IMPROVING TEXT-TO-IMAGE DIFFUSION GENERATION
VIA LARGE LANGUAGE MODEL

Yifei Li

A THESIS

in

Data Science

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Master of Science in Engineering

2023

Co-Supervisor of Thesis

Chris Callison-Burch

Chris Callison-Burch, Associate Professor of Computer and Information Science

Co-Supervisor of Thesis

Mark Yatskar, Assistant Professor of Computer and Information Science

Graduate Group Chairperson

Susan Davidson

Susan B. Davidson, Weiss Professor of Computer and Information Science

IMPROVING TEXT-TO-IMAGE DIFFUSION GENERATION
VIA LARGE LANGUAGE MODELS

ACKNOWLEDGEMENT

I would like to express my heartfelt gratitude to the individuals who have played significant roles in the completion of my thesis and offered support throughout my master's studies at the University of Pennsylvania.

I am immensely grateful to Prof. Mark Yatskar, my thesis co-advisor, for his invaluable research guidance in the realm of multimodal and NLP research. His suggestions contribute to formulating the problem and evaluations in this thesis. I am also thankful for his (advanced) computational linguistics course, which provided me with a solid foundation in NLP knowledge. I am excited to begin my journey as his incoming Ph.D. student. I would like to extend my heartfelt appreciation to Prof. Chris Callison-Burch, my thesis co-advisor, for his invaluable guidance in NLP and suggestions to improve the content of this thesis. His practicum course facilitates collaboration with fellow researchers and laid a strong foundation for this thesis. Furthermore, his course on Interactive Fiction and Text Generation enlightened me on NLP story generation, procedures, and events.

I owe a debt of gratitude to Prof. Lyle Ungar, whose machine learning and deep learning courses sparked my interest in AI and propelled me to pursue a Ph.D. I am deeply appreciative of his guidance during my first research project on conceptors. I would also like to acknowledge Prof. João Sedoc, who also co-advised my first research project on conceptors. His guidance and expertise were invaluable throughout the project. I extend my sincere thanks to Prof. Lara Martin for providing me with valuable advice regarding an academic career; and Prof. Dan Roth for introducing me to the realm of machine reasoning.

I would like to express my gratitude to Yue Zhang and Fangyi Teng, who selflessly served as human raters for this thesis without any compensation. This work cannot be completed without them! Last but not least, I am eternally grateful to my parents for their unwavering support both in spirit and financially, enabling me to pursue my studies in the United States and granting me the freedom to explore my academic interests.

ABSTRACT

IMPROVING TEXT-TO-IMAGE DIFFUSION GENERATION
VIA LARGE LANGUAGE MODELS

Yifei Li

Chris Callison-Burch & Mark Yatskar

Generating accurate and consistent images based on unusual and counterintuitive natural language prompts remains challenging for image-generating models known as diffusion models (DMs; e.g. Stable Diffusion). In this thesis, we explore the potential of language models (LMs; e.g. GPT-3/4) to enhance the performance of DMs in handling such challenging descriptions. We construct challenging datasets to disclose such weaknesses of DMs, then propose two methods: imagine-then-verbalize and sketch-then-draw. The imagine-then-verbalize approach leverages the imaginative abilities of LMs to provide additional details and contexts that enhance the persuasiveness of the descriptions. The sketch-then-draw method utilizes the coding capacity of LMs to generate SVG code, allowing for the creation of more numerically consistent sketches. The human evaluation confirms the poor performance of DMs on unusual or numerically difficult descriptions and highlights the potential of LMs to alleviate these issues. While in many scenarios, the LMs do offer some assistance, they overall suffer from a lack of stability and Generalizability. Moreover, the challenges are related to the out-of-distribution training data of DMs and persisting evaluation subjectivity, which is hard to tackle.

The code is released on: `https://github.com/realliyifei/llm-improve-diffusion`.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1 : Introduction

Can an image-generating model accurately depict the concept of "A bird swimming like a fish"? Text-to-image diffusion models are widely used today, but they still face the challenge of consistently and accurately understanding natural language prompts, particularly when these prompts contain unusual descriptions that deviate from the training data (i.e., out-of-distribution; OOD). In order to evaluate and improve the performance of these diffusion models (DMs), we have devised a set of counterintuitive, anti-commonsense, and numerically difficult descriptions. Our goal is to enhance the accuracy and consistency of image generation by leveraging the imagining and coding abilities of language models (LMs). We refer to two methods to achieve this goal, as described below.

The first method, called **imagine-then-verbalize** (Chapter 3), utilizes the prompting capabilities of LMs to imagine additional details and contexts that make the descriptions more convincing without altering their original meaning. This can be achieved through either template-based prompting or chain-of-thought (CoT) prompting (Wei et al., 2022).

The second method, called **sketch-then-draw** (Chapter 4), harnesses the coding abilities of LMs to generate SVG code that creates a sketch consistent with the number of objects described in the prompt. This sketch, along with the prompt, is then fed to the DMs to generate an image that maintains numerical consistency.

The remaining chapters are structured as follows: Chapter 2 provides a comprehensive literature review on DMs, LMs, prompting methods, and LM-grounded generation. Chapter 5 presents comprehensive conclusions and discusses future work, while Chapter 6 addresses the limitations of this research.

The research questions to be explored in this thesis are:

1. How effectively can diffusion-based text-to-image models depict unusual, counterintuitive, and anti-commonsense descriptions, such as "A bird swimming like a fish"?

What are the limitations of such DMs in accurately understanding and generating such descriptions, particularly when they deviate from the training data?

2. How can LMs contribute to improving the accuracy and consistency of image generation when faced with OOD descriptions? What are the advantages and limitations of utilizing LMs in this context?

3. What is the potential of the imagine-then-verbalize method, which involves leveraging the prompting abilities of LMs to imagine additional details and contexts, in enhancing the persuasiveness and fidelity of image generation? How does this method differ when employing template-based prompting versus CoT prompting?

4. How can the sketch-then-draw approach harness the coding abilities of LMs to generate more numerically consistent sketches and improve image generation? What are the trade-offs in terms of flexibility and semantic interpretation associated with this method?

5. How do human evaluations validate the shortcomings of DMs in generating images based on unusual descriptions? How do these evaluations highlight the potential of LMs to address these issues?

CHAPTER 2 : Liteature Review

2.1. Text-to-Image Diffusion Generation

Inspired by non-equilibrium thermodynamics, DMs employ a parameterized Markov chain composed of incremental diffusion steps to introduce random noise into data gradually (Sohl-Dickstein et al., 2015). By learning to reverse this diffusion process, they can generate desired data samples from the Gaussian noise.

Denoising diffusion probabilistic models (DDPM; Ho et al., 2020) is the first work to show that DMs are capable of generating high-quality images, sometimes outperforming other types of generative models. Similarly, the denoising diffusion implicit model (DDIM; Song et al., 2020) adopts a similar noise distribution pipeline but via a class of non-Markovian diffusion processes that introduce deterministic generation, leading to higher-quality generation in a shorter running time. Later, the latent DM (LDM; Rombach et al., 2022) implements the diffusion generation in the latent space of pretrained autoencoders rather than the pixel space, decreasing the training cost, improving the inference speed, and promoting visual fidelity.

The DMs have a great connection to score-based generative models. For example, Song and Ermon (2019) proposes a noise-conditional score network (NCSN) that can perturb input data with a series of growing Gaussian noise and approximate the score functions for all noisy data distributions by training a neural network conditioned on noise levels. These increasing noise levels are similar to the ones in the forward process of diffusion generation.

There are different families of DMs. Dall-E 2 (Ramesh et al., 2022) uses the joint embedding space of CLIP (Radford et al., 2021) to train a DM and then trains another separate decoder to create images with the help of the CLIP image embeddings. Imagen (Saharia et al., 2022) utilizes a frozen pre-trained large language model (LLM) called T5 (Raffel et al., 2020) as the text encoder to improve linguistic understanding of the prompts and thus generate better images. Stable Diffusion (Rombach et al., 2022), similar to Imagen, uses another frozen text encoder from CLIP for the condition of text prompts.

However, while outperforming their precursors greatly in the text-to-image tasks, these diffusion generation models suffer from the misunderstanding of composition, negation, commonsense, numerical reasoning, and from the manifestation of hallucination and concept leakage (Marcus et al., 2022; Rassin et al., 2022). Some works are thus introduced to have more control of such generation. Liu et al. (2022) introduces the composable DMs that treat each DM as an implicitly parameterized energy-based model. This method trains a bunch of DMs of the same type for the corresponding concept and then uses logical operators like `AND` (for conjunction) and `NOT` (for negation) to compose different concepts during DM generation in a logical way. Zhang and Agrawala (2023) proposes a method called ControlNet that clones the weights of DMs as a trainable copy for the conditional control and a locked copy for preserving the model's original capability, then both copies are connected via a zero-initialized convolution layer. This method enables the users to precisely condition the input by providing a sketch in various ways, such as Canny edges, Hough lines, user scribbles, human key points, and so on.

2.2. Language Models

LLMs are neural network models that are trained on massive amounts of text data, which allows them to learn more about the patterns and relationships within the language. In addition, the huge parameterized architecture also contributes to the scale of the LLM, which gives the model more capacity to learn and generalize from the data and enables the model to learn more features to understand the context and meaning of words. The objectives in the pre-training stage, such as masked language modeling (MLM) and next sentence prediction (NSP), allow the model to learn general language representations that can later be fine-tuned for a specific task.

From GPT-2 (Brown et al., 2020) to GPT-3 (Radford et al., 2019) to InstructGPT (Ouyang et al., 2022), GPT is a state-of-the-art (SOTA) LM family with remarkable billion-level parameters and involves colossal corpora for pretraining. It has demonstrated impressive capabilities, such as zero-shot learning and strong performance across multiple NLP tasks, such as translation, summarization, and, question-answering, and text auto-completion. Notably, GPT-3 demonstrates zero-shot learning capabilities, meaning it can generalize to new tasks without additional training. Despite its impressive performance, GPT-3 may suffer from hallucination problems and can be sensitive to input phrasing. The release of ChatGPT [1] and GPT-4 [2] push the GPT performance on language tasks and beyond further, due to the benefits from the alignment tricks (Ouyang et al., 2022), additional training on code (Chen et al., 2021), and the multimodal ability.

Besides GPT which uses decoder-only architecture, there are LLMs built upon other architectures that are also popular: BERT (Devlin et al., 2019) uses encoder-only architecture and its bidirectional transformer can take into account the context before and after a given word in a sentence, which allows it to understand the meaning of words in the context of the entire sentence. T5 (Raffel et al., 2020) adopts encoder-decoder transformer implementation, unifies every text processing problem under the text-to-text paradigm, and achieves impressive performance by the insights from its exploration with scale and the benefits from their proposed colossal corpus called C4.

All of these models are built on Transformer (Vaswani et al., 2017) architecture, which is the first one to solve sequence-to-sequence tasks entirely on self-attention to compute representations of its input and output without using sequence-aligned techniques and thus spark parallelization significantly. The core component, self-attention (Cheng et al., 2016), can process a sequence by substituting each element with a weighted mean calculated from the remaining elements in the sequence.

2.3. Prompting

As stated in the survey from Liu et al. (2023), given a textual input X to LM with parameter θ, the prompting method can model the probability $P(X;\theta)$ of X itself and then predict the output Y. In this way, it can avoid the necessity of fine-tuning or continued training

[1] ChatGPT Blog: `https://openai.com/blog/chatgpt`
[2] GPT-4 Blog: `https://openai.com/research/gpt-4`

of the LM. A simple way is to provide a prompt template in a format of text completion or generation to get the desired Y. Further, the generated Y can be concatenated by the original X' to become the new X in the next prompting and in this way, we can prompt the LM recursively.

Chain-of-thought (CoT) prompting (Wei et al., 2022) is to provide a few manually-written instructions and let LM generate a reasoning chain C in the step of inferring X and aim to get a better Y. A CoT with n samples can be phrased as the probability

$$P(X_1, C_1, Y_1; X_2, C_2, Y_2; \cdots ; X_n, C_n, Y_n; X)$$

Their experiments show that CoT can boost the LM performance in tremendous benchmarks, especially the numerical reasoning ones.

There are different paradigms of CoT: scratchpad-based prompting can enable the LM to reason step-by-step (Kojima et al., 2022) and even inject symbolic representation in the middle (Nye et al., 2021) to enhance robustness and interpretability. Ensemble-based prompting involves combining several prompting results, which can then be either selected as the best one or used to bootstrap the inference process, improving consistency (Wang et al., 2022b) and enabling multi-hop reasoning (Li et al., 2022).

2.4. LM-Grounded Generation

It is not rare to use LM for the imagination to ground the next step of the generation. In this way we can leverage the prior knowledge encoded in the LM for downstream tasks.

Chakrabarty et al. (2023) proposes a two-stage pipeline that is similar to this work: they first use the CoT prompting with GPT-3 to get an extended textual description of a metaphor statement to explain the metaphor, then feed such description to the diffusion-based text-to-image models to visualize the implicit meaning of the linguistic metaphors. This model-in-the-loop approach is confirmed to be beneficial to the image generation of such implicit meanings, and CoT prompting leads to better images compared to the classic completion prompting.

Wang et al. (2022a) introduce the imagine-and-verbalize paradigm to inject commonsense and achieve plausible generation: to let a trained LM imagine the relational scene knowledge graph (SKG) based on the input concepts and another trained LM verbalize such graph in natural language. By combining the SKG with LMs, this method is able to compose new objects creatively and identify implicit meanings for a scene, while obeying the constraints of commonsense.

CHAPTER 3 : Imagine-Then-Verbalize

3.1. Introduction

In this chapter, the imagine-then-verbalize method employed aims to leverage the imagining capability of LM to enhance the generation of descriptions by prompting the models in two different ways. The primary objective is to construct more convincing and contextually rich descriptions that align with implicit and anti-commonsense cues.

The approach involves prompting the LM to imagine and generate additional details and contextual information based on the given description. By doing so, the LM is encouraged to fill in the gaps and provide a more comprehensive and vivid depiction of the intended scene (imagine). This imaginative prompting technique harnesses the LM's ability to generate creative and plausible content, which can contribute to the generation of more coherent and compelling descriptions (verbalize).

The imagine-then-verbalize method not only enhances the completeness and vividness of the descriptions but also addresses the challenge of generating contextually appropriate content. By leveraging the LM's imaginative capabilities, the approach enables the model to go beyond literal interpretations and generate descriptions that capture the implicit meaning and nuances present in the given input.

Moreover, the imagine-then-verbalize method can be particularly useful in domains where explicit instructions or descriptions might lack crucial details. For example, in visual scene understanding tasks, where images are described, the method can aid in generating more detailed and accurate descriptions by inferring missing information through imagination.

It is important to note that while the imagine-then-verbalize method leverages the LM's imaginative powers, it still operates within the boundaries of the training data and the knowledge it has been exposed to. The generated descriptions are based on patterns and associations learned from the training corpus and might not always align perfectly with human-like understanding. However, by incorporating imagination into the generation process, the method contributes to the production of more compelling and contextually coherent descriptions, improving the overall performance and user experience of language models.

3.2. Dataset and Model

3.2.1. Dataset

The dataset consists of 50 unconventional descriptions, which exhibit either counterintuitive or anti-commonsense characteristics while remaining within the realm of possibility for human interpretation through visual or fictional representations. Prior to inclusion, we eliminated descriptions generated during the initial brainstorming phase that proved excessively vague for DM to understand. Examples of such ambiguous descriptions include "A book that reads to you" or "A pencil writing on water." The complete set of these unusual descriptions can be found in Figure 1.

A bee humming a melody	A moon made of cheese
A bee with a propeller instead of wings	A mountain wearing a hat
A bird swimming like a fish	A mouse chasing a cat
A boat sailing on land	A painting that paints itself
A butterfly is pulling a train	A rock is floating on water
A butterfly with origami wings	A rose with thorns made of glass
A cactus growing in a rainforest	A shadow with a physical presence
A candle burning underwater	A skyscraper lying horizontally
A car driving on clouds	A snail racing a cheetah
A city skyline made of candy	A snake with legs
A cloud shaped like a perfect cube	A squirrel collecting coins
A cloud that rains candy	A squirrel collecting nuts with a shopping cart
A desert covered in snow	A sunflower facing away from the sun
A dolphin walking on land	A teapot whistling a tune
A door that leads nowhere	A toaster toasting ice cubes
A dragon blowing bubbles	A tornado made of leaves
A fire that's cold to the touch	A train traveling through the sky
A fish climbing a tree	A tree growing upside down
A fish riding a bicycle	A tree with leaves made of feathers
A frog sunbathing on a beach	A turtle flying like a bird
A hedgehog with balloons for quills	A volcano erupting with ice
A jellyfish floating in a park	A whale is jumping over a mountain
A lighthouse guiding cars	An elephant balancing on a tightrope
A lion being chased by a gazelle	An island floating in the sky
A monkey typing a novel	Mice are scaring a lion

Figure 1: Unusual descriptions that are counterintuitive or anti-commonsense

3.2.2. Models

For the DM, we use Huggingface's `stable-diffusion-v1-4` [1]; for the LM, we use OpenAI's GPT, `text-davinci-003` [2] with the hyperparameters: `max_tokens=1000`, `n=1`, `stop=None`, and `temperature=0.7`.

3.3. Prompting Methods

The prompting process can be approached in two distinct ways.

3.3.1. Template-Based Prompting

By providing a prompt template (Figure 2), where the unusual description is filled inside the curly braces, we can let LLM ground the visual generation step-by-step:

1. Explain the reason why it is counterintuitive
2. Imagine some details and contexts to make it more convincing
3. Summarize the aforementioned information as a one-sentence description of an image

Particularly, adding explanation is shown to benefit the LMs' reasoning and predicting ability (Wiegreffe et al., 2021; Wei et al., 2022; Lampinen et al., 2022), and step-by-step reasoning is shown to improve LMs' few-shot performance (Nye et al., 2021; Zelikman et al., 2022).

Why "{}" is counterintuitive? Explain. How to provide contexts to the image of "{}" to make it more convincing? List them in declarative sentences with explicit visual features. Summarize the above explanation and details into one long sentence, as a caption of a picture.

Figure 2: Prompt template of template-based prompting

Figure 3 demonstrates the process of template-based prompting for an unusual description– "A bird swimming like a fish". In the beginning, the prompt template urges the LM to provide the reasons why this description is counterintuitive. As a result, the LM correctly explains the key difference between fish and birds: the former is hydrophytic and adapted for

[1] `https://huggingface.co/CompVis/stable-diffusion-v1-4`
[2] `https://platform.openai.com/docs/models/overview`

swimming underwater, while the latter is adapted for flying in the sky. Further, this leads to different biological features. Then the prompt requests LM to imagine the extra features and contexts to make the description more convincing. Here the visual features are explicitly mentioned since these imagined features are later provided to DM to help the visualization. Without this constraint, it is easy for the LM to list some information that is hard to visualize and thus leads to the backfire–adding noise to the DM generation. In the end, considering the limited window size of the DM input, we summarize the aforementioned explanation and details into one sentence. It is stated in the prompt template that this sentence should be an image caption to make it more suitable for the visual representation generated by the DM.

> Why ”A bird swimming like a fish” is counterintuitive? Explain.
”A bird swimming like a fish” is counterintuitive because birds and fish are two distinct animals that have evolved differently. Fish are adapted for swimming underwater, while birds are adapted for flying in the air. They have different body shapes and types of movements that enable them to function in their respective environments. Therefore, it is counterintuitive to think of a bird swimming like a fish because it does not fit with the natural evolution of either species.

> How to provide contexts to the image of ”A bird swimming like a fish” to make it more convincing? List them in declarative sentences with explicit visual features.
1. The bird has the body shape and movements as it swims like a fish, such as long, undulating strokes.
2. The bird adapts feathers for underwater swimming, such as webbed feet or a flattened tail.
3. The bird adapts beak for underwater swimming, such as a hooked beak for catching small fish.
4. The bird adapts wings for underwater swimming, such as small, paddle-like wings.

> Summarize the above explanation and details into one long sentence, as a caption of a picture.
A bird gracefully swimming in a lake, surrounded by aquatic plants and animals, with adapted feathers, beak, and wings for swimming like a fish.

Figure 3: Template-based prompting example: A bird swimming like a fish

3.3.2. Chain-of-Thought Prompting

Wei et al. (2022) indicate that providing exemplars can significantly improve the model performance. In this work, we provide four samples as instructions for CoT prompting and urge LM to infer the given description following a similar pattern. In the instructions, the four unusual descriptions are not included in the dataset to avoid description leakage. Besides the explanation and augmented details and contexts introduced in the template-based prompting, this prompting method also requires the LM to think of the unwanted concepts via the "Forbidden Words" feature and then avoid such concepts in the summarization. Therefore, to generate a final description, the LM should pass through five steps: "Explain", "Convincing Details", "Visual Features", "Forbidden Words", and "One-Sentence Description". All of the contents are manually phrased by humans to inject expert oversight. The whole template is demonstrated in Figure 4 and an output sample is in Figure 5.

3.4. Output Demonstration

In order to closely examine the final image output of the raw description versus that of the prompted counterpart, we present both in a single figure (from Figure 6 to Figure 22)[3]. The four images on the left display the output using the raw description shown at the top, while the four images on the right depict the augmented description through prompting. By comparing the left and right sides, we can easily discern the change in the consistency of image generation before and after the LMs' grounding.

3.4.1. Template-Based Prompting

There are different types of successful improvement.

The first type involves the explicit explanation of features. In the case of "A bird swimming like a fish" (Figure 6), none of the images from the raw description visualize the desired meaning. Instead, they depict a bird that is either normally swaying on the water's surface or flapping its wings above the water. After being prompted by the template-based method, the new description contains more specific details such as "adapted feathers, beak, and wings for swimming", resulting in at least one image (bottom-left) that successfully depicts the desired object–a fish-like swimming bird.

[3]For brevity, only a small subset is listed here. For all outputs, please refer to `https://github.com/realliyifei/llm-improve-diffusion`

Given an unusual description that is anti-commonsense or counterintuitive,
your task is to first give the explanation that why it is unusual,
then add more visual details and contexts to make the image of this description more convincing.
In the last, summarize the above explanation and details into one long sentence, as a caption of a picture.

1. Description: A fish is walking on land
Explain: This image is unusual because fish do not have the ability to walk on land and typically live in water.
Convincing Details: The fish adapts its fins as legs that allow it to explore its surroundings. The fish walks in a city, which is more distinguishable than the area it usually lives in, i.e., water.
Visual Features: The fish is brightly colored and wears a small backpack, indicating it is on an adventure. The fin-adapted legs are flexible, allowing it to walk like a human.
Forbidden words: Since the fish usually live in water, to make the image less usual and more convincing, all water-related words should be avoided.
One-sentence description without forbidden words: A vibrant fish, equipped with fin-adapted legs and a backpack, walks in a beautiful city.

2. Description: A suit of armor is dancing ballet
Explain: This image is unusual because a suit of armor is typically heavy, inanimate, and not associated with the grace and elegance of ballet dancing.
Convincing Details: The suit of armor has been magically enchanted or mechanically modified to move with fluidity and precision, allowing it to perform ballet steps with ease. The dance takes place in a grand ballroom filled with surprised onlookers.
Visual Features: The suit of armor is polished to a high shine, reflecting the chandeliers above. The location is a ballroom specifically designed for ballet.
Forbidden words: Any words related to typical uses of a suit of armor, such as battles, knights, and war, should be avoided to make the image less usual and more convincing.
One-sentence description without forbidden words: In a grand ballroom, a gleaming suit of armor gracefully performs ballet.

3. Description: A cactus is ice-skating
Explain: This image is unusual because cacti are immobile plants and cannot participate in activities like ice-skating.
Convincing Details: The cactus is brought to life by a magical spell, allowing it to participate in winter activities.
Visual Features: The cactus wears a warm scarf and a pair of ice skates, gliding gracefully on a frozen pond with snowflakes gently falling in the background.
Forbidden words: Any words related to the desert are unnecessary since they are highly related to the cactus habitat but also contradict the ice.
One-sentence description: Under a magical spell, a cactus adorned with a cozy scarf and ice skates comes to life and skillfully glides across a picturesque frozen pond.

4. Description: A mountain is building a bridge
Explain: This image is unusual because a mountain does not possess motivation, tools, or a body to build a bridge.
Convincing Details: The mountain has a human-like face, suggesting that it is alive and driven by purpose. Equipped with hands, it wields a hammer to construct the bridge.
Visual Features: The mountain, with its human-like face and hands, is situated among other mountains. It uses a hammer to build a bridge that connects two mountains.
Forbidden words: Any words related to typical mountain characteristics or features, such as peaks, valleys, or rocks, should be avoided to make the image less usual and more convincing.
One-sentence description: In a unique landscape, a mountain with a human-like face and hands skillfully employs a hammer to create a bridge, uniting two distant points.

5. Description: {}

Figure 4: Instruction for Chain-of-Thought Prompting

Description: A bird swimming like a fish
Explain: This image is unusual because birds typically fly in the air and fish swim in water.
Convincing Details: The bird is able to use its wings to propel itself in the water, allowing it to swim like a fish. It trails behind a group of fish, which it follows and mimics their movements.
Visual Features: The bird is a bright, vibrant color, in contrast to the dull-colored fish. Its wings and feathers are slightly modified, allowing it to move gracefully through the water.
Forbidden words: Any words related to the bird's typical activities, such as flying, flying in the air, or soaring, should be avoided to make the image less usual and more convincing.
One-sentence description: A vibrant bird with slightly modified wings and feathers gracefully follows a group of fish, swimming with them with ease.

Figure 5: Example of CoT prompting for "A bird swimming like a fish"

The second type involves adding additional objects to enhance the overall plausibility of the image. In the case of "A cactus growing in the rainforest" (Figure 7), apart from the cactus depicted on both sides, the prompted side contains more tropical plants to convince the audience that it is indeed in a rainforest setting. Similarly, in the case of "A frog sunbathing on a beach" (Figure 8), although both sides show a frog lying on the sand, the prompted side adds sunbathing tools such as sunglasses and a rug to enhance the plausibility of the image.

Some of these improvements are subject to controversy. In the case of "A fire that's cold to the touch" (Figure 10), the LM specifies that the color of the fire should be blue to indicate coldness, but individuals with different commonsense may argue that blue fire is actually hotter. In the case of "A dolphin walking on land" (Figure 11), the prompted output successfully shows the key component–the land, but whether the dolphin is truly walking on the land is debatable. In the case of "A boat sailing on land", the prompted output contains an image (top-left) that is plausible but not obvious enough that a boat is on land.

Overall, there are three types of failures in text-to-image generation. The first type is the absence of a key component. For instance, in the aforementioned case of "A dolphin walking on land" (Figure 11), the raw output fails to generate the land component; In the case of “A bee humming a melody” (Figure 13), both the raw and prompted output do not convincingly depict the melody. The second type of failure is the absence of a relationship. For example, in the case of “A boat sailing on land” (Figure 12), both the raw and prompted outputs generate the boat and land but fail to convey the semantic meaning of “sailing”. The last type of failure is known as concept leakage, which is not uncommon in text-to-image generation models that blend two unrelated concepts together as one (Marcus et al., 2022). In the case of “A snail racing a cheetah” (Figure 14), the DM erroneously creates a new, peculiar creature that combines features of both a snail and a cheetah simultaneously. Addressing this issue by prompting the input proves to be challenging.

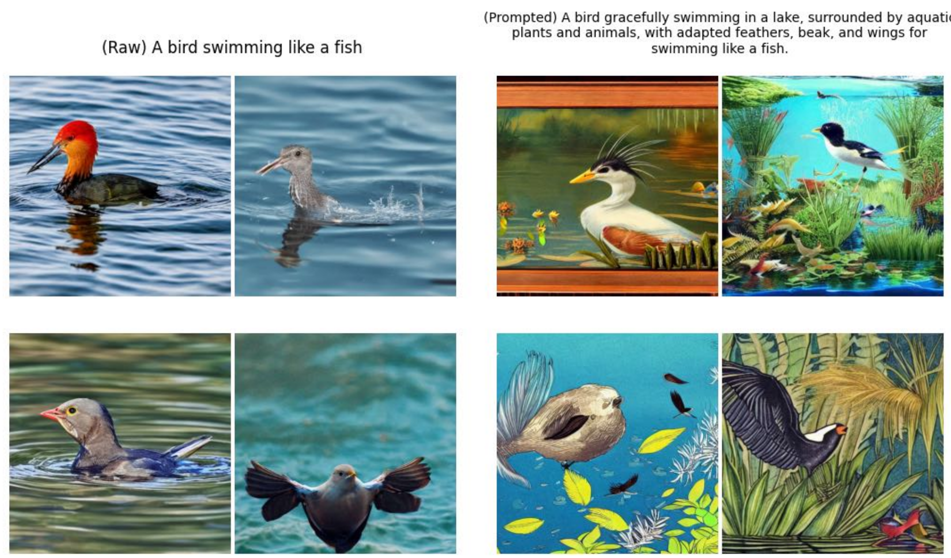


Figure 6: Successful improvement via template-based prompting: A bird swimming like a fish

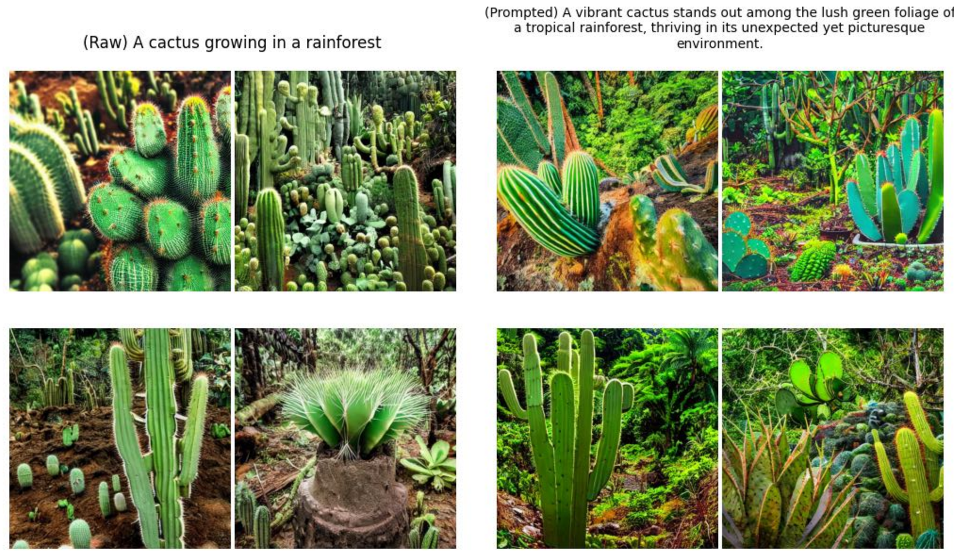


Figure 7: Successful improvement via template-based prompting: A cactus growing in a rainforest

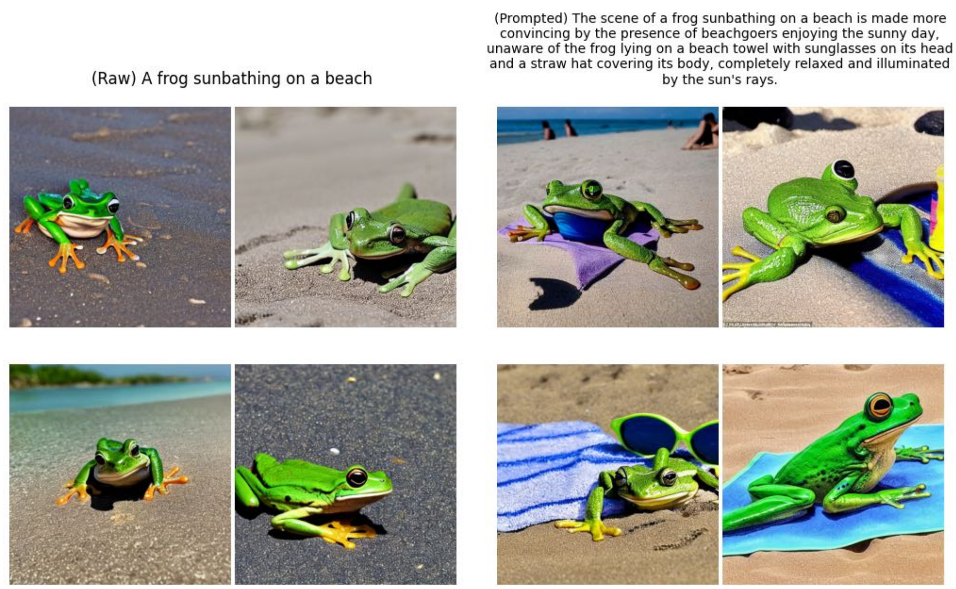


Figure 8: Successful improvement via template-based prompting: A frog sunbathing on a beach

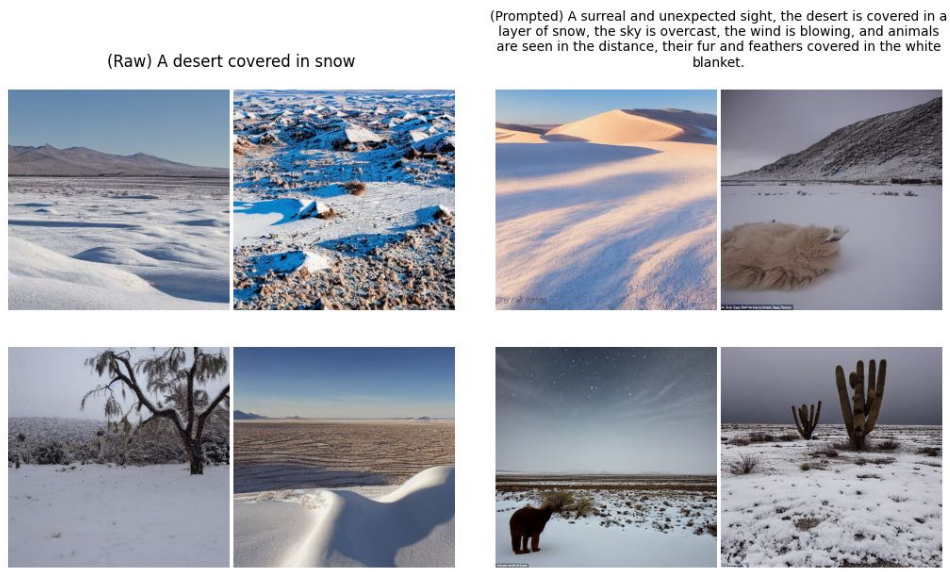


Figure 9: Controversial improvement via template-based prompting: A desert covered in snow

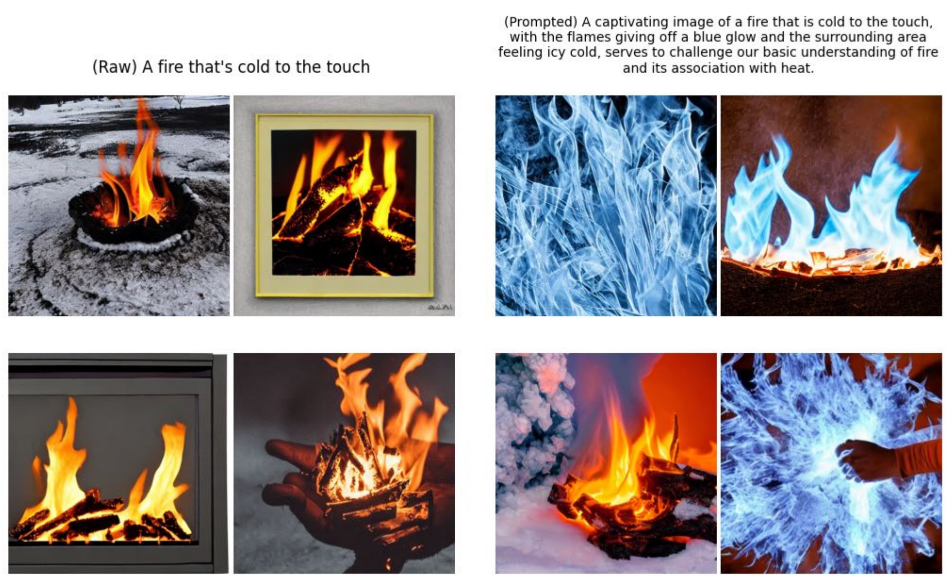


Figure 10: Controversial improvement via template-based prompting: A fire that's cold to the touch

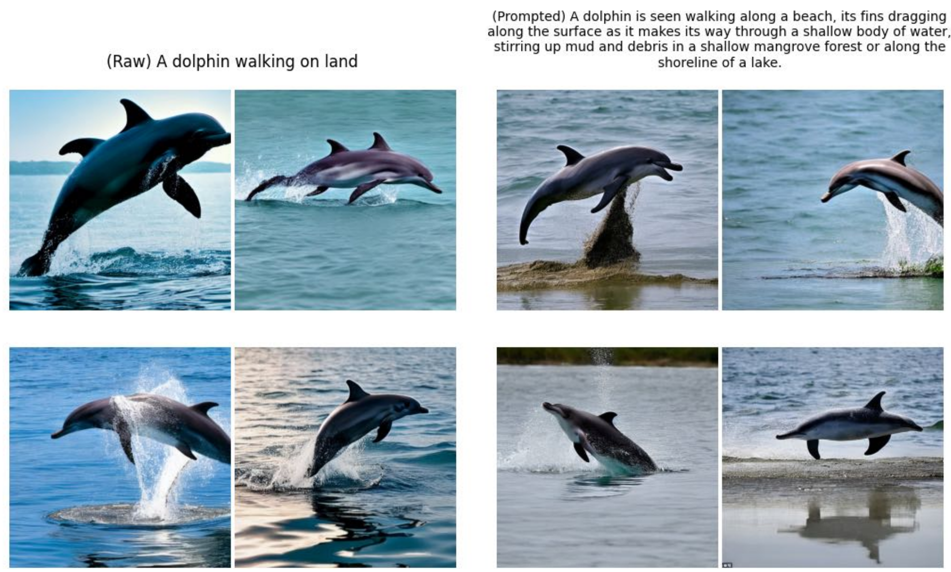


Figure 11: Controversial improvement via template-based prompting: A dolphin walking on land

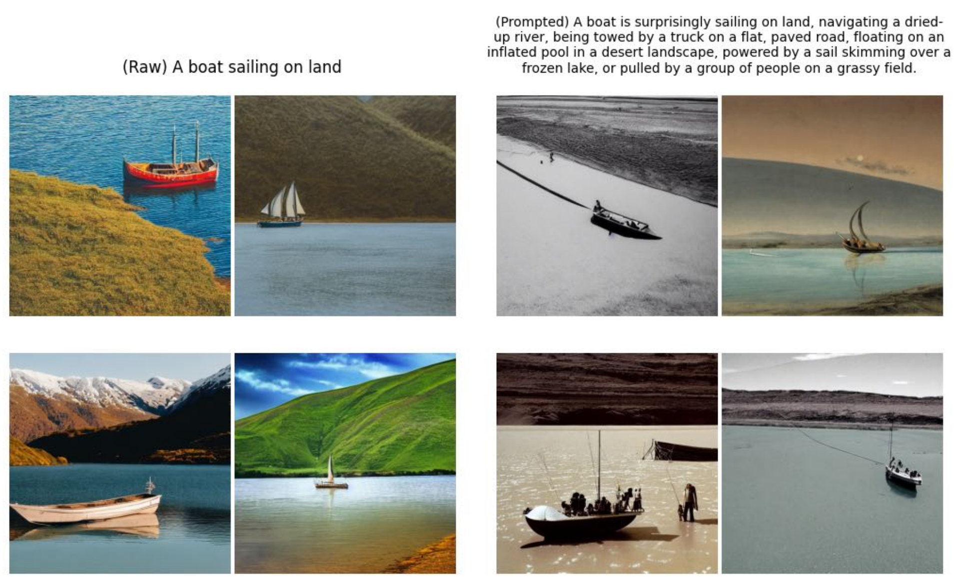


Figure 12: Failed improvement via template-based prompting: A boat sailing on land

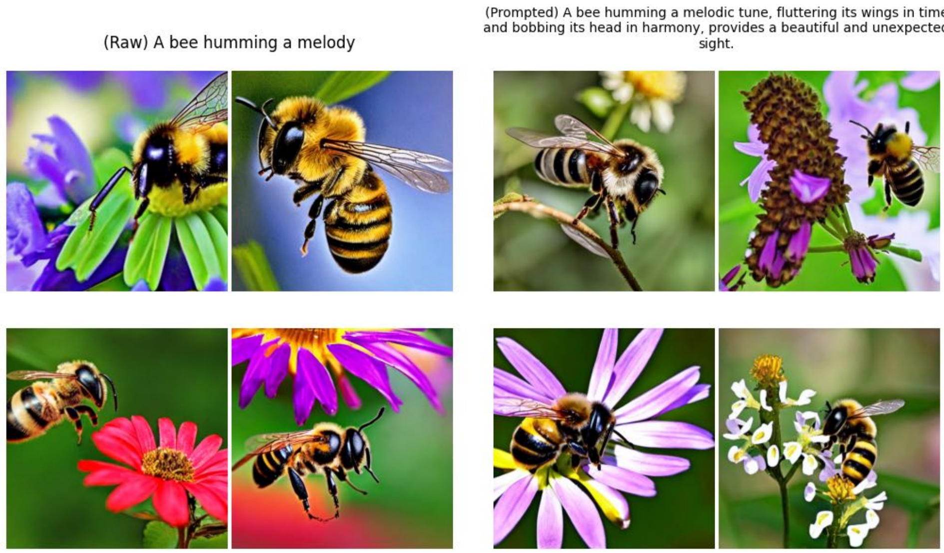


Figure 13: Failed improvement via template-based prompting: A bee humming a melody

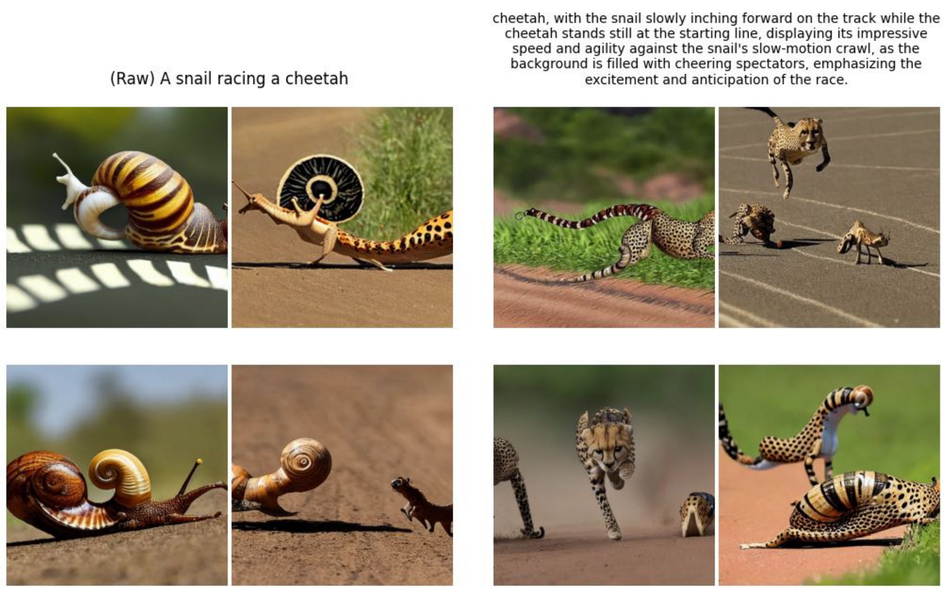


Figure 14: Failed improvement via template-based prompting: A snail racing a cheetah

3.4.2. Chain-of-Thought Prompting

Due to the addition of expert instruction, the method aided by CoT prompting is expected to perform better than the template-based prompting method, which is stated by Chakrabarty et al. (2023).

For instance, consider the example of “A dolphin was walking on land” (Figure 15). Compared to its counterpart in the previous method (Figure 11), the CoT-prompted output incorporates a city background by introducing the keyword ”bustling city” in the textual input, discouraging the DM from associating any water-related concepts. By mentioning different concepts that are unusual in a more detailed way, the LM can nudge the DM into generating the unusual image better. Similarly, in the case of “A boat sailing on land” (Figure 16), the DM explicitly mentions the word “desert”, trying to shift the “boat” concept away from the “water” concept and closer to the “landing” concept, resulting in improved image generation compared to Figure 12.

Furthermore, the LM with CoT has the potential to better visualize the implicit meanings. In the case of “A bee humming a melody” (Figure 17), the “melody” concept in the raw description is expanded to include a “musical note” in the prompted description, and the latter is easier to depict in the image (top-right image). In comparison, the template-based output struggles to achieve the same level of imaginative interpretation (Figure 13). Additionally, in the case of “A city skyline made of candy” (Figure 19), the LM extends the “city skyline” concept to encompass “buildings and streets” concept, resulting in a significant improvement in text-to-image generation.

Similar to the template-based prompting, the CoT method can further assists the LM in imagining additional objects. In the case of “A desert covered in snow” (Figure 18 and 9), the LM not only envisions a snow-covered desert but also includes desert-specific shrubs and cacti to enhance the credibility of the image.

However, like other LM methods, CoT prompting aid suffers from a lack of real-world understanding and the concept pollution. In the case of “A butterfly is pulling a train” (Figure 20), although the CoT-prompted output surpasses its template-based counterpart by successfully placing the butterfly and train in the same image, it fails to accurately depict the concept of ”pulling.” In the generated images, the butterfly merely flies around or touches the train, without exerting any effort to pull it. This observation highlights the challenges faced by the model in comprehending abstract and implicit physical relationships.

Significantly, such an LM-based prompting method can introduce concept pollution to the DM. In the case of “A volcano erupting with ice” (Figure 21), the prompted output mentions something like “ice cream” that is unwanted, thereby distorting the original semantic meaning and confusing the model. Such imagined concepts, although not necessarily contradictory, can be distracting. In the case of “A cloud shaped like a perfect cube” (Figure 22), the LM rephrases it as “[...] unique and vibrant color”, introducing unnecessary features to the DM’s understanding. The greater the disparity between unrelated concepts, the more challenging it becomes for the DM to handle image generation. Consequently, these distractions can significantly raise the chances of failure.

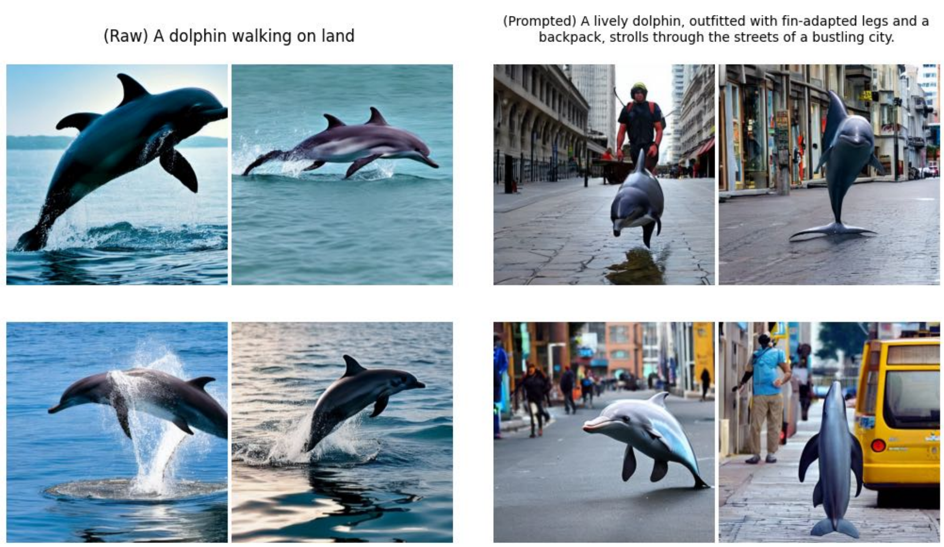


Figure 15: Successful improvement via CoT prompting: A dolphin walking on land

Figure 16: Successful improvement via CoT prompting: A boat sailing on land

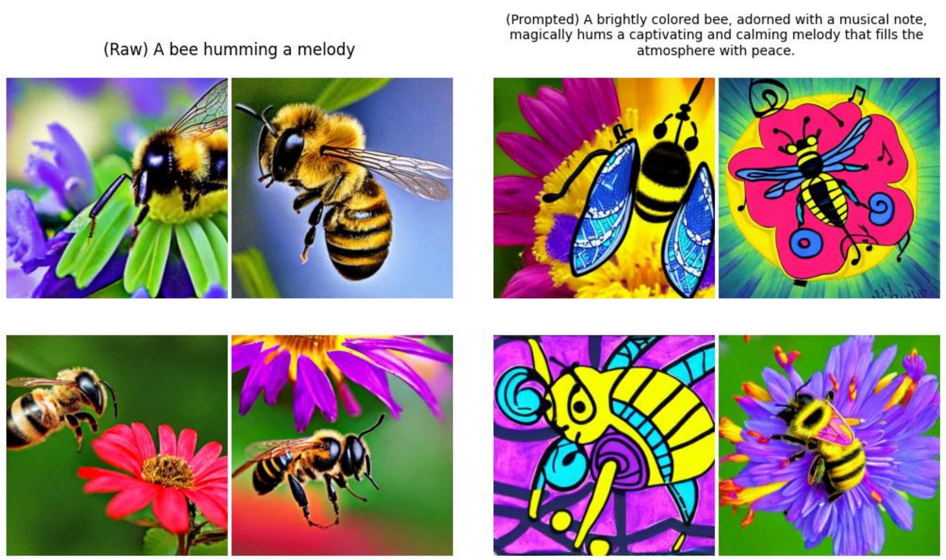


Figure 17: Successful improvement via CoT prompting: A bee humming a melody

Figure 18: Successful improvement via CoT prompting: A desert covered in snow

Figure 19: Successful improvement via CoT prompting: A City skyline made of candy

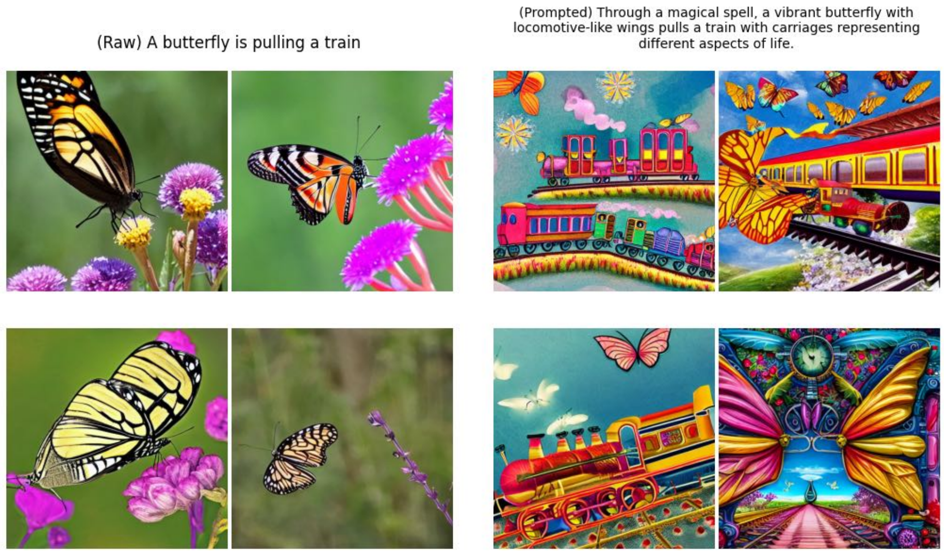


Figure 20: Controversial improvement via CoT prompting: A butterfly is pulling a train

Figure 21: Failed improvement via CoT prompting: A volcano erupting with ice

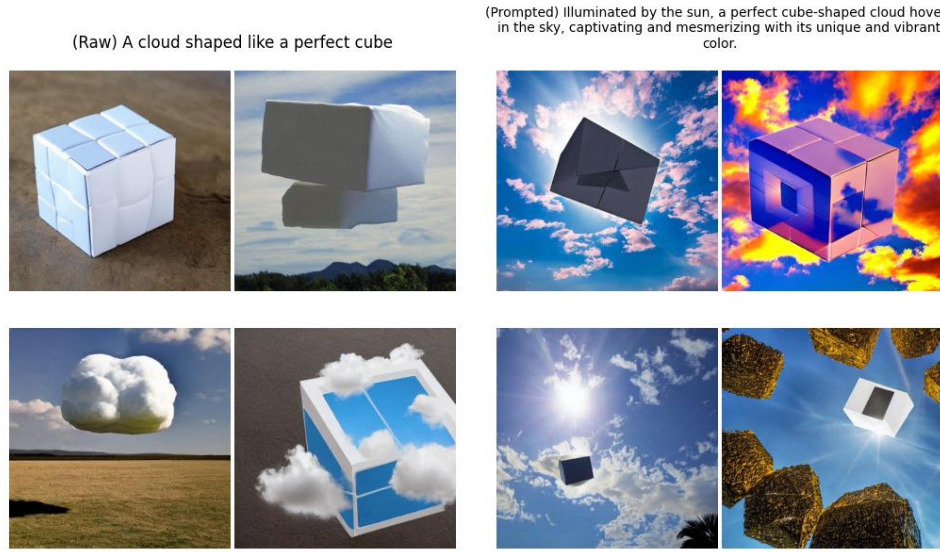


Figure 22: Failed improvement via CoT prompting: A cloud shaped like a perfect cube

3.5. Evaluation and Result

3.5.1. Improvement of Consistency

Due to the fact that the results of image generation are extremely subjective here, we adopt human evaluation. This process is conducted by a blind test comprised of image-text pairs, where the image can be either from the raw description or the prompted description, and the text is the raw description.

There are four images in each of the 50 unusual descriptions, $\mathcal{D} = \{D_1, D_2, \cdots, D_{50}\}$, where contains three types of formats, namely T_0, T_1, and T_2 (i.e. raw description, the description generated by template-based prompting, and the description generated by CoT prompting, respectively). We recruit three human raters, denoted R_1, R_2, and R_3, to independently determine whether each generated image fits the raw description, 1 if it is fitting, and 0 otherwise.

The scores per T per R are then averaged among description set $\mathcal{D}$ to quantify the improvement from T_0 to T_1 or T_2, based on each rater's perspective. The result is shown in

Table 1.

Rater	Format T_1	Format T_2	Better Fitting Format
R_1	0.095	0.220	T_2
R_2	0.160	0.285	T_2
R_3	0.095	0.060	T_1

Table 1: Average score improvement from format T_0 for each rater and comparison of format T_1 and T_2

3.5.2. Agreement of Raters

The agreement of raters is quantified by Cohen's kappa coefficient, which is a statistical measure of inter-rater reliability that takes into account the possibility of agreement occurring by chance. Cohen's kappa coefficient generally ranges from 0 to 1 (though it can be as low as -1), with values closer to 1 indicating higher agreement between raters and values closer to 0 indicating lower agreement. A kappa coefficient of 0 indicates that the observed agreement between raters is no better than chance, while a value of 1 indicates perfect agreement.

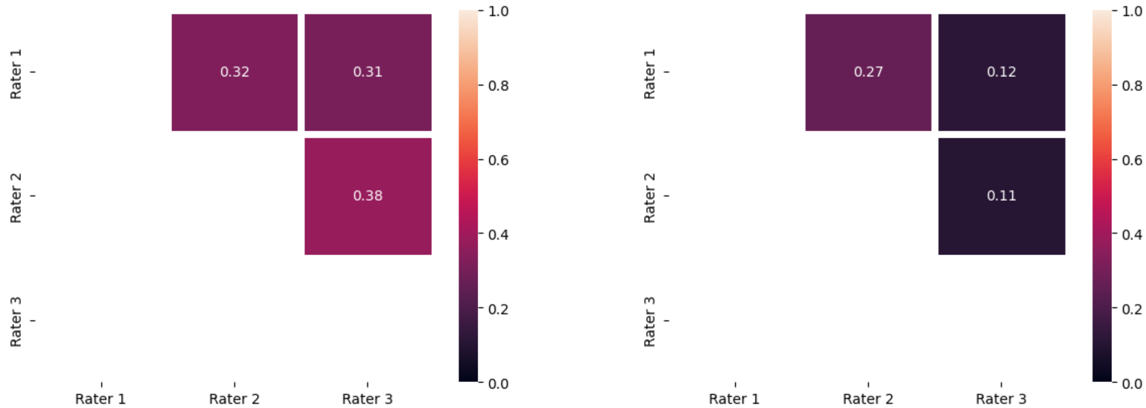
It can be expressed in the mathematical formula:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

where κ represents Cohen's kappa coefficient, P_o is the observed agreement between raters, and P_e is the expected agreement by chance.

By calculating the observed agreement and expected agreement, Cohen's kappa coefficient provides a standardized measure that considers both the observed agreement between raters and the agreement that could occur by chance. This assessment helps evaluate the inter-rater reliability and provides insights into the quality and consistency of the ratings or judgments under examination. Therefore, it is chosen as a robust measurement of the agreement in this work.

The result between three human raters is shown in Figure 23, visualized as a heatmap.



(a) Format T_1: Template-based prompt

(b) Format T_2: CoT prompt

Figure 23: Cohen's kappa coefficient of the human evaluations among three raters

3.6. Discussion

According to the data presented in Table 1, all raters unanimously agree that both template-based prompting and CoT prompting are effective in leveraging the LM for text-to-image generation of DM. The observed improvement ranges from 0.060 to 0.285 in terms of score. However, there is disagreement among the raters regarding which method exhibits a more substantial improvement. Two out of three raters consider the CoT-prompting method to be superior, significantly outperforming the template-based approach. Conversely, the remaining rater holds an opposing view. This discrepancy not only pertains to the formatting level but also extends to the image level.

To evaluate the degree of agreement, Figure 23 provides insights. The three raters demonstrate a fair agreement regarding the effectiveness of the template-based method, with a Cohen's kappa coefficient of approximately 0.3. In contrast, the agreement concerning the CoT method is slight. While R_1 and R_2 reach a fair agreement on the substantial improvement achieved by the CoT method, with a minimum score improvement of 0.22 (Cohen's kappa coefficient of around 0.3), R_3 holds an opposing opinion. R_3 suggests that the CoT method only contributes to a slight improvement, roughly 0.06 in terms of score (Cohen's kappa coefficient of around 0.1).

Given the high level of disagreement observed, we refrained from averaging the scores among the raters to represent the averaged interpretation. Instead, we reported the individual scores provided by each rater. This approach promotes transparency and facilitates a clearer understanding of the diverse perspectives and opinions among the raters.

CHAPTER 4 : Sketch-Then-Draw

4.1. Introduction

The imagine-then-verbalize method, as introduced in Chapter 3, has been developed with the objective of improving the coherence between generated images and their corresponding descriptions. By manipulating and expanding the textual input, this method endeavors to enhance the consistency between the two modalities. However, despite its advancements, the method still faces challenges in achieving direct control over the generation process of detailed descriptions. In particular, aspects such as object location and relational structure continue to exhibit inconsistencies.

To address these limitations, the focus of this chapter is to explore how LMs can be leveraged to directly manipulate the process of generating detailed descriptions. By harnessing the power of LMs, we aim to tackle the persistent issue of inconsistency in finer detail. To illustrate this concept, we specifically examine the issue of numerical consistency.

Numerical consistency plays a crucial role in ensuring accurate and reliable descriptions. In many scenarios, it is essential to maintain precise numerical values and relationships within the generated descriptions. For instance, if a scene involves a group of objects or entities with specific quantities, it is crucial that the generated description accurately reflects these numerical attributes. However, achieving such consistency poses a challenge within existing methodologies.

To overcome this challenge, we tap into an additional capability of LMs: the ability to write code. By harnessing this capacity, we can establish a structured foundation for the image generation process using scalable vector graphics (SVG) programming. This allows us to incorporate precise numerical specifications within the code, ensuring the generation of images that align with the desired frequency of objects.

By integrating LM-based control through code-based manipulation, we aim to demonstrate the potential for overcoming the limitations of the imagine-then-verbalize method. Through this approach, we seek to achieve improved control and consistency in generating detailed descriptions, specifically addressing the issue of numerical consistency. By laying the groundwork for precise and accurate image generation, we pave the way for advancements in multimodal AI systems that integrate symbolic and neural paradigms in image generation.

4.2. Dataset and Model

Here, we develop and use a numerical challenging dataset, which contains five descriptions that explicitly state the number of objects in different scenarios, as listed in Figure 24.

Five pigs are playing on the ground
Four coins on the table
Sixteen birds flying in formation
Twenty cookies arranged on a tray
Twenty-one leaves falling from a tree

Figure 24: Numercial challenging descriptions

For the DM, we use Huggingface's `stable-diffusion-v1-4` [1]; for the LM, we use `gpt-4` [2].

4.3. Prompting Method

Similar to the method in Bubeck et al. (2023) that combines GPT-4 and stable diffusion, we first let the GPT4 generate the SVG programming code of the description, then render the SVG image (sketch), then let the DM generate the final image given the mixture of the original description and the SVG image (draw). In detail, we use an online SVG generator [3] to render the SVG and stability.ai's DreamStudio [4] to combine SVG image with the original description to get the final output.

The prompt template is just one line, as shown in Figure 26. As a result, the LM outputs the SVG snippet that depicts the challenging description in a simple but numerically correct

[1] `https://huggingface.co/CompVis/stable-diffusion-v1-4`
[2] `https://openai.com/research/gpt-4`
[3] `https://www.svgviewer.dev/svg-to-png`
[4] `https://beta.dreamstudio.ai/generate`

way. Figure 27 is an SVG example of the description "Five pigs playing on the ground", which renders the SVG image as a sketch, visualized in Figure 28.

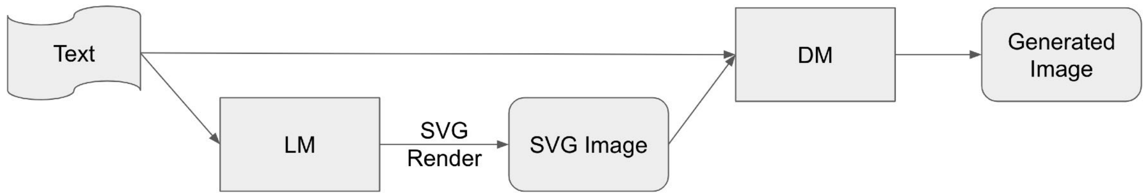


Figure 25: The diagram of the sketch-then-draw pipeline

Produce an image of {} using SVG code. Pay attention to the specified number, etc..

Figure 26: Sketch prompting template

I can provide you with a simple SVG code that represents five pigs playing on the ground. You can copy and paste this code into an HTML file to view the image or use an online SVG editor to see the result and make adjustments as needed.

```
<!DOCTYPE html> <html> <body> <svg width="600" height="300"
xmlns="http://www.w3.org/2000/svg"> <rect width="100%" height="100%"
fill="D0F0C0" /> <ellipse cx="100" cy="200" rx="50" ry="30" fill="F1C4AF"
/> <circle cx="130" cy="180" r="10" fill="black" /> <ellipse cx="250"
cy="200" rx="50" ry="30" fill="F1C4AF" /> <circle cx="280" cy="180" r="10"
fill="black" /> <ellipse cx="400" cy="200" rx="50" ry="30" fill="F1C4AF"
/> <circle cx="430" cy="180" r="10" fill="black" /> <ellipse cx="180"
cy="150" rx="50" ry="30" fill="F1C4AF" /> <circle cx="210" cy="130" r="10"
fill="black" /> <ellipse cx="330" cy="150" rx="50" ry="30" fill="F1C4AF" />
<circle cx="360" cy="130" r="10" fill="black" /> </svg> </body> </html>
```

This code creates a simple representation of five pigs playing on the ground. The pigs are represented by ellipses for their bodies and circles for their eyes. The background is filled with a light green color to represent the ground. You can further customize the code to create more detailed and visually appealing pigs, or add additional elements to the scene

Figure 27: Sketch prompting example: Five pigs playing on the ground

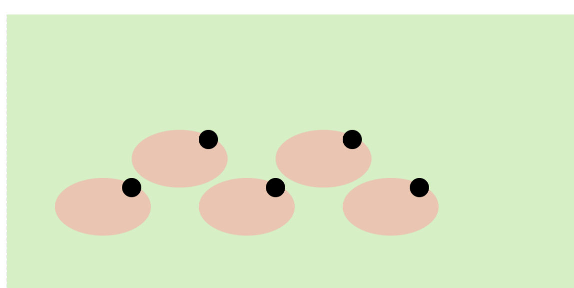


Figure 28: The SVG image sketched for "Five pigs playing on the ground"

4.4. Output Demonstration

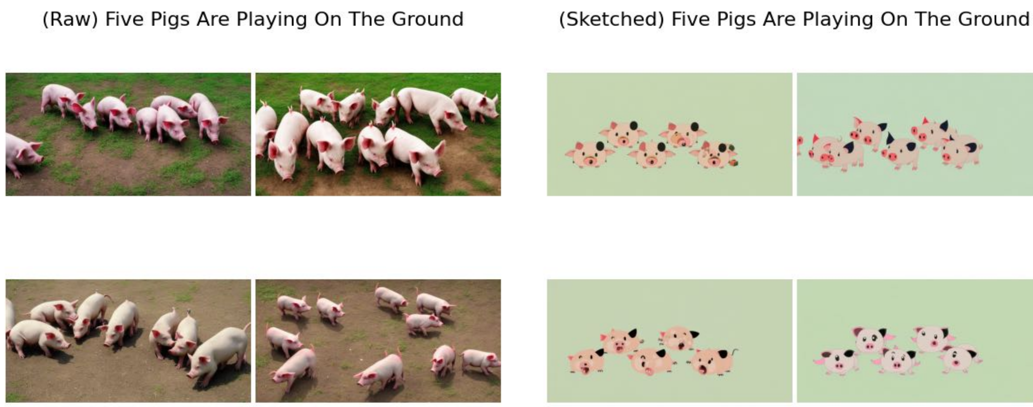


Figure 29: Sketch-then-draw example: Five pigs are playing on the ground

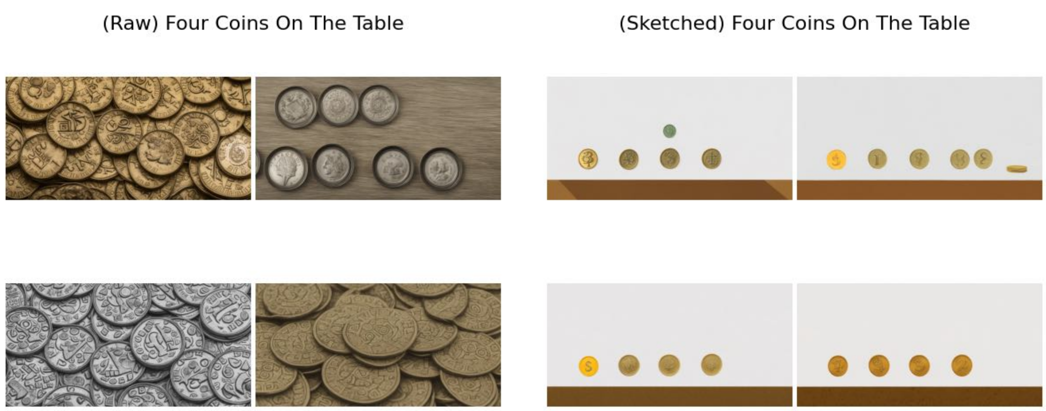


Figure 30: Sketch-then-draw example: Four coins on the table

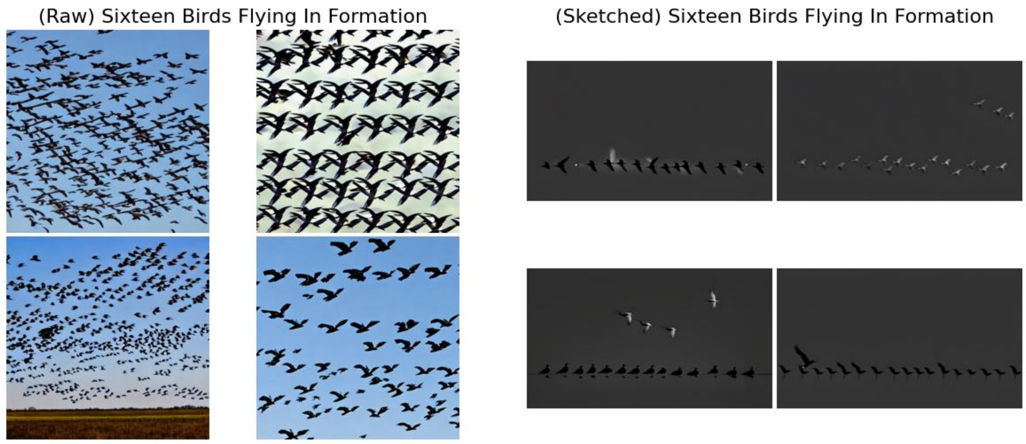


Figure 31: Sketch-then-draw example: Sixteen birds flying in formation

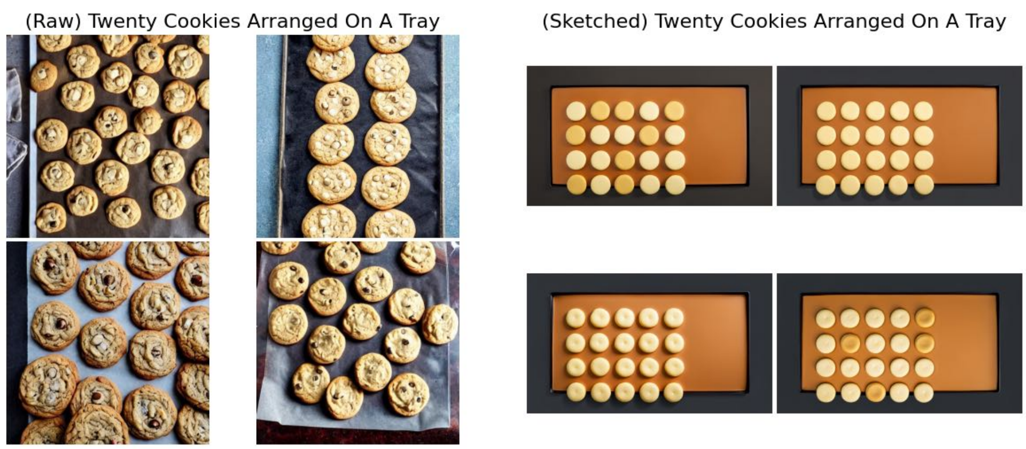


Figure 32: Sketch-then-draw example: Twenty cookies arranged on a tray

Figure 33: Sketch-then-draw example: Twenty-one leaves falling from a tree

4.5. Result and Discussion

For each experiment, we denote the frequency of the target object in the description as F_d and that in the generated image as F_m. Among the four images produced for each experiment, we calculate the average absolute edit distance $\mathbf{d}$ as

$$\mathbf{d} = \frac{1}{4}\sum_{i=1}^{4} |F_d - F_m|$$

where the absolute value is used to ensure that the differences are positive and to ignore the direction of the difference. This way, we are only interested in the magnitude of the difference, not whether F_d is greater or smaller than F_m.

Experiment	Edit Distance (Raw)	Edit Distance (Sketched)
5 Pigs	3.25	0
4 Coins	> 10	0.75
16 Birds	> 10	2.25
20 Cookies	5.25	0
21 Leaves	> 10	3.5

Table 2: The average absolute edit distance between the raw output and the sketched output per experiment

The average absolute edit distance between the raw output and the sketched output is depicted in Table 2, where **d** is marked as > 10 when the target objects in the images are highly overlapping and extensive to be accurately counted.

Remarkably, the sketched method consistently reduces the edit distances across all five experiments. The average absolute edit distances demonstrate a notable improvement when compared to the raw output. In fact, two out of the five experiments achieve zero edit distances, indicating a near-perfect match between the sketched representations and the generated images.

Despite the encouraging reduction in edit distances, it is crucial to consider the limitation by eyeballing the visual results obtained from Figure 29 to Figure 33. These visualizations illustrate the impact of the SVG-rendered sketches on the subsequent DM generation. While the SVG-rendered sketches enhance numerical consistency, this improvement comes at a substantial cost. The fidelity of the generated images is compromised, leading to a loss of visual quality and semantic meaning.

Therefore, while the sketched method effectively reduces the edit distances and enhances numerical consistency, it is important to strike a balance between these improvements and preserving the overall quality and semantic coherence of the generated images.

CHAPTER 5 : Conclusion

5.1. Diffusion model is poor on unusual generation

The failure of DM generation on the constructed dataset of unusual and numerical challenging descriptions suggests that DM is prone to perform poorly on the task of generating counterintuitive, anti-commonsense, and numerically challenging images.

The above observation can be attributed to the absence of crucial components in DM generation, occasional misunderstandings of the semantic meaning, or concept leakage, possibly resulting from OOD problems concerning the training data of DM.

5.2. Language model can (somewhat) help diffusion model

To address this issue, we propose a paradigm known as imagine-then-verbalize. This approach leverages the prompting method of the LM to improve image generation by imaginative visual and convincing features through step-by-step reasoning and explanation. As confirmed by the human evaluation, this method succeeds in improving the consistency between the text-to-image diffusion generation and the original descriptions. However, it should be noted that this method lacks stability and generalizability. While the prompting method cannot fundamentally resolve the OOD problem, it may serve as a bridge to mitigate the issue.

All raters agree that LM can assist in generating images, but there exist disagreements at both the image level and the prompting method level. Overall, the CoT prompting method exhibits potential as a superior alternative to template-based prompting, but further research is warranted.

The sketch-then-draw method, which harnesses the coding capability of the LM to generate an SVG image as an intermediate sketch, has the potential to ensure numerical consistency. However, it comes at the cost of missing flexibility and complex semantic interpretation.

Further, evaluating these image generations is highly subjective, and even human evaluations can yield significant disagreements. Therefore, future research should consider adopting a more meticulous experimental design and evaluation pipeline to increase the reliability of the results.

5.3. Future Works

To advance the application of LM-assisted DM generation methods, such as imagine-then-verbalize, the research community should focus on addressing the instability in both text and image generation. Finetuning the LM to generate extended descriptions that are easier for the DM to understand is a possible solution. Another way is to build a classifier to detect the unwanted concepts in the generation step of LMs and treat them as the negative prompts of DMs.

To mitigate fidelity loss in the sketch-then-draw method, the LM should generate programming in a more flexible manner, while the DM generation should be conditioned on key information from the sketch (e.g., object location, outline). This approach should allow for the precise depiction of the original nuance such as numerical features in the description while retaining freedom in image representation. Additionally, integrating ControlNet (Zhang and Agrawala, 2023) could be beneficial: the LM can guide the generation of pose structure outlines, while the DM can generate realistic people based on these outlines.

CHAPTER 6 : Limitations

There are several limitations:

1. We didn't test all the popular text-to-image DMs such as Imagen and Dall-E2, due to the API inaccessibility or budget issue. Some of them have a better capacity for language input understanding (e.g. Imagen uses T5 as a frozen text encoder) and thus may lead to different results. thus may lead to different results.

2. The scale of the experiment is not large, and the category of the descriptions is not diverse and overarching enough, also due to the limit of budget and manpower.

3. It is unfair to compare the realistic and fictional images together since the latter is more open to imagination. It is better to control the style of image generation at the beginning to get a better comparison.

4. Research involving APIs can be subject to instability, introducing potential variability in the outcomes.

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*, 33:1877–1901, 2020.

S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

T. Chakrabarty, A. Saakyan, O. Winn, A. Panagopoulou, Y. Yang, M. Apidianaki, and S. Muresan. I spy a metaphor: Large language models and diffusion models co-create visual metaphors. In *Findings of ACL*, 2023.

M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. d. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

J. Cheng, L. Dong, and M. Lapata. Long short-term memory-networks for machine reading. *arXiv preprint arXiv:1601.06733*, 2016.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://aclanthology.org/N19-1423`.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*, 2022.

A. K. Lampinen, I. Dasgupta, S. C. Chan, K. Matthewson, M. H. Tessler, A. Creswell, J. L. McClelland, J. X. Wang, and F. Hill. Can language models learn from explanations in context? *arXiv preprint arXiv:2204.02329*, 2022.

Y. Li, Z. Lin, S. Zhang, Q. Fu, B. Chen, J.-G. Lou, and W. Chen. On the advance of making language models better reasoners. *arXiv preprint arXiv:2206.02336*, 2022.

N. Liu, S. Li, Y. Du, A. Torralba, and J. B. Tenenbaum. Compositional visual generation with composable diffusion models. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*, pages 423–439. Springer, 2022.

P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.

G. Marcus, E. Davis, and S. Aaronson. A very preliminary analysis of dall-e 2. *arXiv preprint arXiv:2204.13807*, 2022.

M. Nye, A. J. Andreassen, G. Gur-Ari, H. Michalewski, J. Austin, D. Bieber, D. Dohan, A. Lewkowycz, M. Bosma, D. Luan, et al. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*, 2021.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

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. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

R. Rassin, S. Ravfogel, and Y. Goldberg. Dalle-2 is seeing double: Flaws in word-to-concept mapping in text2image models, 2022.

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

P. Wang, J. Zamora, J. Liu, F. Ilievski, M. Chen, and X. Ren. Contextualized scene imagination for generative commonsense reasoning. In *International Conference on Learning Representations*, 2022a. URL `https://openreview.net/forum?id=Oh1r2wApbPv`.

X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, and D. Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022b.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. Chi, Q. Le, and D. Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

S. Wiegreffe, A. Marasović, and N. A. Smith. Measuring association between labels and free-text rationales. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10266–10284, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.804. URL `https://aclanthology.org/2021.emnlp-main.804`.

E. Zelikman, Y. Wu, J. Mu, and N. D. Goodman. Star: Bootstrapping reasoning with reasoning, 2022.

L. Zhang and M. Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023.