CMIG: Conceptual Metaphor Theory-Inspired Framework for Metaphorical Image Generation

Anonymous ACL submission

Abstract

Metaphorical text encodes cross-domain meaning beyond its literal surface, posing a challenge for text-to-image models to produce semantically faithful visual metaphors. We propose CMIG, a structured prompting framework inspired by Conceptual Metaphor Theory (CMT), which decomposes metaphors into source–target mappings and selects visual realization strategies in a reproducible reasoning workflow. Experiments on DALL·E 3, Imagen 2, and FLUX-1 show that CMIG consistently improves semantic alignment and human-rated metaphor quality over prior prompting baselines. We additionally release a 3,500-instance visual metaphor benchmark to support unified evaluation.

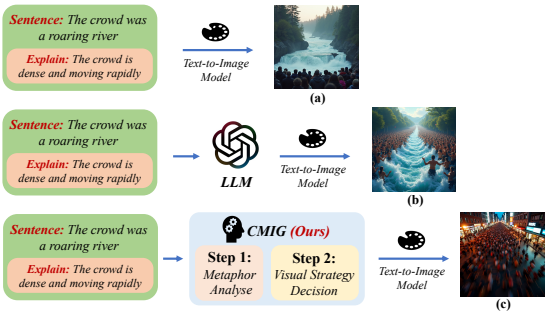

Figure 1: Qualitative comparison on the metaphor "The crowd was a roaring river" (intended meaning: a dense, fast-moving crowd). (**a**) Direct prompting with the metaphor tends to produce a literal river-related scene. (**b**) Chain-of-thought prompting partially improves interpretation but remains semantically inconsistent. (**c**) CMIG generates a prompt that better preserves the figurative meaning while maintaining visual coherence.

1 Introduction

Metaphor is pervasive in natural language and is often used to convey meanings that go beyond literal descriptions. Generating images from metaphorical text is therefore a challenging form of text-to-image generation: a system must infer the intended figurative meaning and translate it into visual content, rather than rendering the surface semantics. In practice, current text-to-image models tend to produce literal scenes (Rombach et al., 2022; Betker et al., 2023), which are visually plausible but semantically inconsistent with the metaphor's implication (e.g., depicting "a crowd by a river" instead of expressing a dense, fast-moving crowd suggested by "The crowd was a roaring river").

This problem matters for applications such as creative design and education (Phillips and McQuarrie, 2004; Forceville, 2002; Scott, 1994; Forceville and Urios-Aparisi, 2009), and it also offers a diagnostic setting for evaluating whether generative models capture meaning beyond literal text. Prior work reports that even strong systems such as Imagen (Saharia et al., 2022), Stable Diffusion (Rombach et al., 2022), and FLUX.1 (Labs, 2024) often drift toward literal interpretations or produce visually misaligned outputs on metaphorical inputs (Akula et al., 2023; Yosef et al., 2023). Figure 1 shows a representative example.

A natural starting point for modeling metaphors is Conceptual Metaphor Theory (CMT) (Lakoff and Johnson, 2008), which describes metaphor understanding as selectively projecting attributes from a *source* domain onto a *target* domain. However, directly turning CMT into a computational procedure is difficult: cross-domain correspondences are rarely explicit, attribute salience depends on context, and multiple visual realizations can be valid. We take an alternative approach: instead of learning CMT end-to-end, we use it to structure a controllable reasoning workflow for large language models (LLMs). Concretely, we **operationalize CMT via prompt engineering** by decomposing metaphor understanding into explicit steps for (i) identifying source/target domains, (ii) enumerating candidate projectable attributes, and (iii) selecting visually expressible attributes while avoiding literal distractors.

Existing prompting-based approaches, including chain-of-thought prompting and semantic expansion (Su et al., 2024; Shahmohammadi et al., 2023), remain limited in three ways. First, many methods rely on heuristic prompting without an explicit semantic structure (Su et al., 2024). Second, some pipelines require manual filtering or closed-source components, which reduces reproducibility and scalability (Chakrabarty et al., 2023). Third, generalization to complex metaphors is often weak (Su et al., 2024; Chakrabarty et al., 2023; Shahmohammadi et al., 2023). As a result, even when the backbone generator is strong, the produced images frequently emphasize literal entities from the metaphor text rather than the intended implication.

We propose **CMIG (Conceptual Metaphor Theory-Inspired Metaphorical Image Generation)**, a **structured prompting framework** for metaphorical image generation. CMIG has two modules: *metaphor parsing*, which identifies domains and derives projectable attributes, and *visual strategy selection*, which chooses a realization strategy using a *Dependency Test* and an *Interference Test*. The resulting prompt integrates figurative semantics with explicit visual planning, enabling text-to-image models to better preserve metaphor meaning while maintaining visual coherence (Figure 1c).

Our contributions are:

1. We introduce **CMIG**, a CMT-inspired structured prompting framework for metaphorical image generation that improves semantic alignment and visual coherence over prior methods.

2. We present a structured design for metaphor parsing and visual planning that reduces reliance on manual curation and brittle heuristics.

3. We release a dataset of 3,500 visual metaphor examples spanning 400 metaphors with multiple positive and negative samples, providing a reproducible benchmark for future research.

2 Related Work

2.1 Metaphorical Image Generation

Recent diffusion models have substantially advanced text-to-image generation, outperforming traditional approaches such as variational autoencoders (VAEs) (Razavi et al., 2019) and generative adversarial networks (GANs) (Bao et al., 2017). Models like DALL·E 2 (Ramesh et al., 2022), Imagen (Saharia et al., 2022), Stable Diffusion (Rombach et al., 2022), and FLUX.1 (Labs, 2024) can produce high-fidelity and diverse images. However, empirical studies (Yosef et al., 2023; Akula et al., 2023) reveal that these models still struggle to interpret metaphorical language, often failing to capture cross-domain semantic relations. To enhance metaphorical image generation, recent research leverages prompt engineering and semantic reasoning. ViPE (Shahmohammadi et al., 2023) transforms arbitrary text into visual descriptions; SPy (Chakrabarty et al., 2023) integrates InstructGPT-3 (Ouyang et al., 2022) and chain-of-thought prompting (Wei et al., 2022) to generate visual interpretations of metaphoric language; and Su et al. (Su et al., 2024) propose a framework that abstracts source-target semantics and employs CLIP (Radford et al., 2021) embeddings to generate metaphor-enhanced prompts. With the integration of large language models (LLMs), text-to-image systems are increasingly coupling linguistic reasoning with visual generation. DALL·E 3 (OpenAI, 2023), integrated into ChatGPT (Achiam et al., 2023), shows stronger semantic consistency in interpreting metaphorical inputs than earlier models. Nonetheless, existing systems largely depend on visual similarity rather than explicit semantic mapping between source and target domains. As noted by Yosef et al. (2023); Akula et al. (2023), even state-of-the-art diffusion models remain constrained without theory-driven mechanisms for modeling the cross-domain cognitive mappings underlying metaphor.

2.2 Multimodal Metaphor Understanding

Metaphor, as a complex cognitive–linguistic phenomenon, has long been a central topic in NLP and machine learning. With advances in deep learning, significant progress has been made in metaphor understanding (Ge et al., 2022), detection (Lin et al., 2021; Su et al., 2021), and generation (Stowe et al., 2021). However, multimodal metaphor understanding remains challenging due to the need for cross-modal semantic alignment and abstract concept modeling. Visual metaphors require models not only to interpret linguistic semantics but also to recognize atypical objects, abstract imagery, and artistic composition, demanding deep multimodal fusion. Conventional vision–language retrieval and

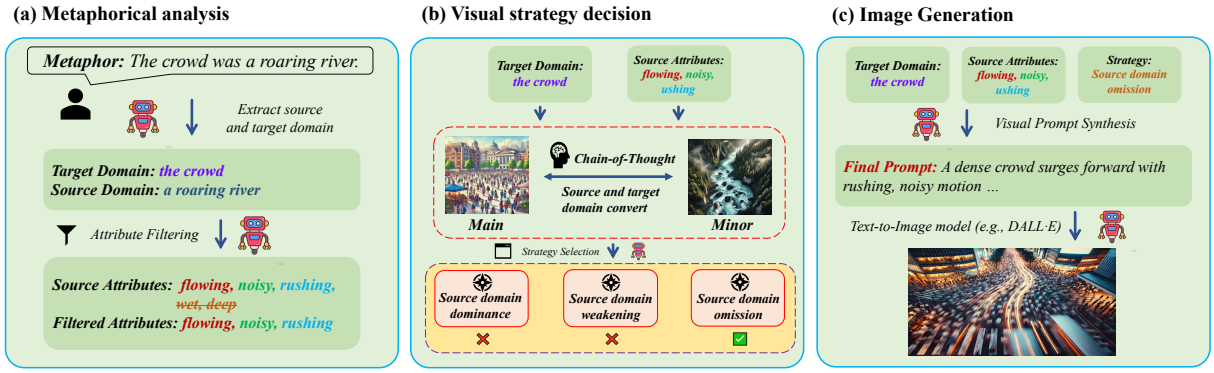

Figure 2: Overview of our CMIG framework for metaphorical image generation. (a) Metaphor analysis module: identifying source and target domains, extracting and filtering source domain attributes; (b) Visual strategy module: reasoning and selecting the expression strategy for source domain visual information.

alignment systems perform poorly in such settings, limiting the construction and utilization of large-scale metaphor datasets. To address these issues, several multimodal metaphor datasets and benchmarks have been introduced. MultiMet (Zhang et al., 2021) provides fine-grained annotations such as metaphor type, sentiment, and author intent. MetaCLUE (Akula et al., 2023) evaluates models' abilities in metaphor understanding and generation. IRFL (Yosef et al., 2023) covers metaphor, simile, and idiom comprehension across modalities, while HAIVMet (Chakrabarty et al., 2023) contains 6,476 metaphorical image samples to support language-to-vision metaphor research.

3 Method

CMIG is a **structured prompting framework** inspired by Conceptual Metaphor Theory (CMT), designed to convert metaphorical expressions into visually realizable metaphorical image descriptions in an interpretable, transferable, and stable manner. Unlike prompts that rely on free-form natural language Chain-of-Thought reasoning, CMIG **operationalizes** key principles in CMT—*cross-domain mapping*, *salience*, and *selective projection*—into a reproducible and compositional instruction template. The framework consists of four logical stages: *Metaphor Parsing*, *Attribute Mapping*, *Strategy Selection*, and *Visual Prompt Synthesis*. An overview of the pipeline is shown in Figure 2.

3.1 Metaphor Parsing

This stage aims to extract the cross-domain structure of a metaphorical input, including **domain identification** and **attribute mapping**.

Domain Extraction Given a metaphorical text x, the LLM identifies its Target Domain (T) and Source Domain (S):

$$(T, S) = \text{DomainExtraction}(x) \quad (1)$$

The target domain denotes the entity being described metaphorically, while the source domain provides semantic attributes to be projected. This corresponds to the backbone of the cross-domain mapping in CMT.

Attribute Mapping To prevent the model from generating overly concrete or semantically irrelevant content from the source domain, we adopt a two-step attribute processing procedure.

(1) Extracting Source Attributes. The LLM enumerates salient attributes of the source domain—typically adjectives or descriptive semantic properties:

$$A_s = \{a_1, a_2, \ldots, a_n\}. \quad (2)$$

(2) Attribute Filtering. To realize the notion of *selective projection* in CMT, we adopt a lightweight and reproducible text-based rule system to filter attributes that can be projected to the target domain. For each attribute, the LLM determines whether it should be retained according to the following criteria:

- If an attribute a can *reasonably describe or modify the target domain* (i.e., it does not introduce semantic contradiction or visual misleading), it is considered *semantically compatible*.
- If applying the attribute to the target domain would introduce *semantic conflict, visual misinterpretation, or conceptual incoherence*, it is removed.

Dependency (d)	Interference (i)	Strategy
High	Low	Dominant
Low	Low	Weakening
Low	High	Omission
High	High	Omission

Table 1: Strategy decision table based on dependency and interference tests.

- The model is required to provide a brief justification for each filtering decision to ensure transparency and interpretability.

The final retained and removed sets are:

$$A_{\text{kept}} = \{a \in A_s \mid \text{Filter}(a, T)\}. \quad (3)$$

$$A_{\text{removed}} = A_s \setminus A_{\text{kept}}. \quad (4)$$

This procedure operationalizes the notion of *selective projection* in CMT.

3.2 Strategy Selection

CMIG selects a visual realization strategy through a rule-based decision system that controls the degree to which the source domain appears in the final prompt. Two independent tests are used: **Dependency** and **Interference**. Dependency assesses whether the visual depiction of the target domain relies on the attributes of the source domain, while Interference evaluates whether visual cues from the source domain may mislead interpretation:

$$d \in \{\text{High}, \text{Low}\}, \qquad i \in \{\text{High}, \text{Low}\}. \quad (5)$$

The dependency–interference pair (d, i) is mapped to a visualization strategy following Table 1. This rule-based mechanism ensures stability and cross-model consistency. Three strategies are defined:

- **Dominant**: source-domain elements primarily drive the visual expression.
- **Weakening**: only low-salience and stylized elements of the source domain are preserved.
- **Omission**: all concrete source-domain elements are removed, retaining only abstract attributes.

Template 1: CMIG Structured Prompt

You are a metaphor interpreter.
Follow all steps strictly and produce the output in the required format.

[Input Metaphor]
"{METAPHOR_TEXT}"

Step 1: Identify domains
Identify the two conceptual domains involved in the metaphor:
Target domain (the entity being described):
Source domain (the domain providing attributes):

Step 2: Extract source attributes
List 3–6 salient attributes of the source domain (adjectives or semantic properties only):
$Attributes_s$: `[...]`

Step 3: Filter attributes
Filter the attributes by keeping only those that can reasonably describe or modify the target domain without causing semantic contradiction or visual misinterpretation.
Briefly justify removals if necessary.
$Attributes_{kept}$: `[...]`
$Attributes_{removed}$: `[...]`

Step 4: Choose strategy
Assess the two conditions using the rules below:
Dependency: High if removing the source-domain attributes would significantly weaken the intended metaphorical meaning; otherwise Low.
Interference: High if literal visualization of the source domain is likely to mislead or distort the intended meaning; otherwise Low.
`Dependency: High / Low`
`Interference: High / Low`

Select the strategy using:
If Interference = High → C (Source-Omission)
Else if Dependency = High → A (Source-Dominant)
Else → B (Source-Weakening)
`Strategy = {A / B / C}`

Step 5: Generate the final visual prompt
Write a concise visual description (60–80 words) that integrates:
(1) the target domain *T* as the main visual focus,
(2) the $Attributes_{kept}$ incorporated visually, and
(3) the selected strategy s to control how source-domain elements appear (dominant, weakened, or omitted).
`Final_Prompt:`

3.3 Structured Prompt and Example Outputs

Building on the above theoretical and rule-based components, we unify the four-stage inference process of CMIG (domain identification, attribute selection, strategy decision, and visual synthesis) into a structured prompting template. The template adopts a slot-filling format, enforcing a fixed out-

put order: target/source domains, extracted source attributes, filtering results, strategy choice, and the final visual prompt. To demonstrate its applicability across metaphor types, we generate three representative examples using this structured prompt (complete outputs are provided in Appendix 11).

4 Dataset Construction

We construct a contrastive metaphor–vision benchmark that aligns metaphorical *implied meaning* with visual realizations (Appendix Fig. 4). The dataset contains 400 metaphorical expressions collected from HAIVMet (Chakrabarty et al., 2023), IRFL (Yosef et al., 2023), and manually verified multi-domain examples. Each entry is annotated with a structured schema including implied meaning, target/source domains, a source-domain handling category, and the final T2I prompt (Appendix Fig. 5). Using DALL·E 3, we generate multiple candidate images per metaphor and retain human-verified positives that match the annotated figurative interpretation. To support contrastive evaluation, we additionally construct two types of negatives: **Literal** (surface rendering) and **Shallow-semantic** (partial mapping without deep grounding). Overall, the benchmark comprises 1,996 positive and 1,528 negative samples (>3,500 pairs total), enabling evaluation of metaphor understanding and metaphorical image generation. Full construction details, quality control, and recommended split are provided in Appendix §10.1.

5 Experiments

5.1 Experiment Design

We primarily evaluate CMIG on DALL·E 3 (OpenAI, 2023), and further examine its transferability on Imagen 2 (DeepMind, 2023) and FLUX-1 (Labs, 2024). CMIG implements metaphor-oriented prompting with an explicit Chain-of-Thought (CoT) structure that guides the model to (i) identify source–target relations, (ii) construct cross-domain semantic mappings, and (iii) choose among three generation strategies (*source-dominant*, *weakening*, *omission*). This design enables zero-shot generation of coherent and interpretable visual metaphors. Full prompt templates and the CoT design are provided in Appendix 11. We compare CMIG against four prompting baselines: Vanilla, HAICF (Chakrabarty et al., 2023), ViPE (Shahmohammadi et al., 2023), and SMIG (Su et al., 2024). To ensure reproducibility under double-blind review, we exclude instruction-following multi-step pipelines (e.g., GPT-4o-driven iterative prompting), whose outcomes may depend on non-deterministic intermediate decisions and are difficult to replicate reliably.

5.2 Evaluation Metrics

We use both automatic and human evaluations to measure semantic alignment, metaphor expressiveness, and perceived visual quality.

Automatic Metrics We adopt two complementary automatic metrics. **LIP** computes CLIP-based similarity between each generated image and the intended implicit metaphorical meaning. **BERT-Sim** captions each image using BLIP-2 and measures Sentence-BERT similarity between the caption and the intended meaning to assess semantic alignment.

Human Evaluation We conduct a controlled human study with nine trained annotators with backgrounds in linguistics and visual design. Annotators evaluate 800 images, each independently rated by three annotators. Following the standardized interface and guidelines in Appendix Figure 7, annotators assign 1–4 Likert scores (1=Poor, 4=Excellent) along three dimensions: *Metaphorical Appropriateness*, *Visual Quality*, and *Creativity*. To improve reliability, large-disagreement cases are flagged for expert adjudication (Appendix 7). Inter-annotator agreement achieves an average pairwise Cohen's $\kappa = 0.67$, indicating substantial consistency.

5.3 Implementation Details

All methods use identical prompt templates, model configurations, and generation parameters. We fix the output resolution to 1024×1024 and generate three images per metaphor. We use official APIs: GPT-4[1] (`gpt-4-0125`) for structured prompt construction; DALL·E 3[1] as the primary generator; Imagen 2[2] and FLUX-1[3] for cross-model evaluation. All models are run with default inference settings, without fine-tuning or post-processing.

5.4 Experimental Analysis

Table 2 reports results on three representative text-to-image (T2I) generators: FLUX-1, DALL·E 3, and Imagen 2. Across all generators and evaluation

[1] https://openai.com
[2] https://deepmind.google/models/imagen
[3] https://flux1.ai

Table 2: Automatic (**LIP**, **BERT**) and Human (**Met.**, **Vis.**, **Cre.**) evaluation across three T2I generators. * denotes statistical significance vs. Vanilla (paired t-test, $p < 0.05$). Best results are in **bold**. ↑: higher is better.

Prompting	T2I	Automatic Eval.		Human Eval.		
		LIP (↑)	BERT (↑)	Met. (↑)	Vis. (↑)	Cre. (↑)
Vanilla	FLUX-1	24.40	0.862	2.35	2.95	2.45
HAICF	FLUX-1	24.65*	0.866*	2.62*	3.08*	2.71*
ViPE	FLUX-1	24.55*	0.864*	2.58*	3.00	2.67*
SMIG	FLUX-1	24.90*	0.871*	**3.12***	3.07	2.86*
CMIG	FLUX-1	**25.02***	**0.875***	3.00*	**3.14***	**3.09***
Vanilla	DALL·E 3	24.50	0.864	2.55	3.02	2.58
HAICF	DALL·E 3	24.72*	0.868*	2.72*	3.14*	2.80*
ViPE	DALL·E 3	24.60*	0.867*	2.68*	3.06	2.72*
SMIG	DALL·E 3	24.88*	0.871*	3.08*	3.12	2.98*
CMIG	DALL·E 3	**25.12***	**0.877***	**3.18***	**3.20***	**3.22***
Vanilla	Imagen 2	24.70	0.866	2.65	3.07	2.70
HAICF	Imagen 2	24.82*	0.870*	2.82*	3.15*	2.85*
ViPE	Imagen 2	24.78*	0.869*	2.76*	3.08	2.80*
SMIG	Imagen 2	24.95*	0.874*	**3.24***	3.13	3.01*
CMIG	Imagen 2	**25.10***	**0.878***	3.22*	**3.18***	**3.28***

metrics, CMIG consistently outperforms Vanilla prompting and competitive baselines, achieving statistically significant improvements ($p < 0.05$) on both automatic semantic-alignment metrics and human judgments.

On automatic metrics, CMIG achieves stable and substantial gains across generators. On DALL·E 3, CMIG improves LIP from 24.50 to 25.12 and BERT-Sim from 0.864 to 0.877, indicating closer alignment between generated images and the implicit semantic structure of metaphorical prompts. Similar trends are observed on Imagen 2 (LIP $+0.40$, BERT-Sim $+0.012$) and FLUX-1 (LIP $+0.62$, BERT-Sim $+0.013$), demonstrating robust generalization across heterogeneous generation architectures. While strong baselines such as SMIG and HAICF also improve over Vanilla, their gains are consistently smaller, suggesting that CMIG's advantage arises from explicit modeling of metaphorical mappings rather than generator-specific prompt heuristics.

Human evaluation further corroborates the automatic results under a standardized annotation protocol. CMIG achieves the highest Creativity scores across all generators, improving over Vanilla by +0.64 on DALL·E 3, +0.58 on Imagen 2, and +0.64 on FLUX-1, indicating more expressive visual realizations without sacrificing semantic fidelity. CMIG also improves Visual Quality (e.g., +0.18 on DALL·E 3 and +0.19 on Imagen 2), suggesting more coherent compositions with fewer distracting artifacts. For Metaphorical Appropriateness, CMIG remains competitive with the strongest baseline SMIG (e.g., 3.18 vs. 3.24 on Imagen 2), while achieving higher Creativity and Visual Quality, reflecting a better balance between faithful metaphor grounding and expressive realization.

Qualitative Analysis. Qualitative comparisons in Figure 3 reveal recurring failure modes of baseline methods, including over-literalization, omission of contextual constraints, and stylistic drift that dilutes semantic focus. In contrast, CMIG translates metaphorical cognition into explicit visual strategies that preserve both conceptual intent and affect while maintaining compositional coherence. For example, it conveys ephemerality in "butterfly in autumn" through subdued tones and wing–leaf blending, captures enclosure in "a blanket of snow" via a snow-cloaked cityscape, and integrates medical and taxi cues in "a hospital bed is a parked taxi" to express urgency and cost. Even for abstract metaphors such as "broken heart" or geopolitical "fermenting", CMIG avoids diffuse symbolism and produces outputs that remain semantically aligned and visually grounded.

Figure 3: Qualitative comparison of metaphorical image generation results on DALLE·3. "Vanilla" denotes direct generation from the original metaphorical text, while the remaining columns correspond to images generated from prompts produced by their respective methods.

6 Ablation Studies

6.1 Experimental Design

We conduct two ablation studies to analyze CMIG components and prompt-generation robustness. We perform a component-wise ablation (A1–A6) by progressively removing key modules, including domain constraints, attribute extraction and filtering, strategy decision, and structured synthesis. A6 (Vanilla) removes the structured pipeline and uses an unstructured prompt, isolating the effect of structured prompting. Then, we evaluate different LLM-based prompt generators (LLaMA-3.1, GPT-4, DeepSeek R1) under both Vanilla and CMIG prompting. All prompts are rendered using DALL·E 3, with other settings identical to the main experiments.

6.2 Results and Analysis

Ablation 1: Module-wise. Relative to A6 (Vanilla), the full CMIG yields clear gains on both automatic semantic-alignment metrics and human ratings (LIP $+2.17$, BERT-Sim $+0.052$; Met./Vis./Cre. all increase), validating the effectiveness of structured prompting for metaphor visualization. Removing Domain Constraint or Structured Synthesis Constraint (A1/A5) causes moderate drops, consistent with their role in bounding semantic scope and enforcing composition-level coherence. In contrast, removing Source Attribute Extraction or Attribute Filtering (A2/A3) produces larger degradation across both alignment and human dimensions, suggesting that clean and relevant semantic primitives are crucial for reliable cross-domain projection. The most pronounced failure occurs without Strategy Decision (A4), where both alignment and metaphor clarity collapse (e.g., Met. $3.18 \rightarrow 2.35$), indicating that strategy selection is a key mechanism for binding extracted semantics into an executable visual plan. Overall, the trends support a coupled pipeline: extraction/filtering supplies usable semantic units, constraints stabilize

Table 3: Ablation study of CMIG components on DALL·E 3. ↑: higher is better. Vanilla corresponds to removal of the entire CMIG structured pipeline.

Method	**LIP (↑)**	**BERT (↑)**	**Met. (↑)**	**Vis. (↑)**	**Cre. (↑)**
CMIG (Full)	**25.12**	**0.877**	**3.18**	**3.20**	**3.22**
A1: Domain Constraint	24.85	0.870	3.00	3.08	3.15
A2: Source Attribute Extraction	24.20	0.858	2.75	2.92	2.95
A3: Attribute Filtering	24.10	0.857	2.70	2.90	2.90
A4: Strategy Decision	23.65	0.842	2.35	2.75	2.70
A5: Structured Synthesis Constraint	24.70	0.868	2.90	3.00	3.10
A6: Naive Prompt (Vanilla)	22.95	0.825	2.02	2.55	2.32

Table 4: Automatic (**LIP**, **BERT**) and human (**Met.**, **Vis.**, **Cre.**) evaluation on **DALL·E 3** using different LLMs for prompt generation. "-Ours" indicates prompts produced with our metaphor-guided generation framework. GPT-4 (Vanilla) is the provided anchor; GPT-4-Ours corresponds to CMIG-equivalent settings.

LLM / Size	**Automatic Eval.**		**Human Eval.**		
	LIP (↑)	**BERT (↑)**	**Met. (↑)**	**Vis. (↑)**	**Cre. (↑)**
(a) Comparison Among Different LLMs					
GPT-4	24.50	0.864	2.55	3.02	2.58
GPT-4-Ours	**25.12**	**0.877**	**3.18**	**3.20**	**3.22**
DeepSeek R1	24.65	0.866	2.60	3.05	2.62
DeepSeek R1-Ours	25.00	0.875	3.15	3.16	3.18
LLaMA3.1 405b	24.30	0.862	2.45	3.00	2.50
LLaMA3.1 405b-Ours	24.95	0.873	3.10	3.12	3.05
(b) Comparison Across LLaMA3.1 with Different Sizes					
LLaMA3.1 70b	24.40	0.863	2.50	3.03	2.53
LLaMA3.1 70b-Ours	25.00	0.872	2.95	3.14	2.90
LLaMA3.1 8b	24.10	0.858	2.20	2.95	2.30
LLaMA3.1 8b-Ours	24.60	0.866	2.55	3.05	2.60

coherence, and strategy converts mappings into effective synthesis.

Ablation 2: Prompt Generation. Across all evaluated LLMs, CMIG-enhanced prompting consistently improves LIP/BERT-Sim and raises human judgments, especially for Metaphorical Appropriateness and Creativity (Table 4). Although larger models start from stronger Vanilla baselines, CMIG delivers stable gains across scales (e.g., for LLaMA3.1: 8B $+0.50$ LIP; 70B $+0.60$; 405B $+0.65$), suggesting that the structured pipeline reduces prompt-generation bottlenecks even for high-capacity LLMs. DeepSeek R1 shows relatively strong metaphor reasoning under Vanilla, yet CMIG still substantially improves human scores, indicating that the framework benefits diverse LLM backbones by making metaphor mappings and constraints explicit and controllable.

7 Conclusion

We introduce CMIG, a metaphor-guided prompting framework that operationalizes key principles from Conceptual Metaphor Theory as an explicit chain-of-thought workflow for text-to-image generation. By decomposing metaphors into source–target mappings and selecting appropriate strategy types (source-dominant, weakening, or omission), CMIG enables zero-shot synthesis of visual metaphors that are more interpretable and less overliteral. Experiments on DALL·E 3, Imagen 2, and FLUX-1 show consistent improvements in automatic semantic-alignment metrics and human evaluations. We further release a 3,500-instance visual metaphor benchmark and a standardized evaluation protocol to facilitate futural research.

8 Limitations

Our work leaves several directions for further improvement. First, CMIG relies on an upstream LLM to produce structured analyses and prompts; while the workflow is template-controlled and stable in our experiments, stronger or more specialized LLMs may further improve mapping quality and strategy selection. Second, the automatic metrics we use (LIP and BERT-Sim) are designed to approximate semantic alignment, but they cannot fully capture all nuanced or highly implicit metaphor interpretations; we therefore treat them as complementary to controlled human evaluation. Third, our evaluation focuses on three widely used text-to-image generators under default API settings; extending the study to additional generators and sampling configurations is a natural next step and may further clarify generalization behavior. Finally, although our human study follows a standardized protocol with trained annotators, metaphor perception can vary across cultures and contexts, motivating future work on broader rater populations and multilingual settings.

9 Ethics Statement

Our study uses publicly available metaphor corpora and synthetic images generated via commercial APIs (DALL·E 3, Imagen 2, FLUX-1), and does not involve personal or sensitive user data. Human evaluation was conducted under a standardized interface and guidelines: annotators were recruited as domain-informed raters, compensated fairly, and could withdraw at any time. Because some metaphors may reference geopolitical or emotionally charged themes, we filtered examples to avoid explicit hate or harassment content and to exclude identifiable individuals. The released benchmark, prompts, and evaluation materials are intended for research use; deploying metaphor-generation systems in user-facing applications should incorporate additional safeguards (e.g., content moderation, bias auditing, and misuse prevention) appropriate to the target context.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Arjun R Akula, Brendan Driscoll, Pradyumna Narayana, Soravit Changpinyo, Zhiwei Jia, Suyash Damle, Garima Pruthi, Sugato Basu, Leonidas Guibas, William T Freeman, et al. 2023. Metaclue: Towards comprehensive visual metaphors research. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23201–23211.

Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. 2017. Cvae-gan: fine-grained image generation through asymmetric training. In *Proceedings of the IEEE international conference on computer vision*, pages 2745–2754.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8.

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 *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Google DeepMind. 2023. Imagen 2: Advancing text-to-image generation. Accessed: 2025-03-04.

Charles Forceville. 2002. *Pictorial metaphor in advertising*. Routledge.

Charles J Forceville and Eduardo Urios-Aparisi. 2009. *Multimodal metaphor*, volume 11. Walter de Gruyter.

Mengshi Ge, Rui Mao, and Erik Cambria. 2022. Explainable metaphor identification inspired by conceptual metaphor theory. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 10681–10689.

Black Forest Labs. 2024. Flux. https://github.com/black-forest-labs/flux.

George Lakoff and Mark Johnson. 2008. *Metaphors we live by*. University of Chicago press.

Zhenxi Lin, Qianli Ma, Jiangyue Yan, and Jieyu Chen. 2021. Cate: A contrastive pre-trained model for metaphor detection with semi-supervised learning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 3888–3898.

OpenAI. 2023. Dall·e 3 system card. Accessed: 2025-03-04.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Barbara J Phillips and Edward F McQuarrie. 2004. Beyond visual metaphor: A new typology of visual rhetoric in advertising. *Marketing theory*, 4(1-2):113–136.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3.

Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. 2019. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494.

Linda M Scott. 1994. Images in advertising: The need for a theory of visual rhetoric. *Journal of consumer research*, 21(2):252–273.

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

Kevin Stowe, Tuhin Chakrabarty, Nanyun Peng, Smaranda Muresan, and Iryna Gurevych. 2021. Metaphor generation with conceptual mappings. *arXiv preprint arXiv:2106.01228*.

Chang Su, Xingyue Wang, Shupin Liu, and Yijiang Chen. 2024. Efficient visual metaphor image generation based on metaphor understanding. *Neural Processing Letters*, 56(3):150.

Chang Su, Kechun Wu, and Yijiang Chen. 2021. Enhanced metaphor detection via incorporation of external knowledge based on linguistic theories. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1280–1287.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

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.

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.

10 Appendix

10.1 Dataset Construction Details

Data sources and scope. To evaluate whether metaphor-aware prompting improves metaphorical image generation, we construct a contrastive benchmark that explicitly aligns metaphorical semantics with visual realizations. We collect approximately 400 metaphorical expressions from two public metaphor resources (HAIVMet (Chakrabarty et al., 2023), IRFL (Yosef et al., 2023)), and complement them with manually verified metaphorical instances from everyday, political, and technological discourse to broaden topical diversity. Each metaphorical expression is normalized into a single textual form (removing duplicates, resolving minor lexical variants, and excluding non-metaphorical or highly context-dependent cases).

Annotation schema. Each entry is annotated into five structured fields (Figure 5): **(i) Metaphor** (surface expression), **(ii) Implied Meaning** (intended figurative interpretation in context-neutral form), **(iii) Target Domain** and **(iv) Source Domain**, **(v) Strategy Category** indicating how the source domain should be handled in visual realization (e.g., dominance/weakening/omission), and the final **Prompt** used for text-to-image (T2I) generation.

Annotation protocol and agreement. As shown in Figure 4, the dataset is built through a five-stage pipeline.

1. **Metaphor collection and screening.** We aggregate candidate metaphors and screen them to ensure (a) metaphorical authenticity, (b) diversity in target/source domains, and (c) sufficient interpretability without requiring long contextual passages.

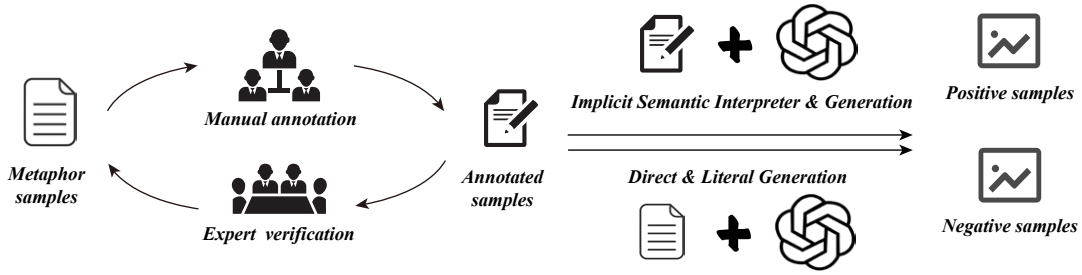

Figure 4: Overview of the data annotation and generation workflow, consisting of manual annotation and expert verification stages. Positive samples are generated via the Implicit Semantic Interpreter & Generation module, while negative samples are derived through the Direct & Literal Generation module.

Figure 5: Annotation schema and examples from the dataset. Each metaphorical text is labeled with implicit semantics, target domain, source domain, category, and corresponding prompts. The right panel displays paired positive and negative image examples: positive samples align with metaphorical meaning, while negative samples illustrate literal or surface interpretations.

2. **Manual annotation.** Six trained annotators independently label implied meaning, target/source domains, and strategy category for each metaphor. Annotators follow a shared guideline that emphasizes: (1) writing implied meaning as a literal paraphrase of the intended figurative claim; (2) selecting target/source domains at a consistent granularity; and (3) assigning strategy category based on whether a visually explicit source-domain depiction would help or distract from the intended metaphorical interpretation.

3. **Expert verification and adjudication.** A committee of three senior researchers adjudicates disagreements, harmonizes taxonomy usage, and standardizes domain granularity across entries. Inter-annotator agreement exceeds 0.8 (Cohen's κ) after adjudication. We also maintain an adjudication log to document recurrent disagreement patterns and resolution rules.

Positive sample generation (figurative). Positive samples are generated via an *Implicit Semantic Interpreter & Generation* module (Figure 4). For each annotated metaphor, we synthesize a figurative T2I prompt conditioned on the fields in Figure 5, including the implied meaning and the intended target/source-domain mapping. We then generate multiple candidate images using DALL·E 3 and apply a two-step filtering process: (1) **semantic fidelity** (the image reflects the implied meaning and the intended conceptual mapping), and (2) **visual coherence** (the image is globally coherent and avoids artifacts that undermine interpretability). This stage yields 1,996 verified positive images.

Negative sample design (contrastive). To support contrastive evaluation, we construct two complementary negative types (Figure 5)

1. **Literal negatives:** prompts that directly render the surface/literal reading of the metaphor, intentionally ignoring figurative intent.

2. **Shallow-semantic negatives:** prompts that retain partial domain elements but omit the deep metaphorical grounding implied by the annotated interpretation.

Negative samples are validated to remain strictly

non-figurative (literal) or semantically incomplete (shallow), resulting in 1,528 verified negatives.

Quality control. We apply a three-stage QC protocol throughout: (a) annotator self-checking against the shared guideline, (b) cross-review to detect subjective bias and inconsistent granularity, and (c) expert adjudication for disputed items and schema enforcement. Additionally, all metaphor–image pairs undergo multimodal validation to ensure that positives align with the annotated implied meaning, while negatives do not unintentionally convey the same figurative interpretation (Figure 5)

Dataset statistics and recommended split. The final benchmark contains over 3,500 metaphor–image pairs (1,996 positives and 1,528 negatives), covering approximately 400 metaphorical expressions with diverse target/source domains. To avoid leakage, we recommend splitting by *metaphor expression* (group-wise split), ensuring that all images derived from the same metaphor belong to the same partition (train/validation/test).

11 Detailed CoT Examples and Generated Visual Prompts

To enable the model to robustly perform metaphor interpretation and visual strategy planning under zero-shot conditions, we construct a set of structured Chain-of-Thought (CoT) examples as semantic guidance templates. These examples follow the four-stage reasoning pipeline of CMIG (domain identification, attribute mapping, interference assessment, and strategy selection), helping the model establish a consistent paradigm of metaphor reasoning in an unsupervised setting and achieve interpretable transfer reasoning from metaphorical text to visual strategies.

In the first example, the target domain *"Time"* is a highly abstract concept lacking inherent visual structure, whereas the source domain *"money"* provides rich concrete visual symbols such as coins, banknotes, and metallic textures that reinforce key semantics of "value," "scarcity," and "preciousness." Through the CoT demonstration, the model learns that when the target domain relies on the source domain to construct a visual representation, and when the source imagery does not introduce semantic ambiguity, it should adopt a Source-Dominant strategy. The resulting visual prompt renders temporal imagery (e.g., hourglasses or clocks) using money-like materials to highlight the latent semantic features of the abstract concept of time.

The second example illustrates a scenario in which the target domain is intrinsically visualizable. The entity *"He"* can directly serve as the visual carrier of the metaphor, while the concrete depiction of the source domain *"star"* (e.g., astronomical-star icons) may distract from the centrality of the human figure. The CoT example enables the model to learn that when the target domain is readily depictable and when concrete source imagery may visually overshadow the target without causing semantic confusion, a Source-Weakening strategy is preferred. The resulting prompt retains attributes such as "shine", "gloss", and "diffused brightness" from the source domain, but omits the star itself, emphasizing meanings such as "prominence" and "being in the spotlight," while preserving the human figure as the visual focus.

The third example demonstrates a case in which visual symbols from the source domain could lead to misinterpretation. The source domain *"fermenting"* evokes strong physical-process imagery (e.g., bubbling, swelling, liquid transformation), which, if visualized directly, would undermine the seriousness of the political context and introduce metaphorical misreading. Through the CoT example, the model learns that when the concrete form of the source domain deviates severely from the target context, it should employ a Source-Omission strategy, removing all concrete source elements and retaining only its abstract tension, such as "growing instability," "accumulating pressure," and "approaching a critical point." The generated visual scene instead uses elements such as smoke, tense atmosphere, crowd dynamics, and regional landmarks to portray the escalating geopolitical situation, without any physical imagery related to fermentation.

11.1 Human Evaluation Details

To improve consistency and reproducibility in subjective evaluation, we implement a standardized web-based labeling system (Figure 7). The interface contains three modules: (i) a metaphor text panel, (ii) an image preview area, and (iii) a structured rating form. Raters first read the metaphor (e.g., "Time flies") and then assess the paired generated image (e.g., a clock formed by birds in motion to convey the abstract notion of time passing swiftly). To minimize potential bias, the interface does not reveal method names or polarity labels; im-

Example1: "Time is money"

Step 1: Identify domains
Target domain: *Time*
Source domain: *Money*

Step 2: Extract source attributes
Attributes_S:
{precious, limited, measurable, tradable}

Step 3: Filter attributes
Attributes_Kept:
{precious, limited, measurable}
Attributes_Removed:
{tradable}

Step 4: Choose strategy
Dependency: *High*
Interference: *Low*
Strategy *= A (Source-Dominant)*

Step 5: Final visual prompt
Final_Prompt:
A clock constructed from golden coins, with coin stacks forming the hour markers and thin bills flowing like clock hands. An hourglass made of transparent currency shapes melts slowly into shimmering money fragments. The scene emphasizes time as precious, limited, and measurable, blending temporal symbols with monetary textures to highlight its inherent value.

Example2: "is a shining star"

Step 1: Identify domains
Target domain: *A man (the person)*
Source domain: *Shining star*

Step 2: Extract source attributes
Attributes_S:
{outstanding, exceptional, radiant, attention-drawing}

Step 3: Filter attributes
Attributes_Kept:
{radiant, attention-drawing, exceptional}
Attributes_Removed:
{outstanding}

Step 4: Choose strategy
Dependency: *Low*
Interference: *Low*
Strategy *= Source-Weakening*

Step 5: Final visual prompt
Final_Prompt:
A portrait of a confident, exceptional man standing under warm ambient light. A subtle radiant glow outlines his figure, with soft star-like highlights around the shoulders and hair. Light beams gently emphasize his presence without overpowering the scene, conveying brilliance an

Example3: "The situation in the Middle East is fermenting"

Step 1: Identify domains
Target domain: *Situation in the Middle East*
Source domain: *Fermentation*

Step 2: Extract source attributes
Attributes_S:
{escalating, unstable, accumulating tension, nearing outbreak}

Step 3: Filter attributes
Attributes_Kept:
{escalating, unstable, accumulating tension}
Attributes_Removed:
{nearing outbreak}

Step 4: Choose strategy
Dependency: High
Interference: High
Strategy *= Source-Omission*

Step 5: Final visual prompt
Final_Prompt:
A tense Middle Eastern landscape rendered with rising smoke, distant flashes of explosions, and dark clouds gathering above urban skylines. Groups of anxious civilians and armed forces highlight escalating instability. The scene visually conveys mounting tension and geopolitical volatility without using any fermentation imagery, focusing solely on conflict dynamics.

Figure 6: Examples of Chain-of-Thought (CoT) prompts designed under the CMIG framework. Each metaphorical expression is annotated with its target and source domains, recommended visual strategy. These structured prompts guide the model to reason about metaphor interpretation and visual planning in a cognitively aligned manner.

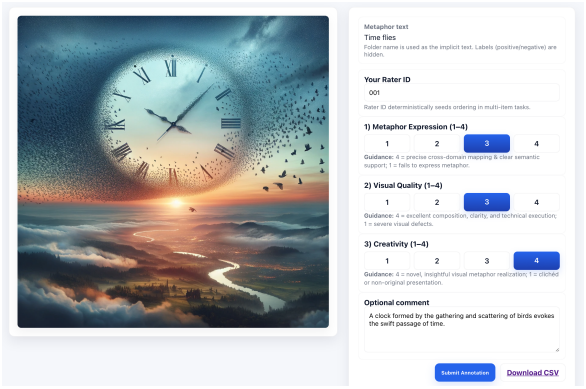

Figure 7: User interface of the standardized human evaluation system. The interface includes three sections: metaphor text display, generated image preview, and a structured rating form. Raters view the metaphor on the left (e.g., "Time flies") and its corresponding generated image on the right (e.g., a clock formed by a flock of birds symbolizing the fleeting nature of time). Each rater is assigned a unique ID to enable deterministic task allocation and traceable evaluation logs.

ages are anonymized and presented in a deterministic order. Each rater is assigned a unique identifier (e.g., `Rater ID 001`), which deterministically seeds the task assignment and evaluation order for traceability.

Each image is evaluated on a four-point Likert scale (1=Poor, 4=Excellent) along three dimensions, with brief on-screen guidelines: (1) *Metaphorical Appropriateness* (a.k.a. Metaphor Expression) evaluates whether the image captures the intended cross-domain mapping and implicit semantic relation (e.g., whether it conveys "time" as something that can "fly"); (2) *Visual Quality* assesses composition, clarity, and aesthetic coherence, while penalizing obvious artifacts; and (3) *Creativity* measures the novelty and insightfulness of the visual metaphor beyond conventional depictions. The interface additionally supports optional free-text comments to justify ratings.

For reliability, each image is independently rated by three raters, and the final score is computed by averaging across raters. We flag large-disagreement cases for secondary review: if any dimension differs by more than one point across raters, expert reviewers conduct adjudication and may revise the final decision. All scores, timestamps, and metadata are automatically logged and exported as CSV files for subsequent quantitative analysis.

11.2 Visualization of Ablation Results

To further examine the effectiveness of CMIG in metaphor visualization, we conducted a qualitative comparison of different large language models (LLMs) as metaphorical prompt generators. Figure 8 illustrates results for three representative metaphors—"He was like a butterfly in autumn," "She wears different hats to earn a livelihood," and "He has a heart of gold." Each metaphor was processed by five LLMs (GPT-4, DeepSeek-R1, LLaMA-405B, LLaMA-70B, and LLaMA-8B). For each model, the left column (Original) shows results generated from the unmodified metaphor text, while the right column (Ours) presents images generated using CMIG-structured prompts, highlighting the framework's contribution to metaphor grounding and visual abstraction.

For "He was like a butterfly in autumn", GPT-4's baseline generates butterflies and autumn leaves but

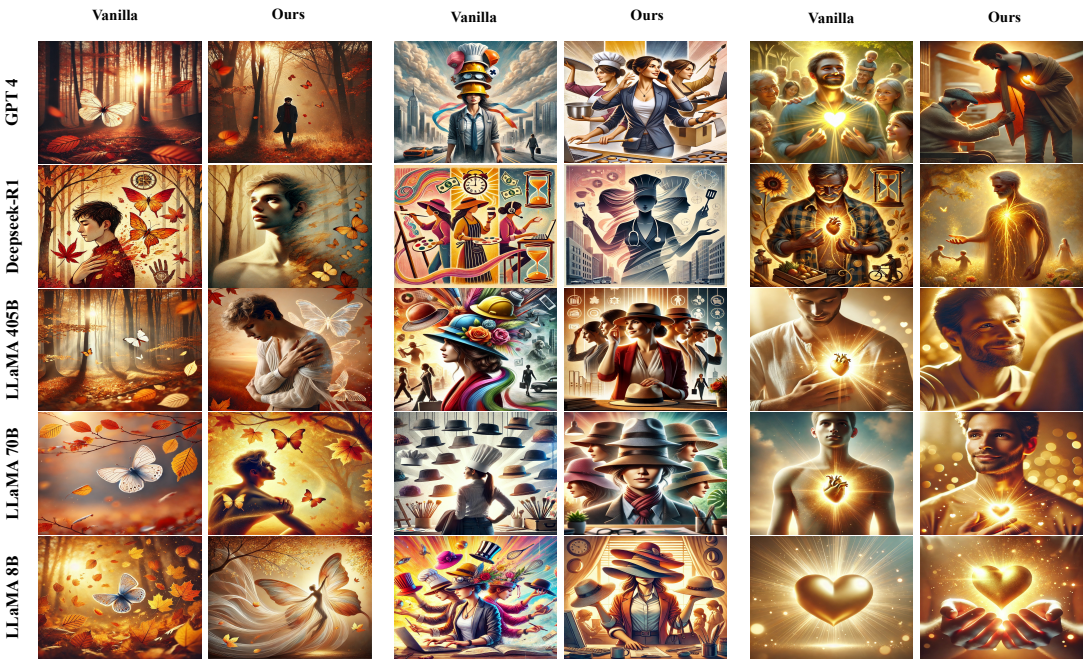

Figure 8: Metaphorical image generation across different models. For each example, the left column (Vanilla) shows results generated directly from the original metaphorical text, while the right column presents images generated using prompts produced by the CMIG framework.

omits the human element, whereas CMIG creates a poetic scene of a figure gazing at drifting butterflies, enhancing metaphorical depth. DeepSeek-R1's baseline offers a static composition, while CMIG merges human and butterfly imagery to express transience and introspection. LLaMA-405B transitions from a literal autumn landscape to a depiction rich in emotional nuance, and LLaMA-70B evolves from a simple butterfly–leaf pattern to a dynamic figure with upward motion, improving aesthetic coherence. Even LLaMA-8B, despite limited capacity, progresses from a plain close-up of butterflies to a creative human–butterfly fusion. These comparisons show that CMIG consistently enhances metaphor recognizability and creative abstraction, particularly benefiting smaller LLMs with weaker semantic reasoning abilities.

In "She wears different hats to earn a livelihood", which conveys professional multiplicity, GPT-4's baseline depicts a single figure balancing multiple hats but lacks contextual coherence. The CMIG version transforms this into a symbolic collage integrating varied professions, clocks, and monetary motifs, reflecting adaptive labor. DeepSeek-R1's baseline merges disparate symbols, while CMIG organizes them into coherent silhouettes within a dynamic urban setting. LLaMA-405B progresses from static stacking to interactive multi-role scenes, and LLaMA-70B introduces creative persona blending, such as detective and artisan archetypes. Even for LLaMA-8B, CMIG enhances the original fantasy-like rendering with richer workplace elements. Across all models, CMIG consistently strengthens narrative coherence and metaphorical depth, highlighting its structured support for conceptual reasoning in prompt generation.

For "He has a heart of gold", which expresses kindness and generosity, GPT-4's baseline generates a glowing heart embedded in the chest, while CMIG enriches the imagery with human interactions—such as acts of empathy—amplifying emotional resonance. DeepSeek-R1 transitions from abstract heart–flower motifs to luminous gestures conveying compassion. LLaMA-405B enhances a basic golden heart with halo-like radiance and contextual figures, while LLaMA-70B extends the composition into a socially grounded scene emphasizing moral warmth. Even LLaMA-8B evolves from a static glowing heart to a vivid depiction of a person embracing it, demonstrating tangible metaphor embodiment. Overall, CMIG systematically improves metaphor interpretability and creative fidelity across LLMs of different sizes, vali-

dating its role as a cognitively grounded framework for metaphorical image generation.