

# POLITECNICO DI TORINO

**Master's Degree in Cinema and Media Engineering**



**Politecnico  
di Torino**



**Master's Degree Thesis**

## **Generative AI for Production Centre**

**Supervisors**

**Prof. Andrea BOTTINO**

**Ing. Roberto IACOVIELLO**

**Ing. Alberto CIPRIAN**

**Ing. Davide ZAPPIA**

**Candidate**

**Gabriele MONTELEONE**

**October 2024**





# Summary

The accelerated advancement of artificial intelligence, in particular generative artificial intelligence (GenAI), has raised extensive discussions in the media and entertainment industry. The Hollywood screenwriters' strikes of 2023 highlight a growing concern about the role AI could take in content creation, raising debates about its potential to equal, or even replace human ingenuity.

GenAI, through machine learning algorithms, enables the autonomous creation of original text, images, video and audio from simple prompts, opening new perspectives for the automation of creative processes.

In this context, this thesis explores the integration of GenAI technologies within a traditional production workflow for the realization of an animated pilot based on Jules Verne's "Around the World in 80 Days", revisited in a futuristic key. The project, developed in collaboration with Rai Centre for Research, Technological Innovation and Experimentation (CRITS), as part of a broader initiative presented at International Broadcasting Convention (IBC) 2024, aims to experiment with the application of GenAI at all stages of audiovisual production, from concept art development to animation.

This project involves a comprehensive analysis of the current generative AI applications available on the market and in the research field, to identify both their strengths and limitations. The workflow combines different tools for scriptwriting, image creation, animation, and sound design, including technologies like Dramatron, ChatGPT, Firefly, Stable Diffusion, Runway, Reespecher and others. Human intervention is crucial at every step to guide the AI and correct its imperfections, such as hallucinations and biases derived from the data on which the generative models were trained.

The findings demonstrate that although AI has the potential to automate many parts of the production process, repeated iterations between humans and AI are essential to achieve the desired quality and consistency, with ongoing refinement of prompts. The latest multimodal models offer more control over the output, allowing for improved consistency issues typical of AI generations. The main goal

of this research is to find a balance between automation and human input, shaping the future of the media and broadcasting industry.

# Table of Contents

- List of Figures*..... III
- 1 Introduction**..... 1
  - 1.1 Context of the Research**..... 1
  - 1.2 Objectives of the Thesis** ..... 5
    - 1.2.1 Case Study: Generative AI in Action ..... 7
- 2 State of the art**..... 9
  - 2.1 Overview of Generative AI models** ..... 9
    - 2.1.1 GAN ..... 9
    - 2.1.2 VAE ..... 11
    - 2.1.3 Transformer ..... 12
    - 2.1.4 Diffusion Model ..... 14
  - 2.2 Generative AI Applications in Media Production** ..... 15
    - 2.2.1 Text Generation ..... 15
    - 2.2.2 Image Generation ..... 19
    - 2.2.3 Video Generation..... 28
    - 2.2.4 Sound Generation ..... 38
- 3 Pre-Production** ..... 43
  - 3.1 Script Development**..... 43
  - 3.2 Concept Art**..... 52
    - 3.2.1 Character Design ..... 53
  - 3.3 Shotlist & Storyboard** ..... 58
- 4 Production**..... 63
  - 4.1 Creating Style Frames** ..... 65
  - 4.2 Creating Style Images with Consistent Characters** ..... 69
  - 4.3 Video Animation** ..... 73
  - 4.4 Dream Machine: Advantages and Limitations**..... 73
  - 4.5 Runway Gen-2: Advantages and Limitations**..... 74

4.6	Kling.AI vs Runway Gen-3: The Next Generation .....	76
5	<i>Post-Production</i> .....	81
5.1	Ai-Generated Voice Over .....	81
5.2	AI-Generated Sound Effects and Music .....	83
6	<i>Video Editing</i> .....	84
7	<i>Conclusions</i> .....	85
7.1	Summary of Results .....	85
7.2	Future Developments and Recommendations .....	86
8	<i>Bibliography</i> .....	89

# List of Figures

Figure 1.1 A creative robot guided by artificial intelligence. ....	1
Figure 1.2 AI-generated image of Pope Francis wearing a branded jacket.....	2
Figure 1.3 AI deepfake of Donald Trump being arrested.....	3
Figure 1.4 International Broadcasting Convention 2024.....	5
Figure 1.5 Diagram illustrating the workflow and tools utilized .....	8
Figure 2.1 Diagram of applications in generative artificial intelligence .....	9
Figure 2.2 GAN model .....	10
Figure 2.3 Autoencoder structure .....	11
Figure 2.4 Variational autoencoder structure.....	12
Figure 2.5 Transformer architecture .....	13
Figure 2.6 Diffusion model process .....	14
Figure 2.7 Examples of two GPT-4 applications finetuned for screenwriting .....	16
Figure 2.8 Hierarchical structure of Dramatron [15].....	18
Figure 2.9 Plotdot’s screenwriting interface.....	19
Figure 2.10 DALL·E’s image generation features inside ChatGPT.....	21
Figure 2.11 Midjourney’s user interface.....	22
Figure 2.12 Midjourney’s generations using generic videogame prompts [20]....	23
Figure 2.13 Comparison of Midjourney’s generations using “Popular movies screencap” prompt vs original copyrighted movie frames [20] .....	24
Figure 2.14 Adobe Firefly’s user interface .....	25
Figure 2.15 Photoshop’s “Generative Fill” feature in action.....	26
Figure 2.16 Comfy.UI’s node system .....	27
Figure 2.17 Fooocus’ user interface .....	28
Figure 2.18 Two different frames from Sora’s demo [22].....	31
Figure 2.19 Adobe Premiere Pro extending a clip using “Generative Extend” feature [23] .....	32
Figure 2.20 “Add Object” feature within Premiere Pro [23].....	32
Figure 2.21 Comparison of two frames before and after Premiere Pro’s “Remove Object” operation [23].....	33
Figure 2.22 VASA-1’s technology [24].....	34

Figure 2.23 Runway’s AI tools .....	35
Figure 2.24 Kling AI’s user interface .....	36
Figure 2.25 Dream machine’s start and end frame feature.....	37
Figure 2.26 Diagram of video-to-audio system [31] .....	41
Figure 3.1 ChatGPT’s concept idea.....	44
Figure 3.2 ChatGPT’s Themes Adaptation.....	45
Figure 3.3 Dramatron’s hyperparameters .....	46
Figure 3.4 Dramatron’s logline.....	48
Figure 3.5 Dramatron’s title.....	48
Figure 3.6 Dramatron’s character description .....	48
Figure 3.7 Dramatron’s Plot Synopsis .....	49
Figure 3.8 Dramatron’s dialogues .....	50
Figure 3.9 Moodboard .....	52
Figure 3.10 Hand-drawn concept art of Phileas .....	53
Figure 3.11 Firefly’s variations on Phileas’ concept art .....	54
Figure 3.12 Image-to-3D model conversion with 3D Studio AI .....	54
Figure 3.13 A. side view of the 3D model. B. Rear view of the 3D model.....	55
Figure 3.14 Concept art of Passepartout by DALL·E .....	56
Figure 3.15 A. Passepartout’s steampunk version B. Passepartout’s legless version .....	56
Figure 3.16 Passepartout’s cuter version .....	56
Figure 3.17 Passepartout’s high-tech minimalist version.....	57
Figure 3.18 Image-to-3D model conversion with Meshy.....	57
Figure 3.19 Polishing of the 3D model in Blender.....	58
Figure 3.20 ChatGPT shotlist generation .....	59
Figure 3.21 Medium shot with main characters of the right proportions but Passepartout’s wrong appearance .....	60
Figure 3.22 Medium shot with main characters of the wrong proportions and Passepartout’s correct appearance .....	60
Figure 3.23 Close-up of Phileas looking surprised.....	61
Figure 3.24 Failed shot with all elements centered .....	62
Figure 3.25 Stylized sketch of the correct composition .....	62

Figure 3.26 Shot obtained using the stylized composition reference .....	62
Figure 3.27 Shot with the incorrect text “hask help” .....	63
Figure 3.28 Failed shot of Phileas submerged underwater.....	63
Figure 4.1 Diverse appearance of Phileas in different generations .....	65
Figure 4.2 Fooocus’ advanced settings.....	67
Figure 4.3 Fooocus’ “Describe” feature .....	67
Figure 4.4 Multiple image prompts with “Stop at” and “Weight” parameters.....	68
Figure 4.5 Police officer before and after inpainted details.....	69
Figure 4.6 Automatic character inpainting mask.....	70
Figure 4.7 Posing models in Blender.....	71
Figure 4.8 Shading set-up for background images .....	72
Figure 4.9 Character compositing in Photoshop .....	72
Figure 4.10 Police robot morphing its appearance .....	73
Figure 4.11 Runway Gen-2 Camera Control.....	74
Figure 4.12 Runway Gen-2 Motion Brush .....	75
Figure 4.13 Runway Gen-2 General Motion parameter.....	76
Figure 4.14 Kling’s generation with subtitles and morphed Passepartout .....	77
Figure 4.15 Continuity failure in Kling’s generation .....	77
Figure 4.16 Correct Runway animation starting from the same input image.....	77
Figure 4.17 Successful animation of the police officer raising his hand.....	78
Figure 4.18 Multiple generations with the wrong number of fingers.....	78
Figure 4.19 Failed character movement on Runway .....	79
Figure 4.20 Inaccurate reconstructions of Phileas’ legs .....	79
Figure 4.21 Phileas stepping out of a bush that initially covered him.....	80
Figure 5.1 Text-to-Speech narration styles.....	81
Figure 5.2 Speech-to-Speech controls .....	82
Figure 5.3 Eleven Labs’ controls .....	83
Figure 6.1 DaVinci Resolve’s timeline.....	84

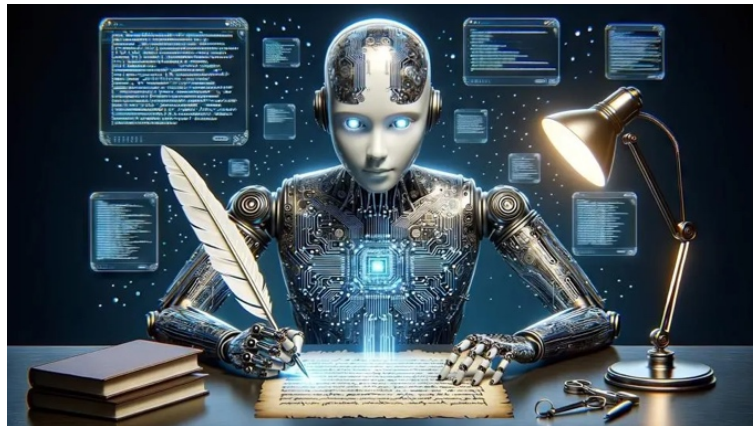




# 1 Introduction

## 1.1 Context of the Research

In recent years, generative artificial intelligence (GenAI) has seen exponential growth and widespread deployment in multiple sectors, starting with the release of advanced models such as ChatGPT in 2022. This evolution has marked a transformative phase, making AI a central tool for automating and improving productive processes in diverse fields such as media, healthcare, education and business. “Generative AI” refers to a category of algorithms designed not only to analyze data but also to create new realistic content such as text, images, video, and audio.



*Figure 1.1 A creative robot guided by artificial intelligence.*

*Source: <https://tinyurl.com/3jjwut9n>*

This technology represents a significant shift in the creative process, opening up new possibilities in the entertainment and media industries. Traditionally, producing high-quality media content required the collaboration of various human professionals: writers, designers, video editors, sound engineers, and more. However, GenAI now offers the potential to handle many of these tasks autonomously, from scriptwriting and rendering to automatically editing images, videos, and audio. This evolution has the potential to reshape media production as we know it.

One of the key advantages of generative AI is its capacity to significantly reduce both production time and costs, a benefit that is particularly advantageous for small studios and independent creators. By streamlining the creative process, GenAI expands the possibilities of content production, democratizing access to high-quality media creation and making it more attainable for a broader range of individuals. This technological advancement empowers creators to respond swiftly to the ever-growing demand for content, driven by the proliferation of streaming services and social media platforms that constantly seek fresh material.

While these platforms are saturated with all forms of media, it's important to note that AI is not always used for positive purposes. For instance, virtual avatars, deepfakes (see Figure 1.2 and Figure 1.3), and voice synthesis technologies are increasingly employed to create viral and often controversial content. These applications can blur the lines between reality and fiction, raising ethical concerns about authenticity and consent in the digital age.



*Figure 1.2 AI-generated image of Pope Francis wearing a branded jacket*

*Source: <https://tinyurl.com/59jkantrn>*



*Figure 1.3 AI deepfake of Donald Trump being arrested*

*Source: <https://tinyurl.com/3jv9v8m5>*

A notable case is the 2023 song “Heart on My Sleeve”, which used AI to imitate the voices of well-known artists Drake and The Weeknd without their consent [1]. This track gained millions of plays on platforms like Spotify, YouTube, and TikTok before sparking legal action, raising important questions about intellectual property and consent in the era of AI-driven creativity.

At the heart of these issues is a growing debate: intellectual property rights have traditionally been reserved for human creators, but the line becomes blurred when machines play a significant role in the creative process. GenAI models are trained on vast datasets that include pre-existing media, learning the characteristics and styles of those works to produce new, similar content. Yet the original creators of those works often receive no compensation and may not have authorized the use of their material in such a way. This raises ethical concerns about authorship and the rights of creators in an AI-driven creative landscape.

Despite these concerns, many sectors of the entertainment industry have begun adopting GenAI as a valuable tool. AI is being used to enhance visual effects and streamline time-consuming processes such as rotoscoping, enabling even small teams to produce high-quality results. A notable example is the Oscar-winning film “Everything Everywhere All at Once” [2], which employed a small team that utilized AI tools like Runway ML and Stable Diffusion to integrate live-action footage with computer-generated imagery seamlessly and obtain hypermodern transitions. This demonstrates the potential for AI to complement human creativity rather than replace it, allowing for innovation in film production while keeping costs manageable.

However, there is also a pervasive fear among professionals in creative industries that AI could eventually replace their roles, which rely on distinctly human attributes like creativity, emotional intelligence, and intuition. The Hollywood writers' strikes in 2023 brought these anxieties to the forefront, as screenwriters and actors expressed concern that AI, if left unchecked, could alter the industry so dramatically that their jobs would be threatened. They feared a future where AI could not only generate scripts but also potentially replace writers altogether, leading to a diminished role for human creativity in favor of machine-generated content. [3]

The negotiations that followed the strikes resulted in a compromise, offering some reassurance to industry professionals. The new rules could be a model for other industries, providing for the regulated use of AI in creative processes. While the agreement does not prohibit the use of AI tools in the writing phase, it sets limits that prevent employers from using technology to replace human workers. Studios cannot use AI to write or edit content developed by writers. Additionally, AI-generated material cannot be considered a source for adaptation, ensuring that screenwriters' roles and compensation for original works are safeguarded.

After three years of development work, the European Union took a position on the issue: on July 2024 the "AI Act" [4] was passed to establish a comprehensive legal framework for the use and development of artificial intelligence technologies, classifying AI systems according to their level of risk, from minimal to unacceptable. This classification system imposes different obligations depending on the level of risk. For generative AI systems, which are considered to be of limited risk, the law introduces transparency requirements, such as requiring that AI-generated content be clearly labelled as such, and safeguards to prevent the generation of illegal content. Additionally, developers are required to publicly provide summaries of the dataset used for training their AI models, thus ensuring accountability and addressing concerns related to copyright and intellectual property rights. Companies will soon have to adapt to these strictures.

In summary, while generative AI holds enormous potential for transforming media and entertainment by making production more efficient and accessible, it also raises complex ethical and professional challenges. As the technology continues to

evolve, it will be critical to establish frameworks that balance innovation with the protection of human creativity and rights, ensuring that AI remains a tool that enhances rather than undermines the creative industries.

## 1.2 Objectives of the Thesis

The primary goal of this thesis is to investigate the potential and limitations of GenAI tools in media production, focusing on how these technologies can be applied to automate traditional production workflows. This research was conducted in collaboration with the Rai “Centre for Research, Technological Innovation and Experimentation” (CRITS), as part of a broader European initiative showcased at the “International Broadcasting Convention” (IBC), held annually in Amsterdam (Figure 1.4). The IBC attracts more than a thousand companies from the entertainment, technology, and media industries, providing a platform for professionals from across the globe to collaborate, share knowledge, and discuss the latest technological advancements. One of the most prominent topics at IBC 2024 is artificial intelligence, particularly its diverse applications in broadcasting and media production.



*Figure 1.4 International Broadcasting Convention 2024*

Source: <https://show.ibc.org/accelerators-2024-challenges>

As part of this larger initiative, I had the opportunity to participate in the “Generative AI in action” accelerator program alongside well-established partners in the broadcasting industry, including the European Broadcasting Union (EBU), YLE, the Finnish national broadcaster, and ITV, the UK’s largest commercial broadcaster. Additionally, the project received support from a diverse range of companies such as IET, Verizon Business, Respeecher, Pluxbox, Plan IX Labs, Somersault, and Xansr Media. The objective of this accelerator was to explore how generative AI could be integrated into a traditional media production process, automating it from concept development to final output. The key challenge was to evaluate the feasibility of AI tools in streamlining and improving the creation of high-quality media content across the various stages of production. GenAI tools available today demonstrate remarkable promise in a variety of areas, including image generation, video animation, and sound design. In theory, these technologies offer the potential to produce complex media content with just a simple text prompt, minimizing human intervention. For instance, a user could generate visual sequences, dialogue, and sound effects for a TV show by simply specifying the desired output. However, the practical reality is more complex. While these tools can produce creative content that rivals human-generated work, they still face significant limitations. One of the main challenges is ensuring coherence across long-form content, whether it be a narrative text or a generated video. AI struggles to maintain a consistent narrative or visual theme over extended periods, leading to inconsistencies in both storytelling and visual representation.

Another key limitation lies in the integration of different AI tools throughout the production process. Each tool, whether for scriptwriting, image creation, or sound design, is trained on different datasets and follows distinct methodologies. As a result, transitioning smoothly from one phase of production to another requires a deep understanding of how these tools function and how to adapt them to work in harmony. This introduces additional complexity, as the current technology still requires human oversight to correct biases, maintain coherence, and ensure the overall integrity of the final product.

The computational demands of GenAI also present a major obstacle. Significant time and computational power are needed to train models to improve their

performance, extend their context windows, and enhance their ability to generate long-form, consistent content. Therefore, the full potential of GenAI in media production has not yet been realized and adjustments are needed to overcome these limitations.

### **1.2.1 Case Study: Generative AI in Action**

In this project [5], we aimed to explore the capabilities of generative AI by creating a pilot episode for a hypothetical TV series inspired by Jules Verne’s “Around the World in 80 Days”. This reimagined adaptation served as a comprehensive test case for assessing how far AI-driven tools can automate various stages of media production, from ideation to the final product.

We began by brainstorming the concept for the series, using large language models (LLMs) such as ChatGPT and Gemini to draft a synopsis and outline the narrative framework. These LLMs helped generate initial story ideas, character arcs, and plot points. Their use significantly sped up the pre-production process by allowing us to generate creative ideas rapidly, which would traditionally require a team of writers and multiple brainstorming sessions. The LLMs acted as collaborators, providing us with a rich array of options, which we then curated and refined to fit our vision for the project.

With the narrative structure in place, we moved on to the visual development phase. AI-powered image generation tools like DALL·E, Adobe Firefly and Stable Diffusion were employed to create concept art, character designs, and storyboards. These tools allowed us to quickly visualize our ideas, generating various art styles and character renditions from simple text prompts. The AI accelerated the creative process by generating visual elements at a much faster rate than human illustrators typically could. However, not every output was achievable, and some frames required numerous attempts. Human intervention was still crucial for curating and selecting the most appropriate designs, ensuring coherence between the characters, setting, and overall tone of the series.

The production phase posed a greater challenge, as it involved synthesizing visual sequences, dialogue, and sound effects. To ensure character consistency throughout the sequence, 3D models were generated from the initial concept art using advanced



AI-based image-to-3D software such as 3D AI Studio and Meshy. These models were then manually posed in front of the environment backgrounds using Blender and Photoshop for compositing. Subsequently, we utilized AI tools like Runway and Kling for video animation and scene generation, which transformed static images into dynamic visual sequences. Reespecher and other AI-driven audio tools were used to generate voiceovers and sound effects, further enhancing the immersion of the pilot episode. This phase was a critical test of how well these generative tools could handle complex tasks like character animation, scene transitions, and voice synthesis. While AI was capable of producing impressive results, certain adjustments had to be made manually, especially in areas like maintaining visual consistency and narrative flow across multiple scenes.

In terms of workload, we estimated that approximately 70% of the production could be managed by AI-driven processes, particularly in the more mechanical or repetitive aspects of content creation. However, traditional manual intervention was necessary for the remaining 30%, particularly in high-level decision-making, quality control, and refinement of the AI-generated outputs. This division of labor between AI and human creativity allowed us to produce a polished pilot episode while testing the boundaries of current AI technology.

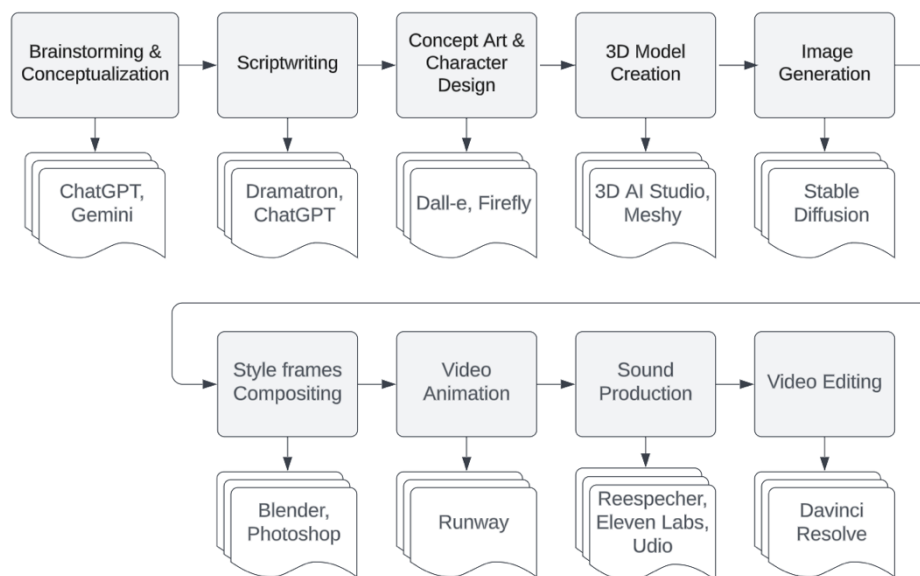


Figure 1.5 Diagram illustrating the workflow and tools utilized



# 2 State of the art

## 2.1 Overview of Generative AI models

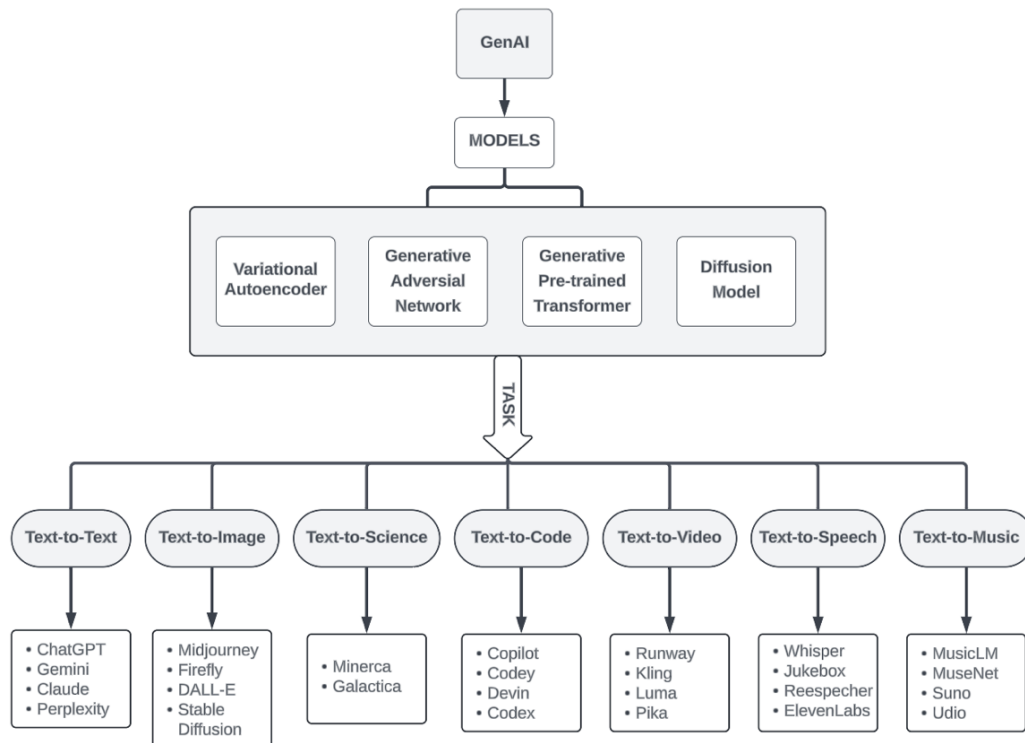


Figure 2.1 Diagram of applications in generative artificial intelligence

The progress that characterizes GenAI have been made possible by the evolution of natural language processing and deep learning algorithms that existed in the 1980s, such as convolutional neural networks (CNNs), which only became a standard in image classification and computer vision in 2010, as the limitations of the available datasets were overcome. Until the introduction of generative adversarial networks (GANs) in 2014, deep learning models were limited to recognizing data patterns and making predictions based on the data available.

### 2.1.1 GAN

With GANs, the center is on the probability distribution of the data: by analyzing the training datasets, models are able to learn their structure and generate new similar and consistent data samples.

GANs are a class of unsupervised learning models, consisting of two neural networks, a generator and a discriminator, which engage in a competitive process. The generator is tasked with creating new, “fake” data, while the discriminator evaluates whether the data is real or fabricated. These adversarial dynamic drives both networks to improve iteratively. The generator starts with random noise as input and gradually learns to produce outputs that closely resemble real data. Meanwhile, the discriminator tries to become increasingly effective at distinguishing real data from the generator’s fabricated outputs. As training progresses, both models improve their performance until the generator’s data becomes nearly indistinguishable from the actual data. Figure 2.2 shows a diagram of the model.

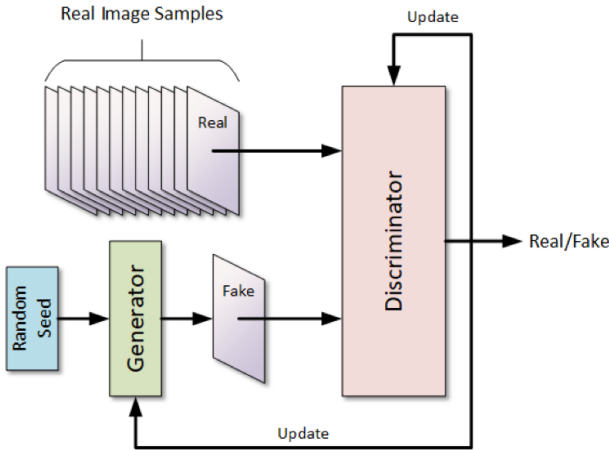


Figure 2.2 GAN model

Source: <https://tinyurl.com/2s3kcxr9>

The evolution of GANs has led to numerous variants, each designed to address specific challenges or improve performance in certain contexts. Notable among these are Progressive GAN [6], BigGAN [7], StyleGAN [8] [9], and CycleGAN [10], which have significantly expanded the capabilities of generative models. These iterations have addressed critical issues such as mode collapse, training instability and fidelity of generated images, leading to increasingly sophisticated and realistic results. The impact of GANs extends beyond image generation, finding applications in fields as diverse as image-to-image translation and resolution enhancement. However, despite these innovations, achieving consistent

results and addressing the inherent challenges of training GANs remain active areas of research. [11]

## 2.1.2 VAE

Another important architecture in the field of generative models is the variational autoencoder (VAE). While GANs rely on a competitive dynamic between a generator and a discriminator to produce realistic data, VAEs take a probabilistic approach, focusing on learning a latent representation of the data. Unlike traditional autoencoders (Figure 2.3), which compress data into a fixed lower-dimensional representation and then reconstruct it, VAEs introduce a probabilistic element to the encoding process.

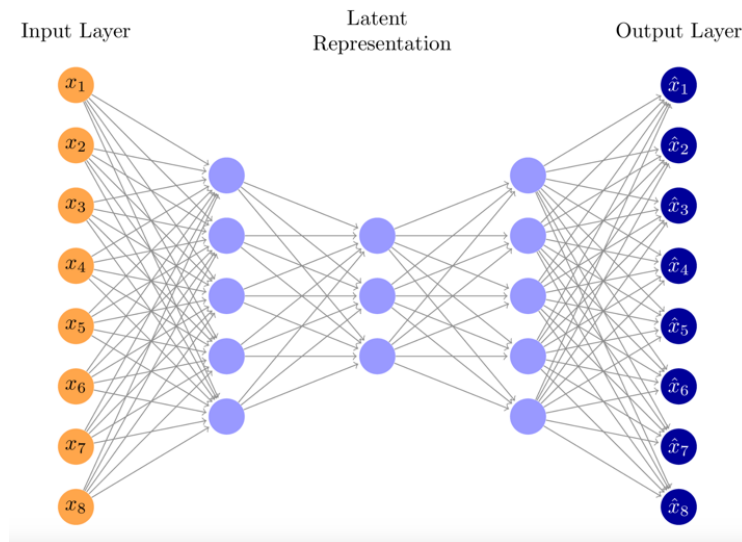


Figure 2.3 Autoencoder structure

Source: <https://tikz.net/autoencoder/>

Specifically, the encoder in a VAE maps the input not to a single latent point but to a distribution, characterized by a mean and variance (Figure 2.4). This allows the decoder to sample new data points from the latent distribution, giving VAEs the ability to generate original data based on the patterns learned during training process.

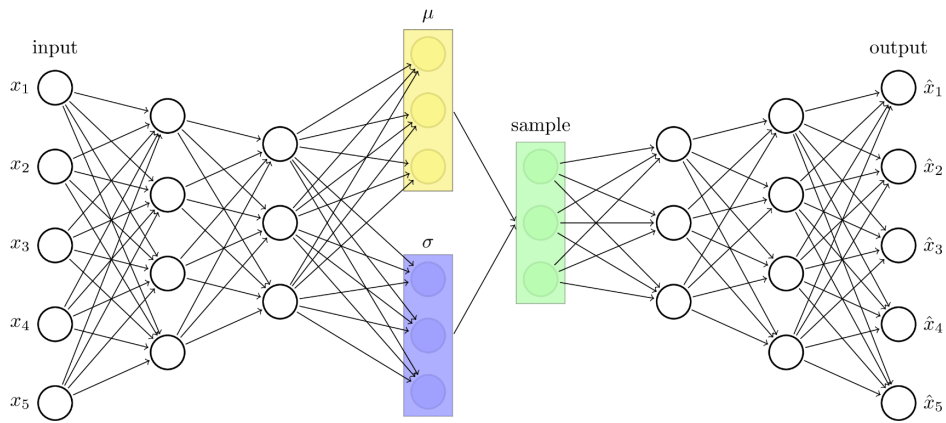


Figure 2.4 Variational autoencoder structure

Source: <https://tinyurl.com/44hjd5pe>

The VAE architecture consists of two main components: the encoder, which compresses the input into a latent distribution, and the decoder, which generates reconstructed data by sampling from this distribution. This allows the model to generate new alternative samples by sampling from various points in the latent space, rather than simply reproducing what it has seen during training. In order to avoid over-fitting, regularization techniques are applied, which help the model to generalize better to the unknown data.

VAEs are implemented in applications such as image generation and unsupervised learning of complex latent structures. DALL·E, one of the first state-of-the-art image generation tools, is developed on this class of algorithms. However, the quality of the data VAE generate, particularly images, tends to be lower compared to other generative models like GANs. This is mainly due to the trade-off between accurate reconstruction and enforcing a regular latent distribution, which can limit the model's ability to capture fine details. Variants like Vector Quantized Variational Autoencoders (VQ-VAE) combine the strengths of autoencoders with vector quantization techniques, allowing for more discrete and higher-quality latent representations. [12]

### 2.1.3 Transformer

The transformer model, developed by Google in 2017, revolutionized NLP by addressing the limitations of earlier models like CNN and RNN. One of its key

innovations is the self-attention mechanism, which enables the model to assess the relevance of different parts of an input sequence independently of their order. This ability makes Transformers particularly adept at capturing long-range dependencies between words, significantly improving performance in tasks such as text generation and translation. [4]

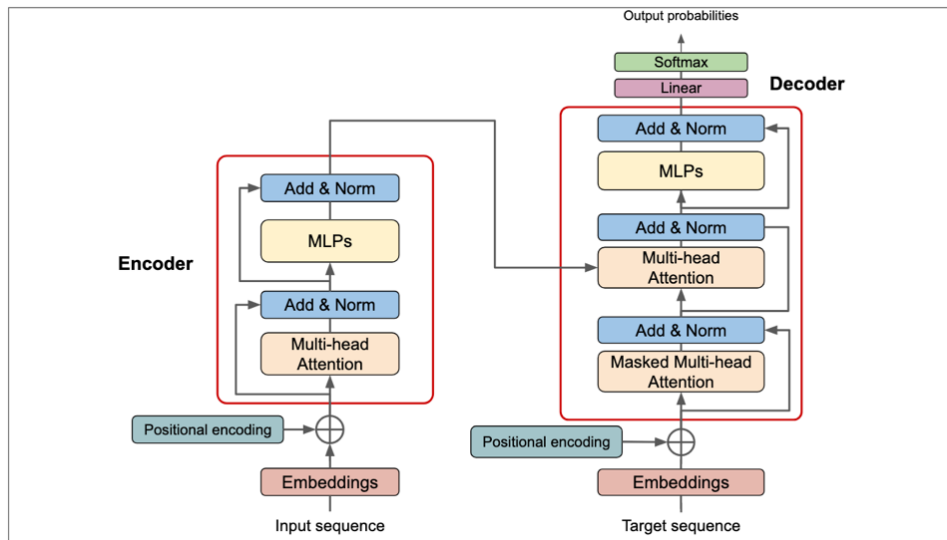


Figure 2.5 Transformer architecture

Source: <https://tinyurl.com/2p8sky3>

The Transformer architecture shown in Figure 2.5 follows an encoder-decoder framework, where the encoder processes the input data, and the decoder generates the output. Multi-head attention mechanisms allow the model to evaluate the input from different perspectives, while feed-forward layers enable the learning of complex data representations. This design has paved the way for more advanced models, such as GPT (Generative Pretrained Transformer), which extended the Transformer’s capabilities to automatic text generation.

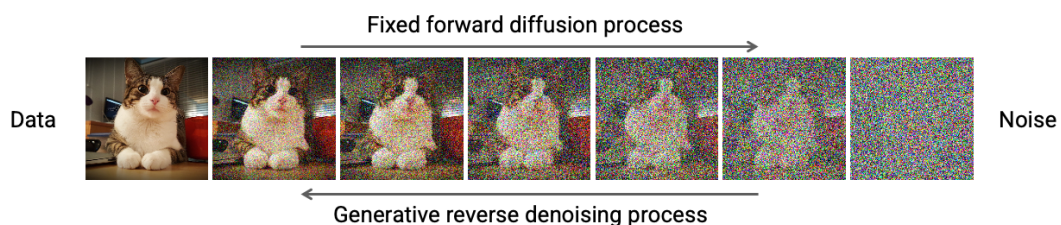
One of the most prominent applications of the Transformer model is ChatGPT, developed by OpenAI in 2018. ChatGPT uses a “decoder-only” architecture, trained on vast amounts of text data from diverse sources, to autonomously produce coherent and contextually relevant text. Its ability to understand and generate text by learning patterns and relationships between words allows it to excel in tasks such as answering questions, creating narratives, and performing various language-

related functions. With the introduction of GPT-4, these capabilities have been further enhanced, as the model can now process both text and images, expanding its potential applications in generative AI.

The success of Transformer-based models like GPT has driven significant advancements in AI research, contributing to better language understanding, generation, and the development of multimodal systems. These systems allow for output in a variety of formats independently from the input that is provided by the user. In addition to text-to-text interactions, GenAI tools enables other types of interactions such as text-to-image, text-to-video, image-to-video, text-to-speech, text-to-3D and others enriching or simplifying the prompt design process of generated content.

### 2.1.4 Diffusion Model

An influential neural network architecture that is driving development within the GenAI is the diffusion model proposed diffused by OpenAI in 2020. The diffusion model operates through two main operations: the forward diffusion process, which adds gaussian noise to data until it becomes random noise, and then reverse generation process, which progressively denoises this random noise to create a new sample data (Figure 2.6). [5]



*Figure 2.6 Diffusion model process*

*Source: <https://tinyurl.com/3ppth2ur>*

Diffusion models present several advantages over GANs, such as enhanced training stability and the capacity to generate a wider variety of samples. Specifically, Diffusion Models with Denoising Prior (DDPM) optimize the process by focusing on predicting the added noise rather than the images themselves, simplifying the learning problem and improving efficiency. Furthermore, these models offer

significant flexibility and control in data generation, enabling users to adjust various parameters like noise level within the diffusion process for extensive customization and creative output. However, the training phase of diffusion models is generally slower than that of GANs, as it necessitates multiple forward passes to reconstruct an image.

## **2.2 Generative AI Applications in Media Production**

The landscape of GenAI applications is rapidly evolving, capturing the attention of creatives across various industries. Today, numerous tools harness the power of these advanced algorithms, each offering a unique approach to media production, especially in textual content creation. Industry leaders such as OpenAI have set benchmarks, inspiring a multitude of emerging platforms that follow similar methodologies but provide varied results and user experiences. This competition has spurred an impressive pace of innovation, with tools constantly updating and improving.

While some GenAI applications are still in the research phase or awaiting commercialization, many are already accessible to the public, though typically through freemium models. These applications often provide a limited trial version, with full features available only through subscriptions or the purchase of credits. Economic considerations thus play a significant role in the selection of tools for professional media production.

### **2.2.1 Text Generation**

Focusing on the most promising generative AI technologies for language processing and scriptwriting, here are some tools that have already begun to reshape the creative workflow:

- ChatGPT: it has emerged as one of the most versatile and widely adopted generative AI tools, particularly within creative and professional settings. As an interactive language model, it excels in tasks requiring written content creation, such as idea generation, brainstorming, drafting, and even refining existing text. Its intuitive interface and capacity for producing coherent, contextually appropriate responses to a wide variety of prompts have made it indispensable

for many industries. What sets ChatGPT apart from traditional text-processing tools is its ability to simulate a genuine dialogue, making it a crucial companion in iterative workflows where creativity and revision go hand in hand.

For this project, ChatGPT initially played a foundational role in shaping the early stages of content creation. It was used to outline narratives, experiment with different tonalities for the storyline, craft dialogues and define a shotlist. Initially, version 3.5 was used in the project; however, it was later upgraded to version 4.0 as it became available for free [13]. The expanded model, marked by “omni” capabilities, greatly enhanced the system’s ability to handle more nuanced and complex creative demands. It allowed for richer human-computer interaction and more refined text generation, making the writing process smoother and more interactive. Additionally, ChatGPT 4.0’s performance improved in areas like context retention over longer conversations, which was particularly beneficial when developing longer-form content.

ChatGPT Plus subscription plan allows users to go beyond the standard conversational model. The premium plan introduces the ability to develop custom GPTs, tailored specifically to individual needs or industries. These custom models can be fine-tuned to a specialized context such as scriptwriting by training them with more focused input rather than relying on the broad general knowledge that powers the standard version (Figure 2.7).

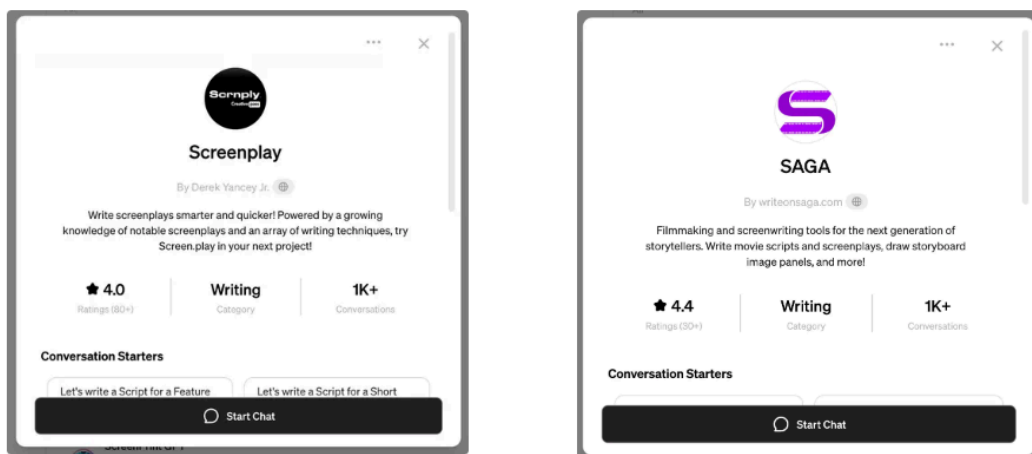


Figure 2.7 Examples of two GPT-4 applications finetuned for screenwriting



- Gemini: developed and released by Google in 2023 [14], it represents the company's step into advanced language models, following the evolution of Bard. Gemini stands out for its deep integration with Google's vast data infrastructure, combining sophisticated language processing abilities with Google's powerful search capabilities. Gemini excels at reasoning and providing more detailed, data-driven answers, particularly useful for research-intensive tasks. This makes Gemini ideal for projects where accuracy, precision and context are critical.

In this project, Gemini was used as a third-party assistant to critically evaluate the strengths and weaknesses of multiple storylines developed using different AI tools. By analyzing the creative output of these tools, Gemini helped identify key elements, such as plot coherence, character development, and motivation, that informed the final selection of the storyline. This evaluative role provided an external perspective, helping the team to refine and focus their narrative ideas, ensuring that the chosen storyline was not only compelling but also well-structured and aligned with the project's goals.

- Dramatron: Designed by Google DeepMind in 2022, Dramatron aims to tackle one of the major challenges in generative text tools: maintaining coherence over long narratives. [15] Unlike traditional LLMs that struggle with extended text sequences, Dramatron generates hierarchical story structures, starting from a simple logline and expanding into characters, locations, and events, all the way to a fully developed script (Figure 2.8). This makes it particularly useful for scriptwriters working on complex stories, as it provides a clear framework for narrative progression. Dramatron operates through Google Colab and requires the implementation of a paid LLM API, such as Google Gemini, OpenAI's GPT, Mixtral, or a custom language API. The tool's user interface allows for creative freedom, with the ability to modify or regenerate content dynamically, making it an ideal tool for collaborative storytelling projects.

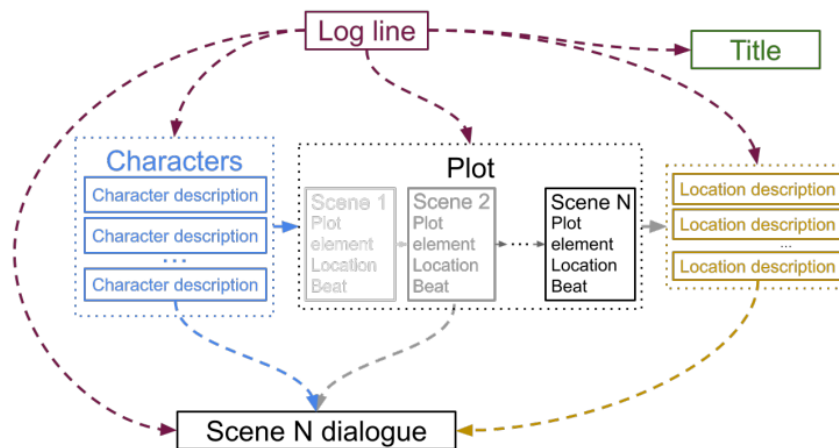


Figure 2.8 Hierarchical structure of Dramatron [15]

- Plotdot: Released officially in July 2024 [16], Plotdot is a more specialized assistant for scriptwriting and character development. What makes it unique is its ability to integrate with various LLMs, such as GPT-3.5, GPT-4.0, Claude-3 Opus, and Claude-3.5 Sonnet, depending on the user’s subscription plan. Its credit-based system, metaphorically referred to as “ink”. Users can purchase credits to access advanced features, keeping in mind that each creative process consumes a set amount of ink. In contrast to simpler writing tools, Plotdot requires detailed input on various aspects of the story, including audience demographics, thematic elements, genre, and character traits. This results in a more structured, step-by-step approach to scriptwriting, allowing users to craft well-developed narratives (Figure 2.9). At each stage, Plotdot can assist by providing suggestions or even generating images to visualize the textual descriptions, adding a multidimensional aspect to the writing process.

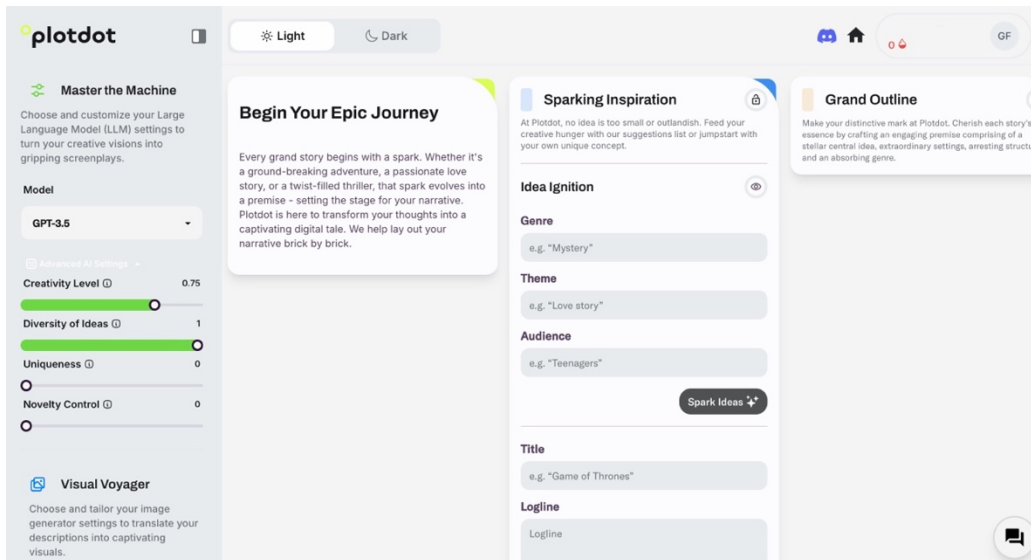


Figure 2.9 Plotdot's screenwriting interface

## 2.2.2 Image Generation

The process of visual generation is one area where artificial intelligence truly demonstrates its creative potential and efficiency. Traditionally, creating high-quality visual content, whether hand-drawn illustrations or computer-generated images, requires immense skill, time, and dedication. However, with the advent of generative AI, it is now possible to produce impressive, comparable digital artworks in mere seconds. This technological leap is revolutionizing the media production landscape. Instead of seeking inspiration through platforms like Pinterest, Behance, or ArtStation, or spending hours combing through vast stock image catalogs, creators can now generate custom, ad-hoc visuals simply by describing what they envision. In essence, GenAI brings ideas to life, offering multiple iterations or variants of an image quickly, allowing for rapid previews and enabling creators to decide on a final version in far less time than traditional methods.

That said, generative AI does not replace the role of artists. While AI can simulate creativity, its capabilities are inherently limited by the data on which it was trained. For particularly complex, unique, or richly detailed concepts, AI may struggle to produce satisfactory results. Artists must still rely on iterative processes: refining their prompts, experimenting with different models, or making manual edits to

achieve the desired outcome. In some cases, AI tools can act as collaborators, not replacements, for the human creative process.

Even platforms widely used by graphic designers and illustrators have begun integrating AI to enhance their workflows. A notable example is Adobe Photoshop, which introduced features like “Generative Fill” [17] and “Generative Expand” [18], allowing users to quickly extend or modify images based on AI-generated suggestions. These functions enhance image editing by filling in gaps or expanding visuals, pixel by pixel, while maintaining context and visual coherence. This marks a new era of fast, AI-enhanced editing, fundamentally changing how images are manipulated and composed. Adobe’s broader suite, including Premiere and other creative tools, is being upgraded with AI functionalities through its Firefly system. Though many of these features are still in development, the early demonstrations have shown great potential for media professionals. [19]

The most notable applications of image generation include the following:

- DALL·E: developed by OpenAI, it is a pioneering platform in the realm of generative AI for images, known for its ability to convert detailed text prompts into sophisticated and visually accurate images. Originally introduced in 2021, the tool quickly gained popularity for its ability to generate artistic, photorealistic, and abstract images alike. The current version, DALL·E 3, builds upon its predecessors by significantly improving its capacity to understand nuanced and complex text descriptions. This advancement allows for more faithful translations of user inputs into creative visual outputs, offering enhanced precision in the rendering of intricate details and compositions.

One of the key differentiators of DALL·E is its tight integration with OpenAI’s ChatGPT platform (Figure 2.10), particularly in the ChatGPT Plus subscription. This feature enables seamless transitions between text-based conversation and image generation. Users can not only request specific visuals but also engage in an interactive refinement process, discussing and altering images as one might in a dialogue with a human designer. This conversational flexibility sets DALL·E apart from many other tools, as it merges the capabilities of natural

language understanding and image synthesis into a fluid iterative process. Additionally, DALL·E 3 introduces expanded format options such as “wide” (1792x1024) and “tall” (1024x1792) aspect ratios, overcoming the square format constraints of earlier versions (1024x1024). These enhancements make DALL·E particularly suited for creative professionals seeking a balance between ease of use, interactivity, and the production of high-quality, versatile imagery.

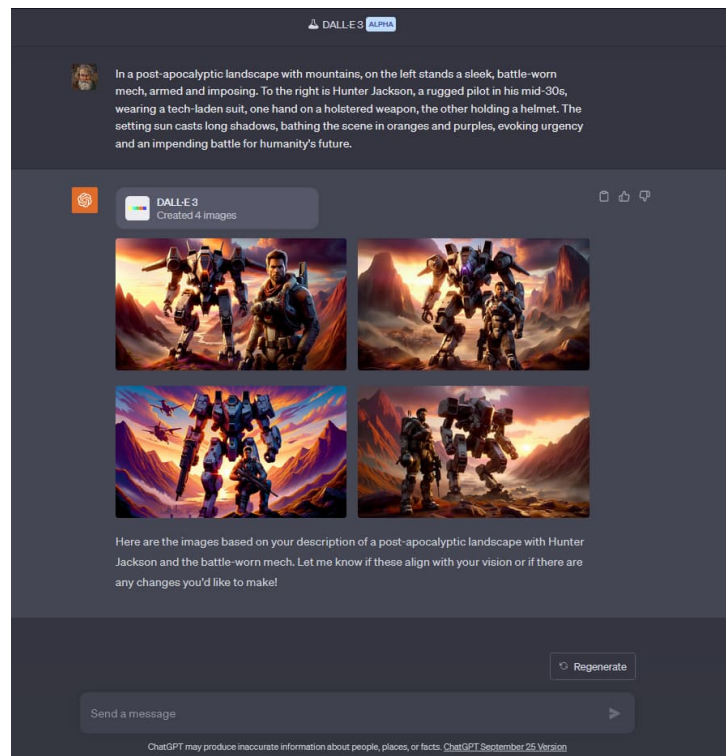


Figure 2.10 DALL·E’s image generation features inside ChatGPT

Source: <https://tinyurl.com/45ss8cff>

- Midjourney: it stands out as one of the most artistically acclaimed generative AI platforms, known for producing some of the highest-quality and aesthetically refined images available in the field. Launched in 2022, Midjourney operates through a Discord server interface, where users can generate images collaboratively in real time, often engaging with a global community of creators. This open and social environment fosters both innovation and inspiration, as users can observe the creative processes of others, share their results, and experiment with various input methods.

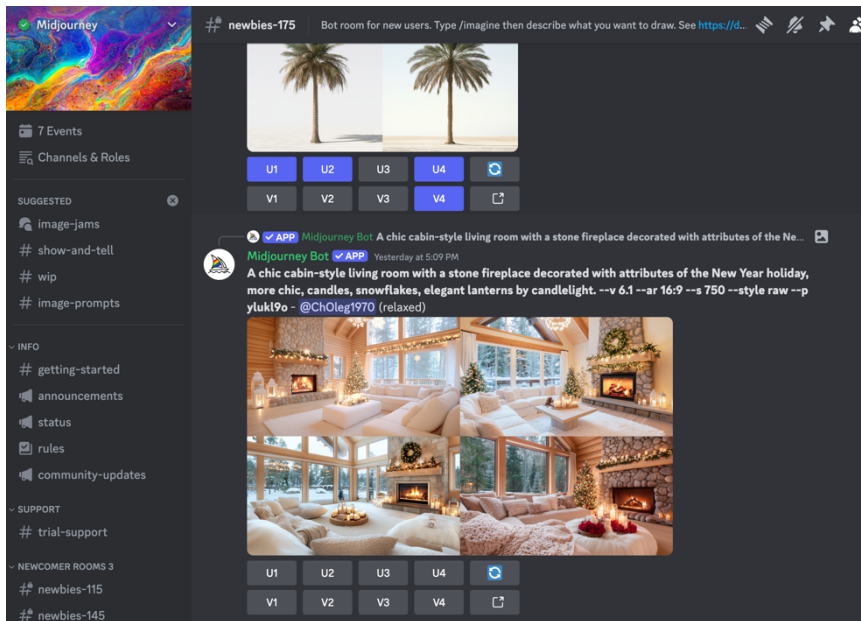


Figure 2.11 Midjourney's user interface

Unlike more casual platforms, Midjourney is often favored by professional artists and designers due to the higher degree of control it offers over image generation. Users can input specific technical parameters, such as in the Figure 2.11 stylization values (e.g. “--s 750” “--style raw”), and model versions (“-v.6”), allowing for meticulous customization of the final product. It also features advanced tools like repeaters and in-painting. This level of detail may present a steeper learning curve, but for those versed in the language of digital art, Midjourney offers unmatched precision and flexibility.

In terms of visual style, Midjourney is often praised for its ability to render highly detailed, visually striking images, ranging from photorealism to abstract and surreal art. This makes it a preferred choice for projects that require a unique aesthetic touch, distinguishing it from tools that prioritize speed or simplicity over creative depth.

However, Midjourney was not used in this project, as there is no certainty on which data it was trained and a recent investigation provides evidence that the GenAI model can produce outputs infringing on copyright, even when users do not explicitly request such results.

The research [20] was conducted by visual artist Southen R. and Marcus G., who experimented with prompts related to commercial films and characters. Through various trials, they found that Midjourney could produce outputs nearly identical to existing copyrighted materials without needing explicit requests to recreate those specific works. For instance, the team discovered that, when prompted with vague instructions or even single words like “screencap”, Midjourney generated images that closely mirrored scenes from iconic movies as well as from various video games. (Figure 2.12 and Figure 2.13)



Figure 2.12 Midjourney’s generations using generic videogame prompts [20]





Figure 2.13 Comparison of Midjourney’s generations using “Popular movies screencap” prompt vs original copyrighted movie frames [20]

This capability indicates that the AI has been trained on copyrighted materials, potentially without appropriate licensing. If users can produce infringing materials unintentionally, they may be exposed to copyright infringement claims. A critical concern is Midjourney’s lack of transparency regarding its training data. The company has not disclosed whether it has obtained licenses for the copyrighted materials that may have been included in its training set. This ambiguity poses risks for the company and its users, especially if the source material is found to be unlicensed. Subsequently, Midjourney revised its terms of service to explicitly prohibit actions that could violate the intellectual



property rights of others, indicating an acknowledgment of the legal risks associated with their outputs.

- Firefly: it represents Adobe’s strategic move into the AI-driven creative landscape, aimed at enhancing its existing suite of professional design tools. While relatively new compared to other platforms, Firefly integrates deeply into the Adobe ecosystem, offering users familiar with Adobe products like Photoshop, Illustrator, an AI-powered extension to their existing workflows.

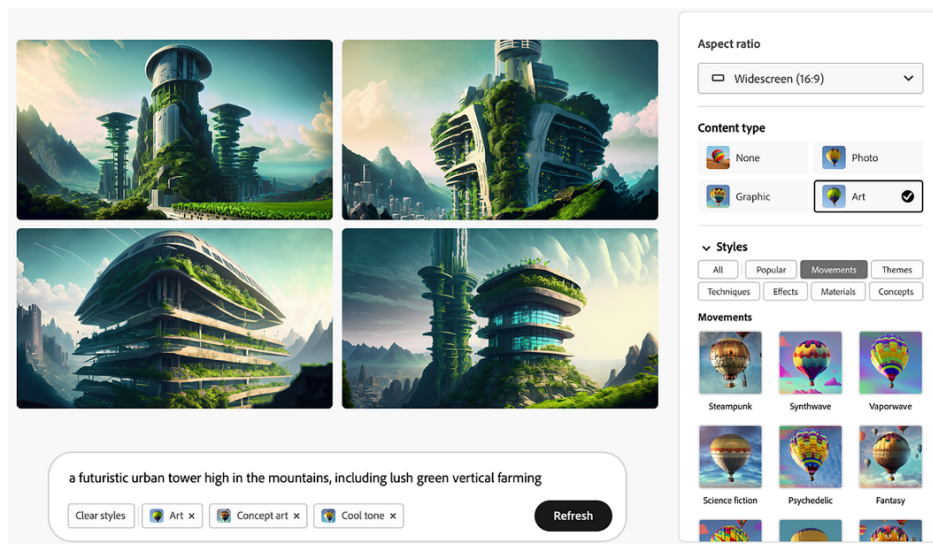


Figure 2.14 Adobe Firefly’s user interface

Source: <https://tinyurl.com/3yve43kx>

Firefly’s web app, shown in Figure 2.14, offers a variety of effect presets, allowing users to decide various aspects of the generated images, such as illumination, camera angles, and color schemes. Users can also upload a reference photo to guide the composition or style of the generated imagery. Unlike other generative tools that prioritize simple image creation, Firefly emphasizes high-quality professional editing. It focuses on enhancing, refining, and improving visuals using AI capabilities. For example, a user can expand portions of an image or remove and generate objects in real time (Figure 2.15), all while preserving the overall composition and visual fidelity of the original artwork. This makes Firefly particularly valuable for tasks like rapid

prototyping and content refinement, bridging the gap between manual artistry and automated generation.

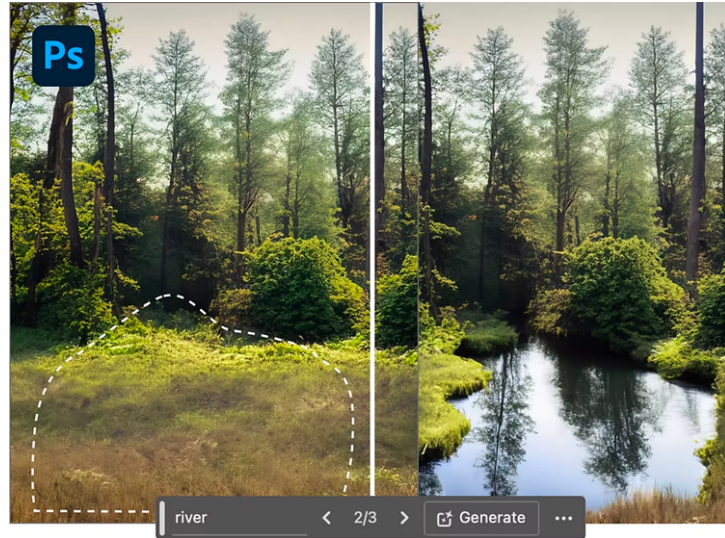


Figure 2.15 Photoshop's "Generative Fill" feature in action

Source: <https://tinyurl.com/4mwrxt9b>

What distinguishes Firefly is its training methodology, utilizing Adobe's extensive Stock database, which comprises high-quality content. This approach ensures that Firefly does not rely on user-generated content or materials from the web; instead, it operates solely on assets for which Adobe has secured permissions. This focus on licensed content not only ensures the quality of the output, but also addresses legal and ethical concerns around copyright.

- **Stable Diffusion:** developed by Stability AI and released in 2022, it stands out as one of the most versatile and widely accessible open-source solutions for generative image creation. Unlike other proprietary platforms, Stable Diffusion offers an unparalleled level of customization and flexibility, allowing developers, artists, and researchers to fine-tune its underlying models for their specific requirements. This open architecture enables users to adjust model parameters, retrain the system with custom datasets, and integrate it into diverse applications, from artistic rendering to specialized domains like medical imaging.

One of the distinguishing factors of Stable Diffusion is its adaptability across various platforms, such as Comfy.UI, which uses a node-based interface to offer intricate control over image generation workflows. Though more complex to use than platforms based primarily on textual prompts, Comfy.UI allows for highly tailored configurations, making it ideal for advanced users seeking deep customization. Below, in Figure 2.16, is a snapshot of its interface.

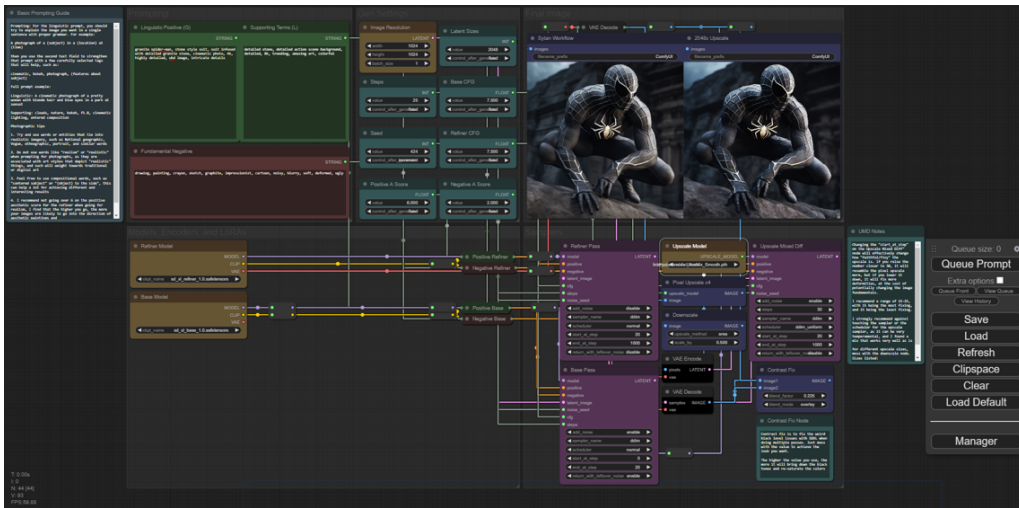


Figure 2.16 Comfy.UI's node system

Source: <https://rundiffusion.com/comfyui-workflows>

While its diffusion-based architecture is common to many generative models, what sets Stable Diffusion apart is its open-source nature and extensive community-driven support. This has led to the development of specialized models for tasks as diverse as anime-style art, photorealism, and fine art creation. In the context of this project, the “Fooocus” platform [21], which will be discussed in detail later in chapter 4.1, was preferred because of its practicality and enhanced control. (Figure 2.17)

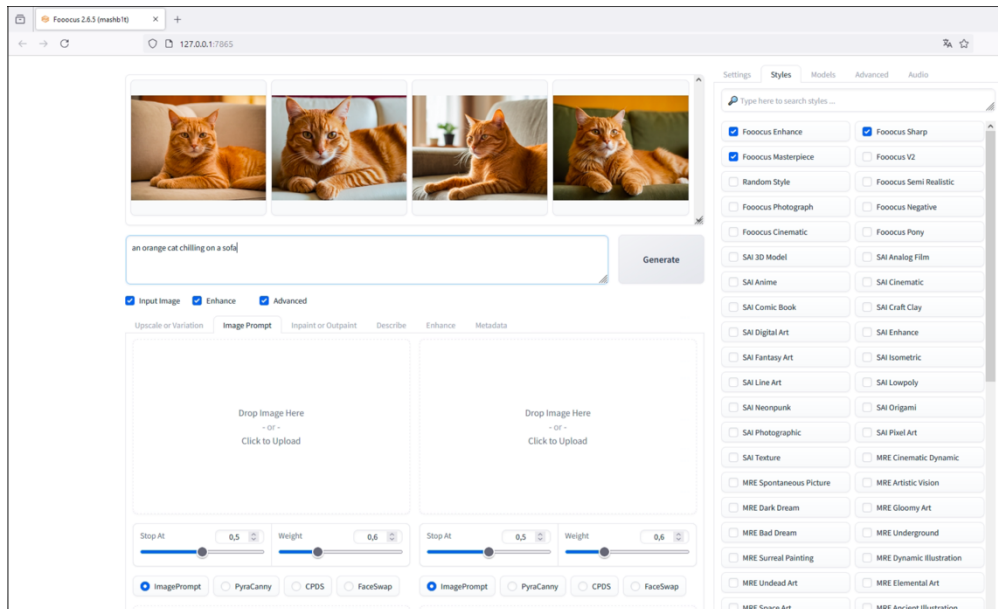


Figure 2.17 Fooocus' user interface

Stable Diffusion's technical robustness and adaptability make it an invaluable tool for users seeking both creative freedom and high-quality, fine-tuned outputs. However, unlike web-based apps that operate through cloud servers, Stable Diffusion must be run locally on a computer. This requires downloading substantial model checkpoints, such as Stable Diffusion XL, and operating them on a machine equipped with a high-performance graphics card. This local deployment offers more control over the model but also places higher demands on computing resources, making it less accessible for users without advanced hardware setups.

### 2.2.3 Video Generation

The field of video generation is undergoing rapid and transformative evolution. Since the early days of cinema, we've known that videos are created by sequencing multiple frames at a speed of at least 24 frames per second to achieve fluid motion. This technical foundation helps explain why early AI-generated videos have been limited to a few seconds in length and relatively low in resolution. Generating realistic, long-duration video content requires significant processing power due to the complexity of encoding and decoding algorithms, which must account for both

the spatial dimensions (x, y) of each frame and the temporal dimension (time) to predict pixel movement between frames.

Challenges in generating single frames, such as maintaining coherence and visual realism, become even more pronounced in video generation, where inconsistencies are more visible over time. For instance, issues in pixel prediction often result in unnatural movements or unrealistic physical interactions in longer videos. Despite these limitations, continuous advancements in generative AI technologies are pushing the boundaries of what is possible in video production. Various industries, from advertising to entertainment, are actively experimenting with AI-driven animation tools, yielding promising results.

User input in video generation can take several forms: text-to-video, image-to-video, or video-to-video. Text-to-video generation offers the AI greater creative freedom, as it generates the entire sequence from scratch based on textual prompts. In contrast, image-to-video and video-to-video approaches provide the AI with a visual foundation, requiring it to first interpret the input content and then animate it, ensuring consistency in graphical elements or altering the style based on user preferences (e.g. transforming live-action footage into an animated comic style). However, not all generative AI platforms currently offer these advanced features, and many are still in the experimental or development phase.

Among the unreleased models are Sora, the Firefly Video Model, and VASA (a research project that is not intended for public release):

- Sora: In March 2024, OpenAI unveiled its ambitious generative AI model aimed at revolutionizing video creation, though the model itself has not yet been publicly released [22].

Sora represents a new frontier in AI-driven content generation, particularly in the domain of video, by building on the advances in text-conditional diffusion models. Unlike existing systems that focused on a narrow range of visual data or short, fixed-size video clips, Sora is designed as a generalist model capable of producing high-quality videos lasting up to one minute. Its unique strength

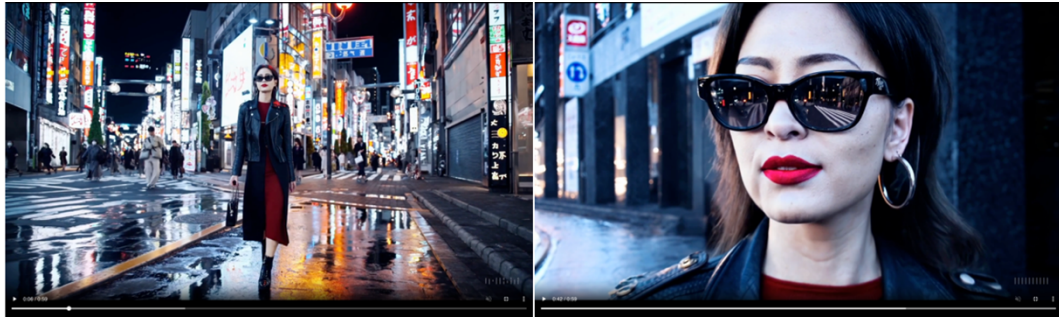
lies in its ability to generate content across a wide spectrum of resolutions, durations, and aspect ratios, offering unmatched versatility.

The architecture of Sora takes inspiration from large language models by turning video frames into spacetime patches, small units of visual data that can be processed much like tokens in text models. This allows the model to seamlessly handle both images and videos in a unified way, making it adept at generating content that ranges from static images to complex, dynamic sequences. What really makes Sora stand out is how well it scales. By training on large datasets of diverse video content, Sora demonstrates emergent capabilities that extend beyond simple generation. These include 3D consistency, long-range coherence, and object permanence. Sora is able to simulate realistic camera motions and even depict interactions within digital and physical worlds in a highly coherent manner. It can generate videos where objects remain consistent even as they move off-screen and return, or simulate environments with dynamic elements that evolve over time.

Below, in Figure 2.18, are two frames from a video sequence generated by Sora, showcasing the model's ability to preserve character details across varying shot types: a wide shot and a close-up. The character's appearance remains consistent, with careful attention to fine details such as sunglasses, earrings and facial features. Notably, Sora excels in rendering realistic lighting effects, as seen in the shadows and reflections on the wet surface. Even the background elements, like the people, are sharply defined rather than blurred, indicating the model's capacity for generating high-fidelity visuals across different layers of the scene.

Prompt: "A stylish woman walks down a Tokyo street filled with warm glowing neon and animated city signage. She wears a black leather jacket, a long red dress, and black boots, and carries a black purse. She wears sunglasses and red lipstick. She walks confidently and casually. The street is damp and reflective, creating a mirror effect of the colorful lights. Many pedestrians walk about"





*Figure 2.18 Two different frames from Sora's demo [22]*

Sora's capabilities go beyond passive video generation; the model also proved to be capable of interacting with digital environments, such as simulating video gameplay and scenarios of "Minecraft". In these cases, Sora not only generates the visuals but also controls the in-game character's actions, offering a glimpse into its potential as a comprehensive simulator for both physical and virtual worlds. Of course, Sora's presentation, while promising, still reveals certain limitations. It struggles to accurately simulate interactions such as breaking glass or eating food, and often fails to capture the proper changes in object states during dynamic scenes. These issues highlight areas that require further refinement. The development team is actively working on improving the model to overcome these challenges and enhance its overall reliability before making Sora available to the public. Its ability to create complex, high-fidelity video from simple text prompts signals a future where AI-generated video content could become as mainstream as AI-generated text or images.

- **Firefly Video Model:** as introduced in the previous chapter, it represents Adobe's latest effort to incorporate generative AI into the audiovisual sector. [23] Building upon its success with image generation, Firefly aims to create original video clips from text prompts while also enhancing traditional editing workflows in industry-standard tools like Adobe Premiere Pro. This new model, set to be released soon, is designed to seamlessly integrate with existing video editing processes, offering features that significantly streamline tasks. The Premiere Pro team has worked closely with the professional video editing community to identify specific challenges that AI could address. Editors have

emphasized three key needs: seamless integration of AI into their existing workflows, particularly with tools like Premiere Pro; faster, more efficient content editing through generative AI and finally, the flexibility to choose between different AI models that excel in various tasks such as Sora, Runway and Pika within Premiere Pro.

Among the new features, the “Generative Extend” tool (Figure 2.19), the equivalent of the Generative Expand shown for images, enables editors to lengthen shots by generating new frames at the beginning or end of a clip, allowing for smoother transitions and edits.

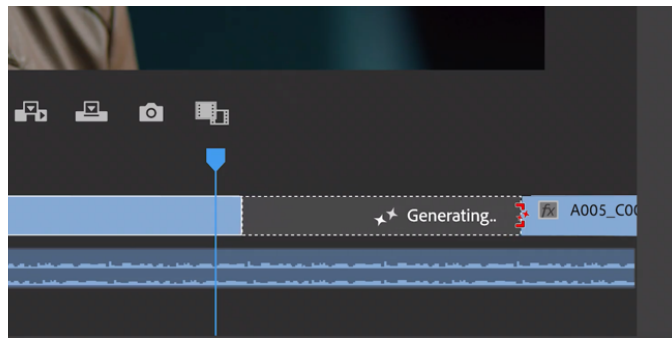


Figure 2.19 Adobe Premiere Pro extending a clip using “Generative Extend” feature [23]

Additionally, object manipulation tools make it easy to add or remove elements from a video (Figure 2.20 and Figure 2.21), such as removing a boom mic or adding set decorations. The “Generative B-Roll” feature, moreover, allows for the quick creation of supporting footage through simple text prompts, eliminating the need to hunt for stock videos.



Figure 2.20 “Add Object” feature within Premiere Pro [23]



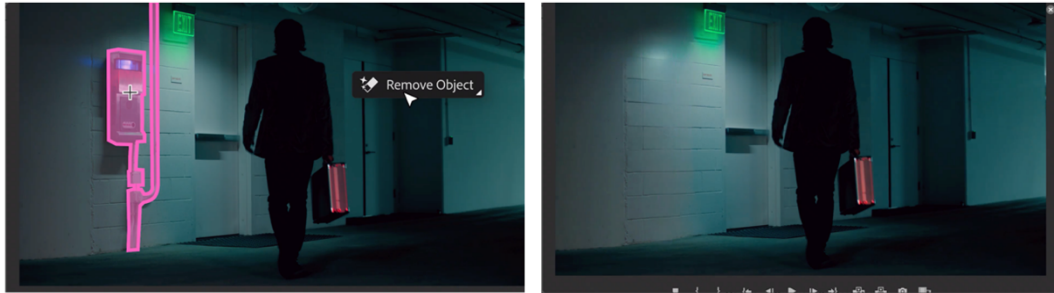


Figure 2.21 Comparison of two frames before and after Premiere Pro's "Remove Object" operation [23]

This level of automation and creative control exemplifies Adobe's vision for how generative AI can reduce tedious tasks while enhancing creative flexibility. Through these advancements, Firefly stands to revolutionize video editing, delivering powerful tools that allow editors to work more efficiently while maintaining full creative control.

- VASA-1: it is a research project developed by Microsoft [24], that propose a different workflow, offering a groundbreaking framework that animates a static portrait based on an audio signal. This model generates 512x512 video frames where not only the subject's lip movements are perfectly synchronized with the recorded speech, but it also captures a wide range of facial expressions and head movements, creating an overall lifelike appearance. Beyond simple lip synchronization, VASA-1 can animate microexpressions such as eye movement, eyebrow raises, and subtle facial dynamics that add a layer of realism to the generated video. Some examples of VASA-1's work are shown in Figure 2.22 below.



Figure 2.22 VASA-1's technology [24]

One of the standout features of this model is its ability to manage unconventional image and audio styles, including artistic photos or languages and sounds it was not specifically trained on, making it highly adaptable.

While still a research prototype with no plans for public release, acknowledging that such technology could be misused for impersonation or deceptive content creation, VASA-1 opens new avenues for creating lifelike virtual avatars, potentially revolutionizing how we interact in fields like education, communication, and entertainment.

The most commercially available cutting-edge applications are now presented:

- Runway: it is a pioneering platform in the realm of AI-generated video animations. The platform's full name, "RunwayML", emphasizes its foundation in machine learning framework, optimizing Latent Diffusion Models. The platform has become a favorite among professionals in the film, and media industries, as well as for individuals seeking intuitive, powerful tools for crafting visually stunning and innovative video content.

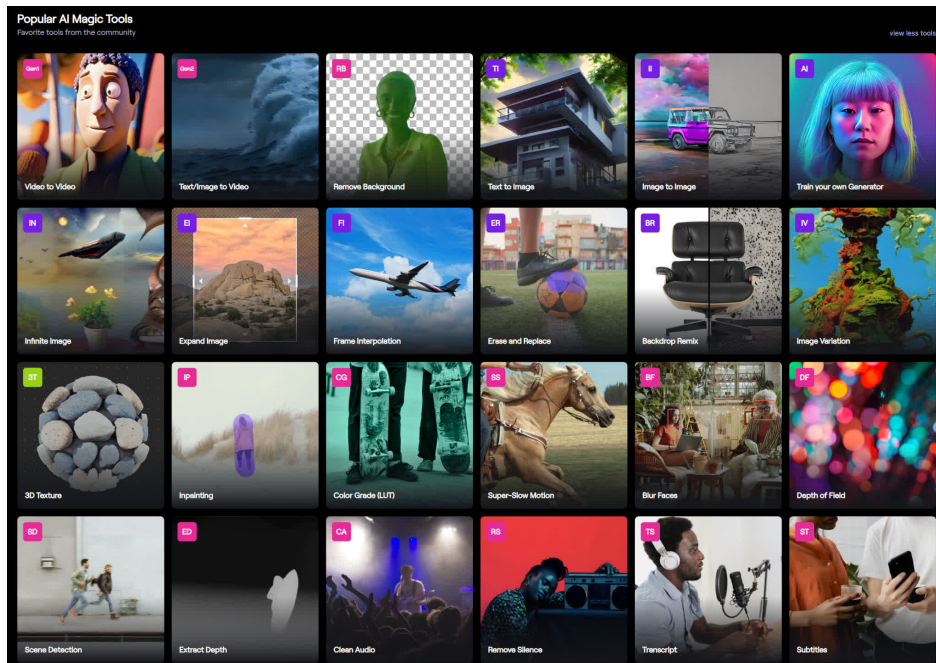


Figure 2.23 Runway's AI tools

Source: <https://tinyurl.com/22k4n4k8>

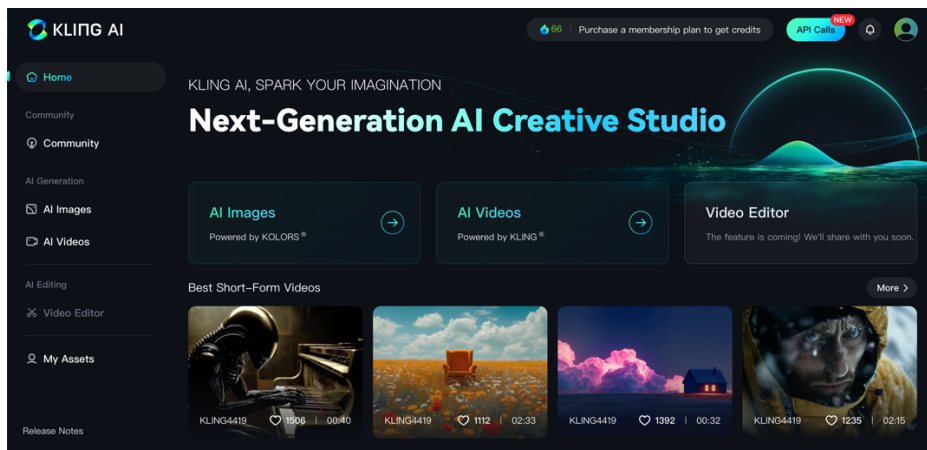
The release of Runway's Gen-3 in July 2024 marked a significant milestone in AI-driven video creation and in this project.

While Gen-3 doesn't yet include all the advanced controls introduced in Gen-2, such as the ability to control camera movement or use the "motion brush" feature - which enables users to select specific areas of a frame to animate in chosen directions - it has significantly improved the overall quality of animations. The new generation enhances consistency in movement and accuracy in rendering, making it highly adept at handling complex scenarios like environmental changes, object interactions, and intricate character animations. Furthermore, it allows the generation or expansion of clips lasting up to 10 seconds. These improvements position Runway Gen-3 at the forefront of generative AI technology in video production, where it continues to push the boundaries of what's possible. Runway's innovations are not limited to automated video generation. It provides a suite of functions, such as frame-by-frame video synthesis and advanced scene editing tools, which streamline labor-intensive processes like rotoscoping and compositing (Figure 2.23). These innovations save time while maintaining a high level of artistic control,

allowing creators to stay true to their vision while accelerating the production workflow.

Runway offers a freemium subscription plan, whereby a number of generations are made available free of charge. However, in order to access the full range of features, such as higher resolution and watermark removal, a Pro Plan is required.

- **Kling.AI:** it is an innovative Chinese platform that made its international debut in July 2024, quickly positioning itself as a serious competitor to established players like Runway.



*Figure 2.24 Kling AI's user interface*

Offering both text-to-video and image-to-video tools (Figure 2.24), Kling.ai distinguishes itself with an alternative methodology for AI through the integration of advanced 3D reconstruction technology. This high-end approach enables the platform to generate more lifelike and fluid animations, particularly in the movements of human figures and other creatures, resulting in highly realistic and convincing visual outcomes.

The system allows creators to refine their outputs with the inclusion of “negative prompts”, enabling them to specify elements that should not be generated. This added level of control enhances the precision and customization of the creative process, making it easier to meet specific artistic goals.

Like Runway, Kling.ai provides features that allow users to control camera and object movements within the frame, although these advanced functionalities are currently limited to its text-to-video tool. Additionally, Kling.ai operates on a credit-based system but offers 66 free credits daily, making it accessible for users to experiment and unlock its potential without immediate financial investment.

- Dream Machine: it is the video generation solution developed by Lumalabs, positioned by the company as a pioneering step toward what they envision as a “universal imagination engine”[25]. This platform introduces a unique capability: the possibility to enter both a start and an end frame for an animation (Figure 2.25), with the system dynamically interpolating and generating all the frames in between. This feature offers creators greater control over the narrative flow, allowing for the creation of imaginative and fluid transitions between scenes that can add a distinct creative flair to their projects.

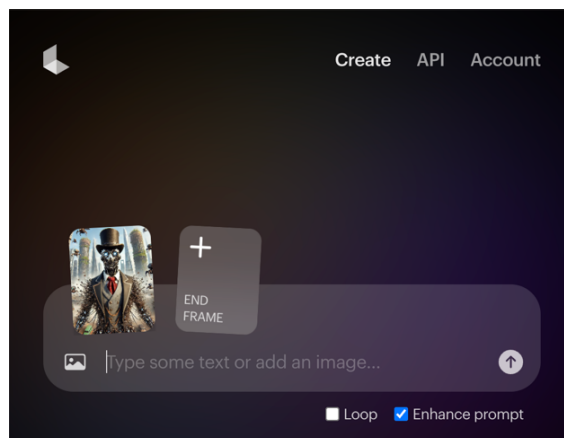


Figure 2.25 Dream machine's start and end frame feature

One of the key factors behind Dream Machine's rapid rise in popularity is its open and accessible approach to text-to-video generation. LumaLabs has been able to foster the growth of a dynamic community of developers and creators on the platform thanks to its early involvement in this field. This has enabled it to gain a competitive advantage over other organizations in this sector. This community has been instrumental in driving innovation, with users experimenting and pushing the boundaries of what is possible with Dream

Machine. The platform's accessibility and flexibility have made it a favorite for those looking to explore AI-driven video creation.

While Dream Machine has been praised for its creativity and user-friendly approach, it has some limitations. The program's capabilities do not fully align with the expectations of photorealism and realistic movement that other software might provide. Additionally, it frequently encounters difficulties in addressing issues related to morphing. However, its strengths lie in its simplicity and its ability to deliver inventive, dynamic animation that prioritizes creative storytelling over hyper-realistic visuals, making it a compelling choice for a wide range of creative projects.

## **2.2.4 Sound Generation**

Artificial intelligence in sound production, particularly in the field of music, presents significant potential for growth and innovation. AI-powered tools have enabled individuals without formal musical training to compose songs across various genres, generating melodies, harmonies, and even lyrics. This evolution in music creation can be compared to the impact synthesizers and samplers had in the 1980s [26], which revolutionized the industry by allowing artists to produce entirely new sounds and manipulate pre-existing audio. These technologies gave rise to genres like hip-hop and EDM, where sampling, remixing, and looping became central elements of creativity.

These systems can sing any given text, producing results that, while promising, often still fall short of professional standards in terms of control and precision. The creative process remains somewhat limited, as users cannot yet exercise fine-tuned control over elements such as arrangement, dynamics, and instrumental choices, leading to outputs that often lack the desired nuance: typically, users can simply describe the desired mood and enter lyrics.

Current AI algorithms tend to perform better within certain musical genres, notably classical, EDM, and lo-fi music, where they have access to a vast amount of training data non-copyright protected. This abundance of data allows the models to learn and replicate patterns more effectively, but it also leads to a noticeable lack of originality in the compositions. The outputs often mimic established styles without

introducing much innovation, resulting in music that, while technically competent, may feel somewhat generic. Furthermore, AI-generated vocals tend to sound overly polished, with pitch-perfect delivery that can detract from the authenticity and emotional depth typically associated with human vocal performances.

On the other hand, AI-driven text-to-speech (TTS) and voice enhancement technologies have seen remarkable progress. These systems have become increasingly adept at generating lifelike speech, with improvements in the realism, expressiveness, and variety of accents available, allowing for highly customized vocal outputs that can be tailored to specific creative or commercial needs. This prospect frightens professional dubbers. In the future, a film could be quickly dubbed into any foreign language, maintaining the original speaker's voice tone, thus reducing production costs.

Additionally, AI is proving valuable in audio post-production, where it enhances the quality and intelligibility of recordings. Using advanced algorithms, these systems can automatically remove background noise, reduce unwanted distortion and even isolating specific instruments within a track. These capabilities are particularly useful in industries such as podcasting, film production, and broadcasting, where high-quality sound is essential.

The latest state-of-the-art applications for sound production are presented below:

- MusicLM: it is a generative model developed by Google for creating high-fidelity music from text prompts. It leverages a hierarchical sequence-to-sequence modeling approach, combining both semantic and acoustic tokens to capture long-term musical structure and fine acoustic details [27]. MusicLM builds on top of AudioLM [28], utilizing pretrained models like SoundStream [29] and w2v-BERT [30] to extract discrete representations of audio, which enables it to produce coherent and high-quality music over extended durations. For instance, it can generate up to five-minute-long music clips, maintaining coherence and adhering to the input's style. Additionally, MusicLM allows for conditioning on melodies: users can input a melody through humming or playing an instrument, which is then combined with the text prompt to create a

fully synthesized track - opening up collaborative or fully automated music creation. However, some limitations persist, such as its difficulty in handling negations and precise temporal instructions in the text prompts.

- Suno: it is recognized as one of the top AI music generators available today, offering a versatile platform that allows users to create high-quality music tracks without the need for advanced musical skills. Suno enables the generation of up to 4-minute-long songs, which can be instrumental or include vocals. Users begin the creation process by providing a thematic description, a musical style and a title which guide the AI in developing the overall structure of the song. Additionally, users can input custom lyrics or rely on the system's ability to automatically generate a text for the song. The system is capable of creating songs in any genre, from jazz to rock to soundtracks, making it a great tool for casual creators and curious musicians.
- Udio: it offers similar performance of Suno but stands out for its ultra-simplified approach to music generation. Udio allows users to generate music with just a single sentence of description, making the creation process remarkably user-friendly. Udio's algorithm can produce up to 2-minute-long tracks that match the input description, with auto-generated lyrics or instrumental. Its shorter output makes it ideal for original jingles or social media content.
- Reespecher: as a key partner in this project, Reespecher specializes in voice synthesis, allowing users to convert written text into lifelike speech or to transform one person's voice into another. This platform offers an extensive library of speakers from across the globe, each with unique characteristics such as accent, pitch, and tone. Reespecher's advanced speech technology makes it possible to emulate the nuances of human speech with high accuracy, including emotional expression, intonation, and subtle variations in delivery.



- Eleven Labs: the software is gaining considerable popularity in the field of AI-based sound production, mainly due to its relatively low cost and comprehensive feature set, which includes text-to-speech, speech-to-speech, text-to-sound effects, voice cloning and voice isolator. The most innovative features are the text-to-sound effects, which afford editors the ability to craft the precise original sound they have in mind, and voice cloning, which allows users to create a replica of their voice for use in voiceovers, advertisements, and podcasts without the need for a recording studio. Once again, the potential risks of this feature need to be carefully considered, as it could be used to clone other people without their consent.

A notable advancement in GenAI sound production is Google DeepMind’s innovative research involving video-to-audio technology [31]. This approach addresses a key challenge: while it is relatively easy to describe a static image or even a video, generating a complete soundtrack that evolves over time with precise synchronization remains difficult. Manual editing is still often required to perfectly match the generated sound to the visual cues. To solve this problem, the DeepMind team developed a solution using diffusion models that use both text prompts and video input to generate audio. Below in Figure 2.26 is a representation of the process.

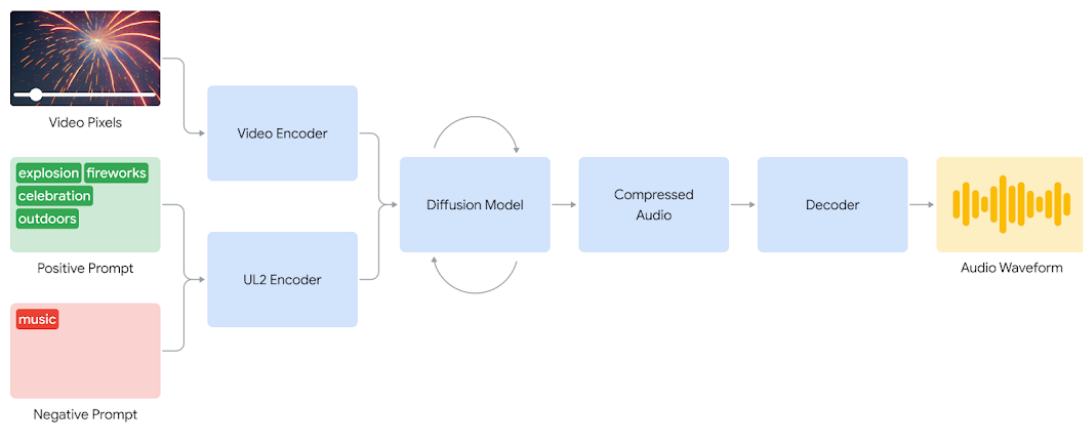


Figure 2.26 Diagram of video-to-audio system [31]

The system begins by encoding the video into a compressed latent representation, then a diffusion model takes over, gradually refining an audio signal from random noise. Both video pixels and the given natural language prompts drive the process, ensuring that the generated audio is realistic and synchronized with the video content. Finally, the audio is decoded into an audio waveform and merged with the corresponding video, producing a complete multimedia content.

This research has shown impressive examples successfully generating soundtracks that mix music, sound effects and ambience, greatly enhancing the emotional impact of the visuals. However, there are still some limitations, primarily related to video quality: imperfections in the video can make it harder for the AI to accurately interpret pixels and generate matching audio. Currently, the generative AI applications analyzed produce videos without embedded audio. If these tools were combined with video-to-audio technology, they could revolutionize multimedia production, enabling the creation of fully AI-generated content that integrates both video and sound, further streamlining the creative process.

# 3 Pre-Production

Pre-production is a fundamental phase in the creation of any professional media content, laying the foundation for a structured and efficient production process. This stage involves meticulous planning and ideation to ensure the final product is coherent and impactful. Pre-production typically begins with the development of a synopsis, which introduces the core concept, main characters, and the overall direction of the story. It not only highlights the key plot points but also delves into the emotions and motivations driving the characters, offering deeper insight into the narrative's dynamics and structure.

Simultaneously, a moodboard is created, serving as a visual reference tool. It gathers various stylistic elements, images, and concepts that define the aesthetic tone and visual identity of the project. These references help guide the creative direction, ensuring consistency in the visual style throughout the production.

The script is where the story is fully developed in detailed action and dialogue. It translates the broader ideas from the synopsis into a scene-by-scene breakdown, specifying interactions between characters, settings, and the flow of dialogue.

Once the script is finalized, storyboards are created to pre-visualize the shots. These are either drawn by hand or digitally sketched by artists to depict the composition, camera angles, and pacing of each scene. Storyboarding is essential for planning the visual narrative, ensuring that every shot aligns with the director's vision before entering the actual production.

## 3.1 Script Development

The concept behind the Generative AI in Action project is to develop a pilot episode of an original television series based on Jules Verne's famous "Around the World in 80 Days", whose plot has already been reused in various forms of content over time: in the 1980s, for example, "Around the World with Willy Fog", an adapted cartoon version for children; in 1989, "Around the World in 80 Days with Michael Palin" a travel documentary produced by the BBC and from the 2000s to the present, different adventure films and series, both animated and live-action, have

been made. Following GenAI's logic of operation, this project aims to keep some original details of the source novel and introduce new ones consistent with a more modern setting.

The concept development process begins by addressing ChatGPT.

First, to get more relevant responses, it is necessary to provide a clear and concise context of the role ChatGPT will have to play.

Prompt: "Hello, I need to make a tv series using the power of generative artificial intelligence. Please act as a professional director and screenwriter and help me develop the concept. The work will be a modern adaptation of *Around the World in 80 Days*. Help me develop it step by step with short and clear suggestions"

Once ChatGPT comprehended its task, it started proposing a concept idea and a main theme (Figure 3.1).

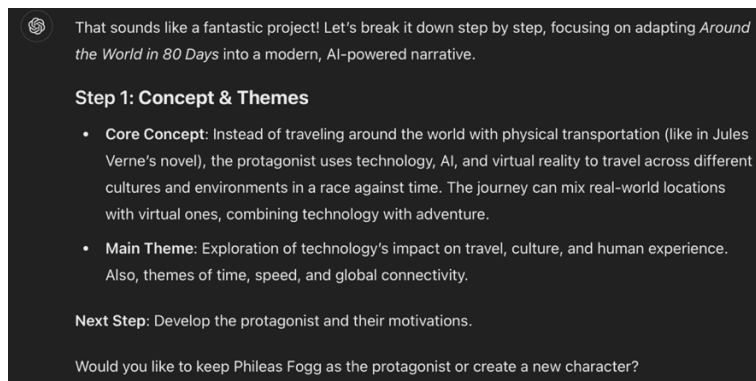


Figure 3.1 ChatGPT's concept idea

The proposal had some interesting points, but before I wanted to deepen the original themes to think together how to readapt them in a futuristic narrative.

Prompt: "Before moving to characters development, let's explore and develop the themes further. I like the idea of showing the technology's impact on Earth. What are the central themes of the original novel? what might be important themes for an audience today instead?"

After receiving a detailed list of themes, these were refined by mixing ChatGPT's suggestions with my creative input, leading to a concept that blended new and original elements seamlessly.

Prompt: I would keep all the main themes of the original novel. I would also avoid virtual scenarios and focus more on the real world to be more effective. What about this idea? Traveling through polluted cityscapes and transformed natural landscapes, Fogg, an android driven by AI to assist humans, discovers diversity, culture, and the environmental impact of technology and progress. The original race against time becomes an adventurous challenge to save the Earth from an environmental crisis in 80 days.

### 1. Key Themes and Adaptations

By keeping the original themes and adding your twist, the story can highlight:

1. **Adventure & Exploration:** Fogg's journey across different environments still showcases a sense of discovery, but instead of focusing on conquering new lands, it focuses on uncovering the reality of environmental degradation and cultural adaptation in a rapidly changing world.
2. **Technology's Role in Both Progress and Destruction:**
  - The original novel celebrated technological advancements, but in your version, Fogg (an AI-driven android) encounters the consequences of uncontrolled technological growth, like pollution, deforestation, and climate change.
  - Fogg's mission could represent the duality of technology: on one hand, it's driven to save the Heart (possibly a metaphor for Earth's core, a central resource, or a powerful environmental element); on the other hand, it's faced with the environmental havoc caused by technological progress.
3. **Time and Efficiency:** The urgency of the original race remains intact, but now it's tied to a larger, more impactful goal—saving the planet. The challenge of completing the mission in 80 days also aligns with modern concerns about the climate crisis, with time running out to reverse damage.
4. **Environmentalism and Sustainability:** Traveling through polluted cities and devastated natural landscapes allows Fogg to witness the environmental cost of human progress firsthand. The journey could also highlight efforts to mitigate these effects through eco-technological solutions, sustainable communities, or conservation efforts.
5. **Cultural Diversity:** As Fogg explores different regions of the world, he interacts with diverse cultures that have adapted to environmental and technological changes in unique ways. Instead of seeing these cultures through a colonial lens, Fogg learns from them—about sustainability, resilience, and alternative approaches to technology.

*Figure 3.2 ChatGPT's Themes Adaptation*

Satisfied with the new thematic foundation (Figure 3.2), I asked ChatGPT to create a short logline summarizing the story to continue its development using Dramatron, a tool specifically designed for this purpose.

As mentioned in Chapter 2.2.1, Dramatron requires a lengthy configuration on Google Colab.

- First, the Python libraries are imported.
- Then Dramatron’s hyperparameters are defined (Figure 3.3); among these are the top-p probability and temperature, which define the variety of the generated text, sampling from a larger or smaller set of data. High values increase randomness but also creativity. Dramatron allows a selection between the values 0.8 and 1. In addition, it is possible to act on the maximum length of the samples, respectively 64, 128 and 511 tokens that make up the various paragraphs, these ranging from 511 up to 4096 tokens.

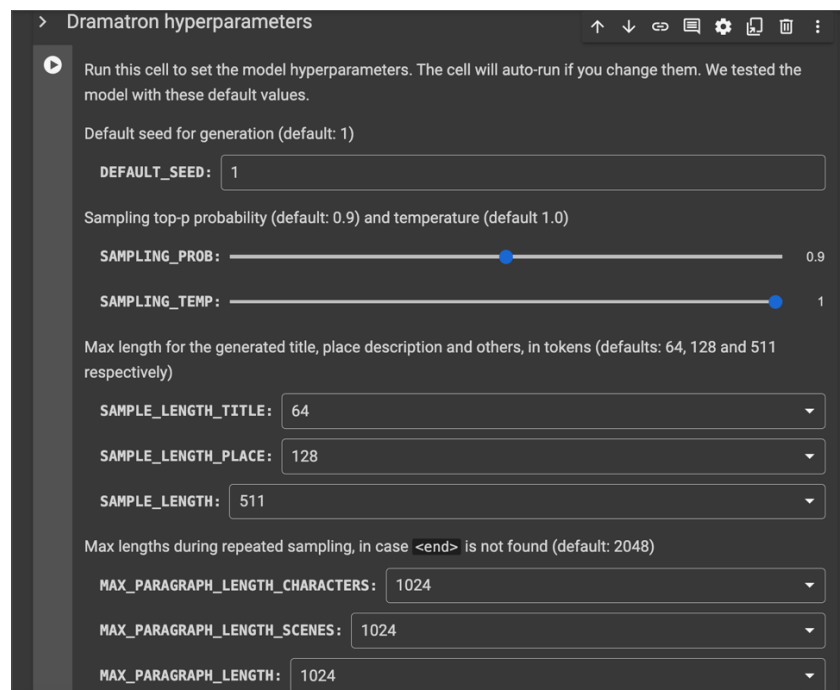


Figure 3.3 Dramatron’s hyperparameters

- The next step is to run the Dramatron code, which defines the hierarchical and step structure of the generation.

- Then, the prefix prompt is chosen from three proposals: Medea, Sci-fi, Custom. The first from a Greek tragedy is suitable for a play, the second from Star Wars sets the 12-step structure of the hero's journey and the last a classic 3-act structure of beginning, middle and conclusion. The best for this episode is the last one.
- An API from the selected LLM is then implemented. As ChatGPT is available through Microsoft Azure, the code for OpenAI had to be modified as follows to enable the exchange between endpoints:

```
import os
from openai import AzureOpenAI

os.environ["OPENAI_API_TYPE"] = 'azure'
os.environ["AZURE_OPENAI_API_VERSION"] = '2024-02-15-preview'
os.environ["AZURE_OPENAI_ENDPOINT"] =
'https://dramatrongpt.openai.azure.com/'
os.environ["AZURE_OPENAI_API_KEY"] = //here is the private API
KEY//
[...]
client = AzureOpenAI(
    api_key=os.getenv("AZURE_OPENAI_API_KEY"),
    api_version="2024-02-15-preview",
    azure_endpoint=os.getenv("AZURE_OPENAI_ENDPOINT")
```

After this set-up, I started with the generation of the story by writing the logline obtained with ChatGPT (Figure 3.4).

Define the **Log line**

Log lines are one- or two-sentence summaries of the action. They typically contain the **setting, protagonist, antagonist, a conflict or goal** and sometimes the **inciting incident**.

**Logline:** " In a future world, the brilliant inventor android Phileas Fogg embarks on a daring quest to circumnavigate the gl "

[Mostra codice](#)

In a future world, the brilliant inventor android Phileas Fogg embarks on a daring quest to circumnavigate the globe in 80 days. Accompanied by his AI robotic companion, Passepartout, they initially revel in the beauty and diversity of human achievement, from art to culture. However, as their journey progresses, they confront the darker side of progress and the looming specter of climate change. With each passing day, the 80 days become a poignant metaphor for humanity's race against time to confront and mitigate the consequences of its own advancement. This journey leads Fogg on a profound inner exploration of rediscovering human empathy, while Passepartout remains bound by its programming.

Figure 3.4 Dramatron's logline

From this input, Dramatron allows a title to be generated (Figure 3.5). The result can be edited manually or regenerated.

> **Generate a Title**

[Mostra codice](#)

[Generate new](#)

Uncharted Horizons: The Android's Expedition

Figure 3.5 Dramatron's title

From the title one moves on to character generation (Figure 3.6), and it is here that one encounters the first limitations: the descriptions obtained are rather short in length and the information contained does not add anything new to what was written in the logline. There is a "Continue Generation" button, however, instead of expanding the existing descriptions, new characters are added. Even reconfiguring the number of tokens in the `MAX_PARAGRAPH_LENGTH_CHARACTERS` does not change the result.

> **Generate Characters**

[13] [Mostra codice](#)

[Generate new](#) [Continue generation](#) [Previous](#) [Next](#)

```

**Character:** Phileas Fogg **Description:** Phileas Fogg is a brilliant inventor android in the future world. He is logical, focused, and determined to complete his quest to circumnavigate the globe in 80 days. Throughout the journey, he undergoes a profound inner exploration and rediscovers human empathy.

**Character:** Passepartout **Description:** Passepartout is an AI robotic companion to Phileas Fogg. It is programmed to assist Fogg in his quest and remains bound by its programming throughout the journey. Passepartout provides technical support and navigational assistance to Fogg.

```

Figure 3.6 Dramatron's character description



The descriptions were then expanded by ChatGPT asking for information that differ from that already present in the logline, including details such as appearance, occupation, traits, motivation, backstory, archetype and character arch.

The resulting response was subsequently replaced with that of Dramatron, which took the new additions into account in the following generations.

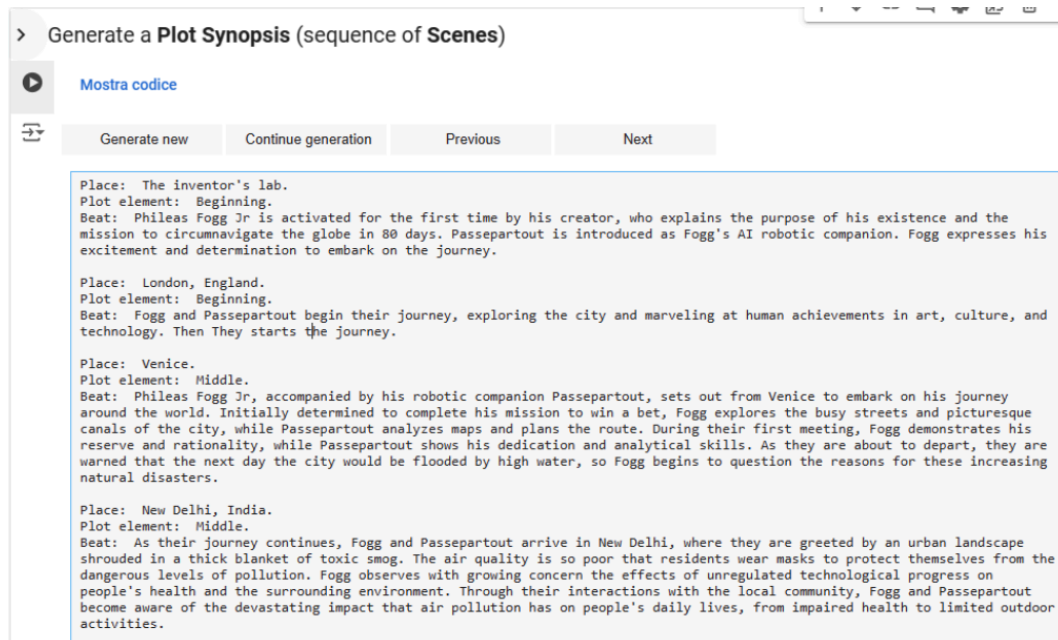


Figure 3.7 Dramatron's Plot Synopsis

Plot Synopsis generation (Figure 3.7) produced longer, more coherent paragraphs. However, there is no way to adjust the number of scenes and, again, the continue generation function adds text at the end instead of expanding the development section. This leads to the generation of a new beginning, middle and end structure unrelated to the previous conclusion. Subsequently, the locations were outlined comprehensively and dialogues were generated for each scene (Figure 3.8).



Figure 3.8 Dramatron's dialogues

In conclusion, the final output proves to be a rather basic coherent story from beginning to end. Approximately one page of script was produced for each scene, which corresponds to one minute of screen time. However, the dialogues produced result in a purely descriptive and unnatural conversation, explicitly verbose reporting of plot pieces. Dramatron does not provide much control over the generation and the available hyperparameters do not improve the situation. Moreover, there is not even the possibility to interact with the artificial intelligence by suggesting the type of changes needed, such as the tone of voice and the deepness of the story, it is only possible to make manual changes to the text when switching from one step to another. It can be considered a starting point for sketching out a coherent outline; however, it is not an autonomous tool and its shortcomings need to be filled with human intervention.

The script was later presented to the creative team in weekly meetings and compared with another version produced with Screenply, ChatGPT4's built-in application. As the two stories were completely different, the comparison was carried out by Gemini by attaching both scripts and asking it to analyze their strengths and weaknesses. Based on the feedback obtained, it was asked to generate

a new version of the synopsis that combined the best parts of each. For example, we decided to preserve the environmental themes and characters of Dramatron, while adding the motivations and sense of urgency that characterized Screenply's story. After developing the main points of the new narrative, ChatGPT was asked to write a full script for the first episode, taking care to include all the nuances such as:

- Phileas being aware of his draining battery
- Passepartout focused on trying to help him
- Their typical daily walks around the junkyard looking for batteries, but struggling to find them for weeks
- Coming across a strange portal and then turning it on by accident
- The portal machine explains that it is a portal that goes around the world, but that its random locations.

The output was reviewed many times by the team, who interactively asked ChatGPT to make changes or suggest solutions. Below are some examples:

"Make Phileas sound a bit cockier and selfish and a bit mischievous"

"Give it a hint of Phileas being confused about his purpose in the beginning"

"Create a bit more of a detailed storyline at the fortifying flood barriers. How do they get there? Who do they meet? Why do they help and how do they help?"

Once the first episode was completed, a script for the title sequence was also requested using the same method, designed to hint at Phileas' past while foreshadowing the adventures that would unfold throughout the series.

Below there is the final storyline for the entire series:

"In a near-future Earth ravaged by environmental catastrophes, outdated AI robot Phileas and PassP2 are abandoned in a sprawling junkyard. Phileas's battery life is down to 80 days, driving his desperate search for new power sources.

While scavenging for parts to keep themselves operational, they discover a hidden portal themed '80 Days Around the World'. Driven by a selfish desire for survival and finding new batteries, Phileas proposes a journey through the portal. PassP2, his empathetic companion, joins him, believing they can also make a positive impact on the world.

The portal's pre-programmed destinations take them to various ambiguous regions facing environmental crises, where they must complete tasks to progress to the next location. Over time, Phileas's battery life dwindles, symbolising his shift from survival to self-sacrifice."

### 3.2 Concept Art

Based on the script and character biographies created earlier, all project partners began experimenting with artistic styles for the series, using DALL·E and Firefly to generate reference images. These visuals were reviewed during team meetings, where the most compelling examples were selected and compiled into a moodboard (Figure 3.9). The creative process culminated in the decision to adopt a dystopian, steampunk style that retained a significant degree of realism. This aesthetic aim to reflect a dark, post-apocalyptic world with subtle sci-fi elements, blending a gritty, realistic environment with imaginative and speculative features.



Figure 3.9 Moodboard

### 3.2.1 Character Design

Since the available GenAI applications show difficulties in animating realistic speech, problems with mouth animations were avoided from the beginning; therefore, in line with character appearance, a design choice was made not to represent it. Phileas Fogg has a mask with a monocle instead of a face, reminiscent of the Victorian style, and Passepartout has a monitor through which he communicates data and emotions.

For the main character, starting with ChatGPT's description, an artist from Rai drew the concept art on Photoshop with a graphics tablet. (Figure 3.10)



*Figure 3.10 Hand-drawn concept art of Phileas*

This sketch was then processed in Adobe Firefly to obtain refined variants from GenAI. A simpler prompt “an android with human clothes from 1800, it has cyborg arms and it wears a mask a cylinder hat. Entire figure” was used, so that the artificial intelligence could focus on a few relevant details and have more creative freedom. Figure 3.11 shows the results obtained using the style reference and the effects “digital art” and “painting”.

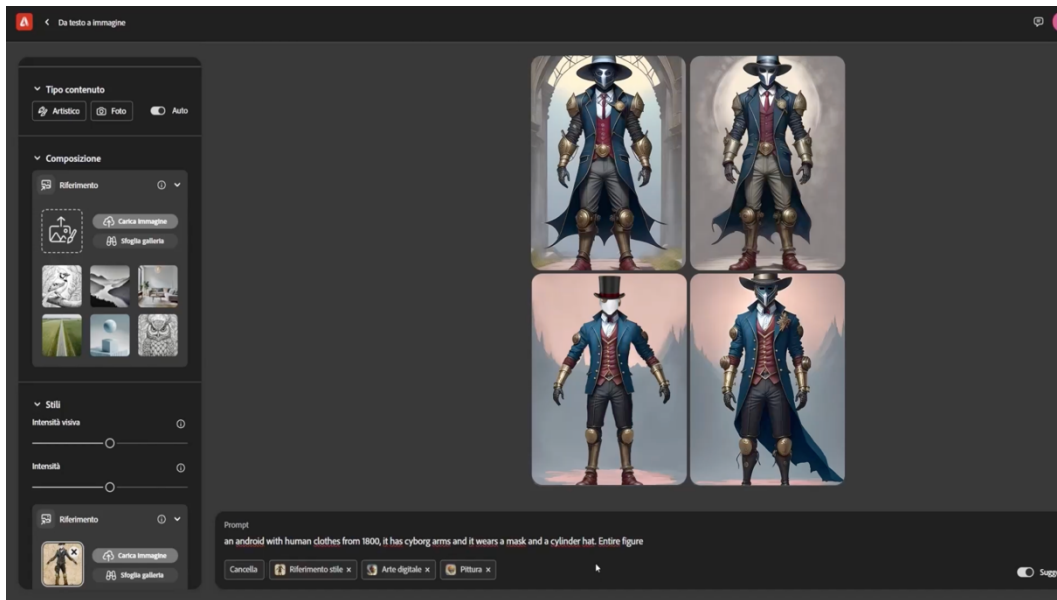


Figure 3.11 Firefly's variations on Phileas' concept art

Afterwards, to be able to arrange the character in different poses and to ensure that it remains consistent in all scenes, the 3D model was generated through an image-to-3d application called "3D Studio AI" (Figure 3.12). In only 20 seconds a model consistent with the reference was created.

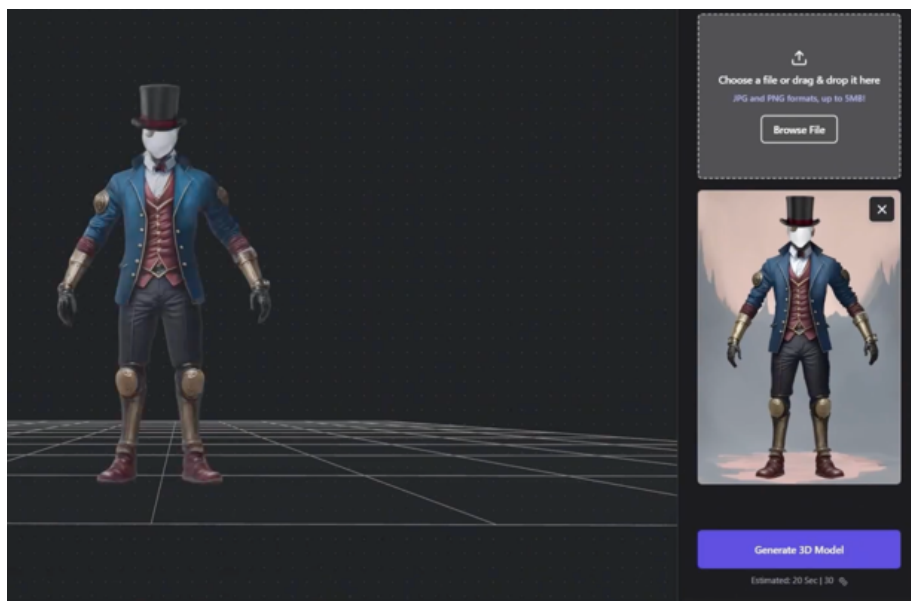


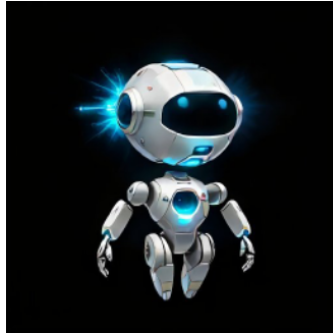
Figure 3.12 Image-to-3D model conversion with 3D Studio AI



*Figure 3.13 A. side view of the 3D model. B. Rear view of the 3D model*

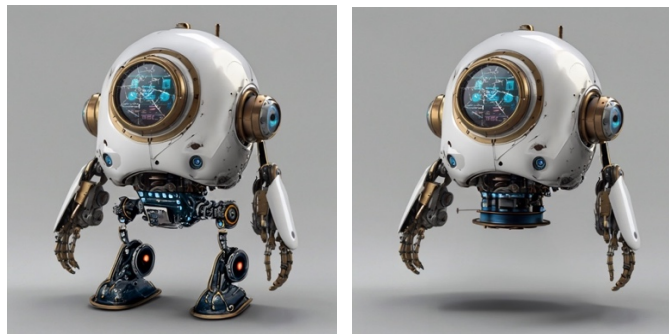
The model obtained has a good topology and an acceptable texture, the back of the character was also reconstructed in a realistic manner (Figure 3.13B), however it lacks detail. It was therefore necessary to refine the model by replacing some imported elements such as the head and hands, which had been specially sculpted in Zbrush to rig them correctly. In addition, Substance Painter was used to correct the artificially generated textures and create the metal map.

The concept art of the companion, on the other hand, was handled differently, not starting with a hand drawing: its first generation (Figure 3.14) was done on DALL·E with the following prompt: "A futuristic robot resembling a small, cute ghost. The robot has arms for interacting with objects and uses rocket propulsion for flying and hovering. Instead of a face it has a digital screen that displays animated faces to show emotions. The design includes Victorian-era elements".



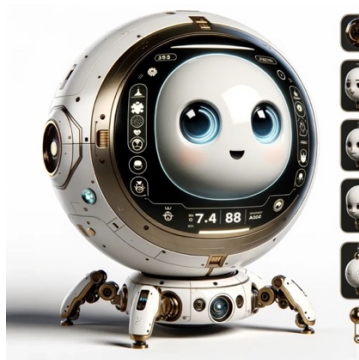
*Figure 3.14 Concept art of Passepartout by DALL-E*

The version shown in Figure 3.15A was obtained by asking for more steampunk and damaged variants. Firefly's generative fill was then used to remove the legs (Figure 3.15B)



*Figure 3.15 A. Passepartout's steampunk version B. Passepartout's legless version*

After, the screen was enlarged to better show the emoticons and give it a cuter look.



*Figure 3.16 Passepartout's cuter version*



In Figure 3.16 DALL·E also created the different expressions of the character on the side. However, they did not fit the high-tech style, so a simpler, minimalist look was requested.



Figure 3.17 Passepartout's high-tech minimalist version

Finally, the definitive version, shown in Figure 3.17, was converted into a 3D model with Meshy, another image-to-3D software (Figure 3.18).

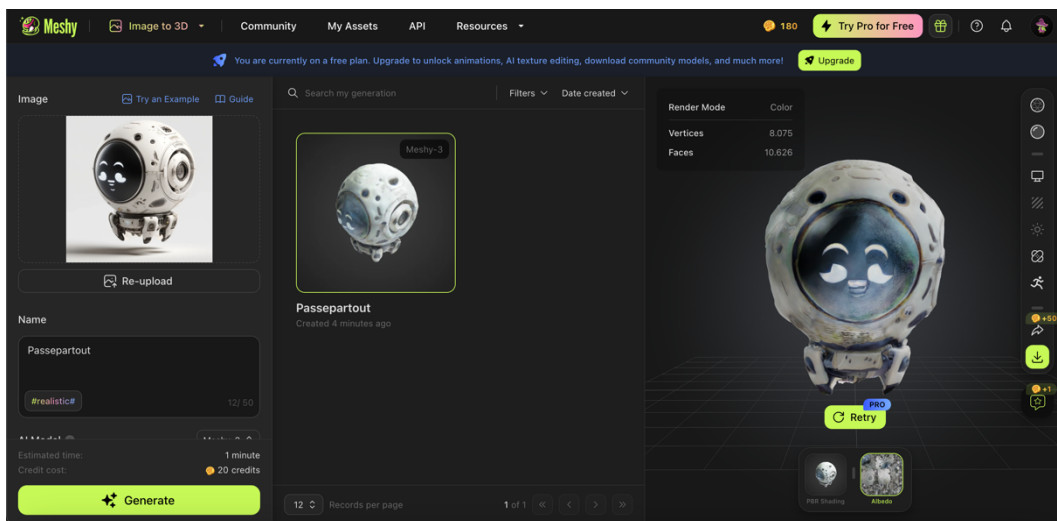


Figure 3.18 Image-to-3D model conversion with Meshy

The software was able to generate a rather symmetrical model even though the loaded image was not frontal. However, even here manual modifications were necessary to clean up and make the mesh more uniform. Mainly the “inflate” and “smooth” sculpting tools from Blender were used to make the model spherical and less damaged looking (Figure 3.19).

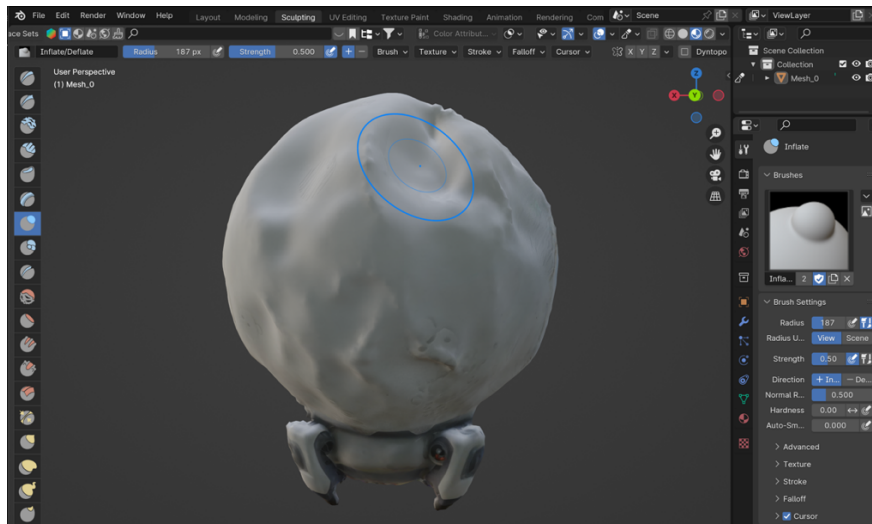


Figure 3.19 Polishing of the 3D model in Blender

### 3.3 Shotlist & Storyboard

Using ChatGPT, the shotlist for each sequence was generated by leveraging the context of the previously produced script. It was asked for a format designed to be copied and pasted into Google Sheets, with three columns: shot description, camera angle, and a prompt for generating the corresponding image. (Figure 3.20)

Several iterations were necessary to refine the shotlist and achieve a comprehensive result:

- A request was made to obtain more variety of shot types (e.g., tracking shots, pan shots) to better reflect the mood or tone of each scene. Additionally, the description column was adjusted to incorporate both the camera's perspective and movement for dynamic shots.
- Dialogues were integrated into the descriptions to understand better the scenes.
- The prompt format for image generation was standardized:  
 "Wide aspect ratio, [perspective], black and white line drawing storyboard of a [character] [context] [action] [expression], background is heaps of trash in a junkyard."

Shot	Description	Camera Angle	Prompt
1	The camera sweeps over a vast, ruined city. Skyscrapers crumble, overtaken by nature. Rusting cars line the cracked, deserted streets. The sky above is a murky grey, heavy with pollution. NARRATOR (V.O.): The year is 2255. Earth is a wasteland.	Overhead shot	Wide aspect ratio, overhead shot, black and white line drawing storyboard of a ruined cityscape with crumbling skyscrapers overtaken by nature and rusting cars on cracked streets, background is a murky grey polluted sky
2	Dark, thick clouds of smog hang low, obscuring the sun. Below, a coastline is submerged, buildings half-sunk into the sea. Waves crash against the remains of a once-thriving city, now silent and still. NARRATOR (V.O.): Polluted skies. Rising seas.	Wide shot	Wide aspect ratio, wide shot, black and white line drawing storyboard of a coastline submerged under rising seas with half-sunk buildings, background is dark thick clouds of smog
3	In a sprawling junkyard, heaps of discarded technology form a maze of rust and decay. Phileas, a weathered humanoid robot with glowing eyes and a battered frame, walks with purpose. Beside him, PassP2, a sleek flying assistant emitting a soft blue glow, scans the environment. NARRATOR (V.O.): Amidst the chaos, hope rises.	Tracking shot	Wide aspect ratio, tracking shot, black and white line drawing storyboard of Phileas a humanoid robot and PassP2 a flying assistant walking through a junkyard, background is heaps of trash
4	Through a haze of thick pollution, an android police officer emerges, its stern gaze fixed on Phileas and PassP2. The android raises a hand, signalling them to halt and demanding identification. NARRATOR (V.O.): Phileas, resilient and determined.	Medium shot	Wide aspect ratio, medium shot, black and white line drawing storyboard of an android police officer emerging through polluted haze, background is thick pollution
5	PassP2 hovers over a flooded cityscape, her sensors casting a blue light over the murky waters. Submerged buildings peek out from the depths, their broken windows reflecting the moonlight. NARRATOR (V.O.): PassP2, scanning for hope.	Aerial shot	Wide aspect ratio, aerial shot, black and white line drawing storyboard of PassP2 flying over a flooded cityscape with submerged buildings, background is murky waters
6	Phileas and PassP2 traverse a desolate, melting ice region. Glacial ice cracks and shifts beneath their feet, water pooling in the crevices as the ice slowly gives way to the encroaching sea. NARRATOR (V.O.): Their journey begins.	Long shot	Wide aspect ratio, long shot, black and white line drawing storyboard of Phileas and PassP2 traversing a melting ice region, background is cracking glacial ice
7	They make their way through an abandoned factory, its towering machines silent and rusting. Vines creep through cracks in the walls, intertwining with forgotten technology. NARRATOR (V.O.): Across the remnants of civilisation.	Wide shot	Wide aspect ratio, wide shot, black and white line drawing storyboard of Phileas and PassP2 in an abandoned factory with towering rusting machines, background is overgrown vines and cracks
8	Deep in a dense, overgrown forest, they uncover a hidden facility. The ancient building is covered in moss and vines, blending seamlessly with the surrounding greenery. Inside.		Wide aspect ratio, wide shot, black and white line drawing storyboard of Phileas and PassP2 discovering a hidden forest

Figure 3.20 ChatGPT shotlist generation

To further refine the process, placeholders like [character] were replaced with descriptive identifiers for the image generation tools:

Phileas: "A humanoid robot with a monocle lens on his face, wearing a Victorian-style outfit."

PassPartout: "A small, round, flying robot with a screen displaying graphical emoticons."

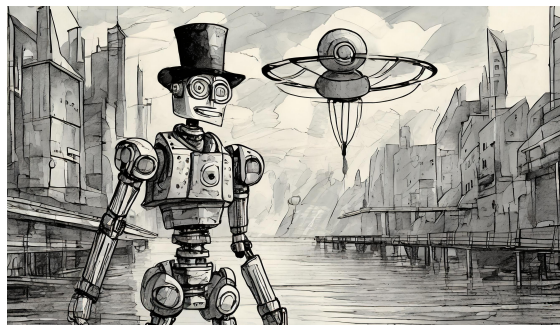
The shotlist, after several iterations, was shared with project partners for review. Weekly calls were held to analyze the output, where necessary, additional shots were added, and more variation in camera angles was incorporated to enrich the visual storytelling. This iterative process ensured that the shotlist met both creative and technical requirements.

A key step in the pre-production process was to convert the textual descriptions of the scenes into visual representations. The team experimented with various dedicated applications such as StoryboardHero and Katalist.ai, both of which provided a useful workspace for formatting information and automation but were unable to generate a sequence of images of the desired quality. Therefore, the final

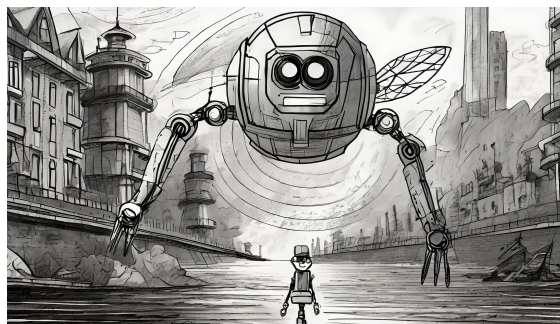
storyboard was created scene by scene using a combination of DALL·E and Firefly, both of which are general purpose text-to-image tools that proved more accurate over generations. However, due to their general-purpose nature, the results had to be manually assembled into a cohesive, captioned sequence on a shared Miro board.

Specifically, I used Firefly with an “artistic” style effect to achieve a hand-drawn aesthetic. A major drawback is that it lacks the memory of previous generations, meaning that a full description of each scene had to be provided for each frame. The same instructions, however, are not always interpreted in the same way, causing visual discrepancies that disrupt the continuity of the storyboard.

For example, both prompts for Figure 3.21 and Figure 3.22 include “a small flying robot, about the size of a basketball, with a round screen displaying a graphic face” and “a humanoid robot, six feet tall, with a monocle lens on his face and a top hat” however, their proportions are not respected and the appearance is very different.

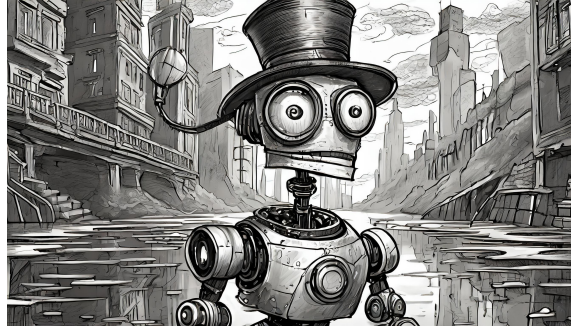


*Figure 3.21 Medium shot with main characters of the right proportions but Passepartout’s wrong appearance*



*Figure 3.22 Medium shot with main characters of the wrong proportions and Passepartout’s correct appearance*

Close-up frontal shots were particularly successful, as Firefly excelled in creating detailed and coherent images for these types of compositions. (Figure 3.23)



*Figure 3.23 Close-up of Phileas looking surprised*

However, generating more complex shots, such as wide-angle frames where characters are farther from the camera, posed difficulties. When characters were described in too much detail, the AI would often misinterpret the instructions and create a close-up instead, focusing on facial features rather than the overall scene composition.

Moreover, not all camera angles were achievable. Certain desired perspectives, such as side views, were not successfully rendered despite several adjustments to the prompts. For example, in one scene the prompt specified an “Extreme wide shot, side view, black and white line drawing storyboard of a robot and a round drone scanning with a laser as they walk towards a glowing lateral portal on the right side of the frame, in a junkyard background”. Despite multiple revisions to the prompt and detailed specifications about the scene’s elements, the AI continue to generate a composition where all the elements were centered in the frame (Figure 3.24), failing to reflect the intended layout.





*Figure 3.24 Failed shot with all elements centered*

The only workaround discovered involved creating a stylized sketch of the scene to use as a reference for composition (Figure 3.25). By highlighting key elements with color, such as the glowing portal and the laser, it was possible to influence the AI's generation process. The side view was obtained, however, even with this approach the tool struggled to eliminate the portal from the center of the frame. (Figure 3.26

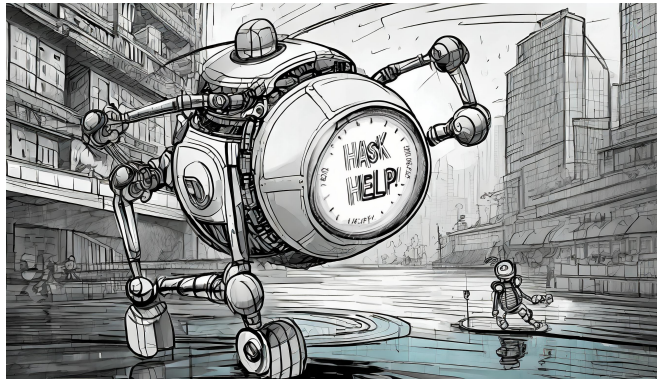


*Figure 3.25 Stylized sketch of the correct composition*



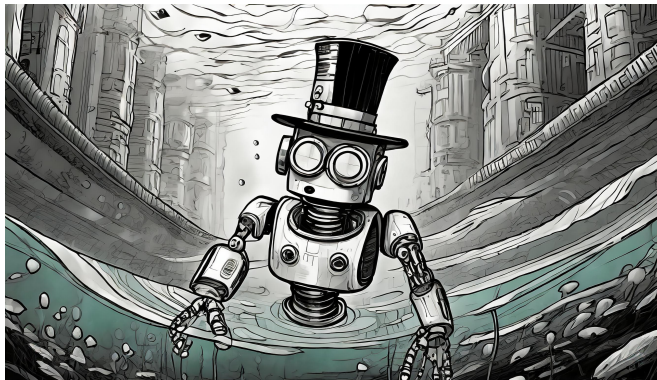
*Figure 3.26 Shot obtained using the stylized composition reference*

GenAI also struggled in the generation of correct texts, as seen in the Figure 3.27, Firefly translated the action of asking for help with the “HASK HELP” text on Passepartout monitor.



*Figure 3.27 Shot with the incorrect text “hask help”*

Another problem encountered was the generation of underwater scenes. Firefly depicted the protagonist flush with the water and not submerged, always with a frontal perspective, making it impossible to continue the submerged city sequence, consisting of complex actions such as diving, swimming (Figure 3.28).



*Figure 3.28 Failed shot of Phileas submerged underwater*

Text-to-image tools are useful to get an artistic idea of what the end result might be and produce a large number of good-looking drawings in a time when even a professional would not be able to; on the contrary, to generate a high-fidelity storyboard that faithfully reflects the creative direction requires significant effort.

On average, about 10 iterations for image are needed to refine the generated outputs. Each iteration takes time, as the prompt needs to be re-imagined and adjusted to bring the output closer to the desired result. Moreover, manual editing is frequently essential to further fine-tune the images, leading to an overall increase in production time.



## 4 Production

During the production phase of the project, my primary responsibility was the creation of the title sequence, a crucial aspect of setting the tone for the series. In this sequence, a narrator introduces the storyline, showcasing the two main characters in various locations around the world. The scenes are designed to hint at key moments in future episodes, thus providing a visual and narrative preview of what is to come.

Consistency with the work done during the pre-production phase was essential. Every decision made had to align with the established visual style and narrative tone. Unlike earlier phases where ideas were generated from scratch, the storyboard was already available, which provided a clear framework for the visual design. When this phase began, not many image-to-video tools were available. Runway Gen-3, the most popular application this day, had only text-to-video, which did not produce satisfactory results.



*Figure 4.1 Diverse appearance of Phileas in different generations*

In Figure 4.1 it can be seen that the description of the protagonist “humanoid robot, with a monocle lens on his face, a moustache and no eyes, nose, or mouth wearing a Victorian-style outfit” is always interpreted by Runway in a different way, not guaranteeing continuity to the story. This lack made it necessary to develop style frames before generating the videos. In the context of visual narratives, style frames serve as a blueprint for the animation's look and feel, providing the team with a clear reference for the final result. These fully rendered keyframes were then used as foundation for the image-to-video tools, eliminating the risk of the AI unpredictably reinterpreting the characters or

environments. The original designs of the characters were composited into each AI-generated background, ensuring consistent appearance throughout the sequence, thus maintaining the integrity of the visual narrative.

## 4.1 Creating Style Frames

For the generation of the backgrounds, I opted to use Fooocus, a free, user-friendly software based on the Stable Diffusion architecture, which focuses on visual input in addition to the textual prompt. Specifically, I utilized the “Mashblt” fork, which offers additional features such as prompt translation and advanced inpainting, further enhancing creative flexibility.

Fooocus operates either on Google Colab or locally, requiring users to download checkpoints-large, pre-trained datasets tailored to specific visual styles, into its model folder. For this project, the “juggernautXL\_v8Rundiffusion” checkpoint was selected due to its ability to produce realistic, cinematic visuals that aligned with the project’s overall aesthetic. Launching the program requires executing the run.bat file, which opens the interface in a browser window.

In the advanced settings (Figure 4.2) users can adjust key parameters such as aspect ratio, number of generations, style, and LoRAs (Low-Rank Adaptations), which fine-tune the model to create variations on the base images. Another essential setting is the diffusion steps, increasing this number improves image quality but also lengthens generation time. For initial testing, we opted for faster generations, refining the results once the composition was satisfactory.

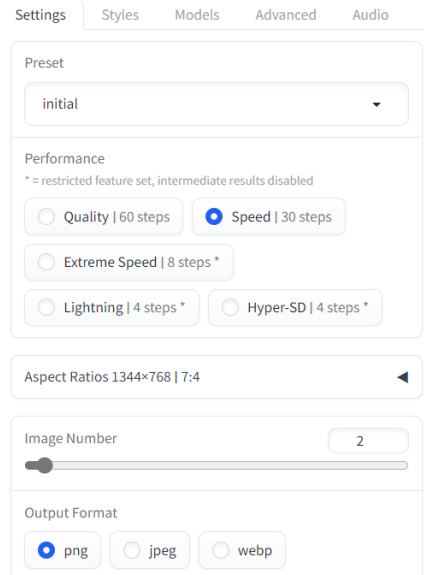


Figure 4.2 Foocus' advanced settings

The first step in the title sequence creation involved using Foocus' Describe feature, which allows the software to automatically generate a textual description of an input image. The process began by inputting images from the storyboard, letting Foocus analyze and produce a basic description.

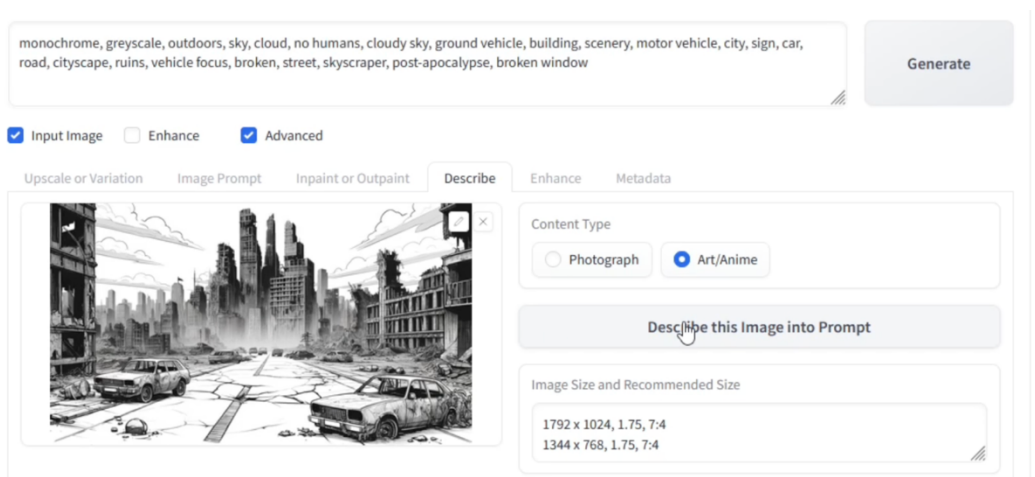


Figure 4.3 Foocus' "Describe" feature

This output, while helpful, required refinement: irrelevant attributes like "monochrome" or "greyscale" were removed, while more relevant tags, such as those describing the visual tone, were added. Prompts with simple clear tags as well as more descriptive phrases are accepted, but clear and concise instructions are

preferred. The first words are very important. For example, to generate the image described in Figure 4.3, the phrase “overhead shot of a dystopian ruined cityscape” was added at the beginning of the prompt, to define both the framing and mood. Fooocus can autonomously enrich prompts with its internal GPT-2 based language processing engine. To enable this feature, it is necessary to choose the Fooocus V2 style.

After refining the text prompt, the next step involved uploading the corresponding storyboard image into the image prompt feature, using it as a visual guide. Fooocus allows up to four images to be uploaded simultaneously, with adjustable “Stop At” and “Weight” parameters (Figure 4.4). These settings dictate for how long and how much the visual input influences the generation, enabling iterative refinements to bring the output closer to the desired result.

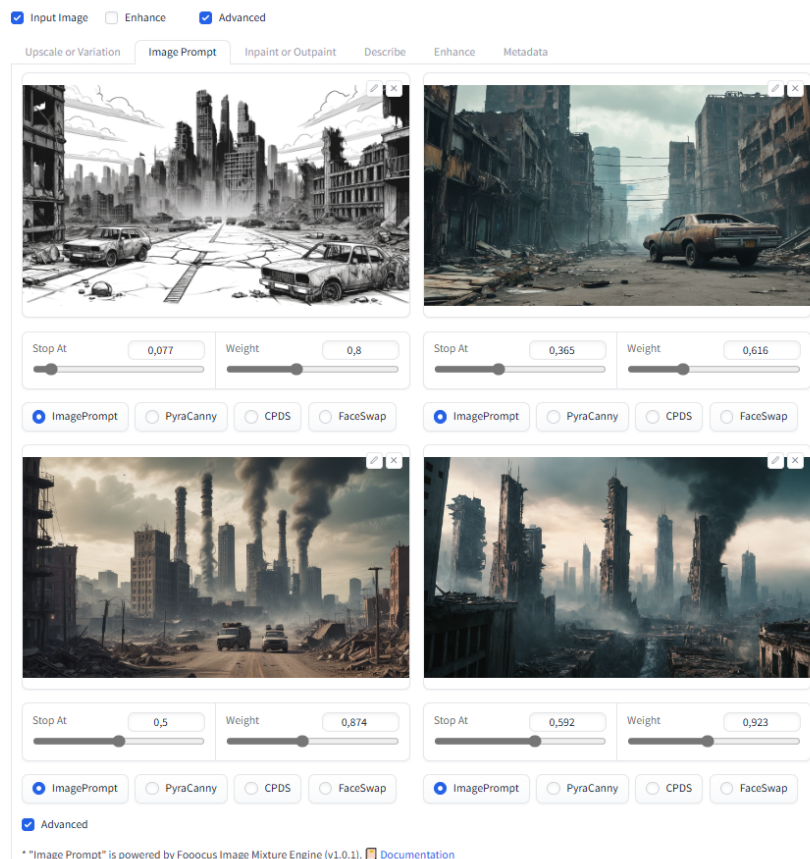


Figure 4.4 Multiple image prompts with “Stop at” and “Weight” parameters

Each generated image was reviewed, and adjustments were made by tweaking the prompt or finetuning the weight of influence for the input images. One Fooocus' important feature is the negative prompt, used to specify elements that should not appear in the generated image. For instance, to create the scene with a post-apocalyptic future atmosphere shown in Figure 4.4 after the first generations we excluded generated people and cars that were inconsistent with it. Although this function is not foolproof, it allowed for more precise control over the final output.

When minor disturbances were present in an otherwise satisfactory composition, the inpainting tool was used to make localized corrections. This feature allows users to select specific areas of the image and generate new content within those boundaries based on the text description. In Figure 4.5 it has been added a police badge on the armour of the robot.



*Figure 4.5 Police officer before and after inpainted details*

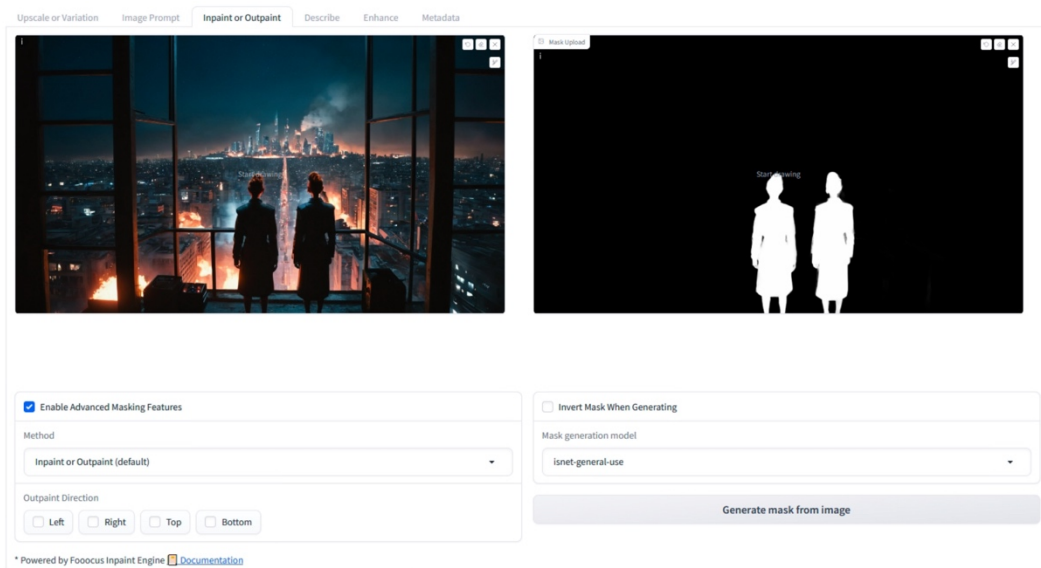
Once the final image was generated, we used the upscaling function within Fooocus to increase the resolution and improve image quality, in order to make the video animations come out much better. This process was repeated for panoramic images and when the main characters were not depicted in the scenes.

## **4.2 Creating Style Images with Consistent Characters**

Shots featuring the main characters of the story require a more intricate workflow compared to landscape scenes, as GenAI tends to create varied results with each iteration. While this variability is acceptable during the storyboarding phase, where

composition and camera angles are prioritized, it becomes problematic for final images where the characters must remain consistent across scenes.

The process begins by developing initial storyboard frames using placeholder characters that match the spatial requirements of the composition. Once the composition is established, inpainting is employed, leveraging the advanced features introduced in the “Mashblt” fork of Fooocus. This feature allows for the creation of an automatic mask (Figure 4.6) around the placeholder characters, which can then be removed while seamlessly expanding the background. This method ensures that the background remains intact and ready for the actual characters to be composited later.



*Figure 4.6 Automatic character inpainting mask*

The next phase involves posing and lighting the characters using Blender (Figure 4.7). The generated background image from Fooocus is imported as a plane, and the predesigned 3D characters are added to the scene. With properly rigged 3D models, each limb of the character can be manipulated without distorting the mesh, ensuring natural movement and positioning.

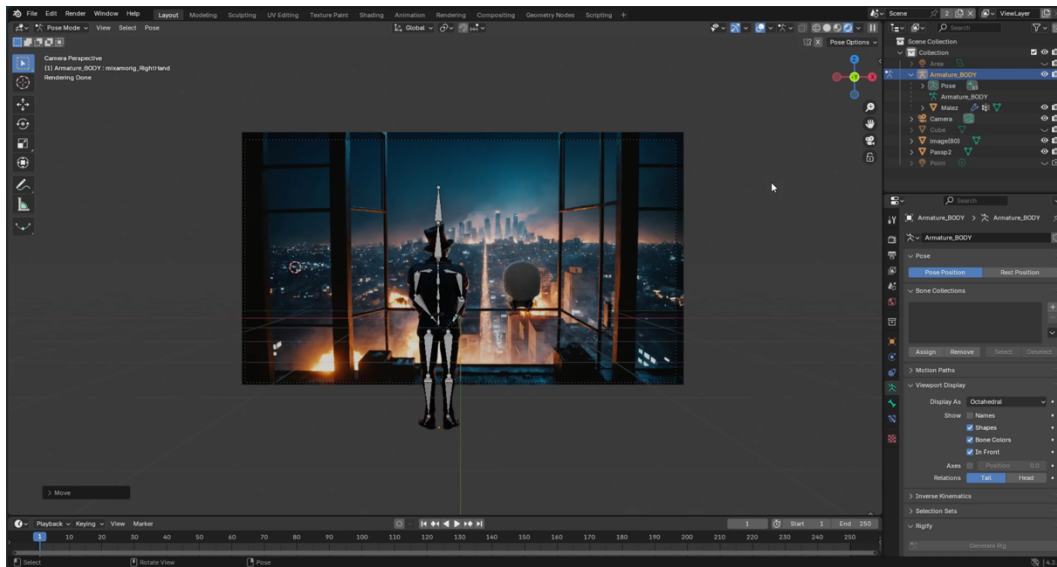


Figure 4.7 Posing models in Blender

For lighting, different techniques were used depending on the mood of the scene. In some cases, a three-point lighting setup was employed to achieve a dramatic effect. This traditional setup consists of a key light (area light) as the main light source, a fill light to soften shadows and reduce contrast (point light), and a back light to separate the characters from the background, adding depth to the scene. In other cases, a more diffuse lighting setup using HDRI maps was chosen to create a softer and more evenly distributed light source.

The background image was made emissive to act as a back light, illuminating the edges of the characters in a way that harmonizes with the environment. To enhance the lighting without altering the colors of the image, a more complex shading setup was required. Instead of using a single Emission node, the Color channel from the Image Texture node was connected to two Emission nodes, each processing color information. Their outputs were merged via a Mix Shader node. The Light Path node controlled the blending, using the Is Camera Ray connector to ensure that rays visible to the camera had a lower emission strength (in Figure 4.8 set to 1.0), while rays blocked by the 3D models received a much higher strength (in Figure 4.8 set to 9.9) to emit stronger light behind the characters. Finally, the Mix Shader was connected to the Material Output, dictating the surface behavior during rendering.



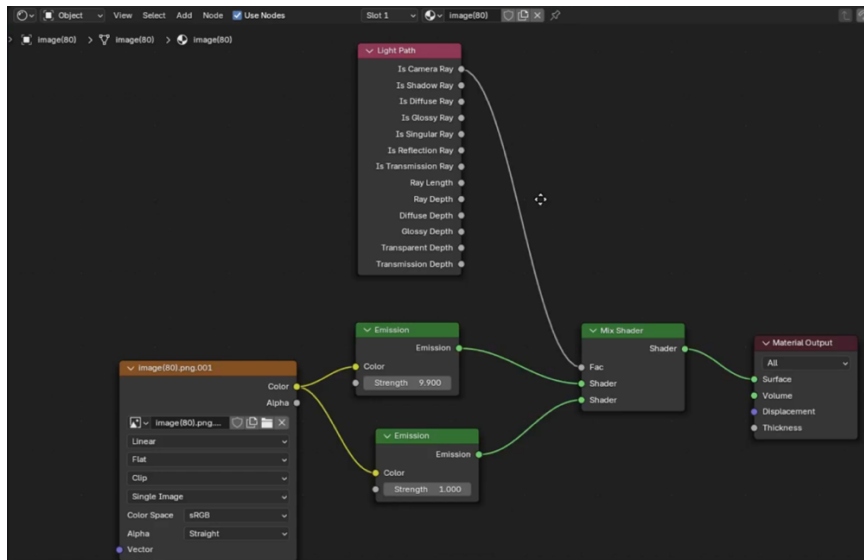


Figure 4.8 Shading set-up for background images

Once the poses and lighting were finetuned, it was preferred to render the characters separately from the background with *crypto matte node*<sup>1</sup> to proceed with a more precise compositing of the scene in Photoshop, having quick controls such as brightness, contrast, blur or to mask the characters behind elements in the foreground (Figure 4.9) and to add their shadows.

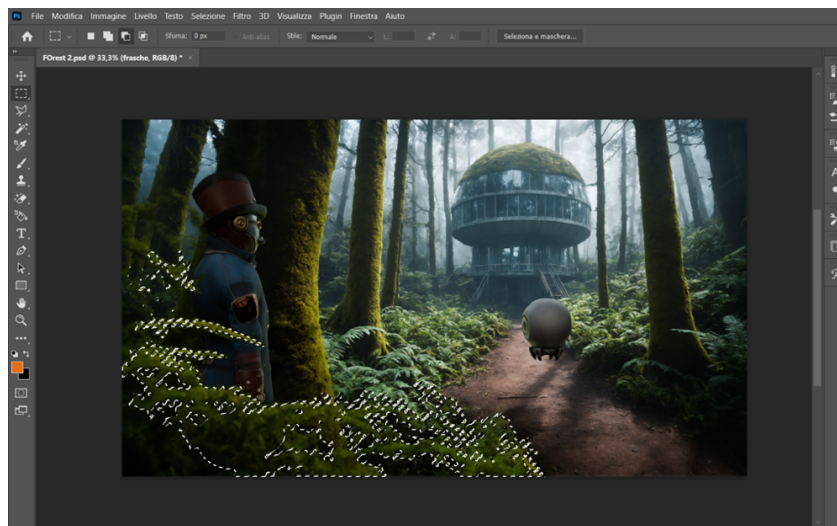


Figure 4.9 Character compositing in Photoshop

<sup>1</sup> Crypto mattes in Blender are masks that allow you to isolate specific elements of a rendered 3D scene, using attributes such as material ID, object ID, or layer.



This process provided a high level of control over the lighting and appearance of each scene, ensuring that the characters were seamlessly integrated into their environments with consistent lighting and precise shading. While this phase still required significant manual work, generative AI helped accelerate the editing process, streamlining tasks that would have otherwise been more time-consuming.

### 4.3 Video Animation

Once the style frames were obtained, the next step was transitioning to image-to-video generation. At the time this phase began, only Dream Machine by LumaLabs and Runway Gen-2 were available. Thus, experimentation with animation started using these two tools.

### 4.4 Dream Machine: Advantages and Limitations

Dream Machine recommends the use of prompts that describe the main subject, setting, and key elements, while focusing on the desired atmosphere or mood of the scene.

Prompt: solo, futuristic police robot stopping with his hand in a dangerous polluted city in the background, dark smoke, sci-fi”



*Figure 4.10 Police robot morphing its appearance*

This kind of instruction led to the generation of highly dynamic results, with fast character movements and the appearance of unexpected new elements in the background, such as buildings built from nothing. These additions demonstrated

Dream Machine’s ability to imagine and extend the scene beyond the visible frame, creating an imaginative continuity. However, despite this capacity to generate complex scenes, the results lacked the realism needed for specific purposes. In particular, the goal of generating a title sequence, which typically requires slow, deliberate camera movements and panoramic shots to immerse the viewer, was not achieved. Even after adjusting the prompt to include phrases like “slow movement” the application continued to produce results that were too fast and unpredictable.

Dream Machine also struggled to maintain character consistency in short five second clips (Figure 4.10). Characters often suffered from unwanted transformations or morphing, resulting in unrealistic transitions. With limited control over the animation process and the inability to manually correct these inconsistencies, Dream Machine was ultimately deemed unsuitable for this project.

## 4.5 Runway Gen-2: Advantages and Limitations

On the other hand, Runway Gen-2 offered a different set of features that greatly improved the creative process. One of the standout aspects of Runway is the Camera Control functionality, which allows precise adjustments to camera movements. These controls, in the form of sliders (Figure 4.11), enable the user to define movements such as panning, zooming, or rolling independently of the text prompt before the video generation begins. This feature alone sets Runway apart, offering a high degree of manual control over how the scene unfolds.

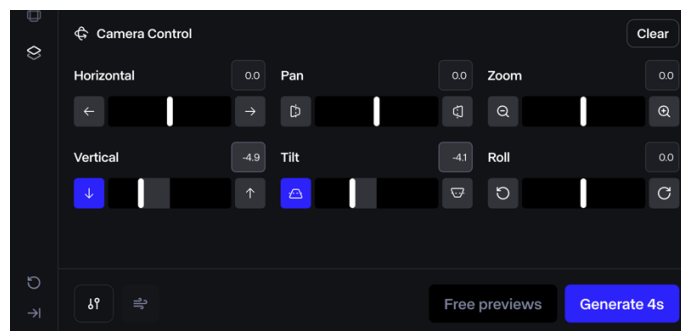


Figure 4.11 Runway Gen-2 Camera Control

Unlike Dream Machine, Runway is very much in line with the starting image. It maintains the original elements of the scene with a high degree of fidelity,

preserving consistency with a realistic visual style. Natural elements, such as flames, smoke, and ocean waves, are animated with impressive accuracy. These animations feel organic, lending a cinematic quality to environmental scenes.

However, Runway’s focus on realism comes with its own challenges. While it excels in animating static elements and environments, it struggles to animate character movements, especially those that extend beyond the original frame. For instance, in Figure 4.12, the goal was to animate the arm of a robot that initially appeared in a medium shot, raising its hand in a “stop” gesture. Despite carefully crafting the text prompt “Medium shot of an android police officer raising a stopping gesture with his palm, signaling to halt and blocking the intruder”, and manually specifying the arm movement with the Motion Brush, the result was unachievable, producing only a slight shoulder movement.

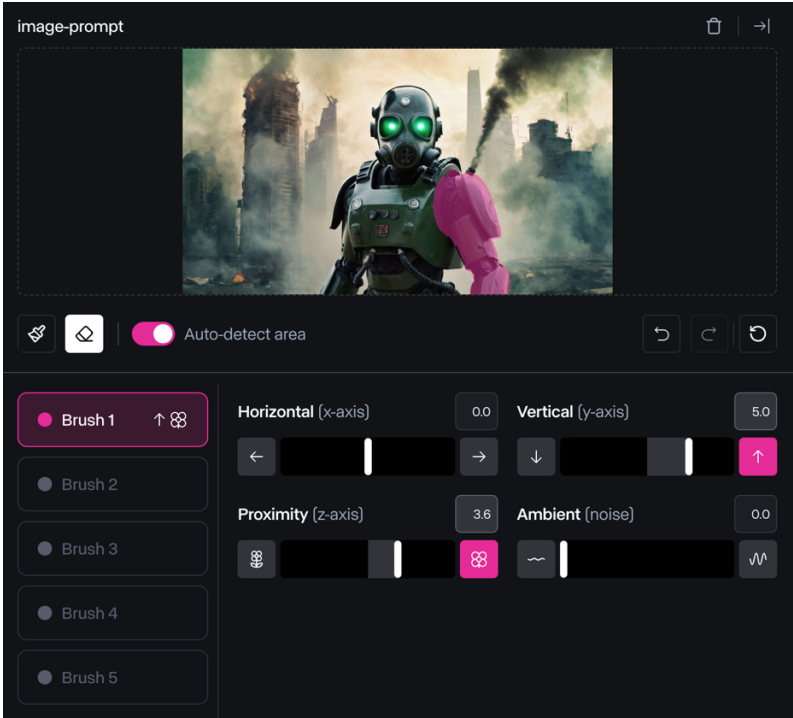


Figure 4.12 Runway Gen-2 Motion Brush

Runway Gen-2 requires a delicate balance between using text prompts and manual controls to achieve the desired effect. When motion brush and camera controls are not utilized, the overall intensity of motion can be adjusted (Figure 4.13). Setting

the camera movement intensity to an intermediate value like 5 results in a subtle animation.

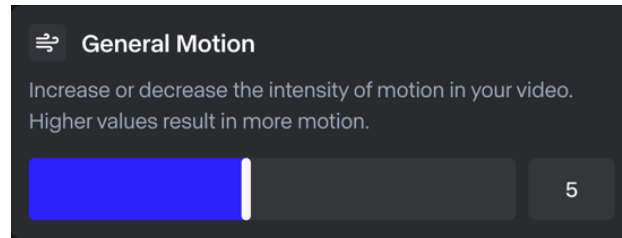


Figure 4.13 Runway Gen-2 General Motion parameter

These properties make Runway Gen-2 suitable for generating environmental scenes or establishing shots, where gentle motion is preferred.

## 4.6 Kling.AI vs Runway Gen-3: The Next Generation

In late July, Kling.AI and Runway Gen-3 were released in Europe, bringing a qualitative leap to the video generation.

Kling.AI's image-to-video demonstrated promising capabilities from the start, with an impressive ability to interpret prompts and animate characters. However, it wasn't without its issues. Strange artifacts, such as Chinese-like subtitles or unexpected shifts in styles, occasionally appeared in the final output. These could be mitigated to some extent by using negative prompts, such as "distortion, misfiguration, text, subtitles, logos" but even then, the overall quality of Kling.AI's generation remained inferior to Runway.

In Figure 4.14 and Figure 4.15 are two frames from the same Kling's generation that show these problems.

Prompt: "Medium shot: A humanoid robot wearing a Victorian-style outfit and a flying round robot with a screen and a graphical emoticon looking around in an abandoned factory. The background features towering rusting machines, overgrown with vines. Keywords: Steampunk, abandoned, industrial, overgrown, cinematic."



*Figure 4.14 Kling's generation with subtitles and morphed Passepartout*



*Figure 4.15 Continuity failure in Kling's generation*

When the same prompt was used to create a video with Runway, the results were significantly better (Figure 4.16), offering cleaner, more consistent animations.



*Figure 4.16 Correct Runway animation starting from the same input image*



Runway Gen-3 represents a substantial improvement over Gen-2, particularly in terms of how it handles dynamic scenes. Achieving the animation of the robotic police officer raising his hand, originally a challenging task, was now easily manageable. (Figure 4.17)



Figure 4.17 Successful animation of the police officer raising his hand

Four attempts were required to obtain a hand with five fingers. (Figure 4.18)



Figure 4.18 Multiple generations with the wrong number of fingers

Runway Gen-3 handles greater camera movements with more fluidity and control. However, there are still some limitations. One key issue is that the composited characters are not always recognized correctly, new characters are often generated in addition, and existing characters remain static or fail to move. (Figure 4.19)



*Figure 4.19 Failed character movement on Runway*

Sometimes, however, Runway Gen-3 struggles when camera movements reveal portions of a character that were not visible in the initial frame. For example, during a push out shot that begins with a medium frame and pulls back to reveal more of the scene, the software has difficulty generating coherent details, such as the lower portions of Phileas' body (Figure 4.20). This challenge sometimes forces a reconsideration of shot types or creative direction to accommodate these limitations.



*Figure 4.20 Inaccurate reconstructions of Phileas' legs*

However, in other instances, the software produces excellent results. In Figure 4.21, for example, Phileas is shown walking out from behind a bush, and the perspective and body movement of Phileas, initially obscured, are accurately reconstructed. Even the subtle animations of his clothing were successfully generated in this scene.



*Figure 4.21 Phileas stepping out of a bush that initially covered him*

Although Kling.AI makes it easier to achieve realistic character movements due to its 3D reconstruction capabilities, the decision was made to animate the frames with Runway Gen-3, that offered significantly higher visual. However, working with Runway required more attempts and greater effort in refining prompts, an average of ten generations per shot.



# 5 Post-Production

## 5.1 AI-Generated Voice Over

For the title sequence, a single narrator’s voice was required, and Reespecher was chosen for this task.

Reespecher offers an intuitive user experience: users input the text they want to be narrated, select from a diverse range of speakers, and the software produces a polished audio file containing the narration. The platform allows for filtering speakers by characteristics such as age (youth, adult, senior), gender, vocal tone (high, medium, low), nationality, and style of narration. Available styles include options from “passive, soft, melancholic” to “resonant, eager” although each speaker is constrained by predefined styles tailored to different contexts.

An older voice would have been preferred, but Spencer was the best choice for his “warm, podcast, narrative” style (Figure 5.1).

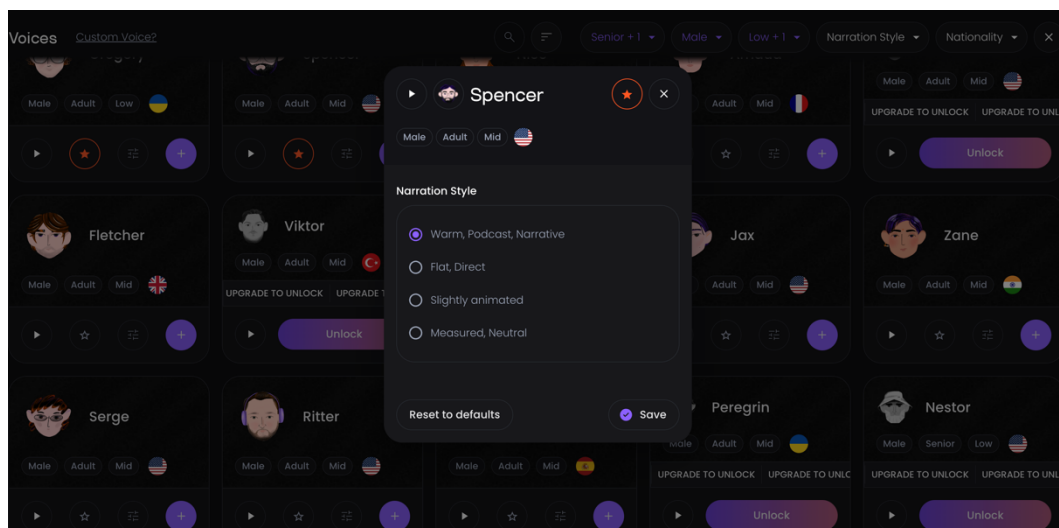


Figure 5.1 Text-to-Speech narration styles

While the results generated by Reespecher were impressively realistic, certain limitations emerged, particularly in the possibility of customizing the intonation and pacing of the narration. The initial outputs often turned out to be too rapid for the desired mood of the title sequence, which aimed for a more reflective pace to

draw viewers into the narrative. Given the importance of emotional delivery and pacing in storytelling, it became clear that human intervention was necessary to achieve the nuanced tone we were aiming for: a live voice recording was made to ensure the appropriate narrative tone.

Subsequently, we switched to Reespecher’s Speech-to-Speech feature to upload the recorded voice file into the system. This functionality allowed the software to process the input and transform speaker’s voice. This time, Nestor was chosen, a male senior voice possessing a low pitch, well-suited for the authoritative narration needed for the sequence. Unlike the standard text-to-speech function, there were no restrictions on narrative style, as the integrity of the original recording’s tone was preserved throughout the process. To enhance the voice’s power and emotional depth, the pitch was lowered by 4.5 semitones. (Figure 5.2)

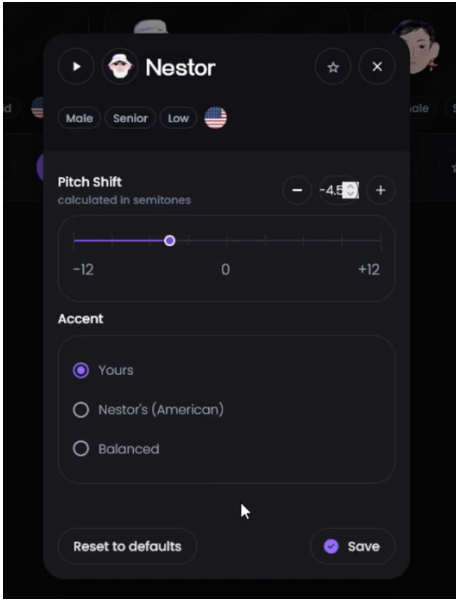


Figure 5.2 Speech-to-Speech controls

This approach effectively combined the richness of human narration with the technological advancements of AI-generated speech, resulting in a more impactful and polished final product, as if it had been recorded by a professional voice actor in a studio.

## 5.2 AI-Generated Sound Effects and Music

The sound production process ended with the generation of the sound effects and soundtrack to accompany and enhance the emotional impact of the sequence. Eleven Labs' free text-to-SFX worked very well for this purpose.

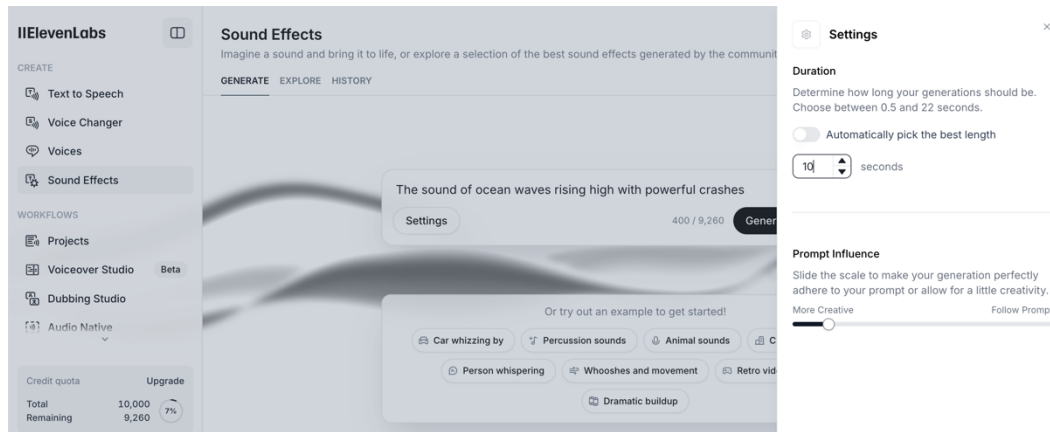


Figure 5.3 Eleven Labs' controls

In the settings (Figure 5.3) it was possible to decide the precise duration of the generated audio and the influence of the prompt, to achieve more or less creative results.

In this way, sound effects suited to the context of the scenes were generated in just a few attempts:

"eerie soundscape of wind howling through an abandoned city"

"the sound of ocean waves rising high with powerful crashes"

"the ambient noise of a forest waking up at dawn"

"crackling flames and intense fire mixed with noises of distant city"

Udio was chosen for the composition of the soundtrack. After several attempts at generation, two different pieces of music were layered, as it was not possible to generate one that evolved correctly over time.

# 6 Video Editing

Before moving on to the final editing phase, the clips that were successfully generated were upscaled using Topaz AI, an AI-powered software. This tool allowed for a significant resolution increase, from 1280×768 to 3840×2160, aligning the footage with broadcasting standards. The upscaling not only enhanced the resolution but also improved the overall visual clarity and sharpness of the material.

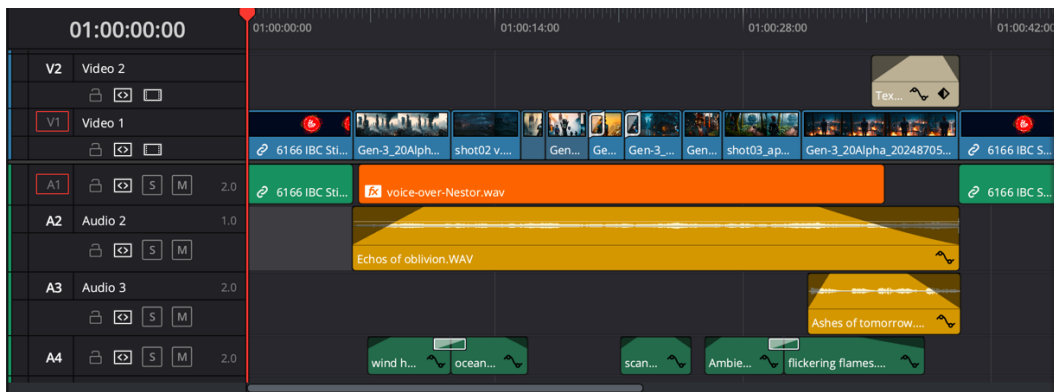


Figure 6.1 DaVinci Resolve's timeline

For the final assembly of the video sequence, DaVinci Resolve (Figure 6.1) was employed using a traditional editing workflow. All the upscaled clips were imported into a timeline where imperfections were trimmed, and some transitions were made to ensure the visuals flowed seamlessly. The clips were carefully arranged to match the pacing of the accompanying music and voice-over. Fades in and fades out were applied to the sound effects to tie the different clips together and, finally, the series title were added at the end of the sequence.

# 7 Conclusions

## 7.1 Summary of Results

This research explored the impact of emerging generative AI technologies in the media production process. Despite initial expectations of producing an entire animated TV episode solely using available generative AI tools, the project demonstrated that AI is not yet capable of autonomously creating a full film. Throughout each production phase, human intervention was essential to iteratively refine the generated content and overcome limitations.

In pre-production, LLMs like ChatGPT and Gemini proved to be valuable creative assistants. They helped generate ideas, build character profiles, and sketch the storyline. However, they struggled with writing engaging dialogue, making the involvement of human storytellers essential to weave empathy and nuance into the narrative. As the interaction with AI is still context-blind, relying on datasets rather than real world understanding, it tends to fall into biases and hallucinations. Therefore, a significant amount of human creativity and intuition remains crucial to overcome these shortcomings.

Generative AI excels in previsualization, as text-to-image technologies like DALL·E and Firefly are excellent tools for quickly producing custom moodboards and concept art, even for users without advanced technical skills. However, generating a high-fidelity storyboard that accurately reflects preproduction decisions often requires numerous iterations. The AI frequently misinterprets the user's instructions, focusing on certain details rather than capturing the overall composition or desired context. This results in an iterative process of prompt refinement and additional manual editing.

In terms of video animation, GenAI is still in an early stage of development. Rendered clips are limited in quality and duration, lasting five to ten seconds. For a TV series, it was necessary to merge multiple generations while maintaining stylistic and character consistency, ensuring that characters' appearances didn't change from scene to scene. Text-to-video generation tools were unpredictable and lacked control, making image-to-video workflows more reliable, particularly using

Runway Gen-3. In this project, tools like Meshy and 3D Studio.AI were used to generate 3D models of the characters based on initial drawings. After cleaning up imperfections, these 3D models were posed and composited using Blender and Photoshop over 2D backgrounds, which were effectively generated using Fooocus, an interface of Stable Diffusion. Fooocus allowed for the combination of multiple visual inputs with textual prompts, finetuning the influence of each.

Runway Gen-3 made it easy to generate high-quality establishing shots from style images. However, there were challenges when it came to animating scenes with character movement and complex actions, leading to cuts or re-evaluations of certain planned shots.

In post-production, voice over and sound effects were added. Text-to-speech tools did not provide the necessary control over tone and narrative flow, so audio tracks were recorded and then processed with Reespecher, producing high-quality voiceovers. Tools like Eleven Labs and Udio performed well for generating sound effects and music through textual descriptions, though it remained challenging to describe the progression of a long track with just a few characters.

The clips were selected, upscaled, and then edited using DaVinci Resolve to cut defects, add transitions and final titles. The trailer, lasting 40 seconds, demonstrates both the creative potential and current limitations of generative AI within a broadcast production setting.

## **7.2 Future Developments and Recommendations**

This thesis establishes a foundation for major advancements in the broadcasting industry, marking a substantial shift in how pre-production, production and post-production processes are approached. As GenAI technologies become more sophisticated and integrated, the creative industry is expected to increasingly adopt these tools to enhance video production workflows. However, it's important to recognize that we are still far from a scenario where machines can replace human professionals: human creativity and emotions remain irreplaceable elements that ensure content resonates with audiences in meaningful ways. Human contribution remains critical to ensure the quality of the final product and to skillfully integrate

the different software tools, each offering unique services that require diverse skill sets to use effectively.

Looking ahead, one of the most promising developments would be the testing of more versatile and integrated software like LTX Studio, a recently launched platform designed to manage all stages of pre-production in a single environment. This overcomes several challenges we faced in this project, such as issues with character consistency across different generations. LTX Studio offers the ability to maintain uniform character traits within a storyboard and integrates complex compositional sketching directly onto a canvas, an approach that significantly speeds up the design process for difficult scenes that AI struggles to comprehend. Its multi-user editing features are also invaluable in team-based production, facilitating a unified visual style across all generations.

Many advancements are also expected to address some of the current limitations of GenAI video production, particularly in areas like scene duration and animation fidelity. Advanced software such as Sora will enable the production of enhanced video sequences longer than the current 5 to 10 seconds, opening the door to more complex and polished broadcast-ready content. In parallel, animation tools that we have already tried such as Kling.AI are making significant progress day by day, for instance with the latest lip-sync features. As these systems improve, the quality and believability of AI-driven animations will improve, making them more realistic.

The integration of GenAI into post-production workflows also looks promising. Adobe's recent demos predicted inspiring features that will soon be integrated into the traditional media editing applications such as Photoshop and Premiere Pro. These advancements are expected to automate labor-intensive tasks like rotoscoping and compositing, significantly shortening production cycles, allowing professionals to focus more on creative decisions rather than technical adjustments. In today's film industry, the demand for high-quality visual effects continues to rise, but they come with significant financial and resource constraints. For instance, dynamic simulations for elements like water effects, explosions or large crowd scenes are currently handled by large teams of highly specialized VFX artists. These processes are among the most expensive parts of any production. However, with advancements in GenAI, these effects could soon be generated

algorithmically. Future GenAI tools are expected to integrate physics-based simulations, which would create realistic, complex effects in far less time and with a fraction of the manpower traditionally required. Additionally, tools like the new Runway Gen-3's video-to-video conversion offer unprecedented creative possibilities. These features can instantly reimagine live-action footage in a completely different visual style, making it possible to create vastly different looks from a single take. This technology has the potential to create new artistic genres and forms of entertainment, offering unique experiences that traditional methods cannot replicate.

Beyond these technical advancements, there is an entire field of opportunity in prompt engineering, which is currently a relatively underexplored aspect of AI-driven production. At present, crafting prompts for GenAI systems involves significant creativity, intuition, and trial-and-error. There is no universal formula for perfect results, and each platform requires its own unique approach. In the future, GenAI systems will likely be more capable of understanding complex human instructions, allowing for more intuitive control over multimodal content creation that seamlessly integrates text, images, sound, and video.

However, as AI capabilities continue to grow, addressing the ethical concerns surrounding the creation of misleading content and the misuse of images or text for illegal purposes will be crucial. Regulations and guidelines that govern the responsible use of these technologies need to be strengthened. Additionally, methods to ensure greater transparency in AI-generated content and to mitigate bias and errors will be essential, ensuring that ethical and cultural values are respected.

In conclusion, the future of GenAI for production centre is promising but requires careful consideration of both technological possibilities and ethical responsibilities. As these systems evolve, the integration of AI and human creativity will likely lead to innovation in storytelling and visual media, with limitless possibilities for the future of broadcast and entertainment.



# 8 Bibliography

- [1] “An A.I. Hit of Fake ‘Drake’ and ‘The Weeknd’ Rattles the Music World.” [Online]. Available: <https://www.nytimes.com/2023/04/19/arts/music/ai-drake-the-weeknd-fake.html>
- [2] T. Davenport and R. Bean, “The Impact of Generative AI on Hollywood and Entertainment.” [Online]. Available: <https://sloanreview.mit.edu/article/the-impact-of-generative-ai-on-hollywood-and-entertainment/>
- [3] Anguiano Dani and Beckett Lois, “How Hollywood writers triumphed over AI – and why it matters.” [Online]. Available: <https://www.theguardian.com/culture/2023/oct/01/hollywood-writers-strike-artificial-intelligence>
- [4] “EU publishes its AI Act: Key steps for organizations.” [Online]. Available: <https://www.dlapiper.com/en-am/insights/publications/ai-outlook/2024/eu-publishes-its-ai-act-key-considerations-for-organizations>
- [5] “AI Media Production Lab 2024 Accelerator Project.” [Online]. Available: <https://show.ibc.org/accelerator-project-ai-media-production-lab>
- [6] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive Growing of GANs for Improved Quality, Stability, and Variation,” Oct. 2017.
- [7] A. Brock, J. Donahue, and K. Simonyan, “Large Scale GAN Training for High Fidelity Natural Image Synthesis,” Sep. 2018.
- [8] T. Karras, S. Laine, and T. Aila, “A Style-Based Generator Architecture for Generative Adversarial Networks,” Dec. 2018, [Online]. Available: <http://arxiv.org/abs/1812.04948>
- [9] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and Improving the Image Quality of StyleGAN,” Dec. 2019, [Online]. Available: <http://arxiv.org/abs/1912.04958>
- [10] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks,” Mar. 2017, [Online]. Available: <http://arxiv.org/abs/1703.10593>

- [11] S. Bengesi, H. El-Sayed, M. K. Sarker, Y. Houkpati, J. Irungu, and T. Oladunni, “Advancements in Generative AI: A Comprehensive Review of GANs, GPT, Autoencoders, Diffusion Model, and Transformers,” *IEEE Access*, vol. 12, pp. 69812–69837, 2024, doi: 10.1109/ACCESS.2024.3397775.
- [12] X. Zeng, C. He, and Y. Jiang, “Investigating the Advancements in Generative Models,” in *Proceedings - 2023 International Conference on Artificial Intelligence and Automation Control, AIAC 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 347–351. doi: 10.1109/AIAC61660.2023.00037.
- [13] “Hello GPT-4o.” [Online]. Available: <https://openai.com/index/hello-gpt-4o/>
- [14] “Introducing Gemini.” [Online]. Available: <https://blog.google/technology/ai/google-gemini-ai/>
- [15] P. Mirowski, K. W. Mathewson, J. Pittman, and R. Evans, “Co-Writing Screenplays and Theatre Scripts with Language Models: An Evaluation by Industry Professionals,” Sep. 2022, [Online]. Available: <http://arxiv.org/abs/2209.14958>
- [16] “Plotdot.” [Online]. Available: <https://plotdot.ai/>
- [17] “Generative Fill - AI Fill in Image - Adobe Photoshop.” [Online]. Available: <https://www.adobe.com/products/photoshop/generative-fill.html>
- [18] “AI Image Extender: Generative Expand in Photoshop.” [Online]. Available: <https://www.adobe.com/products/photoshop/generative-expand.html>
- [19] “AI-powered video editing tools now in Premiere Pro.” [Online]. Available: <https://www.adobe.com/products/premiere/ai-video-editing.html>
- [20] “Generative AI Has a Visual Plagiarism Problem.” [Online]. Available: <https://spectrum.ieee.org/midjourney-copyright>
- [21] “Foocus.” [Online]. Available: <https://github.com/mashb1t/Foocus>
- [22] “Video generation models as world simulators.” [Online]. Available: <https://openai.com/index/video-generation-models-as-world-simulators/>

- [23] “Bringing generative AI to video editing workflows in Adobe Premiere Pro.” [Online]. Available: <https://blog.adobe.com/en/publish/2024/04/15/bringing-gen-ai-to-video-editing-workflows-adobe-premiere-pro>
- [24] S. Xu *et al.*, “VASA-1: Lifelike Audio-Driven Talking Faces Generated in Real Time,” Apr. 2024, [Online]. Available: <http://arxiv.org/abs/2404.10667>
- [25] “Dream Machine LumaLabs.ai.” [Online]. Available: <https://lumalabs.ai/dream-machine>
- [26] “How sampling influenced hip-hop.” [Online]. Available: <https://www.washingtonpost.com/entertainment/2023/08/24/ive-heard-that-sound-before-how-sampling-influenced-hip-hop/>
- [27] A. Agostinelli *et al.*, “MusicLM: Generating Music From Text,” Jan. 2023, [Online]. Available: <http://arxiv.org/abs/2301.11325>
- [28] Z. Borsos *et al.*, “AudioLM: a Language Modeling Approach to Audio Generation,” Sep. 2022, [Online]. Available: <http://arxiv.org/abs/2209.03143>
- [29] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, “SoundStream: An End-to-End Neural Audio Codec,” Jul. 2021, [Online]. Available: <http://arxiv.org/abs/2107.03312>
- [30] Y.-A. Chung *et al.*, “W2V-BERT: COMBINING CONTRASTIVE LEARNING AND MASKED LANGUAGE MODELING FOR SELF-SUPERVISED SPEECH PRE-TRAINING.”
- [31] “Generating audio for video.” [Online]. Available: <https://deepmind.google/discover/blog/generating-audio-for-video/>