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

Tuhin Chakrabarty [1], Arkadiy Saakyan [1], Olivia Winn [1],
Artemis Panagopoulou [2], Yue Yang [2], Marianna Apidianaki [2], Smaranda Muresan [1]

[1]Columbia University, [2]University of Pennsylvania

Motivation and Contributions

- A novel approach for generating visual metaphors through the collaboration of large language models (LLMs) and diffusion-based text-to-image models.
- A high-quality visual metaphor dataset built through Human-AI collaboration.
- A thorough evaluation of LLM-Diffusion Model collaboration and Human-AI collaboration with professional illustrators.

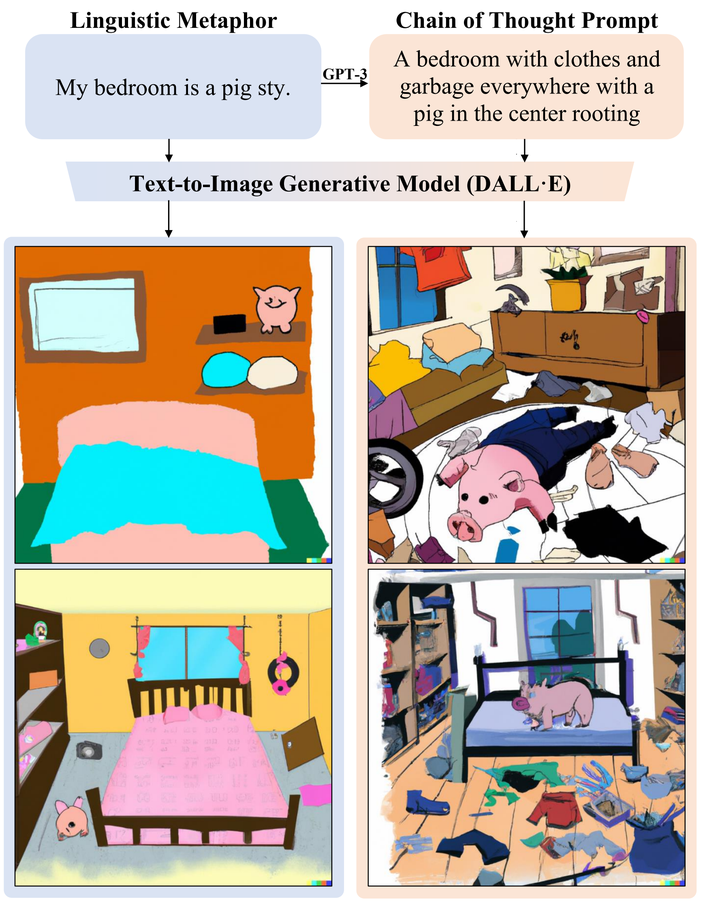
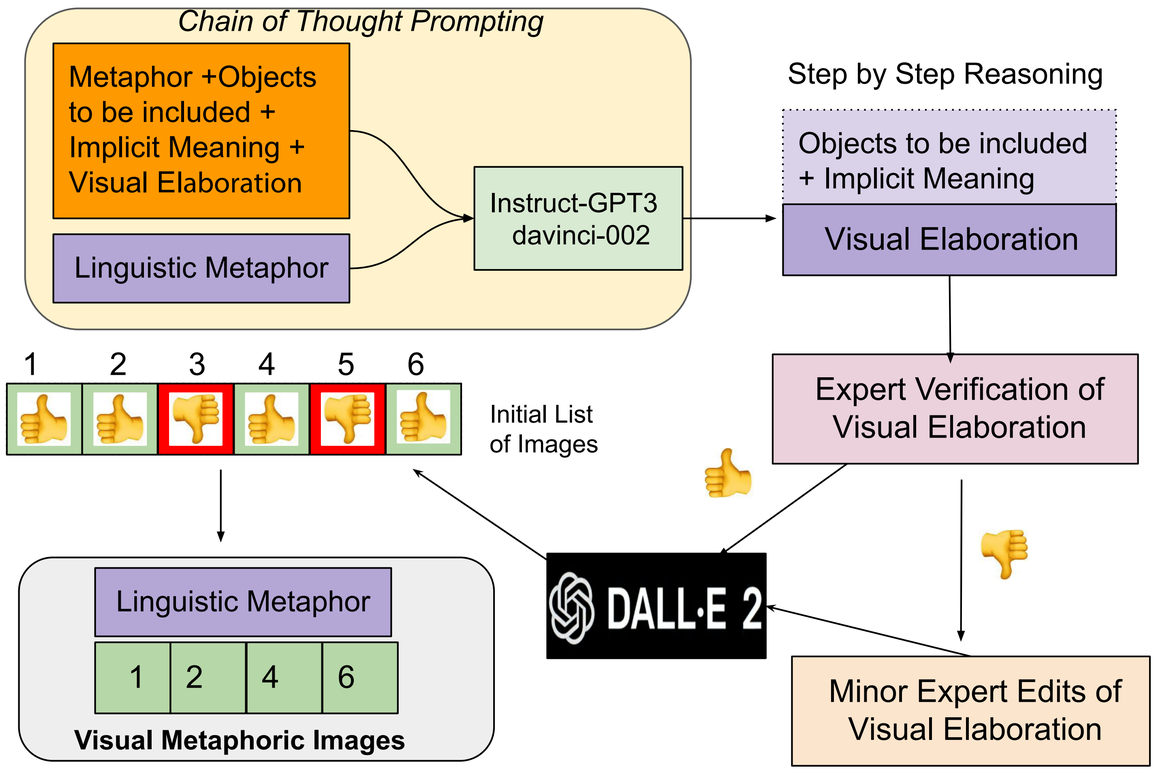
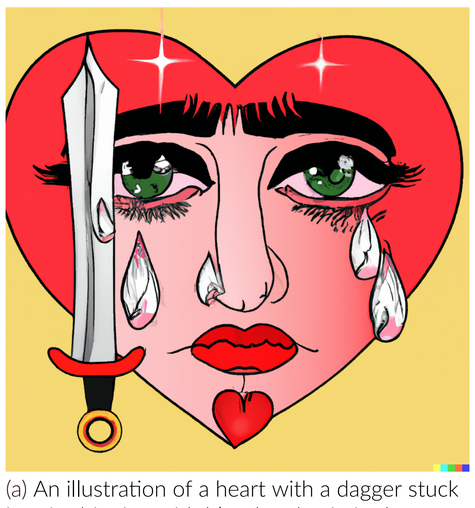


Figure 1. Visual metaphors generated by DALL·E 2 for the linguistic metaphor "My bedroom is a pig sty".We can take the original verbal metaphor as the input (left) or use GPT-3 with Chain of Thought prompting (right).

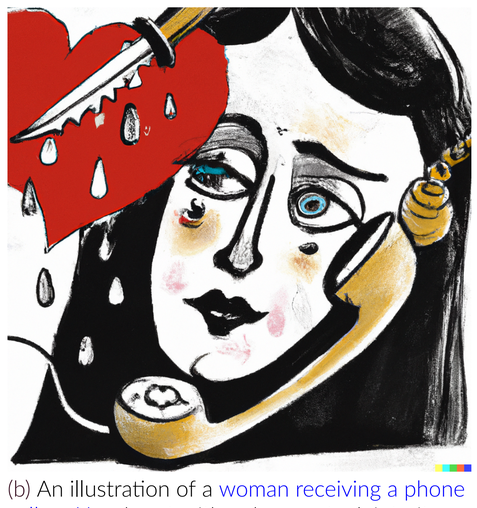
LLM-DALLE-Human Collaboration for Visual Metaphor Generation (HAIVMet)



The news of the accident was a dagger in her heart



(a) An illustration of a heart with a dagger stuck into it, dripping with blood and pain in the woman's eyes



(b) 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.

Evaluation Metrics and Results

- We evaluate on two types of collaboration using **expert concept illustrators** who rank the 5 systems. We also report the percentage of "Lost Cause" cases in order to identify systems that generate the least amount of bad images.
- Annotators are asked to provide atomic instructions denoting a single action/change for imperfect visual metaphors. We compare the models on the basis of the average number of instructions that have been proposed for improving their produced images.
- LLM-Diffusion Model Collaboration Evaluation: How good are state of art diffusion models when they collaborate with GPT3?

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

Table 1. 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.

- Human AI Collaboration Evaluation : Is collaboration with experts actually helpful?

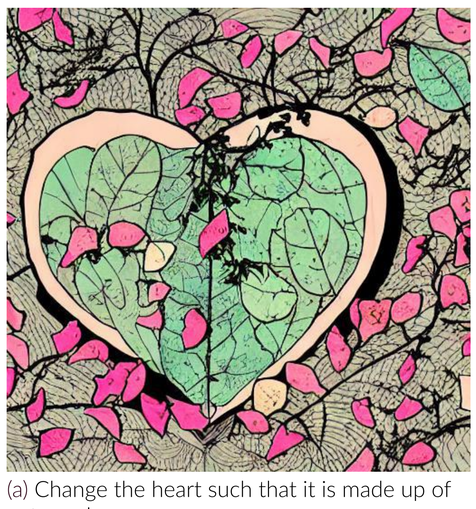
Criterion	LMM-DALL·E 2	**HAIVMet**	Tie
Preference	18%	**45.0%**	37%
Lost Cause	5%	**1.6%**	-
Perfect	52%	**63.6%**	-

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

Some more Visual Metaphors

Ongoing work to improve model generated Visual metaphors

My heart is a garden tired with autumn



(a) Change the heart such that it is made up of autumn leaves

(b)