voice tone and modulation to ElevenLabs, the project team was able to generate a voice clone that not only matched the specified attributes but also conveyed the narrative with an emotional depth that significantly enriched the viewer's experience. This technology's ability to produce a voice that could seamlessly switch between tones—ranging from contemplative to exhilarating—allowed for a dynamic storytelling approach that aligned closely with the video's thematic shifts and narrative pace. The integration of this voiceover into the video project was not just about adding a layer of auditory information; it was about imbuing the narrative with a personal touch that invited the audience into the story, enhancing the engagement and immersion.

The process of integrating the AI-generated voiceover with the video content underscored the importance of precision in timing and emotional congruence. The project team meticulously aligned the voiceover with the visual sequences, ensuring that each spoken word resonated with the unfolding scenes and actions. This alignment was crucial for maintaining the narrative flow and ensuring that the voiceover amplified, rather than distracted from, the visual storytelling. Adjustments were made to the pacing of the voiceover and the timing of key narrative points to achieve a harmonious balance between the audio and visual elements. The capability of ElevenLabs' technology to fine-tune the voiceover in response to feedback allowed for iterative improvements, resulting in a final product where the voiceover acted as a cohesive and integral component of the storytelling experience.

The successful incorporation of the voiceover, generated using ElevenLabs' voice cloning technology, into the video project highlights the expanding role of AI in enhancing multimedia content beyond visual elements. This phase of the project illustrates the potential of AI to add depth and emotion to digital storytelling, transforming the way stories are told and experienced. By leveraging such advanced AI tools, content creators can now craft narratives that are more engaging, expressive, and impactful, marking a significant evolution in digital media production. The use of AI-generated voiceovers, in particular, opens up new possibilities for personalizing and enriching content, making it accessible to a broader audience and further blurring the lines between human and machine-generated creativity. This achievement not only exemplifies the technical capabilities of current AI technologies but also points to a future where AI plays a central role in all facets of content creation, offering innovative solutions for storytelling challenges and setting new standards for narrative excellence.

3.6 Video Quality Enhancement with Topaz AI

The final enhancement phase of the video project, focusing on improving the overall video quality, was adeptly handled using Topaz Video AI. This phase was pivotal in elevating the visual presentation to professional standards, addressing key aspects such as frame rate optimization, noise reduction, and clarity enhancement. Topaz Video AI's advanced deep learn-

ing algorithms analyzed the video clips to identify and rectify common issues that affect video quality, such as blurriness, artifacts, and jitter. By converting the frame rate from 60 FPS to 120 FPS, the tool significantly smoothed motion, lending a more fluid and natural feel to the movements within each scene. This adjustment was particularly effective in scenes with fast-paced action or complex animations, where clarity and smoothness are crucial for viewer comprehension and enjoyment.

Noise reduction was another critical area addressed by Topaz Video AI, which proved essential in maintaining the visual integrity of scenes, especially those generated under less-than-ideal conditions or those requiring a high degree of digital compositing. The AI algorithms worked to selectively remove visual noise without compromising the detail and texture that give each scene its unique character. This delicate balance ensured that the final video maintained a clean and polished look, free from distractions that could detract from the viewer's immersion. Additionally, the tool's ability to enhance the overall sharpness and clarity of the video content helped in accentuating details that might otherwise go unnoticed, making each frame more impactful and engaging.

The integration of Topaz Video AI into the project's workflow underscores the transformative impact of AI on video production, particularly in the post-production phase. By leveraging the tool's capabilities, the project achieved a level of visual quality that rivals professional production standards, showcasing the potential of AI to streamline and enhance the content creation process. This phase not only highlighted the technical proficiency of current AI tools in improving video aesthetics but also demonstrated the practical value of AI in resolving common challenges faced by content creators. The improvements made through Topaz Video AI brought the project to its culmination, with a final product that stood as a testament to the synergy between human creativity and artificial intelligence. This accomplishment illustrates the evolving landscape of digital media production, where AI technologies play an increasingly central role in enabling creators to achieve their vision with greater efficiency and precision, pushing the boundaries of what is possible in storytelling and visual communication.

Chapter 4

DISCUSSION

The roadmap for future development in AI's application to image, video, and now sound generation encapsulates a journey from foundational models to the cusp of groundbreaking innovations, illustrating a trajectory that is as diverse as it is profound. This progression can be dissected into several key stages, each marking significant milestones in the evolution of AI technologies and their applications.

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 [37].

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 [38]. 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 pushing 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 [52].

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] Kyunghyun Cho et al. “On the properties of neural machine translation: Encoder-decoder approaches”. In: *arXiv preprint arXiv:1409.1259* (2014).
- [33] 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).
- [34] Jascha Sohl-Dickstein et al. “Deep unsupervised learning using nonequilibrium thermodynamics”. In: *International conference on machine learning*. PMLR. 2015, pp. 2256–2265.
- [35] Jonathan Ho, Ajay Jain, and Pieter Abbeel. “Denoising diffusion probabilistic models”. In: *Advances in neural information processing systems* 33 (2020), pp. 6840–6851.
- [36] 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).

- [37] Piero Esposito et al. “Mitigating stereotypical biases in text to image generative systems”. In: *arXiv preprint arXiv:2310.06904* (2023).
- [38] 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.
- [39] 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.
- [40] 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.
- [41] Silvia Chiappa et al. “Recurrent environment simulators”. In: *arXiv preprint arXiv:1704.02254* (2017).
- [42] David Ha and Jürgen Schmidhuber. “World models”. In: *arXiv preprint arXiv:1803.10122* (2018).
- [43] Yong Yu et al. “A review of recurrent neural networks: LSTM cells and network architectures”. In: *Neural computation* 31.7 (2019), pp. 1235–1270.
- [44] 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 (MWCAS)*. IEEE. 2017, pp. 1597–1600.
- [45] Wilson Yan et al. “Videogpt: Video generation using vq-vae and transformers”. In: *arXiv preprint arXiv:2104.10157* (2021).
- [46] 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.
- [47] OpenAI. *Whisper*. <https://openai.com/research/whisper>. Accessed: 2024-02-27. 2024.
- [48] Aaron van den Oord et al. “Wavenet: A generative model for raw audio”. In: *arXiv preprint arXiv:1609.03499* (2016).
- [49] Simon Godsill, Peter Rayner, and Olivier Cappé. *Digital audio restoration*. Springer, 2002.
- [50] 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.
- [51] Stephan R Richter, Hassan Abu AlHaija, and Vladlen Koltun. “Enhancing photorealism enhancement”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.2 (2022), pp. 1700–1715.
- [52] Patrick Esser, Peter Michael, and Soumyadip Sengupta. “Towards Unified Keyframe Propagation Models”. In: *arXiv preprint arXiv:2205.09731* (2022).
- [53] Jean-Pierre Briot. “From artificial neural networks to deep learning for music generation: history, concepts and trends”. In: *Neural Computing and Applications* 33.1 (2021), pp. 39–65.
- [54] Gal Greshler, Tamar Shaham, and Tomer Michaeli. “Catch-a-waveform: Learning to generate audio from a single short example”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 20916–20928.

- [55] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).
- [56] Tom Brown et al. “Language models are few-shot learners”. In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.
- [57] Sercan Arik et al. “Neural voice cloning with a few samples”. In: *Advances in neural information processing systems* 31 (2018).
- [58] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

Appendix A

ADDITIONAL MATERIAL

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