Neural Processing Letters (2024) 56:150
https://doi.org/10.1007/s11063-024-11609-w

Efficient Visual Metaphor Image Generation Based on Metaphor Understanding

Chang Su[1] · Xingyue Wang[1] · Shupin Liu[1] · Yijiang Chen[1]

Accepted: 27 March 2024

Abstract
Metaphor has significant implications for revealing cognitive and thinking mechanisms. Visual metaphor image generation not only presents metaphorical connotations intuitively but also reflects AI's understanding of metaphor through the generated images. This paper investigates the task of generating images based on text with visual metaphors. We explore metaphor image generation and create a dataset containing sentences with visual metaphors. Then, we propose a visual metaphor generation image framework based on metaphor understanding, which is more tailored to the essence of metaphor, better utilizes visual features, and has stronger interpretability. Specifically, the framework extracts the source domain, target domain, and metaphor interpretation from metaphorical sentences, separating the elements of the metaphor to deepen the understanding of its themes and intentions. Additionally, the framework introduces image data from the source domain to capture visual similarities and generate visual enhancement prompts specific to the domain. Finally, these prompts are combined with metaphorical interpretation sentences to form the final prompt text. Experimental results demonstrate that this approach effectively captures the essence of metaphor and generates metaphorical images consistent with the textual meaning.

Keywords Metaphor understanding · Metaphorical text-to-image generation · Visual metaphor image · Prompt learning

1 Introduction

Visual metaphor is an artistic technique that uses visual elements to convey and express certain concepts, emotions, or ideas. The task of visual metaphor image generation can intuitively showcase the content conveyed by visual metaphors. J. Hessel et al. evaluated AI's understanding of humor by mapping between comic images and comic title texts[1]. The visual metaphor image generation task we study involves mapping from metaphorical text to metaphorical images, which also demonstrates AI's understanding of metaphors. Visual metaphor image generation is crucial as it not only presents visual metaphors intuitively but

✉ Chang Su
suchang@xmu.edu.cn

[1] School of Informatics, Xiamen University, Xiamen 361005, People's Republic of China

Published online: 16 April 2024

Fig. 1 Stable Diffusion image generation model with different image output generated by different prompt. **a** Directly uses the original sentence "The sunset is fire in the sky". **b** Using prompt "The sunset glow was as red as fire". **c** Is the prompt we generated "The sunset glow is red. The sunset in the style of 'burning', 'edge disturbance', 'flame decorated"'

also reflects AI's understanding of metaphors. Yuri Bizzoni et al. have preliminarily proposed modeling metaphors in the visual space by exploring the similarities and differences between misclassified metaphorical images and visually meaningful metaphors[2]. We aim to further explore visual metaphor image generation.

Large-scale text-to-image models have exhibited exceptional reasoning capabilities when provided with natural language descriptions, enabling the generation of diverse images in various styles[3–7]. These models have found application in artistic creation and design processes. Nonetheless, their utility is constrained by the user's proficiency in articulating the desired target through textual input[8]. Metaphorical expressions are challenging for generating models to directly understand their meanings. Metaphors differ from literal language as they involve mapping from a source domain (conceptual metaphor) to a target domain (target concept), encompassing abstract and non-literal meanings. An illustrative example of visual metaphor is "The sunset is like a flame in the sky." In this example, the target domain is the sunset, and the source domain is the flame, based on their similarity for metaphorical mapping. Metaphorical texts exhibit characteristics of uncertainty and ambiguity.

Due to the uniqueness of metaphorical language, existing models face challenges in understanding and capturing subtle differences in visual metaphors. Therefore, they may struggle to grasp metaphorical mapping relationships, resulting in generated images that contradict the intended meanings, as shown in Figure (a) and (b) in Fig. 1. Tuhin Chakrabarty et al. used Instruct GPT-3 (davinci-002) with Chain-of-Thought (CoT) prompts to generate visual interpretations of linguistic metaphors, using them as inputs for a diffusion model to generate visual metaphors[9]. However, this model fails to capture visual composition and imagery, and it lacks interpretability[10]. To address the issues of poor explanatory power and underutilization of visual features in metaphor image generation, we explore optimization methods for adapting metaphorical features in a multimodal model context.

Prompt optimization has gained traction as a promising method for leveraging large pretrained language models, eliminating the need for costly fine-tuning of the entire model and providing an efficient alternative. Numerous approaches have focused on optimizing soft prompts, such as continuous embedding vectors, utilizing gradient descent methods [11–13]. However, these prompts pose challenges in terms of human comprehensibility and lack compatibility across different language models (LMs) [14]. On the other hand, discrete prompts,

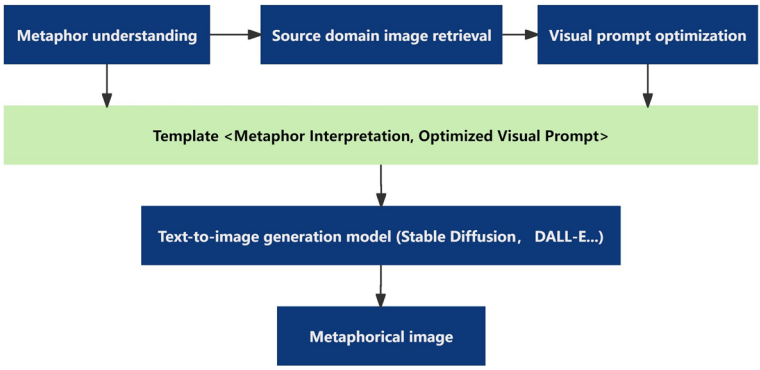

Fig. 2 Frame of our work

while more challenging to optimize, are typically generated using heuristic enumeration-then-selection methods that do not systematically explore the prompt space. Prior methods have relied on manual engineering or selecting from multiple paraphrased/generated prompts[15, 16]. However, these approaches suffer from limitations in terms of effectiveness and practical applicability. AutoPrompt[17] addresses these issues by leveraging gradient information to modify prompt tokens. However, it is susceptible to training instability and faces similar limitations as gradient-based soft prompting methods. Mingkai Deng et al. proposed an efficient approach to discrete prompt optimization employing reinforcement learning (RL)[18]. They formulate a parameter-efficient policy network that generates optimized discrete prompts through training with reward signals. Although the resulting optimized prompts may exhibit ungrammatical or nonsensical text, they retain significant performance and can be transferred between different LMs.

Considering the inherent characteristics of metaphor generation tasks, we propose an improved metaphor-related image generation framework that utilizes prompts for optimization. Our framework focuses on optimizing metaphorical text to enhance the performance of downstream models. Initially, we employ an optimization process that involves extracting the source domain, target domain, and metaphorical interpretation from the given text. This separation helps to deepen our understanding of metaphorical themes and intentions. Additionally, we introduce image data from the source domain to capture visual similarities and generate domain-specific visual enhancement prompts. These prompts are then combined with the metaphorical interpretation sentences to create the final prompt text. For an overview of our approach, please refer to Fig. 2. Our optimized prompts are specifically tailored to the characteristics of visual metaphors, and the results show a significant improvement in effectiveness. The progressive optimization approach used to generate the final text prompts allows for more flexible improvements. The main contributions of this paper are as follows:

- Proposed a visual metaphor image generation framework.
- Integrated understanding of metaphors into image generation.
- The framework introduces image data from the source domain to capture visual similarities and generate visual enhancement prompts.

Table 1 Words with higher specificity in the concreteness rating data set

Word	Specificity rating (1–5)
Sit	4.80
Hum	4.62
Clapping	4.32
Singing	4.20
Sip	4.17
Heartbeat	4.08

2 Visual Metaphor Dataset Collection

We first conducted research on the datasets related to the concept of "visuality" in current natural language studies. Brysbaert et al.[19] presented a publicly available English domain dataset: the concreteness rating dataset, which contains ratings from annotators on the concreteness of nearly 40,000 English words and phrases, with scores ranging from 1 (abstract) to 5 (concrete). In their research, Brysbaert et al. defined concreteness of a word as the extent to which it represents things that can be directly perceived through the five senses, such as "phone" and "moist", while abstract words are concepts that are farther from direct perception or can only be explained by other words, such as "joy" and "truth." Table 1 shows words with higher concreteness ratings in the concreteness rating dataset.

The Visual Genome[20] and ImageNet[21] datasets include annotations for specific objects in images, with Visual Genome providing more detailed object boundary annotations in images. The images in the ImageNet dataset mainly consist of natural scenes and images of real-world entities, while the text modality is used for labeling the entire image. Therefore, we will use the textual data from the object annotation tasks in the Visual Genome and ImageNet datasets to construct our visual metaphor corpus.

When processing the collected text data, we divide it into two parts for processing. Firstly, we extract a large number of adjectives describing objects from the dataset, which are closely related to visual attributes such as color, texture, and shape. By extracting these adjectives, we build a visual attribute word library to better understand and analyze the visual features in visual metaphors. Secondly, we extract nouns and their synonyms to build a visual object noun library. This noun library is used to match and filter metaphorical triplets defined by visual metaphors. When processing these labels, we perform a series of preprocessing steps, including removing punctuation, converting to lowercase, removing stopwords, performing stem extraction, and filtering compound words and abbreviations. Finally, through part-of-speech tagging, we separate the noun part and the modifying adjectives before the noun, automatically removing duplicate words. In the process of building the visual object noun library, to further remove noise, we also use data from the specificity rating dataset to filter out labels with low specificity in the noun library. We found a total of 11,544 object nouns in our dataset, with an average specificity score of 4.08 (rating range 0–5). We calculated the average specificity score for the remaining words in the specificity rating dataset as 2.81, so we use 2.81 as a threshold to further filter our visualized noun objects, removing those with low specificity. This step helps to improve the quality of the noun library and facilitates a more accurate identification and analysis of visual metaphors. Through the above steps, we successfully built a visual attribute word library containing 3,818 adjectives and a visual object word library containing 18,544 nouns, providing corpus data support for better understanding and applying visual metaphors in our future research.

Table 2 Examples of visual metaphor sentences

Metaphor triplets	Metaphor sentences
(Pillow, soft, cloud)	Our pillows are as soft as clouds
(River, flowing, silk)	The river is a ribbon of silk flowing smoothly
(Shadow, long, snake)	Shadows stretched out before them, long as snakes
(Hair, golden, wheat)	Her hair shimmered like golden fields of wheat in the sunlight
(Grass, green, emerald)	The grass spread out like a green field of emeralds
(Snow, white, cotton)	Snow blanketed the ground, as white as cotton

To determine whether a metaphor in the dataset is a visual metaphor, we first automatically check whether its source domain and target domain are both contained in our generated visual object noun library. If both exist, we consider the metaphor to be a visual metaphor. For example, in the metaphor "The sunset is a flame in the sky", people can find clues to "fire" by identifying visual features such as color in the image, all based on the stable visual imagery of "sunset" and "fire" in human cognition. Non-visual words such as "laughter" and "fear" will be filtered out. Words such as "summer", "wind", and "day" have some visual associations and high specificity scores in the specificity rating dataset, but they do not have stable visual imagery and do not meet the definition of visual metaphors based on the variations in human experience and perception. To avoid noise, we applied a filtering process to remove complex metaphorical sentences from multiple datasets[22, 23]. This process involved detecting sentence lengths and conducting manual annotations, resulting in a total of 216 triplets for metaphor image generation. Several examples of these triplets are provided in Table 2 for reference.

3 Visual Metaphor Image Generation Framework

In response to the characteristics of metaphor generation tasks, we adopt a strategy called prompt learning to optimize the performance of downstream models in handling metaphor-related image generation tasks at the text prompt level. We propose a prompt-optimized visual metaphor generation framework. Initially, we optimize metaphorical texts from the perspective of text modality metaphor understanding. To enhance the downstream model's understanding of metaphorical semantics, we extract the source domain, target domain, and metaphor understanding results for each metaphorical sentence. We transform the original metaphorical sentences into metaphorical interpretive sentences with the target domain as the subject and mapped attributes, such as "The sunset is red." This step separates the target domain and the attributes to be mapped explicitly from the sentence, improving the accuracy of grasping metaphorical themes and intentions. Additionally, considering that visual modality metaphor mapping has more subtle and rich visual feature connections, we believe that the textual interpretation is insufficient to fully express the visual similarity between the source domain and the target domain. To highlight the visual features of the source domain, we introduce the image data of the source domain. Based on the metaphorical interpretive text as the initial visual prompt, we generate readable source domain visual enhancement prompts specific to the visual characteristics of the source domain through discrete prompt optimization methods. Then, we concatenate them with the metaphorical interpretive sentences to form the final prompt text. The overall framework of our method is shown in Fig. 2.

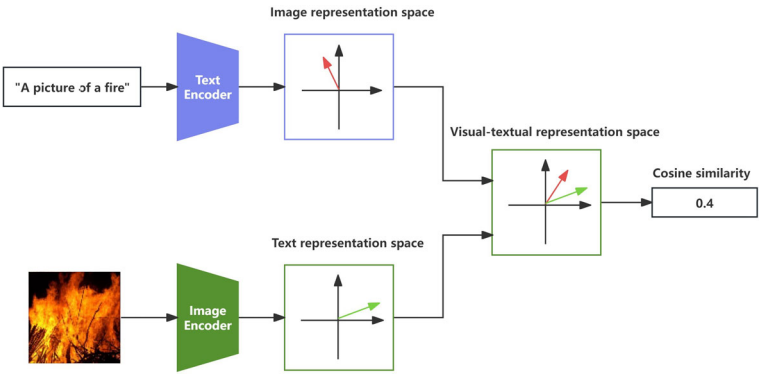

Fig. 3 Image-text similarity calculation of CLIP

Our goal is to generate readable prompts, which require optimizing the discrete space of the text, while the image information of the source domain is encoded in the continuous space. CLIP (Contrastive Language-Image Pre-Training)[24] is a multimodal pre-training model designed to project representations of images and text into the same embedding space for semantic matching, as shown in Fig. 3. CLIP utilizes Transformer architecture and contrastive learning techniques, while employing a large amount of image and text data for pre-training. It consists of two main components: a text encoder and an image encoder. The text encoder uses Transformer to encode natural language sentences into vector representations. The image encoder is a convolutional neural network that encodes images into vector representations. Taking into account the similarity-based visual mapping characteristics of visual metaphors and inspired by the optimization with discrete prompt tokens, we propose to utilize the multimodal representation space of CLIP to gradient optimize text mappings to CLIP's continuous embedding space, and then map the optimized continuous embedding sequence back to the discrete feature space. During this process, we use gradient projection during forward and backward propagation to project the computed gradients back to the discrete space while retaining the precision of storing continuous space weights and accumulated gradients. Although applying this strategy directly to language models poses a problem of not considering token positions, our prompt generation for image generation models is different from general fluent text generation. Generating visual prompts for source domain images mainly involves extracting the semantic information from the images and does not require fluent text. Therefore, we believe that this optimization strategy is suitable for our task.

First, the metaphorical interpretation sentence is generated by the text encoder of CLIP, and is mapped from the discrete feature space to the continuous feature space represented by tensors. Since the feature space of CLIP's pre-training text - image has been aligned, we take the source domain image feature as the target feature, and take the cosine similarity of the randomly sampled source domain image representation and the current prompt representation as the objective function, and carry out gradient update on the current prompt representation in the continuous feature space. In each iteration, we project the updated continuous embedding vector from the continuous feature space back to the discrete feature space, which is a discrete text token sequence. Then in the next iteration, we use the continuous feature corresponding to this text sequence and the target image feature sampled from the source domain to calculate

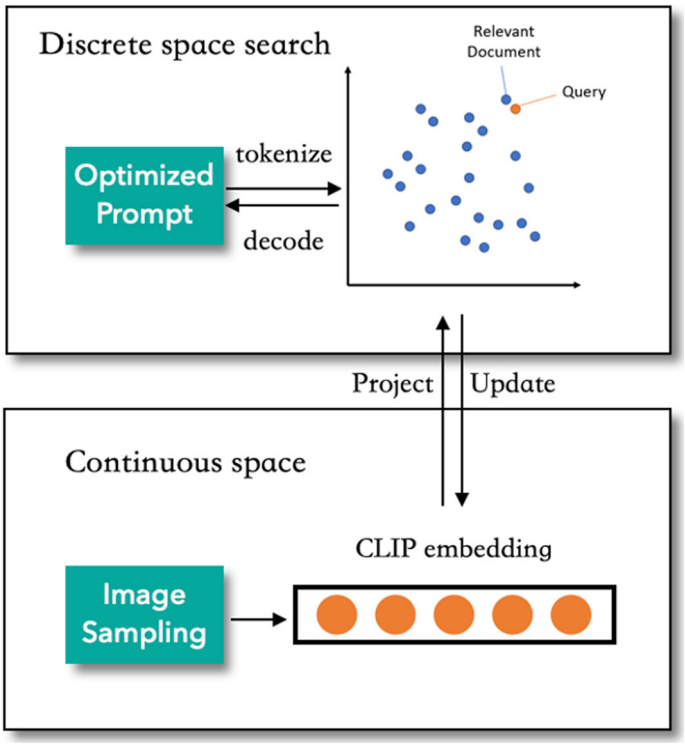

Fig. 4 Prompt Optimization based on source domain images

the cosine similarity score, and use it to update the prompt in the continuous space by gradient optimization. The visualization process is shown in Fig. 4.

Next, a set of source domain images is collected and preprocessed. All collected source domain images are encoded using the ViT image encoder of CLIP. ViT is a Transformer-based visual processing model, which is specifically designed for processing visual input using Transformer architecture. The ViT model divides the image into fixed-size regions and uses self-attention mechanism to capture global relationships in the input image.

After initializing the prompt with source domain words, we perform a limited number of iterations of gradient optimization, and randomly sample source domain image features $I_b \subseteq I$ to form a mini-batch sample (I_b, P_{curr}) in each iteration. Then, we map the continuous prompt embedding to the embedding space of the discrete vocabulary. Specifically, we use a semantic search function to find the nearest neighbor in the embedding of the vocabulary for each prompt embedding. This is a method used in natural language processing to search for text or words with similar semantic meanings in a large vector set by finding the nearest vector similar to the given query vector. Dot product is used as the similarity measurement method for text to search for the nearest neighbor. First, the current embedding and the vocabulary matrix are normalized, and the normalized dot product is equivalent to cosine similarity. Then, the dot product score between the current embedding and each vector in the vocabulary matrix is calculated, the similarity scores are sorted, and the vector with the highest score is taken as the nearest neighbor result. The identified nearest neighbor identifier corresponds to the discrete word token in the word embedding matrix:

$$P_{curr} = [e_1, e_2, e_3, \cdots, e_n], e_i \in \mathbb{R}^d \tag{1}$$

P_curr constitutes a collection of query vectors (dimension: $n \times d$) and a collection of vocabulary vectors V (dimension: $m \times d$), where n is the number of query vectors, m is the size of the vocabulary, and d is the dimension of the vectors. We first normalize the vectors in P and V, then calculate the dot product between the query vectors and the vectors in the corpus to obtain the similarity matrix S:

$$S = P_{norm} \times V_{norm}^T \tag{2}$$

Here, P_norm and V_norm represent the normalized collections of query vectors and corpus vectors, respectively, $(\cdot)$T denotes matrix transposition. For each e_i, we find the vocabulary vector with the highest similarity and its projection prompt embedding in the continuous space. We then calculate the cosine similarity between the projected prompt embedding and the target image embedding:

$$S_{\text{cosim}} = \frac{P_{\text{proj}} \cdot I_{\text{b}}}{\|P_{\text{proj}}\| \|I_{\text{b}}\|} \tag{3}$$

Our objective is to maximize the cosine similarity between the current projected prompt embedding and the target image embedding. Mathematically, this is equivalent to minimizing the negative cosine similarity. To achieve this, we optimize the process by minimizing the following objective function:

$$\mathcal{L}(\mathbf{P}, I_b) = 1 - \mathcal{S}_{cosim} \tag{4}$$

After a fixed number of iterations of gradient updates on the prompt embedding, we select the embedding vector with the highest score as the optimized source domain-enhanced prompt embedding, which exhibits semantic relevance to the source domain images. Subsequently, we decode this embedding into natural language text, which serves as the optimized source domain enhancement prompt.

During both the forward and backward propagation stages, we employ gradient re-projection to adjust the computed gradients back to the discrete space, while preserving the precision of continuous space weights and accumulated gradients. Although applying this strategy directly to language models may have issues with positional information, our text prompt generation for image generation models is different from general fluent text generation. Our focus lies in generating visual prompts that explore the semantic information of images rather than generating fluent text. Hence, we perceive this optimization strategy as suitable for our task.

4 Experiment

To assess the quality of generating metaphorical images with regards to visual metaphors, we combine both manual evaluation methods and automated metric evaluation methods. The image generation experiments in this chapter will be conducted on the Stable-Diffusion-base[3] and DALL-E[4] models to examine the effectiveness of the proposed framework for visual metaphor generation, and analyze the results accordingly.

During the source domain image collection phase, we utilize the Unsplash[1] image library. For each metaphor, we gather a set of source domain images. The method employed in this research involves using the source domain as keywords to construct query requests, invoking

[1] https://api.unsplash.com/.

the Unsplash image library's API to obtain a set of images related to the keywords, sorting them by relevance, and selecting the top N most relevant images.

As for combining the target domain and source domain templates, we conducted preliminary experiments using the Stable Diffusion model with the setting "< **target domain**> in the style of < **source domain enhanced visual prompt**>". This template yielded favorable results in combining the target and source domain content, providing a reliable foundation for our subsequent experiments. This template serves as a variable parameter in our framework, allowing for flexible replacement and adjustment based on the requirements of downstream image-text generation models, catering to the needs of different tasks.

4.1 Experimental Settings

During the phase of visual prompt generation and optimization, we employed the OpenCLIP-ViT-H/14 model. This model shares the same text encoder as our Stable-Diffusion-v2 image generation model, ensuring consistency in prompt optimization. The OpenCLIP-ViT-H/14 model was trained using an English subset of the LAION-5B dataset[25], which comprises 2 billion samples. For image encoding in our study, we utilized the Visual Transformer (ViT)[26] with the ViT-B/32 version. This version of ViT is recognized for its exceptional accuracy and parameter count, achieving an 85.8% top-1 accuracy on the ImageNet dataset.

The detailed parameters of the ViT-B/32 model we utilized are as follows: the input layer consists of 224x224 pixel RGB image data. The feature extractor comprises 12 Transformer modules, each including a multi-head self-attention mechanism and a feed-forward neural network. The multi-head self-attention mechanism consists of 12 heads, with a hidden layer dimension of 768. The outputs of these Transformer modules are flattened into a vector, serving as the feature input for the classifier. The classifier consists of a fully connected layer, with a hidden layer dimension of 3072 and an output layer dimension of the number of categories (which is 1000 in the ImageNet dataset). The entire ViT-B/32 model contains a total of 86 M parameters. In the optimization process for source domain image-enhanced guidance, we utilized a general learning rate of 0.1 and applied a weight decay of 0.1 using the AdamW optimizer[27], running for 3000 optimization steps. For the Stable-Diffusion-v2, we set the guidance scale to 8 and the number of inference steps to 50.

In the testing phase, the image generation model utilized was the Stable-Diffusion-v2 [3]. This model is designed as a generative model that converts textual input into corresponding images. It utilizes a frozen CLIP VIT-L/14 text encoder to adapt the model's output according to textual prompts. Additionally, for generalization testing, we conducted experiments using the DALL-E [4] model.

4.2 Automated Metric Evaluation Methods

To quantitatively assess the efficacy of image generation, we utilized the following two automated evaluation metrics:

- **CLIP Score**: CLIP evaluates image-text consistency using cosine similarity of features, a common method for multimodal generation tasks. This approach involves the extraction and normalization of features from both generated images and text, followed by the calculation of their dot product. Scaling the resultant value provides a correlation score that quantifies the association between the images and the metaphorical explanations. Higher CLIP scores indicate a stronger textual-image correlation.

- **P@k**: To evaluate image topic recognition, we checked if the top-k candidate results matched the correct results in the target domain. This metric calculates the percentage of samples for which the correct matching results were obtained. It quantifies the model's capacity to generate target domain-specific images and evaluate their coherence.

To broaden the scope of the cosine similarity metric utilized by CLIP, we employed a rescaling approach [28] influenced by BertScore[29]. The cosine similarity was adjusted by applying a rescaling operation with a scaling factor of 2.5.

This rescaling technique effectively alleviated the confinement of cosine similarity within the $0-0.4$ range, rendering it more appropriate for comparisons with conventional evaluation metrics.

To measure the semantic coherence between the generated images and the metaphorical target domain, we employed the ImageNet-21K[30] pre-trained Vision Transformer (ViT) model for image topic recognition. Prior to inputting the images into the model, a series of preprocessing steps were applied, including resizing, center cropping, converting to tensor format, and normalization, following the identical protocols used during model training. Subsequently, the preprocessed images were fed into the ViT model for prediction. The model's output provided the top-k predicted categories, representing the k topics with the highest probabilities. These predicted topics were then paired with the target domain using the metaphor sentence as the basis for comparison. For the image topic recognition task, each generated topic was presented in the form of a synonym set, organized according to the WordNet architecture. This arrangement obviated the necessity for synonym expansion during the matching process.

4.3 Manual Evaluation Methods

Considering that there is no absolute standard for evaluating text-to-image generation tasks, the aforementioned multimodal metrics cannot fully represent the specific visual effects of the generated images. These metrics are not entirely applicable when assessing whether the generated images captured the mapping of visual metaphors. Therefore, we introduce manual evaluation as an assessment method. Human subjective evaluation, based on human visual cognition, can directly assess the quality of generated images and the degree of correspondence between images and text, complementing the limitations of automated metrics.

For both the baseline and our method, we constructed 200 test instances consisting of pairs of images to investigate the generation results. We asked human evaluators to answer questions from three aspects: the degree of visual-textual matching, the strength of source domain association, and the degree of thematic matching. In each instance pair, evaluators were required to answer and select from the three questions based on the original metaphorical sentence pair. The collected results were from two well-versed English evaluators. The questions given to the evaluators were as follows:

- Which image better matches the theme of the target domain? (Target domain thematicity)
- Which image exhibits more prominent features from the source domain? (Source domain similarity)
- Which image better corresponds to the metaphorical sentence? (Overall visual-textual matching)

4.4 Analysis of Experimental Results

We set up a contrastive experiment to test the validity of the framework for visualizing our metaphors. In addition to using the original prompt as the baseline method, we also set up an optimization prompt ablation experiment to eliminate source-domain visual enhancement. Finally, the results generated by the original prompt, elimination of the optimization prompt based on source domain visual enhancement, and the final prompt input Stable Diffusion are compared. The experimental results are shown in Table 3(our method is abbreviated as Ours). From the experimental results, it can be seen that the optimized prompt image generated by us performs better in all indicators, especially compared with the original prompt, we increased the CLIP image similarity score by 6.8% and 7.4% in p@1 and p@5 respectively, and 8.3% in the CLIP image similarity score.

From the perspective of the ablation experiment, our method has shown an improvement of 1.9% in CLIP score and is comparable to the P@5 automation indicator when compared to the approach of eliminating the source domain visual-enhancement module. Although there was a slight decrease in the P@1 indicator, our method still had certain advantages when it comes to the human evaluation, especially the proportion of the source domain image effect voting. Given that this module is mainly aimed at enhancing the source domain image effect, and the direction examined by our designed P@k automation indicator mainly focuses on the domain theme highlight of the target domain, we believe that the decrease in this index is acceptable. In summary, the above experimental results indicate that our method has shown a certain improvement effect in enhancing the target domain theme of metaphorical images, highlighting source domain features, and conforming to metaphorical text.

From the perspective of manual evaluation, our method generated images that have obtained higher "visual similarity" scores in human evaluation. Moreover, it has interpretability: Taking "sunset is the flame in the sky" as an example, from the optimized prompt content, our final generated prompts contain very detailed and professional visual rendering prompts for fire, such as "burning", "edge disturbance", "flame decorated", and so on, which indicates that our method has captured more abundant visual features from the source domain object "fire". Partial visual metaphor image results are displayed in Fig. 5, and more results can be found in Appendix A.

The visually metaphorical image generated in Fig. 5 effectively showcases the meaning of the metaphor itself, providing people with new aesthetic experiences. In turn, this indirectly demonstrates that AI has a good understanding of metaphors. For example: The sunset is portrayed as red, giving it a burning flame-like appearance. Through this image, the metaphor "The sunset is fire in the sky" is vividly and powerfully expressed visually, allowing people to intuitively understand and experience its essence. It also indicates that AI's understanding of the metaphor of the sunset being likened to a flame is quite accurate. Regarding the visual expression of the metaphor "Snowflakes are feathers", white snowflakes interweave with feathers in the image. This visual expression not only gives the white color and texture of the snowflake feathers, but also depicts the lightness and delicacy of the snowflake. It also deepens the viewers' understanding of the relationship between snowflakes and feathers through this metaphor. This way of expression undoubtedly enhances the poetic and emotional impact of the entire picture. Through precise, vivid, and poetic visual expression, the imagery and emotions conveyed by the metaphor "snow is feather" are successfully communicated. It also indirectly reflects the similarity in the mapping between snowflakes and feathers in terms of color and texture in AI's understanding.

To evaluate the transferability of our metaphorical visual prompts to other models, we also conducted experiments on another generative model, DALL-E. When comparing the

Table 3 Experimental results

Model	CLIP score	P@1	P@5	Target domain thematicity	Source domain similarity	Overall visual-textual matching
SD-original	0.707	0.718	0.842	–	–	–
Ours w.o. Src	0.770	0.792	0.873	76%	73%	79%
Ours	0.790	0.786	0.889	78%	84%	81%

All three models take as input text to generate images using Stable-Diffusion-v2. The SD-original model directly inputs the original metaphorical sentence. The Ours model takes our optimized prompts as input. The Ours w.o. Src model takes the optimized prompts that remove the visual enhancement based on the source domain as input

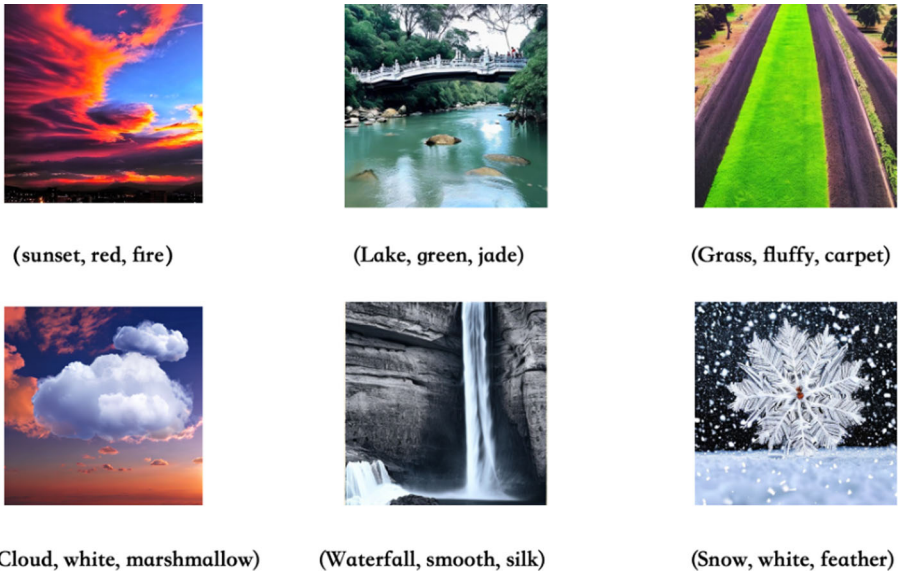

Fig. 5 The metaphor generation results are shown, and the images are labeled as metaphor triplets

Table 4 Universal experimental results

Model	CLIP score	P@1	P@5
Stable Diffusion	0.707–**0.790**	0.718–**0.786**	0.840–**0.889**
DALL-E	0.748–**0.762**	0.734–**0.784**	0.892–**0.905**

The bold part is the result after optimization

results of the original metaphorical phrase input and the results generated by our optimized prompts for the input of the original metaphorical phrase, the experimental results are shown in Table 4, where the bold part represents the optimized results.

According to the experimental results, the metaphorical images generated after optimizing the text prompts for the DALL-E model have also shown some improvement compared to the original metaphorical sentence input, which proves that our optimized prompts are also applicable to DALL-E. However, compared with the Stable Diffusion model, the performance improvement of the DALL-E model on all indicators is relatively limited. This may be due to the fact that during the text prompt optimization stage, we used the CLIP ViT-L/14 text encoder, while the Stable Diffusion model's text-to-image generation process also uses the same frozen text encoder for diffusion adjustment, which makes the text prompts generated by us more suitable for the Stable Diffusion model and thus perform better on the Stable Diffusion model.

5 Conclusion

We propose an interpretable framework tailored for generating visual metaphors. The framework first models the mapping of source domain and target domain for metaphor understanding, and then optimizes the prompts based on visual prompts. Specifically, it separates

metaphorical elements for deeper understanding and gradually enhances the source domain description based on the similarity between image embeddings and prompt representations. This multimodal prompt, which is modeled and optimized based on metaphor understanding, effectively captures the complexity in metaphorical expressions. The final text prompts are generated in a progressively optimized manner, allowing for flexible improvements. The utilization of visual information from the source domain is also a major innovation in this paper.

The proposed framework bridges metaphorical language and visual perceptions, allowing intuitive appreciation of the thought processes and aesthetics. Both quantitative evaluations and sample cases demonstrate noticeable improvements in preserving semantic consistency and realization of the desired imagery. This establishes connections between the comprehension of figurative expressions and the construction of relevant visual representations. Our work highlights new possibilities in explainable AI and multimodal understanding of abstract concepts. The cross-modal generation process reflects a deeper understanding of the intrinsic characteristics of metaphors. Moving forward, expanding the scope and diversity of the metaphor dataset, as well as exploring advanced fusion methods, remain promising directions. We hope this research provides useful inspirations for human-centric AI and transparency in complex cognitive tasks.

5.1 Limitations

In the metaphorical text-to-image generation task, our research primarily focuses on generating images based on visually similar mappings of metaphors, which is relatively limited in scope. In the future, there is a need to explore more diverse and creative types of metaphors. For example: tactile metaphors, auditory metaphors, conceptual metaphors, structural metaphors, and so on. Additionally, our current experiments are based on a limited dataset. Expanding the dataset further is also a direction we are working towards. Furthermore, to enhance the model's generalization ability and stability, we can expand the dataset to cover more types of metaphors as well as text and image data from multiple domains. In the future, we will explore multimodal deep learning methods that can better integrate text and image features and improve image generation models.

Data Availability Data will be made available on request.

Declarations

Conflict of interest The authors declare that they have no Conflict of interest.

Appendix A: Metaphorical Visual Images Generated by Stable Diffusion

See Figs. 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19.

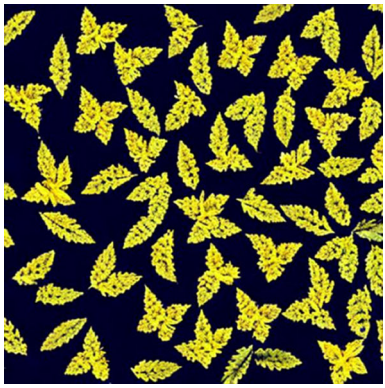

Fig. 6 Leaves are falling butterflies

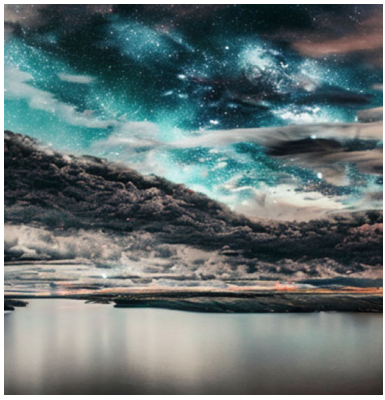

Fig. 7 The night sky is agate

Fig. 8 The lake water is transparent glass

Fig. 9 The bread is as hard as a rock

Fig. 10 The stars in the sky are as numerous as sand

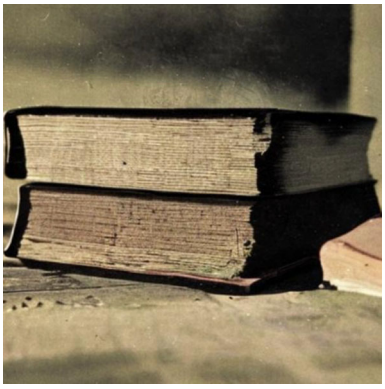

Fig. 11 Books are as thick as bricks

Fig. 12 Rain is a curtain of silver

Fig. 13 The sea is a calm mirror

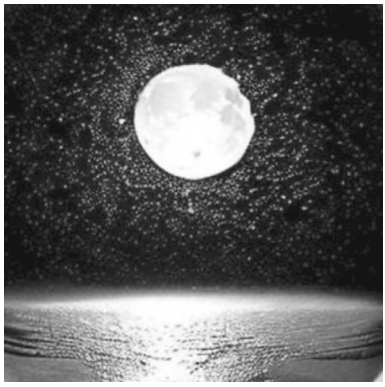

Fig. 14 The moon is a pearl in the night sky

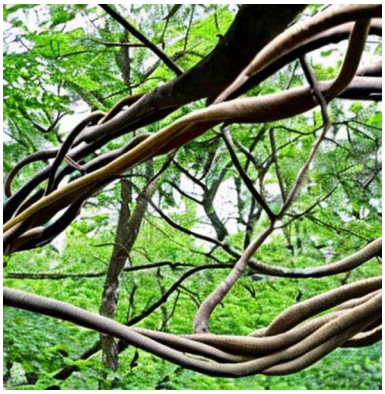

Fig. 15 The branches are like twisted ropes

Fig. 16 The roots dig into the earth like anchors

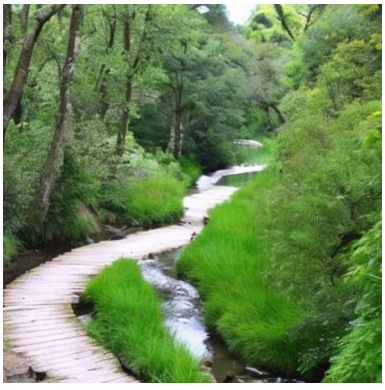

Fig. 17 The path is like a winding river

Fig. 18 Dewdrops are glittering tears

Fig. 19 The leaves under my feet are as crisp as cookies

References

1. Hessel J, Marasović A, Hwang JD, Lee L, Da J, Zellers R, Mankoff R, Choi Y (2023) Do androids laugh at electric sheep? Humor "understanding" benchmarks from the new yorker caption contest
2. Yuri B, Simon D (2020) Sky + fire = sunset. exploring parallels between visually grounded metaphors and image classifiers. In: Beigman KB, Ekaterina S, Patricia L, Smaranda M, Chee W, Anna F, Debanjan G (eds) Proceedings of the second workshop on figurative language processing, pp 126–135, Online. Association for Computational Linguistics
3. Robin R, Andreas B, Dominik L, Patrick E, Björn O (2022) High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10684–10695
4. Aditya R, Mikhail P, Gabriel G, Scott G, Chelsea V, Alec R, Mark C, Ilya S (2021) Zero-shot text-to-image generation. In: International conference on machine learning, pp 8821–8831. PMLR
5. Alex N, Prafulla D, Aditya R, Pranav S, Pamela M, Bob M, Ilya S, Mark C (2022) Glide: Towards photorealistic image generation and editing with text-guided diffusion models arxiv:2205.13168v1
6. Yu J, Xu Y, Koh JY, Luong T, Baid G, Wang Z, Vasudevan V, Ku A, Yang Y, Ayan BK, Hutchinson B (2022) Scaling autoregressive models for content-rich text-to-image generation
7. Saharia C, Chan W, Saxena S, Li L, Whang J, Denton EL, Ghasemipour K, Gontijo Lopes R, Karagol Ayan B, Salimans T, Ho J (2022) Photorealistic text-to-image diffusion models with deep language understanding
8. Gal R, Alaluf Y, Atzmon Y, Patashnik O, Bermano AH, Chechik G, Cohen-Or D (2022) An image is worth one word: Personalizing text-to-image generation using textual inversion

9. Chakrabarty T, Saakyan A, Winn O, Panagopoulou A, Yang Y, Apidianaki M, Muresan S(2023) I spy a metaphor: large language models and diffusion models co-create visual metaphors
10. Adadi A, Berrada M (2018) Peeking inside the black-box: a survey on explainable artificial intelligence (xai). IEEE Access 6:52138–52160
11. Gu Y, Han X, Liu Z, Huang M (2022) Ppt: pre-trained prompt tuning for few-shot learning
12. Qian J, Dong L, Shen Y, Wei F, Chen W (2022) Controllable natural language generation with contrastive prefixes
13. An S, Li Y, Lin Z, Liu Q, Chen B, Fu Q, Chen W, Zheng N, Lou J-G (2022) Input-tuning: adapting unfamiliar inputs to frozen pretrained models
14. Khashabi D, Lyu S, Min S, Qin L, Richardson K, Welleck S, Hajishirzi H, Khot T, Sabharwal A, Singh S, Choi Y (2022) Prompt waywardness: the curious case of discretized interpretation of continuous prompts
15. Petroni F, Rocktäschel T, Lewis P, Bakhtin A, Wu Y, Miller AH, Riedel S (2019) Language models as knowledge bases?
16. Schick T, Schütze H (2021) Exploiting cloze questions for few shot text classification and natural language inference
17. Shin T, Razeghi Y, Logan IV RL, Wallace E, Singh S (2020) Autoprompt: eliciting knowledge from language models with automatically generated prompts
18. Deng M, Wang J, Hsieh C-P, Wang Y, Guo H, Shu T, Song M, Xing EP, Hu Z (2022) Rlprompt: optimizing discrete text prompts with reinforcement learning
19. Brysbaert M, Warriner AB, Kuperman V (2014) Concreteness ratings for 40 thousand generally known english word lemmas. Behav Res Methods 904–911
20. Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, Chen S, Kalantidis Y, Li LJ, Shamma DA, Bernstein MS (2017) Visual genome: Connecting language and vision using crowdsourced dense image annotations. Int J Comput Vis 123:32–73
21. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition
22. Li Y, Lin C, Guerin F (2022) Cm-gen: a neural framework for Chinese metaphor generation with explicit context modelling. In: Proceedings of the 29th international conference on computational linguistics, pp 6468–6479
23. Chang S, Huang S, Chen Y (2017) Automatic detection and interpretation of nominal metaphor based on the theory of meaning. Neurocomputing 219:300–311
24. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, et al (2021) Learning transferable visual models from natural language supervision. In: International conference on machine learning, pp 8748–8763. PMLR
25. Schuhmann C, Beaumont R, Vencu R, Gordon C, Wightman R, Cherti M, Coombes T, Katta A, Mullis C, Wortsman M, et al (2022) Laion-5b: an open large-scale dataset for training next generation image-text models. arXiv preprint arXiv:2210.08402
26. Dosovitskiy Alexey, Beyer Lucas, Kolesnikov Alexander, Weissenborn Dirk, Zhai Xiaohua, Unterthiner Thomas, Dehghani Mostafa, Minderer Matthias, Heigold Georg, Gelly Sylvain, et al (2020) An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929
27. Loshchilov I, Hutter F (2017) Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101
28. Hessel J, Holtzman A, Forbes M, Bras RL, Choi Y (2021) Clipscore: a reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718
29. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y (2019) Bertscore: evaluating text generation with bert. arXiv preprint arXiv:1904.09675
30. Ridnik T, Ben-Baruch E, Noy A, Zelnik-Manor L (2021) Imagenet-21k pretraining for the masses