

THIS IS A THESIS  
OF THE UNDERGRAD VARIETY

Ursula Undergrad

Stanford University  
April 2022

An honors thesis submitted to the department of  
Civil and Environmental Engineering  
in partial fulfillment of the requirements for the undergraduate  
honors program

Advisor: Emmy Noether

\_\_\_\_\_ Date: \_\_\_\_\_

Emmy Noether (Thesis Advisor)

Olga Ladyzhenskaya Professor of Engineering

School of Computer Science

\_\_\_\_\_ Date: \_\_\_\_\_

Ada Lovelace (Thesis Advisor)

Professor

Computer Science

# Abstract

A brief overview of mud

# Acknowledgements

Thank you to water and soil

# Contents

<b>Abstract</b>	<b>iii</b>
<b>Acknowledgements</b>	<b>iv</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Exploring the Problem Space, early October . . . . .	2
<b>2 Data Collection</b>	<b>4</b>
2.1 Overview . . . . .	4
2.2 Version 0 . . . . .	5
2.3 Version 1 . . . . .	7
2.3.1 Overview . . . . .	7
2.3.2 Interface Design . . . . .	8
2.3.3 Deployment Results . . . . .	12
2.4 Version 2 . . . . .	13
2.4.1 Overview . . . . .	13
2.4.2 Results . . . . .	17
2.5 Dataset Summary . . . . .	18
<b>3 Methods</b>	<b>22</b>
3.1 CLIP Finetune . . . . .	22
3.1.1 Image Pre-Processing . . . . .	22
3.1.2 Text Pre-Processing . . . . .	22
3.2 Loss Function . . . . .	23
<b>4 Results &amp; Analysis</b>	<b>24</b>
4.1 Classification Experiments . . . . .	24
<b>5 Related Work</b>	<b>25</b>

<b>6 Conclusion</b>	<b>28</b>
<b>A True Facts</b>	<b>29</b>

# Chapter 1

## Introduction

Robotics have advanced significantly these years, the trend seems to be that while at the beginning, robots are in the far factory working on assembly line tasks, and they are very far from our lives; from the little iRobot roomba to more sophisticated labrador, which that can navigate your home to retrieve objects for you, or []. As the robot moves closer to us and into our lives, we would want the robots to feel not just like a machine that can help us, a device that we command around, but we would like the communication to go both ways so that robots can talk to us and complete tasks with us. In the many tasks that can be done in collaboration with robots, we investigate the task of drawing. The reason being that the creative realm is such a bold challenge, and it seems like creative activities like drawing have always being the defining mark of human intelligence, a sacred part of our experience that is hard to tap into and understand the root of this activity. On this end, numerous works have attempted to replicate the creative process on machines, allowing them to create paintings, write poetry, compose music, etc. One notable creation is GPT-3, and the book [] is co-authored by human writer and GPT-3. GPT-3 can also generate drawings from short poems. DALL-E, GLIDE, DALL-E 2 are all generative models that can generate incredibly creative work that even would make a human go, that's imaginative. Of course, creativity is not a solo activity, and in human society, creative activities are carried out in groups. As humans, we sparked each other's imagination. Song writers compose songs together, painters seek inspiration from their peers. We communicate our ideas to others, and it would be nice if creative machines can communicate with us. The idea that machines have by themselves a creative voice and a way of executing, interpreting, expanding on the short language prompt is enigmatic.

More so, we are motivated by how kids draw. Children start drawing from an early age, even before their language system has fully established, they are using symbols to represent things they experience in this world. It is almost an instinctual activity that is carved into our nature. A clear progression from simple scribbles to sophisticated composition of shapes. What is more inspiring is how children are able to describe to adults their creations and in these exchanges, metaphorical

expressions emerges, indicating that they understand the interaction of abstract shapes and concrete objects. This activity of projecting the high dimensional real-world experience down to the two-dimensional canvas showcases the achievement of human creativity. The completed processes that hide beneath the simple stroke lines of children art are yet to be understood. Inspired by this intriguing activity, we want to build robots that can draw with us, take as input our descriptions of the objects being drawn, and understand sentences like I want to show a large head for the angel. We want humans and machines to all be part of the creative process.

There have not been a lack of work on drawing and creative art. Proliferous amount of work on generative models, not mentioning the countless GAN-based generative models, such as StyleGAN. With the emergence of CLIP, work like StyleCLIP utilizes the rich joint embeddings space of vision and language to create art works. To pinpoint

To study the topic of human-robot collaborative drawing, we start the project by pinpointing how should we define a specific research problem. If we lay out the task, drawing, on a spectrum, we have on one end tasks of constructing still-life sketches that are highly realistic and look as if they are transformed from real images. From a small experiment, where

To shed light on how I got interested in the problem of human-robot collaborative drawing. The topic of text-image synthesis has existed for a very long time. And the academia has also been quite interested in drawing

Which aspect of drawing? Painting like Bob Ross?

Text representation. We have RNN to model the sequence of strokes. Can we still use RNN to model the sequence of words? Does the length of the sentence matter? If we only have adjectives, how can we effectively model these words? What are methods that can model the semantics of these words alone? Where can we find existing representations of these words? How are traditional text-image synthesis methods modeling the words?

Reed et al. [28] obtain the text encoding of a textual description by using a pre-trained character-level convolutional recurrent neural network (char-CNN-RNN). The char-CNN-RNN is pretrained to learn a correspondence function between text and image based on the class labels.

by collecting a dataset of sketches that we can learn such model from.

First meeting with Oliver and Yonatan. Between the two choices of task with long horizon, a long sequence of simple tasks, or short sequence of complicated tasks. World.

## 1.1 Exploring the Problem Space, early October

Master Thesis Brainstorm

The discussion starts in Oliver's office. I am trying to pinpoint where I want to bring in the ideas of natural language processing. What are some ideas that I wanted to explore.

The first question that we want to solve is what data should we collect in order to build a robot



that can draw with humans? Inspired by drawing sessions created by Bob Ross, we have considered painting with robots and experimented briefly with oil painting on canvas with a Franka robot, but precise execution of brush strokes and the technical details surrounding manipulating brushes to create the desired scene are even very difficult for humans, and the focus of our research is to make a step towards human-robot collaboration on tasks, we do not want to be side-tracked by difficulties around robot control, which in itself is an extremely interesting problem that we wish to explore another time.

What inspired us to choose sketch was a session we conducted with another student studying Graphics. Initially, we set up the session to get a sense of how difficult it was to teach someone to draw with only language commands, and along the way, we discovered that generating real-looking sketches are very difficult without visual demonstrations from the instructors, but creating emoji-looking icons is more do-able. What is more interesting is that, icons are intrinsically abstract in that general geometric shapes like circles and rectangles are used for different objects and convey different meanings in different context. For example, a person can draw a circle to represent a left eye or to represent a face. Moreover, a face can be of different shapes: circular, oval, triangular, etc. This abstract nature of icon-like sketches is something interesting that we wish to explore, so we want to collect a dataset of sketches done by humans but, most importantly, we want to track the steps people take to complete them so that we can achieve the goal of drawing together with a robot.

## CLIP

□

# Chapter 2

## Data Collection

### 2.1 Overview

Imagine the following scenario (inspired by the YouTube channel:[]): Today we are going to draw a smiling ice-cream cone. Okay, we are going to first draw a curve as the top of a big scoop of ice-cream. Next, we will draw a sequence of connected U's to represent the bottom of the overflowing ice-cream. Lastly, we will draw a large upside-down triangle as the cone of our ice-cream.

We want to realize this kind of interactions with a robot, as a companion, so we need to collect a dataset that can help us to get closer to this goal. In order to study this problem, we want to collect human sketches, so the first thing we did was designing an web interface. The leading questions of the data collection process. Our goal is to collect a dataset so that we can learn a model that can interactively draw sketches with users. Therefore, we want to collect the drawing for a single step and a person's description of the drawing. Our design of the interface centers around some key questions:

1. Ensure that the drawing responds to the prompt. The underlying assumption here is that the prompt itself will give us some signals in terms of where the objects in the images might be.
2. From the design side, enforce annotators to breaking the sketch generation process into steps. The worst scenarios is for the annotators to
3. How do we make sure that annotators are breaking the sketches they provide into reasonable steps? What we mean by reasonable here is the fact that there should be a good correspondence between some parts of the sketch and the language that is used to describe it. Although in our daily interactions, we might say something like “we now draw this” or “we can do this”, but from a model learning perspective, or more so as a first step, we want there to be little ambiguity in our language and disallowing words like “this”.

Our interface has experienced 2 main versions, and the major difference between the two is that the first one asks users to draw the sketches and annotate each step in their drawings while the second version asks the users to annotate existing sketches. The turning point happens after a pilot deployment of the first version, during which we identified several problems: (1) users take too long to complete one task, and it is outside our budget to collect an ample dataset; (2) users cannot separate the entire sketch into steps consistently, and the annotations either describe more or less than what was done in a single step. In order to shorten the task time and alleviate the burden to think about how to draw certain objects, our second version uses sketches from the Quick,Draw! dataset collected by Google and asks users to provide textual annotations for each part in the sketches. The following sections will walk through each version and discuss the data collected using each version. The following sections will walk through each version and discuss how the design reflects or answers the above N criteria and what in reality happened that caused us to change the design.

Later, we discovered that by simply using existing sketches without asking for users to draw for the prompt would significantly reduce the data collection time, and it would also allow us to put aside DQ 2. In general, if you think about it, classic collection tasks such as assigning label to images/texts or drawing segmentation box, the goal of the task is very clear, and it is easy to determine the quality of the work when you glance at it, or easy to verify. At the beginning, we found it very difficult to describe what should be drawn and what should not be drawn, or what can be written and what cannot be written.

The general trend of the data collection process is that we try to simplify the data collection interface and reduce the number of criteria that we need to satisfy, since each introduces a factor of uncertainty.

## 2.2 Version 0

Since the beginning of data collection, an important question we try to answer is how do we define a semantic unit in the sketch? The end goal is to achieve the kind of interaction shown in the YouTube video *How To Draw A Cute Ice Cream Cone*, and in it, the instructor often uses sentences in the form of “Let’s draw a **X** for **Y**”, where **X** describes the geometric features of the object **Y**. For example, “Let’s draw *small connected U shapes* for the *bottom of the ice-cream cone*.” Therefore, at first, we thought of decomposing the drawing process into a sequence of common geometric shapes, and the objects that they represent become the basic semantic units. At each step, the annotator is first asked to Version 0 was never deployed. I think at this stage of the data collection, we are trying to decide whether there should be a fixed set of primitive that the users could choose from, so learning the model becomes learning to parameterize, for example, the dimensions of the set of primitives.

Functionality:

- Draw the figure and the page will record the sequence
- User can replay its drawing sequence. The original idea was that users will first create the drawing, and then they can replay the sequence as they annotate for each step.

The very first test version: In terms of the main task, I created a test version to confirm that the idea of the drawing board is sensible.

Press *Record*, Draw on the board, Press *Stop* when done with drawing, *Submit* the drawing if one is satisfied with the quality, *Play* to revisit the drawing, *Cancel* to start over.

What was the original motivation behind this functionality was that it will aid the annotators to review the drawing process and divide it into better steps. Responding to DQ 2. However, in this very crude version, we did not really incorporate features for either Responding to DQ 3,

We begin with a very crude version, and then we decide to add features that can allow us to realize the DQ 2 and 3.

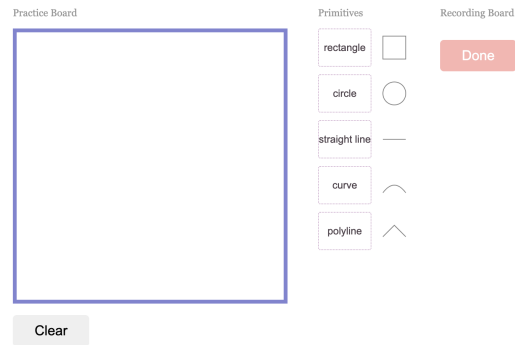
The actual Version 0 has the following flow: There is a practice board, you can try to practice drawing so that the actual drawing submit has good quality and respond to the prompt (reflecting DQ 1). Then hit *Ready to Record*, again baking the sequence into the design of the website will help us to enforce collecting a dataset of steps. Another purpose is to help the annotators decide beforehand what are the necessary primitives used in the process. Why was I so fixated on the primitives, because the abstractness of the icons is what interested me the most. The entire research journey was very explorative, it sorted of started with a sense of *oh, this question or aspect of how humans do things is interesting, I wish robot can do the same*. And what is that thing that I thought was interesting, it was how Rain and I were able to draw the icons and the interactions. The first thing you will do is select a primitive from a list, and then you will draw the step that contains the primitives. Hit *Next* to move on to drawing the next primitive. There are will be a little tag at the bottom showing what is the primitive that corresponds to the step that is drawn on the board. Repeat until finished and hit *Done*. At the end, again, *Play*, *Submit*, or *Cancel* to start over.

**?** Should we use primitive shapes for users to choose from? The reason for considering this aspect is whether during generation we want to learn to change parameters of a fix set of shapes or generate un-constrained strokes. For the first option, we want users to compose a drawing with primitive shapes, much like using In order to learn a more general model, we decided that we want to collect strokes instead of fixed primitive shapes, so we moved onto creating a table that accompanies the drawing board, where the user can choose to annotate each step they draw.

In Figure 2.1.

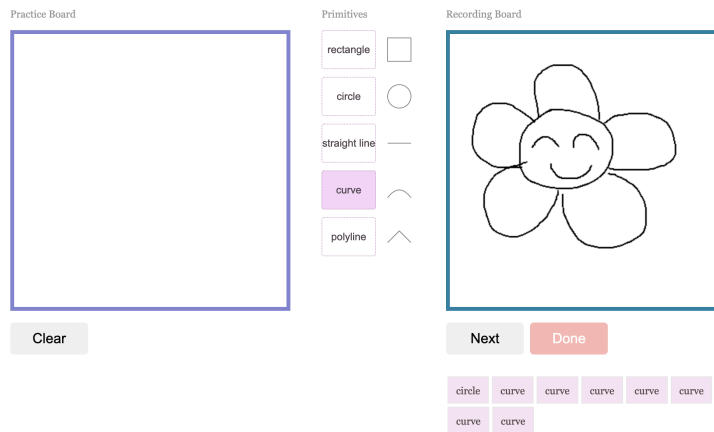
**Annotation Instructions**

Select the primitive used in step No.1.



(a) Design of main task for third pilot.

Draw the component using the primitive selected step No.9. Press `Next` when you are ready to move on to the next step.



(b) Design of main task for final task.

Figure 2.1: Progress of the design two for the main task in version two.

## 2.3 Version 1

### 2.3.1 Overview

For version 1 onward, we decided to host our website on Amazon Mechanical Turk (AMT), which is a crowd-sourcing website that hosts different machine learning annotation tasks. (In the remaining text, we use the word *turker* to refer to annotators that we recruit on AMT.) We have to

design the following sections for the task:

1. an interface containing the main task
2. instructions and requirements to describe the tasks and specify what the annotators should and should not include in the annotations
3. a qualification task accompanying the main task to train the turkers to produce high-quality annotations

Compared to Version 0, which we only dabbled with 1 in the above list, we went through all three stages for Version 1 and eventually deployed a pilot. After deployment, we realized that a few major problems from this design: (1) due to the subjective nature of drawing, it was hard to understand in what ways some annotators are illustrating the given prompts, thus making it difficult to determine the quality of the annotations; (2) turkers are taking more than 30 minutes for each task, showing that providing both sketches and descriptions are inefficient; (3) some turkers are unable to provide descriptions that align with the objects they meant to annotate for; for example, in one step, they drew both eyes and hair, but they only annotate “big eyes”.

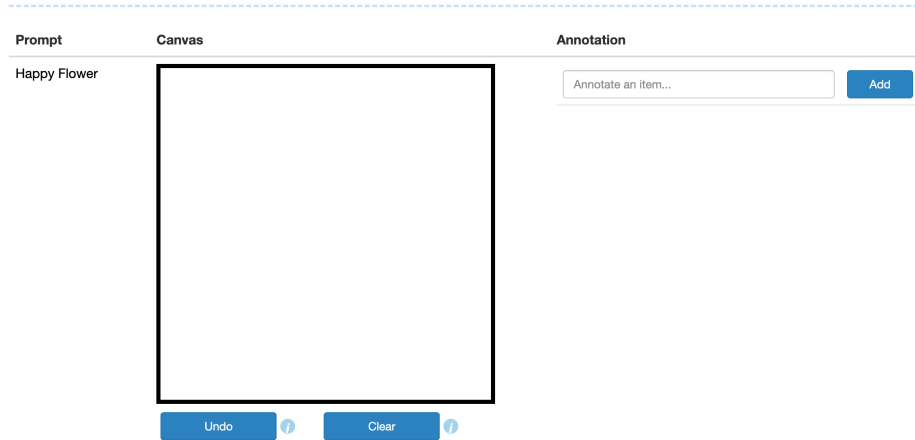
### 2.3.2 Interface Design

#### Main Task

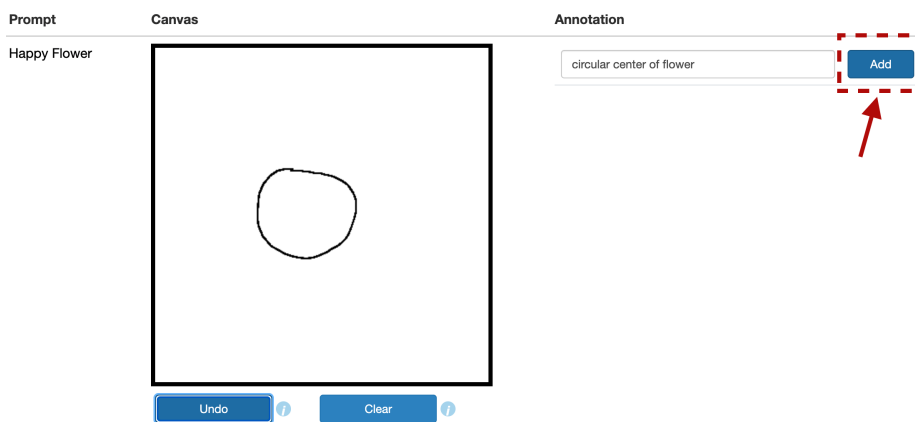
Compare to version 0, we make the following changes to the task interface:

1. Since turkers are paid based on time spent on the task, we decided to forsake the functionality related to the recording and replaying the drawing board.
2. Since we decide to not limit the drawings to be compositions of basic geometric objects, we removed the step to select primitive shape preceding drawing each component.

We illustrate a typical annotation process with Version 1’s interface in Figure 2.2. The annotator starts with an empty canvas and empty table for textual descriptions, as shown in Figure 2.2a. For the annotator’s convenience, we include a *Undo* button and a *Clear* button for erasing strokes and clearing the entire canvas. Then, the annotator draws a step in the sketch, and they would need to enter the text description for this step into the *Annotation* column and hit *Add* to display it as a new row in the annotation table. (Figure 2.2b and 2.2c show what the annotation table looks like before and after adding the texts for a given step.) If the annotator wants to remove an entire step, the objects in the drawing and the text description, they can use the *Delete* button next to the texts to do so. An example is shown in Figure 2.3. Repeat the drawing-and-adding process until the drawing is done. The design of a table for adding and deleting text annotations intends to encourage turkers to break their drawing into a series of semantically meaningful parts, responding



(a) Main task interface at the start of annotation.



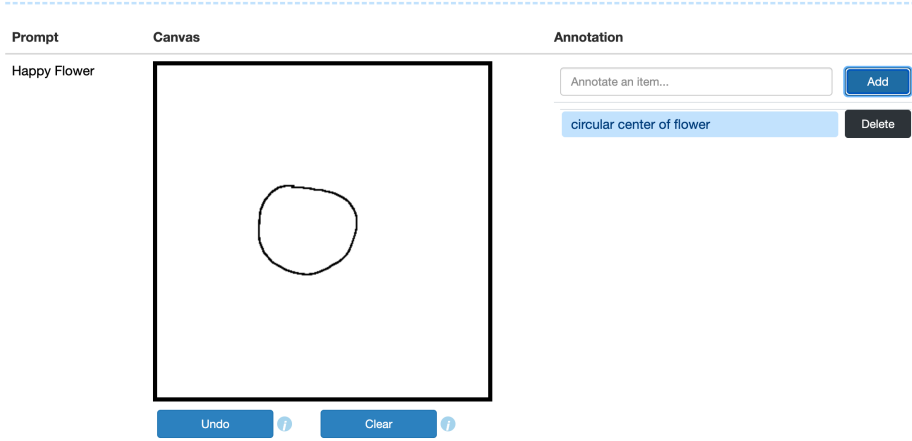
(b) Main task interface before adding text descriptions for the drawing in the given step. Red arrow and box show where to click to add the text.

to principal 3 and 1. Enabling users to be able to delete each component is also demonstrative of these two principals.

We encountered some difficulties when implementing the *Delete* functionality. At the beginning, we treated erasing strokes as drawing the same strokes but in white color; however, when strokes overlap each other, overwriting with white strokes would break other strokes into segments. Therefore, we designed the drawing canvas to use layers like Photoshop, so that deleting strokes would be the same as deleting an entire layer, thus leaving other strokes intact.

### Instruction and Requirement

We illustrated the layout of the main task interface in the instruction, as shown in Figure 2.4. We begin the instruction by giving a short explanation about the motivation behind collecting this



(c) Main task interface after adding text descriptions for the drawing in the given step.

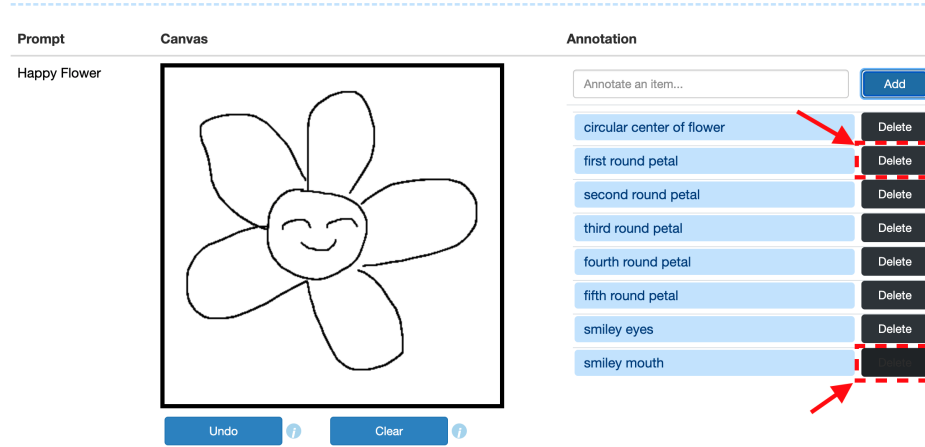
Figure 2.2: The “drawing-and-adding” functionality in Version 1. Repeat the above process for each step in a sketch.

dataset to prime turkers for good-quality annotations, since they would understand more about what the data will be later used for. the better they understand what the data will be used for, the The first thing we spent time considering was how much motivation should we give about the task. Why are we conducting this survey? The second, more important, aspect to consider is what defines a single *step* in the data that we collect. Should the annotator be asked to annotate for each stroke? However, this option is not only time-consuming but also fails to align with how a person would talk when, for example, teaching a child how to draw. Therefore, we decided that the annotator should annotate for each *object* in the drawing. The ambiguity surrounding the word *object* has been the biggest challenge in defining a clear set of requirements for the annotators. For example, when drawing for the prompt *happy face*, one reasonable decomposition is annotating for 4 steps: face, eyes, mouth, and the face contour. However, for someone who draws very detailed eyes, they might want to annotate for the shape of the eye socket and the length of the eye lashes. It seems like there is a wide spectrum of allowed annotations depending on how one would approach drawing for the give prompt. Indeed, the great uncertainty that comes from individuality and personal styles of collecting drawings from turkers would eventually drive us to not collect drawings and simply annotate for sketches found in existing datasets.

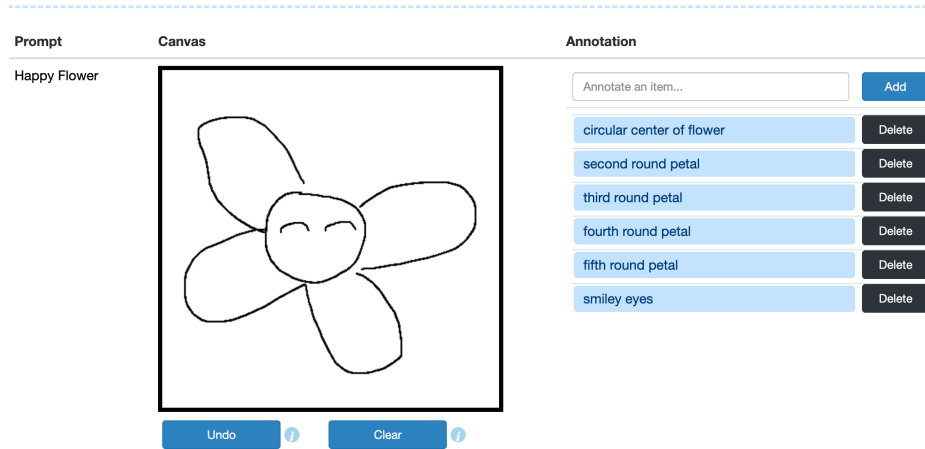
A lot of effort went into defining what is the basic element of the annotation. We tried item, object, component. The goal of the model is to conditioned on previous steps to produce the next part. Figure 3x.a and 3x.b show some previous versions of the instructions. We eventually decided that

There is even more effort that went into defining the requirements of the task. This is really the section that we want to use to enforce all the DQ’s. The final set of instructions is displayed in





(a) A complete annotation for the prompt *Happy Flower*. Red arrows and boxes point to *Delete* buttons that can be used to delete the steps associated with the textual annotations, if the annotator is not satisfied with the steps.



(b) Main task interface after the two steps associated with *first round petal* and *smiley mouth*, respectively.

Figure 2.3: Demonstrating the functionality of the *Delete* button.

Figure x4. [Figure x4: final requirements]

Comparing examples in requirements format:

Methods to help understand the requirements. Specifying which examples demonstrate which requirements, add next to the questions which requirement the question is testing.

### Qualification

A qualification test is setup on AMT to train turkers to have better understanding of the task, and it is also used to select a group of turkers who can provide annotations that satisfy all the requirements. The full test is illustrated in Figure x5.

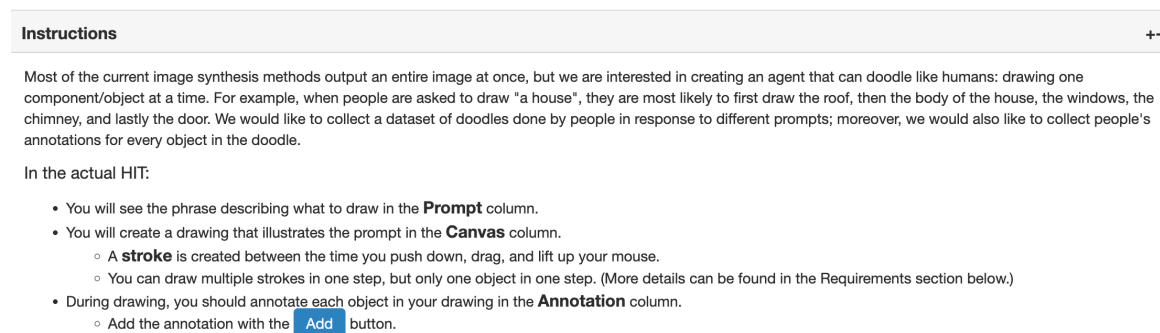


Figure 2.4: Instruction section of Version 1.

[Figure x5: final qualification test]

To come up with a set of questions that have good correspondence with the requirements, we went through several rounds of testing with students in the lab.

The format of the qualification has also went through transformed, as shown in Figure x6. At first, we asked the annotators to select which steps of the annotations satisfy the requirements (Figure x6.a); in order to use repetition to ensure deep understanding of the requirements, we changed to asking a yes/no question for every step, as shown in Figure x6.b.

[Figure x6]

### 2.3.3 Deployment Results

In order to determine how feasible the task is, we first deployed a version among lab members, and we obtained 55 drawings along with their annotations. All 55 drawings are shown in Figure v1.results.2. Some examples of step annotations are shown in Figure v1.results.3.

[Figure v1.results.2: 55 drawings from lab deployment (see jupyter notebook oct\_28\_trial\_analysis)]  
 [Figure v1.results.3: 3 examples?]

To come up with a set of prompts to test for the first pilot, we want to come up with prompts in the forms of *adjective* × *noun*. The list of adjectives includes: *happy, sad, surprised, sleepy, lovestruck, evil*; the list of nouns includes: *person, kid, cat, bear, dog, sheep, jellyfish, cup of boba, apple, burger, sun, moon, star*. We hope to test whether annotators can draw for abstract prompts like. Moreover, we wish to see whether steps can be clearly annotated for these abstract prompts. It was also motivated by the fact that current AI-generated artwork like those produced by DALL-E often test its methods with abstract prompts, and the hope was that our collaborative agent can also respond to these prompts through interactive drawing.

[Figure v1.results.4: drawings from the amt pilot]

What surprised us was the amount of time turkers spent on the task. Histograms of time each annotator spent on the task is illustrated in Figure v1.results.1. Statistics of the distributions are

shown in Table v1.results.1. The discrepancy might be caused by the fact that lab members with their background in computer science have an implicit understandings of what kind of quality data are needed to train ML models.

[Figure v1.results.1: a: oct 28 lab deployment. b: dec 28 amt deployment] [Table v1.results.1: comparing the statistics of lab vs. amt deployment]

In violation of DQ 2. Drawing does not illustrate the prompt well. The quality of the drawings are greatly influenced by how well the annotator can understand the prompts. Drawing is by its nature very subjective, so when we were examining through the sketches that we collected, we were not able to understand in what ways some sketches convey the prompts. [Figure v1.results.4: some examples of sketches that cannot illustrate the prompt from our perspective]

In violation of DQ 3 and 1. Another problem was that annotators often fail to describe every parts they drew in one step, or the descriptions miss some parts in the step, or the description does not align well with the drawings. [Figure v1.results.5: some examples of mis-aligned descriptions]

## 2.4 Version 2

### 2.4.1 Overview

In response to the pilot results, we reconsider the data collection pipeline to reduce the uncertainty around collecting drawings that illustrate the prompts and textual descriptions well-aligned with each step in the drawing. Firstly, in order to alleviate the burden of drawing from the annotators, we examined existing sketch datasets, and information regarding the advantages and disadvantages are shown in Table v2.datasets.1. [Table v2.datasets.1: pros and cons, stats of different sketch datasets]

Between Sketch Perceptual Grouping (SPG) and SketchSeg, both containing annotation for semantically meaningful parts in sketches, SPG annotates for QuickDraw sketches while SketchSeg collects its own sketches. We picked SPG, since it will be easier in the future to extend our datasets given the large QuickDraw reservoir of sketches. Moreover, SketchSeg dataset contains a *fourleg* category that includes many different kinds of animals, such as horse, sheep, and cow, but the QuickDraw categories are more fine-grained, so from a model learning perspective, SPG will also be more generalizable. Therefore, to combine our previous goal, collecting sketches that illustrate certain *adjective*  $\times$  *noun* prompts, we decided to provide annotators with the sketches from QuickDraw and ask them to annotate for each semantically meaningful part provided in the SPG Dataset.

In order to avoid dealing with discrepancies between performance of fellow graduate students and that of the turkers, we deployed a short pilot test of the new version and identified the suitable format and areas that need written requirement for the turkers to avoid these mistakes. Therefore, the transformation of the task format is driven by mistakes we have seen during the pilot trials.

## Main Task



In Figure 2.5, we show the transformation from our first pilot of version to the final task that is deployed to collect the entire dataset. Overall, we used non-gray colors to highlight the parts in the sketch that we want annotations, another design to help with annotation speed. For simplicity, we restricts the annotations from whole sentences (Figure 2.5a) to only adjective phrases (Figure 2.5b, 2.5c, 2.5d). The benefit of juxtaposing two sketches and simultaneous annotate for two sketches is that annotators are implicitly encouraged to provide descriptions that identify features of the objects that differentiate the two sketches. This method is also proposed to facilitate the annotation process and to take less time, since it is easier to differentiate and perform a contrasting task than to generate descriptions from a single sketch. At the beginning, we explicitly mention that the goal of the task is to describe the differences between the objects in the sketches (*Describe differences* in 2.5a and *Compared to Sketch 1/2* in 2.5b), and we received many annotations that contain comparative and superlatives, so we eventually only have a blank without any introductory phrases to overly emphasize that the goal of the tasks is to create a dataset of contrastive pairs of descriptions, and the juxtaposition is meant only as a mental hint to ease annotation.

## Instructions



At first the set of instructions was very restrictive and limit the annotators to pay attention to three types of differences: shapes, size relative to other objects in the same sketch, position relative to other objects in the same sketch. The general trend of the changes to the instructions is that we only require that the annotators fill in the blank with adjective phrases and try to not put too much restrictions on the language, in order to achieve our goal of building a dataset with free-form language instructions.

In this version, the advantage is that since we have greatly simplify the task to only providing the textual descriptions, the turkers do not have to spend time coming up with drawings for a *adjective* $\times$ *noun* prompt, and they do not have to put effort into keeping track of their drawing process to decide how to divide the drawing process into steps and then annotate for each step. Essentially, they only have to do the last step. Therefore, the requirements are much easier to write, and we do not have to specify anything in terms of providing drawings that correspond well with the prompts and providing annotations that align well with the drawings in the each step. One thing we tried was to somewhat rely on the examples to give an idea of what kind of annotations we want. Some examples that we used in the tasks are shown in Figure . However, the downside for doing so is that the vocabularies used in by the annotators are primed by those in the examples, and we see that annotators would tend to repeat these vocabularies. Therefore, we especially added the requirement that states the annotators are not limited to words used in the examples, and they should use any words that can illustrate the parts well. The full set of requirements used in the final version is shown in Figure 2.7.

## Annotation 1

Sketch Category	Sketches	
angel	<div>Sketch 1</div> 	<div>Sketch 2</div> 
Describe differences between the <b>angel bodies</b> (strokes in <b>magenta</b> color) in the two sketches.		
<input type="text"/>		

(a) Design of main task for first pilot.

Sketch Category	Sketches	
angel	<div>Sketch 1</div> 	<div>Sketch 2</div> 
Q1: Compared to Sketch 2, Sketch 1 draws a/an <input type="text"/> <b>angel body</b> (strokes drawn in <b>magenta</b> color).		
Q2: Compared to Sketch 1, Sketch 2 draws a/an <input type="text"/> <b>angel body</b> (strokes drawn in <b>magenta</b> color).		
<input type="checkbox"/> If someone is shown the two sketches, the person can pick out one sketch based on the provided differences.		

(b) Design of main task for second pilot.

The requirement that was a bit challenging for people to understand was the one regarding *Do not use adjectives related to personal opinions, such as random, good, messy, beautiful, and strange, that are hard to achieve consensus if others were to validate your answers..* Since we hope that the model can get signal from the texts about what kind of figures to draw, words that do not directly convey visual properties of the parts are not helpful. We later changed the wording to *Do not use adjectives that fail to describe specific visual properties of the objects in the sketches.* A slight caveat

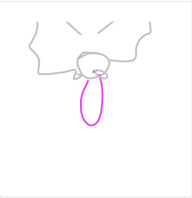

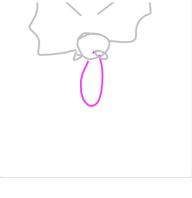
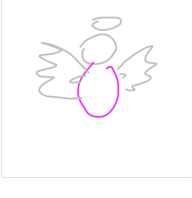
Sketch Category	Sketches
angel	<div style="display: flex; justify-content: space-around;"> <div style="border: 1px solid #ccc; padding: 10px; width: 45%;"> <p style="text-align: center;">Sketch 1</p>  </div> <div style="border: 1px solid #ccc; padding: 10px; width: 45%;"> <p style="text-align: center;">Sketch 2</p>  </div> </div>
<p>Q1: Compared to Sketch 2, Sketch 1 draws a/an <input type="text"/></p> <p>angel body (strokes drawn in magenta color).</p> <p>Q2: Compared to Sketch 1, Sketch 2 draws a/an <input type="text"/></p> <p>angel body (strokes drawn in magenta color).</p> <p><input type="checkbox"/> If someone is shown the two sketches, the person can pick out one sketch based on the provided differences.</p>	
(c) Design of main task for third pilot.	
angel	<div style="display: flex; justify-content: space-around;"> <div style="border: 1px solid #ccc; padding: 10px; width: 45%;"> <p style="text-align: center;">Sketch 1</p>  </div> <div style="border: 1px solid #ccc; padding: 10px; width: 45%;"> <p style="text-align: center;">Sketch 2</p>  </div> </div>
<p>Q1: Compared to Sketch 2, Sketch 1 draws a/an <input type="text"/></p> <p>angel body (strokes drawn in magenta color).</p> <p>Q2: Compared to Sketch 1, Sketch 2 draws a/an <input type="text"/></p> <p>angel body (strokes drawn in magenta color).</p> <p><input type="checkbox"/> If someone is shown the two sketches, the person can pick out one sketch based on the provided differences.</p>	
(d) Design of main task for final task.	

Figure 2.5: Progress of the design two for the main task in version two.

here is that we actually hope to collect descriptions that describe the emotions expressed in the sketches. We know beforehand that we hope to collect a dataset for the *face* category, so it is quite common for faces to express emotions like happy and sad, and we were slightly worried that some turkers might consider these words as not illustrating enough visual properties about the drawings, since they are quite abstract, at least compared to adjectives like *rectangular* or *wide*.

### Qualifications

We prepared 10 qualification questions, all in the style of yes/no questions. We will use the qualification test to filter annotators who have read through all the instructions and examples and have formed a good understanding of the task. The 10 questions are shown in Figure 2.8. We provides hints in each questions that explicitly state which requirement and examples are helpful for solving the questions. The purpose of the qualification test is not to trick annotators but to ensure both quality and speed of the annotations.

We released  $n$  copies of qualifications, and  $n_2$  annotators scored 90 or higher. The average score for the entire test is  $x$ , and the rate of correct answer for each question is shown in Table 2.1. Before releasing the qualification, we have tested the test on

Question Number	1	2	2	2	2	2	2	2	2	2
Correct Rate	1	2	2	2	2	2	2	2	2	2

Table 2.1: Success rate of each question in the qualification test

### 2.4.2 Results

#### Pilot 1

In order to work out the data collection process, we chose the angel category and try to manually examine the sketches and categorize them based on

One purpose of the pilot is to estimate the amount of money that we need to spend for each task, and from Table 2.2, we see that []

	Max.	Min.	Mean	Med.	Std.
Feb 01 Pilot	1	2	2	2	2
Feb 04 Pilot	1	2	2	2	2
Feb 08 Pilot	1	2	2	2	2
Official Collection	1	2	2	2	2

Table 2.2: Comparing time statistics of pilot task

For the data collection process, we decide to collect for the face category of the QuickDraw dataset, and the reason for it was mainly to echo the choice of many SOTA generative modeling works that are done on the CelebA dataset. It seems that face generation is quite a starting point for many of the generative modeling work. We have also surveyed some text-to-image synthesis methods that use datasets like (1) CUB dataset (2) MNIST (3) Omniglot. Several sketch datasets include the one from DoodlerGAN and SketchBirds. A lot of the datasets focus on one or two categories,

so we decide to do the same to ensure that with our budge, we can collect a dataset that contains enough signal to train a generative ML model.

Clustering the faces, we strive to present to the annotators pairs of faces that are distinct as possible in order help them to provide good annotations. It is easier for them to grasp and understand the features of the objects if two sketches are presented in a contrasting way.

If we use CLIP to extract the visual features for the entire face sketch.

The heuristic that we use to choose how to pair up

## 2.5 Dataset Summary

Our dataset comprises of Quick,Draw! sketches and language descriptions of each semantically meaningful part in the sketch. The dataset contains 2 categories: face and angel, and these categories correspond directly to those in the original Quick!Draw! dataset. The part annotation comes from the SPG dataset (Li et al., 2018). For the angel category, we annotate for the parts *halo*, *eyes*, *nose*, *mouth*, *body*, *outline of face*, and *wings*. For the face category, we annotate for the parts *eyes*, *nose*, *mouth*, *hair*, *outline of face*.

	Face	Angel
Number of constrastive pairs	2515	3060
Number of distinct words	833	1107
Number of sketches	572	787

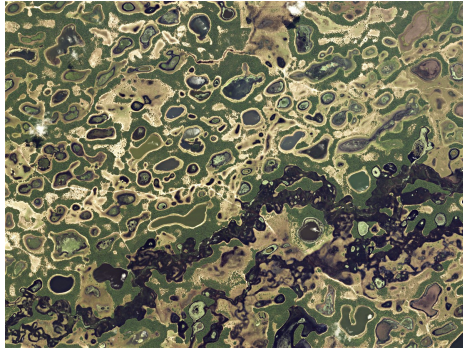
Table 2.3: Statistics of the dataset by category.

	Face					Angel						
	eyes	nose	mouth	hair	face	halo	eyes	nose	mouth	face	body	wings
Number of sketches	334	572	572	104	572	558	114	8	80	732	781	779
Number of distinct words	228	360	325	152	314	365	112	21	88	379	425	534
Number of constrastive pairs	689	401	687	126	612	559	114	8	80	733	785	781

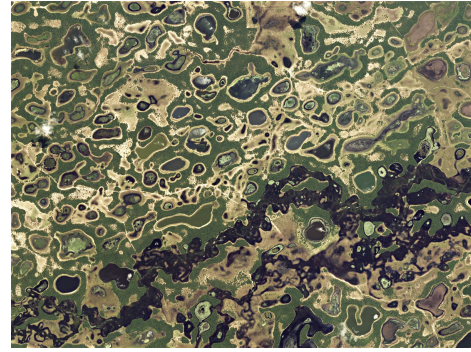
Table 2.4: Statistics of the dataset by sketch parts.

In Table 2.4, we see some statistics about the dataset broken down by sketch parts, while in Table 2.3, we all list out the same statistics for the entire face and angel category. In general, we observe that compared to previous work that tend to have a fixed list of adjectives for each object parts, the descriptions in our dataset are free-form and non-constrained. This characteristics is desirable and aligns with our goal to allow robot to collaborate smoothly with humans, since different people would describe the same things in very diverse ways.

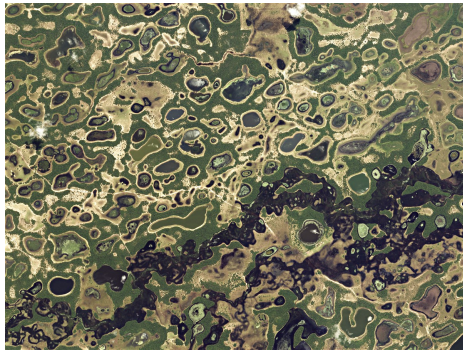




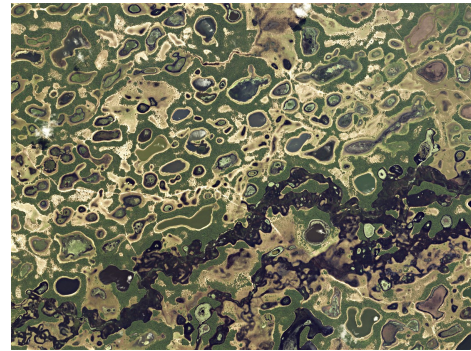
(a) Design of main task for first pilot.



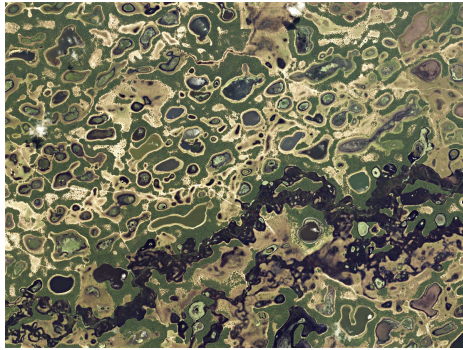
(b) Design of main task for first pilot.



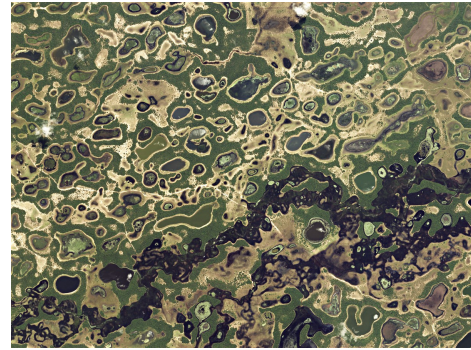
(c) Design of main task for second pilot.



(d) Design of main task for second pilot.



(e) Design of main task for second pilot.



(f) Design of main task for second pilot.

Figure 2.6: Progress of the design two for the main task in version two.

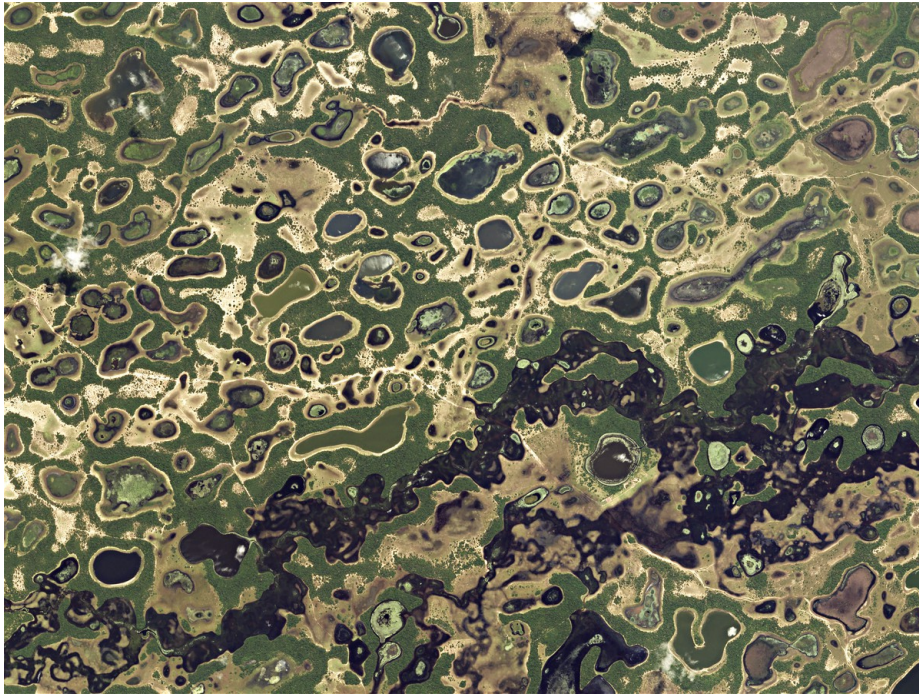


Figure 2.7: The set of requirements used in the final task.

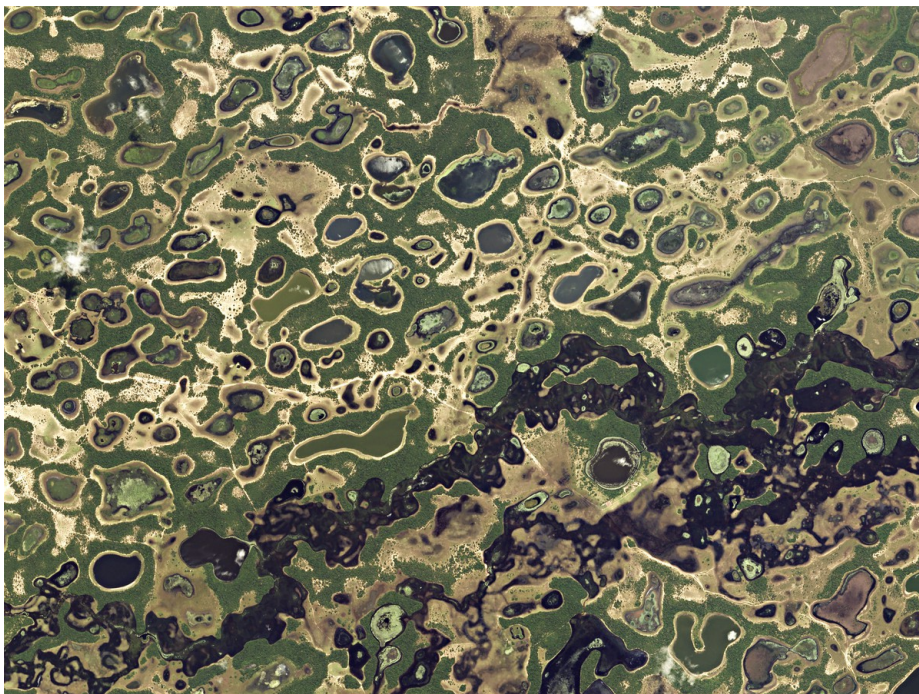


Figure 2.8: The qualification questions.



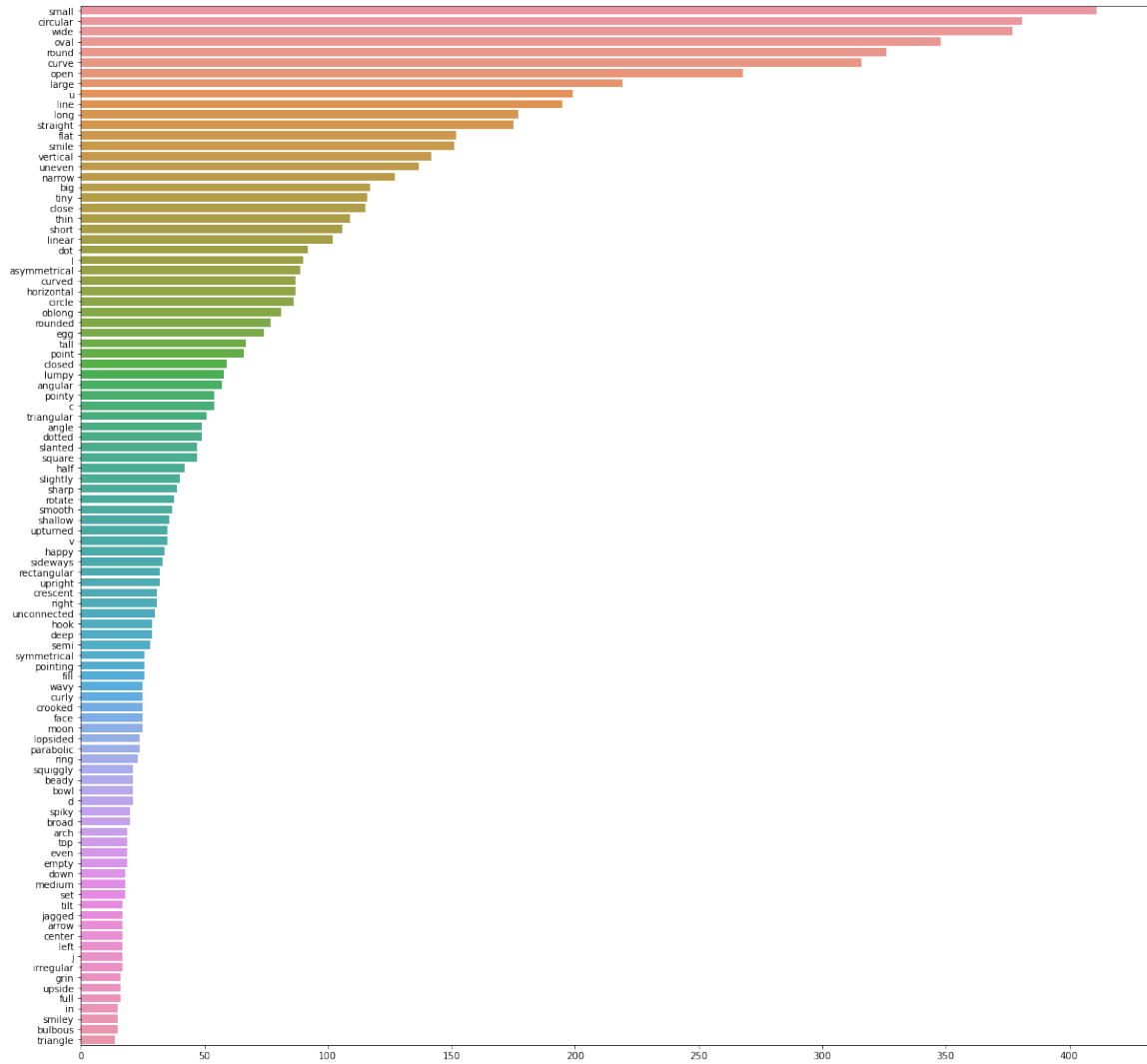


Figure 2.9: Top 100 most frequent words in the dataset corpus.

# Chapter 3

## Methods

In order to gain insights about the nature of our dataset, and whether models like CLIP that have been trained on text-image pairs on the million scale can handle and interpret annotations for sketches that we have collected. We report baseline results on the binary classification of CLIP.

### 3.1 CLIP Finetune

#### 3.1.1 Image Pre-Processing

We use the data provided by SPG (Li et al., 2018), which provides JSON files of the Quick,Draw! sketches in vector format: each sketch is composed of a sequence of  $n$  strokes  $S_i, i \in [n]$ , and  $S_i$  is a sequence of vectors  $(\delta x, \delta y, p, l)$ .  $\delta x$  and  $\delta y$  are changes in the  $x, y$  coordinates with respect to the previous point; for the first point, it is with respect to  $(25, 25)$ . All points are assumed to be drawn on a  $256 \times 256$  canvas.  $p = 1$  if the point is the last point in the current stroke, and  $p = 0$  otherwise. The SPG dataset provides annotation for semantic segmentation of the sketches, so  $l$  is a number representing the semantically meaningful object part.

We obtained the rendered sketches by using `Pycairo`, which is a Python module providing bindings for the cairo graphics library. We use a line width of 5. After rendering, we manually examined the sketches and filter out face sketches that do not have a pair of eyes, a mouth and the face outline; we also filter out angel sketches that are incomplete or have all the parts merged together, possibly due to collection errors in SPG.

#### 3.1.2 Text Pre-Processing

We used the `spacy` package to preprocess the text. `spacy` provides trained natural language processing pipeline and includes models for, for example, token-to-vector and part-of-speech tagging. We use the `en_core_web_sm` pipeline and its lemmatizer to reduce words to their basic forms.

Moreover, we lower-case all words and remove punctuation, a list of which is provided by `Python string` package, `string.punctuation`. We also remove words like *shaped*, *sized*, *and*, *like*, since they act like stop words and do not provide additional visual descriptions of the sketches.

We load the pretrained model from the `Python clip` package, specifically the ViT-B/32 variant, which uses the Vision Transformer (Dosovitskiy et al., 2020) as the image encoder; B stands for BERT Base model, and 32 stands for  $32 \times 32$  input patch size.

## 3.2 Loss Function

## Chapter 4

# Results & Analysis

### 4.1 Classification Experiments

In this set of experiments, what we do is that if we have a pair of sketches  $(s_1, s_2)$ , we use the CLIP image encoder (zero-shot/fine-tuned)  $f_v$  to extract visual features  $v_1$  and  $v_2$  for the two sketches, where  $v_1, v_2 \in \mathbb{R}^{512}$ . We then use the zero-shot/fine-tuned CLIP text encoder to extract the text features for the part descriptions, namely we fill in the template  $t = [\text{ADJ}] [\text{PART NAME}]$ , where  $[\text{ADJ}]$  is filled with the adjective phrases annotations, and  $[\text{PART NAME}]$  is the name of the part in the sketches. For angels,  $[\text{PART NAME}]$  is one of *halo, eyes, nose, mouth, body, outline of face, wings*; for face,  $[\text{PART NAME}]$  is one of *eyes, nose, mouth, hair, outline of face*. After filling in the above template, we obtain the part annotations for the two sketches  $t_1, t_2$ . During data collection, we implicitly juxtapose two sketches, chosen to be as distinct as possible using some heuristic, either from different clusters or whose cosine distance is large, so the process of annotating the two dissimilar sketches is like the annotators are choosing the pair up one annotation with another. Implicitly, the annotators is pairing  $s_1$  with  $t_1$  and  $s_2$  with  $t_2$ , so we would regard the ground truth pairing to be  $(s_1, t_1)$  and  $(s_2, t_2)$ . We want to see how CLIP does on this task, if it is the annotator for the task, would it be able to generate the same pairing. Define cosine similarity to be

	Face		Angel	
	Test	Dev	Test	Dev
zero-shot	1	1	1	1
finetuned on face	1	1	1	1
finetuned on angel	1	1	1	1
finetuned on face + angel	1	1	1	1

Table 4.1: Statistics of CLIP

## Chapter 5

# Related Work

In the space of human-robot collaborative drawing, we are aware of the work by

Eitz et al. (2012) is one of the first works to investigate the characteristics of free-hand sketch and attempts to extract local features from these sketches, which are later used in the task of sketch recognition. It also provides the dataset TU-Berlin that contains 20,000 human sketches, and it includes 250 object categories with 80 samples in each. Quick,Draw! gathers an even larger pool of 50 million sketches, spanning 345 object categories, each containing around 100,000 sketch. A proliferation of work on sketch data followed from this large-scale sketch dataset.

In the space of sketch representation learning. After we have settled on human-robot collaborative sketching, we surveyed the field for existing sketch datasets and what they contain, what they lack, and what gap does our work fill. Sketch representation learning is regarded as a vision task, and it has several tasks associated with it: sketch recognition, sketch generation from image, image generation from sketches, sketch retrieval of 3D objects, sketch retrieval of images, semantic segmentation of sketches, etc. Essentially, one can perform any tasks that are done on images and explore the techniques for sketches. People can refer to the survey paper by P. Xu et al. (2020) for a more comprehensive overview of the subject. There is a wide range of tasks that can be done on sketches, both unimodal and multimodal, and, for each task, a large reservoir of deep learning methods used to solve the tasks. P. Xu et al. (2020) gives a comprehensive review of the task taxonomy, summarized the unique challenges associated with each individual tasks, and evaluated the different deep learning methods on sketch recognition through a library `TorchSketch` the authors wrote, and it contains implementation for CNN, RNN, GNN, and TCN. The sections that are most relevant to us are: sketch generation, sketch segmentation. Sketch generation because we are trying to learn a generative model. Sketch segmentation because we are trying to gain insight about how are semantically meaningful units discovered in sketches and what relationships do the parts have with the whole sketch. Similar to images, sketches have hierarchical structure, and we

The hope is that we can leverage previous work on sketch representation learning to gain insights

about sketches and how to learn good representations of them. What is unique about sketches compared to regular RGB images from, for example, ImageNet is that (1) sketches are abstract characterisation of the objects, and although humans can recognize and understand a sketch perfectly, they do not necessarily bear big resemblance to their image counterpart; therefore, methods that work well on RGB images, especially generative models like GAN that have successfully generated wide range of images from texts, a realm that we care about, it is not necessarily the case that they can generalize well to our dataset.

On the other hand we have interactive drawing, and the seminal work in the realm is the Sketch-RNN work by (Ha & Eck, 2017). There are several interesting aspect about this work. Firstly, it represents the sketches by strokes and the strokes by sampling points on the curve instead of pixel images. This vector representation versus raster representation for sketches is an interesting decision in terms of how to best interpret the sketches. Since Sketch-RNN learns the distribution  $\mathbb{P}$ , it can take in the points from strokes done by users, and then predict the rest of the sketches. This distribution is learnt from the massive dataset Quick!Draw collected from a game hosted by Google. In comparison to Quick!Draw dataset, although our dataset is also based on sketches from Quick!Draw,

In terms of exploring the multimodal sketch generation realm (text-to-sketch synthesis), a recent work is SketchBird (Yuan et al., 2021). This work, similar to ours, deal with the unique challenge of generating sketches from textual descriptions. They setup the task to mimic or as a counterpart to the classic text to image generation on the CUB dataset (Wah et al., 2011). This work is also representative of a line of work that is based on GAN, unique in the way that it is outputting sketches, closer to the domain that we are interested in. The line of text-to-image synthesis work begins with conditional GAN (Reed et al., 2016), which also reports results on the CUB dataset. But what is slightly in lack for the dataset that SketchBirds collected But to examine the line of GAN work, we can see that AttnGAN (T. Xu et al., 2017) (what SketchBirds is based upon or ) One thing we are especially interested in is how these models are able to extract the text features, and how they fuse text features with image features. Moreover what loss is used to encourage the alignment between the image and text domain. In SketchBirds, a bidirectional long short-term memory (Bi-LSTM) network is used as the text encoder. Inspired by AttnGAN, to extract text vectors that are visually aware, SketchBirds trains the text encoder with image-text pairs while minimizing the Deep Attentional Multimodal Similarity Model (DAMSM) loss, proposed in AttnGAN. This loss is calculated based on attention-driven text-image matching score, where matching is between two vectors, one is the vector representing a word in the sentence, and the other is a weighted sum of vectors of image regions, where the weight comes from a matrix of size  $T \times 289$  ( $T$  being the number words in the sentence, and 289 being the number of image regions), calculated using dot-product similarity between word in the sentence and sub-region in the image. It seems like from quite a few papers, such as G. Xu et al. (2021), fuse the visual and textual space by combining the visual



features using weights calculated by dot-similarity between the two modality, or vice versa to achieve cross attention. G. Xu et al. (2021) uses a LSTM+GloVE setup for the unimodal text embeddings.

The SketchCUB dataset collected by SketchBird contains sketches that are more similar to still-life portrait sketches and are very realistic, but sketches in the Quick,Draw! dataset are more similar to icons. This is due to how SketchCUB is transferred from RGB images in the CUB dataset by using open-source holistically-nested network (HED). The SketchCUB dataset contains 200 bird categories with 10,843 images. It includes a training set with 8,326 images in 150 categories and a test set with 2,517 images in the remaining 50 categories.

What are some other ways that we can extract visually informed text embeddings.

StyleGAN-NADA: CLIP Guided Domain Adaptation of Image Generators

Of course, there are other techniques to generate images from texts, namely, leverage large pretrained model such as GPT-3. GPT-3 and DALL-E are particular nowadays for researchers to replicate on their own and try to query the immense feature space for creative art pieces. However, the abstract art style work is not our focus, and while creativity is an interesting future direction, we emphasize the collaborative aspect more than creativity.

Another recent work done with GAN is DoodlerGAN.

In the larger realm of RGB images: Therefore, our dataset will be a good benchmark for how well these models work at capturing the individual semantic components of an object. The reason that we claim this is that some work on GAN's have try to look at how to manipulate certain regions in the images by manipulating the latent space. While this line of work also try to look at how .This area of the work is around facial feature editing. Work such as Semantic Photo Manipulation with a Generative Image Prior (Bau et al., 2019), has an interactive interface where the user can use stroke to indicate where in the image they would want a certain object, and the GAN will generate the objects in that location. "semantic image editing tasks, including synthesizing new objects consistent with background, removing unwanted objects, and changing the appearance of an object". semantic edit on an object. They would apply a semantic vector space operation in the latent space. How our work is different from this work is that: how well the methods can work on sketches and how well can the edits can done through language. [?] Moreover, it seems like we need to have an image already in order to do the manipulation, but for our ideal tasks, we start from a blank canvas.

## Chapter 6

# Conclusion

We learned so much from this project.

**Appendix A**

**True Facts**

# Bibliography

- Bau, D., Strobel, H., Peebles, W., Wulff, J., Zhou, B., Zhu, J., & Torralba, A. (2019). Semantic photo manipulation with a generative image prior. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH)*, 38(4).
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... Houlsby, N. (2020). *An image is worth 16x16 words: Transformers for image recognition at scale*. arXiv. Retrieved from <https://arxiv.org/abs/2010.11929> doi: 10.48550/ARXIV.2010.11929
- Eitz, M., Hays, J., & Alexa, M. (2012). How do humans sketch objects? *ACM Trans. Graph. (Proc. SIGGRAPH)*, 31(4), 44:1–44:10.
- Ha, D., & Eck, D. (2017). *A neural representation of sketch drawings*.
- Li, K., Pang, K., Song, J., Song, Y.-Z., Xiang, T., Hospedales, T. M., & Zhang, H. (2018). *Universal perceptual grouping*. arXiv. Retrieved from <https://arxiv.org/abs/1808.02312> doi: 10.48550/ARXIV.1808.02312
- Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016). *Generative adversarial text to image synthesis*. arXiv. Retrieved from <https://arxiv.org/abs/1605.05396> doi: 10.48550/ARXIV.1605.05396
- Wah, C., Branson, S., Welinder, P., Perona, P., & Belongie, S. (2011). *The Caltech-UCSD Birds-200-2011 Dataset* (Tech. Rep. No. CNS-TR-2011-001). California Institute of Technology.
- Xu, G., Kordjamshidi, P., & Chai, J. Y. (2021). *Zero-shot compositional concept learning*. arXiv. Retrieved from <https://arxiv.org/abs/2107.05176> doi: 10.48550/ARXIV.2107.05176
- Xu, P., Hospedales, T. M., Yin, Q., Song, Y.-Z., Xiang, T., & Wang, L. (2020). *Deep learning for free-hand sketch: A survey*. arXiv. Retrieved from <https://arxiv.org/abs/2001.02600> doi: 10.48550/ARXIV.2001.02600
- Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., & He, X. (2017). *AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks*. arXiv. Retrieved from <https://arxiv.org/abs/1711.10485> doi: 10.48550/ARXIV.1711.10485

- Yuan, S., Dai, A., Yan, Z., Guo, Z., Liu, R., & Chen, M. (2021). Sketchbird: Learning to generate bird sketches from text. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)* (p. 2443-2452). doi: 10.1109/ICCVW54120.2021.00277