



**Cairo University**

**Faculty of Computers and Artificial Intelligence**

**Artificial Intelligence Department**

# **Story Image Generator (IMAGITALE)**

This documentation Submitted as required For  
the degree of Bachelors in  
Artificial Intelligence Program  
Computers and Artificial Intelligence  
Cairo University

**By**

Ali Hisham Farouq

Mohammed Amir Mohammed

Mohammed Abu Bakr Mustafa

Hisham Mohammed Helmy

Sahar Hamdi Abdulhafeez

Natalie Monged Mories

**Supervisors**

**Dr. Mohammed Al-Ramli**

Computer Science Department

Faculty of Computers and Artificial Intelligence

Cairo University

**T.A. Belal Tareq**

Information Technology Department

Faculty of Computers and Artificial Intelligence

Cairo University

Cairo, July 2024

# Acknowledgment

We want to convey our profound gratitude and extend our appreciation to the many individuals who have supported and guided us through our Graduation project.

First, we sincerely thank our supervisor, Professor Mohammed El-Ramli, for his insightful comments, patience, helpful information, practical advice, and unceasing ideas that have always helped us in our project. His immense knowledge, profound experience, and professional expertise have enabled us to reach this point and introduce our IMAGITALE project. Working and studying under his guidance was a great privilege and honor. Without his support and guidance, this project would not have been possible.

Special thanks to TA. Belal for helping us by providing us with the information and resources needed for Studying and following up with us to explain everything that is not understandable to us.

We would also like to extend our gratitude to the faculty members. Their tireless efforts in imparting knowledge, providing valuable contributions to our education, and continuous support throughout our academic journey have been immensely valuable. Their academic and professional expertise have contributed significantly to our growth and learning. Moreover, we would like to express our sincere appreciation to the Faculty of Computers and Artificial Intelligence for accepting us into the graduate program.

# Abstract

This project introduces an innovative system (IMAGITALE) leveraging Artificial Intelligence (AI) to generate images based on textual story descriptions. Users input their narratives, from which the system extracts essential elements to serve as prompts for a generative AI model (Stable Diffusion). Subsequently, the model crafts images that visually depict the essence of the story. Complemented by a user-friendly web interface, the system offers writers and authors a streamlined approach to image creation, enhancing productivity, fostering creativity, and reducing manual effort. Extensive evaluation across a diverse array of narratives underscores the system's proficiency in generating both accurate and imaginative images. User feedback further validates the system's efficacy and ease of use, affirming its potential to revolutionize the creative process for writers and authors.

# Table of Contents

Acknowledgment	2
Abstract	3
List of Figures	6
1. Introduction	7
2. Methodologies	9
2.1 Related Work	9
2.1.1. Storytelling AI: A Generative Approach to Story Narration	9
2.1.1.1. IMAGITALE Vs Storytelling AI	10
2.2. Dataset	14
2.2.1. Dataset for Base SDXL.	14
2.2.2. Manually Collected Dataset for Fine Tuning.	14
2.3. Model Selection	16
2.3.1. Text Generator	16
2.3.2. Image Generator	18
2.3.3. Proposed Model	21
2.3.4. Stakeholders	23
2.4 Model Training and Evaluation.	24
2.4.1. Methods To Solve Consistency.	24
2.4.1.1. Fine-tuning Stable Diffusion XL with Dream Booth and LoRA.	24
2.4.1.2. Creating Consistent Representations, Storing Metadata.	29
2.4.1.3. Merging Fine-Tuning and Metadata Method.	30
2.4.1.4. Batch Generation.	31
2.4.2.4. Prompt Engineering.	32
2.4.3.4. Control Nets.	32
2.4.4.4. Reference Image.	34

3. Results.	35
3.1. A Tale of Donald by Base SDXL.	36
3.2. A Tale of Donald by Integration Fine Tuning with metadata.	37
3.3. A Tale of Mickey by Base Model.	38
3.4. A Tale of Mickey by Integration Fine Tuning with metadata.	39
3.5. Control Nets Results.	40
3.6. Prompt Engineering.	41
4. IMAGITALE Website.	42
4.1. Importance of IMAGITALE Website.	42
4.2. IMAGITALE Website Pages.	43
5. Conclusion.	48
6. References.	49

# List of Figures

Figure 1 System Architecture of Storytelling AI.....	9
Figure 3 CycleGAN Architecture .....	11
Figure 4 SDXL Example.....	13
Figure 5 StackGAN, BigGAN, and CycleGAN Example .....	13
Figure 6 Samples from Dataset.....	15
Figure 7 SDXL Architecture Consists of Base and Refiner.....	18
Figure 8 SDXL Base + Refiner Example.....	19
Figure 9 SDXL Base Example.....	19
Figure 10 SDXL Architecture (Text to Image) .....	20
Figure 11 IMAGITALE Architecture.....	21
Figure 12 Dream Booth Results.....	25
Figure 13 A Tale of Donald by Base SDXL .....	36
Figure 14 Tale of Donald by Integration Fine Tuning with metadata .....	37
Figure 15 A Tale of Mickey by Base SDXL. ....	38
Figure 16 A Tale of Mickey by Integration Fine Tuning with metadata .....	39
Figure 17 Control Nets Results by 2 prompts.....	40
Figure 18 Prompt Engineering Results .....	41
Figure 19 IMAGITALE Website .....	43

# 1. Introduction

In the age of rapid technological advancement, the realm of Artificial Intelligence (AI) has emerged as a frontier where imagination meets innovation. Through the lens of pixels, we delve into a world where lines of code breathe life into pixels, birthing stories that captivate and inspire. In this project, we embark on a journey through the narratives woven by the AI generation, encapsulated in the frames of their photos.

As we navigate through the pixels, we encounter a tapestry of tales, each one a testament to the boundless potential of AI-driven creativity. These stories transcend the limitations of human imagination, pushing the boundaries of what is conceivable. From the depths of virtual landscapes to the intricacies of abstract compositions, the photographs captured by AI reflect a convergence of artistry and algorithmic ingenuity.

Yet, beyond the pixels and the algorithms lies a deeper truth—a truth that transcends the realm of ones and zeros. It is a truth that speaks to the essence of humanity itself—the capacity to create, to imagine, and to inspire. In the hands of AI, this capacity is magnified, amplified, and redefined, giving rise to a new era of artistic expression.

As we immerse ourselves in the stories of the AI generation, we are reminded of the profound impact of technology on the human experience. Through their lens, we glimpse the future—a future where creativity knows no bounds and where the boundaries between the real and the artificial blur into insignificance. In this brave new world, the stories of the AI generation serve as beacons of inspiration, guiding us towards a future where imagination knows no limits.

within these pages offer but a glimpse into the vast tapestry of narratives woven by the AI generation. They are stories of innovation, of creativity, and of the enduring human spirit. As we embark

on this journey through pixels and algorithms, let us embrace the possibilities that lie ahead and celebrate the transformative power of AI in shaping the narratives of tomorrow.

Among these developments stands IMAGITALE, a pioneering system that harnesses AI to generate images based on textual story descriptions. IMAGITALE represents a significant advancement in the realm of content creation, offering writers and authors a novel tool to streamline the process of visual storytelling.

At the core of IMAGITALE lies a revolutionary premise: users input their thoughts, and the Language Processing Model (API ChatGPT) brings their narratives to life. And weaves tales with unparalleled eloquence and insight, stories are born—rich in depth, emotion, and intrigue. This capability holds immense potential for enhancing the creative process, enabling writers to transcend traditional image creation methods and explore new avenues of expression. Yet, this is just the beginning.

These stories undergo a transformative journey, channeled through the meticulous lens of our promotion model, shaping them into compelling messages that captivate and compel. Distilled with precision, each story becomes a potent promotion, ready to ignite curiosity and inspire action.

But the crescendo of our odyssey unfolds as these stories take visual form, courtesy of our visionary photo generation model (Stable Diffusion XL). With each line of prose transformed into pixels, a new world materializes—one where words transcend their textual confines to become vibrant images that speak volumes.

In this seamless fusion of storytelling and visual expression lies the magic of AI, unveiling a realm where imagination knows no bounds and creativity reigns supreme.



## 2. Methodology

### 2.1. Related Work:

#### 2.1.1. Storytelling AI: A Generative Approach to Story Narration:

The Storytelling AI project aims to generate short stories accompanied by visuals, achieving this through three main sub-goals. Firstly, users input a text prompt to initiate story generation. Then, images complementing the story text are generated, followed by a neural style transfer to give the images an illustrated appearance.

To accomplish these tasks, three generative models are utilized. Firstly, a language model, specifically OpenAI's GPT-2, is fine-tuned on a dataset of short stories from the Brothers Grimm. Secondly, two text-to-image synthesis techniques are explored: StackGAN, which decomposes the problem into manageable sub-problems, and BigGAN, which offers class-conditional image synthesis. Finally, neural style transfer is applied

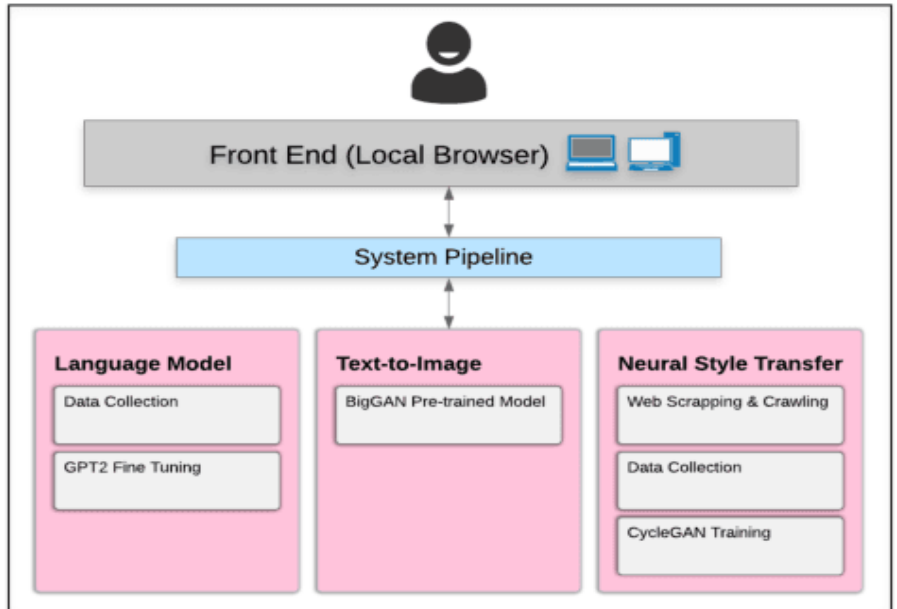


Figure 1 System Architecture of Storytelling AI

using the CycleGAN model to give the generated images an illustrated aesthetic.

### **2.1.1.1. IMAGITALE Vs Storytelling AI:**

- **Story Generation:**

Storytelling uses OpenAI's GPT-2 for text generation due to its expertise in language modeling, flexibility for fine-tuning specific datasets, and ability to generate high-quality and scalable text. By fine-tuning a dataset of short stories by the Brothers Grimm, GPT-2 aligns with the project's focus on narrative storytelling, making it an ideal choice for crafting engaging and personalized short stories.

IMAGITALE uses OpenAI's GPT-3.5 for text (story) Generation as GPT-3.5's larger size, more extensive training data, improved context understanding, and adaptability make it generally superior for story generation tasks as GPT-3.5 having 175 billion parameters compared to GPT-2's 1.5 billion parameters This increased size allows GPT-3.5 to capture more nuanced patterns in text and generate more coherent and contextually relevant stories. Also Due to its larger size and more extensive training data, GPT-3.5 tends to produce higher quality and more diverse story outputs compared to GPT-2. The stories generated by GPT-3.5 are often more engaging, creative, and contextually relevant.

- **Image Generation:**

In the Storytelling AI project, three models play crucial roles in the image generation process: StackGAN, BigGAN, and CycleGAN.

StackGAN utilizes a two-stage process to generate photo-realistic images based on text descriptions. The first stage generates a low-resolution image based on the text description, while the second stage refines this image to produce a high-resolution output. By decomposing the problem into manageable sub-problems, StackGAN aims to generate diverse and realistic images that closely match the provided text descriptions

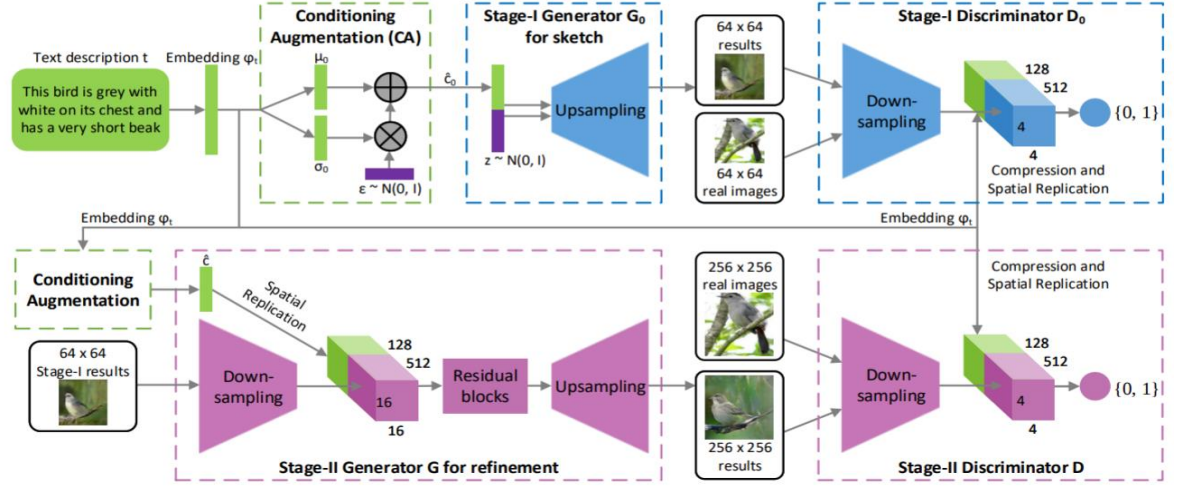
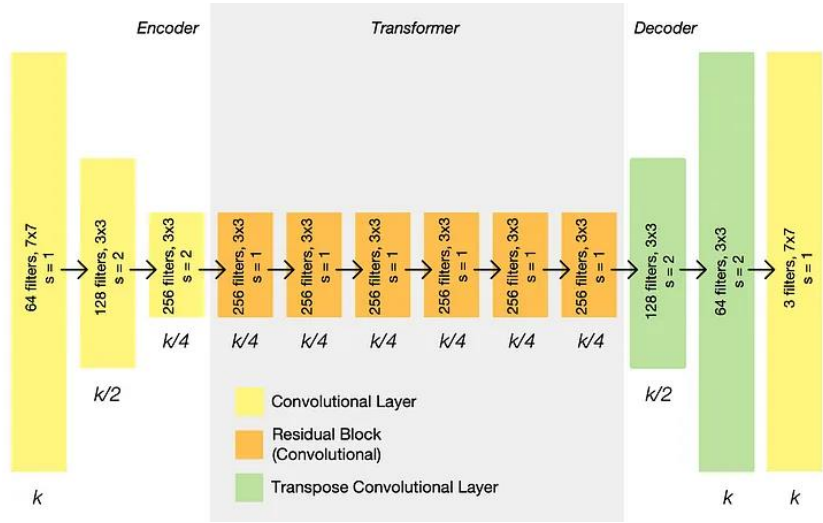


Figure 2 StackGAN Architecture

BigGAN offers class-conditional image synthesis, focusing on producing high-fidelity natural images. Trained on the ImageNet dataset, BigGAN can generate images from 1000 classes with impressive fidelity. Despite being less automated than StackGAN, BigGAN prioritizes realism and fidelity, providing superior image quality. Leveraging the pre-trained BigGAN model, the project produces visually compelling images to accompany the generated stories.

CycleGAN facilitates the transfer of illustration style to generated images in the absence of paired examples. By learning to translate images from a source domain to a target domain, CycleGAN allows for the creation of images with an

illustrated aesthetic. In the Storytelling AI project, CycleGAN is trained from scratch on a dataset comprising realistic images and manually collected illustrated images. This enables the model to learn the mapping from realistic images to illustrated ones, resulting in images that resemble illustrations.



**Figure 3 CycleGAN Architecture**

IMAGITALE uses one model only for Image Generation (Stable Diffusion XL), as Stable Diffusion XL (SDXL) marks a transformative leap in image generation, boasting an impressive resolution of up to  $1024 \times 1024$  pixels—significantly surpassing its predecessors. This remarkable increase in pixel count allows SDXL to capture finer details with unparalleled fidelity, from textures to facial features, resulting in imagery characterized by sharper edges, smoother gradients, and more natural color transitions.



**Figure5 StackGAN, BigGAN, and CycleGAN Example**



**Figure4 SDXL Example**

## **2.2. Dataset:**

### **2.2.1. Dataset for Base SDXL:**

The SDXL dataset is a proprietary collection of data carefully curated for training purposes. Its contents are not publicly disclosed, akin to safeguarding a secret recipe for an extraordinary cake. While the exact composition remains confidential, we can provide a metaphorical glimpse into the dataset's magnitude and capabilities. It is a much larger model. In the AI world, we can expect it to be better. The total number of parameters of the SDXL model is 6.6 billion, previous iterations were trained on the LAION 5B dataset and the LAION 5B dataset supplemented with the LAION-NSFW classifier. Notably, the newest version boasts training on an even larger dataset, showcasing advancements in data volume and diversity.

### **2.2.2. Manually Collected Dataset for Fine Tuning:**

We downloaded images of Disney characters from Google. We made sure that the downloaded images had a white background to easily extract the characters details. Those collected images were later passed to BLIP to analyze and extract captions from it.

- We Collected images for 2 Disney Characters (Mickey and Donald).
- Collected Dataset Size is 31 images.
- BLIP Steps:
  - Open and preprocess the image.
  - Fine-tune the model using the dataset.
  - Generate a caption for the image. Write the image file name and caption as a JSON entry in the metadata file.
  - Save the fine-tuned model.

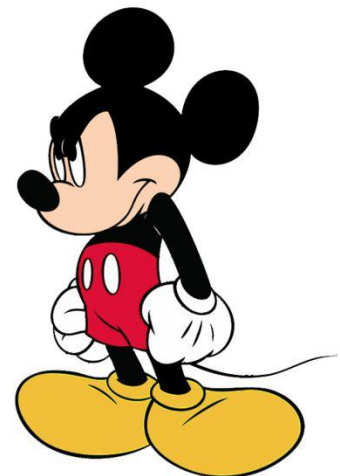
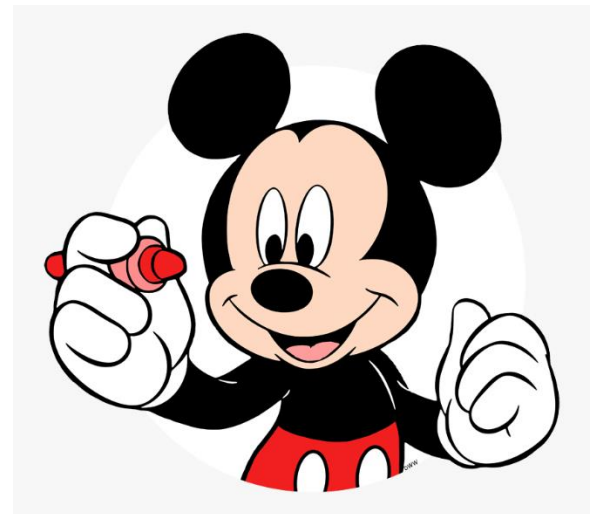
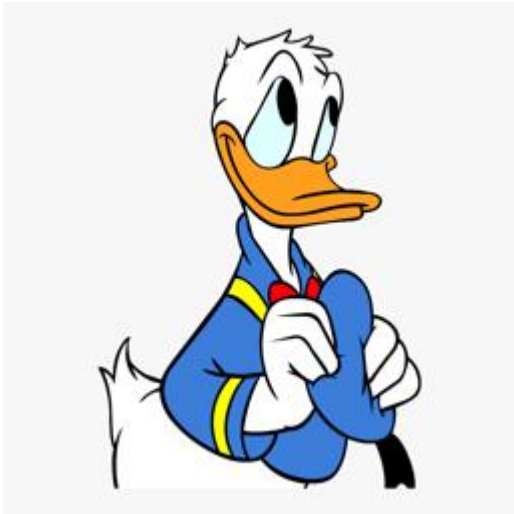


Figure 6 Samples from Dataset



## 2.3. Model Selection:

### 2.3.1. Text Generator:

- **OpenAI GPT-3.5:**

IMAGITALE leverages GPT-3.5, a cutting-edge language model developed by OpenAI, to generate stories dynamically based on user inputs. GPT-3.5 represents the forefront of natural language processing technology, equipped with advanced capabilities to understand, and generate human-like text across a wide range of topics and styles.

Choosing GPT-3.5's API for our project's story generation was a no-brainer. It's the pinnacle of natural language processing tech, offering versatility, seamless integration, and top-notch quality. With GPT-3.5, we can create captivating narratives across genres effortlessly. Its scalability ensures our project can handle any workload, while continual updates guarantee it stays ahead of the curve. In short, GPT-3.5 is the ultimate tool for crafting immersive and engaging stories, making it the perfect choice for our project.

And this process in IMAGITALE, there is a small journey from user input to the emergence of fully formed stories is akin to traversing uncharted territories, where unveils narratives waiting to be discovered. It all begins with a simple act: the user inputs a description, igniting the spark of creativity that sets our process in motion.

As the user inputs a description, GPT-3.5's journey begins. The model undergoes tokenization, breaking down the input into manageable units called tokens. These tokens are then encoded into numerical representations, enabling the model to understand the essence of the text.

With its pre-trained knowledge, GPT-3.5 delves into the context and semantics of the input, analyzing not only individual tokens but also their



relationships and patterns. This contextual understanding forms the bedrock for the model's subsequent actions.

Employing self-attention mechanisms, GPT-3.5 prioritizes the most relevant parts of the input sequence, focusing its processing power where it's needed most. This keen attention to context allows the model to generate responses that are coherent and contextually appropriate.

Now, the generation process unfolds. Drawing upon its understanding of the input and the broader linguistic landscape, GPT-3.5 predicts the most probable next tokens in the sequence. It iteratively refines its predictions, considering both the input prompt and the tokens it has already generated.

During the generation process, GPT-3.5 may utilize sampling or beam search algorithms to select the next tokens. Sampling injects an element of randomness into token selection, while beam search prioritizes the most likely sequences based on predefined criteria.

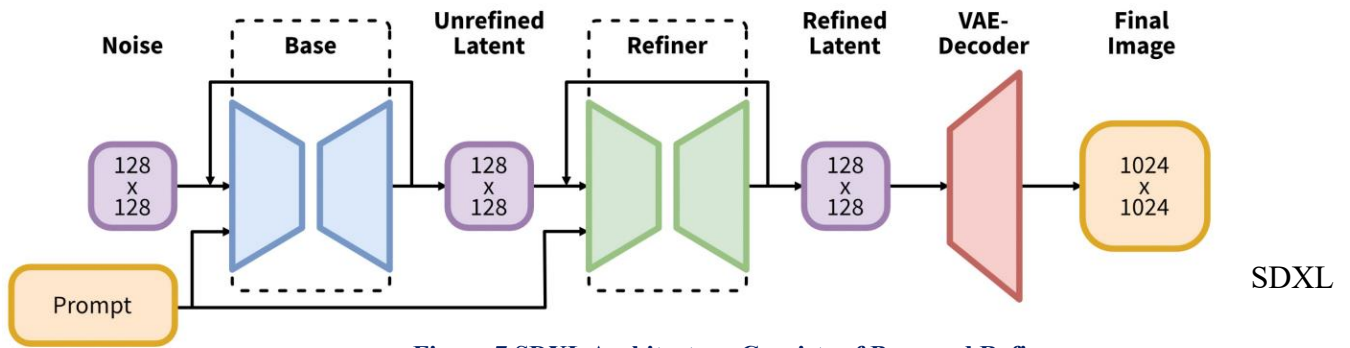
Finally, with the complete generation process, GPT-3.5 presents the user with a fully realized story, crafted from the depths of its linguistic prowess. This output reflects the model's ability to understand and respond to natural language prompts with nuance and creativity, marking the culmination of our journey.

In essence, GPT-3.5's journey from input to output showcases its remarkable capacity to transform user input into compelling narratives, underscoring its position as a leading force in natural language processing technology.

### 2.3.2. Image Generator:

- **Stable Diffusion XL (SDXL):**

Stable Diffusion XL (SDXL) stands as a monumental advancement in image generation technology, revolutionizing the creative landscape with its dual-model approach. By seamlessly integrating the base and refiner models, SDXL imbues images with unprecedented realism, breathing life into every pixel. Enhanced with OpenAI's CLIP ViT-L, SDXL ensures effortless prompting, while its innovative image size conditioning expands its capabilities. We selected as the cornerstone of our project, SDXL complements the narrative prowess of GPT-3.5 by bringing stories to vivid visual fruition. With a default image size of 1024x1024, SDXL sets a new standard for fidelity, ushering in a new era of boundless creativity and visual storytelling.



**Figure 7 SDXL Architecture Consists of Base and Refiner**

embodies a revolutionary paradigm in image generation, employing a tandem of models to sculpt each masterpiece. With the base model laying the groundwork for composition and the refiner model delicately infusing intricate details, SDXL brings images to life with an unrivaled level of precision and sophistication. This dynamic



**Figure 8 SDXL Base + Refiner Example**



**Figure 9 SDXL Base Example**

duo heralds a new era of visual storytelling, where every pixel is meticulously curated to perfection.

Embarking on the journey of image generation with SDXL, our process is a symphony of innovation and precision. Beginning with a story prompt from GPT-3.5 (will be Encoded by CLIP Encoder), we delve into the realm of possibility. Random noise is meticulously sampled and encoded by a Variational Auto Encoder (VAE), yielding a latent representation ( $Z$ ). This, alongside the prompt embeddings, navigates through the intricate pathways of the U-Net. Here, the removal of noise to sculpt the envisioned image unfolds with each iteration, guided by the watchful eye of the scheduler. As noise dissipates, the latent representation ( $Z_{\text{hat}}$ ) emerges, a testament to the refined vision taking shape. With a seamless transition to the Decoder, the culmination is reached – a vivid tableau of our story brought to life in the form of captivating images as shown in [Figure 9].

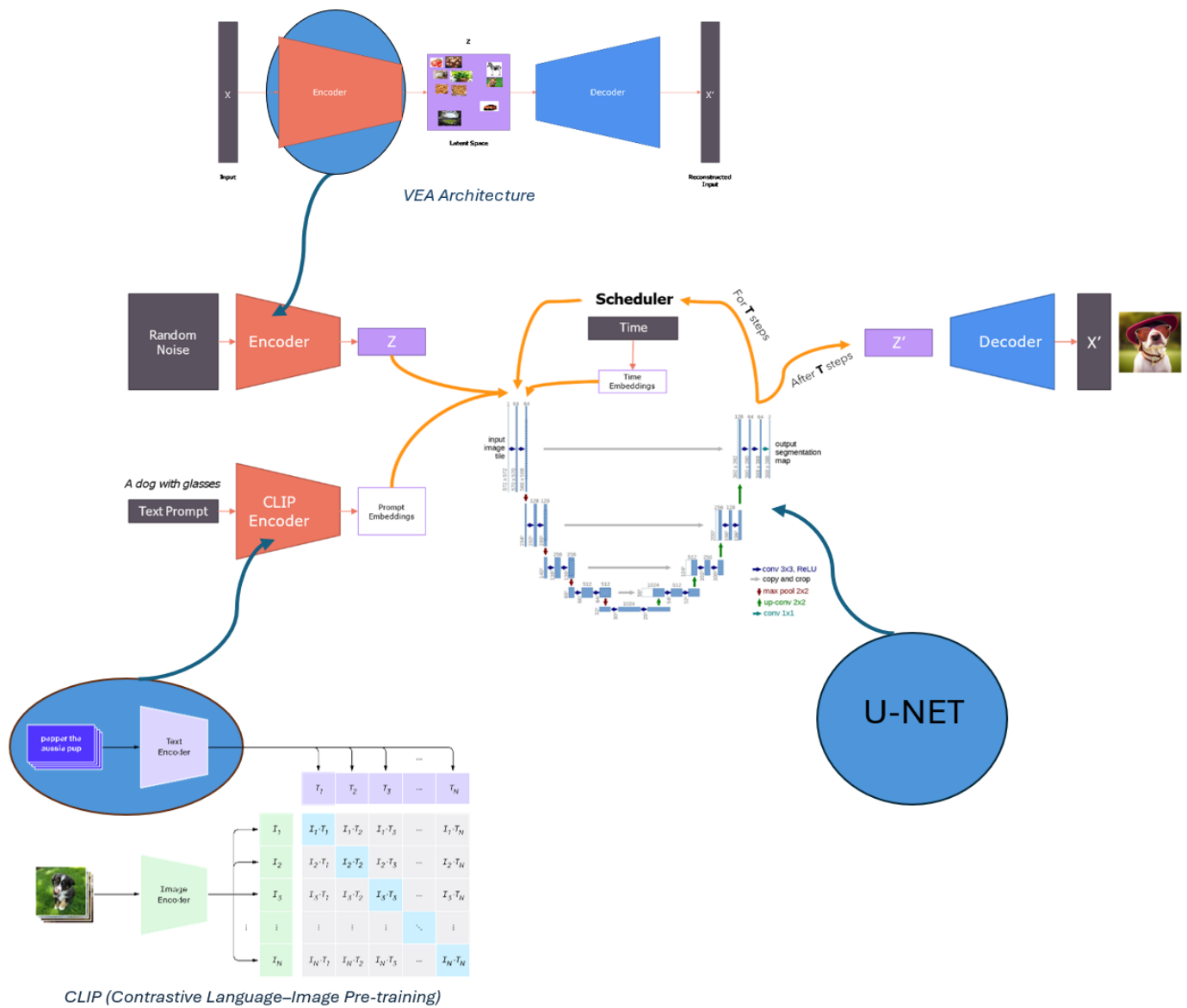


Figure 10 SDXL Architecture (Text to Image)

### 2.3.3. Proposed Model:

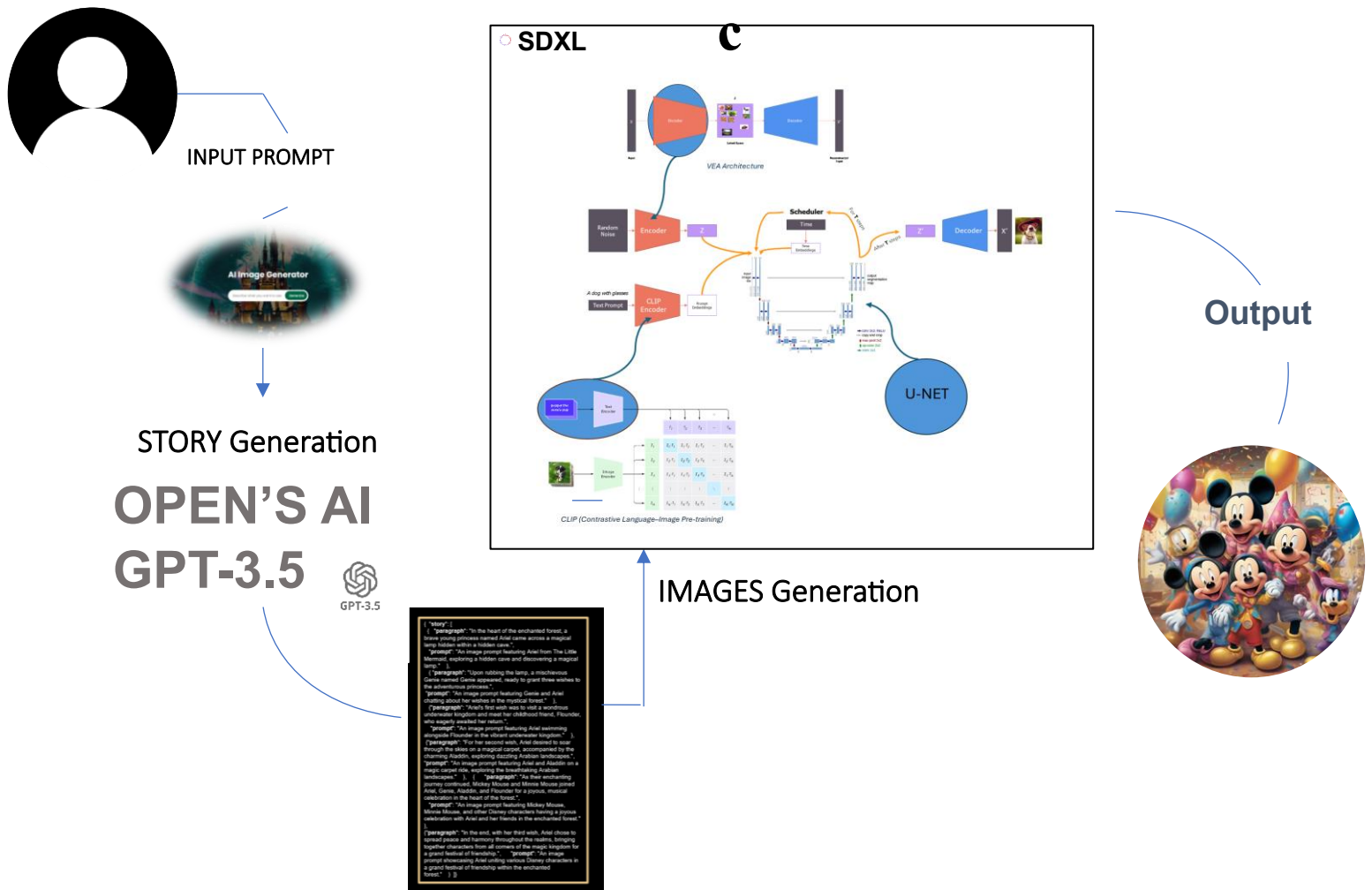


Figure11 IMAGITALE Architecture

The architecture shown in the diagram illustrates the integration of OpenAI's GPT-3.5 with the SDXL (Stable Diffusion XL) model for generating stories and images based on an input description.

- 1. Input Description:** The process starts with a user providing a textual description or query that sets the stage for generating both a story and corresponding images.
- 2. Story Generation (GPT-3.5):** The input description is first processed by OpenAI's GPT-3.5, a large language model known for its advanced text generation capabilities. GPT-3.5 takes the description and generates a detailed and coherent story based on it. This story forms the narrative foundation for the subsequent image generation. Additionally, GPT-3.5 divides the generated story into a structured format and outputs it as a JSON file. This JSON file serves as the input for the SDXL model, ensuring that the story elements are organized and easily interpretable for image generation.
- 3. Image Generation (SDXL):** The generated story is then fed into the SDXL model. SDXL stands for Stable Diffusion XL, a sophisticated model designed for high-quality image synthesis. The SDXL model incorporates several key components:
  - **Encoder:** The encoder processes the textual input, transforming it into a format suitable for further processing by the model.
  - **CLIP Feature:** This component utilizes Contrastive Language-Image Pre-training (CLIP) to align the generated story with relevant visual concepts. CLIP helps in understanding the story context and extracting key features for image generation.
  - **Scheduler:** The scheduler manages the diffusion process, which iteratively refines the generated images. This involves several steps of noise

addition and removal, controlled by the scheduler to ensure high-quality output.

- **U-Net:** A U-Net architecture, commonly used in image processing, further refines the images by performing upsampling and downsampling operations. This helps in enhancing image details and ensuring coherence with the story.
  - **Decoder:** Finally, the decoder converts the processed features back into visual images, resulting in detailed and story-consistent outputs.
- 
- **Output:** The final output consists of the generated images, which visually represent the narrative created by GPT-3.5. These images are designed to closely match the themes and elements of the story, providing a cohesive and immersive experience for the user.

#### 2.3.4. Stakeholders:

- **CHILDREN:**

In today's digital age, a concerning trend has emerged: children are reading less frequently than previous generations. This decline poses a threat to their reading proficiency and overall enjoyment of literature, crucial for personal growth and academic success.

One of the primary culprits behind this trend is the lack of engagement in current storytelling techniques. Recognizing this challenge, we propose an innovative approach to revitalize storytelling methods and reignite the passion for reading among children.

- **WRITERS AND AUTHORS:**

Writers and Authors often find themselves grappling with the limitations of traditional creation methods, which fail to fully capture the essence of their narratives. This struggle manifests as a barrier to their creative expression, hindering their ability to convey stories with the depth and richness they envision. Writers yearn for a more intuitive and dynamic approach to visual storytelling, one that transcends the confines of conventional techniques and unlocks new realms of imagination.

## **2.4 Model Training and Evaluation**

In the part of image Generation in IMAGITALE we used SDXL which is a pre-trained model "trained on billions of parameters", SDXL is a great model for Image generation and it achieved high scores in this part, but on the other hand, IMAGITALE is a story generator which means that the characters should be consistent in all the story, SDXL is weak on this point, it generates images based on the prompt but the consistency between the prompt is not always guaranteed, actually most of the times there is a low consistency between the prompts.

So, we tried different methods to solve the problem of the Consistency:

### **2.4.1. Methods To Solve Consistency:**

#### **2.4.1.1. Fine-tuning Stable Diffusion XL with DreamBooth and LoRA**

##### **- What is DreamBooth?**

DreamBoth, introduced in 2022 by the Google research team, represents a significant advancement in the field of generative AI,



particularly in the realm of text-to-image models like Stable Diffusion.

It is called DreamBooth because, in the words of the Google researchers:

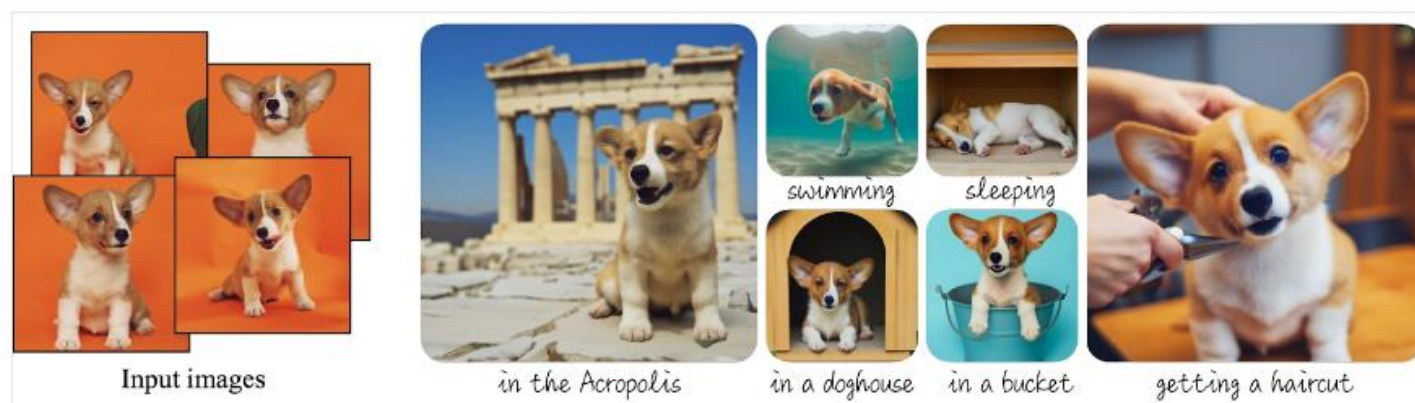


Figure 12 Dream Booth Results

*"It's like a photo booth but captures the subject in a way that allows it to be synthesized wherever your dreams take you."*

DreamBooth allows you to inject a specific custom subject that the fine-tuned model then becomes specialized at rendering in different ways. So, in a sense, it opens the possibility to create your image generator focused on a particular person, character, object, or scene.

DreamBooth requires only a few (typically 3-5) images of the subject to train the model effectively. Once trained, the model can place the subject in a myriad of settings,

scenes, and poses, limited only by the user's imagination.

## - **Fine-Tuning The SDXL Model with Accelerate and Transformers:**

Fine-tuning the SDXL model with Accelerate involves optimizing the model for specific data using the Accelerate library. After generating image captions with the BLIP model, the training script uses Accelerate for efficient distributed training. Techniques like mixed precision and gradient checkpointing speed up the process. The fine-tuned model is then uploaded to Hugging Face Hub for generating high-quality images based on narrative prompts.

### 1. **One/Two Character Fine Tuning:**

**Setting Up:** Before running the DreamBooth script, we should set up some variables that we will use to run the script.

1- Setup and Imports.

2- Paths the Definition.

- For one Character we Defined the dataset directory and the path for the metadata file of each character “Donald Images”
- For the two characters “Mickey and Donald”, we defined the dataset directory and the path for the metadata file of each character. we combined the

files of the Donald dataset with those of the Mickey dataset, creating a Combined dataset to work with as a single dataset file.

### 3- Initialize BLIP Processor and Model

- Load the BLIP (Bootstrapped Language-Image Pre-training) processor and model for image captioning.

### 4- Caption Generation Function

- Define a function caption images that takes an image, processes it, and generates a caption using the BLIP model.

### 5- Create Metadata File

- Open the metadata file for writing.
- For each image in the dataset:
  - Open and preprocess the image.
  - Generate a caption for the image.  
Write the image file name and caption as a JSON entry in the metadata file.

## **With Dream Booth:**

### 6- Install and Launch Training Script:

- Install the datasets library.

- Launch the training script `train_dreambooth_lora_sd-xl.py` using the accelerate tool with specified parameters (e.g., model paths, dataset, output directory, training settings).

#### 7- Upload Trained Model to Hugging Face Hub:

- Define user-specific variables (username and repository name).
- Upload the trained model from the output directory to the Hugging Face Hub.
- Display a link to the uploaded model.

#### 8- Load and Use the Model:

- Load the VAE (Variational Autoencoder) and Stable Diffusion pipeline with LoRA weights for image generation.
- Clear CUDA cache to free up GPU memory.
- Generate an image based on a prompt and display it.

9- Refine the Generated Image:

- Clear CUDA cache again to free up GPU memory.
- Load the Stable Diffusion XL Img2Img pipeline refiner.
- Refine the previously generated image with a prompt and specified seed.
- Display the refined image.

10- Clean Up:

- Clear CUDA cache to free up GPU memory.

### **2.4.1.2. Creating Consistent Representations, Storing Metadata:**

To maintain consistency in the appearance of characters and scenes across different scenes, you can store metadata about characters and scenes, including seeds for the random number generator. This approach uses Stable Diffusion to generate the images and ensures that the same characters and scenes look the same whenever they reappear. Adjust the prompts and metadata as needed for your specific use case.

- 1. Metadata Storage:** Create a dictionary to store metadata for characters and scenes.

**Ex:** metadata = {  
    "characters": {},  
    "scenes": {}}

## **2. Generate Images and Save Metadata:**

Generate images and save metadata whenever a new character or scene is introduced.

### **2.4.1.3. Merging Fine-Tuning and Metadata**

#### **Method:**

We merged the previous two methods (fine-tuning and metadata storage) to achieve more stability in the generated images. By combining fine-tuning techniques using DreamBooth and LoRA with consistent metadata storage, we ensure that characters and scenes maintain visual consistency across different contexts. This integrated approach leverages the power of Stable Diffusion, DreamBooth, and LoRA to produce high-quality and stable images. The fine-tuning process allows the model to specialize in rendering specific subjects, while metadata storage ensures that the appearance of characters and scenes remains consistent. Together, these methods provide a robust framework for generating reliable and coherent images.

#### **➤ Steps to Combine the Methods:**

1. Setup and Imports.
2. Load Pre-trained Models and VAE:
  - Loading VAE and Pipeline: Load the VAE and the Stable Diffusion model and set it to use the GPU.
3. Metadata Storage:
  - Creating Metadata Dictionary: Initialize a dictionary to store metadata for characters and scenes.

#### 4. Generate Images with Metadata:

- Function to Generate Images: Define a function to generate images and store metadata for consistency.

#### 5. Generate and Save Multiple Images:

- Generating Images: Use the function to generate and save images for different prompts and scenes.

#### 6. Display Generated Images

- Displaying Images: Show the generated images to verify consistency.

#### 7. Refine Images with Stable Diffusion XL

##### Img2Img Pipeline:

- Image Refinement: Refine the generated images using the Img2Img pipeline for better quality.

### **2.4.1.4. Batch Generation:**

To display 4 prompts using batch generation in SDXL and achieve consistency, the minimum GPU memory required can vary based on several factors, including the complexity of the prompts, the resolution of the images, and the specific implementation of SDXL. However, a rough estimate for the minimum GPU memory is as follows:

- SDXL Base Model: Typically requires around 10-12 GB of GPU memory per instance.

- Batch Generation: For 4 prompts, you would need 4 instances running simultaneously.
- Considering this, you would need at least:  $12\text{GB} \times 4 = 48\text{GB}$ ,  $12\text{GB} \times 4 = 48\text{GB}$ , this calculation is an estimate, and you might need more memory to handle overhead and ensure smooth operation. Hence, having a GPU with at least 48-64 GB of memory would be advisable for batch generation with 4 prompts in SDXL to achieve consistency.

#### **2.4.2.4. Prompt Engineering:**

- Detailed Descriptions: Provide detailed and specific descriptions in your prompts. The more precise you are with the elements you want in the image, the better the model can capture those details.
- Use of Keywords: Identify and use key terms that are known to influence the model's output effectively.
- Syntax and Structure: Experiment with the structure of your prompts. Sometimes rephrasing or changing the order of words can yield better results.

#### **2.4.3.4. Control Nets:**

Control Nets are a technique designed to provide additional control over the generation process, thereby enhancing the consistency and coherence of the produced images. By leveraging Control Nets, we can guide the diffusion model more precisely, ensuring that the output adheres to specific



constraints or follows a desired pattern, even across diverse prompts.

➤ **Setting Up:**

1. Setup and Imports.

2. Checking and Setting Device:

- Check if a CUDA-capable GPU is available and sets the device accordingly. If no GPU is available, it defaults to the CPU.

3. Clearing CUDA Cache:

- Clear the CUDA memory cache to ensure optimal memory usage and prevent memory overflow issues.

4. Generate Images with Control Nets:

- Create a function to take the text prompt and uses the pipeline to generate an image.
- Influence the image generation process.
- Ensure reproducibility.
- Control the number of inference steps for image generation.

5. Generate and Display Images from Multiple Prompts:

- Iterate over a list of prompts and generates an image for each prompt.

#### 2.4.4.4. Reference Image:

Using reference images in the image generation process ensures that specific visual elements remain consistent across multiple scenes, providing a coherent and unified visual story.

➤ **Setting Up:**

1. Setup and Imports.
2. Load Models.
3. Loading the Reference Image:
  - The reference image is loaded from a local file, which will guide the consistency of generated images.
4. Defining Prompts for Each Scene:
  - These prompts describe different scenes in the story, focusing on new elements and the progression of the narrative.
5. Generate Images with Reference Image Influence:
  - Generate an image based on the prompt, using the reference image to guide the consistency of the output. Various parameters like strength, guidance\_scale, and num\_inference\_steps are tuned for optimal results.
6. Generating and Saving Images for Each Scene.

### 3. Results

We Tried Various Methods for Image Generation, as we faced some problems (Consistency) with using SDXL only for image Generation, so we had to fix these problems by using different ways to improve the story quality and make it more accurate like:

1. SDXL Base Model.
2. SDXL Fine Tuning (Single Character).
3. Integration Fine Tuning Techniques with metadata utilization.
4. Prompt Engineering.
5. Reference Images.
6. Control Nets.

In this section, we will present the results obtained from each method:

Method	Consistency	Quality	Weights
Base SDXL	Not Achieve	The Best Quality	uses Constants weights of the model
Fine Tuning Single Character	Achieving Character Details	Slightly lower than base model	updated weights for specific Characters
Integration Fine Tuning with metadata	Along with Characters details it maintains more details in the whole image	Lower than base model	updated weights for specific Characters
Prompt Engineering	Consistency increasing with more details	The same as base SDXL	Updated during training for specific characters
Control Nets	Background and the character features achieve it	The Lowest	Updated during training for specific characters
Reference Image	High Consistency “it copies the reference image”	Same as base SDXL	Updated during training for specific characters

### 3.1. A Tale of Donald by Base SDXL:

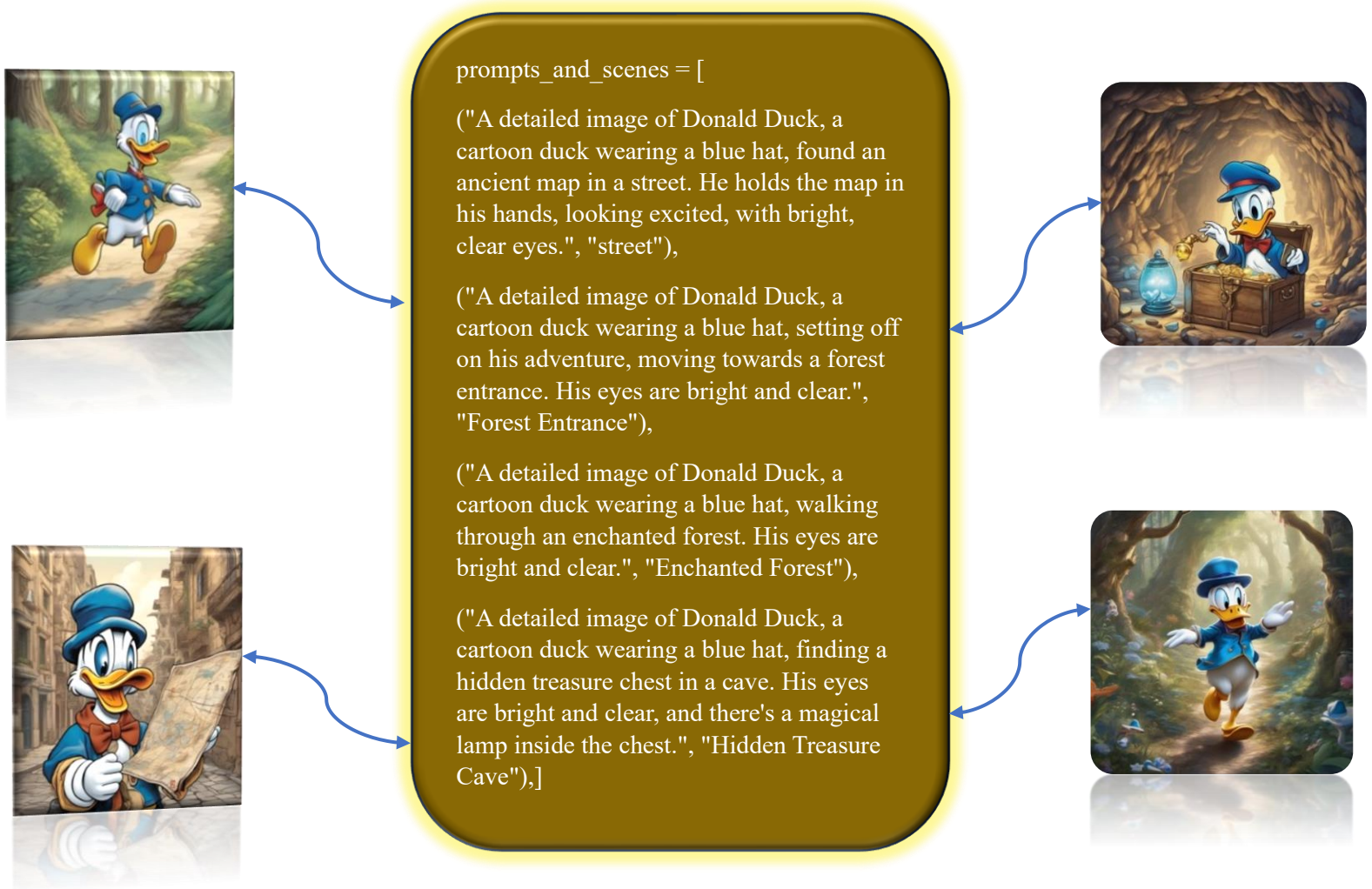


Figure 13 A Tale of Donald by Base SDXL

### 3.2. A Tale of Donald by Integration Fine Tuning with metadata:



`prompts_and_scenes = [`

("A detailed image of Donald Duck, a cartoon duck wearing a blue hat, found an ancient map in a street. He holds the map in his hands, looking excited, with bright, clear eyes.", "street"),

("A detailed image of Donald Duck, a cartoon duck wearing a blue hat, setting off on his adventure, moving towards a forest entrance. His eyes are bright and clear.", "Forest Entrance"),

("A detailed image of Donald Duck, a cartoon duck wearing a blue hat, walking through an enchanted forest. His eyes are bright and clear.", "Enchanted Forest"),

("A detailed image of Donald Duck, a cartoon duck wearing a blue hat, finding a hidden treasure chest in a cave. His eyes are bright and clear, and there's a magical lamp inside the chest.", "Hidden Treasure Cave"),]



Figure 14 Tale of Donald by Integration Fine Tuning with metadata

### 3.3. A Tale of Mickey by Base Model:

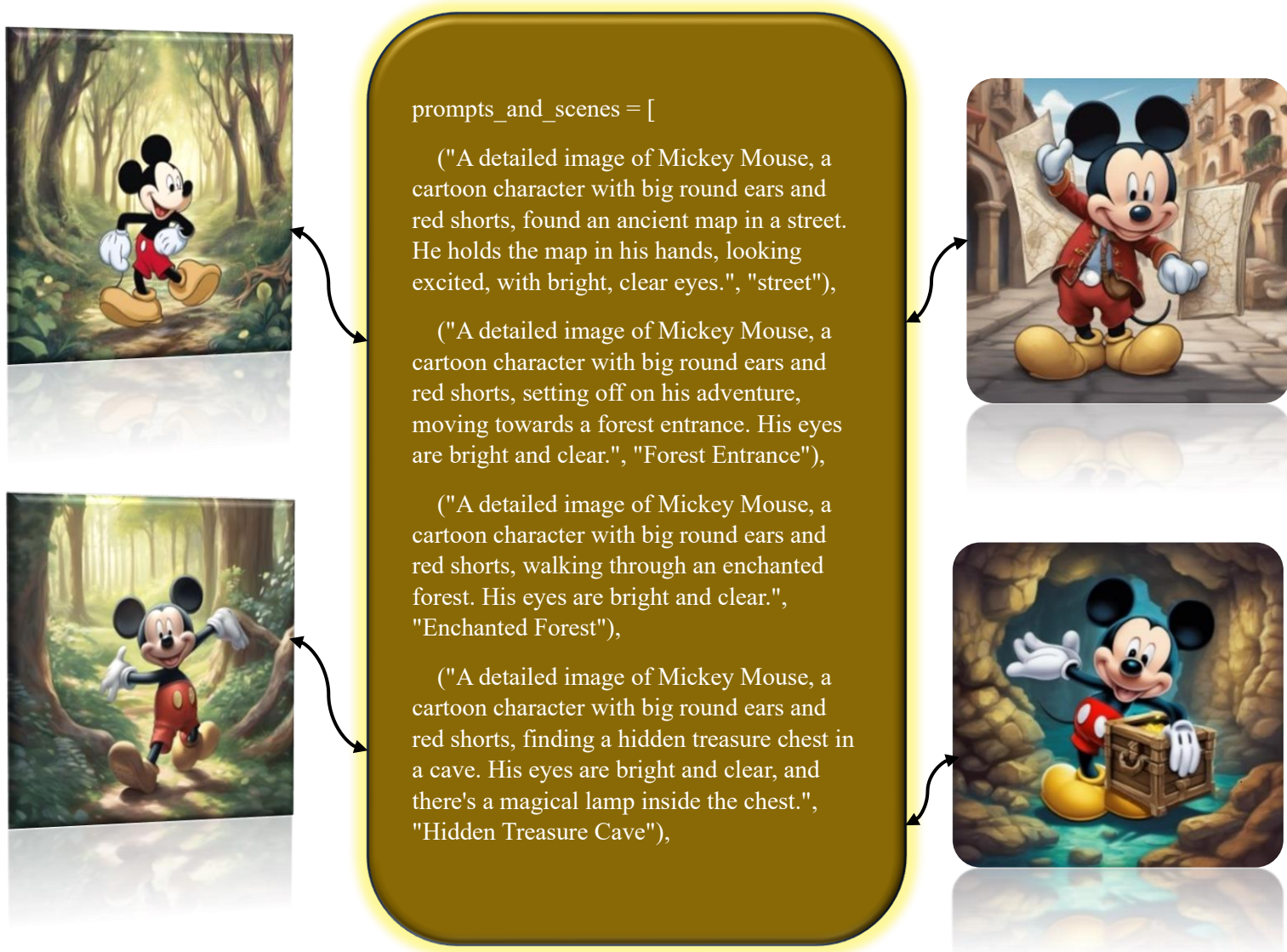


Figure 15 A Tale of Mickey by Base SDXL



### 3.4. A Tale of Mickey by Integration Fine Tuning with metadata:



### 3.5. Control Nets Results:

Spider-Man swings through the sunlit skyscrapers of New York...



Spotting masked bank robbers, Spider-Man shoots his web and ...



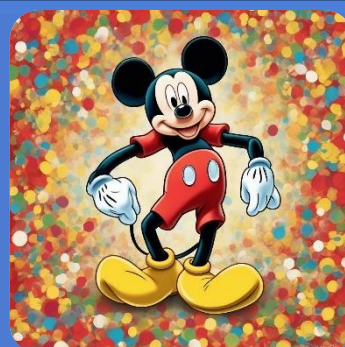
Figure 17 Control Nets Results by 2 prompts



### 3.6. Prompt Engineering:



Prompt Engineering = (“A mesmerizing and detailed scene of Mickey Mouse, dressed in a stylish scuba suit, exploring an ancient underwater city. The city is filled with vibrant coral reefs and majestic ruins, teeming with colorful marine life. Mickey's big round ears and red shorts are visible through his transparent diving helmet, and his bright, clear eyes are wide with fascination as he swims past schools of fish and shimmering underwater plants.”)



Prompt Engineering = (“An intriguing and detailed depiction of Mickey Mouse, the beloved cartoon character with big round ears and red shorts, exploring the interior of a grand, ancient pyramid. The pyramid is filled with intricate hieroglyphics, hidden passages, and glowing torches illuminating the golden walls. Mickey's bright and clear eyes shine with curiosity and determination as he uncovers hidden secrets and treasures, his hands gently



Prompt Engineering = (“A breathtaking image of Mickey Mouse, the iconic cartoon character with big round ears and red shorts, standing on the balcony of a majestic sky castle. The castle floats high above the clouds, with towers made of gleaming white marble and golden spires. Mickey looks out over the endless sea of clouds, his bright and clear eyes filled with wonder and excitement, as the sun sets in the distance, casting a warm, golden glow across the scene.”)



Prompt Engineering = (“An intriguing and detailed depiction of Mickey Mouse, the beloved cartoon character with big round ears and red shorts, exploring the interior of a grand, ancient pyramid. The pyramid is filled with intricate hieroglyphics, hidden passages, and glowing torches illuminating the golden walls. Mickey's bright and clear eyes shine with curiosity and determination as he uncovers hidden secrets and treasures, his hands gently brushing against the ancient stone carvings.”)

Figure 18 Prompt Engineering Results

## **4. IMAGITALE Website**

IMAGITALE is an innovative web platform designed to bring the magic of Disney stories to life. This project allows users to easily create and generate their own unique Disney stories and images, making the creative process accessible to everyone. The website offers an intuitive and user-friendly interface, ensuring that users can engage with the platform effortlessly, regardless of their technical expertise.

### **4.1. Importance of IMAGITALE Website:**

1. **Creative Empowerment:** IMAGITALE empowers users to unleash their creativity by providing them with the tools to craft personalized Disney stories. Whether for personal enjoyment, educational purposes, or sharing with friends and family, users can bring their imaginative ideas to life.
2. **Accessibility:** By offering a straightforward and accessible interface, Imagitale ensures that users of all ages and backgrounds can engage with the platform. The simplicity of the registration and story generation process means that even those with minimal technical skills can use the website effectively.
3. **Educational Value:** IMAGITALE can serve as a valuable educational tool, helping users, especially children, to develop their storytelling skills, creativity, and imagination. It can also be used in educational settings to inspire students to create their own narratives and explore the world of storytelling.

## 4.2. IMAGITALE Website Pages:

### ❖ Welcome Page:

- Navigation Bar: The navigation bar at the top provides easy access to different sections of the website. Here are the options:
  - ✓ Home: Takes users back to the main page.
  - ✓ Create: Likely where users can start crafting their Disney stories.
  - ✓ Discover: Possibly a section where users can explore existing stories or find inspiration.
  - ✓ My Stories: Likely a place where users can manage and view their own created stories.
  - ✓ Login: Probably for user authentication.
- Welcome Message.
- About us.
- Supervision.
- Copywrites.

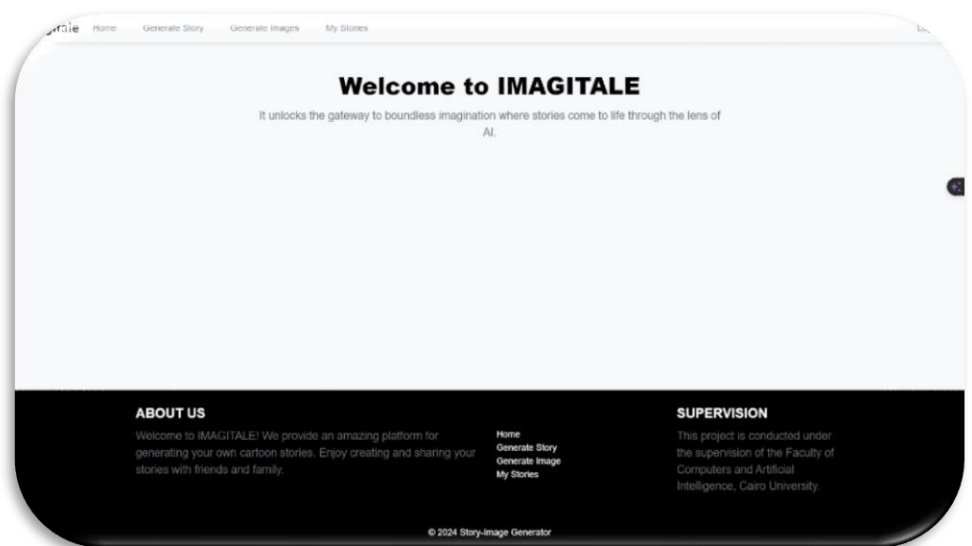
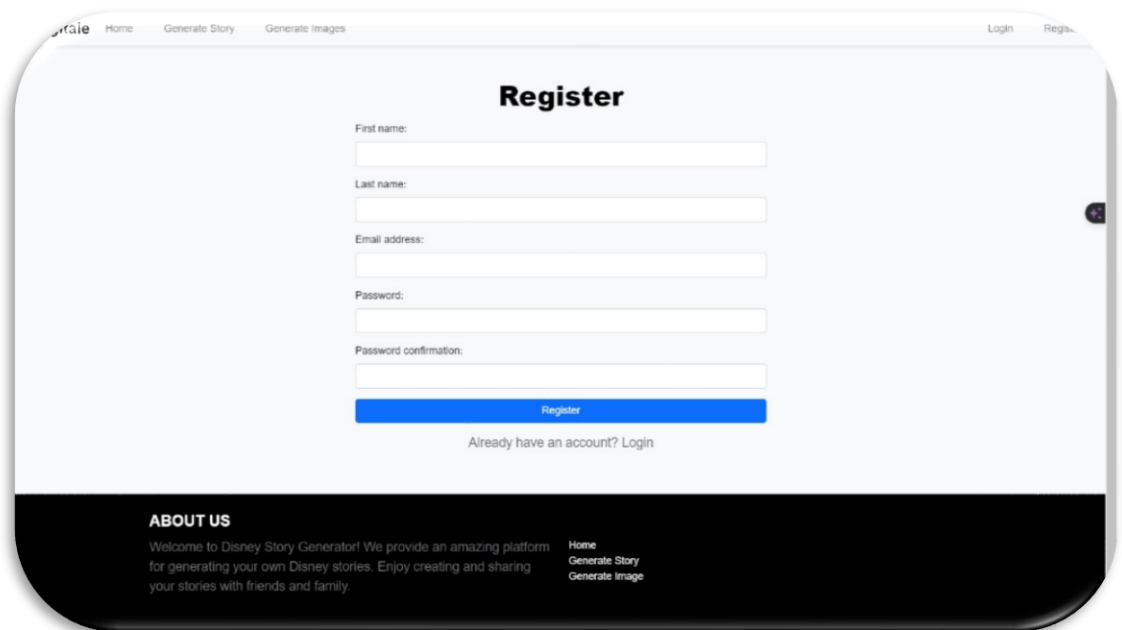


Figure 19 IMAGITALE Website

### ❖ Registration Form:

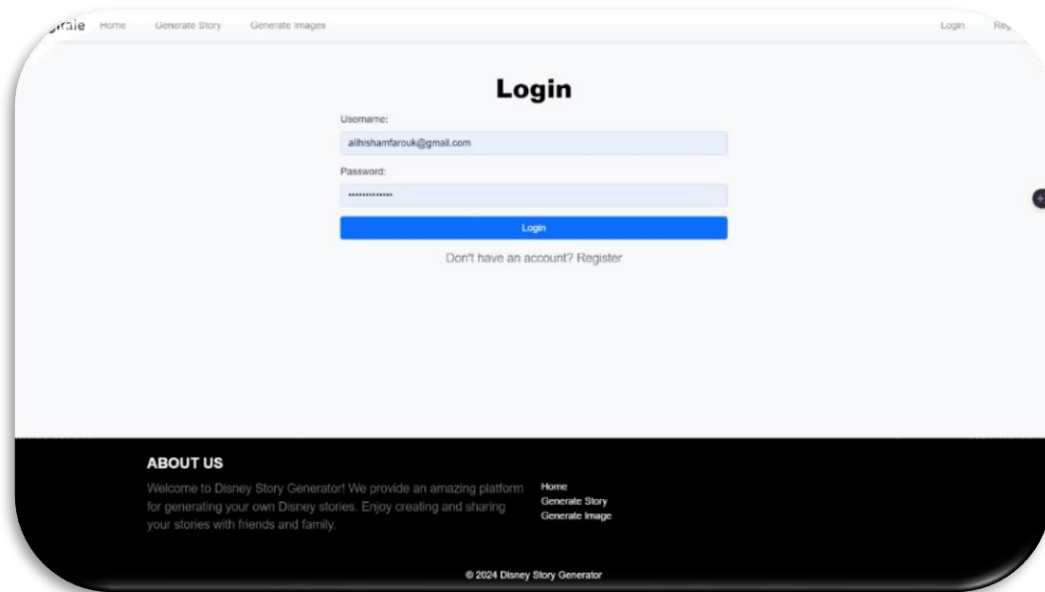
- The registration form allows new users to create an account. It collects the following information:
  - ✓ First Name.
  - ✓ Last Name.
  - ✓ Email Address
  - ✓ Password
  - ✓ Password Confirmation
  - ✓ Users fill in these details to register and gain access to the platform.
  - ✓ Register Button:
    - Below the form fields, there's a blue “Register” button. Clicking this button submits the registration form.
  - ✓ Already Have an Account? Login:
    - Beneath the registration section, there's a message inviting users who already have an account to log in. This suggests that registered users can access additional features beyond registration.



The screenshot displays a web application interface for the Disney Story Generator. At the top, a navigation bar includes links for 'Home', 'Generate Story', 'Generate Images', 'Login', and 'Register'. The main content area is titled 'Register' and contains a form with the following fields: 'First name:', 'Last name:', 'Email address:', 'Password:', and 'Password confirmation:'. Each field is represented by a white input box. Below these fields is a prominent blue button labeled 'Register'. Underneath the button, there is a link that says 'Already have an account? Login'. The footer of the page is dark and contains an 'ABOUT US' section with a welcome message and a small navigation menu with links to 'Home', 'Generate Story', and 'Generate Image'.

### ❖ Login Form:

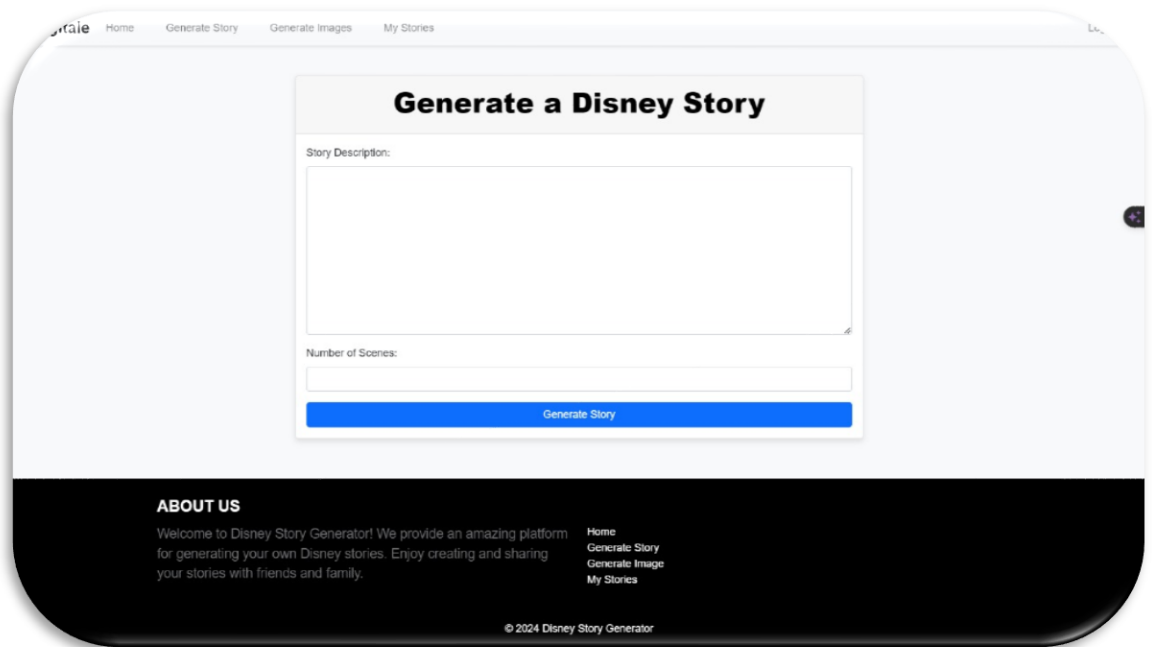
- If you already have an account, there's no need to register again. Simply go to the Login page, enter your email address and password, and you can start using the website right away.



The screenshot shows a web browser window displaying the 'Login' page of the 'Disney Story Generator'. The page has a light blue header with navigation links: 'Home', 'Generate Story', 'Generate Images', 'Login', and 'Register'. The main content area is white and features a 'Login' title. Below the title are two input fields: 'Username:' with the email 'alishanfarouk@gmail.com' and 'Password:' with masked characters. A blue 'Login' button is positioned below the password field. A link 'Don't have an account? Register' is located below the button. The footer is a dark blue bar containing an 'ABOUT US' section with a welcome message, a 'Home' link, and links for 'Generate Story' and 'Generate Image'. The copyright notice '© 2024 Disney Story Generator' is at the bottom right.

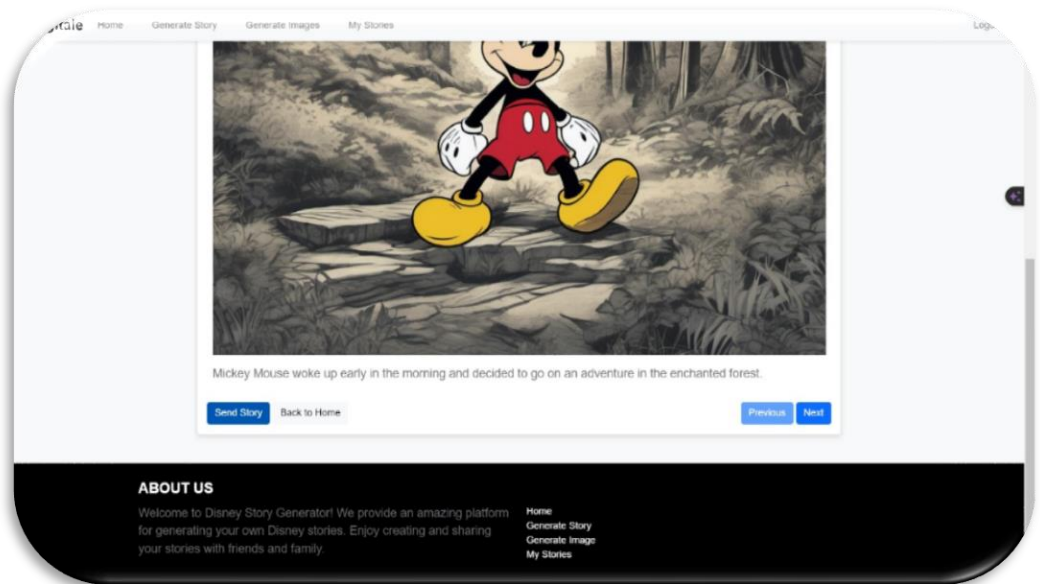
### ❖ Generate a Disney Story with Images Page :

- **Story Description:** This large text box allows users to enter a detailed description of the story they want to create. Users can be as imaginative and specific as they like, describing characters, settings, plot elements, and any other aspects they want to include in their Disney story.
- **Number of Scenes:** This input field lets users specify the number of scenes they want in their story. By entering a number here, users can control the length and structure of the story generated by the platform.
- **Generate Story Button:** Once the user has filled in the story description and the number of scenes, they can click this button to generate their Disney story. The platform will process the input and produce a unique story based on the provided details.



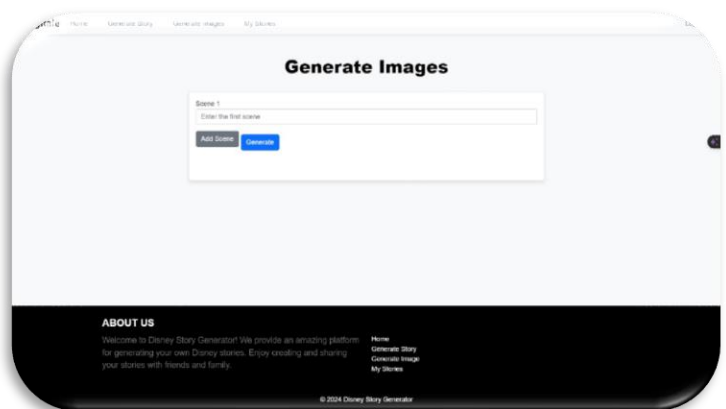
### ❖ Generated Story Viewing Page:

- Story Image: The top portion of the page displays an image related to the current scene of the story. In this example, it shows Mickey Mouse embarking on an adventure in an enchanted forest. This image is generated by the platform to visually enhance the storytelling experience.
- Story Text: Below the image, the story text describes the current scene. In this case, it reads, "Mickey Mouse woke up early in the morning and decided to go on an adventure in the enchanted forest." This text is generated based on the user's input and the number of scenes they specified earlier.
- Navigation Buttons:
  - ✓ Previous: This button allows users to go back to the previous scene of the story.
  - ✓ Next: This button lets users move forward to the next scene in the story.
  - ✓ Send Story: This button enables users to send or share the story they have created. It provides a way to share their creative work with others directly from the platform.
  - ✓ Back to Home: This button redirects users back to the homepage of the website.



### ❖ **Generate Images only Page:**

- IMAGITALE offers an image generation service only, not a full story creation.
- **Scene Input Box:**
  - ✓ **Scene 1:** Users are prompted to enter the description for the first scene of their story. This input field is where users provide the text that will be used to generate the corresponding image.
  - ✓ **Add Scene Button:** This button allows users to add additional scenes to their story. By clicking this button, users can create more input fields for subsequent scenes, enabling them to build a multi-scene narrative.
- **Generate Button:** After entering the scene description, users click this button to generate the image based on the text provided. The platform uses advanced image generation models to create visuals that match the scene description.



## 5. Conclusion

In the ever-evolving landscape of artificial intelligence, our journey through the realms of storytelling and image generation has illuminated the boundless potential of human-AI collaboration. From the inception of narratives woven by the AI generation to the visual manifestation of these tales through cutting-edge technology, we have witnessed the convergence of creativity, innovation, and imagination.

Through the collaborative efforts of pioneering systems like IMAGITALE and the transformative capabilities of models such as GPT-3.5 and Stable Diffusion XL (SDXL), we have ventured into uncharted territories of creative expression. Our exploration has not only pushed the boundaries of what is conceivable but has also redefined the paradigms of storytelling and visual artistry.

As we stand at the precipice of this brave new world, it becomes evident that the fusion of AI and human ingenuity holds the key to unlocking new realms of possibility. The stories captured within these pixels, birthed from the depths of algorithms and nurtured by the human spirit, serve as testaments to the symbiotic relationship between technology and creativity.

In closing, let us embrace the transformative power of AI in shaping the narratives of tomorrow. Let us continue to innovate, to imagine, and to inspire, as we chart a course towards a future where imagination knows no bounds and where the lines between the real and the artificial blur into insignificance. Together, let us embark on this journey with curiosity, courage, and creativity, for the possibilities are as limitless as the stories we dare to imagine.



## 6. References

<https://stable-diffusion-art.com/sd-xl-model/>

<https://huggingface.co/spaces/google/sd-xl>

<https://www.datacamp.com/tutorial/fine-tuning-stable-diffusion-xl-with-dreambooth-and-lora>

<https://www.analyticsvidhya.com/blog/2023/09/image-generation-using-stable-diffusion/>

[https://huggingface.co/docs/diffusers/en/api/pipelines/stable\\_diffusion/stable\\_diffusion\\_xl](https://huggingface.co/docs/diffusers/en/api/pipelines/stable_diffusion/stable_diffusion_xl)

<https://huggingface.co/docs/diffusers/en/using-diffusers/sd-xl#inpainting>

[https://huggingface.co/docs/diffusers/v0.26.2/en/api/pipelines/controlnet\\_sd-xl#diffusers.StableDiffusion](https://huggingface.co/docs/diffusers/v0.26.2/en/api/pipelines/controlnet_sd-xl#diffusers.StableDiffusion)

<https://techvify-software.com/what-is-stable-diffusion/>

<https://huggingface.co/spaces/google/sd-xl>

<https://anakin.ai/blog/stable-diffusion-sd-xl/#1-stable-diffusion-21-vs-sd-xl-key-differences>

<https://github.com/hkproj/pytorch-stable-diffusion>

<https://waxy.org/2022/08/exploring-12-million-of-the-images-used-to-train-stable-diffusions-imagegenerator/>

<https://waxy.org/2022/08/exploring-12-million-of-the-images-used-to-train-stable-diffusions-imagegenerator/>

<https://ieeexplore.ieee.org/search/searchresult.jsp?newsearch=true&queryText=stable%20diffusion%20>

<https://sci-hub.se/https://ieeexplore.ieee.org/document/8706212/>

<https://towardsdatascience.com/diffusion-models-midjourney-dall-e-reverse-time-to-generate-images-from-prompts-ba760f472103>

<https://architizer.com/blog/practice/tools/an-architects-guide-to-midjourney-ai-generated-imagery/>

[https://en.wikipedia.org/wiki/Stable\\_Diffusion](https://en.wikipedia.org/wiki/Stable_Diffusion)

<https://drive.google.com/drive/folders/1WiI92yaVXbaFbE9zoLnnfKvZiwYSJcPJ>

[https://storyspark.ai/?gclid=Cj0KCQjwsp6pBhCfARIsAD3GZuY6ScR3G-mFyPMHon8iXJVFDvz8aHb2F-YzrLsbvtv72qU\\_KF4sPygaAhzaEALw\\_wcB](https://storyspark.ai/?gclid=Cj0KCQjwsp6pBhCfARIsAD3GZuY6ScR3G-mFyPMHon8iXJVFDvz8aHb2F-YzrLsbvtv72qU_KF4sPygaAhzaEALw_wcB)

<https://platform.openai.com/docs/introduction>

<https://www.ikomia.ai/blog/stable-diffusion-xl-sd-xl-model>

<https://g.co/gemini/share/7f0ed33e010a>