Text-to-Image Generation for Abstract Concepts

Jiayi Liao[1*†], Xu Chen[2*‡], Qiang Fu[2], Lun Du[2], Xiangnan He[3§], Xiang Wang[3§], Shi Han[2], Dongmei Zhang[2]

[1] University of Science and Technology of China
[2] Microsoft
[3] MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition, University of Science and Technology of China
ljy0ustc@mail.ustc.edu.cn, {xu.chen, qifu, lun.du, shihan, dongmeiz}@microsoft.com,
{xiangnanhe, xiangwang1223}@gmail.com

Abstract

Recent years have witnessed the substantial progress of large-scale models across various domains, such as natural language processing and computer vision, facilitating the expression of concrete concepts. Unlike concrete concepts that are usually directly associated with physical objects, expressing abstract concepts through natural language requires considerable effort since they are characterized by intricate semantics and connotations. An alternative approach is to leverage images to convey rich visual information as a supplement. Nevertheless, existing Text-to-Image (T2I) models are primarily trained on concrete physical objects and often struggle to visualize abstract concepts. Inspired by the three-layer artwork theory that identifies critical factors, **intent**, **object** and **form** during artistic creation, we propose a framework of **T**ext-to-**I**mage generation for **A**bstract **C**oncepts (**TIAC**). The abstract concept is clarified into a clear intent with a detailed definition to avoid ambiguity. LLMs then transform it into semantic-related physical objects, and the concept-dependent form is retrieved from an LLM-extracted form pattern set. Information from these three aspects will be integrated to generate prompts for T2I models via LLM. Evaluation results from human assessments and our newly designed metric **concept score** demonstrate the effectiveness of our framework in creating images that can sufficiently express abstract concepts.

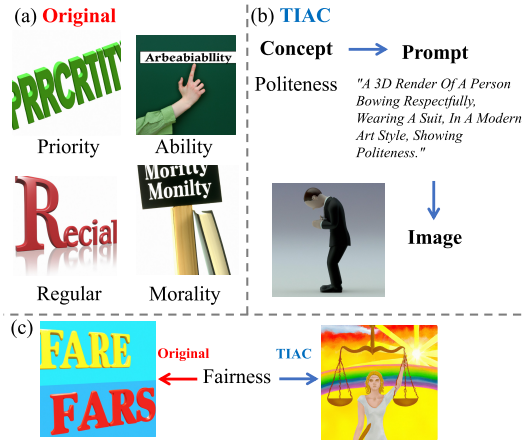

Figure 1: (a) Unsatisfactory cases generated by DALL·E 2 for abstract concept: priority, ability, regular and morality. (b) An illustration of "politeness" in the context of introducing cooperation skills. The prompt guiding the visualization is generated by TIAC, and the image is generated by DALL·E 2. (c) Image on the left is generated by taking word "fairness" as input, and the right one is produced with prompt from TIAC (LLM+PE), consisting of physical objects like a scale to express "fairness".

Introduction

Concepts are cognitive representations that encapsulate ideas. The expression of concepts plays a pivotal role in communication, especially within the context of profound intellectual discourse. Recent advancements in large-scale models in the field of natural language processing and computer vision (Cao et al. 2023; Zhang et al. 2023b; Wu et al. 2023) have demonstrated remarkable performance in conveying concrete concepts. These concrete concepts are perceptible entities or occurrences and are often associated with physical objects, such as animals and planets. On the contrary, how to express abstract concepts that encompass rich and intricate connotations is less explored (Recchia and Jones 2012; Liao, Chen, and Du 2023). Abstract concepts serve a crucial role in accurately conveying philosophical thoughts, moral perspectives, and emotional states in everyday life. Moreover, they can present themes in a deeper and multi-dimensional manner in domains like art, literature, and music, rich with creativity and aesthetic value. On the other hand, abstract concepts are mentally constructed ideas that

*These authors contributed equally.
†This work is done during the internship in Microsoft.
‡Corresponding author.
§Xiangnan He and Xiang Wang are also affiliated with Institute of Dataspace, Hefei Comprehensive National Science Center.

usually lack physical forms, and their definitions are often quite abstract as well. Therefore, when expressing abstract concepts through natural language, speakers encounter challenges in providing clear explanations, while recipients also face difficulties in understanding (Schwanenflugel 2013), leaving it a great barrier between human communications.

As an important channel of communication, conveying abstract concepts through vision can effectively alleviate the aforementioned challenges, enhancing the intuitive and vivid nature of their expression. For example, when using slides to introduce "Cooperation Skills" that include bullet points like "Politeness" and "Fairness', incorporating relevant images to represent these abstract concepts can greatly improve communication efficiency: an image with a person bowing respectfully for "Politeness" and a person holding a scale for "Fairness" as shown in Figure 1 (b-c).

While it is of great potential to express abstract concepts through the visual channel, current text-to-image generation (T2I) models face obstacles in realizing this purpose. Existing T2I models such as Stable Diffusion (Rombach et al. 2022), DALL·E (Ramesh et al. 2022), NUWA (Wu et al. 2022) and Imagen (Saharia et al. 2022) have achieved impressive improvements in generating realistic and eye-catching images with input texts (Zhang et al. 2023a). However, these models are primarily designed for concrete concepts, and high-quality datasets (Lin et al. 2014; Young et al. 2014; Wah et al. 2011; Nilsback and Zisserman 2008) of text-image pairs used for their training only focus on physical objects, leading to the unsatisfactory generalization capabilities of T2I models for abstract concepts naturally (Ramesh et al. 2022; Saharia et al. 2022; Nichol et al. 2022; Yu et al. 2022). When an abstract concept is directly inputted into a T2I model, it often generates images with distorted English letters resembling the input, as seen in the four images in Figure 1 (a) and the left image of Figure 1 (c). An analysis of prompts submitted by users on T2I servers (Xie et al. 2023) also reveals that images generated from vague and abstract prompts are often scored lower by users.

To tackle the above challenges, we draw inspiration from a 3-layer artistic creation hierarchy (Ocvirk et al. 1968; Xie et al. 2023), which is illustrated in Figure 2. The hierarchy is organized in a top-down manner, including: (1) **Intent** layer that reveals the high-level purpose the creator intends to express. (2) **Object** layer, which denotes physical objects and their spatial relationships. (3) **Form** layer that refers to the basic elements of artistic styles, such as line, color, shape, and texture, along with their arrangement. With the hierarchy in art, we can provide a new perspective to explain why T2I models perform worse for abstract concepts. When creating images for concrete concepts, information from the intent layer and the object layer are highly similar, i.e., what I think is what I will draw on the picture, and the corresponding form information is easy to obtain with plenty amount of images drawing physical objects. However, for abstract concepts, the connection between the concept and the information on the three layers is not obvious for current T2I models. Therefore, the key to effectively expressing abstract concepts lies in building connections to the three layers.

In this paper, we propose a framework of **T**ext-to-**I**mage Generation for **A**bstract **C**oncepts **TIAC** that aims to bridge the gap between human input abstract concepts and the generated images by leveraging the knowledge stored in LLMs and their comprehension ability. Specifically, TIAC links abstract concepts to nodes in WordNet so that the abstract concepts can be bonded to unambiguous definitions, which contributes to clarifying the user intents. Abstract concepts are then transformed by LLMs to related physical objects to represent their connotations. Additionally, the concept-dependent form patterns extracted from a prompt dataset are retrieved to enrich the information from the form layer. By integrating the above information from three layers, LLMs can generate prompts that tangibly describe abstract concepts, enabling T2I models to create satisfactory images.

Through conducting experiments on the abstract branch of WordNet, we compare different approaches and design a new metric called **concept score**. The results indicate that prompts generated using our framework facilitate effective visualization of abstract concepts. Furthermore, the concept score demonstrates better consistency with human preferences compared to existing metrics for assessing the alignment between abstract text inputs and generated images. Our framework, TIAC, optimizes prompts directly without necessitating model fine-tuning, making it adaptable to various T2I models. The main contributions are as follows:

- We introduce a novel task of text-to-image generation for abstract concepts, aiming to fill the gap in abstract concept expression in the area of image generation.
- We design TIAC leveraging LLMs to integrate the enriched information of abstract concepts in three layers.
- We propose concept score, a new metric that is more aligned with human cognition for evaluating images generated for abstract concepts. Experimental results demonstrate the effectiveness of TIAC in this task.

Related Work

Concept Expression in Image Generation. In the exploration of image generation, there has been a recent upsurge in interest regarding concept expression. For example, concept customization (Gal et al. 2022; Ruiz et al. 2022; Kumari et al. 2022; Wei et al. 2023) aims to integrate existing T2I models with new concepts. However, these concepts are either specific objects like newly created objects "moongates" or customized objects in our daily lives like someone's pet dog. Furthermore, concept disambiguation (Mehrabi et al. 2022) also focuses on the syntactic equivocation inherent in human input, which leads to ambiguity concerning the referential relationships of physical objects, rather than delving into the subtle distinctions within abstract concepts. In general, current research in the field of image generation predominantly emphasizes the depiction of physical concepts rather than abstract ones. Consequently, we aim to bridge the gap in the study of abstract concepts within this domain.

Abstract Concepts in Computer Vision. One exemplary application of abstract concepts in computer vision is ad images. Ad images are creative artworks that convey abstract concepts, incorporating a wealth of knowledge such

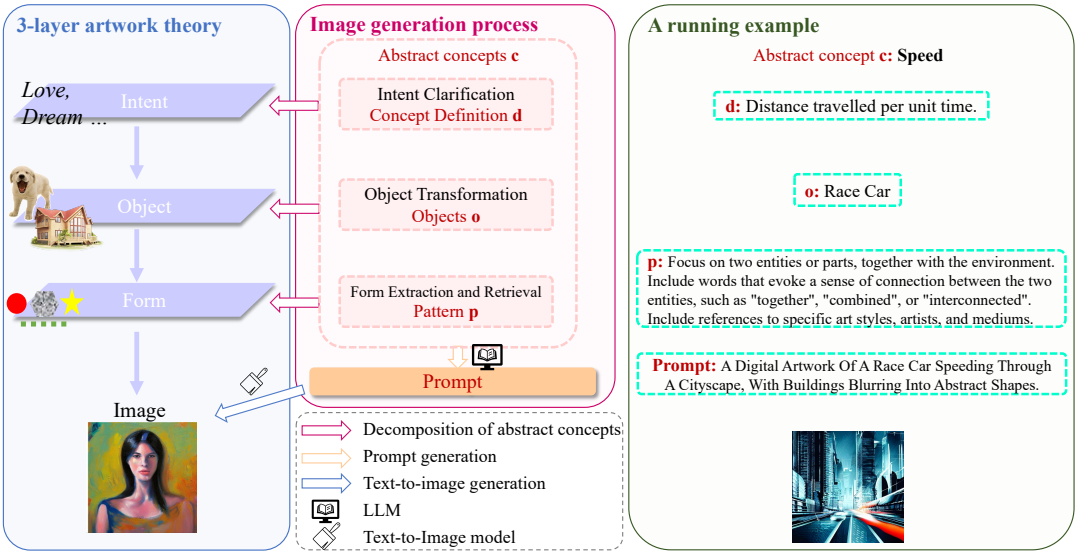

Figure 2: Framework of TIAC. The left section demonstrates the intent/object/form layer considered when creating artwork. The middle part denotes how to decompose abstract concepts into 3 layers and corresponding prompt and image generation process. The right section is a running example showing how TIAC depicts the concept "speed".

as common-sense reasoning, cultural context, and symbolism. Hence, tasks associated with ad images pose significant challenges for machines. Several datasets (Hussain et al. 2017; Akula et al. 2022; Xu et al. 2022) and works of ads understanding (Ye and Kovashka 2018) and generation (Chilton, Petridis, and Agrawala 2019) have been proposed for visual ads. The underlying idea of the paper mentioned before is combining objects in a logical manner to convey messages. This suggests that they also share the belief that the essence of abstract concept research lies in selecting objects and forms that can simultaneously express information. Visual ads are a subset of our research, highlighting the potential application of abstract concept studies in the realm of artistic creation. In addition, our applications also encompass illustrations of slides, decorative paintings, and more.

Prompt Optimization for T2I Models. While the performance of large-scale models driven by textual inputs is progressively advancing, the resource-intensive nature of training and fine-tuning these models has posed challenges for researchers to afford. Consequently, improving the model input directly, known as prompt optimization, emerges as a commendable choice. It can enhance image quality without altering the model structure or requiring extensive training and fine-tuning of T2I models. This field is relatively new, resulting in a lack of comprehensive research. Experiences of writing good T2I prompts manually are shared through blog posts and user guidebooks(Oppenlaender 2022; Smith 2022; Pavlichenko and Ustalov 2022). Basic elements for a good T2I prompt and prompt terms (*i.e.,* modifiers) describing various perspectives of image style are summarized by a taxonomy survey (Oppenlaender 2022) and DALL·E 2 prompt book (Parsons 2022). As prompt optimization can be conducted in either text space or embedding space (i.e., soft tuning (Lester, Al-Rfou, and Constant 2021)), some studies also train a prompt optimization model for soft tuning, which results in a high degree of coupling with the T2I model (Hao et al. 2022). However, current T2I prompt optimization (Ge et al. 2022) primarily aims to improve the style and aesthetics of synthesized images. In contrast, our research will prioritize the visual comprehension of concepts.

Preliminary

Definition 1 (Concept, concrete concept and abstract concept)**.** *Concepts are mental representations of coherent classes of entities (Schwanenflugel 2013); they can be divided into concrete concepts and abstract concepts. Concrete concepts are perceivable objects or occurrences, whereas abstract concepts are those that cannot be directly perceived through senses (Zdrazilova, Sidhu, and Pexman 2018; Katja Wiemer-Hastings and Xu 2005).*

As intuitive examples, concrete concepts can be a tiger, a keyboard, or a T-shirt, while abstract concepts can be dream, happiness, or love. Based on the above definition, the new task is further defined as:

Definition 2 (Text-to-image generation for abstract concept)**.** *Given a human text input that intends to express an abstract concept* $\boldsymbol{c}$*, the task of text-to-image generation for abstract concepts requires a mapping* $\boldsymbol{f}$ *to produce images that can reveal the meaning of* $\boldsymbol{c}$*.*

$$\boldsymbol{Image} = \boldsymbol{f}(\boldsymbol{c}). \tag{1}$$

Method: TIAC

To deal with this task, we propose TIAC which is inspired by the 3-layer artwork theory. The framework consists of 4 stages and is demonstrated in Figure 2. (1) **Intent Clarification** stage aims to clarify the human intent. The input text will be linked to an existing entity in the WordNet

knowledge base to retrieve its detailed definition as the definite intent. (2) **Object Transformation** stage is designed to decompose the abstract concept in the object layer. Here, concrete objects related to the intent will be obtained with the help of external knowledge and comprehension ability of LLMs. (3) **Form Extraction** stage will enrich the form information conditioned on the intent, which will be accessed through pre-extracted form patterns from a high-quality human-submitted prompt dataset. (4) **Prompt Generation and Image Generation** is the final stage and the above information will be integrated to generate prompts for T2I models so that more desirable images can be produced. Design in each stage will be elaborated on in this section.

Intent Clarification

When visualizing a human input abstract concept, its potential multiple connotations undoubtedly increase the difficulty of expression. For example, "energy" can mean (1) *enterprising or ambitious drive* or (2) *a thermodynamic quantity equivalent to the capacity of a physical system to do work* and so on. Hence, it is imperative to clarify the precise semantics of the input abstract concept so that there is an exact drawing intent in the intent layer. To achieve this goal, an abstract concept is linked to a synset in WordNet (Miller 1995; Du et al. 2021) with a definite meaning. WordNet is a lexical database where semantically similar words are grouped into a set of cognitive synonyms called a synset; thus each synset represents a unique concept with corresponding definition in the database. Here, we focus on nouns in WordNet that are organized into a hierarchical tree with the root node of "entity.n.01"[1]. There are two main branches under it: one is a subtree rooted at "abstraction.n.06" and the other rooted at "physical_entity.n.01.". The former is the main focus of this paper as it represents abstract concepts.

Based on the subtree of abstract concepts $\boldsymbol{T}$, the intent clarification stage can be mathematically expressed as

$$\boldsymbol{d} = \boldsymbol{f}_{IC}(\boldsymbol{c}; \boldsymbol{T}). \tag{2}$$

For each human input abstract concept $\boldsymbol{c}$, we link it to a specific node under the abstraction subtree $\boldsymbol{T}$ and retrieve the definition $\boldsymbol{d}$ of this node in WordNet. The intent clarification mapping $\boldsymbol{f}_{IC}$ builds a bridge between input abstract concepts and detailed drawing intent to mitigate potential ambiguity. Note that we assume the human input can be precisely mapped to a node in WordNet, whereas in the real scenario, it requires more efforts to determine the corresponding node in WordNet for input abstract concepts without other contexts. But the key idea is to identify the drawing intent of users, and we can achieve this by simply retrieving the definitions of all relevant WordNet nodes and asking users to decide from this candidate intent set. Overall, the definition $\boldsymbol{d}$ is now regarded as the exact intent in this task.

Object Transformation

The WordNet-based intent clarification alleviates the burden of ambiguous intent, but the intricate and abstract definition of abstract concepts can still be a barrier for this task. As seen in the case study, directly using the abstract concept or its definition as the input of T2I model both fail to yield satisfactory results, and the corresponding images are usually characters of input words with random noise instead of meaningful objects. On the other hand, there is a strong correlation between the intent layer and the object layer for concrete concepts; thus, they can be easily transformed into concrete objects and then illustrated by current T2I models. Hence, the critical step is to transform the abstract concept from the intent layer to the object layer.

Actually, abstract concepts and concrete objects are not entirely irrelevant; in WordNet, the entity "abstraction" is defined as *extracting common features from specific examples*, indicating that abstract concepts can be the summarization of properties or states of physical objects, or interactions between them. Therefore, concrete objects and their interactions can serve as instantiations of abstract concepts in reverse.

To fulfill the transformation from the abstract concept to relevant physical objects or their interactions, we utilize LLMs for their knowledge and understanding ability. More specifically, LLMs have been trained with numerous corpus and should establish associations between concepts and relevant objects. With appropriate instructions, it can assist in the object transformation for given abstract concepts. This process can be formulated as:

$$\boldsymbol{o} = \boldsymbol{f}_{OT}(\boldsymbol{c}, \boldsymbol{d}; \boldsymbol{i}). \tag{3}$$

Here given an abstract concept $\boldsymbol{c}$ and its definition $\boldsymbol{d}$ as well as the object transformation instruction $\boldsymbol{i}$, the function $\boldsymbol{f}_{OT}$ (i.e., LLMs) is able to ground the human input to concrete objects $\boldsymbol{o}$. As an example, the abstract concept "shrinkage" can be transformed into objects such as "deflating balloon", with $\boldsymbol{i}$ that employ words prompting LLMs to contain concrete objects which exemplify the meaning of concepts.

Form Extraction and Retrieval

Besides information from the object layer, incorporating information from the form layer also aids in expressing abstract concepts. The form layer depends on the intent layer, while LLMs may have difficulty generating form information for abstract concepts directly due to limited training data available about it. Hence, we introduce a dataset Simulacra Aesthetic Captions (SAC) (Pressman, Crowson, and Contributors 2022) to enhance LLMs in building the connection between intent and form. Moreover, abstract concepts in the same class are supposed to share common form information (Du et al. 2021; Chen et al. 2020; Chen, Wang, and Xie 2021; Chen et al. 2022; Du et al. 2022). Therefore, our approach involves initially extracting patterns pertaining to the form of a concept class. Subsequently, for a given concept, the corresponding form pattern is retrieved for the following usage. Details will be illustrated as follows.

Form Extraction. In form extraction, form patterns that describe the artistic properties of objects, how they are organized and the style of the picture are extracted with in-context learning from the SAC dataset. It consists of over

[1] "entity.n.01" is the notation of a synset in WordNet where "entity" is the synset name and "n" means it is a noun. Synsets with the same name are distinguished by a number like "01".

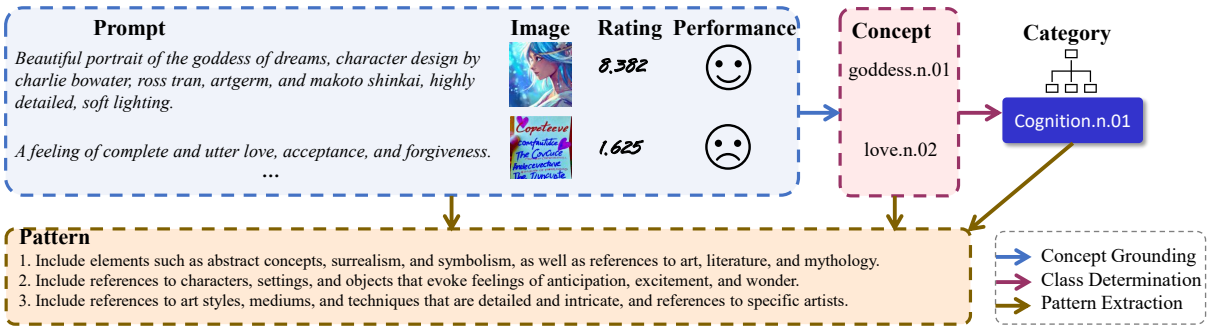

Figure 3: Details for 3 steps of Form Extraction in TIAC.

33k user-submitted T2I prompts and over 238k images generated from these prompts along with user ratings (1∼10). In preprocessing, we assign the average rating as the score for prompts and remove prompts with fewer than 3 user ratings. The performance of prompts with scores below 3 is viewed as "bad" and above 8 is "good". Form extraction is conducted through three steps: concept grounding, class determination and pattern extraction, as shown in Figure 3.

Concept Grounding. Human-submitted prompts contain other information besides the desired abstract concepts; thus, we ground the main concept in the prompt data $\boldsymbol{P}$ to a WordNet node $\boldsymbol{c}'$ by:

$$\boldsymbol{c}' = \boldsymbol{f}_{\text{ground}}(\boldsymbol{P}; \boldsymbol{T}). \tag{4}$$

Here $\boldsymbol{f}_{\text{ground}}$ is an LLM that completes the tasks of (1) extracting the main concept of a prompt and ignoring artwork-related descriptions like pictures, images and so on; (2) grounding the concept to the most relevant synset in the abstract subtree $\boldsymbol{T}$ to obtain the intended abstract concept $\boldsymbol{c}'$.

Class Determination. Given the inadequate amount of prompt data for any single concept while abstract concepts under the same ancestor share a common form pattern, we group the mapped nodes into different concept classes. For an abstract concept $\boldsymbol{c}'$, its class can be obtained by

$$\boldsymbol{C} = \boldsymbol{f}_{\text{classify}}(\boldsymbol{c}'). \tag{5}$$

$\boldsymbol{f}_{\text{classify}}$ determines the class by finding a common ancestor as the class label for a group of nodes and ensuring each class has enough nodes while maintaining distinctiveness.

Pattern Extraction. Based on our pre-defined classes, the form pattern can be extracted with an LLM summarizing over a batch of prompts belonging to the same class in a contrastive manner as

$$\boldsymbol{p} = \boldsymbol{f}_{\text{extract}}(\boldsymbol{C}, \boldsymbol{D}, \boldsymbol{P}_{good}, \boldsymbol{P}_{bad}). \tag{6}$$

For a class $\boldsymbol{C}$, both good and bad prompts whose main concept is under $\boldsymbol{C}$ are put together. We instruct LLM $\boldsymbol{f}_{\text{extract}}$ to extract helpful form patterns from the good prompts $\boldsymbol{P}_{good}$ while avoiding harmful factors from bad ones $\boldsymbol{P}_{bad}$ based on the class and corresponding definition $\boldsymbol{D}$.

Form Retrieval. With the extracted form pattern, we can acquire the form information for an intended abstract concept. Given that it has been mapped to $\boldsymbol{c}$ through intent clarification, the corresponding class $\boldsymbol{C}$ can be known following Eq. 5 even if it does not appear in the SAC. Then the shared form pattern $\boldsymbol{p}$ can be retrieved from class $\boldsymbol{C}$, which has been extracted and stored during the form pattern extraction.

Prompt Generation and Image Generation

In cognitive psychology (Neisser 2014), the cognitive processes of humans are defined as "all mental processes by which the sensory input is transformed, reduced, elaborated, stored, recovered, and used". In our framework, the process of intent clarification, object transformation as well as form extraction and retrieval are similar to the transformation, reduction, elaboration and storage of abstract concepts. Accordingly, prompt generation and image generation are the final steps, recovery and utilization.

From the above stages, We have obtained $\boldsymbol{d}$ from the intent layer, $\boldsymbol{o}$ from the object layer and $\boldsymbol{p}$ from the form layer. Furthermore, considering the benefits of in-context learning for LLMs, we select eight good prompts from the guidebook of DALL·E 2 (Parsons 2022) as few-shot examples $\boldsymbol{e}$. Also, the task description and the consideration of the token limit of downstream models are both incorporated. Hence, by utilizing the abilities of LLMs in information integration and content generation, the prompt for T2I model is obtained by

$$\textbf{Prompt} = \text{LLM}(\boldsymbol{c}, \boldsymbol{d}, \boldsymbol{o}, \boldsymbol{p}, \boldsymbol{e}). \tag{7}$$

With T2I prompts from our designed framework, downstream T2I models are capable of generating images that can better express the intended abstract concepts:

$$\textbf{Image} = \text{T2I-Model}(\textbf{Prompt}), \tag{8}$$

where T2I-Model can be any text-to-image generation model like Stable Diffusion 2 and DALL·E 2.

Experiments

Experiment Settings

Datasets. We construct two datasets based on abstract concepts in WordNet with different scales. The small-scale one contains 57 abstract concepts and the large-scale one contains 3,400 abstract concepts. (1) For each subclass with more than 100 nodes and under the seven classes on WordNet, we sample 100 abstract concepts to constitute the large-scale dataset. (2) Based on the large-scale dataset, we further select three abstract concepts from each subclass whose number of prompts in SAC is over 10 to construct the small-scale dataset for human evaluation.

	HE	CS	IS	VS
W	1.86±1.23	0.91±0.02	2.58±0.09	0.24±0.00
W+D	2.60±1.56	1.05±0.03	2.66±0.12	**0.29±0.00**
LLM	3.34±1.54	1.14±0.01	2.75±0.12	0.27±0.00
LLM+P	3.52±1.40	1.32±0.01	2.62±0.11	0.25±0.00
LLM+PE	**3.80±1.30**	**1.50±0.01**	**2.76±0.07**	0.25±0.00

Table 1: Evaluation Metrics on the Small-Scale Dataset. Bold and underline indicate the best and the second best performance, respectively. HE: human evaluation, CS: concept score, IS: inception score, VS: visual-semantic similarity.

Implementation Details. T2I prompts generated from five different approaches are compared to verify the effectiveness: (1) **W** denotes taking **W**ords of the abstract concept as the prompt. (2) **W+D** further concatenates the abstract concept name and its **D**efinition as the prompt. (3) **LLM** means using the concept name and its definition as the input of **LLM** for prompt generation. (4) **LLM+P** introduces information of the transformed objects and extracted form **P**atterns compared with **LLM**. (5) **LLM+PE** represents utilizing Eq. 7 to obtain the prompts. Compared with LLM+P, few-shot **E**xamples $\boldsymbol{e}$ are added in LLM+PE.

We use GPT-3.5 (Brown et al. 2020) (text-davanci-003) as the LLM in our framework, and Stable Diffusion v2 (Rombach et al. 2022) (v2-inference and checkpoint of 512-base-ema) as the T2I model. Performance of explicitly imposing object transformation is slightly off, so we use a more intuitive instruction in the last two baselines.

Evaluation. We conduct two types of evaluations: human evaluation and image-generation metrics evaluation. (1) ***Human Evaluation.*** We make a survey using images generated from small-scale dataset. Given an abstract concept in the dataset, for each type of prompt, we generate three images for each prompt type and arrange them in a row. The five types of images are randomly shuffled, with both the concept name and definition provided. Respondents are asked to rate each row on a scale of 1 to 5 (1 is the worst and 5 is the best), indicating the perceived relevance between the images and the concept. We collect feedback from three participants. (2) ***Image-generation Metrics Evaluation.*** In the field of image generation, Inception Score (Salimans et al. 2016) (IS) and Fréchet Inception Distance (Heusel et al. 2017) (FID) are used for evaluating image quality and fidelity, while R-precision (Xu et al. 2018) and Visual-Semantic similarity (Radford et al. 2021) (VS) are employed to measure the relevance between the image and the text. Here we adopt IS and VS in our experiments for evaluation.

Results on the Small-Scale Dataset

Experimental results of both human evaluation and image-generation metrics on the small-scale dataset are reported in Table 1. Table 1 demonstrates that LLM+P and LLM+PE achieve the top-2 results on human evaluation, showcasing significant improvements compared to taking raw human input or the concept definition as the prompts.

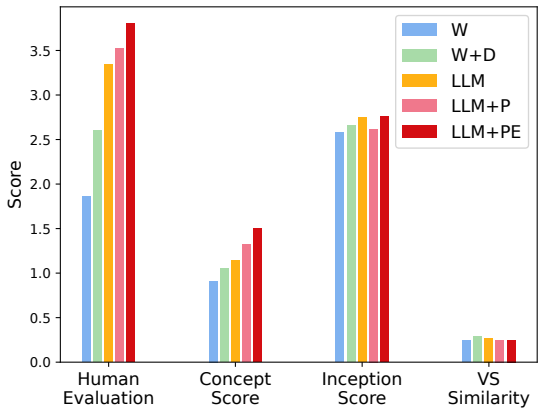

Figure 4: Evaluation Metrics on small-scale dataset. Concept scores are more consistent with human ratings.

Discussion on Image-Generation Metrics. Although our framework obtains the best performance in human assessment, Table 1 and Figure 4 also seem to demonstrate that LLM+PE and LLM+P are not so powerful on the image-generation metrics like IS and VS. In other words, the machines seem to be against the conclusion that our design is remarkably better than simply inputting the desired abstract concept. What is the reason for this gap? As a matter of fact, the task of image generation for abstract concepts is different from traditional image generation task that focuses on concrete concepts. For concrete concepts, there is a large number of training data and it is much easier to identify whether a concrete concept is correctly drawn on the image.

However, to understand the underlying abstract concepts, it requires deeper processing progress on the combinations of objects and forms (recalling viewing pieces of art in the exhibition), which is beyond the existing image-generation metrics. Actually, the calculation of the IS utilizes Inception Network to classify the images into the classes in ImageNet 1k, and the computation of VS encodes images and input text into a shared latent space using a pre-trained encoder (here we use CLIP for both image encoding and text encoding), and then calculates the distance between two vectors to measure the alignment between image and text. They are both designed for concrete objects, leaving the emergency of designing an evaluation metric for abstract objects.

To this end, we introduce the Aesthetic score (Schuhmann 2022) that is proposed to evaluate a generated image from the perspective of art considering abstraction to some extent. Hence, we consider both factors and take the aesthetic score as a coefficient for VS similarity as the measure of alignment between texts and images with abstraction:

$$\textbf{Concept score} = \textbf{VS similarity} \times \textbf{Aesthetic score} \quad (9)$$

By revisiting Figure 4, we can find that the concept score is more consistent with human preference. Hence, it enables us to evaluate the five approaches on the large-scale dataset.

Results on the Large-Scale Dataset

We further validate the effectiveness of different prompts on large-scale dataset with concept score. The calculation

Category	W	W+D	LLM	LLM+P	LLM+PE
Attribute.n.02	0.78±0.00	0.81±0.01	1.06±0.01	**1.42±0.00**	1.40±0.00
Cognition.n.01	0.83±0.01	0.92±0.01	1.07±0.00	1.27±0.01	**1.52±0.01**
Communication.n.02	0.87±0.00	0.90±0.01	1.01±0.01	1.29±0.01	**1.38±0.00**
Event.n.01	0.92±0.00	1.01±0.01	1.17±0.01	1.38±0.00	**1.50±0.00**
Group.n.01	0.99±0.01	1.17±0.01	1.29±0.00	1.44±0.01	**1.49±0.01**
Measure.n.02	0.98±0.01	1.12±0.01	1.21±0.00	1.26±0.01	**1.31±0.00**
Relation.n.01	0.85±0.01	0.93±0.01	1.13±0.01	1.13±0.02	**1.34±0.01**
Average	0.89±0.00	0.98±0.00	1.14±0.00	1.31±0.01	**1.42±0.00**

Table 2: Concept Score on the Large-Scale Dataset. Bold indicates the best performance.

(a) Healthfulness.n.01

(b) Government.n.03

(c) Auditory hallucination.n.01

Figure 5: Images from left to right are generated by five approaches: W, W+D, LLM, LLM+P and LLM+PE. Healthfulness.n.01 means the quality of promoting good health. Government.n.03 means the system or form by which a community or other political unit is governed. Auditory hallucination.n.01 means illusory auditory perception of strange nonverbal sounds.

is conducted over 3400 abstract concepts from 34 subclasses under the seven classes. The statistical results on the seven classes are organized in Table 2. Similarly, LLM+P and LLM+PE generally obtain the highest concept score, which illustrates that even being evaluated on a more diverse dataset with more abstract concepts, our design can still outperform other baselines and generate better images.

Case Study

We randomly select three abstract concepts and generate images with five different types of prompts to demonstrate the performance. As shown in Figure 5, LLM+P and LLM+PE generated images are much more meaningful and beautiful than the left ones. Concretely, LLM+P and LLM+PE convey the concept of "healthfulness" through running in the green forest or practicing yoga on the grass, and express "government" through a magnificent castle and a king sitting on the throne in the palace. For "auditory hallucination", LLM+P seems to produce a less relevant image, while LLM+PE guided image containing a person wearing headphones with an ethereality background is closer to the concept.

Conclusion

In this paper, we delve into a novel task text-to-image generation for abstract concepts. Intricate connotations associated with abstract concepts pose a great challenge for explanation and comprehension. Hence, we propose a framework TIAC to leverage the comprehension and generation abilities of LLMs in this task. TIAC enriches information from intent/object/form layer based on artwork theory so that LLMs can construct effective T2I prompts for better image generation. We design concept score inspired by our human assessment for comprehensive evaluations, and the results show our superiority in the task over baselines.

Acknowledgements

This paper was in part supported by the computing resources funded by National Natural Science Foundation of China (92270114).

References

Akula, A. R.; Driscoll, B.; Narayana, P.; Changpinyo, S.; Jia, Z.; Damle, S.; Pruthi, G.; Basu, S.; Guibas, L. J.; Freeman, W. T.; Li, Y.; and Jampani, V. 2022. MetaCLUE: Towards Comprehensive Visual Metaphors Research. *CoRR*, abs/2212.09898.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In *NeurIPS*.

Cao, Y.; Li, S.; Liu, Y.; Yan, Z.; Dai, Y.; Yu, P. S.; and Sun, L. 2023. A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT. *CoRR*, abs/2303.04226.

Chen, X.; Qiu, Q.; Li, C.; and Xie, K. 2022. GraphAD: A Graph Neural Network for Entity-Wise Multivariate Time-Series Anomaly Detection. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2297–2302.

Chen, X.; Wang, J.; and Xie, K. 2021. TrafficStream: A Streaming Traffic Flow Forecasting Framework Based on Graph Neural Networks and Continual Learning. In *International Joint Conference on Artificial Intelligence*.

Chen, X.; Zhang, Y.; Du, L.; Fang, Z.; Ren, Y.; Bian, K.; and Xie, K. 2020. Tssrgcn: Temporal spectral spatial retrieval graph convolutional network for traffic flow forecasting. In *2020 IEEE International Conference on Data Mining (ICDM)*, 954–959. IEEE.

Chilton, L. B.; Petridis, S.; and Agrawala, M. 2019. VisiBlends: A Flexible Workflow for Visual Blends. In *CHI*, 172. ACM.

Du, L.; Chen, X.; Gao, F.; Fu, Q.; Xie, K.; Han, S.; and Zhang, D. 2022. Understanding and Improvement of Adversarial Training for Network Embedding from an Optimization Perspective. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, 230–240.

Du, L.; Gao, F.; Chen, X.; Jia, R.; Wang, J.; Zhang, J.; Han, S.; and Zhang, D. 2021. TabularNet: A neural network architecture for understanding semantic structures of tabular data. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 322–331.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. *CoRR*, abs/2208.01618.

Ge, T.; Hu, J.; Dong, L.; Mao, S.; Xia, Y.; Wang, X.; Chen, S.; and Wei, F. 2022. Extensible Prompts for Language Models. *CoRR*, abs/2212.00616.

Hao, Y.; Chi, Z.; Dong, L.; and Wei, F. 2022. Optimizing Prompts for Text-to-Image Generation. *CoRR*, abs/2212.09611.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *NIPS*, 6626–6637.

Hussain, Z.; Zhang, M.; Zhang, X.; Ye, K.; Thomas, C.; Agha, Z.; Ong, N.; and Kovashka, A. 2017. Automatic Understanding of Image and Video Advertisements. In *CVPR*, 1100–1110. IEEE Computer Society.

Katja Wiemer-Hastings, K.; and Xu, X. 2005. Content differences for abstract and concrete concepts. *Cognitive science*, 29(5): 719–736.

Kumari, N.; Zhang, B.; Zhang, R.; Shechtman, E.; and Zhu, J. 2022. Multi-Concept Customization of Text-to-Image Diffusion. *CoRR*, abs/2212.04488.

Lester, B.; Al-Rfou, R.; and Constant, N. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In *EMNLP (1)*, 3045–3059. Association for Computational Linguistics.

Liao, J.; Chen, X.; and Du, L. 2023. Concept Understanding in Large Language Models: An Empirical Study.

Lin, T.; Maire, M.; Belongie, S. J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *ECCV (5)*, volume 8693 of *Lecture Notes in Computer Science*, 740–755. Springer.

Mehrabi, N.; Goyal, P.; Verma, A.; Dhamala, J.; Kumar, V.; Hu, Q.; Chang, K.; Zemel, R. S.; Galstyan, A.; and Gupta, R. 2022. Is the Elephant Flying? Resolving Ambiguities in Text-to-Image Generative Models. *CoRR*, abs/2211.12503.

Miller, G. A. 1995. WordNet: A Lexical Database for English. *Commun. ACM*, 38(11): 39–41.

Neisser, U. 2014. *Cognitive psychology: Classic edition*. Psychology press.

Nichol, A. Q.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, 16784–16804. PMLR.

Nilsback, M.; and Zisserman, A. 2008. Automated Flower Classification over a Large Number of Classes. In *ICVGIP*, 722–729. IEEE Computer Society.

Ocvirk, O. G.; Stinson, R. E.; Wigg, P. R.; Bone, R. O.; and Cayton, D. L. 1968. *Art fundamentals: Theory and practice*. WC Brown Company.

Oppenlaender, J. 2022. A Taxonomy of Prompt Modifiers for Text-to-Image Generation. *arXiv preprint arXiv:2204.13988*.

Parsons, G. 2022. The DALL·E 2 Prompt Book.

Pavlichenko, N.; and Ustalov, D. 2022. Best Prompts for Text-to-Image Models and How to Find Them. *CoRR*, abs/2209.11711.

Pressman, J. D.; Crowson, K.; and Contributors, S. C. 2022. Simulacra Aesthetic Captions. Technical Report Version 1.0, Stability AI. url https://github.com/JD-P/simulacra-aesthetic-captions .

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, 8748–8763. PMLR.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *CoRR*, abs/2204.06125.

Recchia, G.; and Jones, M. N. 2012. The semantic richness of abstract concepts. *Frontiers in human neuroscience*, 6: 315.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*, 10674–10685. IEEE.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2022. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. *CoRR*, abs/2208.12242.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.; Ghasemipour, S. K. S.; Ayan, B. K.; Mahdavi, S. S.; Lopes, R. G.; Salimans, T.; Ho, J.; Fleet, D. J.; and Norouzi, M. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *CoRR*, abs/2205.11487.

Salimans, T.; Goodfellow, I. J.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved Techniques for Training GANs. In *NIPS*, 2226–2234.

Schuhmann, C. 2022. LAION-AESTHETICS.

Schwanenflugel, P. J. 2013. *The psychology of word meanings*. Psychology Press.

Smith, E. 2022. A Traveler's Guide to the Latent Space.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset.

Wei, Y.; Zhang, Y.; Ji, Z.; Bai, J.; Zhang, L.; and Zuo, W. 2023. ELITE: Encoding Visual Concepts into Textual Embeddings for Customized Text-to-Image Generation. *CoRR*, abs/2302.13848.

Wu, C.; Liang, J.; Ji, L.; Yang, F.; Fang, Y.; Jiang, D.; and Duan, N. 2022. NÜWA: Visual Synthesis Pre-training for Neural visUal World creAtion. In *ECCV (16)*, volume 13676 of *Lecture Notes in Computer Science*, 720–736. Springer.

Wu, J.; Gan, W.; Chen, Z.; Wan, S.; and Lin, H. 2023. AI-Generated Content (AIGC): A Survey. *CoRR*, abs/2304.06632.

Xie, Y.; Pan, Z.; Ma, J.; Jie, L.; and Mei, Q. 2023. A Prompt Log Analysis of Text-to-Image Generation Systems. In *WWW*, 3892–3902. ACM.

Xu, B.; Li, T.; Zheng, J.; Naseriparsa, M.; Zhao, Z.; Lin, H.; and Xia, F. 2022. MET-Meme: A Multimodal Meme Dataset Rich in Metaphors. In *SIGIR*, 2887–2899. ACM.

Xu, T.; Zhang, P.; Huang, Q.; Zhang, H.; Gan, Z.; Huang, X.; and He, X. 2018. AttnGAN: Fine-Grained Text to Image Generation With Attentional Generative Adversarial Networks. In *CVPR*, 1316–1324. Computer Vision Foundation / IEEE Computer Society.

Ye, K.; and Kovashka, A. 2018. ADVISE: Symbolism and External Knowledge for Decoding Advertisements. In *ECCV (15)*, volume 11219 of *Lecture Notes in Computer Science*, 868–886. Springer.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguistics*, 2: 67–78.

Yu, J.; Xu, Y.; Koh, J. Y.; Luong, T.; Baid, G.; Wang, Z.; Vasudevan, V.; Ku, A.; Yang, Y.; Ayan, B. K.; Hutchinson, B.; Han, W.; Parekh, Z.; Li, X.; Zhang, H.; Baldridge, J.; and Wu, Y. 2022. Scaling Autoregressive Models for Content-Rich Text-to-Image Generation. *CoRR*, abs/2206.10789.

Zdrazilova, L.; Sidhu, D. M.; and Pexman, P. M. 2018. Communicating abstract meaning: concepts revealed in words and gestures. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1752): 20170138.

Zhang, C.; Zhang, C.; Zhang, M.; and Kweon, I. S. 2023a. Text-to-image Diffusion Models in Generative AI: A Survey. *CoRR*, abs/2303.07909.

Zhang, C.; Zhang, C.; Zheng, S.; Qiao, Y.; Li, C.; Zhang, M.; Dam, S. K.; Thwal, C. M.; Tun, Y. L.; Huy, L. L.; Kim, D. U.; Bae, S.; Lee, L.; Yang, Y.; Shen, H. T.; Kweon, I. S.; and Hong, C. S. 2023b. A Complete Survey on Generative AI (AIGC): Is ChatGPT from GPT-4 to GPT-5 All You Need? *CoRR*, abs/2303.11717.