

I Spy a Metaphor: Large Language Models and Diffusion Models Co-Create Visual Metaphors

Tuhin Chakrabarty^{1*}, Arkadiy Saakyan^{1*}, Olivia Winn^{1*}, Artemis Panagopoulou²
Yue Yang², Marianna Apidianaki², Smaranda Muresan¹

¹Columbia University ²University of Pennsylvania
{tuhin.chakr,a.saakyan,olivia}@cs.columbia.edu

Abstract

Visual metaphors are powerful rhetorical devices used to persuade or communicate creative ideas through images. Similar to linguistic metaphors, they convey meaning implicitly through symbolism and juxtaposition of the symbols. We propose a new task of generating visual metaphors from linguistic metaphors. This is a challenging task for diffusion-based text-to-image models, such as DALL-E 2, since it requires the ability to model implicit meaning and compositionality. We propose to solve the task through the collaboration between Large Language Models (LLMs) and Diffusion Models: Instruct GPT-3 (davinci-002) with Chain-of-Thought prompting generates text that represents a visual elaboration of the linguistic metaphor containing the implicit meaning and relevant objects, which is then used as input to the diffusion-based text-to-image models. Using a human-AI collaboration framework, where humans interact both with the LLM and the top-performing diffusion model, we create a high-quality dataset containing 6,476 visual metaphors for 1,540 linguistic metaphors and their associated visual elaborations. Evaluation by professional illustrators shows the promise of LLM-Diffusion Model collaboration for this task. To evaluate the utility of our Human-AI collaboration framework and the quality of our dataset, we perform both an intrinsic human-based evaluation and an extrinsic evaluation using visual entailment as a downstream task.

1 Introduction

Visual metaphors are rhetorical devices that serve to communicate a message through an image. They are often used as a means of persuasion in advertising (Phillips, 2003; Phillips and McQuarrie, 2004), as their use leads to more favorable attitude toward the ad (McQuarrie and Mick, 1999). Similarly to linguistic metaphors (Lakoff, 1993), a visual metaphor takes a concept from a source domain

*Equal contribution.



Figure 1: Visual metaphors generated by DALL-E 2 for the linguistic metaphor “My bedroom is a pig sty”. We can take the original verbal metaphor as the input (left) or use GPT-3 with Chain of Thought prompting (right).

and applies it to a target domain. In the case of visual metaphors, these domains need to be in some way visually grounded.

Large diffusion-based text-to-image models, such as DALL-E 2 (Ramesh et al., 2022a), PARTI (Yu et al., 2022), Stable Diffusion (Rombach et al., 2022), or IMAGEN (Saharia et al., 2022), can generate visually compelling images conditioned on input texts. However, in order to generate visual metaphors from linguistic metaphors, models are required to first identify the implicit meaning as well as the objects, properties, and relations involved, and then find a way to combine them in the generated image. For instance, given the linguistic metaphor “*My bedroom is a pig sty*”, as

shown in Figure 1, a model would ideally need to extract the implicit meaning of the bedroom being “messy”, and then compose the concepts “Bedroom”, “Messy” & “Pig”. However, as shown in the left two images, when presented just with the linguistic metaphor, DALL·E 2 generates images of a bedroom where pink is the prevalent color (perhaps due to pig’s skin color), sometimes with the presence of a pig as a toy in a corner, and with little indication of a mess in the room.

The visual metaphor generation task is greatly impacted by two common challenges in text-to-image models, namely *under-specification* and *attribute-object binding* (Hutchinson et al., 2022; Ramesh et al., 2022a; Saharia et al., 2022). Under-specification refers to the fact that finite and reasonable-length linguistic descriptions of real-world scenes by necessity omit a great deal of visual information (Hutchinson et al., 2022). Attribute Binding is the task of binding the attributes to the correct objects, and is a fundamental problem for a more complex and reliable compositional generalization. Our proposed contributions address these challenges:

- **A novel approach for generating visual metaphors through the collaboration of large language models (LLMs) and diffusion-based text-to-image models.** Our LLM — Instruct GPT-3 (davinci-002) (Ouyang et al., 2022) with **Chain-of-Thought** (CoT) prompting (Wei et al., 2022) — generates a **visual elaboration** of the linguistic metaphors. To design our CoT prompting elements, we take inspiration from prior work on VisualBlends (Chilton et al., 2019) that put an emphasis on the objects to be represented in the visual metaphor. In addition, we also consider the implicit meaning to finally generate a visual elaboration that contains the essential objects and the implicit meaning of the linguistic metaphor. For our linguistic metaphor “*My bedroom is a pig sty*”, the visual elaboration generated by Instruct GPT-3 (davinci-002) with CoT prompting is “*A bedroom with clothes and garbage everywhere with a pig in the center rooting around.*” (See Table 1). The generated visual elaboration becomes the input to diffusion-based text-to-image models such as DALL·E 2 or Stable Diffusion to generate visual metaphors (see Figure 1 right).
- **A high-quality visual metaphor dataset built through Human-AI collaboration.** We propose a collaboration between humans, LLM,

and the top-performing diffusion-based model (DALL·E 2) to create a high-quality dataset of **6,476** visually metaphoric images. These represent **1,540** distinct linguistic metaphors and their associated visual elaborations generated through CoT prompting. We call our dataset **HAIVMet** (**H**uman-**A**I **V**isual **M**etaphor) (Section 3).

- **A thorough evaluation of LLM-Diffusion Model collaboration and Human-AI collaboration.** In order to evaluate the power of LLM-Diffusion Model collaboration, we recruit professional illustrators and designers and ask them to compare the output of DALL·E 2 and Stable Diffusion v2.1 when the input corresponds to the linguistic metaphor alone, or to the LLM-produced visual elaboration. Our evaluation shows the power of the LLM-Diffusion Model collaboration and the superiority of DALL·E 2 compared to Stable Diffusion v2.1 (Section 4.1). To evaluate the utility of Human-AI collaboration and the quality of our dataset, we perform an intrinsic evaluation using the same expert evaluators and an extrinsic evaluation using a downstream task (Section 4.2). For the latter, we choose the Visual Entailment task: given an image and a hypothesis sentence, the model is asked to predict whether the sentence is implied by the image. We show that fine-tuning a state-of-the-art vision-language model on our dataset leads to ~23-points improvement in accuracy compared to when it is only finetuned on SNLI-VE (Xie et al., 2019), a large-scale visual entailment dataset.

We release our dataset, code, prompts, and illustrator annotations at <https://github.com/tuhinjubcse/VisualMetaphors>.

2 Related Work

Generative Art. There has recently been a huge surge of AI-generated artwork and imagery with the new diffusion-based models being substantially better than previous Variational Autoencoders (VAE) and Generative Adversarial Networks (GANs). Some of the most popular current models are DALL·E 2 (Ramesh et al., 2022b), MidJourney,¹ Craiyon,² and Stable Diffusion (Rombach et al., 2021). These image generation models are able to handle a wide variety of prompts, though recent work has shown that there are still

¹<https://www.midjourney.com/>

²<https://www.craiyon.com/>

aspects of accurate depiction that these models fail to capture (Leivada et al., 2022). Recently, Kleinlein et al. (2022) showed that diffusion models can handle language that is content-based and aimed at a neutral description of the scene, and fail to capture the underlying abstraction of figurative language. Recent work has also explored cutting-edge systems showcasing the power of large language models and text-to-image models in aiding creative processes across various applications. Wang et al. (2023) present PopBlends, a system that leverages traditional knowledge extraction methods and large language models to automatically generate conceptual blends for pop culture references, significantly increasing the number of blend suggestions while reducing mental demand for users. Similarly, Liu et al. (2023) introduce Generative Disco, an AI system that generates music visualizations using large language models and text-to-image models, offering an enjoyable, expressive, and easy-to-use tool for professionals in the creative field. Wang et al. (2023) present ReelFramer, a system where GPT4 and DALLE2 collaborate in order to assist journalists in transforming written news stories into engaging short video narratives, by generating scripts, character boards, and storyboards. The proposed user study shows ReelFramer’s effectiveness in easing the process and making framing exploration rewarding for journalism students.

Visual Metaphor. Visual metaphors are often abstract and can be challenging to interpret. Petridis and Chilton (2019) test several theories about how people interpret visual metaphors. They find that visual metaphors are interpreted correctly, without explanatory text, with 41.3% accuracy. Indurkha and Ojha (2013) highlight the important role of perceptual similarity between the source and the target image (in terms of color, shape, etc) in metaphor comprehension and creative interpretation.

Achlioptas et al. (2021) propose the ArtEmis dataset which contains emotion attribution and explanation annotations for 80K artworks from WikiArt, including several visual metaphors and similes. Their dataset serves to train captioning systems to express emotions and associated explanations derived from visual stimuli, instead of generating images conditioned on text. Zhang et al. (2021) collect a multimodal metaphor dataset from Twitter posts and advertisement posters that contain a metaphor in the caption, in the image, or both. However, they do not generate any new data

and, as of yet, the data has not been publicly released. Liu et al. (2022b) release Opal, a system that guides users in generating diverse and relevant text-to-image illustrations for news articles by utilizing structured exploration. Unlike research on generating textual metaphors (Yu and Wan, 2019; Chakrabarty et al., 2020, 2021; Veale, 2016; Abe et al., 2006; Terai and Nakagawa, 2010), visual metaphor generation has received less attention. Akula et al. (2023) proposed MetaCLUE, a set of vision tasks that serve to evaluate the metaphor understanding and generation capabilities of state-of-the-art vision and language models. Their results show that most tested state-of-the-art models struggle to produce satisfactory results, in both a zero-shot and a finetuning setting. Hwang and Shwartz (2023) focus on building a dataset for captioning and interpreting memes that are a widely popular tool for web users to express their thoughts using visual metaphors. More recently, Yosef et al. (2023) present the Image Recognition of Figurative Language dataset, designed to evaluate vision and language models’ understanding of figurative language, including metaphors, similes, and idioms. The dataset features multimodal examples and introduces two novel benchmark tasks, aimed at promoting the development of models that can effectively comprehend figurative language. Current baseline models have shown significantly poorer performance compared to human understanding, highlighting the challenges this domain poses for machine learning.

3 Human-AI Collaboration for Visual Metaphor Dataset Creation

We propose a three-step Human-AI collaboration approach for generating visual metaphors from linguistic metaphors. This process involves 1) selecting linguistic metaphors that are visually grounded; 2) using large language models to generate visual elaborations of linguistic metaphors that capture relevant objects and implicit meaning, with expert edits when required; 3) using diffusion-based models to generate visual metaphors from visual elaborations, with filtering of low quality samples by experts. A detailed pipeline diagram for our dataset creation is shown in Figure 2.

We source our linguistic metaphors from six resources, removing any duplicates: **FLUTE** (Chakrabarty et al., 2022b), **Advertisements** (Hussain et al., 2017), **CoPoet** (Chakrabarty

et al., 2022a), FigQA (Liu et al., 2022a), Figure-of-Speech,³ CrossLing Metaphors (Tsvetkov et al., 2014) and Metaphor Paraphrase (Bizzoni and Lappin, 2018).

<p><i>Your task will be to elaborate a metaphor with rich visual details along with the provided objects to be included and implicit meaning. Make sure to include the implicit meaning and the objects to be included in the explanation</i></p>	
1. Metaphor: My lawyer is a shark.	<p>Objects to be included: Lawyer, Shark</p> <p>Implicit Meaning: fierce</p> <p>Visual elaboration: A shark in a suit with fierce eyes & a suitcase & a mouth open with pointy teeth.</p>
2. Metaphor: I've reached my boiling point.	<p>Objects to be included: Person, Boiling Pot</p> <p>Implicit Meaning: anger</p> <p>Visual elaboration: A boiling pot of water with a person's head popping out of the top, steam coming out of their ears, and an angry expression on their face.</p>
3. Metaphor: Joe: that's because you're like a snail surfing on molasses.	<p>Objects to be included: Person like a snail, Snail on molasses</p> <p>Implicit Meaning: slow</p> <p>Visual elaboration: A person with a snail shell on their back slowly sliding down a hill of molasses.</p>
4. Metaphor: Absence is the dark room in which lovers develop negatives	<p>Objects to be included: Darkroom, Negative Film Strip with a red heart, Person</p> <p>Implicit Meaning: ominous and lonely</p> <p>Visual elaboration: An ominous dark room with a film strip negatives hanging and a red heart in the center with a person in the corner looking sad and lonely</p>
5. Metaphor: My heart is a rose thorn	<p>Objects to be included: Heart, Thorn</p> <p>Implicit Meaning: prickly</p> <p>Visual elaboration: A heart with a prickly thorn coming out of the center and barbs going outwards.</p>
6. Metaphor: My bedroom is a pig sty	<p>Objects to be included: Messy bedroom, Pig</p> <p>Implicit Meaning: dirty</p> <p>Visual elaboration: A bedroom with clothes & garbage everywhere with a pig in the center rooting around.</p>

Table 1: Chain-of-Thought (CoT) prompt to elicit a visual elaboration for a given metaphor. We provide the first five examples in a few-shot learning setting and the model jointly generates Objects to be Included, Implicit Meaning, and Visual elaboration (highlighted in brown) step-by-step.

1) Visually Grounded Linguistic Metaphors:

Given that not all linguistic metaphors can be rendered as visual metaphors, we manually select those that are visually grounded. Concrete subjects can clearly be visually grounded, but some abstract subjects can be visually grounded as well through their usual representations in media. For

³<https://www.kaggle.com/datasets/varchitalalwani/figure-of-speech>

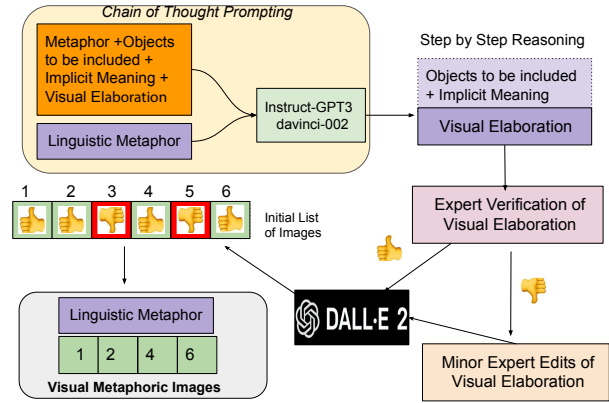


Figure 2: Human-AI collaboration framework (LLMs-Diffusion Model-Humans). Instruct GPT-3 with CoT prompting generates visual elaborations from linguistic metaphors, which are then validated and possibly edited by humans, if necessary. Visual elaborations are then used as input to DALL-E 2 to generate visual metaphors. Experts filter poor-quality visual metaphors. For example, images 3 and 5 in the figure are discarded by experts.

example, “love” can be represented as two people holding hands with hearts above them, “confusion” as question marks, or “idea” as a lightbulb over someone’s head. Linguistic metaphors that describe non-visual phenomena (e.g., a smell, a sound) are removed unless the act of experiencing the sense is the subject of the sentence, which can be visualized with, e.g., a facial expression. We consider emotional phenomena as visual since often emotions and feelings are expressed through facial expression and/or body posture which can be visualized.

2) Visual Elaboration Generation with Chain-of-Thought Prompting:

Existing text-to-image generation models do not perform well when their input contains linguistic metaphors, since they lack the ability to model implicit meaning and compositionality. Recently, Wei et al. (2022) proposed a prompting method for improving the reasoning abilities of language models. This method, called **Chain-of-Thought** (CoT) prompting, enables models to decompose multi-step problems into intermediate steps. We take advantage of CoT prompting by using the relevant objects and implicit meaning of the metaphors as our intermediate steps, to then elicit detailed textual visualizations of linguistic metaphors using Instruct GPT-3 (davinci-002). We refer to this detailed textual visualization as a **visual elaboration**. We hypothesize that these visual elaborations obtained from CoT prompting

will help text-to-image models create better visual metaphors, as the objects and implicit meaning will be explicitly contained in the input.

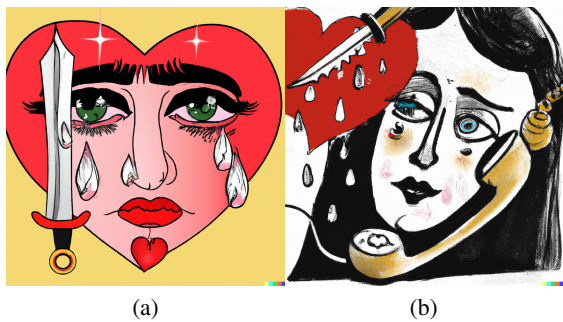


Figure 3: Visual metaphors obtained using DALL-E 2 for the linguistic metaphor “*The news of the accident was a dagger in her heart*”. The image on the left is obtained using the visual elaboration “*An illustration of a heart with a dagger stuck into it, dripping with blood and pain in the woman’s eyes*”, and the one on the right using the edited prompt “*An illustration of a woman receiving a phone call and her heart with a dagger stuck into it, dripping with blood and pain in the woman’s eyes.*”

Table 1 shows the instruction and CoT prompt used to elicit a visual elaboration for a given linguistic metaphor. The first five examples are given as few-shot examples and the model (Instruct GPT-3 (davinci-002)) then jointly generates the objects to be Included, implicit meaning, and visual elaboration (highlighted in brown) step-by-step. As our prompts follow a certain structure for step-by-step reasoning, a zero-shot approach would not work well. We found that using five few-shot examples was sufficient to generate elaborations of good quality. We selected five representative examples of visualizable metaphors for the prompt. We used the same examples for generation of every visual elaboration.

While this approach leads to good-quality outputs, not all generated visual elaborations are perfect. We recruit three expert annotators with multiple years of experience in figurative language research and ask them to validate the generated visual elaborations and to slightly edit them if needed, in order to make sure they accurately represent the implicit meaning and the objects involved. Our pipeline is illustrated in Figure 2. As can be seen in Figure 3, for the given linguistic metaphor “*The news of the accident was a dagger in her heart*”, the first visual elaboration is almost correct but it misses the crucial information about the

metaphoric source, i.e. “*the news of the accident*”. An expert performs a minor edit by adding the phrase “*woman receiving a phone call*” in order to convey the metaphoric source which leads to a perfect visual metaphor. Experts performed minor edits on 29% of the generated visual elaborations.

3) Visual Metaphor Generation using Diffusion-based Models and Human Quality Check: For this part of the data curation process, we first prompt DALL-E 2 to generate multiple images⁴ for a single visual elaboration (cf. Figure 2). Post-generation, each set of generated images is examined jointly by three experts to determine whether they accurately and fully represent the meaning of the original linguistic metaphor. The experts need to validate whether the image contains the relevant objects and whether the objects are positioned correctly or have the appropriate indicators of movement or action, also referred to as **Attribute Binding** (Ramesh et al., 2022a; Saharia et al., 2022). For example, for the phrase “*Her eyes were like peonies*”, the image would need to depict both a face and peonies and the peonies would need to be in the place of the eyes rather than around the head (which was the case in some images). Images that do not meet the above criterion were discarded.

The dataset curated in this way contains **1,540** unique linguistic metaphors (and their associated visual elaborations) and **6,476** unique images. Each linguistic metaphor has **four** associated images, on average. We call our data **HAIVMet (Human-AI Visual Metaphor)**.

4 Evaluation

Our goal is to assess the impact of the LLM-Diffusion Model collaboration (Section 4.1), and of the Human-AI collaboration on building a high-quality dataset.

4.1 LLM-Diffusion Model Collaboration

Models. Diffusion models are trained to recover the original version of an image after random noise has been applied to it (Ramesh et al., 2022a). Both DALL-E 2 and Stable Diffusion are diffusion-based text-to-image models. Stable Diffusion is open source; DALL-E 2 is not. Note that in this evaluation, there is no human intervention (no editing of the output of Instruct GPT-3 with CoT prompting, nor filtering of images produced by diffusion-based

⁴DALL-E 2 automatically generates four images per prompt.

models). We use the following LLM-Diffusion Model collaboration setups, where the input to the diffusion models is the visual elaboration of the linguistic metaphor generated using Instruct GPT-3 (davinci-002) with CoT prompting:

- **LLM-DALL·E 2:** DALL·E 2 (Ramesh et al., 2022a) with the LLM-generated visual elaboration as input.
- **LLM-SD:** The Stable Diffusion (Rombach et al., 2022) v2.1 model, with the same input as LLM-DALL·E 2.
- **LLM-SD_{Structured}** We use the diffusion method of Feng et al. (2022) which combines the structured representations of prompts (for example, their constituency tree) with the diffusion guidance process, using the same input as LLM-DALL·E 2.

We also use DALL·E 2 and Stable Diffusion (SD) with the linguistic metaphor given directly as input (no collaboration with the LLM). This comparison allows us to assess the benefit that can be drawn from LLM-Diffusion Model collaboration.

Human Evaluation Setup. Among the popular automatic evaluation metrics, both Fréchet Inception Distance (FID) (Heusel et al., 2017) and CLIP (Radford et al., 2021) scores are not tailored towards metaphorical images, and are not reliable in assessing whether the generated images capture the essence of visual metaphors (Akula et al., 2023). We also chose not to rely on non-expert crowdworkers as even with training they have been found to be unreliable for open-ended tasks (Karpinska et al., 2021). Following the recommendation from Karpinska et al. (2021), we recruit three professional artists with experience in concept illustration and visual arts through the Upwork⁵ platform. We ask them to evaluate the visual metaphors that are generated by the five approaches described above for a subset of 100 randomly selected linguistic metaphors from our dataset. For each metaphor, we ask to rank the five generated images on the basis of how well they represent the metaphor.

Additionally, we collect targeted feedback by asking the raters to provide natural language instructions for improving the images. Five text fields are shown under each image, and the annotators are invited to make up to five recommendations. In the occasional case where the image is “Perfect”

⁵<https://www.upwork.com>

or absolutely not worthy of transformation (“Lost Cause”), the annotators do not need to provide any feedback for improvement. The suggested types of instructions are the following: 1) Add an object; 2) Remove an object; 3) Move an object; 4) Replace an object with another object; 5) Change an object’s property (e.g., color, size). The annotators are encouraged to supply whatever type of change they believe is required to improve the visual metaphor; the only stipulation to the instructions is that each one must denote a single action/change. We identify the average rank assigned to a model across metaphors and annotators. We also report the percentage of “Lost Cause” cases in order to identify systems that generate the least amount of bad images. Additionally, we compare the models on the basis of the average number of instructions that have been proposed for improving their produced images. The number of suggested changes acts as a proxy for how close the image is to the perfect representation of the metaphor. “Perfect” images are considered to have 0 edits, and images that are a “Lost Cause” are considered to have 5 edits to ensure fairness in this computation.

Model	Avg Rank	% Lost Cause	Avg # of Instructions
SD	3.82	31.6	2.25
LLM-SD	3.40	23.3	1.83
LLM-SD _{Structured}	3.05	18.3	1.57
DALL·E 2	2.76	16.6	1.44
LLM-DALL·E 2	1.96	6.0	0.76

Table 2: Human evaluation results: the average ranking given by three human raters to the output of each model for 100 test metaphors; the percentage of images labeled as “Lost Cause”; and the average number of edits needed to make the image perfect otherwise.

4.1.1 Results and Analysis

Table 2 shows that without collaboration with a LLM (i.e., just with the linguistic metaphor as input), DALL·E 2 performs better than SD (line 4 vs. line 1). The main take away is that LLM-Diffusion Model collaboration outperforms simple Diffusion Models (LLM-DALL·E 2 vs. DALL·E 2, LLM-SD and LLM-SD_{Structured} vs. SD). That is, using Instruct-GPT3 with CoT prompting to produce visual elaborations as input to diffusion models consistently improves the performance over providing the diffusion models directly with linguistic metaphors. Overall, LLM-DALL·E 2 emerges as the best system. Only 6% are “Lost Cause” images, affirming our choice for using

Metaphor	HAIVMet [Gold]	LLM-Dall·E 2	LLM-SD	LLM-SD-Structured	Dall·E 2	SD
<i>My whole mind is a leaking black hole</i>						
<i>I feel like a lily in February</i>						
<i>Books are the mirror to the soul</i>						

Figure 4: Examples of output from each model described in Section 4.1 for three randomly chosen metaphors. **HAIVMet** is our gold standard. More examples are available in Figure 8 in the Appendix.

From its blue vase the rose of evening drops



1. Change background to sunset colors

He was like a butterfly in autumn, waiting to be destroyed by the first frost



1. Add male figure to background

I'm like a drug user being searched



1. Put arms down
2. Change expression to sadness

It was a moonless night, the air was still and the crickets were like living shadows



1. Remove the moon
2. Add shadow to crickets

Figure 5: Experts suggested image edits in the form of natural language instructions for images generated from the CoT visual elaborations of linguistic metaphors.

LLM-DALL·E 2 to create **HAIVMet**. Rank 1 (best) was assigned to LLM-DALL·E 2 in 44.6% of cases, followed by 24.0% for DALL·E 2, 14.0% for LLM-SD_{Structured}, 10.0% for LMM-SD, and 7.3% for SD. Using the same prompts as for LMM-DALL·E 2, we still observe an improvement in LLM-SD over the original SD output. Finally, as expected, LLM-SD_{Structured} improves over LMM-SD.

In Figure 4, we show examples of visual metaphors generated using the linguistic metaphors or their visual elaborations as CoT prompts. We observe that the latter, where CoT prompting is involved, are of higher quality. For instance, a good visual metaphor for the metaphorical expression “*Books are the mirror to the soul*” would require

books, a mirror, and superimposing the mirror with some approximate depiction of a soul (usually illustrated as a person). However, the images that DALL·E 2 and Stable Diffusion generate (columns 3 and 5, respectively), just contain books. This problem is fixed with CoT prompting, as seen in columns 2, 4, and 6. The observations are similar for the metaphor “*I feel like a lily in February*”, where the implicit meaning of being *out of place* is depicted by lilies blooming in February over a snowy (instead of sunny) landscape.

How do expert illustrators perceive model-generated visual metaphors? One of the goals of our evaluation, besides obtaining a subjective ranking of the tested models, was to analyze some

of the flaws in the output. As stated above, for every image that was not considered “Perfect” or “Lost Cause”, we collected suggestions from experts about changes that would improve the image as a visual metaphor. Examples are given in Figure 5. This helps us understand where models might still be lacking, and the extent to which future interaction with illustrators might shape model-generated outputs to be acceptable. We find that issues in the output may be due to a model not being able to accurately depict a prompt, due to under-specification in terms of the objects to be represented or to the implicit property not being properly depicted. For instance, the CoT prompt for the metaphor “*It was a moonless night, the air was still and the crickets were like living shadows*” accurately describes it as “*An illustration of a moonless night sky with still air and crickets crawling around as living shadows.*”. However, the model fails to understand the word moonless and adds a moon to the picture. Additionally, while it adds the crawling crickets to the picture, there are no shadows. This affects the way we perceive the metaphor since its implicit meaning is “dark and creepy”. However, the rest of the image is high quality in terms of depiction. On the contrary, for the metaphor “*He was like a butterfly in autumn, waiting to be destroyed by the first frost*”, the CoT prompt “*An illustration of a butterfly perched on an autumn leaf with the first frost starting to form around it*” misses out on the source ‘He’ (ideally a fragile man) but the model depicts it perfectly.

Table 2 shows that nearly all models have room for improvement. Future work can use these suggestions in the form of natural language instructions to edit model-generated images, as demonstrated in recent work by Brooks et al. (2022).

4.2 Human-AI Collaboration Evaluation

Intrinsic Evaluation. To better understand if Human-AI collaboration leads to better quality visual metaphors, we conduct another round of evaluation with the same group of professional artists. Our experimental setup is the same as in our previous evaluation, except that instead of five images, we provide them with two visual metaphors for the same input: one from the **HAIVMet** corpus and the other from LLM-DALL-E 2 used in the previous round of evaluation (with their order shuffled). We then ask them to objectively provide a ranking between the two systems or tie them if they are both

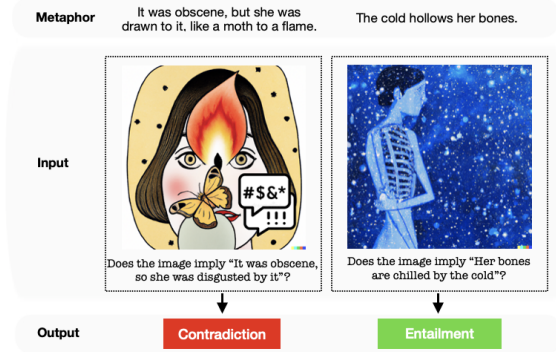


Figure 6: Visual Entailment task. Given an image and a prompt containing a hypothesis, predict whether the hypothesis entails, contradicts, or is neutral to the image.

of the same quality. They are also asked to provide instructions for improving them (unless they are Perfect or Lost Cause). We get the final verdict using majority voting. We obtain an inter-annotator agreement of 0.57 based on Fleiss’s kappa (Fleiss, 1971) (“moderate agreement”). Our results in Table 3 show that while 37% of the images are of similar quality, from the remaining images professionals preferred instances from **HAIVMet** 45% of the time compared to LLM-DALL-E 2 18% of time. Finally, the **HAIVMet** data has an almost negligible number of Lost Causes, providing further evidence of its high quality.

Criterion	LMM-DALL-E 2	HAIVMet	Tie
Preference	18%	45.0%	37%
Lost Cause	5%	1.6%	-
Perfect	52%	63.6%	-

Table 3: Proportion of Preference, Lost Cause, and Perfect cases from LMM-DALL-E 2 and **HAIVMet** for metaphors in our blind test set.

Extrinsic Evaluation: Visual Entailment Task.

Apart from being a rich source of visual metaphors, our dataset can also be useful in downstream applications. We showcase this by using it in a Visual Entailment (VE) task, where a vision-language model needs to predict whether a hypothesis is entailed by an image (cf. Figure 6). We use OFA (Wang et al., 2022), a state-of-the-art VE model finetuned on SNLI-VE (Xie et al., 2019). SNLI-VE only contains real-world images, but OFA is pre-trained on $\sim 20M$ image-text pairs some of which are synthetic. We extract 958 metaphors from our dataset that are associated with **literal** natural language entailment pairs from FLUTE (Chakrabarty et al., 2022b), CrossLing Metaphors (Tsvetkov

Model	Dev	Test
OFA _{SNLI-VE}	25.25	27.81
OFA _{SNLI-VE+HAIVMet}	49.90	51.15

Table 4: Visual Entailment Results. OFA (Wang et al., 2022) fined-tuned on SNLI-VE (Xie et al., 2019) vs. SNLI-VE+HAIVMet. Bold indicates best performance.

et al., 2014) and Metaphor Paraphrase (Bizzoni and Lappin, 2018) (see Appendix C for details on the data construction procedure). We split the data into train, validation and test sets, which contain 708, 100 and 150 metaphors (3686/506/831 image-text pairs), respectively. We fine-tune OFA-base (182M parameters) for 10 epochs with learning rate $6e-5$ and polynomial decay ($\text{weight}=0.01$), and batch size 8 on an NVIDIA RTX A6000 GPU for 8 hours. We select the model that has best performance on the development set. We show that accuracy on the test set improves by ~ 23 points compared to OFA’s performance when it is only finetuned on SNLI-VE. This result is indicative of the quality and usefulness of our dataset which can help vision-language models capture metaphoric meaning

5 Compositionality in Visual Metaphors

In prior work, Gutiérrez et al. (2016) showed that metaphorical meaning is not only a property of individual words but arises through cross-domain composition. Gal (2019) further argues that a metaphor is a visual material rather than conceptual. It is a mechanism of syntactic structure, forms, and material composition, which goes along with the perception of structures and compositions. Many images from our HAIVMet data showcase the compositional nature of visual metaphors, as can be seen in Table 5. For example, to visualize the metaphor “*Love is a crocodile in the river of desire*” the model needs to show both a human and a crocodile while depicting a sense of desire by embodying love as a concept. Similarly, for “*He froze with fear when he saw it*”, the metaphor needs to not only depict fear but also combine it with the state of being frozen. We can successfully achieve these difficult compositional visualizations through efficient human-AI collaboration.

6 Conclusion

We show that using Chain-of-Thought prompting for generating visual elaborations of linguistic metaphors leads to significant improvements

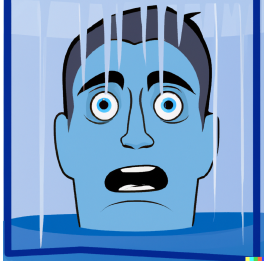


He froze with fear when he saw it.	
Beauty is a fading flower.	
Love is a crocodile in the river of desire.	

Table 5: Visual Metaphors from the HAIVMet dataset demonstrating compositional generalization.

in the quality of visual metaphors generated by diffusion-based text-to-image models. These models excel at depicting literal objects and actions, but cannot make the leap from figurative phrases to visual depiction without a detailed explanation of the implicit meaning. Though there are still particular aspects of visual composition and figurative imagery that current models fail to capture, the breadth of information collected in this dataset not only allows us to understand the current limitations of image generation but also provides the data necessary to improve visual metaphor generation in the future. We plan to further examine the effect of prompt phrasing on the quality of the generated visual metaphors, and how that effect differs across different models.

Limitations

While the results of Human-AI collaboration for visual metaphor generation are very promising, such a procedure might be time-consuming but at the same time necessary for maintaining quality. We want to acknowledge that both our LLM and best-forming Diffusion models are released through a paid API and are not open-sourced. While our

best-performing system uses Chain Of Thought Prompting, there are several other prompting or task decomposition techniques that we did not perform an extensive comparison with. Last but not least, there is still enough room for potential improvement in generating visual metaphors which can be achieved by designing better prompts or by improving the compositional generalization of diffusion models. We also recognize the inherent limitation of an English-only basis for our visual metaphors and hope in the future to expand to other languages for source material.

Ethics Statement

The use of text-to-image generation models is subject to concerns about intellectual property and copyrights of the images generated since the models are trained on web-crawled images. Our task is restricted to generating visual metaphors from linguistic metaphors, and the human-AI collaboration setup should be considered as a creative aid tool. All data collected by human respondents were anonymized and only pertained to the data they were being shown. We do not report demographic or geographic information, given the limited number of respondents, so as to maintain full anonymity. Workers on UpWork were informed that that the work they were doing was going to be used for research purposes. They were paid a wage of 20\$ per hour as decided by the workers themselves. Workers were paid their wages in full immediately upon the completion of their work.

References

- Keiga Abe, Sakamoto Kayo, and Masanori Nakagawa. 2006. [A computational model of the metaphor generation process](#). In *Proceedings of the 28th Annual Meeting of the Cognitive Science Society*, pages 937–942, Vancouver, Canada. Psychology Press.
- Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas Guibas. 2021. [ArtEmis: Affective Language for Visual Art](#). In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11564–11574.
- Arjun R. Akula, Brendan Driscoll, Pradyumna Narayana, Soravit Changpinyo, Zhiwei Jia, Suyash Damle, Garima Pruthi, Sugato Basu, Leonidas Guibas, William T. Freeman, Yuanzhen Li, and Varun Jampani. 2023. [Metaclue: Towards comprehensive visual metaphors research](#). In *CVPR 2023*.
- Yuri Bizzoni and Shalom Lappin. 2018. [Predicting human metaphor paraphrase judgments with deep neural networks](#). In *Proceedings of the Workshop on Figurative Language Processing*, pages 45–55, New Orleans, Louisiana. Association for Computational Linguistics.
- Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2022. [Instructpix2pix: Learning to follow image editing instructions](#). *arXiv preprint arXiv:2211.09800*.
- Tuhin Chakrabarty, Smaranda Muresan, and Nanyun Peng. 2020. [Generating similes effortlessly like a pro: A style transfer approach for simile generation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6455–6469, Online. Association for Computational Linguistics.
- Tuhin Chakrabarty, Vishakh Padmakumar, and He He. 2022a. [Help me write a poem: Instruction tuning as a vehicle for collaborative poetry writing](#). *arXiv preprint arXiv:2210.13669*.
- Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022b. [FLUTE: Figurative language understanding through textual explanations](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Tuhin Chakrabarty, Xurui Zhang, Smaranda Muresan, and Nanyun Peng. 2021. [MERMAID: Metaphor generation with symbolism and discriminative decoding](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4250–4261, Online. Association for Computational Linguistics.
- Lydia B. Chilton, Savvas Petridis, and Maneesh Agrawala. 2019. [Visiblends: A flexible workflow for visual blends](#). In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19*, page 1–14, New York, NY, USA. Association for Computing Machinery.
- Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. 2022. [Training-free structured diffusion guidance for compositional text-to-image synthesis](#). *arXiv preprint arXiv:2212.05032*.
- Joseph L Fleiss. 1971. [Measuring nominal scale agreement among many raters](#). *Psychological bulletin*, 76(5):378.
- Michalle Gal. 2019. [Visual metaphors and cognition: Revisiting the non-conceptual](#). In Kristof Nyiri and Andras Benedek, editors, *Perspective on Visual Learning, Vol. 1. The Victory of the Pictorial Age*, pages 79–90. PhilPapers.

- E. Dario Gutiérrez, Ekaterina Shutova, Tyler Marghetis, and Benjamin Bergen. 2016. [Literal and metaphorical senses in compositional distributional semantic models](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 183–193, Berlin, Germany. Association for Computational Linguistics.
- Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.
- Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. 2017. Automatic understanding of image and video advertisements. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1100–1110.
- Ben Hutchinson, Jason Baldridge, and Vinodkumar Prabhakaran. 2022. [Underspecification in scene description-to-depiction tasks](#). In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1172–1184, Online only. Association for Computational Linguistics.
- EunJeong Hwang and Vered Shwartz. 2023. [Memecap: A dataset for captioning and interpreting memes](#).
- Bipin Indurkha and Amitash Ojha. 2013. [An Empirical Study on the Role of Perceptual Similarity in Visual Metaphors and Creativity](#). *Metaphor and Symbol*, 28(4):233–253.
- Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. [How can we know what language models know?](#) *Transactions of the Association for Computational Linguistics*, 8:423–438.
- Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. [The perils of using Mechanical Turk to evaluate open-ended text generation](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1265–1285, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Ricardo Kleinlein, Cristina Luna-Jiménez, and Fernando Fernández-Martínez. 2022. Language does more than describe: On the lack of figurative speech in text-to-image models. *arXiv preprint arXiv:2210.10578*.
- George Lakoff. 1993. [The Contemporary Theory of Metaphor](#). In Andrew Ortony, editor, *Metaphor and Thought*, pages 202–251. Cambridge University Press.
- Evelina Leivada, Elliot Murphy, and Gary Marcus. 2022. [Dall-e 2 fails to reliably capture common syntactic processes](#).
- Emmy Liu, Chen Cui, Kenneth Zheng, and Graham Neubig. 2022a. [Testing the ability of language models to interpret figurative language](#).
- Vivian Liu and Lydia B Chilton. 2022. Design guidelines for prompt engineering text-to-image generative models. In *CHI Conference on Human Factors in Computing Systems*, pages 1–23.
- Vivian Liu, Tao Long, Nathan Raw, and Lydia Chilton. 2023. Generative disco: Text-to-video generation for music visualization. *arXiv preprint arXiv:2304.08551*.
- Vivian Liu, Han Qiao, and Lydia Chilton. 2022b. Opal: Multimodal image generation for news illustration. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pages 1–17.
- Edward F. McQuarrie and David Glen Mick. 1999. [Visual Rhetoric in Advertising: Text-Interpretive, Experimental, and Reader-Response Analyses](#). *Journal of Consumer Research*, 26(1):37–54.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.
- Savvas Petridis and Lydia B. Chilton. 2019. [Human errors in interpreting visual metaphor](#). In *Proceedings of the 2019 on Creativity and Cognition, C&C '19*, page 187–197, New York, NY, USA. Association for Computing Machinery.
- Barbara J Phillips. 2003. Understanding visual metaphor in advertising. *Persuasive imagery*, pages 304–317.
- Barbara J. Phillips and Edward F. McQuarrie. 2004. [Beyond Visual Metaphor: A New Typology of Visual Rhetoric in Advertising](#). *Marketing Theory*, 4(1-2):113–136.
- Adam Poliak, Aparajita Haldar, Rachel Rudinger, J Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2018. Collecting diverse natural language inference problems for sentence representation evaluation. In *BlackboxNLP@ EMNLP*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning transferable visual models from natural language supervision](#). In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

- Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022a. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.
- Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022b. Hierarchical text-conditional image generation with clip latents.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-resolution image synthesis with latent diffusion models.
- Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*.
- Asuka Terai and Masanori Nakagawa. 2010. A computational system of metaphor generation with evaluation mechanism. In *International Conference on Artificial Neural Networks*, pages 142–147, Thessaloniki, Greece. Springer.
- Yulia Tsvetkov, Leonid Boytsov, Anatole Gershan, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 248–258, Baltimore, Maryland. Association for Computational Linguistics.
- Tony Veale. 2016. Round up the usual suspects: Knowledge-based metaphor generation. In *Proceedings of the Fourth Workshop on Metaphor in NLP*, pages 34–41, San Diego, California. Association for Computational Linguistics.
- Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR.
- Sitong Wang, Samia Menon, Tao Long, Keren Henderson, Dingzeyu Li, Kevin Crowston, Mark Hansen, Jeffrey V Nickerson, and Lydia B Chilton. 2023. Reelframer: Co-creating news reels on social media with generative ai. *arXiv preprint arXiv:2304.09653*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.
- Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*.
- Ron Yosef, Yonatan Bitton, and Dafna Shahaf. 2023. Irfi: Image recognition of figurative language. *arXiv preprint arXiv:2303.15445*.
- Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*.
- Zhiwei Yu and Xiaojun Wan. 2019. How to avoid sentences spelling boring? Towards a neural approach to unsupervised metaphor generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 861–871, Minneapolis, Minnesota. Association for Computational Linguistics.
- Dongyu Zhang, Minghao Zhang, Heting Zhang, Liang Yang, and Hongfei Lin. 2021. MultiMET: A multimodal dataset for metaphor understanding. *ACL/IJCNLP 2021 - 59th Annu. Meet. Assoc. Comput. Linguist. 11th Int. Jt. Conf. Nat. Lang. Process. Proc. Conf.*, pages 3214–3225.

A Appendix

A.1 Hyperparameters for chain of Thought Prompting

We use the Instruct GPT-3 (davinci-002) model for Chain-of-Thought (CoT) prompting. To generate **Objects to be Included**, **Implicit Meaning** and **Visual Elaboration** we use the following hyperparameters: temperature=0.7, max tokens=256, top p=1.0, best of=1, frequency penalty=0.5, presence penalty=0.5.

In this task you will judge the quality of visual metaphors. The metaphors are written in bold, and they are followed by two images. Write [1] in the box next to the metaphor if the left image describes the metaphor better than the right one, and [2] otherwise. If both are equally good write [0].

And now they were leaking all kinds of documents. []



Figure 7: Annotators were provided with a list of metaphors along with two images generated by DALL-E 2 using our two different prompting methods, CoT and Completion. The order of the images is random.

B Does better prompting lead to better images?

Language models are sensitive to prompting (Jiang et al., 2020), as are text-to-image diffusion-based models (Liu and Chilton, 2022). We employ CoT prompting to generate visual elaborations of linguistic metaphors using Instruct GPT-3 (davinci-002). The alternative to CoT would be classic Completion prompting, which would require Instruct GPT-3 (davinci-002) to provide visual elaborations for the metaphors without first reasoning about objects and implicit meaning.

We evaluate whether or not requiring Instruct GPT-3 (davinci-002) to reason about both the included objects and the implicit meaning **before** providing a visual elaboration improves the quality of the generated visual metaphor, by comparing to Completion prompting where the visual elaboration is directly predicted without the intermediate reasoning steps. For a fair comparison, we require the prompts to be as similar in content as possible, and use the same 5 few-shot examples as for CoT, only removing the intermediate information (objects to be included, implicit meaning) for the Completion prompt.

We verify the hypothesis that CoT improves image quality through a small-scale human evaluation. We consider 50 metaphors for this experiment and generate visual descriptions using the prompt template shown in Table 6, which replicates the metaphors and visual elaborations in Table 1 but without the instruction section or the step by step reasoning used in CoT. The resulting prompts are passed to DALL-E 2 to generate images.

We provide 3 annotators with the list of 50 metaphors, as well as the two images that are generated by Instruct GPT-3 (davinci-002) using CoT and Completion prompting without any further post-processing. Figure 7 shows the instructions provided to the annotators and an annotation example. To mitigate the subjectivity of the task, which is confirmed by a fair average pairwise Cohen’s Kappa score ($\kappa=0.26$), we consider the majority vote selection for each example. Our results show that annotators select 27/50 images that are generated using CoT prompts, 11/50 using Completion prompts, and 12/50 images are judged to be of equal quality regardless of the prompting strategy used. Our results indicate that prompting can significantly improve the quality of the generated images suggesting that future work should investigate ways to further improve the quality of the generated visual metaphors by extracting more detailed specifications from LLMs.

C Visual Entailment Data

In order to perform the visual entailment task, we require metaphors that are associated with literal hypotheses and their corresponding labels (entailment, contradiction, neutral). **FLUTE** (Chakrabarty et al., 2022b) offers such data without any further processing. For the metaphors in **CrossLing Metaphors** (Tsvetkov et al., 2014) and **Metaphor Paraphrases** (Bizzoni and Lappin, 2018) we employ **recasting**, namely “*leveraging existing datasets to create NLI examples*”, (Poliak et al., 2018) to convert them into textual entailment data.

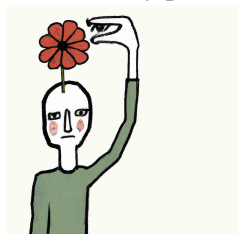
The metaphors in **Metaphor Paraphrases** (Bizzoni and Lappin, 2018) are each associated with four ranked candidate literal sentence. Each sentence is annotated with a value from 1 to 4, indicating the degree to which the sentence is a paraphrase of the original metaphoric sentence, where 4 stands for exact paraphrase. We consider each sentence and each of the candidate paraphrases as a sentence pair for a textual entailment classification problem

1. Metaphor: My lawyer is a shark. <i>An illustration of a shark in a suit with fierce eyes and a suitcase and a mouth open with pointy teeth</i>
2. Metaphor: I've reached my boiling point. <i>An illustration of a boiling pot of water with a person's head popping out of the top, steam coming out of their ears, and an angry expression on their face.</i>
3. Metaphor: Joe: that's because you're like a snail surfing on molasses. <i>An illustration of a person with a snail shell on their back slowly sliding down a hill of molasses.</i>
4. Metaphor: Absence is the dark room in which lovers develop negatives. <i>An illustration of an ominous dark room with a film strip negatives hanging and a red heart in the center with a person in the corner looking sad and lonely.</i>
5. Metaphor: My heart is a rose thorn. <i>An illustration of a heart with a prickly thorn coming out of the center and barbs going outward.</i>
6. Metaphor: My bedroom is a pig sty <i>An illustration of a messy bedroom with clothes and garbage strewn about and a pig in the center rooting through the mess.</i>

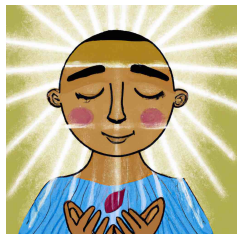
Table 6: Simple Completion prompt to elicit visual elaboration for a given metaphor, using the same 5 few shot examples as in the CoT prompting strategy, but without the objects to be included and the implicit meaning.

CoT

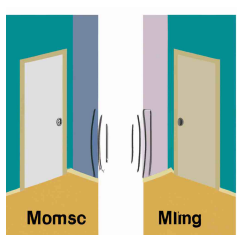
I had already planted the idea in her mind.



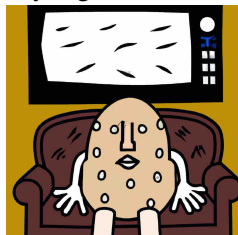
Completion



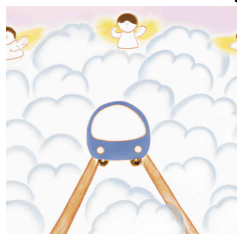
The rooms communicated.



My big brother is a couch potato.



You will love the new train. It is a heavenly ride



and manually annotate them.

The **CrossLing Metaphors** (Tsvetkov et al., 2014) dataset consists of 200 metaphoric English sentences, 200 literal English sentences, and their Russian translations. For the purposes of this study we were only concerned with using the 200 English metaphoric sentences to construct entailment pairs. We manually created three literal hypotheses with corresponding labels (entailment, contradiction, and neutral).

The data was presented to 3 annotators to verify the quality of the labels. The annotators were presented with both the metaphoric premise and the literal hypothesis, and had to decide whether the hypothesis was entailed, contradicted, or neutral to the statement. The mean pairwise annotator agreement for the labels was .79. The gold label for the data was assigned by majority vote.

C.1 Evaluation Interface

Figure 9 and 10 show the evaluation interface for LLM-Diffusion Model collaboration and Human AI collaboration respectively. For the LLM-Diffusion Model 5 images are presented in randomly shuffled order while for Human AI collaboration 2 images are presented one from LLM-DALLE and the other from HAIVMet.

Table 7: Examples of images generated by DALL-E 2 prompted with CoT (left) and Completion (right).












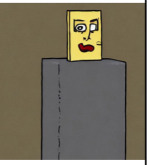

















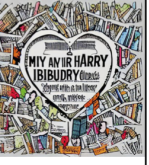



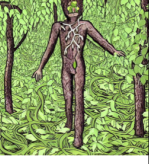
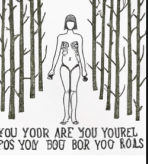
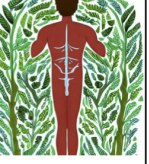


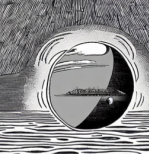







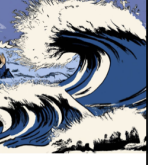



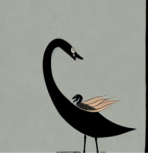


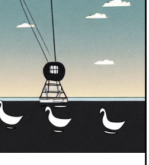
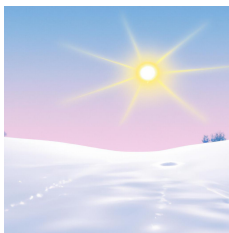
Metaphor	HAIVMet [Gold]	LLM-Dall·E 2	LLM-SD	LLM-SD-Structured	Dall·E 2	SD
<i>They had entered their autumn years</i>						
<i>He was like a block of cement</i>						
<i>Consumed by the thoughts that grew in the head</i>						
<i>My brain is like a box of crayons</i>						
<i>My heart is a library</i>						
<i>Your body is a forest</i>						
<i>The planet is a sinking ship</i>						
<i>The stormy ocean was a raging bull</i>						
<i>My gondola is a black sea-swan</i>						

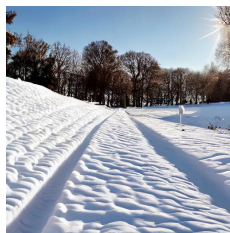
Figure 8: Additional examples of output from the models described in Section 4.1 for randomly chosen metaphors. HAIVMet is our gold standard.

Visual Metaphor: All was like a winter morning after it had snowed all night

Enter your ranking among the images, separated by a comma. Example: 5,4,2,3,1



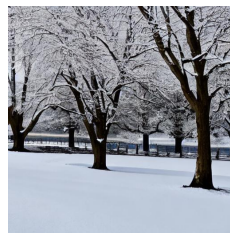
1



2



3



4



5

<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>

submit

(a)

Figure 9: Evaluation interface for LLM-Diffusion Model collaboration with five systems, as described in Section 4.

Visual Metaphor: Love is a warrior's yearning

Enter your ranking among the images between 2 images separated by a comma. Example: 1,2. If they are both are same just type "Tie"



1



2

submit

(a)

Figure 10: Evaluation interface for Human AI collaboration with two systems, as described in Section 4.

ACL 2023 Responsible NLP Checklist

A For every submission:

- A1. Did you describe the limitations of your work?
Un-numbered section after Conclusion
- A2. Did you discuss any potential risks of your work?
Un-numbered section after Limitations
- A3. Do the abstract and introduction summarize the paper’s main claims?
Abstract and Section 1
- A4. Have you used AI writing assistants when working on this paper?
Left blank.

B Did you use or create scientific artifacts?

Sections 1, 3, and 4

- B1. Did you cite the creators of artifacts you used?
Sections 1, 3, and 4
- B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section 4.1
- B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Our dataset is free to use without restriction, and as such does not require specification for the intended use.
- B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Un-numbered Code of Ethics section after the Conclusion and Limitations
- B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Sections 1 and Limitations
- B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Sections 1, 3, and 4

C Did you run computational experiments?

Section 4

- C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 4.2

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

- C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4.2
- C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4.2
- C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Sections 3 and 4
- D** **Did you use human annotators (e.g., crowdworkers) or research with human participants?**
Sections 3, 4; Appendix A and B
- D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Sections 3, 4; Appendix A and B
- D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section 4
- D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Code of Ethics section
- D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
There was no IRB, and as such no protocol needed to be approved.
- D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
We did not collect such information from our annotators as they were not the subjects of our research.