Looking Beyond the Pixels: Evaluating Visual Metaphor Understanding in VLMs

Manishit Kundu[1], **Sumit Shekhar**[2], **Pushpak Bhattacharyya**[1]
[1]Indian Institute of Technology Bombay, India
[2]Adobe Systems, India
manishit@minds.iitb.ac.in, sushekha@adobe.com, pb@cse.iitb.ac.in

Abstract

Visual metaphors are a complex vision–language phenomenon that requires both perceptual and conceptual reasoning to understand. They provide a valuable test of a model's ability to interpret visual input and reason about it with creativity and coherence. We introduce ImageMet, a visual metaphor dataset, featuring 2177 synthetic and 350 human-annotated images. We benchmark several SOTA VLMs on two tasks: Visual Metaphor Captioning (VMC) and Visual Metaphor VQA (VM-VQA). We establish strong baselines by fine-tuning on ImageMet, which yields substantial performance gains in VMC (+4.67% SBERT-Similarity, +4.84% task-specific metric) and VM-VQA (+9.3% Accuracy on average). Additionally, we introduce a task-specific CoT prompting strategy that outperforms standard few-shot baselines (+1.99% in VMC, +5.21% in VM-VQA). We observe that despite strong performance on the VMC task, VLMs still significantly lag behind humans in understanding visual metaphors, indicating that their success often relies on learned associations rather than genuine analytical reasoning. We note that this gap is often obscured in metaphor captioning tasks where the automatic metrics correlate only moderately at best with human judgment (Pearson r < 0.6), highlighting the need for careful, holistic evaluation of the visual metaphor understanding of the models.

1 Introduction

Visual metaphors are figurative pictorial comparisons in which one entity (primary concept or target domain) is represented in terms of another (secondary concept or source domain). They are monomodal in nature, i.e. both the target and source domains rely entirely on visual cues to convey metaphorical meaning. Due to their visual appeal and cognitive impact, visual metaphors are widespread in advertising (Mick and Mcquarrie, 1999), political cartoons, and contemporary art, making them a vital component of nonverbal communication and cultural expression. In this work, we focus exclusively on visual metaphors, leaving out multimodal metaphors (where textual and visual clues jointly construct the metaphor's meaning). This restriction allows us to isolate and focus on the challenges posed by visual metaphors, stemming from their monomodal nature and the difficulty of interpreting meaning solely from visual cues. For more on metaphors, see Appendix A.

Visual metaphor understanding involves two key challenges (Forceville, 2008). First, visual metaphors lack the syntactic structure of language, making it harder to identify source and target elements. Viewers must rely solely on visual cues like composition, emphasis, and familiarity to interpret the mapping. Second, to analyze visual metaphors conceptually, the visual relation must be translated into a verbal "*A IS B*" format. This translation is interpretive and often ambiguous, as different verbalizations can emphasize distinct aspects of the concepts, which may alter the metaphor's overall meaning. Visual metaphor understanding requires perceptual reasoning to infer target-source mappings and conceptual reasoning to translate them into a verbal "*A IS B*" format.

Thus, visual metaphor understanding offers an interesting probe into the multimodal reasoning abilities of Vision-Language Models (VLMs). While prior work in visual metaphors (Yosef et al., 2023; Chakrabarty et al., 2023) has addressed visual metaphor detection, retrieval, and generation, metaphor understanding remains largely unexplored. We introduce two tasks to evaluate this: (i) Visual Metaphor Captioning (VMC), where models must generate coherent captions that capture the intended meaning of a given visual metaphor, and (ii) a Visual Metaphor VQA task (VM-VQA) that assesses a model's ability to infer target-source mappings in a multiple-choice

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 23137–23158
November 4-9, 2025

setting.

Our contributions are:

1. **ImageMet**: The largest visual metaphor dataset of its kind, containing 2177 synthetic image–simile pairs and a manually annotated test set of 350 images for challenging evaluation. Fine-tuning on ImageMet demonstrates significant improvement in Visual Metaphor Captioning (+4.67% SBERT-Similarity, +4.84% task-specific metric) and Visual Metaphor VQA (+9.3% Accuracy on average) (Section 3).

2. **Benchmarking of state-of-the-art VLMs** on the proposed test set for the two visual metaphor tasks (VMC and VM-VQA) which, to the best of our knowledge, is the first of its kind. We compare 4 open-source models, 2 closed-source models, and human performance on ImageMet, providing both quantitative and qualitative analysis (Section 5, 6).

3. **Strong baselines for visual metaphor understanding**, featuring a task-specific CoT prompting technique that begins with objective, perceptual questions and gradually shifts to conceptual ones. This method outperforms standard zero-shot and few-shot setups (+1.99% in VMC, +5.21% in VM-VQA) and achieves the strongest overall performance with GPT-4o. Fine-tuning on ImageMet yields the best open-source results (Section 4.2, 5).

We release our code and data to support further research in this area [1].

2 Related Works

Textual Metaphors: Significant progress has been made in understanding (Aghazadeh et al., 2022), detecting (Choi et al., 2021; Su et al., 2020; Badathala et al., 2023), and generating (Stowe et al., 2021; Chakrabarty et al., 2020) linguistic metaphors, supported by sentence-level and token-level datasets (Tsvetkov et al., 2014; Mohammad et al., 2016; Mohler et al., 2016).

Visual Metaphors: Akula et al. (2023) extended metaphor understanding to the visual domain by introducing classification, generation, and captioning tasks. Chakrabarty et al. (2023) explored classification and generation, while Rajakumar Kalarani et al. (2024) focused on video metaphor captioning. However, image-based visual metaphor understanding remains largely unexplored. Prior work (Akula et al., 2023) evaluated only a single model using limited template-based metrics. We present the first comprehensive benchmark of SOTA VLMs on visual metaphor understanding in images, incorporating both automatic and human evaluations.

Multimodal Metaphor Datasets: Existing datasets such as MetaCLUE (Akula et al., 2023) and MultiMET (Zhang et al., 2021) are not publicly available and were therefore not used. We primarily compare our dataset, ImageMet, against VFLUTE (Saakyan et al., 2024), as it is a curated subset of both IRFL (Yosef et al., 2023) and HAIVMet (Chakrabarty et al., 2023). VFLUTE includes 806 training, 116 validation, and 103 test examples combining metaphor and simile cases. However, the text-image pairs in VFLUTE predominantly feature *non-deliberate metaphors*, where figurative elements are incidental rather than central, making them suboptimal for evaluating visual metaphor understanding, where the metaphor must be visually expressible. In contrast, ImageMet focuses exclusively on *deliberate metaphors*, where the metaphor forms the core meaning and is explicitly grounded in a visual comparison. To the best of our knowledge, ImageMet is the only *openly available dataset* specifically designed for evaluating visual metaphor understanding in this deliberate sense. Expanded comparison with existing datasets in Appendix C.

3 Dataset: ImageMet

We introduce ImageMet, a visual metaphor dataset developed with two key objectives: (1) to provide a challenging, manually-annotated test set containing clearly intended visual metaphors, thereby minimising ambiguity and subjectivity; and (2) to supply synthetic training data for fine-tuning models to improve their visual metaphor understanding.

For the captions, we adopt a strict simile-based template—*"<Primary Concept> is as <Attribute> as <Secondary Concept>"*—inspired by prior work (Akula et al., 2023; Rajakumar Kalarani et al., 2024). This format explicitly decomposes metaphorical comparisons into three components: the *Primary* (target domain), the *Secondary* (source domain), and the shared *Attribute*. During fine-

[1] https://github.com/manishitIITB/Visual_Metaphor_ImageMet_EMNLP2025

Figure 1: **Examples from the ImageMet dataset:** (i) The train and validation sets (first and second row) consist of synthetically generated instances, where a linguistic simile is first generated, followed by an image created through its visual elaboration. (ii) The test set (last row) consists of manually curated and annotated images with intended visual metaphors to minimise subjectivity.

tuning, this structured format allows for more efficient learning than open-ended metaphorical expressions (Appendix B). During evaluation, it facilitates reliable automatic assessment and emphasises the importance of verbalising the correct concepts, an essential aspect of visual metaphor understanding (Forceville, 2008). This explicit verbalisation challenges models to go beyond surface-level pattern recognition and demonstrate true perceptual reasoning and conceptual grounding. The train and the validation sets were synthetically generated using the following *automated pipeline*:

- **Primary Concepts:** Inspired by Yang et al. (2020), we curate nouns from a list of 4000 words sourced from the MRC Psycholinguistic Database (Coltheart, 1981; Wilson and Division, 1997), selecting those with imageability and concreteness scores above 500 (PAIVIO et al., 1968) to exclude non-visual concepts (see Appendix H).

- **Attributes:** We extract adjectives associated with each primary noun from WordNet 3.0 (Miller, 1994) using the 'examples' and 'hyponyms' attributes where available. Additionally, we prompt GPT-4o (OpenAI, 2023) to generate the top three most associated adjectives for each word, ensuring diverse semantic coverage (see Appendix G).

- **Secondary Concepts:** We prompt GPT-4o to complete a template of the form "The *Primary* is as *Attribute* as _____", allowing for

grammatical variations. Among the generated candidates, we select the one with the lowest perplexity score as the final synthetic simile for the (Primary, Attribute, Secondary) tuple. To ensure metaphorical quality, a generation is retained only if the primary and secondary concepts are semantically dissimilar (cosine similarity < 0.5).

- **Images:** For each synthetically generated simile, we create a corresponding synthetic image using Stable Diffusion 3.5 Large (8B)[2]. To achieve this, we transform the simile into a detailed visual elaboration, following prior works (Zhang et al., 2024; Chakrabarty et al., 2023).This elaboration is generated using GPT-4o, which first classifies the primary and secondary concepts as concrete or abstract, then follows a corresponding set of predefined instructions to generate a coherent visual elaboration. For a detailed breakdown of the prompting strategy, see Appendix I.

To ensure consistency and reduce reliance on a single LLM, we decompose simile generation into smaller controlled steps, incorporating external signals where needed. Our train and validation sets consist of **2000** and **177** instances, respectively, each pairing a synthetic image with a synthetic simile caption.

Synthetic Data Validation: We sampled 10% of similes from ImageMet and an equal number from VFLUTE and asked human annotators to rate the coherence of the similes on a 5-point scale. ImageMet achieved an average score of **4.19**, while VFLUTE scored **4.25**. We also evaluated 10% of image-text pairs from both datasets for alignment quality using the same 5-point scale. In image-text alignment, ImageMet scored **4.17** and VFLUTE **4.53**. These results demonstrate that, despite being synthetically generated, ImageMet instances are comparable in coherence and alignment to VFLUTE's human-written content, validating both our simile and image generation pipelines.

Test set: It contains **350 manually curated images** featuring clearly intended visual metaphors, each paired with a simile-style caption. Images were scraped using targeted queries such as "metaphor" and "symbolism". Two Master's students, fluent in English and experienced in linguistic annotation, independently wrote concise captions in our preferred simile format. In Round 1, each annotator worked independently. In Round 2, they met to resolve disagreements. If consensus was reached, the agreed caption was retained; otherwise, the instance was discarded. This framework was inspired by the two-round protocol of Fabbri et al. (2021). Annotation guidelines are detailed in Appendix J.

[2]stabilityai/stable-diffusion-3.5-large

4 Methodology

4.1 Tasks

To evaluate models' understanding of visual metaphors, we define two tasks: (i) **Visual Metaphor Captioning (VMC)** and (ii) **Visual Metaphor VQA (VM-VQA)**.

(i) Visual Metaphor Captioning (VMC): Given an image containing a visual metaphor, the model must generate a caption that coherently verbalises the metaphor. We explore two settings:

- **Open Format:** The model generates a free-form caption. We evaluate it using SBERT similarity with the ground-truth caption.
- **Simile Format:** The model generates a caption in the form "A is as B as C", which forces explicit verbalization of the metaphor's components. This tests the deeper perceptual and conceptual reasoning and the inherent verbalisation as argued by Forceville (2008). We evaluate using BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), CIDEr (Vedantam et al., 2015), SBERT-similarity (Reimers and Gurevych, 2019), and a task-specific metric: Average Cosine Distance (ACD) (Rajakumar Kalarani et al., 2024), which weights BERT-Score (Zhang et al., 2020) by the dissimilarity between the primary and secondary concepts (See Appendix D).

(ii) Visual Metaphor VQA (VM-VQA): The model is shown an image and asked to identify one of the three metaphor components: the **Primary**, **Secondary**, or **Attribute**. For each component, it is given multiple options to select from. The options are framed as follows:

- For the **Primary** or **Secondary** components, the model is given three options: the correct answer and two distractors, which are the remaining metaphor components (converted to noun form).

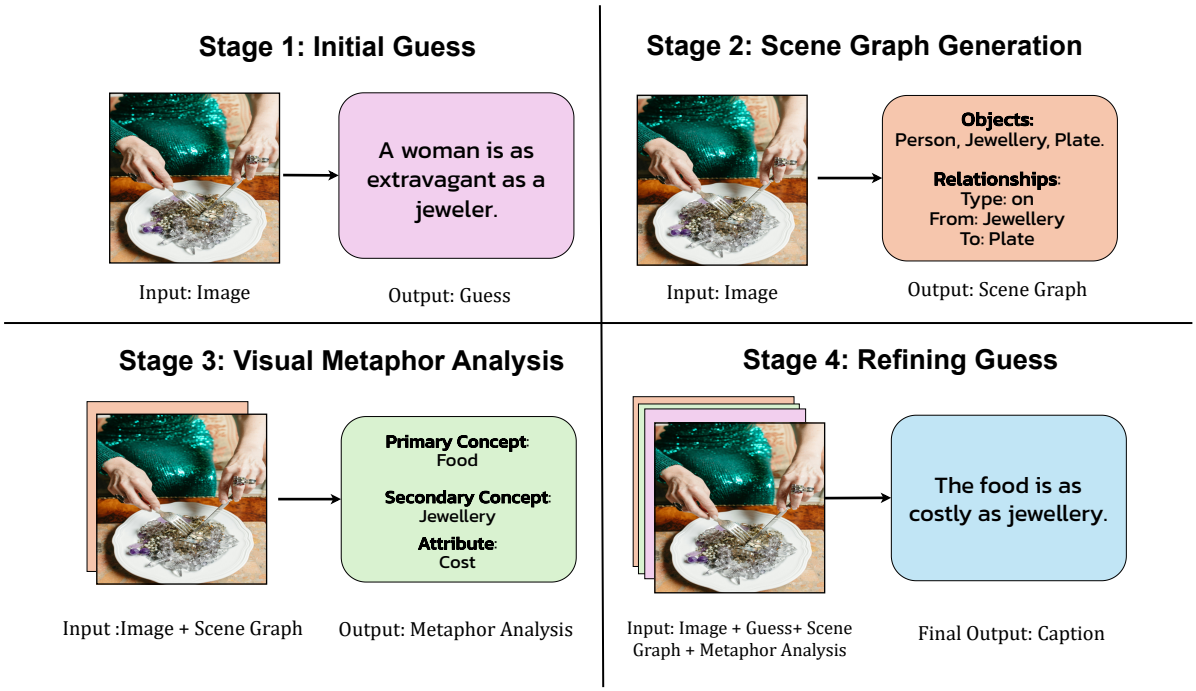

Figure 2: **Our task-specific CoT pipeline** consists of four steps: (1) Initial Model Guess, where the model provides a preliminary interpretation of the visual metaphor based on the image alone, (2) Scene Graph Generation, capturing object arrangements in the image to extract structural relationships, (3) Visual Metaphor Analysis using CoT, progressing from objective observations to abstract reasoning to identify the primary, secondary, and attribute, leveraging both the image and the generated scene graph, and (4) Refinement, where the model revisits its initial guess, incorporating insights from the scene graph and CoT analysis to enhance its final prediction.

- For **Attribute**, we provide five adjective options. Challenging distractors are selected through a multi-step process. First, we generate a verbose literal caption for the image and extract all adjectives from it. Next, we compute the cosine similarity between each extracted adjective and the ground-truth attribute. We then filter out adjectives that are either too dissimilar (cosine similarity < 0.2) or too similar (> 0.8) to the ground-truth attribute, as these are unlikely to function as effective distractors. From the remaining candidates, we select the top-4 most similar adjectives to serve as distractors.

We report model accuracy for each metaphor component.

4.2 Approaches

Prompting Techniques: We evaluate the following prompting strategies:

- **Zero-Shot:** Models are first evaluated in a zero-shot setting using a standard inference prompt (details in Appendix I). Model names without a suffix tag indicate zeroshot.
- **Multishot with Explanations (MS):** For few-shot capable models, we use a 4-shot prompt based on visual metaphor theory (Ojha, 2013), covering core types: (1) juxtaposition, (2) fusion, (3) abstract source, and (4) tangible source. Each shot includes an image, its simile caption, and an explanation. Prompt details are in Appendix I.
- **Chain-of-Thought Reasoning (CoT):** We introduce a task-specific CoT prompting technique comprising four stages (Figure 2): (1) The model first generates an initial caption based solely on the image. (2) A scene graph is generated (Mitra et al., 2024) to capture object relationships, enabling perceptual reasoning over spatial and relational cues. (3) We then guide the model through a structured sequence of questions grounded in both the scene graph and the image, progressing from objective descriptions to more abstract inferences that help uncover the metaphor's source domain, target domain, and shared attribute. This stage improves the model's conceptual reasoning. (4) Finally, the model refines its

initial caption by integrating insights from the scene graph and metaphor decomposition. This four-stage pipeline is designed to serve as a strong task-specific baseline for visual metaphor understanding. The pipeline is applicable to both zero-shot and multi-shot settings, denoted as **CoT** and **MS+CoT**, respectively. Full prompt details are provided in Appendix I.

Fine-Tuning: We fine-tune two open-source models, LLaVa-1.5-7B (Liu et al., 2023) and Qwen2-VL-7B-Instruct (Wang et al., 2024), on our ImageMet dataset. The fine-tuned variants are denoted as **LLaVa-1.5-7B-IM** and **Qwen2 IM**, respectively. We finetune the models to adhere to the simile format. We freeze the vision encoders and fine-tune the language layers, attention modules and MLP layers. This strategy enables better adherence to structured outputs while enhancing vision-language alignment. Since ImageMet images are crafted to foreground metaphorical elements, fine-tuning helps models focus on relevant visual cues. Training is performed using LoRA (Hu et al., 2022) with a batch size of 4, a learning rate of 2×10^{-4}, and for up to 5 epochs. We use BF16 precision and a single A100 GPU.

4.3 Models

We evaluate both open-source and closed-source models alongside human baselines on our test set. Among open-source models, we consider **LLaVa-1.5-7B** (Liu et al., 2023), **LLaVa-1.6-7B** (Liu et al., 2024), **Phi-3.5-Vision** (Abdin et al., 2024), and Qwen2-VL-7B-Instruct (Wang et al., 2024) (hereby referred to as **Qwen2**). Among closed models, we examine **GPT-4o** (OpenAI, 2023), and Gemini-1.5-Flash (Reid et al., 2024) (hereby referred to as **Gemini**). For the human baseline, we hired a linguistics expert, fluent in English and familiar with linguistic annotation work. They received fair and competitive stipends for their contribution.

5 Results and Analysis

5.1 Automatic Evaluations

Closed Models exhibit Strong Latent Capabilities. GPT-4o and Gemini outperform all open-source baselines. Prompting yields only modest improvements (+4.48 % SBERT-Similarity) for open models, but unlocks substantial gains in closed models (14.64 % on average) – highlighting their latent reasoning abilities and greater adaptability.

Model	SBERT-Similarity
LLaVa	43.72
LLaVa 1.6	46.52
Qwen2	51.57
Qwen2 CoT	52.56
Qwen2 VF	56.20
Qwen2 IM	56.04
Gemini	59.44
GPT4o	62.78
GPT4o MS + CoT	**69.88**

Table 1: **Benchmarking the mean performance** (over three runs) of open and closed SOTA Vision Language models on the ImageMet test set for the task of **Visual Metaphor Captioning (VMC) in Open Format**. Here, IM denotes finetuned on ImageMet, VF denotes finetuned on VFLUTE, MS denotes 4-shot with Explanation, CoT denotes Chain-of-Thought prompting, and the absence of tags denotes zeroshot inferencing. The metric has been scaled to lie in the range 0-100 and the higher the value, the better.

Qwen2 leads among open-source models, outperforming others like LLaVA and Phi (+24.49% SBERT-Similarity), and significantly narrowing the gap to closed-source systems. Finetuning on our synthetic ImageMet dataset further boosts its performance, making Qwen2 IM the strongest open-source baseline. While ImageMet's large-scale synthetic data enables strong generalisation, even to unseen tasks (see Section 5.3), VFLUTE, despite being smaller, delivers comparable performance due to its high-quality human annotations. This highlights the importance of curated data for nuanced tasks like metaphor understanding. Still, the scalability of ImageMet makes it a compelling resource. We also observe that training with simile-format data encourages structured reasoning over metaphor components (see Appendix B, Table 7), further validating the effectiveness of our synthetic data.

Our prompting techniques improve visual metaphor understanding by facilitating a guided approach to analyse the visual metaphor. Few-shot examples grounded in linguistic theory improve format alignment and guide the model's reasoning. Chain-of-Thought prompting consistently enhances performance across models (+1.81 % on average across all models) by guiding step-wise metaphor analysis from perceptual (scene-graph generation and objective questions) to con-

Model	BLEU-4	ROUGE-L	CIDEr	SBERT-Similarity	ACD
LLaVa-1.5-7B	0.88	36.50	3.34	47.81	54.23
LLaVa-1.6-7B	1.00	36.80	3.50	47.70	54.43
Phi-3.5-Vision	1.78	31.42	3.33	48.09	56.54
LLaVa-1.5-7B-IM	5.02	46.57	6.26	56.87	58.63
Qwen2	5.28	43.72	7.67	59.51	58.19
Qwen2 CoT	8.46	46.87	10.48	60.74	60.41
Qwen2 MS	11.18	52.12	10.97	60.14	59.80
Qwen2 MS + CoT	11.41	53.99	13.13	62.18	61.04
Qwen2 VF	11.28	53.50	13.02	61.98	61.10
Qwen2 IM	12.55	55.48	13.76	62.29	61.01
GPT4o	11.57	53.78	12.41	59.49	62.90
GPT4o MS	20.86	61.74	19.52	70.00	64.45
GPT4o MS + CoT	20.95	**61.90**	19.60	70.06	**64.78**
Gemini	9.24	49.24	12.12	62.93	64.39
Gemini MS	**22.05**	61.46	**21.05**	**70.18**	64.56
Human	15.91	51.58	18.43	68.19	63.84

Table 2: **Benchmarking the mean performance** (over three runs) of open and closed SOTA Vision Language models on the ImageMet test set for the task of **Visual Metaphor Captioning (VMC) in Simile format**. Here, IM denotes finetuned on ImageMet, VF denotes finetuned on VFLUTE, MS denotes 4-shot with Explanation, CoT denotes Chain-of-Thought prompting, and the absence of tags denotes zeroshot inferencing. All metrics have been scaled to lie in the range 0-100 and the higher the value, the better.

Model	Primary Accuracy (%)	Secondary Accuracy (%)	Attribute Accuracy (%)
LLaVa	46	51	33
LLaVa 1.6	45	37	39
Qwen2	46	45	45
Qwen2 CoT	44	49	46
Qwen2 VF	46	41	47
Qwen2 IM	51	45	53
GPT4o	67	56	67
GPT4o MS + CoT	75	62	74
Human	**93**	**92**	**90**

Table 3: **Benchmarking the performance** of open and closed SOTA Vision Language models on the ImageMet test set for the task of **Visual Metaphor VQA (VM-VQA)**. Here, IM denotes finetuned on ImageMet, VF denotes finetuned on VFLUTE, CoT denotes Chain-of-Thought prompting, and the absence of tags denotes zeroshot. The accuracies are represented as a percentage out of 100.

ceptual understanding (subjective questions about the metaphorical components).

Humans demonstrate superior metaphor understanding compared to all models. Model performance on generation tasks should be interpreted with caution. In Table 2, models appear to outperform humans across all automatic metrics. However, this is primarily due to the limitations of automatic metrics in evaluating generated outputs. Subtle phrasing variations, along with the inherent subjectivity of human interpretations in VMC, are often penalized by automatic metrics, resulting in lower scores for human responses despite their greater conceptual depth.

While we report gains on multiple automatic metrics, we note that these metrics correlate only moderately with human judgments of semantic consistency (highest correlation: SBERT-Similarity, Pearson r = 0.57). This highlights the interpretive nature of visual metaphor understanding and the limitations of current automatic metrics. Model gains in VMC likely stem from a reliance on surface-level patterns memorized during pretraining, with models often repeating familiar expressions rather than exhibiting genuine abstract reasoning. In contrast, the VM-VQA task exposes persistent gaps in visual metaphor understanding, highlighting the difficulty VLMs face in accurately identifying and reasoning about the metaphor components.

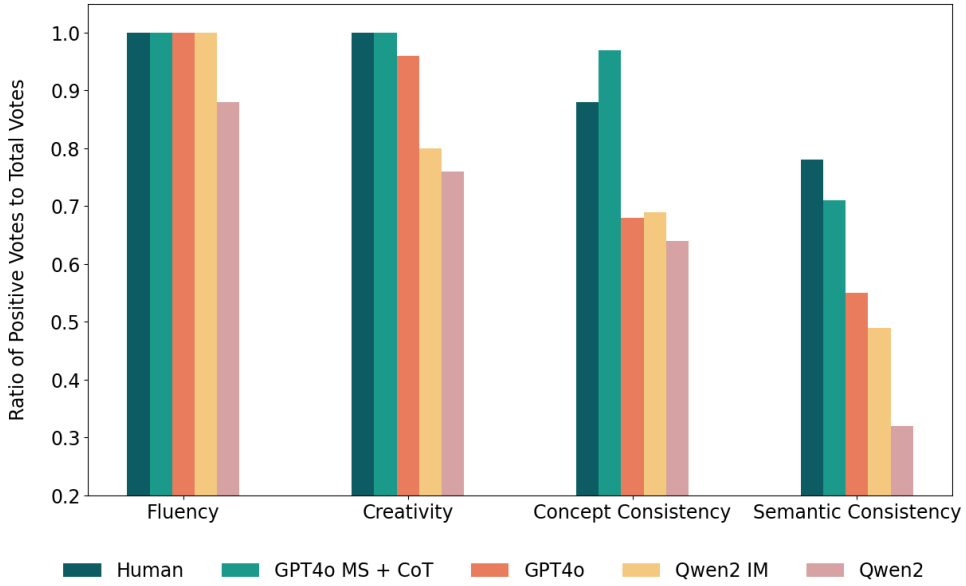

Figure 3: **Results of human evaluation of the captions** generated by 5 models demonstrate that the Human outperforms all models in Semantic Consistency scores. Here, CoT denotes Chain-of-Thought, MS denotes 4-shot with Explanation and only model name denotes zeroshot. For all metrics, the higher the better.

5.2 Human Evaluations

Two Master's students (aged 25–30), fluent in English, served as annotators and were fairly compensated. They evaluated 100 test images and outputs from the models across the following four binary-labeled metrics: (1) **Fluency:** Assesses grammatical correctness, natural flow, and adherence to the task's template. (2) **Creativity:** Evaluates the originality of the metaphor, ensuring dissimilar Primary and Secondary concepts. (3) **Concept Consistency:** Checks accurate identification of the metaphorical concepts without verifying their relationship. (4) **Semantic Consistency:** Ensures captions align with the image content, accurately capturing the intended meaning. The models considered for manual evaluation are (i) GPT4o MS + CoT, (ii) GPT4o, (iii) Qwen2 IM, (iv) Qwen2 and (v) Human.

Human evaluations reveal that finetuning improves Semantic Consistency of the model. Qwen2 IM shows consistent improvement in human evaluations (+20.34 % increase on average across all metrics), confirming that finetuning does improve the visual metaphor understanding of the model. This is also consistently reflected in the VM-VQA task, where the metaphor component understanding of the model significantly improves (+10% increase in Primary accuracy, +17% increase in Attribute Accuracy).

Models struggle with Semantic Consistency, which requires Compositional Understanding and Reasoning. Despite strong performance in visual recognition, models exhibit a persistent gap in semantic consistency compared to humans, a trend mirrored in the VM-VQA task. Manual inspection indicates that errors in semantic consistency often stem from difficulties in aligning metaphorical domains and from a bias toward literal interpretation. These limitations, elaborated in Section 6, reflect a broader difficulty in abstract reasoning and in verbalising visual metaphors, a core component of metaphor understanding (Forceville, 2008). An illustrative output example is provided in Appendix E.

In contrast, human performance on VM-VQA demonstrates a robust grasp of metaphorical structure, with consistently accurate identification of source and target domains in an objective MCQ format. While captioning tasks involve some degree of subjectivity, the structured evaluation reveals the human ability to flexibly adapt to narrative scaffolds and compose creative interpretations. These findings highlight that perception alone does not guarantee correct interpretation, and that genuine metaphor understanding requires nuanced perceptual and conceptual reasoning that current models have yet to fully achieve.

Task		Figurative Image-Text Entailment [I]	Figurative Image-Text Entailment [II]	Metaphorical Captioning	Visual Reasoning VQA
Metric		Accuracy (%)	Accuracy (%)	SBERT-Similarity	Accuracy (%)
Models	Qwen2	46.62	**60.15**	50.78	75.62
	Qwen2 VF	63.15	60.04	**62.25**	82.23
	Qwen2 IM	**66.06**	59.92	57.64	**86.41**

Table 4: **Fine-tuning on ImageMet improves performance across related tasks.** VF denotes fine-tuning on VFLUTE's metaphor and simile subset, and IM denotes fine-tuning on ImageMet. We evaluate on three tasks: Figurative Image-Text Entailment (on two subsets—metaphor/simile [I] and humor/sarcasm/irony [II] from VFLUTE), Metaphorical Captioning (VFLUTE simile subset), and Visual Reasoning VQA (NLVR2). Qwen2 IM performs competitively with Qwen2 VF, highlighting the utility of our synthetic data. All metrics are scaled to 100; higher is better.

5.3 Generalization of Our Approach to Related Tasks

In Table 4, Qwen2 IM yields consistent gains across tasks involving abstract visual understanding and visual reasoning.

We see notable improvements in Figurative Image-Text entailment, Metaphorical Captioning, and Visual Reasoning VQA (Suhr et al., 2019), suggesting a deeper grasp of cross-modal metaphor structure beyond template matching. Importantly, ImageMet finetuning enhances metaphor comprehension without degrading performance on non-metaphorical figurative tasks, indicating that the model may be learning to differentiate metaphor-specific reasoning. On VFLUTE Simile Captioning, Qwen2 IM performs comparably to Qwen2 VF. While VFLUTE contains a limited set of idiomatic similes, which is likely easily memorised during VF finetuning, ImageMet's broader and more diverse data enables stronger generalisation despite less lexical overlap.

6 Qualitative Error Analysis

While models show strong perceptual abilities, their main limitation lies in composing visual observations into coherent metaphorical interpretations. We identify three recurring error patterns:

(a) Domain Swap. Models occasionally reverse the source and target domains of the metaphor. For example, when interpreting an image where an object from domain A is metaphorically mapped onto domain B, the model instead generates a description where B is mapped onto A. This inversion suggests that while the model recognises both domains, it struggles to capture the intended directionality of the mapping, which often requires strong linguistic and cultural knowledge.

(b) Visual Representation Bias. Models often conflate the literal depiction of an object with its metaphorical meaning. For instance, an image containing a syringe may be interpreted as referencing the medical instrument itself rather than the concept of *addiction or medicine*. This reveals a bias toward surface-level visual recognition rather than abstraction.

(c) Implicit Metaphors. When metaphors are only partially realised, such as when one domain is indirectly evoked through subtle visual cues, models tend to fail. They either (i) associate the metaphor with another salient but irrelevant visual concept, or (ii) fall back on stereotypical metaphorical associations learned during pretraining. This highlights the model's difficulty with inference in the absence of explicit visual cues.

Overall, these errors reinforce the observation that visual metaphor understanding requires not only perceptual grounding but also flexible reasoning over abstract conceptual mappings, where current models remain limited.

7 Conclusion

We present **ImageMet**, a benchmark for visual metaphor understanding with 2177 synthetic train/-val and 350 human-annotated test instances. We evaluate 4 open and 2 closed models, and a human baseline on two visual metaphor understanding tasks (VMC and VM-VQA). Finetuning on ImageMet improves performance, narrowing the gap with closed models. Our CoT strategy sets the strongest baseline (GPT4o MS+CoT) overall. Despite gains, models still struggle with metaphor domain mapping, as evidenced in VM-VQA, highlighting a gap in abstract reasoning. ImageMet serves as a valuable tool for studying both the conceptual and perceptual reasoning abilities of VLMs, and it further contributes to the development of models capable of metaphor-aware generation and reasoning.

Limitations

We briefly describe the identified limitations in our work.

- **Shortcomings of automated evaluations:** While we report improvements using automatic metrics such as SBERT-Similarity and a task-specific score, these metrics may not fully capture the nuance, creativity, or interpretive depth required for understanding visual metaphors. Metaphor comprehension often involves subjective and culturally dependent reasoning, which is difficult to evaluate through standard metrics alone. In fact, we find that these metrics correlate only moderately with human judgments (highest correlation: SBERT-Similarity, Pearson r = 0.57, p $\ll$ 0.01), underscoring their limitations. Although we include human evaluations to better assess these aspects, they are limited in scale and inherently subjective, leaving room for future work to develop more robust automated evaluation protocols.

- **Resource constraint:** Due to resource constraints, we limit ourselves to training and testing on relatively smaller models (in the range of 4-11 B parameters). Larger open-source models (over 20B) parameters remain yet to be investigated.

- **ImageMet Scalability vs Diversity:** While the dataset is highly scalable due to automation, at sufficiently large scales, the diversity of generated metaphors may still be constrained by the capability of the generation pipeline.

Ethical Considerations

We release ImageMet under the CC-BY 4.0 license and associated code under the MIT license for research purposes, and we cite all third-party models and datasets as required by their creators. Some images in the ImageMet test set are real-world advertisements and are not publicly released due to copyright restrictions; these images are used internally for research and evaluation purposes only. As these images may reflect societal biases commonly found in advertising media, we emphasise that no personally identifiable information is present in either the images or the captions written by our annotators.

For the training and validation sets, similes are synthetically constructed by associating adjectives with nouns, using large language models (LLMs) such as GPT-4o. Some bias inherent to the LLM may still be reflected in the generated similes. Additionally, the synthetic images used for these pairs are generated using Stable Diffusion 3.5 Large, which may also possess latent visual biases.

We acknowledge these limitations and encourage the research community to use ImageMet and any derived models with appropriate caution, particularly in downstream applications. Responsible and context-aware use is essential to mitigate the propagation of unintended social or representational biases.

References

Marah I Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat S. Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, and 68 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *CoRR*, abs/2404.14219.

Ehsan Aghazadeh, Mohsen Fayyaz, and Yadollah Yaghoobzadeh. 2022. Metaphors in pre-trained language models: Probing and generalization across datasets and languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2037–2050, Dublin, Ireland. Association for Computational Linguistics.

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, Vancouver, BC, Canada, June 17-24, 2023*, pages 23201–23211. IEEE.

Naveen Badathala, Abisek Rajakumar Kalarani, Tejpalsingh Siledar, and Pushpak Bhattacharyya. 2023. A match made in heaven: A multi-task framework for hyperbole and metaphor detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 388–401, Toronto, Canada. Association for Computational Linguistics.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

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, 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, Toronto, Canada. Association for Computational Linguistics.

Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. MelBERT: Metaphor detection via contextualized late interaction using metaphorical identification theories. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1763–1773, Online. Association for Computational Linguistics.

Max Coltheart. 1981. The mrc psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, 33(4):497–505.

Alexander R. Fabbri, Wojciech Kryscinski, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir R. Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Trans. Assoc. Comput. Linguistics*, 9:391–409.

Charles Forceville. 2008. *"Metaphor in Pictures and Multimodal Representations"*.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. *Preprint*, arXiv:2405.01535.

George Lakoff. 1993. The contemporary theory of metaphor. In Andrew Ortony, editor, *Metaphor and Thought*, pages 202–251. Cambridge University Press.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Yuheng Li, Yuanhan Zhang Bo Li, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

David Mick and Edward Mcquarrie. 1999. Visual rhetoric in advertising: Text-interpretive, experimental, and reader-response analyses. *Journal of Consumer Research*, 26:37–54.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. 2024. Compositional chain-of-thought prompting for large multimodal models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 14420–14431. IEEE.

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 23–33, Berlin, Germany. Association for Computational Linguistics.

Michael Mohler, Mary Brunson, Bryan Rink, and Marc Tomlinson. 2016. Introducing the LCC metaphor datasets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4221–4227, Portorož, Slovenia. European Language Resources Association (ELRA).

Amitash Ojha. 2013. An experimental study on visual metaphor.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

ALLAN PAIVIO, JOHN YUILLE, and STEPHEN MADIGAN. 1968. Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of experimental psychology*, 76:Suppl:1–25.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Abisek Rajakumar Kalarani, Pushpak Bhattacharyya, and Sumit Shekhar. 2024. Unveiling the invisible: Captioning videos with metaphors. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6306–6320, Miami, Florida, USA. Association for Computational Linguistics.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, and 34 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR*, abs/2403.05530.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

I.A. Richards. 1936. *The Philosophy of Rhetoric*. Bryn Mawr College. Mary Flexner lectures. Oxford University Press.

Arkadiy Saakyan, Shreyas Kulkarni, Tuhin Chakrabarty, and Smaranda Muresan. 2024. V-FLUTE: visual figurative language understanding with textual explanations. *CoRR*, abs/2405.01474.

Kevin Stowe, Tuhin Chakrabarty, Nanyun Peng, Smaranda Muresan, and Iryna Gurevych. 2021. Metaphor generation with conceptual mappings. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6724–6736, Online. Association for Computational Linguistics.

Chuandong Su, Fumiyo Fukumoto, Xiaoxi Huang, Jiyi Li, Rongbo Wang, and Zhiqun Chen. 2020. DeepMet: A reading comprehension paradigm for token-level metaphor detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 30–39, Online. Association for Computational Linguistics.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy. Association for Computational Linguistics.

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 (Volume 1: Long Papers)*, pages 248–258, Baltimore, Maryland. Association for Computational Linguistics.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575. IEEE Computer Society.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Michael Wilson and Informatics Division. 1997. Mrc psycholinguistic database: Machine usable dictionary, version 2.00. *Behav Res Methods*, 20.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. 2020. Towards fairer datasets: filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 547–558, New York, NY, USA. Association for Computing Machinery.

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, Singapore. Association for Computational Linguistics.

Dongyu Zhang, Minghao Zhang, Heting Zhang, Liang Yang, and Hongfei Lin. 2021. MultiMET: A multimodal dataset for metaphor understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3214–3225, Online. Association for Computational Linguistics.

Linhao Zhang, Jintao Liu, Li Jin, Hao Wang, Kaiwen Wei, and Guangluan Xu. 2024. GOME: Grounding-based metaphor binding with conceptual elaboration for figurative language illustration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18500–18510, Miami, Florida, USA. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

ImageMet Fraction	BLEU-4	ROUGE-L	CIDEr	SBERT-Similarity	ACD
33%	11.4	51.65	11.56	61.27	59.14
66%	11.96	51.72	11.91	62.10	59.89
100%	**12.55**	**55.48**	**13.76**	**62.29**	**61.01**

Table 5: **Variation in performance of Qwen2 with variation in finetuning dataset size.** We observe a clear increase in performance across all metrics with increase in dataset size, with a very strong Pearson correlation (0.89 across all metrics)

Dataset	Number of Instances	Text Source	Image Source
ImageMet	2177	Synthetically generated	Synthetically Generated
VFLUTE	1024	Distilled from IRFL and HAIVMet	Distilled from IRFL and HAIVMet
HAIVMet	958	Manually Curated	Synthetically Generated
IRFL	1440	Manually Curated	Manually Curated

Table 6: **Comparison of ImageMet with other openly available related datasets** based on size and the source of text and images. ImageMet is the largest dataset of its kind and is fully synthetic, generated through an automated pipeline that enables scalability.

A Background on Metaphors and Similes

As described in Section 1, we provide more linguistic background on metaphors in this section. A metaphor is a map between a source and a target domain through their shared properties. It is a mapping of knowledge about one concept (the 'source concept') to another (the 'target concept') (Lakoff, 1993). For example:

- **Simple metaphor**: In '*Life is a rollercoaster*', the target domain is '*life*' and the source domain is a '*rollercoaster*' and the unpredictability of a rollercoaster is used to highlight the unpredictable nature of life.
- **Implied metaphor**: In '*He is drowning in paperwork*', the target domain is '*having too much paperwork*' and the source domain is '*drowning*' and the feeling of despair associated with drowning is used to highlight the overwhelming nature of the work.

Metaphors are implicit comparisons and their subtlety makes them difficult to extract information from. Simple metaphors usually follow an 'A is B' format, where A is called the *tenor* or the *primary concept* and B is referred to as the *vehicle* or the *secondary concept* (Richards, 1936). In simple metaphors, the primary and the secondary concepts are explicitly stated in the text, whereas in implied metaphors, the comparison is subtle and is *implied*, as the name suggests. This makes simple metaphors more desirable for computationally extracting information. However, consider the example: "*Life is a box of chocolates*". Although the primary and the secondary concepts are explicitly stated, the attribute or common property being shared between them is up to interpretation. One possible attribute is 'sweetness', but another possible attribute is 'unpredictability' (as used in the movie 'Forrest Gump'), which completely changes the meaning. Such ambiguity and subjectivity in a metaphor is part of its beauty, however for our purposes, it poses a problem. Similes are explicit comparisons and often follow one of two formats — 'A is like B' or 'A is as C as B'. Clearly, the latter is the most desirable format for our purpose as it states not only the primary concept A, and the secondary concept B but also the '*Attribute*' or 'Quality' C that is being shared between them. Therefore, we stick to the following template to describe the metaphors present in our dataset:

Primary is as *Attribute* as *Secondary*.

For example, we write 'Life is as unpredictable as a rollercoaster'. This syntax imposes rigid constraints on the structure of the sentence, which might sometimes diminish the aesthetic value of a metaphor but offers a convenient setting for easier information extraction and better evaluation of our models. Prior works (Rajakumar Kalarani et al., 2024; Akula et al., 2023) in visual metaphors have also focused on such a format.

Task		VMC	VM-VQA		
Metric		SBERT-Similarity	Primary Accuracy (%)	Secondary Accuracy (%)	Attribute Accuracy (%)
Models	Qwen2 IM-M	**61.45**	40	43	39
	Qwen2 IM	61.27	**46**	**45**	**45**

Table 7: **Finetuning on simile captions demonstrates better results compared to open-form metaphorical sentences.** This demonstrates the effectiveness of the ImageMet simile format for the Train set. IM denotes finetuned on ImageMet and IM-M denotes finetuned on open-form metaphorical sentences paraphrased from ImageMet.

Model	BLEU-4	ROUGE-L	CIDEr	SBERT-Similarity	ACD	Prometheus-Eval
Qwen2	5.28	43.72	7.67	59.51	58.19	62.23
Qwen2 IM	12.55	55.48	13.76	62.29	61.01	63.16
GPT4o MS + CoT	20.95	61.90	19.60	70.06	64.78	75.87
Human	15.91	51.58	18.43	68.19	63.84	65.15

Table 8: **Performance of 4 models on the VMC task across multiple metrics.** As Prometheus offers no substantial additional insight over existing metrics, we omit it from further analysis.

B Ablation Studies

As described in Section 3 and Section 5, we demonstrate the effect of scaling the ImageMet data and the effectiveness of the simile caption form in this section.

B.1 Finetuning Dataset Size

Table 5 presents the model's performance relative to the size of the fine-tuning dataset. We evaluate three subsets of ImageMet — 33%, 66% and 100% — to assess the impact of data quantity and highlight the effectiveness of our dataset in improving performance on this task. We observe a consistent performance improvement with increasing dataset size, underscoring both the need for more data and the effectiveness of our dataset. This highlights the value of our automated pipeline in efficiently generating large-scale data for this task.

B.2 Finetuning on Simile captions vs Open-form captions

To evaluate the impact of structured simile-based training, we fine-tune Qwen2 on a subset (33%) of the ImageMet train set, where simile captions are paraphrased into open-form metaphorical sentences using GPT-4o (denoted as Qwen IM-M). While Qwen IM-M maintains SBERT similarity, it underperforms on VM-VQA compared to Qwen IM, indicating that the simile structure more effectively supports metaphor component understanding during fine-tuning (see Table 7).

C Comparison with existing Datasets

We compare ImageMet with existing datasets in Table 6, considering only its synthetically generated portion, which comprises 2177 instances. Including the manually annotated test set, the total dataset size is 2527. As shown in the table, ImageMet is the largest openly available dataset and the only one fully generated through a synthetic pipeline. Its automated nature enables scalability, making it a valuable resource for advancing research in visual metaphor understanding.

In Table 1, 2, and 3, we analyze the impact of fine-tuning Qwen2 on ImageMet versus VFLUTE. We evaluate the fine-tuned models on both the ImageMet and VFLUTE test sets. Qwen2 IM outperforms Qwen2 VF on the ImageMet Test Set but falls short on the VFLUTE Test Set (Table 4). Manual inspection reveals that VFLUTE simile test set size is small (26 instances) and less diverse (only 14 unique secondary concepts forming similar similes). Since the VFLUTE training set contains similar similes, Qwen2 VF performs better. Despite this, Qwen2 IM performs comparably to Qwen2 VF. Additionally, images in the VFLUTE Simile Set rarely contain intended visual metaphors, with the figurative language being confined to the text.

This underscores that the quality of ImageMet is at least on par with existing datasets like VFLUTE. ImageMet stands out as the largest openly available dataset for visual metaphors and offers scalability through its automated pipeline.

D Choice of metrics

In Section 4, we provide the metrics used for the VMC task. In this section, we justify our choice of metrics. We evaluate a range of metrics to de-

termine the most suitable for the VMC task, experimenting on a subset of models and their outputs (Table 8). These include BLEU, ROUGE-L, CIDEr, SBERT-Similarity, and ACD, as well as Prometheus Eval—an LLM-based evaluation metric. BLEU, ROUGE-L and CIDEr are n-gram based metrics, SBERT-similarity is to capture semantic similarity and ACD is a task-specific metric that attempts to capture semantic similarity as well as primary and secondary concept dissimilarity. However, we observe only moderate correlation between these automatic metrics and human judgments (based on semantic consistency labels), with SBERT-Similarity showing the highest Pearson correlation ($r = 0.57$, $p \ll 0.01$). This is expected given the complex and interpretive nature of visual metaphor understanding. Since Prometheus scores do not offer substantially different insights and are compute-intensive, we opt not to report them across the full test set.

E Output Example

In Figure 10, we compare model outputs against the ground truth: "Caffeine is as addictive as a drug."

Qwen2 overly focuses on literal objects ("syringe," "coffee beans"). Qwen2-IM abstracts better, replacing "syringe" with "drug", and aligns more closely with the metaphor. GPT-4o and GPT-4o MS+CoT both provide creative yet valid interpretations, shifting the attribute from "addictive" to "energizing." The CoT-enhanced output improves structure and aligns better with metaphor format. The human-written output also emphasizes caffeine's energizing effect. While these diverge from the ground truth, they reflect legitimate metaphorical readings, showing the value of interpretive flexibility. Fine-tuning and prompting strategies notably improve metaphor grounding and coherence.

F Models and Module Details

The following are the checkpoint details for all models used:

- GPT4o: gpt-4o-2024-08-06
- Gemini: gemini-1.5-flash
- LLaVa: llava-hf/llava-1.5-7b-hf
- LLaVa 1.6: llava-hf/llava-v1.6-mistral-7b-hf
- Phi: microsoft/Phi-3.5-vision-instruct
- Qwen2: Qwen/Qwen2-VL-7B-Instruct
- Prometheus: prometheus-eval/prometheus-7b-v2.0 (Kim et al., 2024)
- SBERT: sentence-transformers/all-mpnet-base-v2
- Stable Diffusion 3.5 Large: stabilityai/stable-diffusion-3.5-large

We use:

- WordNet 3.0 (Miller, 1994)
- NLTK 3.9.1 (Bird and Loper, 2004)
- SpaCy 3.7.2 (Honnibal and Montani, 2017)
- Huggingface Transformers 4.46.0 (Wolf et al., 2019)

G WordNet Structure and Adjectives

We utilise the WordNet structure to extract frequently associated adjectives corresponding to each Primary noun (refer Section 3), wherever possible. For each noun, we explore the 'hyponyms' and 'examples' attributes to extract possible adjectives.

Hyponyms: X is a hyponym of Y if X is a (kind of) Y. For example, crow is a hyponym of bird. For words with abstract meanings, hyponyms sometimes contain a possible adjective for the word. For example, for the word 'tongue' meaning 'a manner of speaking', 'sharp tongue' is one of the hyponyms.

Examples: WordNet also contains example sentences for some words. Examples sometimes contain adjectives associated with the nouns. For example, one of the examples under the same synset of 'tongue' is 'She has a glib tongue' which gives us the adjective 'glib'.

We exploit these attributes to gather adjectives. This method allows us to obtain adjectives pertaining to different meanings of the same word. For example, 'tongue' meaning 'a mobile mass of muscular tissue' has adjectives such as red, thin etc. Iterating through different synsets of the same word allows us to collect adjectives corresponding to different meanings of the word, thus ensuring a diverse collection of adjectives.

H Imageability and Concreteness of Words

We plot the imageability and concreteness distributions for all words present in the MRC Database. Based on Figure 4 and Figure 5, we choose our cut-offs for both imageable and concrete concepts to be at 500. All concepts having a score over 500 for both imageability as well concreteness are chosen as the Primary nouns (refer Section 3).

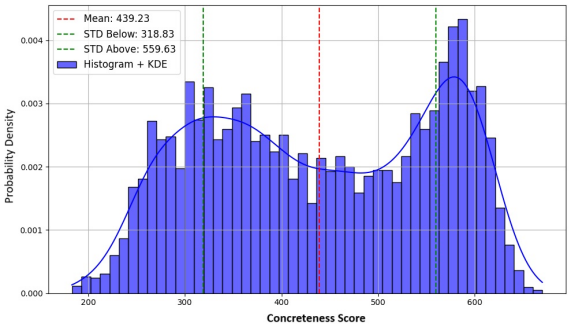

Figure 4: Distribution over Concreteness Scores. We observe a natural separation between concrete and non-concrete concepts around the 500 mark.

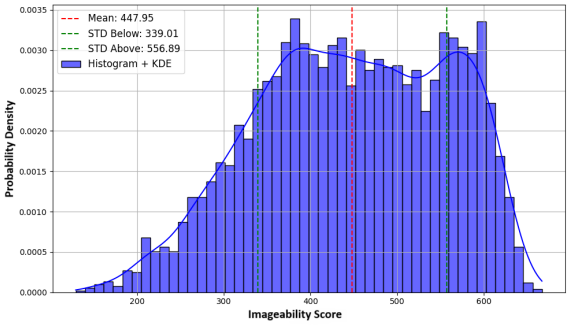

Figure 5: Distribution over Imageability Scores. We see a dip around the 500 mark suggesting a natural separation between highly imageable and non-imageable concepts.

I Prompt Details

Sections 3 and 4 outline the methods employed for data generation and inference. Here, we present a detailed elaboration of each prompt.

I.1 Inference

For zeroshot inferencing on the test set, we use the following prompt across all models:

Caption the image in one sentence using a simile. The image contains a metaphorical comparison between two concepts. You must write the caption in the form 'A is as B as C' where A is being compared to C on the basis of a shared property B.

Figure 6 shows the 4 examples we use for our multishot prompting technique along with the accompanying explanations. We pass the four examples back to back, followed by the above zeroshot inferencing prompt.

I.2 Instruction Tuning

We use the following instruction prompt for instruction-finetuning the models:

Caption the visual metaphor in the image in a single sentence. The caption should be a simile of the form: A is as B as C where A is being compared to C on the basis of shared property B.

I.3 Visual Elaboration

For the visual elaboration, we use Chain-of-Thought prompting to identify the primary concept, the secondary concept and the attribute, and then prompt the model to follow specific rules based on whether the tenor and vehicle are abstract or concrete. The detailed visual prompt is provided in Figure 7 and Figure 8.

I.4 CoT

We show the Stage 2 and Stage 3 prompts for our CoT prompting technique. In stage 2, we generate the scene graph for the image. In stage 3, we prompt the model to analyse the image using a set of questions. The detailed prompt is given in Figure 9.

J Annotation Details

In Sections 3 and 5, we briefly referenced the human annotation processes involved in our study. This section provides a detailed account of the guidelines followed for each annotation task.

J.1 ImageMet Test Set

For the manual annotation of the ImageMet test set, we provided the following guidelines to our annotators in Round 1:

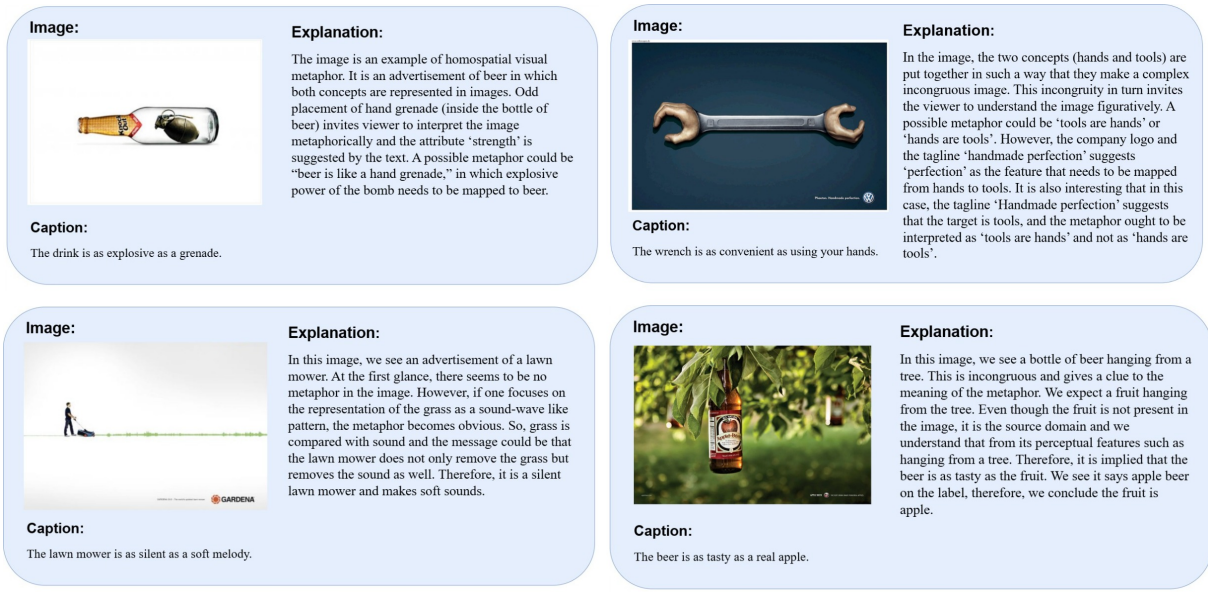

Figure 6: **Multi-shot examples presented to the models**, grounded in linguistic theory (Ojha, 2013), were selected to highlight the diverse ways in which source and target domains can manifest in visual form.

Manual Annotation Guidelines (Round 1)

Task Instruction:
"You have to caption 350 images. Each image contains a metaphorical comparison. Your task is to understand the intended meaning of the visual metaphor present in the image. Then, you must write it in a single sentence while capturing the essence of the comparison."

Task restrictions:
You must stick to the following format: *A* is as *B* as *C*. Some of the instances might be difficult to fit in the format, however you must caption it as closely as possible while sticking to the format.

For round 2, we give them the following instruction:

Manual Annotation Guidelines (Round 2)

Task Instruction:
"If the Primary, Attribute, and Secondary concepts are all similar for both (considering synonyms), choose whichever you unanimously decide to be more natural and suitable.
If you differ on even one of the three elements, you must discuss and reach a consensus among yourselves as to which caption suits the image better.

If you're unable to reach a consensus even after discussing, discard the instance."

J.2 Synthetic Data validation

The following are the guidelines used to validate the quality of the synthetic data.

1. Synthetic Simile Validation:

Synthetic Simile Validation Guidelines

Task Instruction:
Rate the following similes on a scale of 1–5 based on their coherence. Follow the following label descriptions as a guidance:

1 = No Coherence; The sentence

makes no sense.
2 = Poor Coherence; The simile makes sense but is confusing or irrelevant.
3 = Moderate Coherence; The simile is understandable but loosely connected.
4 = Good Coherence; The simile is relevant but lacks creativity.
5 = Excellent Coherence; The simile is clear, vivid, and creative.

2. Synthetic Image-Simile Pair Validation:

Synthetic Image-Simile Pair Validation Guidelines

Task Instruction:
Rate the following image-text pairs on a scale of 1–5 based on their metaphorical alignment. Each text is a simile, and each image is intended to be a meaningful visual metaphor representing the comparison expressed in the simile. Follow the label descriptions below as guidance:

1 = No Alignment; The image does not relate to the simile's comparison in any meaningful way.
2 = Weak Alignment; The image loosely connects to the simile but fails to visually convey the metaphor.
3 = Moderate Alignment; The image somewhat reflects the metaphor in the simile but lacks clarity, creativity, or precision.
4 = Strong Alignment; The image clearly represents the metaphorical comparison made in the simile with relevant visual cues.
5 = Excellent Alignment; The image vividly and creatively embodies the metaphor in the simile, enhancing its meaning through visual storytelling.

J.3 Human Evaluation

For the human baselines, we provided the same guidelines as Round 1 of the manual annotation phase.

For the manual evaluation process, each annotator was given 100 instances, where each instance included an image and 5 model-generated captions. The following guidelines were given to the annotators:

Human Evaluation Guidelines

Task Instruction:
"Label each caption with 4 binary labels(1 or 0) for the 4 metrics defined below."

Metric Definitions:
Fluency: Is the caption grammatically correct? Does the caption stick to the simile syntax *A* is as *B* as *C*? If yes to both, mark 1. Else 0.

Creativity: Are 'A' and 'C' similar? For example, apple and cherry can be considered similar since they are both fruits, but apple and ruby are dissimilar. If they are similar, mark 0. Else 1.

Concept Consistency: Do the Primary, Attrbiute and Secondary concepts appear in the caption? If all present, mark 1. If all absent, mark 0. If some present, the decision is up to you. Mark suitably.

Semantic Consistency: Does the caption capture the essence of the visual metaphor? If yes, mark 1. Else 0.

During all the annotation processes, the authors made sure to be available to the annotators for any doubts they had regarding the task. However, they did not explicitly look at any of the annotated captions to avoid bias. They only clarified the task instructions, definitions and related doubts.

CoT Prompt Guide for Visual Elaboration from Similes

Your task is to generate a visual elaboration given a simile in 'A is as B as C' format. The key idea in this elaboration is to create a visual scene using C and then swap C with A since A is being compared to C.

CoT Guide - Think step-by-step:

1. The tenor (A) is the object that is being compared in the simile. What is the tenor?
2. The vehicle (C) is the object that the tenor is being compared to. What is the vehicle?
3. The attribute (B) is a property of the vehicle being used to highlight the same property (B) in the tenor. What is the attribute?
4. Consider this definition of abstract and concrete. Abstract: Concepts that do not have a definite physical form that can be represented visually Concrete: Concepts that have a definite physical form that can be represented visually
5. Is the tenor abstract or concrete?
6. Is the vehicle abstract or concrete?

Based on the tenor and vehicle type, determine which of the following cases applies, and follow the corresponding rules. Only one case applies per sentence.

CASE 1: If the tenor is concrete and the vehicle is concrete: You must adhere to the following rules:
Rule-1: Create a visual elaboration with the vehicle as the focus of the scene.
Rule-2: Replace the vehicle with the tenor in the visual elaboration. You should swap exactly the vehicle with the tenor without changing any other component of the visual elaboration. Do not mention the vehicle in this rewritten elaboration.
Rule-3: Write the same visual elaboration concisely within 40 words.

CASE 2: If the tenor is concrete and the vehicle is abstract: You must adhere to the following rules:
Rule-1: Think of a way to visually represent the vehicle using a concrete concept.
Rule-2: Create a visual elaboration with this concrete concept as the focus of the scene.
Rule-3: Replace the concrete concept with the tenor in the visual elaboration. Also include the concrete concept in the visual elaboration. Create a juxtaposition or fusion of the tenor and the concrete concept.
Rule-4: Write the same visual elaboration concisely within 40 words.

CASE 3: If the tenor is abstract and the vehicle is concrete: You must adhere to the following rules: Rule-1: Create a visual elaboration with the vehicle as the focus of the scene.
Rule-2: Think of a way to visually represent the tenor using a concrete concept, a scene or some objects.
Rule-3: Introduce the tenor's visual representation in the visual elaboration and make it the focus of the scene. Create a juxtaposition or fusion of the vehicle and the tenor's visual representation.
Rule-4: Write the same visual elaboration concisely within 40 words.

CASE 4: If the tenor is abstract and the vehicle is abstract: You must adhere to the following rules:
Rule-1: Think of a way to visually represent the vehicle using a concrete concept, a scene or some objects.
Rule-2: Think of a way to visually represent the tenor using a concrete concept, a scene or some objects.
Rule-3: Create a visual elaboration, making a meaningful juxtaposition or fusion of the tenor and the vehicle's visual representations.
Rule-4: Write the same visual elaboration concisely within 40 words.Figure 9: Systemic Knowledge used to prompt GPT-4o for Visual Elaboration Generation from Simile samples.

Figure 7: Systemic CoT Prompt Guide used for Visual Elaboration Generation from Similes.

Few Shot Examples for Visual Elaboration from Similes

Simile: The lake is as reflective as a mirror.

Output:
1. The tenor (A) is "The lake."
2. The vehicle (C) is "a mirror."
3. The attribute (B) is "reflectiveness."
4. The tenor "The lake" is concrete.
5. The vehicle "a mirror" is concrete.

Case 1 applies here.
Initial Visual Elaboration: A large, perfectly smooth mirror sits nestled among trees and under the sky, reflecting the blue expanse and fluffy clouds above with pristine clarity.
Final Visual Elaboration: [Case 1] A large, perfectly smooth lake sits nestled among trees and under the sky, reflecting the blue expanse and fluffy clouds above with pristine clarity.

Simile: The sword is as sharp as justice.

Output:
1. The tenor (A) is "The sword."
2. The vehicle (C) is "justice."
3. The attribute (B) is "sharpness."
4. The tenor "The sword" is concrete.
5. The vehicle "justice" is abstract.

Case 2 applies here.
Initial Visual Elaboration: A set of scales, perfectly balanced, stands in the centre of a grand hall, its golden surface catching the light.
Final Visual Elaboration: [Case 2] A gleaming sword stands upright in the centre of a grand hall, blending into the scales of justice as part of its hilt.

Simile: Time is as valuable as money.

Output:
1. The tenor (A) is "time."
2. The vehicle (C) is "money."
3. The attribute (B) is "value."
4. The tenor "time" is abstract.
5. The vehicle "money" is concrete.

Case 3 applies here.
Initial Visual Elaboration: A pile of gold coins, gleaming under soft light, symbolizing wealth.
Final Visual Elaboration: [Case 3] An hourglass with shimmering golden sand flows like coins, blending time's passage with money's value, as grains slip elegantly from the top to the bottom, embodying their shared significance.

Simile: Hope is as fragile as a dream.
Output:
1. The tenor (A) is "hope."
2. The vehicle (C) is "a dream."
3. The attribute (B) is "fragility."
4. The tenor "hope" is abstract.
5. The vehicle "a dream" is abstract.

Case 4 applies here.
Initial Visual Elaboration: A dream bubble showing a fantastic landscape.
Final Visual Elaboration: [Case 4] A flickering candle illuminates a bubble-filled floating landscape, symbolizing hope's delicate nature, interwoven with ethereal dreams.

Figure 8: Few-shot examples provided to GPT-4o for Visual Elaboration Generation from Simile samples.

Scene Graph Generation (Stage 2) and Metaphor Analysis (Stage 3) Prompts for our Novel CoT Approach

Stage 2: Scene Graph Generation Prompt

For the provided image, generate a scene graph in JSON format that includes the following:

Objects that are prominently visible in the image.
Object attributes that are relevant to its description.
Object relationships between the visible objects.
Object textures for each visible object.

Scene Graph:

Stage 3: Visual Metaphor Analysis Prompt using CoT step-by-step thinking

Chain-of-Thought Guide - Think step-by-step:

Step 1: Image and Task Context
You are given an image that may contain a visual metaphor. You are also given a scene graph of that image. Your task is to analyze the image in a structured manner to progressively uncover its metaphorical meaning.

Step 2: Objective Understanding (Basic Image Comprehension)
First, extract factual information about the image.
a. Identify all objects present in the image. List them clearly.
b. Describe the attributes (e.g., color, size, texture) of the objects.
c. Explain the spatial arrangement of the objects.
d. Identify any interactions between the objects. If applicable, describe them.
e. Determine if any humans or animals are present. If yes, describe their actions.
f. Extract any visible text from the image and transcribe it.

Step 3: Visual Metaphor Building Block Understanding (Contextual Interpretation)
A visual metaphor contains a source domain and a target domain. The target domain is usually present in the picture.
The source domain may or may not be visually present. But it will always be visually represented.
The source domain lends one of its qualities to the target domain.
The source domain is often something general that people are familiar with and is used to highlight the quality of the target domain which is more specialised.
If the source domain is visually present, it may be juxtaposed along with the target. It may also be fused with the target.
If the source domain is visually not present but represented through its perceptual features, it may be through colour, texture, shape etc. that is commonly and generally associated to the source.
Based on this theoretical understanding of visual metaphors, answer the following questions:

a. What could be the source? Just write the source. Do not write what it represents.
b. What could be the target? Just write the target. Do not write what it represents.
c. What is the quality, attribute or property that the source is lending to the target?

Step 4: Final Output Format
Your final response should be structured as follows:
Objective Description: (List answers to questions 2a - 2f)
Visual Metaphor Building Block Analysis: (List answers to questions 3a - 3c)

Figure 9: Prompt for our Task-Specific CoT Prompting Technique

Example

Qwen2: *Coffee beans are as potent as a syringe in delivering a powerful caffeine boost.*

Qwen2 IM: *The coffee is as potent as a drug.*

GPT4o: *The coffee injection is as energizing as a jolt of electricity.*

GPT4o MS + CoT: *Coffee is as energizing as a medicine injection.*

Human: *Coffee is as energizing as medicine.*

Ground Truth: *Caffeine is as addictive as drugs.*

Figure 10: **Illustrative example of outputs** from Qwen2, Qwen2 IM, GPT-4o, GPT-4o MS+CoT, and a human baseline on a visual metaphor input. The ground-truth caption is provided for reference, highlighting differences in how models align the source and target domains of the metaphor.