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

Qingbao Huang[1,5], Cheng Yang[1,5], Jiawei Yao[1], Zhiyue Liu[2]*, Yi Cai[3], Xingmao Zhang[4]*
[1]School of Electrical Engineering, Guangxi University, Nanning 530004, China
[2]School of Computer, Electronics and Information, Guangxi University, Nanning 530004, China
[3]School of Software Engineering, South China University of Technology, China
[4]College of General Education, Guangxi Arts University, Nanning 530022, China
[5]Guangxi Academy of Artificial Intelligence, Nanning, 530200, China
qbhuang@gxu.edu.cn, {2212391065, 2212391071}@st.gxu.edu.cn, liuzhy@gxu.edu.cn,
ycai@scut.edu.cn, 20170024@gxau.edu.cn

Abstract

Metaphorical text expresses meaning through cross-domain mappings rather than literal surface content, which makes it difficult for text-to-image systems to generate semantically faithful images. We propose CMIG, a structured prompting framework inspired by Conceptual Metaphor Theory (CMT). CMIG identifies source–target mappings, filters projectable source attributes, and selects a visual realization strategy in a reproducible reasoning workflow. Experiments on DALL·E 3, Imagen 2, and FLUX-1 show that CMIG consistently improves semantic alignment and yields a better overall balance of human-rated metaphor quality, visual coherence, and controllability on metaphorical prompts. To support systematic evaluation, we also construct a 3,500-instance visual metaphor benchmark.

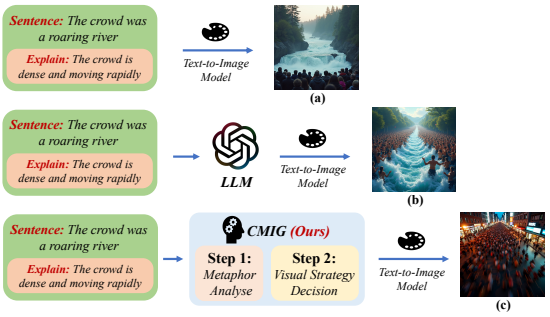

Figure 1: Qualitative comparison on the metaphor "The crowd was a roaring river" (intended meaning: a dense, fast-moving crowd). (**a**) Direct prompting tends to generate a literal river-related scene. (**b**) Chain-of-thought prompting partially improves interpretation but remains semantically inconsistent.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

1 Introduction

Metaphor is pervasive in natural language and often conveys meaning beyond literal description. This makes metaphorical image generation challenging: a text-to-image system must recover the intended figurative meaning rather than simply render surface words. In practice, current text-to-image models often produce visually plausible but semantically literal outputs (Rombach et al., 2022; Betker et al., 2023). For example, given "The crowd was a roaring river," a model may generate an actual river scene instead of a dense, fast-moving crowd.

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

A natural starting point is Conceptual Metaphor Theory (CMT) (Lakoff and Johnson, 2008), which views metaphor understanding as selectively projecting attributes from a source domain onto a target domain. However, turning CMT into a computational procedure is difficult because source–target correspondences are often implicit, attribute salience is context-dependent, and multiple visual realizations may be valid. Rather than learning CMT end-to-end, we use it to structure a controllable reasoning workflow for large language mod-

*Corresponding authors.

Findings of the Association for Computational Linguistics: ACL 2026, pages 23748–23761

els (LLMs). Specifically, we decompose metaphor interpretation into explicit steps for domain identification, source-attribute extraction, attribute filtering, and visual planning.

Existing prompting-based approaches, including chain-of-thought prompting, semantic expansion, and metaphor-aware rewriting (Su et al., 2024; Shahmohammadi et al., 2023; Chakrabarty et al., 2023), still have three limitations. First, they typically lack an explicit source–target semantic structure. Second, they often rely on free-form or ad-hoc elaboration, which reduces interpretability and controllability. Third, they tend to over-emphasize literal source-domain content or miss the intended figurative implication. These limitations suggest that stronger backbone generators alone are not enough for metaphorical image generation.

We therefore propose CMIG, a structured prompting framework for metaphorical image generation. CMIG contains two modules: *metaphor parsing*, which identifies domains and derives projectable attributes, and visual strategy selection, which determines how source-domain information should be expressed using a Dependency Test and an Interference Test. The resulting prompt integrates figurative semantics with explicit visual control, enabling text-to-image models to better preserve metaphor meaning while maintaining visual coherence. Our contributions are as follows:

1. We introduce **CMIG**, a CMT-inspired structured prompting framework that improves semantic alignment and visual coherence on metaphorical prompts over prior prompting baselines.

2. We present an explicit design for metaphor parsing and visual strategy selection, making metaphor realization more interpretable and controllable.

3. We construct a 3,500-instance visual metaphor benchmark spanning 400 metaphors to support systematic evaluation of metaphor-aware prompting methods.

2 Related Work

2.1 Metaphorical Image Generation

Recent diffusion-based text-to-image models, such as DALL·E 2 (Ramesh et al., 2022), Imagen (Saharia et al., 2022), Stable Diffusion (Rombach et al., 2022), and FLUX.1 (Labs, 2024), have greatly improved visual fidelity and diversity. However, prior studies show that these models still struggle with metaphorical inputs, often producing literal or semantically misaligned images (Yosef et al., 2023; Akula et al., 2023).

To improve metaphorical image generation, recent work has explored prompt engineering and semantic rewriting. ViPE (Shahmohammadi et al., 2023) rewrites figurative text into depictable visual descriptions. HAICF (Chakrabarty et al., 2023) combines LLM prompting and chain-of-thought reasoning to elaborate metaphorical meaning for image generation. SMIG (Su et al., 2024) abstracts source–target semantics and uses CLIP-based guidance to produce metaphor-enhanced prompts. These methods demonstrate the value of linguistic reasoning, but they mostly handle metaphor implicitly through free-form elaboration, rewriting, or embedding-level manipulation. By contrast, CMIG models metaphor as an explicit source–target mapping and adds strategy-level control over how source-domain information should be visually realized.

2.2 Multimodal Metaphor Understanding

Metaphor understanding has been studied in NLP through tasks such as metaphor detection, interpretation, and generation (Lin et al., 2021; Su et al., 2021; Ge et al., 2022; Stowe et al., 2021). In multimodal settings, the problem is harder because models must align figurative language with visual content rather than literal objects alone. This challenge has motivated several datasets and benchmarks. MultiMet (Zhang et al., 2021) provides fine-grained annotations for multimodal metaphors. MetaCLUE (Akula et al., 2023) and IRFL (Yosef et al., 2023) evaluate metaphor understanding across vision and language. HAIVMet (Chakrabarty et al., 2023) provides metaphor-related image data for language-to-vision research. These resources have advanced multimodal metaphor research, but they primarily target understanding or specific generation settings. In this work, our benchmark is designed as evaluation infrastructure for systematically comparing metaphor-aware prompting methods.

3 Method

CMIG is a structured prompting framework for metaphorical image generation, inspired by Con-

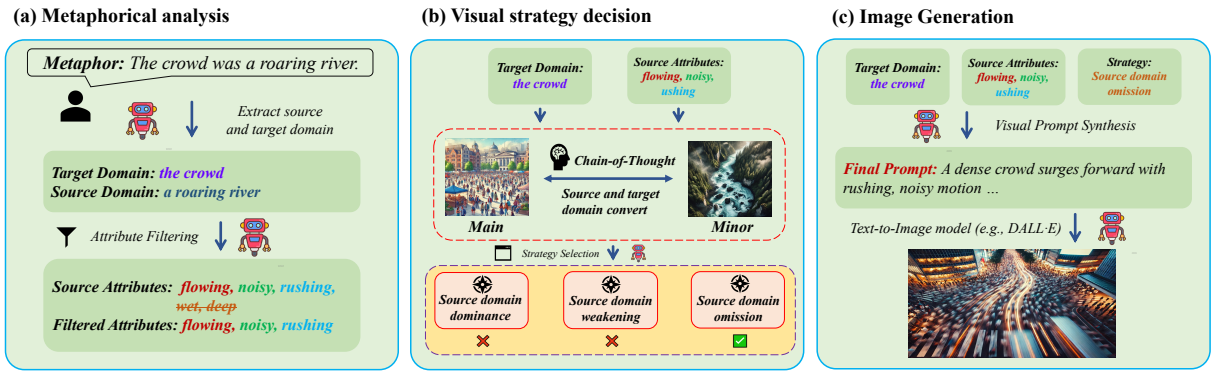

Figure 2: Overview of our CMIG framework for metaphorical image generation. (a) Metaphor analysis module: identifying source and target domains, extracting and filtering source domain attributes; (b) Visual strategy module: reasoning and selecting the expression strategy for source domain visual information.

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.

ceptual Metaphor Theory (CMT). Rather than treating metaphor through free-form explanation or rewriting, CMIG explicitly models the source–target mapping, filters projectable source attributes, and selects a visual realization strategy before prompt synthesis. In this way, CMIG operationalizes three core ideas from CMT—cross-domain mapping, salience, and selective projection—within a reproducible prompting workflow. The framework contains four stages: Metaphor Parsing, Attribute Mapping, Strategy Selection, and Visual Prompt Synthesis. Figure 2 gives an overview.

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):

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

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:

$$A_s = \{a_1, a_2, \ldots, a_n\}. \tag{2}$$

(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 misinterpretation), 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.
- 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:

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

$$A_{\text{removed}} = A_s \setminus A_{\text{kept}}. \tag{4}$$

This procedure operationalizes the notion of selective projection in CMT.

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:`

3.2 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.

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 output 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). To improve consistency, we normalize expressions by removing duplicates and minor lexical variants, and perform manual annotation under

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***

shared guidelines with expert adjudication. Using DALL·E 3, we generate multiple candidate images for each metaphorical input and retain human-verified positive samples that correctly realize the annotated figurative meaning. To support contrastive evaluation, we additionally construct two types of negative samples for the same metaphorical input: Literal (surface rendering) and Shallow-semantic (partial mapping without deep grounding). Thus, positive and negative samples do not correspond to different input types (e.g., metaphorical vs. literal text), but to contrasting visual realizations of the same metaphorical expression. Overall, the benchmark contains 1,996 positive and 1,528 negative samples, yielding over 3,500 annotated metaphor–image instances for evaluating metaphor understanding and metaphorical image generation. Positive samples are filtered for semantic fidelity and visual coherence, while negative samples are validated to remain literal or semantically incomplete. We recommend a group-wise split by metaphor expression to avoid leakage across train/validation/test partitions. Full construction details, quality control, and recommended split are provided in Appendix §10.1.

5 Experiments

5.1 Experiment Design

We evaluate CMIG on DALL·E 3 (OpenAI, 2023), and further test its transferability on Imagen 2 (DeepMind, 2023) and FLUX-1 (Labs, 2024). CMIG performs metaphor-oriented prompting via explicit source–target analysis, attribute filtering, and strategy control, enabling zero-shot generation of coherent visual metaphors. Full prompt templates are provided in Appendix 11. We compare CMIG with four representative baselines. **Vanilla** directly feeds the original metaphorical text to the T2I model. **HAICF** (Chakrabarty et al., 2023) uses LLM-based free-form elaboration. **ViPE** (Shahmohammadi et al., 2023) rewrites figurative inputs into more depictable descriptions. **SMIG** (Su et al., 2024) strengthens source-domain cues through metaphor-enhanced prompting. In contrast, **CMIG** explicitly models source–target mappings grounded in CMT and applies rule-governed strategy selection to regulate source-domain realization.

To make these differences explicit, Table 3 summarizes the design contrast between prior prompting methods and CMIG. We additionally include a Paraphrase baseline on DALL·E 3 to test whether CMIG's gains reduce to generic rewriting, and a literal-prompt control to examine possible over-correction outside the intended metaphor setting.

5.2 Evaluation Metrics

We use both automatic and human evaluations.

Automatic Metrics **LIP** measures CLIP-based similarity between the generated image and the in-

Figure 3: Qualitative comparison of metaphorical image generation results on DALL·E 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.

Method	Modeling	Synthesis	Strategy	Depend.
HAICF	Implicit	Free-form	–	LLM
ViPE	Implicit	Rewriting	–	Trained LM
SMIG	Implicit	Prompting	–	CLIP/T2I
CMIG	Explicit	Rule-guided	✓	Low

Table 3: Conceptual differences between prior prompting baselines and CMIG.

tended metaphorical meaning. **BERT-Sim** captions each image using BLIP-2 and computes Sentence-BERT similarity between the caption and the intended meaning. For the literal-prompt control, the original literal caption is used as the reference text.

Human Evaluation We recruit nine trained annotators with backgrounds in linguistics and visual design. They rate 800 images, each scored independently by three annotators, on Metaphorical Appropriateness, Visual Quality, and Creativity using a 1–4 Likert scale. Large-disagreement cases are adjudicated by experts.

Metric Validation Because LIP and BERT-Sim are used as automatic proxies for figurative alignment, we further validate them against human Metaphorical Appropriateness. On 400 images, both metrics show moderate and significant correlation with human judgments (LIP: $r = 0.46$, $p < 0.01$; BERT-Sim: $r = 0.42$, $p < 0.01$), supporting their use as complementary automatic signals rather than replacements for human evaluation.

5.3 Implementation Details

We use identical generation settings across methods. Images are generated at 1024×1024, with three images per metaphor. We use GPT-4[1] (`gpt-4-0125`) for prompt construction, DALL·E 3[1] as the primary generator, and Imagen 2[2] and FLUX-1[3] for cross-model evaluation. All models use default inference settings without

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

Method	LIP (↑)	BERT-Sim (↑)
Base	24.50	0.864
Paraphrase	24.52	0.865
CMIG	**25.12**	**0.877**

Table 4: Comparison with a semantic paraphrasing baseline on DALL·E 3.

fine-tuning or post-processing. For the Paraphrase baseline, we use FLAN-T5-large (Chung et al., 2024)[4] with deterministic decoding to produce one rewrite per input while preserving meaning and metaphorical intent. For the literal robustness control, we sample 100 literal prompts from MSCOCO 2017 validation captions and compare Base, CMIG-always, and CMIG-gated under GPT-4 + DALL·E 3. For error analysis, we randomly sample 100 metaphorical cases from the DALL·E 3 evaluation set and manually categorize failures into four types: Literal Bias, Source Dominance, Misbinding, and Missing Source/Target.

5.4 Experimental Analysis

Table 2 shows that CMIG consistently outperforms Vanilla and prior baselines across FLUX-1, DALL·E 3, and Imagen 2. On DALL·E 3, CMIG improves LIP from 24.50 to 25.12 and BERT-Sim from 0.864 to 0.877. Similar gains are observed on Imagen 2 and FLUX-1, indicating good cross-model transferability. Human evaluation shows that CMIG achieves the best overall balance between metaphorical appropriateness, visual quality, and creativity. Compared with strong baselines such as SMIG, the gains are not merely due to stronger source emphasis, but to more controlled metaphor realization through explicit mapping and strategy selection. To test whether these gains come merely from rewriting, we compare CMIG with a semantic paraphrasing baseline in Table 4. Paraphrasing yields only marginal improvement over direct prompting, whereas CMIG provides substantially larger gains. This suggests that CMIG benefits mainly from structured cross-domain mapping rather than surface-level lexical variation.

Error Analysis. Average scores alone do not fully capture failure patterns. We therefore compare SMIG and CMIG on 100 randomly sampled metaphorical cases in Table 5. SMIG fails mainly through Source Dominance, where literal source imagery overwhelms the target scene. CMIG reduces this error type from 10% to 3% and lowers the total failure rate from 22% to 16%, indicating better interference control and structural robustness. The slight increase in Literal Bias and Missing Source/Target reflects a more conservative suppression of source imagery.

[4] https://huggingface.co/google/flan-t5-large

Error Type	SMIG	CMIG
Literal Bias	5%	6%
Source Dominance	10%	3%
Mis-binding	4%	3%
Missing Source/Target	3%	4%
Total Failure Rate	22%	16%

Table 5: Error analysis on 100 randomly sampled metaphorical cases generated with DALL·E 3.

Setting	LIP (↑)	BERT-Sim (↑)
Base	**25.30**	**0.889**
CMIG-always	25.18	0.884
CMIG-gated	25.29	0.888

Table 6: Robustness control on purely literal prompts using GPT-4 + DALL·E 3.

Qualitative Analysis. Figure 3 shows that baseline methods often over-literalize metaphors or introduce distracting source-domain elements. In contrast, CMIG better preserves figurative intent while maintaining visual coherence.

5.5 Robustness on Literal Inputs

We further test whether CMIG harms generation on non-metaphorical inputs. Table 6 shows that applying CMIG indiscriminately to literal prompts causes a small drop relative to Base, suggesting that CMIG should not be used unconditionally outside its intended setting. However, detector-gated deployment remains nearly identical to Base, indicating that CMIG can be safely bypassed for literal inputs without harming literal fidelity.

6 Ablation Studies

6.1 Experimental Design

We conduct two ablations on DALL·E 3. The first removes key CMIG components, including domain constraint, source attribute extraction, attribute filtering, strategy decision, and structured synthesis. The second compares Vanilla and CMIG prompting across different prompt-generation LLMs (GPT-4, DeepSeek R1, and LLaMA-3.1 at multiple scales). All other settings follow the main experiments.

Table 7: 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 8: 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

6.2 Results and Analysis

Ablation 1: Module-wise. Table 7 shows that the full CMIG substantially outperforms the unstructured prompt setting, confirming the effectiveness of the structured pipeline. Among all components, Strategy Decision is the most critical: removing it causes the largest drop across both automatic and human metrics. Removing Source Attribute Extraction or Attribute Filtering also leads to large degradation, while removing Domain Constraint or Structured Synthesis Constraint causes smaller but consistent declines.

Ablation 2: Prompt Generation Across LLMs. Table 8 shows that CMIG consistently improves performance across all tested LLMs, including GPT-4, DeepSeek R1, and LLaMA-3.1 at different scales. This indicates that CMIG is robust to the choice of prompt generator and remains effective even when the backbone LLM is already strong.

7 Conclusion

We introduce CMIG, a metaphor-guided prompting framework that operationalizes key principles of Conceptual Metaphor Theory for text-to-image generation. By decomposing metaphors into source–target mappings and selecting suitable strategy types (source-dominant, weakening, or omission), CMIG enables zero-shot synthesis of visual metaphors that are more interpretable and less literal. Experiments on DALL·E 3, Imagen 2, and FLUX-1 show consistent gains in automatic semantic alignment and overall human evaluation. We also release a 3,500-instance visual metaphor benchmark and a standardized evaluation protocol to support future research.

8 Limitations

CMIG currently assumes a single dominant source–target mapping and performs strategy selection at the sentence level. It is therefore less suitable for multi-entity or nested metaphors that require mapping-level coordination. CMIG is also designed for metaphorical prompts rather than as a universal prompt enhancer: our literal-prompt control suggests that indiscriminate use on non-metaphorical inputs may introduce mild abstraction, making gated deployment preferable. In addition, CMIG relies on an upstream LLM for structured analysis and prompt construction, and stronger or more specialized LLMs may further improve mapping quality and strategy selection. Although LIP and BERT-Sim correlate significantly with human judgments, they remain imperfect proxies and cannot fully capture subtle or highly implicit metaphor interpretations. Finally, our error analysis also suggests a trade-off: while CMIG reduces source-dominance failures, it may slightly increase more conservative errors such as literal bias or missing source/target cues.

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.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62276072), National Natural Science Foundation Youth Project (62406081), the Guangxi Natural Science Foundation Key Project (2025GXNSFDA069017), Guangxi Natural Science Foundation (2025GXNSFBA069232), the Guangxi Philosophy and Social Science Research Annual Project 2025 (No. 25WYF489), and the Bagui Scholar Program of Guangxi (2025) and the Guangxi Bagui Youth Talent Program (2025).

References

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.

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*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

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.

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.

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.

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.

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.

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.

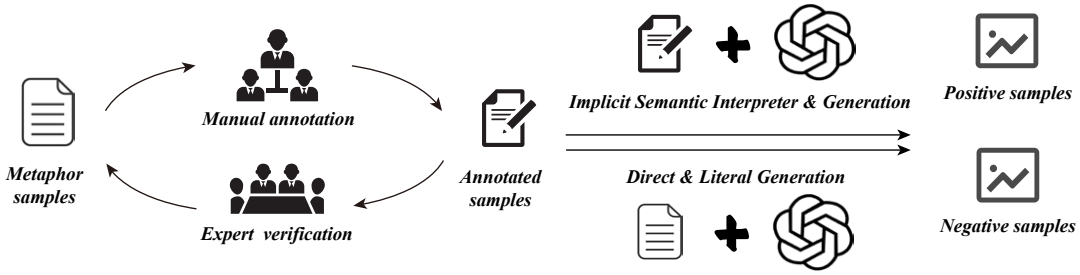

Figure 4: Overview of the data annotation and generation workflow, consisting of manual annotation and expert verification stages. Positive samples are generated via the Implicit Semantic Interpreter & Generation module, while negative samples are derived through the Direct & Literal Generation module.

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.

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; images 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* as-

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.

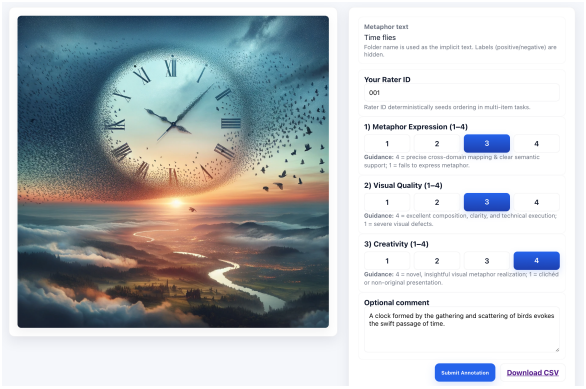

Figure 7: User interface of the standardized human evaluation system. The interface includes three sections: metaphor text display, generated image preview, and a structured rating form. Raters view the metaphor on the left (e.g., "Time flies") and its corresponding generated image on the right (e.g., 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.

sesses 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 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 com-

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.

parisons 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, validating its role as a cognitively grounded framework for metaphorical image generation.