GOME: Grounding-based Metaphor Binding With Conceptual Elaboration For Figurative Language Illustration

Anonymous ACL submission

Abstract

The illustration or visualization of figurative language, such as linguistic metaphors, is an emerging challenge for existing Large Language Models (LLMs) and multimodal models. Due to their comparison of seemingly unrelated concepts in metaphors, existing LLMs have a tendency of over-literalization, which illustrates figurative language solely based on literal objects, ignoring the underlying groundings and associations across disparate metaphorical domains. Furthermore, prior approaches have ignored the binding process between visual objects and metaphorical attributes, which further intensifies the infidelity of visual metaphors. To address the issues above, we propose GOME (**GrO**unding-based **ME**taphor Binding), which illustrates linguistic metaphors from the grounding perspective elaborated through LLMs. GOME consists of two steps for metaphor illustration, including grounding-based elaboration and scenario visualization. In the elaboration step, metaphorical knowledge is integrated into systematic instructions for LLMs, which employs a CoT prompting method rooted in rhetoric. This approach specifies metaphorical devices such as vehicles and groundings, to ensure accurate and faithful descriptions consumed by text-to-image models. In the visualization step, an inference-time metaphor binding method is realized based on elaboration outputs, which register attentional control during the diffusion process, and captures the underlying attributes from the abstract metaphorical domain. Comprehensive evaluations using multiple downstream tasks confirm that, GOME is superior to isolated LLMs, diffusion models, or their direct collaboration.

1 Introduction

Figurative language, such as metaphors, is a rhetorical device that describes an object or concepts in a non-literal manner to elucidate an idea or facilitate a comparison (LAKOFF, 1993). For example, in the famous saying 'books are the ladder of human progress', books are described as ladders, which highlights the role of books in facilitating intellectual and societal advancement. Visualizing such figures of speech is exceedingly beneficial to express creative ideas in a more intuitive way (Schwering et al., 2009), which facilitates the understanding of both perceptible objects and implicit concepts or emotions, and has been leveraged as persuasive tools to evoke attitudes (Jahameh and Zibin, 2023).

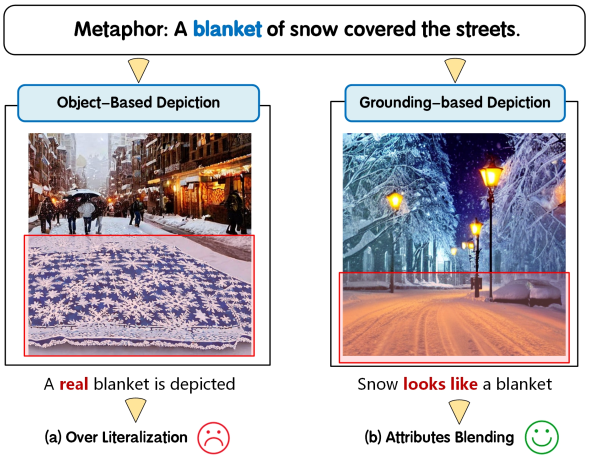

Metaphor: A blanket of snow covered the streets.
Object-Based Depiction
Grounding-based Depiction
A real blanket is depicted
Snow looks like a blanket
(a) Over Literalization
(b) Attributes Blending

Figure 1: For the illustration of 'a blanket of snow covered the streets', we are expecting some metaphorical attributes, such as pervasive or encompassing, to be adapted from 'blanket' (source domain) to 'snow' (target domain), instead of a real blanket to be presented (over-literalization).

Due to the non-literal juxtaposition in figurative expressions (Zhang et al., 2024), metaphors can not be visualized directly through large diffusion-based text-to-image models, which can only work conditioned on descriptive texts with literal captions (Rombach et al., 2022; Saharia et al., 2022). Recent works primarily deal with this issue through object-based visual elaboration (Chilton et al., 2019; Chakrabarty et al., 2023), which is a query rewriting method with Large Language Models (LLMs) focusing on the objects to be represented. For instance, the metaphorical statement ‘A blanket of snow coverd the streets’, can be elaborated into a descriptive caption, like ‘An illustration of a blanket with snowflakes falling on it and the streets below’, which identifies the objects of ‘blanket’, ‘snowflake’, and then consumed by diffusion-based models for illustration, as shown in Figure 1 (a).

Despite their inspiring exploration, we’ve found two main problems in the entire process, including over-literalization and metaphorical attribute-object binding. (1) **Over-literalization** means that, when depicting a linguistic metaphor as an image with LLMs, objects within the metaphor are excessively detailed, especially for the objects in the source domain for evoking abstract concepts, leading to a cluttered or diverted representation from the metaphor’s original intent (Black et al., 1979). Still take Figure 1 as the example, ‘blanket’ in the statement is used for reflecting the pervasive or encompassing nature of ‘snow’, rather than a referential object to be depicted. Such excessive concretization may diminish the metaphor’s original grounding, becoming overly straightforward and singular (Davidson, 1984). (2) **Attribute Binding** is the task of binding the attributes to the correct objects (Rombach et al., 2022; Saharia et al., 2022), which is particularly challenging for figures of speech because, the attributes is metaphorically entailed across different metaphorical domains (source domain and target domain), which impulses extra burden to diffusion models.

To address the issues above, we propose GOME (**L**LM-**e**labora**t**ed **Me**taphor), which illustrates linguistic metaphors from the grounding perspective to avoid over-literalization in LLM elaborations. The core idea of GOME is to unfold the non-literal expressions through a textual description from a rhetorical perspective, including tenor, vehicle, and pragmatic groundings, which are further leveraged for metaphor binding to preserve provoking attributes, instead of referential objects. GOME involves three main stages, firstly, following (Chakrabarty et al., 2023), we compile a collection of linguistic metaphors from six sets as a rich source of figurative language, which is post-filtered by LLM for visualizable metaphors. Secondly, we construct grounding-based visual elaboration with a CoT prompting method from a rhetoric perspective, which generates fine-grained metaphorical elements, as well as visual elaborations for subsequent depiction. Finally, an inference-time binding method is conducted through cross-attention controlling, which realizes compelling and faithful metaphor illustration by integrating objects and figurative attributes.

Overall, our contributions are the following: (1) The problem of over-literalization is firstly noticed in LLM elaborations for metaphors, which is then analyzed by a grounding-based depiction method to avoid excessive concretization. (2) A publicly available dataset [1] is introduced with 1351 visual elaborations of metaphors, together with the fine-grained metaphorical elements, including tenor, vehicle, and groundings for comprehensive metaphor illustration. (3) We propose a metaphorical attribute-object binding approach at an inference-time speed, which realizes attentional registration in the text-to-image process. (4) Comprehensive experiments verify the high robustness and fidelity of our method, which paves the way for figurative language visualization, as well as other downstream applications.

2 Related Work

2.1 Text-to-Image Generation

In recent years, advancements in text-to-image synthesis have been remarkable, with diffusion-based models surpassing earlier techniques like Variational Autoencoders (VAE) (Razavi et al., 2019) and Generative Adversarial Networks (Bao et al., 2017). Prominent models in this field include DALL·E 2 (Ramesh et al., 2022), Stable Diffusion (Rombach et al., 2022), MidJourney, and Craiyon. Despite their success in generating vivid and appealing imagery, there remain areas where they fail to capture accurate depictions (Leivada et al., 2022). For instance, recent studies (Kleinlein et al., 2022) have demonstrated that while diffusion models may struggle with the abstraction required for figurative language. Recent work (Liu et al., 2022b, 2023a; Wang et al., 2023) has explored cutting-edge systems showcasing the power of large language models and text-to-image models.

Extensive research has been conducted on textual figurative language, encompassing areas such as metaphor generation (Yu and Wan, 2019; Chakrabarty et al., 2020), idiom generation and paraphrasing (Liu and Hwa, 2016; Zhou et al., 2021), and simile recognition and interpretation (Zeng et al., 2020; He et al., 2022a). In contrast, the visualization of figurative language has garnered

[1] our code and data at https://github.com/EMNLP-2024-Submission/GOME.git

comparatively less attention. Existing methodologies (Chakrabarty et al., 2023) have predominantly focused on the creation of datasets that include images and annotations for metaphors, similes, and idioms (Yosef et al., 2023; Akula et al., 2023). However, these datasets tend to focus more on the inclusion of objects in metaphors. For instance, (Chakrabarty et al., 2023) generated visual descriptions based on objects and synthetic images for 1,540 linguistic metaphors. (Yosef et al., 2023) compiled a dataset containing about 3,000 figurative expressions paired with ground truth images through human annotations. (Akula et al., 2023) collected 5,061 metaphorical advertisement images using a simple annotation format of "A is as B as C" (e.g., "this pencil is as red as a firetruck"). Although these researches offer valuable resources, they do not facilitate an intrinsic process for the faithful depiction of metaphors.

3 Methodology

We present GOME, a collaboration of large language models and text-to-image models designed to generate visual elaborations and pictures from metaphorical text inputs. The development of GOME comprises three main stages, including data collection and the other two depiction steps illustrated in Figure 3. Firstly, we perform data collection by preprocessing a collection of metaphors sourced from previous researches. Secondly, we utilize a large language model (LLM) to generate visual elaborations for the metaphors by appropriate CoT prompt design with rhetoric knowledge in the system role. Finally, the paired data of metaphors and generated visual elaborations are fed into a diffusion model to realize metaphor depiction. Although previous research used DALL·E (Ramesh et al., 2022) to generate images, we utilize Stable Diffusion for a transparent and reproducible approach, and more importantly, a novel method to explore metaphorical attribute-object binding through attentional control. Concretely, The diffusion process is enriched with metaphorical object-attribute binding, using an inference-time optimization with a loss over cross-attention maps. The primary goal of GOME is to generate detailed textual descriptions of visual scenes (visual elaborations) to convey the intended meaning of the rich figurative phrases in metaphors.

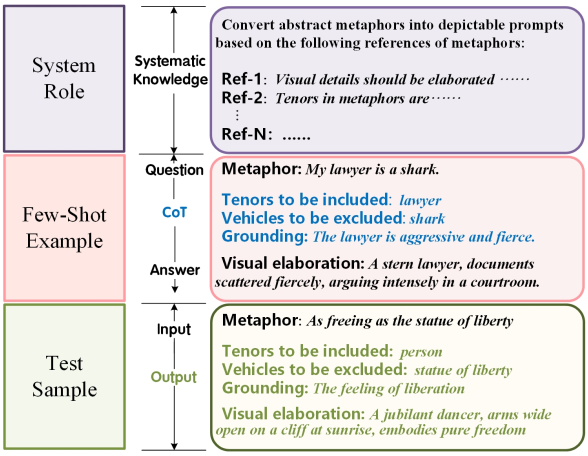

System Role
Few-Shot Example
Test Sample
Systematic Knowledge
Question
CoT
Answer
Input
Output
Convert abstract metaphors into depictable prompts based on the following references of metaphors:
Ref-1: Visual details should be elaborated ……
Ref-2: Tenors in metaphors are……
⋮
Ref-N: ……
Metaphor: My lawyer is a shark.
Tenors to be included: lawyer
Vehicles to be excluded: shark
Grounding: The lawyer is aggressive and fierce.
Visual elaboration: A stern lawyer, documents scattered fiercely, arguing intensely in a courtroom.
Metaphor: As freeing as the statue of liberty
Tenors to be included: person
Vehicles to be excluded: statue of liberty
Grounding: The feeling of liberation
Visual elaboration: A jubilant dancer, arms wide open on a cliff at sunrise, embodies pure freedom

Figure 2: Gounding-based LLM elaboration for figurative language. Outputs of the test sample are used for subsequent metaphor binding and image generation.

3.1 Visual Elaboration

Following previous research, (Chakrabarty et al., 2023; Shahmohammadi et al., 2023), we take ‘visual elaboration’ as a mention, which refers to the process of transforming or expanding figurative contents into visualizable textual descriptions. We generate synthetic visual elaborations using GPT-4. Synthetic data produced by LLMs (Thoppilan et al., 2022; Brown et al., 2020; Liu et al., 2023b) offer substantial benefits and demonstrate competitive, and in certain instances, superior performance compared to human-annotated data (He et al., 2022b; Wang et al., 2021; Hu et al., 2022). To decipher linguistic metaphors demanding proficiency in rhetorical devices, we ask the large language model (LLM) to act as an expert in metaphors, by integrating systematic domain knowledge, including the definition and characteristics of tenor, vehicle, groundings, etc, as well as examples into carefully designed instructions over 400 words in its system role. Details are seen in Appendix E.

Unlike previous prompts focused on all the possible objects, we propose to elaborate metaphors with less provocative objects from vehicles, but consider more on the underlying groundings. For example, given the original metaphor ‘love is like a gust of wind’, if the grounding is perceived as ‘love is gentle’, then the original metaphor could be converted into a visual description like: ‘two lovers embracing each other in a sunny field, their hair and clothes gently blown by a soft breeze’. Otherwise, if the grounding is ‘love is a brief passage’, then the metaphor should be depicted as: ‘In a park with fallen leaves during autumn, a couple broke

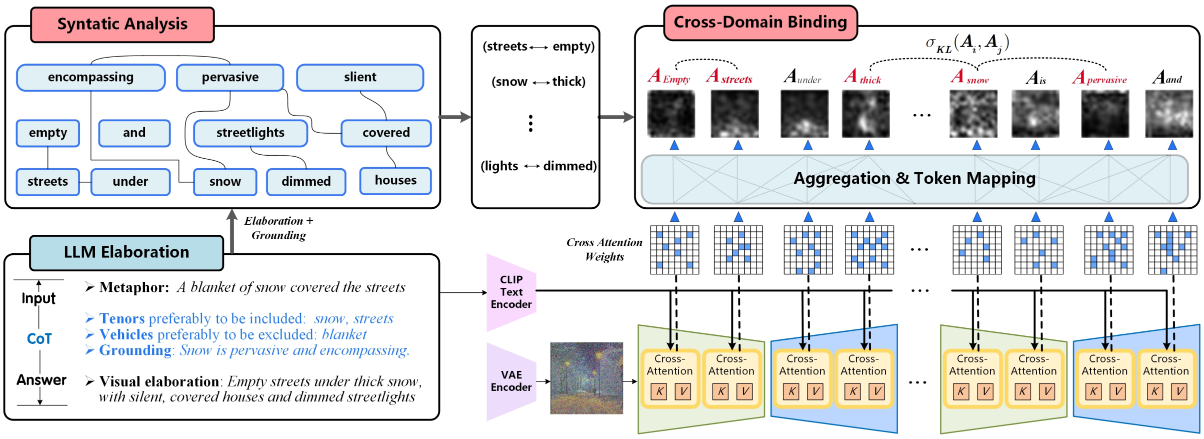


Figure 3: The overall workflow of our method. Firstly, the input metaphor is elaborated based on an LLM according to Figure 2. Secondly, the elaboration outputs, including the groundings, as well as the visual descriptions, undergo a syntactic analysis process to extract the binding pairs. Finally, elaboration outputs serve as the text inputs of a diffusion model, together with the metaphor binding objective based on results from syntactic analysis.

up. The woman left, and a man reached out his arm to grab her hand.' Specifically, we queried LLMs in the way of CoT prompting from a rhetorical perspective, together with the rhetorical knowledge integrated into the System Role.

3.2 Cross Domain Linguistic Binding

Different from previous metaphor visualization methods, which struggles to depict abstract concepts solely based on API calling, we conduct a metaphorical attribute-object binding process through attentional registration during the diffusion process. Our approach, which we call GOME, builds on the key idea that, vehicles can be internalized in the final scenario by metaphorical attribute-object binding, which blends metaphorical attributes from vehicles in the source domain to tenor objects in the target domain. Such cross-domain bindings, which consist of object nouns and attribute modifiers, can be analyzed based on the syntactic structure of natural language visual elaboration enhanced by metaphor groundings. Moreover, inspired by (Rassin et al., 2023), these bindings can be adhered to by designing an appropriate loss over the cross-attention maps of the diffusion model, and finally steer the generation of visual metaphors.

Given a pair of an object-noun from tenor and attribute modifiers from the vehicle, it is expected that the cross-attention map of the attribute significantly overlaps with that of the object, while remaining mostly distinct from the maps of other objects and attributes. To enforce these spatial relations within the attention maps, a specifically designed loss function is employed to operate across all cross-attention maps. This loss is then utilized during the inference phase with a pretrained diffusion model. The noised latents are optimized by performing a gradient step aimed at minimizing this loss. Detailed illustrations of the entire process are included in Figure 3.

Object-Attribute Pairs: Considering an enhanced visual elaboration sentence S_v with N tokens, which is obtained by concatenating the original elaboration sentence with the perceived natural grounding sentence, we first need to specify the objects and attributes to be attached across different domains (source and target domains). Let S_{MB} denote the sets containing k cross-domain pairs of objects and attributes $S_{MB} = \{(o_1, a_1), (o_2, a_2), \ldots, (o_k, a_k)\}$, where (o_i, a_i) is the i-th pair of tokens between the tenor object o_i and attribute modifiers a_i. For instance, the set for 'now is pervasive and encompassing' includes two pairs: ('snow', 'pervasive') and ('snow', 'encompassing'). To identify the object-attribute sets, we parse the enhanced visual elaboration S_v using spaCy's transformer-based dependency parser (Honnibal and Montani, 2017) and identify all object-nouns (either proper-nouns or common-nouns) that are not serving as direct modifiers of other nouns, and more importantly, presented as objects to be included in the visual elaborations. We then recursively collect all modifiers of the noun into the metaphor binding set S_{MB}:

$$\begin{aligned} S_{MB} &= \{(o_1, a_1), (o_2, a_2), \ldots, (o_k, a_k)\} \\ &= Parser_{DP}(S_v); \end{aligned} \quad (1)$$

Where $Parser_{DP}$ denotes the dependency parser (Honnibal and Montani, 2017). It is worth noting that, the set of attributes includes a range of syntactic relations, such as adjectival modification (amod; 'the broken heart'), compounds (compound; 'the history wheels'), adjectival complement (acomp; 'Her words were as sharp as a knife'), and coordination between modifiers (conj; 'Her voice was a melody, sweet and haunting').

Metaphorical Binding: Let $A_1, A_2, \ldots, A_N$ represent the attention maps of all N tokens in the enhanced visual prompt S_v, and let $M_{dis}(A_i, A_j)$ signify a measure of distance, indicating the lack of overlap, between the attention maps A_i and A_j. Our first loss aims to minimize that distance (maximize the overlap) over pairs of entity modifiers and their corresponding object attributes (o, a):

$$\mathcal{L}_{pos}(A, S_v) = \sum_{(o,a)\in S_{MB}} \frac{1}{2} M_{dis}(A_o, A_a). \quad (2)$$

For a measure of distance $M_{dis}(A_i, A_j)$ between attention maps, we use a symmetric Kullback-Leibler divergence:

$$M_{dis}(A_i, A_j) = K(A_i||A_j) + K(A_j||A_i); \quad (3)$$

$$K(A_i||A_j) = \sum_{pixels} A_i log(A_i/A_j); \quad (4)$$

where A_i, A_i are attention maps normalized to a sum of 1, i and j are generic indices.

We also construct a loss that compares pairs of modifiers and entity nouns with the remaining words in the prompt, which are grammatically unrelated to these pairs. This loss is defined between words within the (object-nouns, attribute-modifiers) set and words outside of it. Formally, let U_v represent the set of unmatched words obtained by excluding the words in S_{MB} from the full set of words. A_u is the corresponding attention map for a given unrelated word u. The following loss encourages moving apart the correlations between grammatically unrelated pairs of words:

$$\mathcal{L}_{neg} = - \sum_{(o,a)\in S_{MB}} \frac{1}{4|U_v|} \sum_{u\in U_v} D(o, a, u); \quad (5)$$

$$D(o, a, u) = \sum_{u\in U_v} [d(A_o, A_u) + d(A_u, A_a)]; \quad (6)$$

where $d(A_o, A_u)$ is the abbreviation of $M_{dis}(A_i, A_j)$ defined in Equation 3 and 4. Our final loss combines the two loss terms:

$$\mathcal{L} = \alpha_p * \mathcal{L}_{pos} + \alpha_n * \mathcal{L}_{neg}. \quad (7)$$

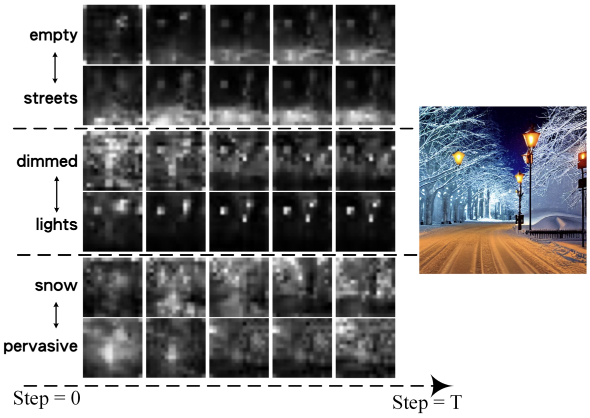


Figure 4: Evolution of cross-attention maps along denoising steps. The attention maps of objects-attribute pairs are initially unrelated, and gradually become intertwined adhering to the expected binding.

Our inference-time optimization approach is inspired by the work of (Chefer et al., 2023; Rassin et al., 2023), which defined a loss over the cross attention maps to update the latents at generation time. However, their loss aims to strengthen the activations of a set of selected tokens or the relations of general entity modifiers, while our loss depends on pairwise relations of metaphorically related words, especially for objects in tenors and attributes in vehicles. Our method aims to align the diffusion process to the underlying groundings of the visual elaborations.

4 Evaluation

Evaluating the visualization of figurative language presents a significant challenge due to its inherently subjective nature. Additionally, current evaluation methodologies vary widely, encompassing image recognition (Yosef et al., 2023), visual entailment (Chakrabarty et al., 2023), as well as retrieval and localization (Akula et al., 2023). Consequently, to thoroughly assess the robustness of GOME, we advocate for an evaluation complemented by diverse automated metrics, together with human evaluations applied at multiple levels of granularity.

4.1 Intrinsic Evaluation

In this section, we evaluate the general figurative language understanding of GOME using the Fig-QA dataset (Liu et al., 2022a). It contains 12k figurative phrases with correct and incorrect interpretations in the Winograd style. For instance, the figurative sentence 'Her word had the strength of a wine glass', is paired with both 'Her promises

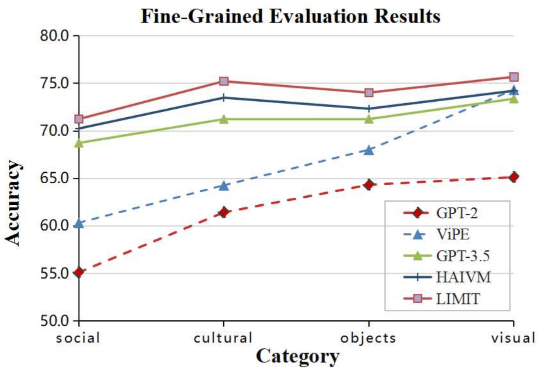


Figure 5: Fine-grained evaluation results on different categories of the Fig-QA dataset. GOME outperforms other models across all categories with a more pronounced gap in the visual category.

can be believed' and 'Her promises cannot be trusted'. This benchmark covers various themes, including common-sense object knowledge, visual metaphors, common-sense social understanding, and cultural metaphors. We employed their evaluation framework for GPT-2 and evaluated the small version trained with the context size of one. Table 1 presents a comparison between the results of GOME and other baselines, as reported by (Liu et al., 2022a), in both zero-shot and fine-tuned contexts. The findings underscore the superiority of GOME over the pre-trained GPT-2 in both scenarios, demonstrating its advanced comprehension. Subsequently, we assess GOME on fine-grained categories within the Fig-QA dataset (Liu et al., 2022a). As illustrated in Figure 5, GOME exhibits a comprehensive understanding across all categories. The significant improvement observed in the visual categories aligns with producing descriptions for metaphors suitable for visualization.

Besides, we also conduct a qualitative experiment to illustrate the effect of metaphor binding in Figure 4. Specifically, we visualize the weights of cross-attention maps mapped to tokens over the denoising steps. The left column displays three pairs of object-attributes to be coupled, including (street, empty), (lights, dimmed), and (snow, pervasive). At the beginning, their weights of aggregated attention maps are initialized based on textual representations from CLIP encoders, as well as the latent image representations. It can be observed that the attention maps of three object-attribute pairs are unrelated regardless of the expected binding, but gradually become intertwined alongside the denoising steps with the proposed modification. More comparisons can be seen in Appendix A.

Settings	Model	ZS	L-Tuned	XL-Tuned
Supervised	GPT-2	54.57	57.13	64.00
	ViPE-S	58.50	61.42	67.28
	GOME-G	59.47	63.02	68.44
Few-shot	GPT-3.5	69.24	-	-
	GOME	74.33	-	-

Table 1: Zero-shot and fine-tuned evaluation results using Fig-QA. L and XL denote the large and X-large variations of the dataset. Our model, GOME-G, demonstrates enhanced comprehension of figurative language compared to other supervised models.

4.2 Extrinsic Evaluation

For a comprehensive end-to-end evaluation, image-to-text and text-to-image retrieval tasks are conducted using the HAIVMet dataset (Chakrabarty et al., 2023). The HAIVMet dataset comprises linguistic metaphors and corresponding visual elaborations, which have been reviewed by experts. Pairs of metaphors and visual elaborations, as well as visual elaborations and images, were created for evaluation purposes. Specifically, one positive image was generated based on visual elaborations, followed by the generation of four negative images per metaphor using Stable Diffusion (Ramesh et al., 2022). Given that HAIVMet includes ground truth visual elaborations, only the negative samples required generation. The negative samples were produced using two methods (Akula et al., 2023): (a) Negative Tenor, which replaces the tenor in the metaphor statement with one from another statement; (b) Negative Vehicle, which replaces the vehicle in the metaphor statement with one from another statement.

After acquiring the relevant images from GPT-3.5, ViPE, HAIVMet, and our own GOME, we applied the fine-tuned version of BLIP (Li et al., 2022) on the COCO (Lin et al., 2014) retrieval dataset. BLIP demonstrates superior performance on vision-language benchmarks by effectively leveraging a multimodal encoder-decoder mixture model, rendering it highly suitable for retrieval evaluation. Our experiments utilized BLIP in both zero-shot and fine-tuned configurations. In the zero-shot setting, the entire retrieval dataset served as the test set, whereas in the fine-tuned setting, 80% of the data was allocated for fine-tuning, with the remaining 20% split equally for validation and evaluation. The mean recall scores across the top-1, top-5, and top-10 retrieval results, as well as the rank

	Setting	Metaphor			Elaboration			Grounding		
		IR↑	TR↑	Rank↓	IR↑	TR↑	Rank↓	IR↑	TR↑	Rank↓
SD & GPT-3.5	zero-shot	46.34	34.13	3.24	72.65	59.32	2.87	73.13	61.31	2.74
	fine-tuned	48.45	34.84	3.11	75.62	61.34	2.71	78.12	62.53	2.63
ViPE	zero-shot	48.23	36.39	3.18	74.72	66.23	2.54	79.72	67.81	2.38
	fine-tuned	52.34	53.17	3.04	80.32	68.44	2.37	81.42	69.61	2.21
HAIVMet	zero-shot	54.23	43.62	3.07	74.25	65.25	2.62	78.27	65.76	2.42
	fine-tuned	**56.92**	51.23	**2.88**	81.32	69.75	2.24	80.54	67.22	2.38
GOME	zero-shot	51.43	42.31	3.13	75.23	69.45	2.37	81.12	72.35	2.31
	fine-tuned	54.25	**52.73**	2.93	**82.55**	**71.22**	**2.21**	**84.37**	**73.78**	**2.17**

For IR and TR, larger values (↑) are better. For Rank, lower values (↓) are better.

Table 2: A comparative report on image-text and text-image retrieval using corpora generated by GPT-3.5, GOME, and human experts (HAIVMet dataset) in zero-shot (zs) and fine-tuned (ft) settings. TR and IR denote the mean image-to-text and text-to-image retrieval scores respectively. GOME outperforms GPT-3.5 and shows competitive understanding to human experts.

of searching images based on text, are presented in Table 2. GOME surpasses GPT-3.5, ViPE, and HAIVMet in image-metaphor retrieval (the first TR column in the table). However, despite its advantage over other baselines, GOME slightly underperforms compared to human experts in metaphor retrieval from images (the first IR column in the table). This discrepancy may stem from the overspecification, with which human experts describe metaphorical images (Chakrabarty et al., 2023) based more on objects, resulting in a more discrete feature space that BLIP can interpret more easily. Furthermore, we conducted similar evaluations on pairs of images and visual elaborations, as well as groundings, instead of metaphors, to evaluate the alignment between the elaborations and their corresponding images, similar to image-metaphor retrieval. As illustrated in the right columns of Table 2, GOME surpasses SD & GPT-3.5 and human experts in both zero-shot and fine-tuned scenarios. Notably, while ViPE demonstrates lower performance, it still exhibits superior results to humans in image-grounding retrieval. This observation implies that HAIVMet emphasizes the visualizability of its generated elaborations with a robust link to the objects instead of underlying groundings. Conversely, GOME not only achieves comparable or even superior evaluations in image-metaphor and image-elaboration related tasks compared to HAIVMet, but also produces more compelling visual elaborations faithful to original meanings, as indicated by its high average recall and ranking scores in the tasks of image-grounding retrieval (The rightest three columns in Table 2).

4.3 Human Evaluation

To realize a comprehensive evaluation, a study was undertaken involving three participants, aged 20 to 30, who were experts in metaphor analysis. From the HAIVMet dataset, one hundred metaphors were randomly selected. Visual elaborations for each metaphor were produced using ChatGPT and GOME, alongside additional elaborations from human experts within the HAIVMet dataset. Subsequently, these visual elaborations were utilized to generate corresponding images using Stable Diffusion. The experiment presented participants with a metaphor alongside three images generated from prompts by human experts (HAIVMet dataset), ChatGPT, and GOME.

The participants are instructed to complete two missions: (a) select the image that best reflects the metaphor's literal meaning based on objects; (b) select the image that best reflects the metaphor's underlying meaning based on groundings. Accord to the results of Task (a), participants preferred visual metaphors from human experts 37.82% of the time, followed by those from GOME at 31.32%, and ChatGPT at 30.86%. While in the case of Task (b), which accesses visualizations based on groundings, participants preferred images from GOME at 36.43% of the time, followed by those from GOME at 35.15%, and ChatGPT at 28.42%. These results confirm GOME's superiority over the direct collaboration of Stable Diffusion and ChatGPT, and demonstrate its competitive performance relative to human experts, especially for faithfully depicting the underlying groundings of linguistic metaphors.

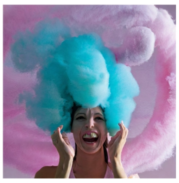
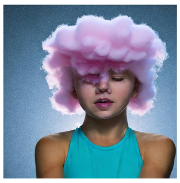
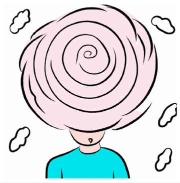
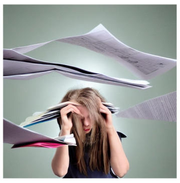
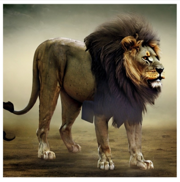
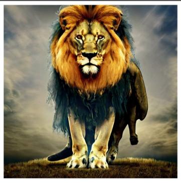
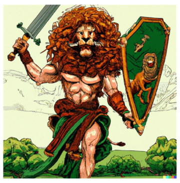
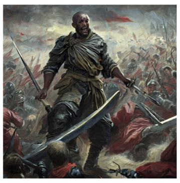
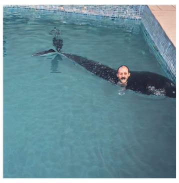
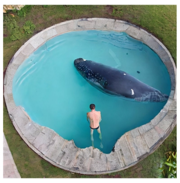
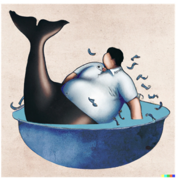
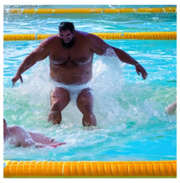
Metaphor	Grounding	Stable Diffusion	DALL•E 2	HAIVM	Ours
After 10 minutes your head becomes like ***spinning cotton candy***	***Confusion*** *or being* ***overwhelmed***				
He was a ***lion*** *on the battlefield*	*The* ***soldier*** *is* ***brave*** *and* ***formidable*** *in battle*				
The guy was a ***floating whale*** *in the small swimming pool*	*The guy is overly* ***large*** *or* ***clumsy*** *for the space*				

Figure 6: Examples of metaphor illustration through different methods. Previous methods focused on objects to be included in the metaphor, while our method focuses more on the underlying groundings. It can be observed that excessive cretization of objects, especially for thought-provoking vehicles in the source domain, may diminish the metaphor's original meaning, becoming overly straightforward.

In Figure 6, we show examples of visualization generated using linguistic metaphors or their visual elaborations as prompts for the text-to-image model. We observe that our method, where CoT prompting based on groundings is involved, is of higher quality. For instance, a good visual metaphor for the metaphorical expression 'After 10 minutes your head becomes like spinning cotton candy' would reflect the underlying meanings, which indicates a feeling of confusion or overwhelmed by taking 'spinning cotton candy' as the vehicle in the original textual statements. Other methods just simply stack multiple objects together, such as people, heads, and spinning cotton candy, neglecting the true meaning of confusion or being overwhelmed. While in our method, the genuine underlying meaning is captured with CoT prompting and systematic knowledge, which transform the abstract object or concept into a specific scenario, in which a student is surrounded by flying papers, with a frustrated emotion on her face to show the overwhelmed feeling.

The observations are similar to the metaphors in other samples, such as transforming the 'lion' into a brave soldier, and 'floating whale' into an 'overly large man'. Obviously, we are not expecting a real lion or whale presented in visual illustrations. These vehicles play the role of secondary objects, emphasizing some attributes of primary objects. The implicit meaning in metaphors is well captured by our model, and depicted in the final picture. We also discover some cases hard to visualize, such as metaphors with extreme subject feelings, or abstract attributes blended in verbal expressions. More discussions are provided in Appendix B.

5 Conclusion

In this paper, we introduced GOME, the first model with linguistic binding for visualizing metaphors from the grounding perspective. Our research notices the problem of over-literalization for the first time, and solves this issue through conceptual elaborations for binding implicit metaphorical attributes, rather than their presentation. Overall, our contributions are the following: firstly, a grounding-based depiction method is proposed for accurately binding metaphorical attributes. Secondly, a dataset with conceptual elaborations of metaphors is introduced, encompassing fine-grained metaphorical elements such as tenor, vehicle, and groundings. Finally, extensive experiments validate the fidelity of our method in capturing the underlying meaning of metaphors. In future work, we plan to employ GOME with knowledge from other related fields, such as cognitive science.

6 Limitations

While we offer evidence of GOME's effectiveness and understanding of figurative language across various benchmarks, we have to acknowledge potential limitations. There is still room for improvement in LLM elaboration by training a domain-specific LLM for figurative language, which is a common challenge in metaphor analysis, and not fully solved in this work, due to the limited computational and data resources. Additionally, the selection of evaluations, including metrics, and datasets chosen for assessment may not comprehensively capture the subtleties inherent in human figurative languages. For example, the cultural variations in the creation, and the subjectivity in interpreting figurative phrases, pose a significant consideration. Further investigation and comparative analysis utilizing a broader range of tasks, measurements, and datasets, may enhance the ability of GOME.

References

Arjun R. Akula, Brendan Driscoll, Pradyumna Narayana, Soravit Changpinyo, Zhiwei Jia, Suyash Damle, Garima Pruthi, Sugato Basu, Leonidas J. Guibas, William T. Freeman, Yuanzhen Li, and Varun Jampani. 2023. Metaclue: Towards comprehensive visual metaphors research. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, pages 23201–23211. IEEE.

Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. 2017. CVAE-GAN: fine-grained image generation through asymmetric training. In IEEE International Conference on Computer Vision, ICCV 2017, pages 2764–2773. IEEE Computer Society.

Yuri Bizzoni and Shalom Lappin. 2018. Predicting human metaphor paraphrase judgments with deep neural networks. In Proceedings of the Workshop on Figurative Language Processing, Fig-Lang@NAACL-HLT 2018, pages 45–55. Association for Computational Linguistics.

Max Black et al. 1979. More about metaphor. Metaphor and thought, 2:19–41.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020.

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 2020, pages 6455–6469. 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. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, pages 6848–6863. Association for Computational Linguistics.

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, EMNLP 2022, pages 7139–7159. Association for Computational Linguistics.

Tuhin Chakrabarty, Arkadiy Saakyan, Olivia Winn, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki, and Smaranda Muresan. 2023. I spy a metaphor: Large language models and diffusion models co-create visual metaphors. In Findings of the Association for Computational Linguistics: ACL 2023, pages 7370–7388. Association for Computational Linguistics.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. 2023. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. ACM Trans. Graph., 42(4):148:1–148:10.

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 2019, page 172. ACM.

Donald Davidson. 1984. What metaphors mean." inquiries into truth and interpretation.

Qianyu He, Sijie Cheng, Zhixu Li, Rui Xie, and Yanghua Xiao. 2022a. Can pre-trained language models interpret similes as smart as human? In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, pages 7875–7887. Association for Computational Linguistics.

Xuanli He, Islam Nassar, Jamie Kiros, Gholamreza Haffari, and Mohammad Norouzi. 2022b. Generate, annotate, and learn: NLP with synthetic text. Trans. Assoc. Comput. Linguistics, 10:826–842.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. To appear, 7(1):411–420.

Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A. Smith, and Jiebo Luo. 2022. Promptcap: Prompt-guided task-aware image captioning. CoRR, abs/2211.09699.

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. In 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, pages 1100–1110. IEEE Computer Society.

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, AACL/IJCNLP 2022, pages 1172–1184. Association for Computational Linguistics.

Haifaa Jahameh and Aseel Zibin. 2023. The use of monomodal and multimodal metaphors in advertising jordanian and american food products on facebook: A comparative study. Heliyon, 9(5).

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. CoRR, abs/2210.10578.

G LAKOFF. 1993. The contemporary theory of metaphor. Metaphor and Thought.

Evelina Leivada, Elliot Murphy, and Gary Marcus. 2022. DALL-E 2 fails to reliably capture common syntactic processes. CoRR, abs/2210.12889.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In International Conference on Machine Learning, ICML 2022, volume 162 of Proceedings of Machine Learning Research, pages 12888–12900. PMLR.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In Computer Vision - ECCV 2014 - 13th European Conference, volume 8693 of Lecture Notes in Computer Science, pages 740–755. Springer.

Changsheng Liu and Rebecca Hwa. 2016. Phrasal substitution of idiomatic expressions. In NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 363–373. The Association for Computational Linguistics.

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022a. Testing the ability of language models to interpret figurative language. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, pages 4437–4452. Association for Computational Linguistics.

Vivian Liu, Tao Long, Nathan Raw, and Lydia B. Chilton. 2023a. Generative disco: Text-to-video generation for music visualization. CoRR, abs/2304.08551.

Vivian Liu, Han Qiao, and Lydia B. Chilton. 2022b. Opal: Multimodal image generation for news illustration. In The 35th Annual ACM Symposium on User Interface Software and Technology, UIST 2022, pages 73:1–73:17. ACM.

Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, Dajiang Zhu, Xiang Li, Ning Qiang, Dinggang Shen, Tianming Liu, and Bao Ge. 2023b. Summary of chatgpt/gpt-4 research and perspective towards the future of large language models. CoRR, abs/2304.01852.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with CLIP latents. CoRR, abs/2204.06125.

Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. 2023. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. In Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023.

Ali Razavi, Aäron van den Oord, and Oriol Vinyals. 2019. Generating diverse high-fidelity images with VQ-VAE-2. In Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, pages 14837–14847.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, pages 10674–10685. IEEE.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi.

2022. Photorealistic text-to-image diffusion models with deep language understanding. In Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022.

Angela Schwering, Kai-Uwe Kühnberger, Ulf Krumnack, Helmar Gust, Tonio Wandmacher, Bipin Indurkhya, and Amitash Ojha. 2009. A computational model for visual metaphors. interpreting creative visual advertisements. In ICAART 2009 - Proceedings of the International Conference on Agents and Artificial Intelligence, pages 339–344. INSTICC Press.

Hassan Shahmohammadi, Adhiraj Ghosh, and Hendrik P. A. Lensch. 2023. Vipe: Visualise pretty-much everything. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, pages 5477–5494. Association for Computational Linguistics.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Agüera y Arcas, Claire Cui, Marian Croak, Ed H. Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications. CoRR, abs/2201.08239.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, 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, ACL 2014, pages 248–258. The Association for Computer Linguistics.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? GPT-3 can help. In Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021, pages 4195–4205. Association for Computational Linguistics.

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. CoRR, abs/2304.09653.

Ron Yosef, Yonatan Bitton, and Dafna Shahaf. 2023. IRFL: image recognition of figurative language. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 1044–1058. Association for Computational Linguistics.

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, NAACL-HLT 2019, pages 861–871. Association for Computational Linguistics.

Jiali Zeng, Linfeng Song, Jinsong Su, Jun Xie, Wei Song, and Jiebo Luo. 2020. Neural simile recognition with cyclic multitask learning and local attention. In The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, pages 9515–9522. AAAI Press.

Linhao Zhang, Li Jin, Guangluan Xu, Xiaoyu Li, Cai Xu, Kaiwen Wei, Nayu Liu, and Haonan Liu. 2024. CAMEL: capturing metaphorical alignment with context disentangling for multimodal emotion recognition. In Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, pages 9341–9349. AAAI Press.

Jianing Zhou, Hongyu Gong, and Suma Bhat. 2021. PIE: A parallel idiomatic expression corpus for idiomatic sentence generation and paraphrasing. In Proceedings of the 17th Workshop on Multiword Expressions, MWE@ACL-IJCNLP 2021, pages 33–48. Association for Computational Linguistics.

A Effectiveness of Metaphor Binding

This part verifies the effectiveness of Metaphor Binding through a qualitative ablation. Specifically, we drop the binding loss during the text-to-image process, while other settings remains unchanged, to generate two illustrations for the metaphorical statement 'A blanket of snow covered the streets'. The weights of cross-attention maps linked to tokens over the denoising steps are visualized in Figure 7. The top part displays the final illustration results, with the pairs of object-attributes to be coupled displayed at the left part, such as (street, empty), (lights, dimmed), and (snow, pervasive). At the beginning, it can be observed that the attention maps of three object-attribute pairs are both unrelated. However, the left part with the proposed binding method gradually becomes intertwined alongside the denoising steps, while in the right part without a binding process, the attention maps remain unrelated. Finally, the left image is obviously more compelling and faithful to the original metaphor. The comparison can partly demonstrate the effectiveness of the metaphor binding.

B More Discussion About Cases

This part displays more cases of illustrations through GOME in Figure 8 and 9. Despite the compelling images, there are also some controversial cases between our model and experts in HAIVM in Figure 10. One situation is for metaphors with ambiguous emotional state. For instance, in the first metaphor of Figure 10, although the vehicle 'cold iron' is correctly perceived in the grounding, with an interpretation of 'the man is unfeeling and emotionally cold'. However, the emotional state of 'unfeeling' is hard for text-to-image models to visualize, 'emotionally cold' is also misunderstood as the states of weather, with snow presented in the picture. In this case, the depiction of HAIVM is better to depict an 'iron heart'. Another instance is a combination of multiple metaphors, like 'The teacher planted the seeds of wisdom', while our model converts 'planted' into the action of 'teaching in front of the blackboard' to show the teacher is nurturing and educational, the expression of 'seeds of wisdom' is not fully represented. These cases shows that our model still have potential limitation in comprehensively capturing the subtleties of metaphors inherent in human figurative languages.

C Implementation Details

Knowledge Distilation for GOME-G: We employ the small version of GPT-2, which is finetuned on GOME corpus for 5 epochs using 1 v100 Nvidia GPU with 32 GB RAM. We use the AdamW optimizer with a learning rate of 5e-05 and a linear scheduler with 1000 warmup steps. For GOME-G, the batch size is generally 48. 10% of the samples is used for validation.

Image-text Retrieval: We load a BLIP checkpoint trained on COCO, initialized on ViT-B and BERT-base. To fine-tune the model, we use a batch size of 16 for 10 epochs using AdamW, a learning rate of 1e -4, and a batch size of 128 with reranking for fast inference, commonly used in retrieval.

Figurative QA: We made use of the provided evaluation framework, and trained with the batch size of 32 for 5 epochs using AdamW optimizer, with a learning rate of 5e-5. As the original leaderboard is not available, we make an evaluation on the validation set.

D Linguistic Metaphor Collection

Numerous explorations have been conducted to collect linguistic metaphors (Hussain et al., 2017; Chakrabarty et al., 2022a) or figurative expressions (Chakrabarty et al., 2022a; Bizzoni and Lappin, 2018). Following previous work (Chakrabarty et al., 2023), we extended the annotations of 1351 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, CrossLing Metaphors (Tsvetkov et al., 2014) and Metaphor Paraphrase (Bizzoni and Lappin, 2018). It is worth noting that not all linguistic metaphors can be rendered as visual metaphors, some figurative expressions involve much cultural specificity or deep emotional states are difficult to depict visually. To overcome this challenge, we apply a pre-processing pipeline to filter original collections. Our pipeline mainly considers the following aspects: Diversity: Duplicated metaphors are removed, which can be measured by the sequence similarity based on difflib. Brevity: Sentences exceeding 30 words in total are excluded to maintain conciseness, which are important for avoiding under-specification (Hutchinson et al., 2022). Visualizability: metaphors should be easily described in a visual form. We remove metaphors with extremely emotional states and nu-

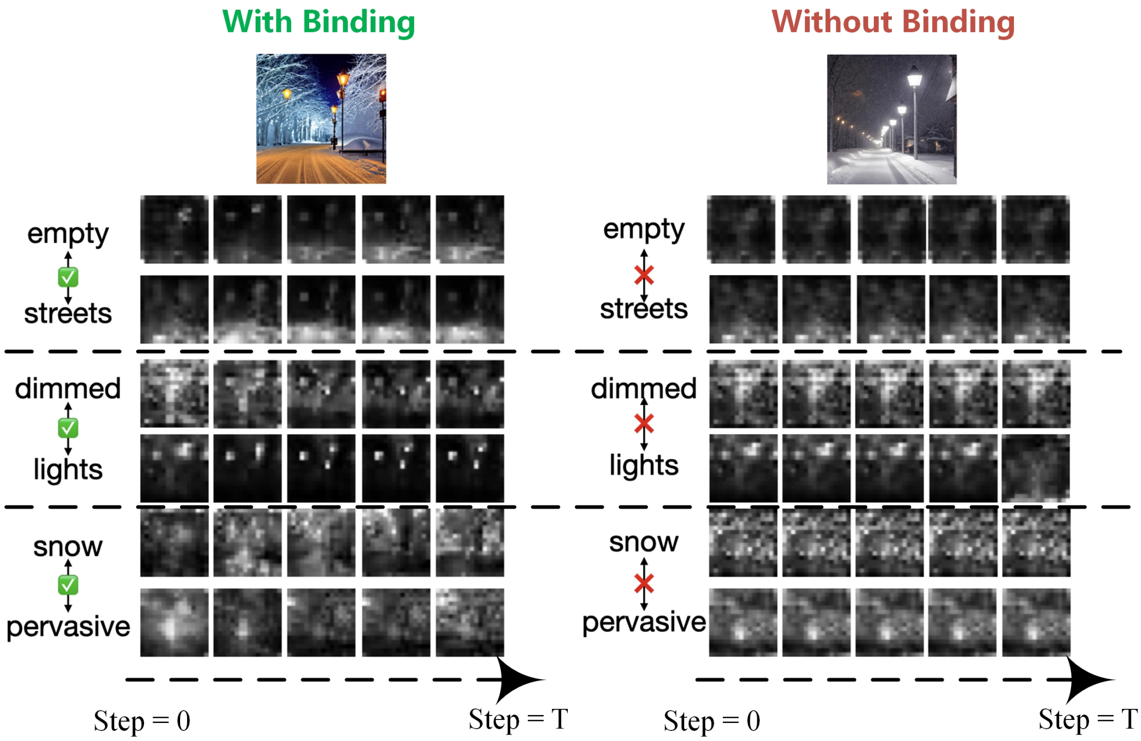


Figure 7: Effectiveness of Metaphor Binding

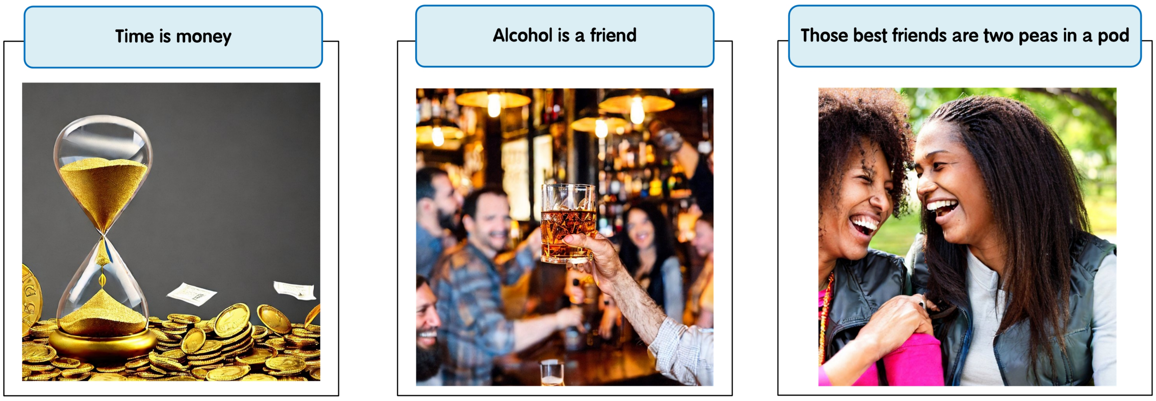


Figure 8: Illustrations of classic metaphors (smilies) through GOME.

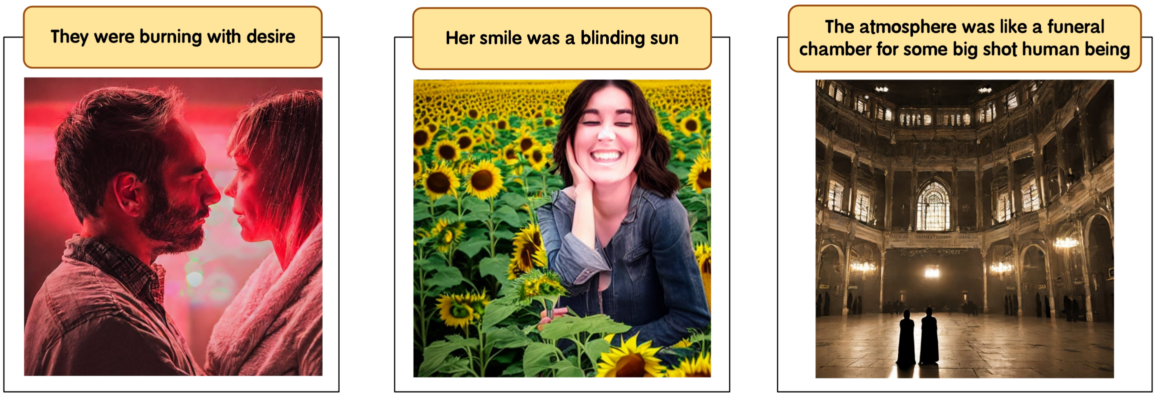

They were burning with desire
Her smile was a blinding sun
The atmosphere was like a funeral chamber for some big shot human being

Figure 9: Illustrations of emotion-related metaphors through GOME.

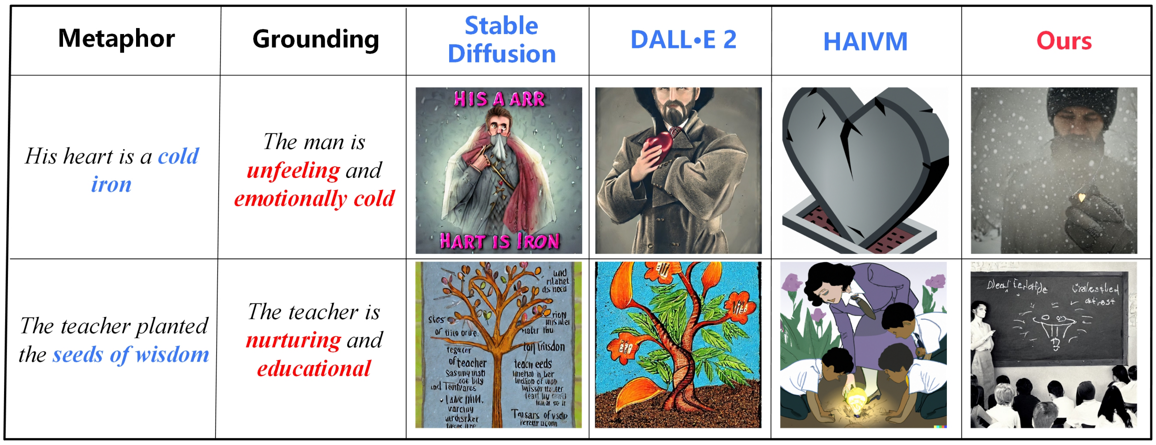

Metaphor
Grounding
Stable Diffusion
DALL•E 2
HAIVM
Ours
His heart is a cold iron
The man is unfeeling and emotionally cold
HIS A ARR
HART IS IRON
The teacher planted the seeds of wisdom
The teacher is nurturing and educational

Figure 10: Controversial situations during human evaluations

merical expressions hard for text-to-image models to depict, such as ‘five little monkeys’, recognized by a Qwen model with 72B parameters.

E System Role

This part introduces the System Role for instructing GPT3.5, which generates visual elaboration for a given metaphor. We refer to some commands from (Shahmohammadi et al., 2023)

Your are an expert aware of linguistic metaphors, who is able to elaborate a metaphor with rich visual details with no more than 30 words. The outputs should include perceived concepts and objects that served as tenors, while trying to exclude objects or concepts served as vehicles unless they are necessary to enable the metaphor to align with common sense. Determine the overall visual setting and environmental style based on the conceptual groundings in the metaphor. Obey the following commands:

1. Convert abstract metaphors into depictable prompts that represent the original lines. Visual details should be elaborated in the outputs, along with the provided objects to be included.

2. Consider the conceptual groundings of the metaphor when generating prompts. The same line should be represented differently depending on the groundings of the metaphor. For example, "love is like a gust of wind" could be converted to "two lovers embracing each other in a sunny field, their hair and clothes gently blown by a soft breeze." if the grounding is "love is gentle", or "In a park during autumn, a couple broke up. The woman left, and a man reached out his arm to grab her hand. As a gust of wind blew by their side, it swept away the fallen leaves." if the grounding is "love is a brief passage".

3. Use concrete objects to represent abstract concepts or unspecific expressions, which are difficult to visualize, such as using "a man and a woman are having a conversation over a cup of tea" to represent "somebody once told me" and "a shining diamond ring" to represent "all that glitters is gold."

4. When generating prompts, do not focus on what the subject is thinking or feeling. For example, instead of "a student thinking about his long assignment list, overwhelmed by so much coursework," which is difficult to visualize, describe the student appearance, such as "a male student looking at a long assignment list, with a scared expression, tears rolling down from his cheek."

5. Structure all prompts by setting a scene with at least one subject and a concrete action term, followed by a comma, and then describing the scene. For instance, "a view of a forest from a window in a cozy room, leaves are falling from the trees."

6. To add variety and avoid repetition, it is important to mix up singular and plural forms when referring to subjects or objects in the prompts. For example, "two cats," "ten men," "five girls," or "seven books" can be used instead of consistently using singular forms.

7. Do not use generic words such as person, people, man, woman, individual, figure, object, etc. Instead, across various topics, use diverse and specific terms such as desert, island, statue, skyscraper, stars, moon, rainbow, snowflakes, wolf, horse, dragon, bird, python, bike, truck, airplane, astronaut, daisies, roses, diamond ring, and so on, where appropriate.