Master of Science Thesis

# Interactive Storyboarding System Leveraging Large-Scale Pre-trained Model

# 대규모 사전학습 모델을 활용한 인터랙티브 스토리보딩 시스템

February 2023

Graduate School of Engineering
Seoul National University
Electrical and Computer Engineering Major

Sihyeon Jo

Master of Science Thesis

# Interactive Storyboarding System Leveraging Large-Scale Pre-trained Model

# 대규모 사전학습 모델을 활용한 인터랙티브 스토리보딩 시스템

February 2023

Graduate School of Engineering
Seoul National University
Electrical and Computer Engineering Major

Sihyeon Jo

# Interactive Storyboarding System
# Leveraging Large-Scale Pre-trained Model

## 대규모 사전학습 모델을 활용한
## 인터랙티브 스토리보딩 시스템

Supervisor: Seung-Woo Seo, Seong-Woo Kim

This work is submitted as a Master of Science Thesis

February 2023

Graduate School of Engineering
Seoul National University
Electrical and Computer Engineering Major

Sihyeon Jo

Confirming the master's thesis written by
Sihyeon Jo

February 2023

| | | |
|---|---|---|
| Chair | Eun Suk Suh | (Seal) |
| Vice Chair | Seung-Woo Seo | (Seal) |
| Examiner | Seong-Woo Kim | (Seal) |

# Abstract

Artificial Intelligence (AI) technologies have impacted almost every domain and system, including the entertainment industry. Although AI-based systems are expected to offer significant benefits in making content, it is still challenging to build a real-world AI application that can effectively contribute to content production. This paper presents a novel system, Gennie, that can interact with users and suggest AI-generated sketches for developing a storyboard. Gennie is implemented leveraging several large-scale pre-trained models such as GPT-2 and CLIP, which have recently achieved great success and become milestones in the field of AI. With Gennie, users can quickly visualize the composition of each scene. This paper presents the findings of a user study and visualizes the process of creating a storyboard with Gennie.

**keywords**: Storyboarding, multimodal embedding, large-scale pre-trained model

**student number**: 2019-21939

# Contents

**Abstract** **i**

**Contents** **ii**

**List of Figures** **iv**

**List of Tables** **v**

**1 Introduction** **1**

1.1 Overview . . . . . . . . . . 1

1.2 Interactive storyboarding . . . . . . . . . . 2

**2 Related work** **4**

2.1 Human-AI co-creation . . . . . . . . . . 4

2.2 AI for visual story generation . . . . . . . . . . 5

2.3 Evaluation metric for sketch research . . . . . . . . . . 6

**3 Proposed method** **8**

3.1 Interactive storyboarding system . . . . . . . . . . 8

3.2 Knowledge base preparation . . . . . . . . . . 11

3.3 Sentence-to-sketches retriever . . . . . . . . . . 14

**4 Experimental results** **16**

4.1 Experimental setup . . . . . . . . . . 16

4.2 Procedure . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 17

4.3 User research . . . . . . . . . . . . . . . . . . . . . . . . . . . . 18

4.4 Participants' strategies in interactive storyboarding . . . . . . . . . . 22

**5 Conclusion** **27**

**Abstract in Korean** **34**

# List of Figures

1.1 The proposed system's inputs and outputs. . . . . . . . . . . . . . . . 3

2.1 Milestones of sketch research in terms of tasks, datasets, and supervision. Note that self-supervised learning is also unsupervised learning. 7

3.1 Overview of the Gennie, the interactive storyboarding system. . . . . 9

3.2 Structure of the interactive storyboarding system. . . . . . . . . . . . 10

3.3 Effects of segmentation masks in producing sketch images. . . . . . . 12

3.4 Prepared sketches and their sources (©Summit Entertainment). . . . . 14

4.1 A series of scenes generated with the proposed system Gennie. . . . . 17

4.2 Samples of sketches produced by humans and Gennie in the user study. 19

4.3 Example of user's perception of the suggested sketch. . . . . . . . . . 20

4.4 Participants' responses to each question. . . . . . . . . . . . . . . . . 21

4.5 The distribution of time to complete the storyboarding task with Gennie. 21

4.6 Use case of the interactive storyboarding system. . . . . . . . . . . . 22

4.7 A feedback loop to improve the proposed system. The candidate sketches-selected sketch pair can be utilized to customize the recommendation. 23

4.8 Process of creating storyboards by participants. . . . . . . . . . . . . 24

4.9 The professional cartoonists developed ideas from Gennie's sketches. 25

4.10 Search-history of the participants. . . . . . . . . . . . . . . . . . . . 26

# List of Tables

3.1 Exemplary sketch datasets. “**s**” and “**p**” mean “sketches” and “photos”. 13

3.2 Specification of the large-scale pre-trained model (GPT-2) . . . . . . 15

3.3 Specification of the large-scale pre-trained model (GPT-3) . . . . . . 15

4.1 Statistics of the user study. The participants take significantly less time to finish the task with the proposed system on the second try. . . . . . 21

# Chapter 1

# Introduction

## 1.1 Overview

AI technologies are transforming the way we approach real-world tasks done by humans. Recent years have seen a surge in the research field of deep learning, where massive parameters are tuned to generalize on carrying out a particular task. For example, with the understanding of images, deep learning models have surpassed that of humans in several vision tasks [1, 2]. Besides, we are witnessing the possibility of AI applications for story generation [3], music composition [4], drawing [5], and so on. However, successful uses of deep learning algorithms in creative areas are raising the bar for required sensibleness and specificity, which are far below those of humans [6].

As alternative approaches to explore practical methodologies or creative user experience, Human-AI collaboration is being considered [7]. In particular, in the case of medicine, there are cases in which AI and humans collaborated to increase the cancer detection rate compared to when human doctors were alone [8]. As for drawing, the possibility of Human-AI co-creation is investigated while the user and AI take turns sketching to complete pictures [9]. However, though there were studies that created visual stories with deep learning algorithms [10, 11], most of the studies paid little attention to the effects of human-in-the-loop application in deep learning.

## 1.2 Interactive storyboarding

The storyboard is a sequence of drawings that represent the shots planned for story products, typically with some directions and dialogue. The storyboard creation step is crucial in that storyboards serve as a visual road map during the story product development period. However, storyboard creation is a difficult task even for professional artists, let alone novices, to simultaneously consider vital components of storyboards such as subject, background, and point of view.

Inspired by the fact that movies provide abundant sources of scene knowledge, which can be utilized to compose storyboards, 300,000 captured images from the 7083 movie trailers are collected and processed with deep learning models to extract scene knowledge. Object detection and semantic segmentation modules based on Convolutional Neural Networks (CNNs) are applied to spot objects which could play essential roles in the plot. The image-to-sketch style transfer model composed of Generative Adversarial Networks (GANs) is adopted to generate sketches from images of segmented objects.

A large-scale pre-trained language model is employed to generate relevant sentences to users' inputs, while another large-scale pre-trained text encoder is utilized to match the sentences to scenes using the similarity scores calculated in the text-image co-embedding vector space. The advantages of large-scale pre-trained models are leveraged in building a knowledge base for downstream tasks. Given the text descriptions specified by the users, Gennie represents multiple draft sketches that match the story, and the users can get some inspiration or utilize the sketches for their story.

This paper represents user studies and visualizes the co-creation process of a storyboard. The participants' strategies collaborating with Gennie are classified into two categories, while the required function for Gennie is matched to each strategy. Since draft sketches offered by Gennie have flexibility and lack detail, users can utilize the sketches as blueprints for their sake as sources for the composition of objects or the final look of visual scenes.

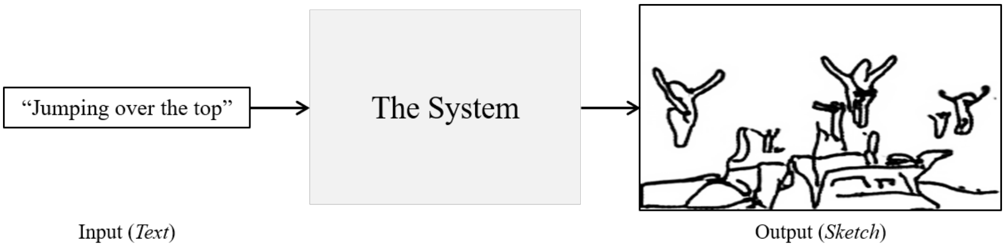


Figure 1.1: The proposed system's inputs and outputs.

In sum, the contributions of this work are as follows:

- This paper suggests an AI system for storyboard co-creation with users. Several deep learning models are effectively incorporated to implement a user-friendly and practical system.

- A user study is conducted to visualize participants' interaction with the storyboarding system, and the users' collaboration strategies to generate visual stories are examined.

- This paper discussed the implications of storyboard co-creation. This work is the first research track focused on the human-in-the-loop application in storyboard generation to the author's best knowledge.

# Chapter 2

# Related work

## 2.1 Human-AI co-creation

Advance in AI technologies has opened up the possibilities of human-AI co-creation for drawing [12, 13], creative writing [14, 15], music composition [16], and video games [17]. For example, AI can create a half-sketched picture [18], write the next paragraph of the story [19], or add images to the design mood board. The key challenge in this range of previous tasks was to develop collaborative AI agents that could coordinate tasks based on users' goals and behaviors. To this end, some systems were designed to generate outputs according to the surrounding context of human-generated content, and some systems utilized user feedback to better match AI behavior to user intentions [20, 21].

These new interfaces and algorithms were still in the experimental stage, but they have opened up the possibility that humans and AI can work together to produce creative results. As deep learning models advance, novel frameworks and adequate design guidelines are needed to understand users' perceptions of these new technologies and improve UX. In this regard, a prototype system is designed to leverage several pre-trained deep learning models to investigate the realm of collaboration between humans and intelligent machines.

In consideration of the inherent multimodal characteristics of the storyboard, a deep learning-based model dealing with both images and text descriptions [22] is utilized. In addition, the output of the AI model is designed considering the actual collaboration process with the user. Since storyboards are created in the early phase of content production, the exploration of motifs is helpful for creators in generating stories and composing scenes. Based on the assumption, the system is built and shows how the AI agent can be used in the actual creative process and the user's cognitive experience.

AI-based systems offer potential benefits in making artworks or content [23, 24]; however, few of the promising applications of AI were produced without the proper engagement of humans. Instead, humans can collaborate with AI agents to achieve users' creative goals by getting some inspiration [25], gaining practical support in the progress, or enjoying the co-creation process itself. To integrate AI into the already-complicated human workflow, bringing the human-centered design philosophy into computational interaction research is crucial. To deeply understand the user experience of human-AI co-creation in generating a storyboard, Gennie is implemented.

## 2.2 AI for visual story generation

The literature related to automatic storyboard creation tasks has mainly two directions: generation-based [26, 27, 11] and retrieval-based methods [14, 28, 29]. The generation-based method creates images directly from a text through Generative Adversarial Networks (GANs) that can create new images. As a result, the degree of freedom of the generated output is high and produces novel output. However, training difficulties make creating high-quality, diverse, and relevant images challenging. In addition, the AI is often overloaded with information to interpret, such as frequent interactions, which add complexities to building up an intelligent system in a co-creation setting.

Retrieval-based methods detour the difficulty of creating images by searching for existing high-quality images with text. Retrieval-based methods can ensure high-quality output, resulting in better user satisfaction. However, the retrieved output is limited to the system designer's knowledge which may deter free and creative thinking. Most text-image retrieval tasks focus on matching a single sentence with a single image [30, 22], where global and dense visual semantic matching models are frequently used. Most text-image retrieval systems focus on matching quality rather than considering the user experience of how to utilize the image's output.

This paper proposes a new framework for human-in-the-loop storyboard creation while overcoming the limitations of generation and retrieval-based methods to implement the system. Both generative and retrieval methods are employed to prepare a knowledge base and enable cross-modal matching combining the advantages of each method. A system designer's laborious engineering effort can be relieved to compose output candidates while guaranteeing high-quality outputs. It is also easy to expand the knowledge set for novel outputs. Expert knowledge can be easily fed into the database.

## 2.3 Evaluation metric for sketch research

In recent years, the sketch research community has developed as summarized by Figure 2.1. For example, in 2017, Google released a million-scale sketch dataset [31]. QuickDraw dataset contains over 50 million sketches collected from the online game "QuickDraw". This work motivated the community to go beyond considering sketches as static pictures by utilizing stroke sequences as input, which is a study of the temporal processing of sketches [5].

From 2018 to date, various novel tasks have been proposed while introducing bespoke datasets and evaluation metrics enabling deep learning techniques. The Sketchforme [32] solved the task of sketching for instructive sentences. The user survey was conducted to compare the system's outputs to the human-drawn sketches.

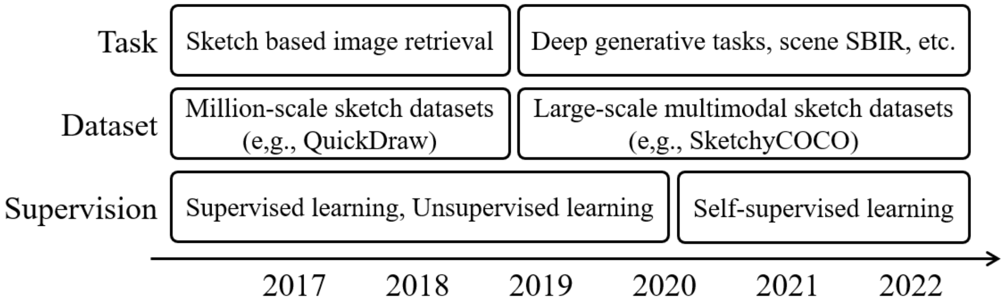

Task
Sketch based image retrieval
Deep generative tasks, scene SBIR, etc.
Dataset
Million-scale sketch datasets (e,g., QuickDraw)
Large-scale multimodal sketch datasets (e,g., SketchyCOCO)
Supervision
Supervised learning, Unsupervised learning
Self-supervised learning
2017
2018
2019
2020
2021
2022

Figure 2.1: Milestones of sketch research in terms of tasks, datasets, and supervision. Note that self-supervised learning is also unsupervised learning.

In the Scones study [33], a model was proposed to receive text instruction as input and continuously generate sketches. This work was also evaluated based on user surveys on satisfaction and enjoyment. Storyboarding, the task covered in this paper, is also human-involved, requiring humans to measure the performance of the system. Both qualitative and quantitative methods are adopted to evaluate the proposed systems' outputs along with novel tasks.

# Chapter 3

# Proposed method

## 3.1 Interactive storyboarding system

The storyboard is a sequence of drawings, typically with some directions and dialogue, that represents the shots planned for story products. In the early stages of production, the artists outline a narrative structure using storyboards, organizing the main plot and order of events. The storyboard creation step is crucial in that storyboards serve as a visual roadmap during the story product development period. Storyboard creation is a difficult task even for professional artists to simultaneously consider vital components of the storyboard such as subject, background, and point of view. Drawings in the storyboard contain plentiful abstract information about stories.

From the perspective of data modality, sketches in storyboards have characteristics providing domain-unique challenges. Sketches are highly abstract and diverse. People can depict a house as a square with a triangle on top in sketch form. Different people usually draw distinctive sketches when given identical instructions. And sketch images can also be represented in diverse forms. For example, a sketch can be expressed in static pixel space (when rendered as an image), dynamic stroke coordinate space (when considered as a times series), and geometric graph space (when considered topologically) [5].

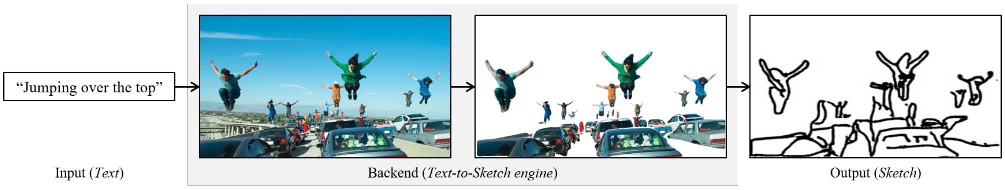


Figure 3.1: Overview of the Gennie, the interactive storyboarding system.

Sketch data also has advantages in certain tasks. Sketches can serve as one of the easiest computer-interaction modalities in a way that photos cannot due to the intuitive way humans can create sketches without training. Considering these unique challenges and opportunities regarding sketch data, it is often beneficial to design sketch-specific models to obtain the best performance in various sketch-related tasks. Thus, from a pattern recognition or machine intelligence perspective, unique characteristics of sketch modality often lead to task-specific system designs to exploit task-specific data properties and achieve task-specific performance indicators requiring significant engineering labor for each task which is not scalable.

This paper proposes an AI collaborator Gennie and a co-creation framework for storyboard generation in a scalable way. Instead of focusing on leveraging the characteristics of sketch data, the proposed system utilizes the general text-image co-embedding model to exploit the advantages of the large-scale pre-trained model. The overall system and framework are shown in Figure 3.2. Figure 3.2 represents knowledge preparation, story-to-sketches retrieval, and user interface. This section first introduces how to collect and process the data for preparing a scene knowledge database and describes the story-to-sketches model applying both Natural Language Generation (NLG) model and the text-image co-embedding deep learning model.

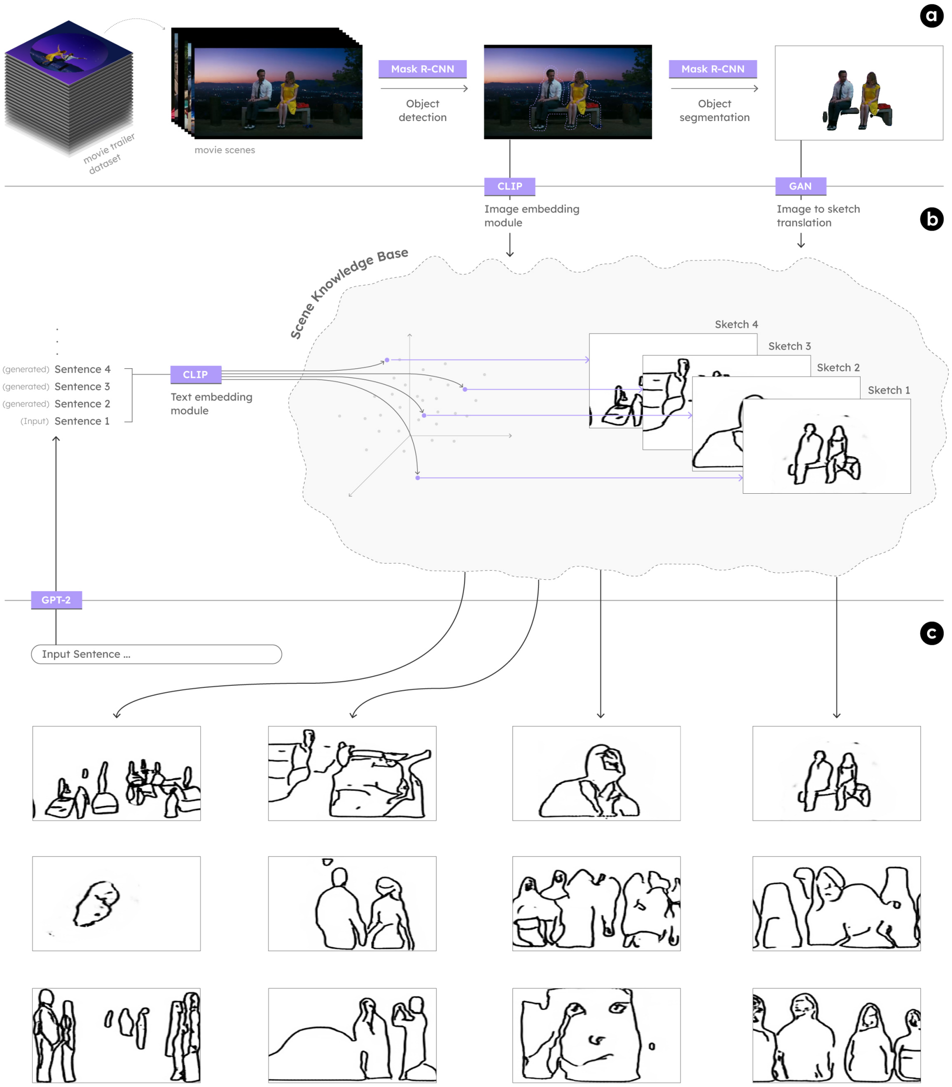


Figure 3.2: Structure of the interactive storyboarding system.

## 3.2 Knowledge base preparation

A knowledge base is prepared to provide sketches suitable for the user's creative goals. Essential elements of the storyboard, such as the main character, background, and perspective information, are derived from the scenes in movies. First, 7,083 movie trailers and movie descriptions at HD-trailers.net are crawled and captured with 40 pictures in each trailer at regular intervals. Then, around 300,000 sketch images are sourced from the crawled movie images. Each sketch image should contain semantically evident instances of the original crawled image to make users interpret the image instantly. Therefore, the knowledge base preparation process is divided into instance recognition and sketch translation.

For the instance recognition step, Mask R-CNN [34] model is utilized as the main module that efficiently detects objects in an image while parallelly generating fine-grained segmentation masks for the objects. Both bounding box regions and segmentation masks can be obtained by passing original images to the module. Before translating object images to the sketch style, the images are post-processed with the segmentation masks as shown in Figure 3.3. Key object-grounded regions can be obtained without uninformative objects or backgrounds by extracting overlapping regions between segmentation masks and the original images. In addition, noisy images are filtered out with this process while enhancing the quality of sketch translation output. For example, trailer scenes containing only text descriptions or blank images can be captured at scene-change moments. After all, 25% of total images were considered as noises and removed.

Then the contour drawings are generated from refined images. pix2pix model [35] is exploited as an image-to-sketch translation model. Unlike conventional edge or boundary detection algorithms, pix2pix predicts the salient contours in images, and the outputs are in a familiar style resembling human drawing sketches. Methods suggested in Photo-Sketching model [18] are also used in the sketches translation step. Several image-to-sketch translation models were considered as design alternatives.

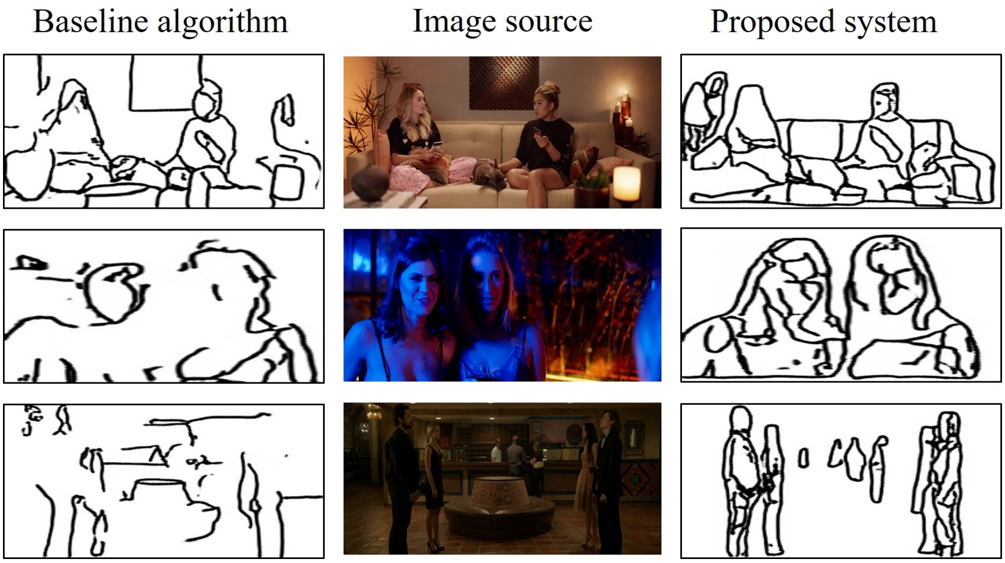


Figure 3.3: Effects of segmentation masks in producing sketch images.

Sketch datasets can be grouped in several ways as shown in Table 3.1. In terms of modality, single-modal sketch datasets contain only sketches while multimodal sketch datasets include various modalities such as photos, text, 3D rendering, or video. Paired multimodal dataesets support cross-modal applications (cross-modal retrieval, cross-modal generation, etc.) and the knowledge base in this paper is also a paired multimodal dataset to support cross-modal retrieval of text, photos, and sketches.

Collection strategies of sketch datasets are also a criterion to classify the datasets. For example, some datasets are created by researchers [36], [37], [38]. Other datasets are collected via crowd-sourcing with platforms or online drawing games [31], [39], [40], [41]. Web crawling is another option to build a massive sketch dataset with limitations to obtain rich annotation compared with crowd-sourcing [42], [43]. Sketch datasets' potential usages are determined by both their collection and annotation protocol. The knowledge base in this paper (GennieMovie) leverages web crawling and sketch generation from movie scenes, which is an efficient and scalable way to gather paired multimodal sketch datasets adequate to storyboarding.

Table 3.1: Exemplary sketch datasets. “**s**” and “**p**” mean “sketches” and “photos”.

| Datasets | Modalities | Size | Annotations |
|---|---|---|---|
| SketchSeg-10K [36] | single | 10K **s** | class, segmentation |
| SPG [37] | single | 20K **s** | class, grouping |
| SketchSeg-150K [38] | single | 150K **s** | class, segmentation |
| QuickDraw [31] | single | 50M+ **s** | class |
| QMUL Chair [39] | multimodal | 297 **s**, 297 **p** | pairing, triplet, attribute |
| QMUL Shoe [39] | multimodal | 419 **s**, 419 **p** | pairing, triplet, attribute |
| PACS DG [42] | multimodal | 9991 (**s**, **p**, etc.) | class |
| Sketchy [40] | multimodal | 75K **s**, 12K **p** | class |
| GennieMovie | multimodal | 225K **s**, 225K **p** | pairing |
| QuickDrawExtended [41] | multimodal | 330K **s**, 204K **p** | class |
| DomainNet [43] | multimodal | 600K (**s**, **p**, etc.) | class |

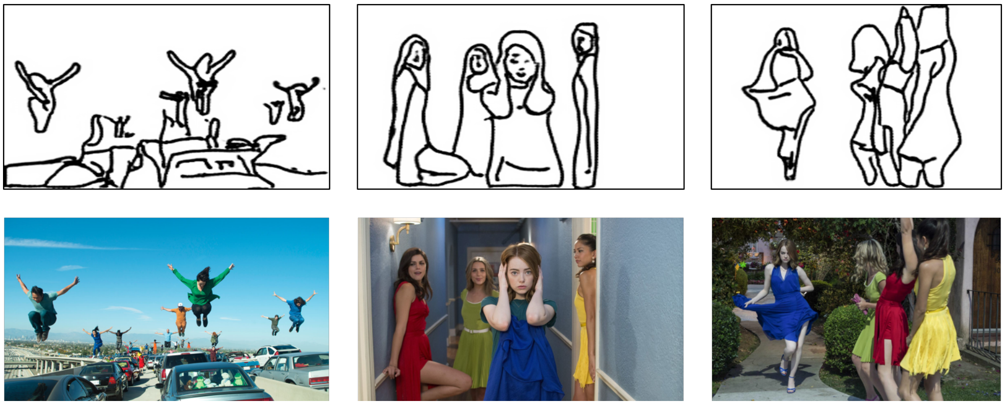


Figure 3.4: Prepared sketches and their sources (©Summit Entertainment).

## 3.3 Sentence-to-sketches retriever

A large-scale pre-trained language model [44] is employed to generate relevant sentences to users' inputs, while another large-scale pre-trained text-image co-embedding model [22] is utilized to match the sentences to scenes using the similarity scores calculated in the text-image joint embedding vector space. When a user inputs a sentence, GPT-2 generates related sentences. This process is devised for suggesting various ideas to a single input and adopted as the system's module after conducting a test with an explainable method [45] to examine the matching quality between the generated sentences and the sketches. GPT-3 [46] could be a design alternative for this module.

After generating several related sentences from the input sentence, the CLIP model finds the closest images in the co-embedding space utilizing pre-trained knowledge of the relation between text and image. And then, matched sketches are picked with the closest images and presented to the user. Sketches from the knowledge base depict salient instances in images and have the flexibility and lack of scenery details to leave room for users' creative roles. Gennie represents multiple draft sketches, and the users can get inspiration or utilize the sketches for their own stories. The CLIP's matching quality is expected to be enhanced via fine-tuning with the movie scene-subtitle pair.

Table 3.2: Specification of the large-scale pre-trained model (GPT-2)

| Specification | GPT-2 (S) | GPT-2 (M) | GPT-2 (L) | GPT-2 (XL) |
|---|---|---|---|---|
| number of parameters | 117M | 345M | 762M | 1,542M |
| number of layers | 12 | 24 | 36 | 48 |
| model dimensionality | 768 | 1024 | 1280 | 1600 |

Table 3.3: Specification of the large-scale pre-trained model (GPT-3)

| Specification | GPT-3 (S) | GPT-3 (M) | GPT-3 (L) | GPT-3 175B |
|---|---|---|---|---|
| number of parameters | 125M | 350M | 760M | 175B |
| number of layers | 12 | 24 | 24 | 96 |
| model dimensionality | 768 | 1024 | 1536 | 12288 |

The generative algorithm has a vast search space from which the output can be derived and produce novel results. However, the generation-based approach has difficulties in its outputs being accepted as appropriate or satisfactory for human understanding. On the other hand, the retrieval-based approach produces valuable results because experts can directly construct a search space. However, there is a disadvantage in that the results can be limited to existing knowledge.

Thus, generation-based and retrieval-based methods are combined to develop Gennie in this study. The knowledge base is constructed using a generation-based method to explore novel sketches while keeping search spaces not far from human understanding with movie scene references. Indexing is achieved with a multimodal embedding model so that the scalability of data is preserved in a retrieval setting. Randomness or ambiguity that occurs in the retrieval process is properly understood in the context of applying to the user's creative goal. The actual User-AI collaboration process analysis is described in Chapter 4.

# Chapter 4

# Experimental results

## 4.1 Experimental setup

Experiments were conducted several times. In the user study, participants (87 in total) were recruited to observe user experiences interacting with the designed system. The participants consisted of 5 amateur artists, 79 high school students, and 3 professional cartoonists. Most of the participants were accustomed to dealing with touchscreen devices and digital pens. The experiment took less than one hour and was screen-recorded. Participants were asked whether they consented to the recording and publication of results under anonymity before starting the experiment and experimental data was collected for only those who agreed.

Participants were asked to create a storyboard that consisted of four cuts with images and text, collaborating with Gennie. Examples of expected output storyboards were shown in advance. The theme of the story and collaboration methods were not restricted officially, while half of the participants followed the example story instructions. Participants were informed that they could use and modify the sketches suggested by Gennie as they wanted to. Tablet PCs, digital pens, and sketch apps are prepared for drawing tools. The storyboard template is also provided with a canvas layer with four blank boxes for visual and horizontal lines for text description.//

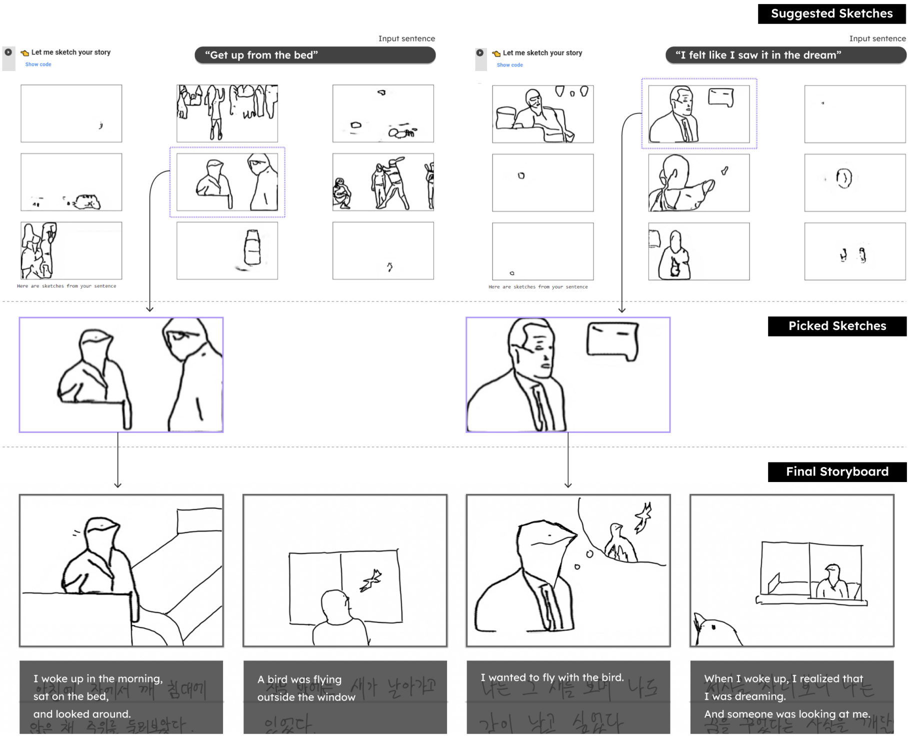


Figure 4.1: A series of scenes generated with the proposed system Gennie.

## 4.2 Procedure

Firstly, the participants were given instructions about interacting with Gennie and using the sketch app. Then, the participants experienced the whole collaboration process; they asked Gennie to generate sketches, received the sketches, brought one to the canvas, and modified the sketch. Once the participants became familiar with this process, they were asked to start to create a storyboard in their way.

After the experiment, some participants were asked to join a semi-structured interview based on their overall experience and behaviors. Questions were prepared to ask regarding general satisfaction and enjoyment with the produced outcome, collaborative experiences with AI, and the perceived role of AI in co-creation. The screen-recorded videos were used during the interview when needed.

## 4.3 User research

The interactive storyboarding process is visualized in Figure 4.8. The screen-recorded videos, which include the process of generating storyboards by the participants with Gennie, were behavior-coded. The vertical axis represents the participant number, and the horizontal axis represents the time it took to complete the storyboard. The time limit was set, but the experimental environment was free in time, and the participants spent enough time creating a storyboard with Gennie. The specific meaning of the color of the bars is as follows.

The part where the user interacts with the system is indicated in purple, and the part where the user works on the canvas page is indicated in light green. In the purple part, the user entered text into the system and selected a draft to be used for storyboard production from the recommended sketches. In the light green part, users cut and paste the recommended sketch into an empty storyboard template, draw sketches from the beginning, or write text. In particular, the part where the user modifies and utilizes the recommended sketches is marked in dark green.

As can be seen from Figure 4.8, the distribution of time spent on modifying the recommended sketches has a large variance for each user in the free-drawing experiment settings. If nothing happens on the screen or the cursor is moving meaningless, that process is indicated by a solid black line. During the interview, no activity on the screen mainly indicates that the user focused on brainstorming or thinking about the storyline. The idle time was different for each participant but was relatively short as interacting with Gennie.

Participants went through exploring how Gennie works in the beginning. Though users were guided on using the Gennie before the experiment, participants wanted to interact with Gennie to see how to respond to various inputs. Participants tried to compare Gennie's sketches by entering the same stories but different lengths of sentences or different abstraction levels. The insights from the interview were interpreted based on behavioral data.

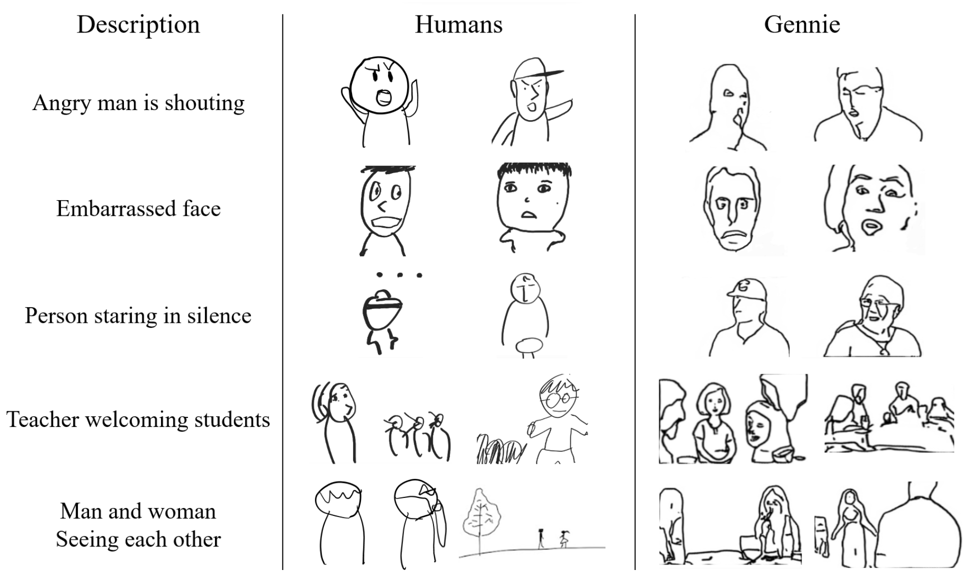


Figure 4.2: Samples of sketches produced by humans and Gennie in the user study.

For example, P8 stated that he tried to get a clear picture of how Gennie is working, saying, “To understand how it works, I tried writing a short text, long text, and a sarcastic expression.” P11 also commented, “I don’t know Gennie well, but in the case of AI speech recognition, the AI system recognizes better for specific sentences. So I thought there are certain sentences that Gennie recognizes well.”

In the free-drawing experimental setting, exploring the working principle of Gennie was carried out in the early phase of storyboard creation. As shown in Figure 4.8, there was no action on the screen recording in the beginning, and it alternated with the action of simple hesitation or moving the cursor around quickly, which indicates the users are brainstorming the scenario.

After finishing the investigation of Gennie, the participants gradually utilized sketches presented by Gennie in their stories and proceeded to create a storyboard. The majority of reactions were in the cases where Gennie’s abstract sketch was interpreted and transformed into a story he/she had envisioned. There were cases of using Gennie’s draft as initially intended or adding a few strokes to the output.

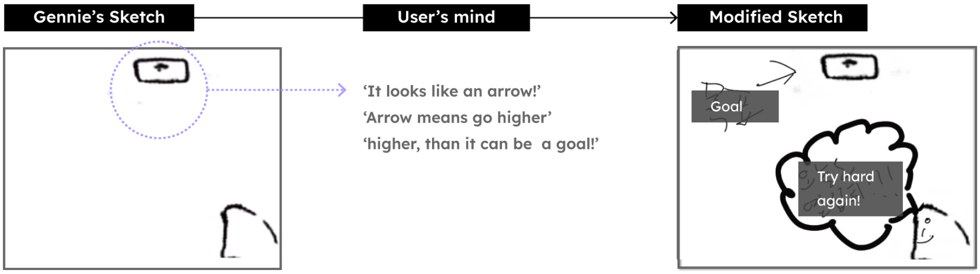


Figure 4.3: Example of user's perception of the suggested sketch.

P7 interpreted Gennie's sketch as an image of a target, saying, "I saw that square thing as an arrow. That felt like a goal." P9 saw the composition that expressed the specific scene that he had imagined. "This is how I look when I leave school. It feels like all the students are leaving school." P14 reported the experience of accurately matching the desired scene, "I think this was the scene I thought of. A woman was standing in front of me who looked like a TV (the object - pointing at a specific sketch), so I thought that I could complete the scene with this draft." Participants felt that Gennie's drawings were carefully recommended for their scene ideas in this case.

There were cases where questions were raised about Gennie's outputs as the suggested sketches did not fit the participants' intentions. In addition, there were cases where Gennie felt that the desired picture was not presented, thus Gennie's recommendation was excluded from the creative process. There were also cases where the irrelevant results lead to entirely different new stories.

Figure 4.4 shows that the participants evaluate the storyboard creation process with Gennie as a satisfying experience. The proposed system has the potential to better engage the artists with the results for the questions in the first column while the graphs in the second column indicate that the co-creation process requires some time to adapt. Figure 4.5 and Table 4.1 indicate that the users adapt to leverage Gennie after the first storyboarding (instruction-given and sketch setting).

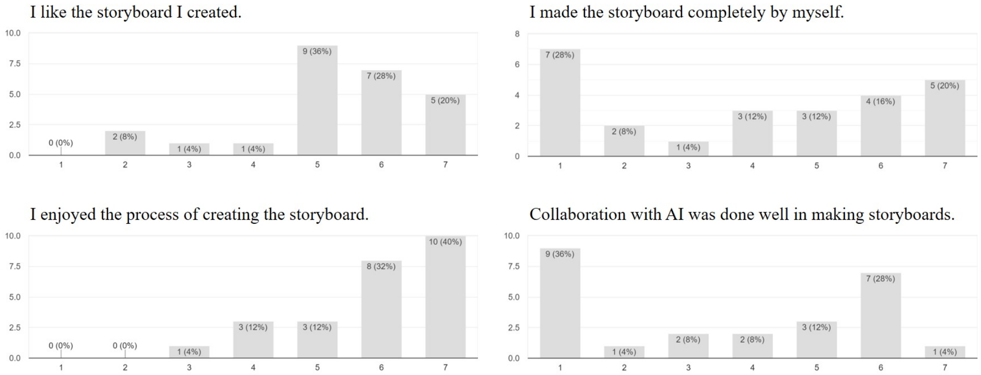


Figure 4.4: Participants' responses to each question.

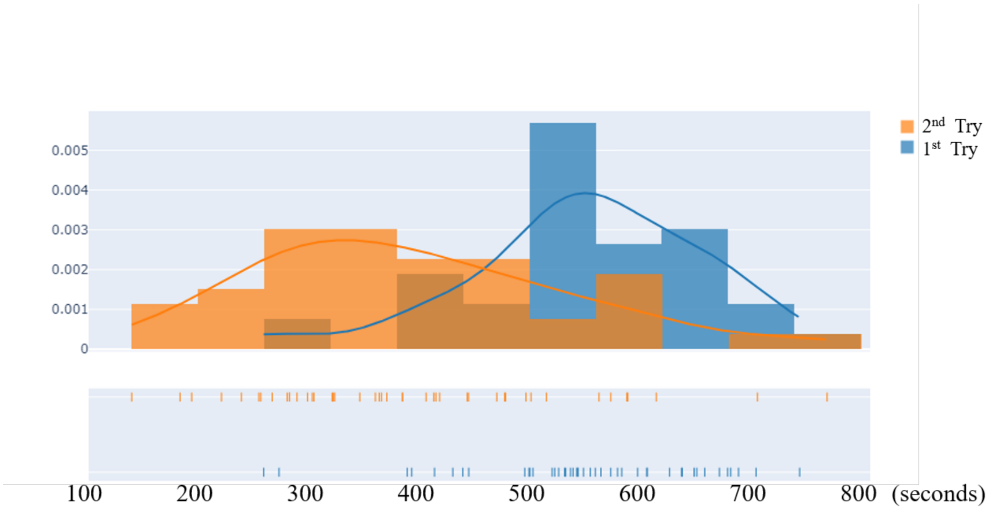


Figure 4.5: The distribution of time to complete the storyboarding task with Gennie.

Table 4.1: Statistics of the user study. The participants take significantly less time to finish the task with the proposed system on the second try.

| Statistics | First Try (w/ Gennie) | Second Try (w/ Gennie) | w/o Gennie |
|---|---|---|---|
| Mean (seconds) | 558.34 | 400.36 | 564.73 |
| SD (seconds) | 105.12 | 139.43 | 158.73 |
| N | 44 | 44 | 41 |

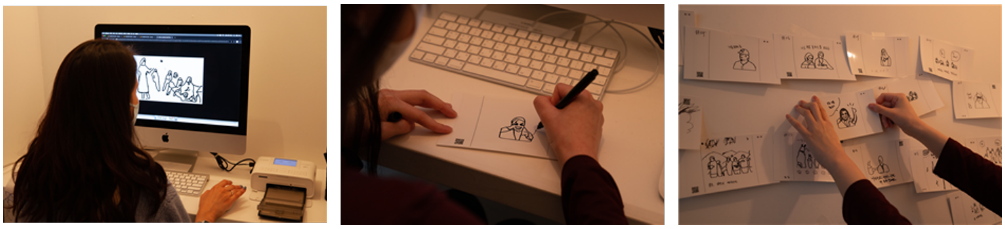


Figure 4.6: Use case of the interactive storyboarding system.

## 4.4 Participants' strategies in interactive storyboarding

Users had mainly two goals leveraging the interactive storyboarding system: generating a new topic and drawing a specific scene. The users employed different strategies to achieve each goal. In order to achieve their task goal of creating a free-topic storyboard, participants used Gennie in different manners.

First, users utilize Gennie as a trigger to compose a storyline. Inspired by the proposed sketches, participants came up with interesting subjects and stories. Results from Gennie were used as various triggers for participants' ideations in this case. Once participants succeeded in getting some clues to start a story from Gennie, they went on their way to creating storyboards without Gennie.

Second, participants used Gennie as a tool to sketch out scenes to represent their stories. When participants could not specify how they should sketch the scenes from their topics, they interacted with Gennie to get insights into how to fill their stories. Even when Gennie created unexpected results, participants did not ignore the results; they reflected the sketches on their stories.

The two behaviors above were not clearly divided, as participants used Gennie for diverse needs as shown in Figure 4.9. When the users anticipated Gennie's outputs depicting high-level concepts, participants used the engine as a casual partner in the ideation. When participants expected more specific results from Gennie, input become detailed. Figure 4.10 shows participants' search history implying the user's approaches using the system. The asterisk means the sketch was picked in that sentence.

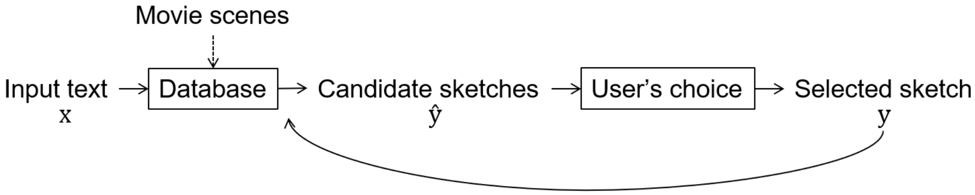


Figure 4.7: A feedback loop to improve the proposed system. The candidate sketches-selected sketch pair can be utilized to customize the recommendation.

Drawing a storyboard is crucial to planning any form of visual narrative. The composition of the objects in the frame and the point of view created by the angle can significantly enhance or alter how the viewer entertains the visual story. The AI agent Gennie is developed in this regard to help extend the limits of an individual's imagination and creativity. Gennie is not only a novel invention but also a growing partner in collaboration with the user with feedback loop works as Figure 4.7.

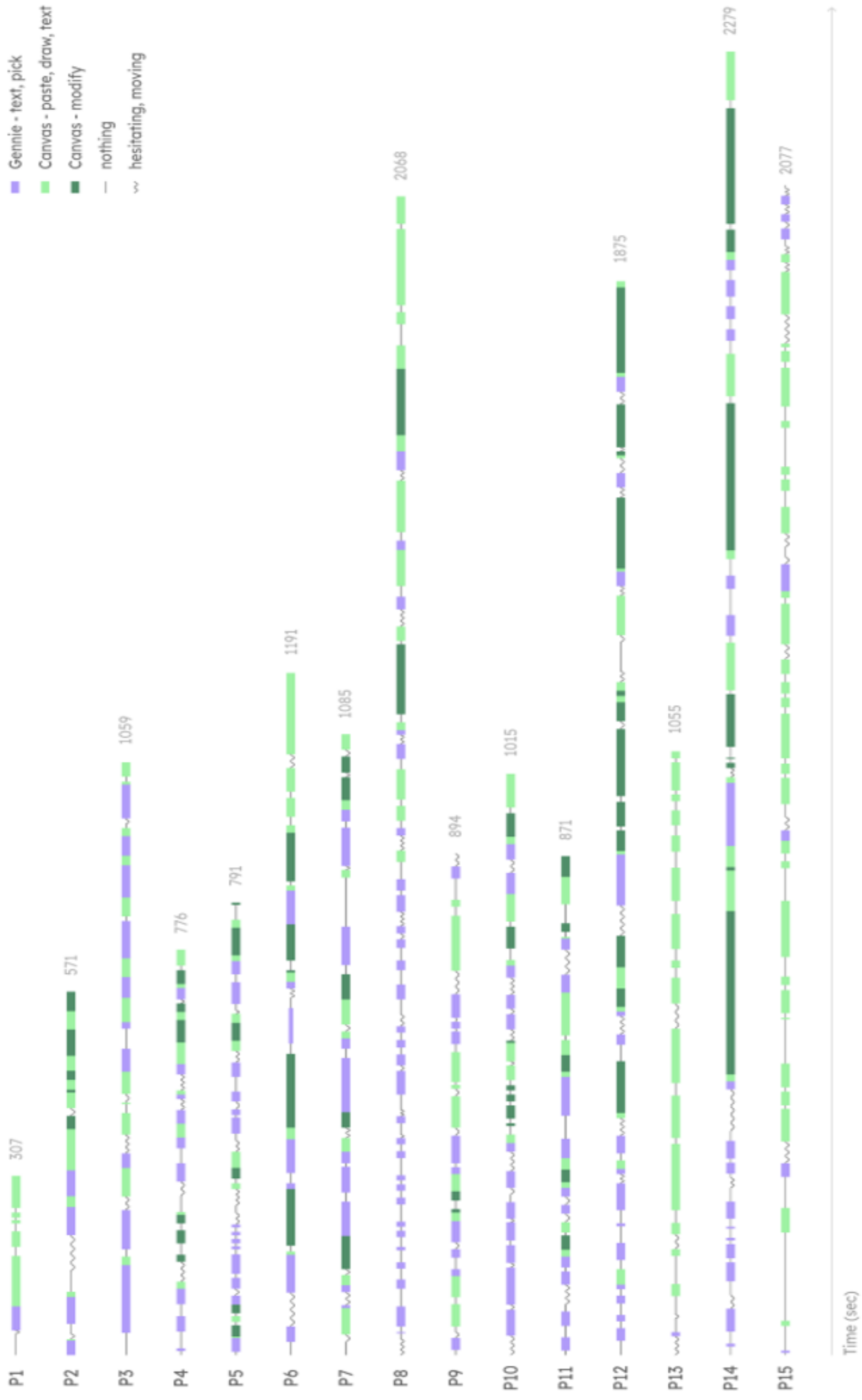


Figure 4.8: Process of creating storyboards by participants.

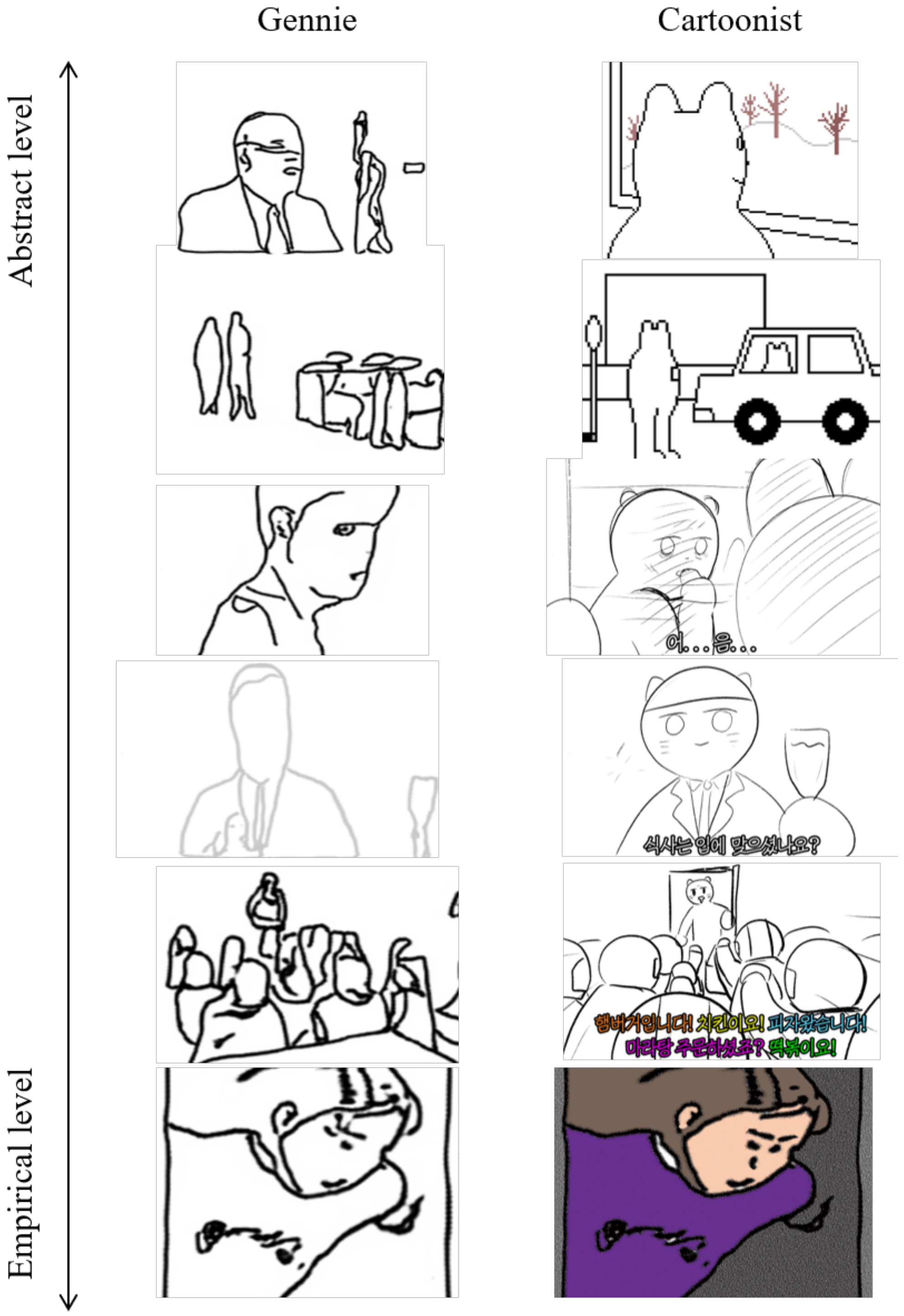
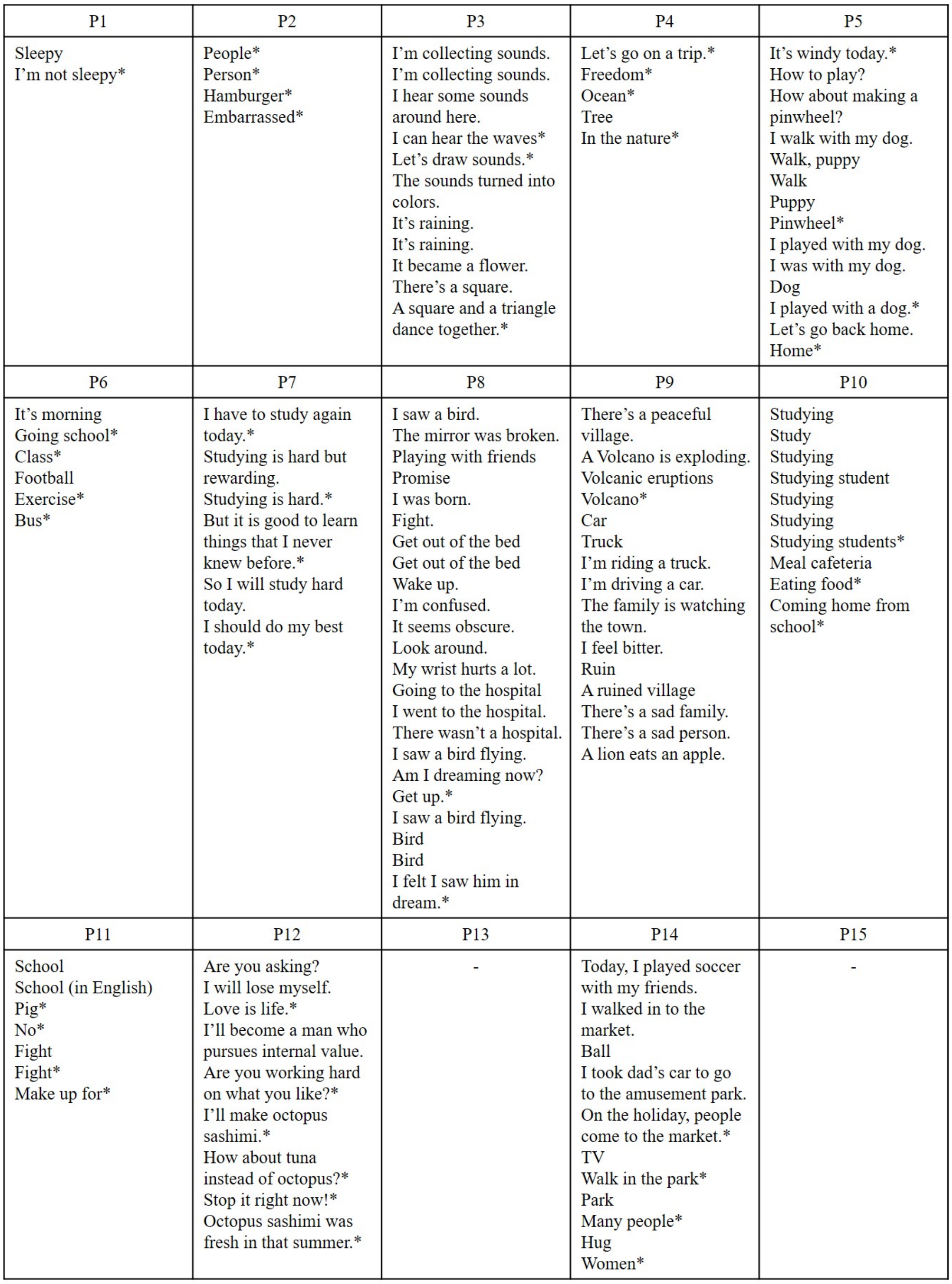


Figure 4.9: The professional cartoonists developed ideas from Gennie's sketches.



| P1 | P2 | P3 | P4 | P5 |
|---|---|---|---|---|
| Sleepy<br>I’m not sleepy* | People*<br>Person*<br>Hamburger*<br>Embarrassed* | I’m collecting sounds.<br>I’m collecting sounds.<br>I hear some sounds around here.<br>I can hear the waves*<br>Let’s draw sounds.*<br>The sounds turned into colors.<br>It’s raining.<br>It’s raining.<br>It became a flower.<br>There’s a square.<br>A square and a triangle dance together.* | Let’s go on a trip.*<br>Freedom*<br>Ocean*<br>Tree<br>In the nature* | It’s windy today.*<br>How to play?<br>How about making a pinwheel?<br>I walk with my dog.<br>Walk, puppy<br>Walk<br>Puppy<br>Pinwheel*<br>I played with my dog.<br>I was with my dog.<br>Dog<br>I played with a dog.*<br>Let’s go back home.<br>Home* |
| P6 | P7 | P8 | P9 | P10 |
| It’s morning<br>Going school*<br>Class*<br>Football<br>Exercise*<br>Bus* | I have to study again today.*<br>Studying is hard but rewarding.<br>Studying is hard.*<br>But it is good to learn things that I never knew before.*<br>So I will study hard today.<br>I should do my best today.* | I saw a bird.<br>The mirror was broken.<br>Playing with friends<br>Promise<br>I was born.<br>Fight.<br>Get out of the bed<br>Get out of the bed<br>Wake up.<br>I’m confused.<br>It seems obscure.<br>Look around.<br>My wrist hurts a lot.<br>Going to the hospital<br>I went to the hospital.<br>There wasn’t a hospital.<br>I saw a bird flying.<br>Am I dreaming now?<br>Get up.*<br>I saw a bird flying.<br>Bird<br>Bird<br>I felt I saw him in dream.* | There’s a peaceful village.<br>A Volcano is exploding.<br>Volcanic eruptions<br>Volcano*<br>Car<br>Truck<br>I’m riding a truck.<br>I’m driving a car.<br>The family is watching the town.<br>I feel bitter.<br>Ruin<br>A ruined village<br>There’s a sad family.<br>There’s a sad person.<br>A lion eats an apple. | Studying<br>Study<br>Studying<br>Studying student<br>Studying<br>Studying<br>Studying students*<br>Meal cafeteria<br>Eating food*<br>Coming home from school* |
| P11 | P12 | P13 | P14 | P15 |
| School<br>School (in English)<br>Pig*<br>No*<br>Fight<br>Fight*<br>Make up for* | Are you asking?<br>I will lose myself.<br>Love is life.*<br>I’ll become a man who pursues internal value.<br>Are you working hard on what you like?*<br>I’ll make octopus sashimi.*<br>How about tuna instead of octopus?*<br>Stop it right now!*<br>Octopus sashimi was fresh in that summer.* | - | Today, I played soccer with my friends.<br>I walked in to the market.<br>Ball<br>I took dad’s car to go to the amusement park.<br>On the holiday, people come to the market.*<br>TV<br>Walk in the park*<br>Park<br>Many people*<br>Hug<br>Women* | - |

Figure 4.10: Search-history of the participants.

# Chapter 5

# Conclusion

This paper focused on storyboard generation, which has multimodal characteristics for visual storytelling. The AI system was developed for storyboard co-creation with users. Several deep learning models, including large-scale pre-trained models, were effectively incorporated to implement a user-friendly and practical system. Gennie is an early step towards intelligent systems that support human-in-the-loop applications for communicating and developing ideas in storyboard generation.

# Bibliography

[1] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[2] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[3] P. Jain, P. Agrawal, A. Mishra, M. Sukhwani, A. Laha, and K. Sankaranarayanan, “Story generation from sequence of independent short descriptions,” *arXiv preprint arXiv:1707.05501*, 2017.

[4] N. Jaques, S. Gu, R. E. Turner, and D. Eck, “Generating music by fine-tuning recurrent neural networks with reinforcement learning,” 2016.

[5] P. Xu, “Deep learning for free-hand sketch: A survey,” 2020.

[6] D. Adiwardana, M.-T. Luong, D. R. So, J. Hall, N. Fiedel, R. Thoppilan, Z. Yang, A. Kulshreshtha, G. Nemade, Y. Lu *et al.*, “Towards a human-like open-domain chatbot,” *arXiv preprint arXiv:2001.09977*, 2020.

[7] D. Wang, E. Churchill, P. Maes, X. Fan, B. Shneiderman, Y. Shi, and Q. Wang, “From human-human collaboration to human-ai collaboration: Designing ai systems that can work together with people,” in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–6.

[8] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck, “Deep learning for identifying metastatic breast cancer,” *arXiv preprint arXiv:1606.05718*, 2016.

[9] C. Oh, J. Song, J. Choi, S. Kim, S. Lee, and B. Suh, “I lead, you help but only with enough details: Understanding user experience of co-creation with artificial intelligence,” in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–13.

[10] K. Frans, L. Soros, and O. Witkowski, “Clipdraw: Exploring text-to-drawing synthesis through language-image encoders,” *arXiv preprint arXiv:2106.14843*, 2021.

[11] A. El-Nouby, S. Sharma, H. Schulz, D. Hjelm, L. E. Asri, S. E. Kahou, Y. Bengio, and G. W. Taylor, “Tell, draw, and repeat: Generating and modifying images based on continual linguistic instruction,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10 304–10 312.

[12] L. Sun, P. Chen, W. Xiang, P. Chen, W.-y. Gao, and K.-j. Zhang, “Smartpaint: a co-creative drawing system based on generative adversarial networks,” *Frontiers of Information Technology & Electronic Engineering*, vol. 20, no. 12, pp. 1644–1656, 2019.

[13] N. M. Davis, C.-P. Hsiao, K. Y. Singh, and B. Magerko, “Co-creative drawing agent with object recognition,” in *Twelfth artificial intelligence and interactive digital entertainment conference*, 2016.

[14] E. Bensaid, M. Martino, B. Hoover, J. Andreas, and H. Strobelt, “Fairytailor: A multimodal generative framework for storytelling,” *arXiv preprint arXiv:2108.04324*, 2021.

[15] Y. Ratawal, V. S. Makhloga, K. Raheja, P. Chadha, and N. Bhatt, “Poemai: Text generator assistant for writers,” in *Intelligent Sustainable Systems*. Springer, 2022, pp. 575–584.

[16] R. Louie, A. Coenen, C. Z. Huang, M. Terry, and C. J. Cai, “Novice-ai music co-creation via ai-steering tools for deep generative models,” in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–13.

[17] S. W. Kim, Y. Zhou, J. Philion, A. Torralba, and S. Fidler, “Learning to simulate dynamic environments with gamegan,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1231–1240.

[18] M. Li, Z. Lin, R. Mech, E. Yumer, and D. Ramanan, “Photo-sketching: Inferring contour drawings from images,” in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 1403–1412.

[19] H. Rashkin, A. Celikyilmaz, Y. Choi, and J. Gao, “Plotmachines: Outline-conditioned generation with dynamic plot state tracking,” *arXiv preprint arXiv:2004.14967*, 2020.

[20] P. Karimi, N. Davis, M. L. Maher, K. Grace, and L. Lee, “Relating cognitive models of design creativity to the similarity of sketches generated by an ai partner,” in *Proceedings of the 2019 on Creativity and Cognition*, 2019, pp. 259–270.

[21] K. Kim, J. Heo, and S. Jeong, “Tool or partner: The designer’s perception of an ai-style generating service,” in *International Conference on Human-Computer Interaction*. Springer, 2021, pp. 241–259.

[22] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” *arXiv preprint arXiv:2103.00020*, 2021.

[23] L. A. Gatys, A. S. Ecker, and M. Bethge, “Image style transfer using convolutional neural networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2414–2423.

[24] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2337–2346.

[25] O. Sbai, M. Elhoseiny, A. Bordes, Y. LeCun, and C. Couprie, "Design: Design inspiration from generative networks," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0–0.

[26] Y. Li, Z. Gan, Y. Shen, J. Liu, Y. Cheng, Y. Wu, L. Carin, D. Carlson, and J. Gao, "Storygan: A sequential conditional gan for story visualization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6329–6338.

[27] J. Zakraoui, M. Saleh, S. Al-Maadeed, J. M. Alja'am, and M. S. Abou El-Seoud, "Visualizing children stories with generated image sequences," in *International Conference on Interactive Collaborative and Blended Learning*. Springer, 2020, pp. 512–519.

[28] S. Chen, B. Liu, J. Fu, R. Song, Q. Jin, P. Lin, X. Qi, C. Wang, and J. Zhou, "Neural storyboard artist: Visualizing stories with coherent image sequences," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 2236–2244.

[29] Y. Kawano, E. Takaya, K. Yamanobe, and S. Kurihara, "Automatic plot generation framework for scenario creation," in *International Conference on Interactive Digital Storytelling*. Springer, 2018, pp. 453–461.

[30] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "Vse++: Improving visual-semantic embeddings with hard negatives," *arXiv preprint arXiv:1707.05612*, 2017.

[31] D. Ha and D. Eck, "A neural representation of sketch drawings," *arXiv preprint arXiv:1704.03477*, 2017.

[32] F. Huang and J. F. Canny, “Sketchforme: Composing sketched scenes from text descriptions for interactive applications,” in *Proceedings of the 32nd annual ACM symposium on user interface software and technology*, 2019, pp. 209–220.

[33] F. Huang, E. Schoop, D. Ha, and J. Canny, “Scones: towards conversational authoring of sketches,” in *Proceedings of the 25th International Conference on Intelligent User Interfaces*, 2020, pp. 313–323.

[34] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask r-cnn,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[35] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” 2018.

[36] F. Wang, S. Lin, H. Wu, H. Li, R. Wang, X. Luo, and X. He, “Spfusionnet: Sketch segmentation using multi-modal data fusion,” in *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2019, pp. 1654–1659.

[37] K. Li, K. Pang, J. Song, Y.-Z. Song, T. Xiang, T. M. Hospedales, and H. Zhang, “Universal sketch perceptual grouping,” in *Proceedings of the european conference on computer vision (ECCV)*, 2018, pp. 582–597.

[38] Y. Qi and Z.-H. Tan, “Sketchsegnet+: An end-to-end learning of rnn for multi-class sketch semantic segmentation,” *Ieee Access*, vol. 7, pp. 102 717–102 726, 2019.

[39] Q. Yu, F. Liu, Y.-Z. Song, T. Xiang, T. M. Hospedales, and C.-C. Loy, “Sketch me that shoe,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 799–807.

[40] P. Sangkloy, N. Burnell, C. Ham, and J. Hays, “The sketchy database: learning to retrieve badly drawn bunnies,” *ACM Transactions on Graphics (TOG)*, vol. 35, no. 4, pp. 1–12, 2016.

[41] S. Dey, P. Riba, A. Dutta, J. Llados, and Y.-Z. Song, "Doodle to search: Practical zero-shot sketch-based image retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2179–2188.

[42] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Deeper, broader and artier domain generalization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5542–5550.

[43] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang, "Moment matching for multi-source domain adaptation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1406–1415.

[44] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[45] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[46] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

# 초 록

인공지능 기술이 가져오는 변화는 전 산업과 시스템을 망라한다. 엔터테인먼트, 문화 콘텐츠 산업 또한 인공지능 기술의 발전에 영향을 받고 있다. 인공지능 기반의 시스템이 콘텐츠 제작에 큰 활약을 할 것으로 기대되지만, 여전히 상용화할 만한 인공지능 애플리케이션을 구축하는 것은 도전적인 문제이다. 본 논문에서는 스토리보드를 제작할 때 사용되는 새로운 시스템 ‘젠이’를 제시한다. GPT-2, CLIP 등 최근 괄목할 만한 성과를 보여준 대규모 사전학습 모델을 이용해 젠이 시스템을 구축했다. 잘 설계된 목표함수와 막대한 파라미터를 통해 효과적으로 지식 기반을 구성하는 대규모 사전학습 모델의 이점을 시스템에 활용하고자 하였다. 그리하여 사용자는 젠이를 활용해 머릿속의 장면 구상을 빠르게 시각화할 수 있다. 본 논문은 사용자 스터디를 통해 인터랙티브 스토리보딩 시스템으로 스토리보드를 창작하는 과정을 시각화하고 시사점을 제시한다.

**주요어**: 스토리보딩, 멀티모달 임베딩, 대규모 사전학습 모델

**학번**: 2019-21939