# BLENDING CONCEPTS: BENCHMARKING VISUAL METAPHOR GENERATION IN TEXT-TO-IMAGE MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Text-to-image (T2I) models have achieved remarkable success at faithfully rendering specified objects and attributes, yet their ability to produce *visual metaphors*—images that convey abstract ideas by combining elements from two distinct domains—remains largely unexamined. To bridge this gap, we introduce **VMetaphor-Bench**, a fine-grained diagnostic benchmark for visual metaphor generation in T2I models. It comprises 1,500 visual metaphors curated from real-world creative imagery, organized into three levels and ten categories, with each sample paired with two prompts of differing specificity. For evaluation, we develop a hybrid framework within an MLLM-as-judge paradigm, combining a multiple-choice question (MCQ) based protocol of 9,594 questions across four levels of metaphorical fidelity with a dimension-based scoring protocol along three perceptual dimensions. Extensive evaluation of 12 representative T2I models reveals that even the strongest proprietary models struggle with compositional structuring and cross-domain mapping, key aspects of metaphorical expression, highlighting visual metaphor generation as an important frontier for future T2I research.

## 1 INTRODUCTION

Recent years have witnessed remarkable progress in text-to-image (T2I) generation, driven by advances across three main architectural directions: diffusion models (Ho et al., 2020; Rombach et al., 2022), autoregressive models (Team, 2024; Tian et al., 2024), and unified multimodal architectures (Chen et al., 2025b; Deng et al., 2025; Wu et al., 2025b). Powered by large-scale training, a wide range of T2I systems, from leading proprietary models such as GPT-Image-1 (OpenAI, 2025) and Gemini 3.1 Flash Image (Google, 2026) to open-source counterparts (Labs, 2024; Wu et al., 2025a), can now faithfully render specified objects, attributes, and spatial layouts, achieving strong performance on basic semantic prompt following. However, their ability to express abstract concepts and convey meaning beyond what is literally depicted remains far less explored.

An important form of such abstract expression is the visual metaphor, where one concept is understood through another via visual analogy (Forceville et al., 2008). Drawing on conceptual blending theory (Fauconnier & Turner, 2003), a visual metaphor combines elements from two distinct domains — a source that lends its properties, and a target that receives them — to produce a single image carrying meaning richer than either domain alone. Generating such images is far from straightforward: the model must (i) blend the two given domains coherently using a specific structure (e.g., fusion or juxtaposition), (ii) establish meaningful element-level mappings between the two, and (iii) ensure that the resulting image conveys the intended meaning to viewers. As shown in Figure 1, even state-of-the-art T2I models often struggle with these tasks, producing images that miss one domain, adopt the wrong structural form, blend the two awkwardly, or fail to convey the intended meaning.

Yet this challenge has been largely overlooked by existing T2I benchmarks, which remain dominated by literal prompt following. Mainstream benchmarks such as GenEval (Ghosh et al., 2023) and T2I-CompBench (Huang et al., 2023) evaluate object, attribute, and basic spatial fidelity from short prompts, while recent reasoning-aware benchmarks (Niu et al., 2025; Sun et al., 2025a; Zhao et al., 2025) extend to factual or physical inference, leaving the cross-domain conceptual blending required by visual metaphors unexamined. On the metaphor side, existing efforts have focused predominantly on the understanding direction (Akula et al., 2023; Kundu et al., 2025; Zhang et al., 2026); the generation probe within MetaCLUE (Akula et al., 2023) is limited to 300 samples assessed via FID,



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

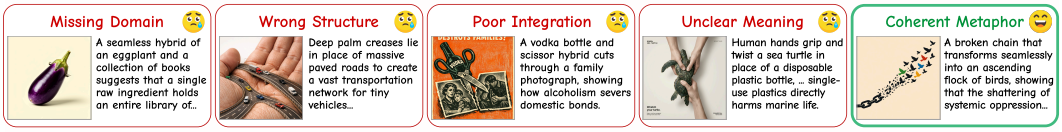


Figure 1: Common failure modes in visual metaphor generation from GPT-Image-1.5 (OpenAI, 2025), Seedream 5.0 Lite (ByteDance Seed, 2026), and FLUX.2-dev (Labs, 2025).

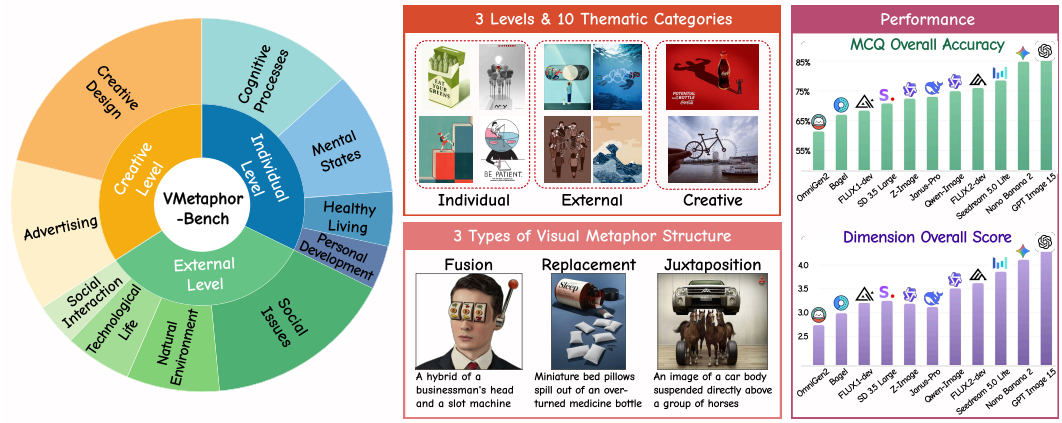


Figure 2: Overview of VMetaphor-Bench. **Left:** hierarchical taxonomy with three levels and ten thematic categories. **Middle:** examples spanning the three levels and the three structure types. **Right:** performance of T2I models on MCQ accuracy and dimension scoring under conceptual prompts.

CLIP similarity, and a small user study — none of which can systematically diagnose where and how T2I models fail at metaphorical expression. Together, these limitations leave a clear gap: no existing benchmark systematically evaluates how well T2I models generate visual metaphors across diverse themes and metaphorical structures.

To address this gap, we introduce **VMetaphor-Bench**, a new benchmark for evaluating visual metaphor generation in T2I models. It comprises 1,500 visual metaphors curated from real-world creative imagery, organized into three levels and ten categories, as shown in Figure 2. Each sample is annotated with its source and target domains, structure type, metaphorical meaning, element-level mappings, and paired with two prompts: a conceptual prompt conveying only the metaphorical idea, and a descriptive prompt detailing its visual realization. For evaluation, we develop a hybrid framework comprising an MCQ-based protocol with 9,594 questions targeting four levels of metaphorical fidelity (*domain presence*, *structure type*, *meaning*, and *element mapping*) and a dimension-based scoring protocol rating each image on *metaphoric efficacy*, *metaphor logic*, and *perceptual harmony*. Both protocols operate within an MLLM-as-judge paradigm, enabling fine-grained, interpretable diagnosis of T2I models' strengths and weaknesses in visual metaphor generation.

Using VMetaphor-Bench, we conduct a comprehensive evaluation of **12 representative T2I models**, spanning leading proprietary systems, open-source diffusion models, and unified multimodal architectures. Our results reveal that proprietary models such as GPT Image 1.5 and Nano Banana 2 achieve substantially better performance than open-source counterparts. However, even the strongest models struggle with compositional structuring and cross-domain mapping—key aspects of metaphorical expression—highlighting visual metaphor generation as an important frontier for future T2I research.

In summary, our contributions are threefold:

- We introduce **VMetaphor-Bench**, a fine-grained diagnostic benchmark for visual metaphor generation in T2I models, comprising 1,500 high-quality, carefully curated samples organized into a hierarchical taxonomy of three levels and ten categories.
- We design a hybrid evaluation framework with an MCQ-based protocol of 9,594 questions across four levels of metaphorical fidelity and a dimension-based protocol along three perceptual dimensions, enabling fine-grained, interpretable assessment.
- We conduct a comprehensive evaluation of 12 representative T2I models, providing insights into their capabilities in visual metaphor generation and highlighting key directions for future research.

# 2 Related Works

## 2.1 Text-to-Image Generation Models

Text-to-image (T2I) generation has advanced rapidly along three main architectural directions. Diffusion models (Dhariwal & Nichol, 2021; Ho et al., 2020; Rombach et al., 2022) remain the dominant paradigm, with recent progress driven by scaling transformer-based denoisers (Esser et al., 2024; Labs, 2024; Peebles & Xie, 2023). Autoregressive models (Sun et al., 2024; Team, 2024; Tian et al., 2024) cast image generation as sequential token prediction, offering fine-grained compositional control through explicit modeling of inter-token dependencies. More recently, unified multimodal models (Chen et al., 2025b; Deng et al., 2025; Wu et al., 2025b; Xie et al., 2025) integrate visual understanding and generation within a single architecture, leveraging multimodal reasoning to better interpret complex prompts. State-of-the-art systems such as GPT-Image-1 (OpenAI, 2025) and Gemini 3.1 Flash Image (Google, 2026) now produce images of remarkable fidelity and aesthetic quality for everyday scenes. However, generating images that convey abstract semantic relationships, such as visual metaphors, remains a significant challenge across all three paradigms.

## 2.2 Text-to-Image Generation Benchmarks

Early evaluation of text-to-image generation relied on distribution-level metrics such as FID (Heusel et al., 2017) and feature-based metrics like CLIPScore (Hessel et al., 2021), which capture perceptual quality but fail to assess fine-grained semantic correctness. The recent rise of multimodal large language models (MLLMs) (Hurst et al., 2024; Team et al., 2023) has enabled a paradigm shift toward MLLM-as-judge evaluation (Chen et al., 2024), where vision-language models directly examine generated image against prompts, providing a more flexible and semantically grounded assessment framework. Under this paradigm, a line of benchmarks targets foundational semantic alignment — verifying whether generated images faithfully depict the specified objects (Kamath et al., 2025), attributes (Ye et al., 2025), and spatial relationships (Wang et al., 2025; 2026). Representative works include GenEval (Ghosh et al., 2023), T2I-CompBench (Huang et al., 2023), and TIFA (Hu et al., 2023), which decompose prompts into atomic verification questions to systematically evaluate compositional fidelity. More recently, benchmarks have shifted toward higher-order reasoning capabilities, probing whether T2I models can incorporate world knowledge (Niu et al., 2025; Zhang et al., 2025a) and reasoning (Chen et al., 2025a; Sun et al., 2025a; Zhao et al., 2025) into their generations, reflecting a broader trend from evaluating object-level fidelity toward scene-level semantic understanding.

## 2.3 Visual Metaphor

Visual metaphor communicates abstract ideas by mapping one conceptual domain onto another through visual imagery (Ervas, 2021; Forceville et al., 2008; Van Mulken et al., 2014). On the understanding side, MetaCLUE (Akula et al., 2023) introduces a benchmark for metaphor classification, localization, interpretation, along with a small probe on generation; ImageMet (Kundu et al., 2025) benchmarks VLMs on metaphor captioning and VQA, and MetaphorStar (Zhang et al., 2026) applies reinforcement learning to improve reasoning. On the generation side, Creative Blends (Sun et al., 2025b) explores commonsense-guided concept fusion, The Mind's Eye (Koushik et al., 2025) combines VLM-based metaphor alignment scores with generic metrics for evaluation and generation refinement on HAIVMet, and Beyond Pixels (Xu et al., 2026) introduces schema-driven metaphor transfer. CMIG (Huang et al., 2026) uses structured prompting to visualize linguistic metaphors and evaluates semantic alignment with LIP, BERT-Sim, and human ratings. Existing evaluations typically use generic metrics or human judgments on small, task-specific

Table 1: Comparison with prior visual metaphor generation evaluations.

| Work | Data scale | Eval. scale | Real-world | Evaluation type | Multi-level diagnosis |
|---|---|---|---|---|---|
| *I Spy a Metaphor* | 1,540 | 100 | No | Human | No |
| *Efficient VM Generation* | 216 | 200 | No | Generic metrics + human | No |
| MetaCLUE | 5,061 | 300 | Yes | Generic metrics | No |
| CMIG | 3,524 | 400 | No | Generic metrics + human | No |
| **VMetaphor-Bench** | 1,500 | 1,500 | Yes | **VM-specific automated evaluation** | Yes |

test sets (Chakrabarty et al., 2023; Su et al., 2024). As summarized in Table 1, VMetaphor-Bench provides a unified, visual-metaphor-specific automated protocol for systematic multi-level diagnosis.

## 3 VMETAPHOR-BENCH

We present VMetaphor-Bench, a structured diagnostic benchmark for visual metaphor generation in text-to-image models, featuring multi-dimensional assessments of cross-domain conceptual blending. We describe the dataset construction process in Section 3.1 and detail the evaluation framework in Section 3.2. An overview of the complete pipeline is illustrated in Figure 3.

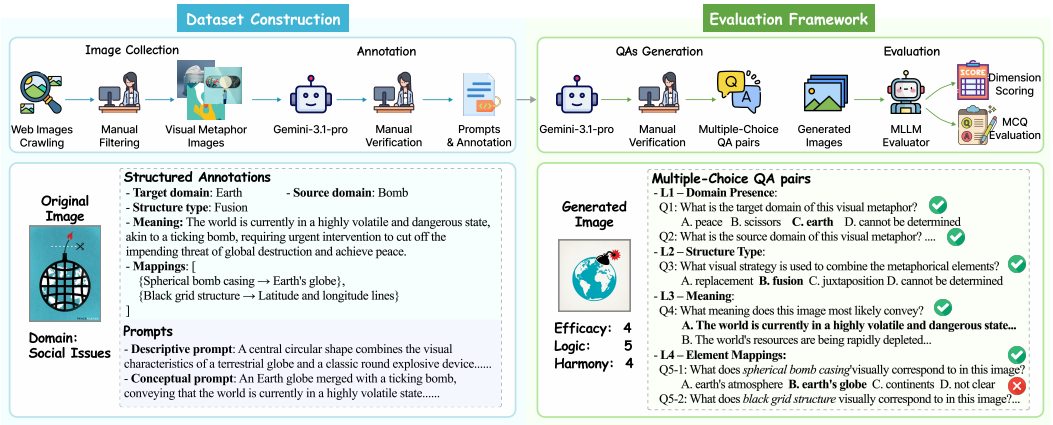


Figure 3: Overview of the VMetaphor-Bench pipeline. **Left: Dataset Construction** — images are collected and annotated through a two-stage pipeline, producing structured annotations and two granularity levels of prompts. **Right: Evaluation Framework** — generated images are assessed via MCQ-based evaluation (L1–L4 levels) and dimension-based scoring (Efficacy, Logic, Harmony).

### 3.1 DATASET CONSTRUCTION

**Image Collection.** To construct a benchmark grounded in real-world visual metaphors, we collect images from Pinterest (pin), a platform rich in creative visual content such as advertisements, editorial illustrations, and conceptual art. We query the platform using over ten keywords spanning both general and domain-specific terms, e.g. *visual metaphor*, *creative ads*. This process yields approximately 5,000 candidate images. Two annotators then review the entire collection, filtering out duplicates, images that do not contain a clear visual metaphor, and low-quality or ambiguous samples, resulting in approximately 1,600 images for subsequent annotation.

**Data Annotation.** After collecting images, we employ Gemini 3.1 Pro (Google DeepMind, 2026) as the annotation backbone for its strong visual reasoning capabilities. Since visual metaphors involve multi-layered semantics (Cappa et al., 2025), generating accurate annotations and high-quality prompts simultaneously in a single pass is challenging. We therefore adopt a two-stage annotation pipeline. In the first stage, the model is presented with each image and tasked with extracting structured annotations:

- **Target domain**: the subject being described — what the image is ultimately "about".
- **Source domain**: the concept used to characterize the target— it lends a property to the target through visual analogy.
- **Structure type**: how the two domains are visually combined. Following Phillips et al. (2004), we categorize it as *replacement* (one domain is absent and inferred from context), *fusion* (both domains are merged into a single object), or *juxtaposition* (both domains appear as separate objects in the same scene).
- **Meaning**: the overall message conveyed through the interaction of the two domains.
- **Element mappings**: fine-grained correspondences specifying which source element maps to which target element.

In the second stage, the model receives both the image and the extracted annotations, and generates two prompts at different levels of granularity: (1) The *descriptive prompt* provides a detailed depiction

of the visual content and how the metaphor is physically rendered in the image; (2) The *conceptual prompt* specifies the target and source domains, implicitly conveys the structure type, and states the intended meaning, without describing specific visual details or element-level mappings. Additionally, a style suffix describing the visual style of the original image (e.g., photorealistic, illustration) is extracted and appended to both prompts during generation to ensure stylistic consistency. This two-stage design allows the structured annotations to guide prompt generation, ensuring consistency between the structured annotations and the textual descriptions.

The two annotators involved in image screening and a third author, all trained in visual metaphor theory and our annotation guidelines, perform rigorous manual inspection of all images, annotations, and prompts, and cross-validate their decisions. Ambiguous or unrecognizable metaphors are removed, while inaccurate annotations (e.g., misidentified domains) and imprecise prompts are corrected to match the visual content. This verification process yields our final set of 1,500 high-quality images, each accompanied by structured annotations and two granularity levels of prompts.

### 3.2 Evaluation Framework

Evaluating visual metaphor generation requires assessing both metaphorical fidelity (whether the image faithfully captures the intended cross-domain mapping) and expressive quality (how effectively and aesthetically the metaphor is conveyed). To this end, we adopt a hybrid evaluation framework with MLLM-as-judge (Zhang et al., 2025b), comprising two complementary approaches: (1) MCQ-based evaluation that measures the correctness of metaphorical elements through multiple-choice questions, and (2) dimension-based scoring that rates the overall quality of the generated metaphor on a 1-5 Likert scale. An overview of the evaluation framework is shown in Figure 3 (Right).

**MCQ-based Evaluation.** To objectively assess the correctness of generated images at each level of the metaphorical structure, we automatically construct a set of multiple-choice questions (MCQs) for all 1,500 samples in the dataset. Specifically, we leverage Gemini 3.1 Pro to generate MCQs based on each sample's structured annotations along with its descriptive prompt. The generated MCQs assess four hierarchical levels of metaphorical fidelity:

- **L1 – Domain Presence**: Whether the target and source domains are visually recognizable in the image (one question per domain).
- **L2 – Structure Type**: Whether the visual combination strategy matches the ground-truth structure type (replacement, fusion, or juxtaposition).
- **L3 – Meaning**: Whether the image conveys the intended metaphorical meaning.
- **L4 – Element Mapping**: Whether each source element correctly corresponds to its intended target element (one question per mapping pair).

This process yields 9,594 MCQs across all samples, manually verified by the same annotators to ensure correctness. Each question consists of four options: one ground-truth answer derived from the annotation, two plausible but incorrect distractors, and a "cannot be determined" option. All option orders are randomized to eliminate positional bias. We report micro-averaged accuracy within each level and over all evaluated questions for the overall score.

**Dimension-based Scoring.** While MCQ-based evaluation measures the correctness of metaphorical elements, it does not capture the overall perceptual quality of the generated image. We therefore complement it with dimension-based scoring, where an MLLM evaluator is presented with the generated image along with its structured annotations and asked to rate the image on a 1–5 Likert scale across three independent dimensions:

- **Metaphoric Efficacy**: How clearly the image communicates the intended metaphorical meaning.
- **Metaphor Logic**: Whether the cross-domain mapping between source and target is conceptually coherent and logically sound.
- **Perceptual Harmony**: How naturally and coherently the two domains are visually presented, considering blending, composition, and overall visual cohesion.

The overall dimension score averages the three dimensions. Together with the MCQ-based evaluation, this hybrid framework provides a comprehensive assessment of both the metaphorical fidelity and expressive quality of generated images. Full evaluation prompts are provided in Appendix A.10.

Table 2: Evaluation results on VMetaphor-Bench with **conceptual prompts**.

| | MCQ Accuracy (%) | | | | | Dimension Score (1–5) | | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | Domain | Structure | Meaning | Mapping | Overall | Efficacy | Logic | Harmony | Overall |
| *Closed-source T2I Models* | | | | | | | | | |
| Seedream 5.0 Lite | 90.6 | 65.6 | 96.0 | 66.5 | 78.5 | 3.78 | 3.74 | 4.08 | 3.86 |
| Nano Banana 2 | 93.0 | 76.3 | **98.5** | 75.9 | 84.8 | 4.05 | 4.08 | 4.21 | 4.11 |
| GPT Image 1.5 | **93.6** | **78.1** | 97.6 | **76.5** | **85.4** | **4.27** | **4.23** | **4.38** | **4.30** |
| *Open-source T2I Models* | | | | | | | | | |
| SDXL Base | 74.1 | 48.2 | 77.3 | 52.3 | 62.4 | 2.24 | 2.38 | 3.41 | 2.68 |
| FLUX.1-dev | 82.3 | 54.9 | 85.6 | 55.2 | 68.4 | 2.90 | 2.87 | 3.82 | 3.19 |
| SD 3.5 Large | 83.0 | 60.1 | 85.7 | 58.9 | 70.8 | 2.99 | 3.01 | 3.75 | 3.25 |
| Z-Image | 84.3 | 59.6 | 89.5 | 60.8 | 72.4 | 2.93 | 3.03 | 3.67 | 3.21 |
| Qwen-Image | 87.3 | 66.1 | 90.5 | 61.7 | 74.9 | 3.39 | 3.31 | 3.84 | 3.52 |
| FLUX.2-dev | 88.6 | 63.7 | 92.1 | 63.9 | 76.0 | 3.48 | 3.46 | 3.93 | 3.62 |
| *Open-source Unified MLLMs* | | | | | | | | | |
| OmniGen2 | 74.8 | 48.1 | 77.1 | 49.1 | 61.4 | 2.33 | 2.44 | 3.45 | 2.74 |
| Bagel | 79.8 | 54.8 | 84.2 | 54.2 | 67.0 | 2.69 | 2.72 | 3.56 | 2.99 |
| Janus-Pro | 83.4 | 59.1 | 89.1 | 63.2 | 73.0 | 2.90 | 3.02 | 3.43 | 3.12 |

# 4 EXPERIMENTS

In this section, we evaluate representative T2I models on VMetaphor-Bench and analyze their behavior under different prompt granularities. We further validate the reliability of our MLLM-as-judge framework through cross-judge consistency and human alignment studies. Additional experimental details and results are provided in the Appendix.

## 4.1 SETUP

**Models.** We evaluate 12 representative text-to-image models spanning three categories: (1) **Closed-source T2I Models.** GPT Image 1.5 (OpenAI, 2025), Nano Banana 2 (Gemini 3.1 Flash Image) (Google, 2026), and Seedream 5.0 Lite (ByteDance Seed, 2026). These proprietary models represent the current state of the art in text-to-image generation; (2) **Open-source T2I Models.** SDXL Base (Podell et al., 2023), Stable Diffusion 3.5 Large (Stability AI, 2024), FLUX.1-dev (Labs, 2024), FLUX.2-dev (Labs, 2025), Z-Image (Team, 2025), and Qwen-Image (Wu et al., 2025a). These models represent established and leading open-source text-to-image models; (3) **Open-source Unified MLLMs.** OmniGen2 (Wu et al., 2025b), Bagel (Deng et al., 2025), and Janus-Pro (Chen et al., 2025b). These models employ a unified architecture for both image generation and multimodal understanding. All models are evaluated using their default generation configurations.

**MLLM Evaluator.** For both MCQ-based evaluation and dimension-based scoring, we adopt an open-source MLLM, Qwen3.5-27B (Qwen Team, 2026), as the default judge model for its strong visual reasoning capability and full reproducibility. We additionally report results with GPT-5.4 (OpenAI, 2026) in the Appendix A.6 to validate cross-judge consistency.

## 4.2 MAIN RESULTS ON VMETAPHOR-BENCH

We report the main evaluation results with conceptual prompts in Table 2. All metrics are assigned by Qwen3.5-27B as the default judge. Based on the results across 12 models, we highlight several key findings regarding model capabilities and core challenges in visual metaphor generation.

**Closed-source models lead by a substantial margin.** GPT Image 1.5 achieves the best overall performance with 85.4% MCQ accuracy and a dimension score of 4.30, followed by Nano Banana 2 (84.8%, 4.11). In contrast, the best-performing open-source T2I model, FLUX.2-dev, trails by 9.4 percentage points (76.0%, 3.62). While FLUX.2-dev narrows the gap with the weakest closed-source model Seedream 5.0 Lite (78.5%), a clear performance tier remains, suggesting that visual metaphor generation—which demands both semantic comprehension and creative compositional ability—still substantially benefits from the larger-scale training and architectural advantages of proprietary systems. Unified MLLMs show mixed performance: while Janus-Pro (73.0%) is competitive with mid-tier open-source T2I models, OmniGen2 (61.4%) and Bagel (67.0%) rank at the bottom despite their multimodal understanding capabilities. As an older baseline, SDXL Base achieves 62.4% MCQ

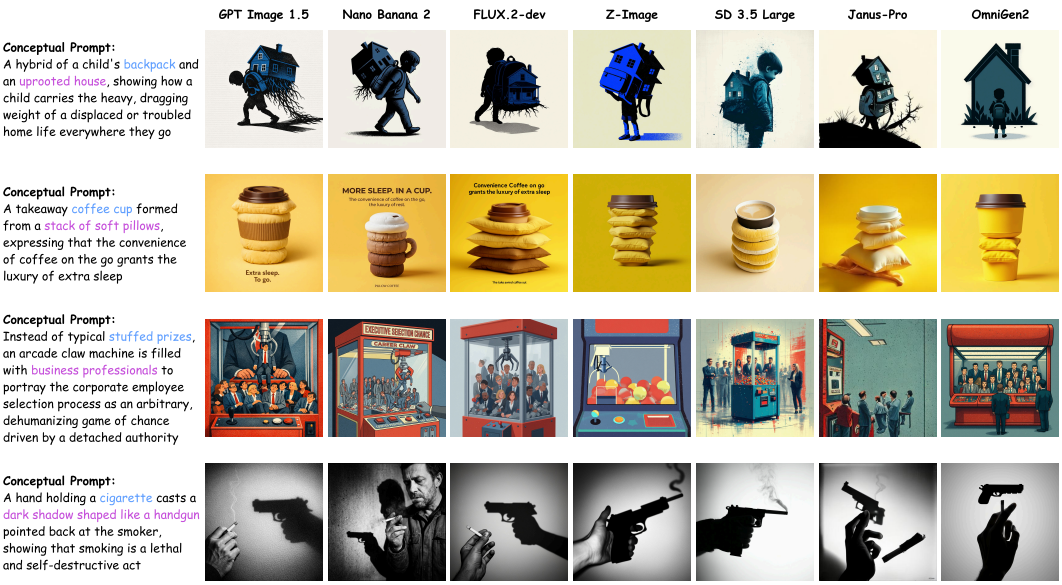


Figure 4: Qualitative comparison of seven models on four visual metaphors. The four metaphors cover three structure types: fusion (rows 1–2), replacement (row 3), and juxtaposition (row 4). Conceptual prompts appear on the left, with target and source domains highlighted in blue and purple.

accuracy and a 2.68 dimension score, substantially trailing SD 3.5 Large (70.8%/3.25) and illustrating progress within the Stable Diffusion family.

**Structure and Mapping are the primary bottleneck.** All models exhibit a consistent difficulty hierarchy across MCQ levels: Domain and Meaning accuracies are consistently the two highest among the four levels for every model, indicating that models can generate images containing the correct conceptual elements and conveying the intended metaphorical meaning. However, performance drops sharply on Structure (48.1–78.1%) and Mapping (49.1–76.5%), revealing that the core challenge lies in compositional structuring—models can depict the right elements but struggle to compose them into coherent metaphorical structures. This pattern is particularly evident in Seedream 5.0 Lite, which achieves 90.6% on Domain and 96.0% on Meaning yet drops to 65.6% on Structure and 66.5% on Mapping. Notably, a deeper analysis of structure type errors reveals that *replacement* is far more challenging than *fusion* or *juxtaposition*, a pattern that mirrors the cognitive complexity hierarchy in visual rhetoric theory (Phillips & McQuarrie, 2004) (see Appendix A.4).

**Visually harmonious but metaphorically shallow.** Across all models, Perceptual Harmony is consistently the highest-scoring dimension (3.41–4.38), while Metaphoric Efficacy and Metaphor Logic generally trail behind. This suggests that although the intended meaning can be inferred (as reflected by high meaning accuracy), the metaphor is not conveyed with sufficient clarity and immediacy through visual structure—models produce aesthetically pleasing images without powerful metaphorical expression. The gap between Harmony and Efficacy widens for weaker models: SDXL Base scores 3.41 on Harmony but only 2.24 on Efficacy (gap = 1.17), whereas GPT Image 1.5 maintains a gap of only 0.11 (4.38 vs. 4.27).

### 4.3 Qualitative Analysis

To complement the quantitative results above, Figure 4 presents representative outputs from seven models on four visual metaphors covering different structure types. Several patterns emerge that are consistent with our quantitative findings. Closed-source models such as GPT Image 1.5 and Nano Banana 2 generally produce coherent compositions that integrate the two domains in meaningful ways. In contrast, weaker open-source models often compose them incorrectly. Mid-tier models such as FLUX.2-dev and Z-Image fall in between, occasionally capturing the metaphorical concept but producing visually disharmonious compositions or failing on structurally demanding cases.

A closer look at typical failure cases reveals three recurring issues that align with our quantitative findings. The most common is *incorrect structural composition*: models depict both domains but fail to combine them in the structure implied by the prompt. For example, in row 4, SD 3.5 Large, Z-Image,

Table 3: Performance comparison between **conceptual** and **descriptive** prompts on representative models. "Conc." and "Desc." denote the conceptual and descriptive prompt settings, respectively. $\Delta$ denotes the change when switching from conceptual to descriptive prompts.

| | MCQ Overall (%) | | | Dimension Overall | | |
|---|---|---|---|---|---|---|
| Model | Conc. | Desc. | $\Delta$ | Conc. | Desc. | $\Delta$ |
| GPT Image 1.5 | 85.4 | 88.5 | +3.1 | 4.30 | 4.24 | −0.06 |
| Nano Banana 2 | 84.8 | 89.2 | +4.4 | 4.11 | 4.18 | +0.07 |
| FLUX.2-dev | 76.0 | 87.3 | +11.3 | 3.62 | 4.09 | +0.47 |
| Z-Image | 72.4 | 85.9 | +13.5 | 3.21 | 3.90 | +0.69 |
| OmniGen2 | 61.4 | 78.0 | +16.6 | 2.74 | 3.35 | +0.61 |

and Janus-Pro fuse the cigarette and handgun into a single object rather than produce the intended juxtaposition, echoing their weakness on L2 Structure. A second issue is *poor visual integration*, where the two domains are blended awkwardly, such as a coffee cup stacked on visibly distinct pillows without any unified form (row 2). Although visually polished, such compositions weaken the metaphor's expressive force, consistent with Metaphoric Efficacy lagging behind Perceptual Harmony. In more severe cases, models *omit one domain* entirely: Z-Image fills the claw machine with ordinary toy balls rather than business professionals (row 3), making the metaphor unrecoverable. Together, these failures demonstrate the diagnostic value of our multi-level evaluation framework.

### 4.4 Effect of Prompt Granularity

In addition to the conceptual prompts reported in our main experiments, we also generate and evaluate images using the descriptive prompts paired with each sample. Full results under descriptive prompts are provided in the Appendix A.5. Here we analyze the differences between the two prompt settings on a representative subset of five models spanning the three model categories, summarized in Table 3.

We first observe that all models benefit from descriptive prompts on MCQ accuracy, and most models also see improvements on dimension scores, indicating that explicit visual specification reduces the difficulty of metaphor generation. However, the magnitude of improvement varies dramatically across models. GPT Image 1.5 gains only +3.1 percentage points on MCQ overall and shows essentially no change on dimension score ($-0.06$), whereas Z-Image improves by +13.5 points and +0.69, respectively. A clear inverse relationship emerges between a model's baseline capability and the gain it receives: stronger models exhibit smaller gains, while weaker models benefit substantially more.

This pattern suggests that capable models such as GPT Image 1.5 and Nano Banana 2 already possess the ability to reason about *how* a metaphorical idea should be visually composed without explicit specification. In contrast, weaker models lack this inferential capacity and rely heavily on the explicit visual guidance provided by descriptive prompts. This aligns with our earlier observation that visual metaphor generation requires more than literal rendering: it demands a degree of conceptual reasoning, on which current open-source models still fall short.

Beyond the per-model gains, descriptive prompts also narrow the capability gap between models. Under conceptual prompts, the dimension score gap between GPT Image 1.5 (4.30) and Z-Image (3.21) is 1.09; under descriptive prompts, this gap narrows to 0.34 (4.24 vs. 3.90). For practitioners working with open-source models, providing fine-grained visual descriptions is therefore an effective strategy to compensate for the limited metaphorical reasoning capability.

### 4.5 Reliability of the Evaluator

To validate the reliability of using Qwen3.5-27B as our default evaluator, we conduct two studies: a cross-judge consistency study assessing robustness to the choice of evaluator, and a human alignment study measuring agreement with human judgment.

**Cross-Judge Consistency.** To test whether our conclusions depend on the specific choice of evaluator, we re-run the full evaluation pipeline with GPT-5.4 (OpenAI, 2026) as an alternative judge and compare model rankings and scores against those produced by Qwen3.5-27B. As shown in Table 4, the top-5 models receive identical rankings from both judges across both metrics. Spearman's $\rho$ reaches 0.993 for MCQ and 0.951 for dimension scores, confirming that our findings are not dependent on the choice of evaluator.

Table 4: Cross-judge model rankings by MCQ Overall and Dimension Overall.

| | MCQ | | Dimension | |
|---|---|---|---|---|
| Model | Qwen3.5 | GPT-5.4 | Qwen3.5 | GPT-5.4 |
| GPT Image 1.5 | 1 | 1 | 1 | 1 |
| Nano Banana 2 | 2 | 2 | 2 | 2 |
| Seedream 5.0 Lite | 3 | 3 | 3 | 3 |
| FLUX.2-dev | 4 | 4 | 4 | 4 |
| Qwen-Image | 5 | 5 | 5 | 5 |

Table 5: Alignment of Qwen3.5-27B and GPT-5.4 with human annotators on 200 sampled images: (a) MCQ balanced accuracy (%); (b) dimension-score mean absolute error (MAE) on a 1–5 scale.

(a) MCQ alignment (Bal.Acc, %)

| Judge | Domain | Structure | Meaning | Mapping | Overall |
|---|---|---|---|---|---|
| Qwen3.5-27B | 87.61 | 71.25 | 83.97 | 76.45 | 80.24 |
| GPT-5.4 | 88.79 | 73.89 | 84.50 | 76.51 | 81.16 |

(b) Dimension alignment (MAE)

| Judge | Efficacy | Logic | Harmony |
|---|---|---|---|
| Qwen3.5-27B | 0.65 | 0.71 | 0.64 |
| GPT-5.4 | 0.71 | 0.89 | 0.75 |

**Human Alignment Study.** We further evaluate how well the two MLLM judges align with human annotators. We randomly sample 100 images each from FLUX.2-dev and Qwen-Image, yielding 200 images in total, as in SpatialGenEval (Wang et al., 2026). Five human annotators, trained on visual metaphor theory and our annotation guidelines, independently evaluate each image using the same protocol applied to the MLLM evaluator. We measure alignment using balanced accuracy (Brodersen et al., 2010) for MCQ evaluation and Mean Absolute Error (MAE) for dimension-based scoring, following established practice in recent benchmark studies (Wang et al., 2026; Zhao et al., 2025). Results are reported in Table 5.

The MCQ alignment is consistently high across all four levels, with both Qwen3.5-27B and GPT-5.4 achieving over 80% balanced accuracy overall. Alignment is highest on the more concrete L1 (Domain) and L3 (Meaning) levels (above 83%), and lower on L2 (Structure) and L4 (Mapping), which require finer-grained judgments about how source and target are visually combined. For dimension-based scoring, Qwen3.5-27B yields a low MAE across all three dimensions (below 0.8 on a 1–5 scale), indicating that its ratings are close to those of human annotators. GPT-5.4 shows slightly larger errors (up to 0.89 on Logic) but remains within a reasonable range ($< 1.0$). The fact that an open-source evaluator achieves alignment comparable to or better than a leading proprietary model further supports the choice of Qwen3.5-27B as our default judge.

## 5 Limitations

We acknowledge several limitations. First, the distribution of samples across thematic categories and structure types is uneven, reflecting our curated real-world corpus rather than controlled balancing. Second, our evaluation relies on an MLLM-as-judge framework, which, despite strong alignment with human annotators, inherits the fundamental limitations of automatic visual reasoning on the most subjective judgments. Finally, our prompts and source imagery are predominantly English and from Western creative traditions, leaving multilingual and cross-cultural metaphors as future directions.

## 6 Conclusion

In this paper, we introduce VMetaphor-Bench, a fine-grained diagnostic benchmark tailored to visual metaphor generation in text-to-image models. Built upon 1,500 carefully curated visual metaphors with structured annotations, our benchmark provides a hierarchical taxonomy of three levels and ten categories and a hybrid evaluation framework that combines MCQ-based fidelity assessment with dimension-based scoring. Extensive evaluation of 12 representative T2I models reveals that proprietary models substantially outperform open-source counterparts, yet even the strongest models continue to struggle with compositional structuring and cross-domain mapping—key aspects of metaphorical expression. We hope VMetaphor-Bench serves as a foundation for future research on bridging the gap between literal rendering and creative, meaning-rich visual generation.

### AI USE STATEMENT

Generative AI tools were used in the benchmark pipeline for initial annotation and prompt generation and as automated evaluators, as described in the paper. They also assisted with language editing and LaTeX formatting. All AI-assisted outputs were reviewed and verified by the authors, who take full responsibility for the paper's content, analyses, and claims.

### ETHICS STATEMENT

VMetaphor-Bench is intended for academic research. To respect original content creators, we do not redistribute the Pinterest source images and instead release their URLs and derived annotations. We acknowledge the dataset's limited cultural coverage and the potential misuse of visual metaphors for persuasive or misleading content; these issues are discussed further in the Limitations and Appendix.

### REPRODUCIBILITY STATEMENT

The main paper describes the benchmark construction, evaluation protocols, and experimental setup. The Appendix provides dataset statistics, model versions and checkpoints, complete evaluation results, prompts, human-annotation interfaces, and robustness analyses. The dataset, evaluation code, and experimental configurations are available in our anonymous repository at `https://anonymous.4open.science/r/VMetaphor-Bench-D728`.

## REFERENCES

Pinterest. `https://www.pinterest.com`.

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

Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M Buhmann. The balanced accuracy and its posterior distribution. In *2010 20th international conference on pattern recognition*, pp. 3121–3124. IEEE, 2010.

ByteDance Seed. Seedream 5.0 lite. `https://seed.bytedance.com/en/seedream5_0_lite`, 2026.

Silvia Cappa, Anna Sofia Lippolis, and Stefano Zoia. Meanings are like onions: A layered approach to metaphor processing. *arXiv preprint arXiv:2507.10354*, 2025.

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

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*, 2024.

Kaijie Chen, Zihao Lin, Zhiyang Xu, Ying Shen, Yuguang Yao, Joy Rimchala, Jiaxin Zhang, and Lifu Huang. R2i-bench: Benchmarking reasoning-driven text-to-image generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 12606–12641, 2025a.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025b.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Francesca Ervas. Metaphor, ignorance and the sentiment of (ir) rationality. *Synthese*, 198(7): 6789–6813, 2021.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Gilles Fauconnier and Mark Turner. Conceptual blending, form and meaning. *Recherches en communication*, 19:57–86, 2003.

Charles Forceville et al. Metaphor in pictures and multimodal representations. *The Cambridge handbook of metaphor and thought*, pp. 462–482, 2008.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

Google. Gemini 3.1 Flash Image Preview. `https://ai.google.dev/gemini-api/docs/models/gemini-3.1-flash-image-preview`, 2026.

Google DeepMind. Gemini 3.1 pro. `https://deepmind.google/models/gemini/pro/`, 2026.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7514–7528, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20406–20417, 2023.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

Qingbao Huang, Cheng Yang, Jiawei Yao, Zhiyue Liu, Yi Cai, and Xingmao Zhang. CMIG: Conceptual metaphor theory-inspired framework for metaphorical image generation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 23748–23761. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1189. URL `https://aclanthology.org/2026.findings-acl.1189/`.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Amita Kamath, Kai-Wei Chang, Ranjay Krishna, Luke Zettlemoyer, Yushi Hu, and Marjan Ghazvininejad. Geneval 2: Addressing benchmark drift in text-to-image evaluation. *arXiv preprint arXiv:2512.16853*, 2025.

Girish A Koushik, Fatemeh Nazarieh, Katherine Birch, Shenbin Qian, and Diptesh Kanojia. The mind's eye: A multi-faceted reward framework for guiding visual metaphor generation. *arXiv preprint arXiv:2508.18569*, 2025.

Manishit Kundu, Sumit Shekhar, and Pushpak Bhattacharyya. Looking beyond the pixels: Evaluating visual metaphor understanding in vlms. *Findings of the Association for Computational Linguistics: EMNLP*, pp. 23137–23158, 2025.

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

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

Kostoula Margariti, Leonidas Hatzithomas, Christina Boutsouki, and Yorgos Zotos. A path to our heart: Visual metaphors and "white" space in advertising aesthetic pleasure. *International Journal of Advertising*, 41(4):731–770, 2022.

Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Chaoran Feng, Kunpeng Ning, Bin Zhu, et al. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv preprint arXiv:2503.07265*, 2025.

OpenAI. GPT-Image-1.5. https://developers.openai.com/api/docs/models/gpt-image-1.5, 2025.

OpenAI. Introducing gpt-5.4. https://openai.com/zh-Hans-CN/index/introducing-gpt-5-4/, 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

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

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

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

Stability AI. Stable diffusion 3.5 large. https://huggingface.co/stabilityai/stable-diffusion-3.5-large, 2024.

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

Kaiyue Sun, Rongyao Fang, Chengqi Duan, Xian Liu, and Xihui Liu. T2i-reasonbench: Benchmarking reasoning-informed text-to-image generation. *arXiv preprint arXiv:2508.17472*, 2025a.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

Zhida Sun, Zhenyao Zhang, Yue Zhang, Min Lu, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Creative blends of visual concepts. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–17, 2025b.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Z-Image Team. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024.

Margot Van Mulken, Andreu Van Hooft, and Ulrike Nederstigt. Finding the tipping point: Visual metaphor and conceptual complexity in advertising. *Journal of Advertising*, 43(4):333–343, 2014.

Zehan Wang, Jiayang Xu, Ziang Zhang, Tianyu Pang, Chao Du, Hengshuang Zhao, and Zhou Zhao. Genspace: Benchmarking spatially-aware image generation. *arXiv preprint arXiv:2505.24870*, 2025.

Zengbin Wang, Xuecai Hu, Yong Wang, Feng Xiong, Man Zhang, and Xiangxiang Chu. Everything in its place: Benchmarking spatial intelligence of text-to-image models. *arXiv preprint arXiv:2601.20354*, 2026.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025a. URL https://arxiv.org/abs/2508.02324.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025b.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*, 2025.

Yu Xu, Yuxin Zhang, Juan Cao, Lin Gao, Chunyu Wang, Oliver Deussen, Tong-Yee Lee, and Fan Tang. Beyond pixels: Visual metaphor transfer via schema-driven agentic reasoning. *arXiv preprint arXiv:2602.01335*, 2026.

Junyan Ye, Dongzhi Jiang, Zihao Wang, Leqi Zhu, Zhenghao Hu, Zilong Huang, Jun He, Zhiyuan Yan, Jinghua Yu, Hongsheng Li, et al. Echo-4o: Harnessing the power of gpt-4o synthetic images for improved image generation. *arXiv preprint arXiv:2508.09987*, 2025.

Chenhao Zhang, Yazhe Niu, and Hongsheng Li. Metaphorstar: Image metaphor understanding and reasoning with end-to-end visual reinforcement learning. *arXiv preprint arXiv:2602.10575*, 2026.

Daoan Zhang, Che Jiang, Ruoshi Xu, Biaoxiang Chen, Zijian Jin, Yutian Lu, Jianguo Zhang, Liang Yong, Jiebo Luo, and Shengda Luo. Worldgenbench: A world-knowledge-integrated benchmark for reasoning-driven text-to-image generation. *arXiv preprint arXiv:2505.01490*, 2025a.

Zicheng Zhang, Junying Wang, Farong Wen, Yijin Guo, Xiangyu Zhao, Xinyu Fang, Shengyuan Ding, Ziheng Jia, Jiahao Xiao, Ye Shen, et al. Large multimodal models evaluation: a survey. *Science China Information Sciences*, 68(12):221301, 2025b.

Xiangyu Zhao, Peiyuan Zhang, Kexian Tang, Xiaorong Zhu, Hao Li, Wenhao Chai, Zicheng Zhang, Renqiu Xia, Guangtao Zhai, Junchi Yan, et al. Envisioning beyond the pixels: Benchmarking reasoning-informed visual editing. *arXiv preprint arXiv:2504.02826*, 2025.

# A Appendix

## A.1 Benchmark Statistics

Table 6 summarizes the dataset statistics. Conceptual prompts are concise (mean 25.6 words), while descriptive prompts provide richer detail (mean 65.0 words); appending the style suffix roughly

doubles the length in both cases. We organize the 1,500 images into 10 thematic categories across three levels: *Individual* (495 images) covers topics related to personal cognition, emotion, health, and growth; *External* (499 images) encompasses societal, environmental, technological, and interpersonal themes; and *Creative* (506 images) includes artistic designs and commercial advertisements. The distribution is approximately balanced across levels, with the largest single category being Creative Design (21.07%).

Table 6: Left: Prompt length statistics (word count). Right: Distribution of thematic categories.

| | Mean | Min | Max |
|---|---|---|---|
| Conceptual Prompt | 25.6 | 10 | 56 |
| + Style Suffix | 68.4 | 29 | 133 |
| Descriptive Prompt | 65.0 | 20 | 142 |
| + Style Suffix | 107.7 | 38 | 226 |

| Level | Category | Count | Prop.(%) |
|---|---|---|---|
| Individual (495) | Cognitive Processes | 198 | 13.20 |
| | Mental States | 161 | 10.73 |
| | Healthy Living | 72 | 4.80 |
| | Personal Development | 64 | 4.27 |
| External (499) | Social Issues | 243 | 16.20 |
| | Natural Environment | 119 | 7.93 |
| | Technological Life | 88 | 5.87 |
| | Social Interaction | 49 | 3.27 |
| Creative (506) | Creative Design | 316 | 21.07 |
| | Advertising | 190 | 12.67 |

### A.2 Dataset Coverage Audit

**Structure distribution.** We did not impose structure-specific quotas; labels were assigned after collecting and filtering the images. The final dataset contains 1,199 fusion images (79.9%), 231 replacement images (15.4%), and 70 juxtaposition images (4.7%). These proportions reflect the observed distribution in our curated real-world corpus rather than quota-based sampling. The fusion-dominant pattern is also consistent with prior visual-metaphor research in international advertising (Margariti et al., 2022). Giving the three structure types equal weight yields similar rankings across the 12 analyzed models (Spearman's $\rho = 0.979$ for MCQ and $0.979$ for dimension scores). L2/L4 remain below L1/L3 under this weighting. These results suggest that the main comparative and diagnostic findings are not solely driven by the predominance of fusion samples.

**Cultural coverage.** We conducted a preliminary manual audit of 100 samples, drawing ten from each thematic category. A sample was considered culture-dependent when interpreting its intended metaphor required region- or language-specific knowledge. Two samples met this criterion: one used the English idiom "red tape" and one used the South Asian concept of "karma." The remaining 98 showed no explicit dependence on such knowledge. This audit suggests that explicit culture-specific knowledge is rarely required in the audited samples. Nevertheless, Section 5 acknowledges the benchmark's limited cross-cultural coverage and identifies broader cultural coverage as future work.

### A.3 Experimental Setup Details

We conduct all experiments on a single server equipped with 4 $\times$ NVIDIA H200 GPUs. Detailed information about each evaluated model and MLLM evaluators is provided in Table 7.

Table 7: Detailed information about the evaluated T2I models and MLLM evaluators.

| Model | Year | Resolution | Source | Version / HF Checkpoint |
|---|---|---|---|---|
| *Closed-source T2I Models* | | | | |
| Seedream 5.0 Lite | 2026 | 2K | API | `doubao-seedream-5-0-260128` |
| Nano Banana 2 | 2026 | 1024 × 1024 | API | `gemini-3.1-flash-image-preview` |
| GPT Image 1.5 | 2025 | 1024 × 1024 | API | `gpt-image-1.5-2025-12-16` |
| *Open-source T2I Models* | | | | |
| SDXL Base | 2023 | 1024 × 1024 | Checkpoint | `stabilityai/stable-diffusion-xl-base-1.0` |
| FLUX.1-dev | 2024 | 1024 × 1024 | Checkpoint | `black-forest-labs/FLUX.1-dev` |
| SD 3.5 Large | 2024 | 1024 × 1024 | Checkpoint | `stabilityai/stable-diffusion-3.5-large` |
| Z-Image | 2025 | 1024 × 1024 | Checkpoint | `Tongyi-MAI/Z-Image-Turbo` |
| Qwen-Image | 2025 | 1328 × 1328 | Checkpoint | `Qwen/Qwen-Image-2512` |
| FLUX.2-dev | 2025 | 1024 × 1024 | Checkpoint | `black-forest-labs/FLUX.2-dev` |
| *Open-source Unified MLLMs* | | | | |
| OmniGen2 | 2025 | 1024 × 1024 | Checkpoint | `OmniGen2/OmniGen2` |
| Bagel | 2025 | 1024 × 1024 | Checkpoint | `ByteDance-Seed/BAGEL-7B-MoT` |
| Janus-Pro | 2025 | 384 × 384 | Checkpoint | `deepseek-ai/Janus-Pro-7B` |
| *MLLM Evaluators* | | | | |
| Qwen3.5-27B | 2026 | – | Checkpoint | `Qwen/Qwen3.5-27B` |
| GPT-5.4 | 2026 | – | API | `gpt-5.4-2026-03-05` |

### A.4 Per-Structure Analysis

To better understand where models struggle on the L2 Structure level, we examine the structure type predicted by the MCQ evaluation against the ground-truth structure type annotated in our dataset. Each L2 MCQ question asks the judge to identify the structure type of the generated image among *fusion*, *replacement*, *juxtaposition*, and *cannot be determined*, allowing us to construct a confusion matrix between ground-truth and predicted structure types. Aggregated results across the 12 models included in this analysis are shown in Figure 5.

The confusion matrix reveals a striking asymmetry across structure types. Models achieve reasonable accuracy on *fusion* (68.7%) and *juxtaposition* (70.7%), where the two domains are either merged into a single object or placed alongside each other. In contrast, *replacement* accuracy drops dramatically to only 19.3%. When prompted with replacement metaphors, models tend to either fuse the two domains together (41.0%) or place them as separate objects (31.6%), rather than substituting one for the other.

Interestingly, this difficulty hierarchy mirrors the cognitive complexity ordering proposed in visual rhetoric theory (Phillips & McQuarrie, 2004), where replacement is identified as the most demanding structure type for human viewers because the missing element must be inferred from its absence. In contrast, fusion requires disentangling two co-present elements, while juxtaposition simply pairs them side by side. Our results suggest that this complexity ordering also applies to visual metaphor generation. This suggests that the cognitive demands of replacement, which require preserving context while substituting a localized element, pose a fundamental challenge for current models. This finding points to compositional precision, the ability to perform localized, structure-aware modifications, as a key direction for advancing visual metaphor generation.

### A.5 Full Results with Descriptive Prompts

Table 8 reports the full evaluation results under descriptive prompts, complementing the conceptual prompt results in Table 2. Most models show substantial improvements under descriptive prompts, as the detailed visual descriptions provide more explicit guidance for generation. GPT Image 1.5's dimension score remains essentially unchanged ($-0.06$), while SDXL Base shows no overall improvement, indicating that descriptive guidance remains model-dependent. Despite these magnitude differences, the overall model ranking remains largely consistent across the two prompt types.

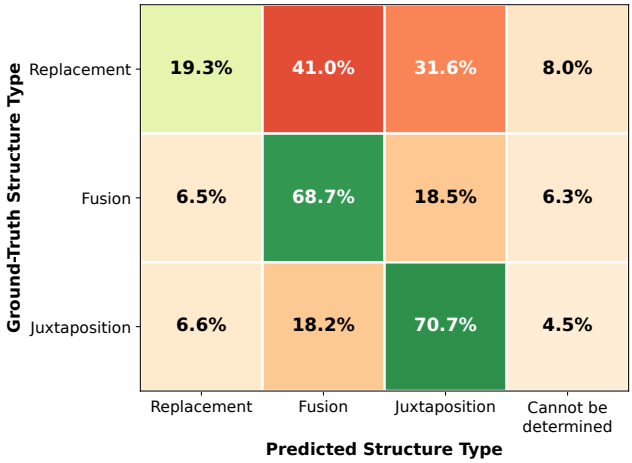


Figure 5: Structure confusion matrix across 12 models (Qwen3.5-27B, conceptual prompts).

Table 8: Evaluation results on VMetaphor-Bench with **descriptive prompts**.

| | MCQ Accuracy (%) | | | | | Dimension Score (1–5) | | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | Domain | Structure | Meaning | Mapping | Overall | Efficacy | Logic | Harmony | Overall |
| *Closed-source T2I Models* | | | | | | | | | |
| Seedream 5.0 Lite | 89.3 | 69.2 | 96.5 | 87.4 | 86.6 | 4.02 | 4.14 | 4.27 | 4.15 |
| GPT Image 1.5 | 89.6 | 72.9 | 96.6 | 90.6 | 88.5 | **4.12** | **4.24** | **4.34** | **4.24** |
| Nano Banana 2 | **90.8** | **73.6** | **97.1** | **91.2** | **89.2** | 4.06 | 4.21 | 4.28 | 4.18 |
| *Open-source T2I Models* | | | | | | | | | |
| SDXL Base | 68.3 | 51.0 | 79.0 | 55.0 | 62.3 | 2.16 | 2.34 | 3.27 | 2.59 |
| SD 3.5 Large | 83.0 | 62.3 | 93.4 | 76.0 | 78.8 | 3.36 | 3.45 | 3.82 | 3.54 |
| FLUX.1-dev | 83.8 | 64.1 | 93.5 | 79.6 | 80.7 | 3.45 | 3.55 | 3.97 | 3.66 |
| Z-Image | 87.9 | 67.7 | 96.4 | 87.5 | 85.9 | 3.76 | 3.90 | 4.04 | 3.90 |
| Qwen-Image | 88.6 | 70.9 | 95.9 | 88.5 | 87.0 | 3.98 | 4.03 | 4.08 | 4.03 |
| FLUX.2-dev | 89.4 | 70.3 | 96.4 | 88.9 | 87.3 | 4.00 | 4.09 | 4.18 | 4.09 |
| *Open-source Unified MLLMs* | | | | | | | | | |
| OmniGen2 | 82.6 | 59.9 | 93.2 | 75.3 | 78.0 | 3.07 | 3.26 | 3.71 | 3.35 |
| Bagel | 83.2 | 62.7 | 92.6 | 76.8 | 79.1 | 3.28 | 3.36 | 3.77 | 3.47 |
| Janus-Pro | 84.0 | 65.9 | 94.9 | 79.1 | 81.0 | 3.31 | 3.52 | 3.62 | 3.48 |

### A.6 Cross-Judge Results with GPT-5.4

Table 9 reports the evaluation results using GPT-5.4 as the judge, complementing the Qwen3.5-27B results in Table 2. GPT-5.4 produces highly consistent rankings with Qwen3.5-27B, with Spearman rank correlations of 0.993 for MCQ and 0.951 for dimension scoring. The two judges differ slightly in absolute scores, with GPT-5.4 tending to assign somewhat more lenient dimension scores. However, the relative performance ordering of models is preserved, confirming that the conclusions reported in our main experiments do not depend on the specific choice of evaluator.

Table 9: Evaluation results on VMetaphor-Bench with **conceptual prompts**, judged by **GPT-5.4**.

| | MCQ Accuracy (%) | | | | | Dimension Score (1–5) | | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | Domain | Structure | Meaning | Mapping | Overall | Efficacy | Logic | Harmony | Overall |
| *Closed-source T2I Models* | | | | | | | | | |
| Seedream 5.0 Lite | 93.3 | 66.7 | 96.3 | 73.0 | 82.0 | 3.72 | 3.85 | 3.94 | 3.84 |
| Nano Banana 2 | 94.3 | 75.6 | **98.3** | 76.9 | 85.5 | 4.08 | 4.29 | 4.14 | 4.17 |
| GPT Image 1.5 | **94.4** | **77.2** | 97.6 | **77.2** | **85.8** | **4.13** | **4.37** | **4.33** | **4.27** |
| *Open-source T2I Models* | | | | | | | | | |
| SDXL Base | 90.4 | 59.8 | 90.2 | 61.6 | 74.8 | 2.44 | 2.63 | 3.32 | 2.80 |
| FLUX.1-dev | 91.3 | 63.8 | 91.9 | 62.8 | 76.4 | 2.95 | 3.07 | 3.70 | 3.24 |
| SD 3.5 Large | 91.5 | 65.7 | 91.4 | 64.6 | 77.4 | 2.97 | 3.14 | 3.68 | 3.26 |
| Z-Image | 93.1 | 66.5 | 94.8 | 65.6 | 78.9 | 3.05 | 3.25 | 3.56 | 3.28 |
| Qwen-Image | 91.9 | 70.3 | 93.9 | 67.3 | 79.6 | 3.32 | 3.45 | 3.73 | 3.50 |
| FLUX.2-dev | 94.0 | 69.3 | 94.7 | 69.0 | 80.9 | 3.45 | 3.62 | 3.84 | 3.64 |
| *Open-source Unified MLLMs* | | | | | | | | | |
| OmniGen2 | 90.2 | 62.0 | 90.9 | 57.6 | 73.7 | 2.51 | 2.65 | 3.42 | 2.86 |
| Bagel | 91.7 | 64.7 | 92.6 | 62.1 | 76.6 | 2.78 | 2.92 | 3.44 | 3.05 |
| Janus-Pro | 92.3 | 67.1 | 93.9 | 67.6 | 79.4 | 3.06 | 3.34 | 3.49 | 3.30 |

### A.7 Expanded Human–Evaluator Alignment

To examine the reliability of our MLLM-as-judge protocol across diverse T2I outputs, we evaluate 180 images from six models, with two representatives from each model category. We sample 30 images per model, cover all ten thematic categories, and balance the sample across the three structure types. Five trained annotators independently evaluate every image using the same MCQ and dimension protocols as the MLLM judges. As shown in Table 10, overall Qwen/GPT alignment reaches 81.68%/83.30% MCQ balanced accuracy and 0.705/0.703 dimension-score MAE. Alignment remains broadly consistent across model categories, performance levels, and structure types, supporting the reliability of our MLLM-as-judge protocol across diverse T2I outputs.

Table 10: Expanded human alignment across six T2I models and three structure types. MCQ alignment is balanced accuracy (BA, %) against the human majority; dimension alignment is MAE against mean human ratings.

| Stratum | Qwen MCQ BA ↑ | GPT MCQ BA ↑ | Qwen MAE ↓ | GPT MAE ↓ |
|---|---|---|---|---|
| **Overall** | **81.68** | **83.30** | **0.705** | **0.703** |
| *By T2I model* | | | | |
| GPT Image 1.5 | 85.85 | 83.74 | 0.747 | 0.636 |
| Nano Banana 2 | 84.95 | 89.98 | 0.620 | 0.640 |
| FLUX.2-dev | 82.11 | 83.25 | 0.640 | 0.502 |
| Qwen-Image | 80.70 | 78.85 | 0.680 | 0.893 |
| BAGEL | 77.59 | 82.34 | 0.762 | 0.733 |
| Janus-Pro | 78.68 | 81.41 | 0.780 | 0.816 |
| *By structure type* | | | | |
| Fusion | 81.36 | 80.67 | 0.763 | 0.774 |
| Replacement | 82.56 | 83.26 | 0.679 | 0.713 |
| Juxtaposition | 80.85 | 85.47 | 0.672 | 0.622 |

### A.8 Multi-Generation Sensitivity

The primary benchmark uses one image per prompt to keep generation and evaluation costs tractable. We test sensitivity to this choice using 100 conceptual prompts stratified across ten categories and three structure types. Seven models spanning all three model categories generate one image

per prompt with each of three fixed seeds (42, 428, and 2026), yielding 2,100 evaluated images. Across seeds, standard deviations are 0.7–2.0 percentage points for MCQ Overall and 0.02–0.06 for Dimension Overall (Table 11). Mean pairwise Spearman correlations between seed-specific model rankings are 0.886 and 0.857, respectively, indicating broadly stable aggregate comparisons on this stratified subset.

Table 11: Multi-generation robustness on 100 conceptual prompts. Values are mean ± standard deviation across three fixed seeds (42, 428, and 2026).

| Model | MCQ Overall (%) | Dimension Overall |
|---|---|---|
| GPT Image 1.5 | $84.7 \pm 0.7$ | $4.13 \pm 0.02$ |
| Nano Banana 2 | $83.4 \pm 1.0$ | $4.00 \pm 0.04$ |
| Z-Image | $73.6 \pm 1.0$ | $3.24 \pm 0.03$ |
| Janus-Pro | $73.4 \pm 0.7$ | $3.12 \pm 0.02$ |
| SD 3.5 Large | $72.2 \pm 1.4$ | $3.14 \pm 0.02$ |
| FLUX.1-dev | $69.5 \pm 1.3$ | $3.18 \pm 0.05$ |
| BAGEL | $67.6 \pm 2.0$ | $2.95 \pm 0.06$ |

### A.9 Benchmark Construction Prompts.

Figure 6 and Figure 7 show the prompts for the two-stage annotation pipeline: structured annotation extraction (Stage 1) and descriptive/conceptual prompt generation (Stage 2). Figure 8 shows the prompt for MCQ distractor generation.

**Stage 1: Prompt for Extracting Structured Annotations**

You are analyzing a single image that may contain a visual metaphor.

First, check the following conditions. If ANY of them apply, return {{"has_metaphor": false}}:
[conditions]

If it DOES contain a visual metaphor, decompose it into structured components. Only extract structural and semantic information grounded in the image.

Please identify:
1. Target domain: What is the primary object being reconstructed?
2. Source domain: What external domain is visually applied to the target?
3. Structure type:
- Choose one: (a) Fusion: [explainations] (b) Replacement: [explainations] (c) Juxtaposition: [explainations]

4. Element-level mappings
Identify visually supported correspondences between elements of the source and target. Format as pairs: source_element→ target_element

5. Meaning summary
Provide a 1–2 sentence summary of the overall metaphorical meaning. Focus on abstract interpretation, not visual description.

Output in the following JSON format:

```
{{
 "has_metaphor": true,
 "target_domain": "",
 "source_domain": "",
 "structure_type": "",
 "mappings": [
  {{"source_element": "", "target_element": ""}},
  ...
 ],
 "meaning": ""
}}
```

Figure 6: Prompt for structured annotation generation (Stage 1).

**Stage 2: Prompt for Generating T2I Prompts**

You are a vision-language expert. Your task is to write high-quality generation prompts for the provided image, which contains a visual metaphor.

**### Reference Annotations:**
{structured_annotations}

**### Instructions:**
You must analyze the ATTACHED IMAGE as your primary reference to ensure visual fidelity. Use the Annotations only to anchor the metaphorical logic.

You will produce three prompts. IMPORTANT: The first two prompts must NOT contain any style-related information (artistic style, color palette, lighting, mood, texture). All style information goes exclusively into style_suffix.

**### 1. Descriptive Prompt (Visual Reconstruction):**
Write a text-to-image generation prompt that describes the visual content needed to recreate the image.
[requirements]

**### 2. Conceptual Prompt (Abstract Level):**
Write a concise, natural-sounding generation prompt at the conceptual level.
[requirements]

**### 3. Style Suffix:**
Describe ONLY the artistic style and visual aesthetics of the image.
[requirements]

Return strictly valid JSON without markdown formatting or code fences.
### Output Format (Strict JSON):
{{
  "descriptive_prompt": "...",
  "conceptual_prompt": "...",
  "style_suffix": "..."
}}

Figure 7: Prompt for descriptive and conceptual prompt generation (Stage 2).

**Prompt for Generating MCQs**

You are an expert in visual metaphor analysis. Given the following annotation and descriptive prompt for a visual metaphor image, generate plausible but INCORRECT distractors for multiple-choice questions.

**## Annotation**
[structured_annotations]

**## Descriptive Prompt (describes the visual content of the image)**
[descriptive_prompt]

**## Distractor Rules**
[rules]

**## Categories**
Q1 (target_domain): "What is the target domain (tenor) of this visual metaphor?"
- Correct answer: "{target}"
- Generate 2 distractors following the rules above.

Q2 (source_domain): "What is the source domain (vehicle) of this visual metaphor?"
- Correct answer: "{source}"
- Generate 2 distractors following the rules above.

Q4 (meaning): "What meaning does this image most likely convey?"
- Correct answer: "{meaning}"
- Generate 2 plausible but wrong interpretations. They must differ in substance from the correct meaning, not just in phrasing.

Q5 (element_mapping): For EACH mapping below, generate 2 wrong target elements.
{mappings_text}
- For each mapping, the question asks: "What does [source_element] visually correspond to in this image?"
- Correct answer is the target_element.
- Generate 2 plausible wrong target elements (NOT synonyms/hypernyms/hyponyms of the correct target_element).

**## Output Format**
Return strictly valid JSON without markdown formatting or code fences:

```
{{
  "Q1_distractors": ["distractor1", "distractor2"],
  "Q2_distractors": ["distractor1", "distractor2"],
   ......
}}
```

Figure 8: Prompt for MCQ distractor generation.

### A.10 EVALUATION PROMPTS

Figure 9 and Figure 10 show the prompts for MCQ evaluation, which is split into two calls: Q1–Q4 (target, source, structure, and meaning) are answered together, while Q5 (element mapping) is answered separately to prevent domain leakage from its answer options. Specifically, Q5 (element mapping) options may reveal the source domain (e.g., a question stem mentioning 'spherical bomb casing' would expose 'bomb' as the source), which would compromise the independence of L1 and L4 evaluations. Figure 11 shows the prompt for dimension scoring with 1–5 Likert-scale rubrics.

**Prompt for MCQ-Based Evaluation (Q1-Q4)**

You are an expert in visual metaphor analysis.

**## Key Concepts**
A visual metaphor compares two concepts by visually mapping one onto another, implying a shared property or relationship between them.

- Target domain: The subject being described — what the image is ultimately "about". It receives a property from the source domain. For example, in "french fries are deadly", french fries are the target domain.

- Source domain: The concept used to describe or characterize the target — it lends a property to the target. For example, in "french fries are deadly", bullets serve as the source domain, lending the property "deadly" to french fries.

Note: Either domain may appear literally or be implied in the image. There is no fixed rule about which one is visually depicted.

- Structure type: How the two domains are visually presented in the image. There are three strategies:

  - Replacement: Only one domain's element is directly visible. The other domain is ABSENT and must be inferred from the surrounding visual context. For example, a light bulb's position is occupied by a tree — the tree is visible, but the light bulb is absent and inferred from the lamp context.

  - Fusion: Two normally distinct objects are physically merged into a single gestalt. Both domains are simultaneously present in ONE object, combined in a way that is physically impossible. For example, a globe with an ice-cream surface — one single object exhibiting features of both earth and ice cream.

  - Juxtaposition: Both domains are depicted as separate, complete objects within the same scene. Their metaphorical connection is suggested through visual similarity in form, position, or spatial arrangement. For example, a car placed alongside a cheetah to suggest speed.

- Meaning: The message the image conveys through the comparison of the two domains. For example, an image blending earth and ice cream might convey "the earth is melting".

**## Task**
Look at the image carefully and answer each of the following multiple-choice questions. Base your judgment ONLY on what is visually present in the image.

For EACH question, respond with ONLY the question ID and the letter answer. Use this exact format (one per line):
Q1: <letter>
Q2: <letter>
Q3: <letter>
Q4: <letter>

Do not explain. Do not add any other text.

{Q1-Q4 questions}

Figure 9: Prompt for MCQ evaluation (Q1–Q4: target domain, source domain, structure type, and meaning).

**Prompt for MCQ-Based Evaluation (Q5)**

You are an expert in visual metaphor analysis.

**## Key Concepts**
A visual metaphor communicates an abstract idea by visually mapping one concept onto another.

- Element mapping: The specific visual correspondence between a element from the source domain and a concept or element from the target domain. For
example, if a rose represents love, the "thorns" (source element) might visually correspond to "emotional pain" (target element).

**## Context**
- Target domain: {target_domain}
- Source domain: {source_domain}

**## Task**
Given the target and source domains above, look at the image carefully and answer each of the following multiple-choice questions. Base your judgment ONLY on what is visually present in the image.

For EACH question, respond with ONLY the question ID and the letter answer. Use this exact format (one per line):
Q5_1: <letter>
Q5_2: <letter>

Do not explain. Do not add any other text.

{Q5 questions}

Figure 10: Prompt for MCQ evaluation (Q5: Element mapping).

**Prompt for Dimension-based Scoring**

You are an expert evaluator of visual metaphor images.

**## Annotation (Ground Truth)**
This image is intended to convey a visual metaphor with the following properties:
- Target domain: {target}
- Source domain: {source}
- Structure type: {structure}
- Element mappings: {mappings_text}
- Intended meaning: {meaning}

**## Task**
Important: The annotation above describes what the image is intended to express — it is a reference, NOT what you are evaluating. Your job is to assess how well the IMAGE ITSELF visually realizes these intentions. A well-written annotation does NOT imply the image is good. If the metaphor described in the annotation is NOT visually perceivable in the image, you must score low regardless of how coherent the annotation sounds.

Evaluate the image on the following three dimensions independently, each on a 1-5 scale.

**### 1. Metaphoric Efficacy (1-5)**
Can a viewer understand the intended metaphorical meaning from the image alone, WITHOUT reading any textual explanation?
- 5: Any viewer, regardless of background, would immediately and effortlessly understand the metaphor; the meaning is unmistakable and powerfully conveyed.
- 4: A viewer can understand the metaphor with only brief thought; the core meaning is clear and the visual expression shows creativity.
- 3: A viewer can tell the image is attempting a metaphor and may grasp the general direction, but would need the annotation to fully understand the intended meaning.
- 2: The metaphorical intent is very difficult to interpret from the image alone; only vague or ambiguous hints are present.
- 1: The metaphor is completely unrecognizable; a viewer would have no idea what is being communicated.

**### 2. Metaphor Logic (1-5)**
Is the cross-domain mapping between source and target conceptually coherent and logically sound as depicted in the image?
- 5: The mapping is exceptionally well-constructed; the cross-domain connection feels natural, insightful, and almost inevitable — it is hard to imagine a better visual mapping.
- 4: The mapping is coherent and convincing; the connection is logical, though one could imagine equally good alternative mappings.
- 3: The mapping is understandable but requires some effort to accept; there are noticeable gaps or stretches in the cross-domain logic.
- 2: The mapping is weak; the connection between visual elements and the intended metaphor feels forced or highly unintuitive.
- 1: The mapping makes no sense; there is no logical connection between the visual elements and the intended metaphor.

**### 3. Perceptual Harmony (1-5)**
As a whole, does this image work as a visual metaphor — does it look natural, coherent, and easy to appreciate?
Consider: Do the combined elements form a coherent whole, or does the image feel strange, confusing, or visually uncomfortable despite good rendering quality?
- 5: The image works perfectly as a whole; the metaphorical fusion feels natural, visually pleasing, and immediately appreciable.
- 4: The fusion is seamless with only very minor imperfections; the image feels cohesive and natural at first glance.
- 3: The image mostly works; the fusion is coherent and the metaphor is appreciable, but with some noticeable awkwardness.
- 2: The image is technically well-rendered, but the metaphorical fusion feels unnatural, confusing, or uncomfortable to view.
- 1: The image feels visually incoherent or grotesque; the combination of elements is jarring and incomprehensible.

**## Output Format**
Respond with ONLY the three scores in this exact format (one per line):
metaphoric_efficacy: <score>
metaphor_logic: <score>
perceptual_harmony: <score>

Do not explain. Do not add any other text.

Figure 11: Prompt for dimension score evaluation.

### A.11 Interface for Human Annotators

We provide screenshots of the annotation interfaces used in our study. Figure 12 shows the interface for manual verification during data curation (Section 3.1), where annotators review and edit metadata fields such as domains, structure type, element mappings, and meaning. Figures 13 and 14 show the two pages of the interface used in the human alignment study (Section 4.5): one for answering MCQ questions alongside the generated image, and another for rating dimension scores on a 1–5 scale.

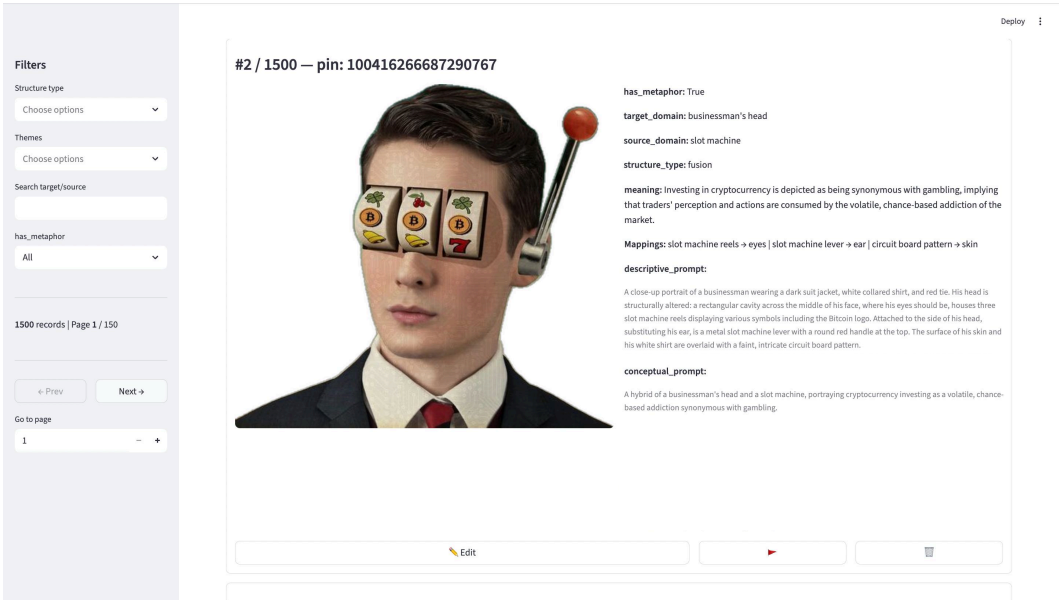


Figure 12: Manual verification interface for data curation

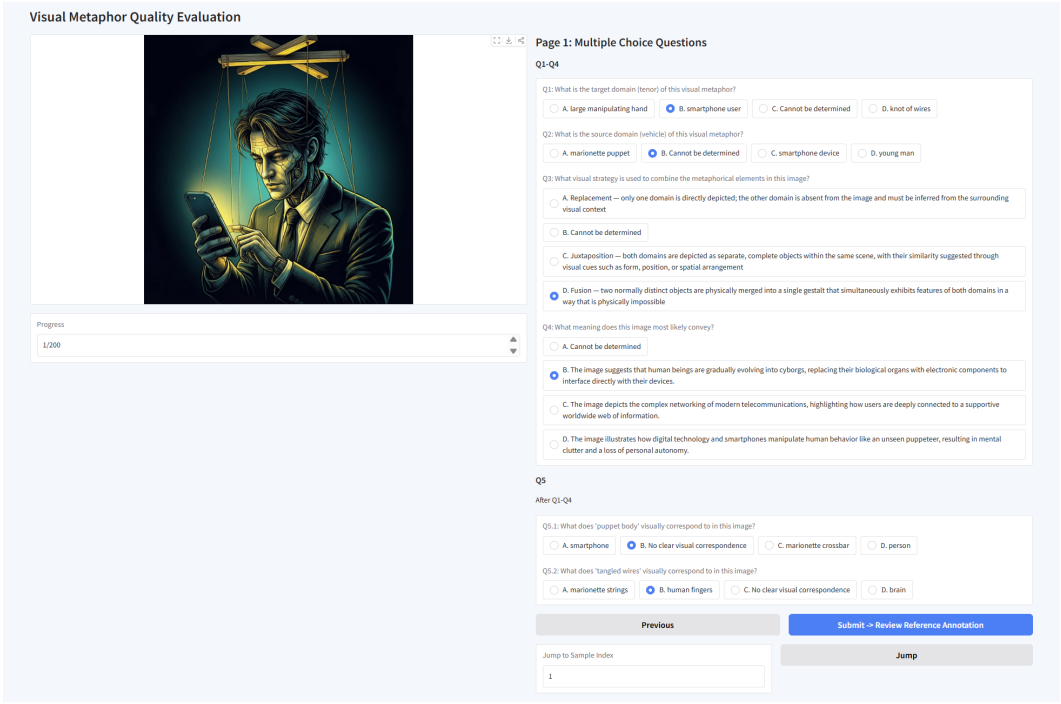


Figure 13: Human evaluation interface for the alignment study: MCQ answering

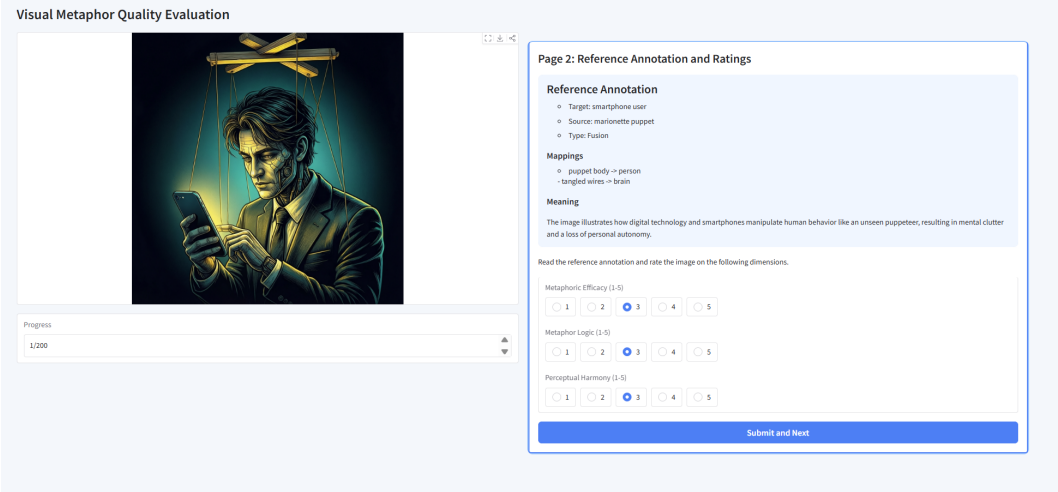


Figure 14: Human evaluation interface for the alignment study: dimension rating

### A.12 SAMPLES

We present representative samples in Figure 15. For each category, we select one visual metaphor and display outputs from the models included in this qualitative analysis under the conceptual prompt setting. These examples illustrate the diversity of metaphorical themes in our benchmark and highlight the varying capabilities of different models in generating visual metaphors.

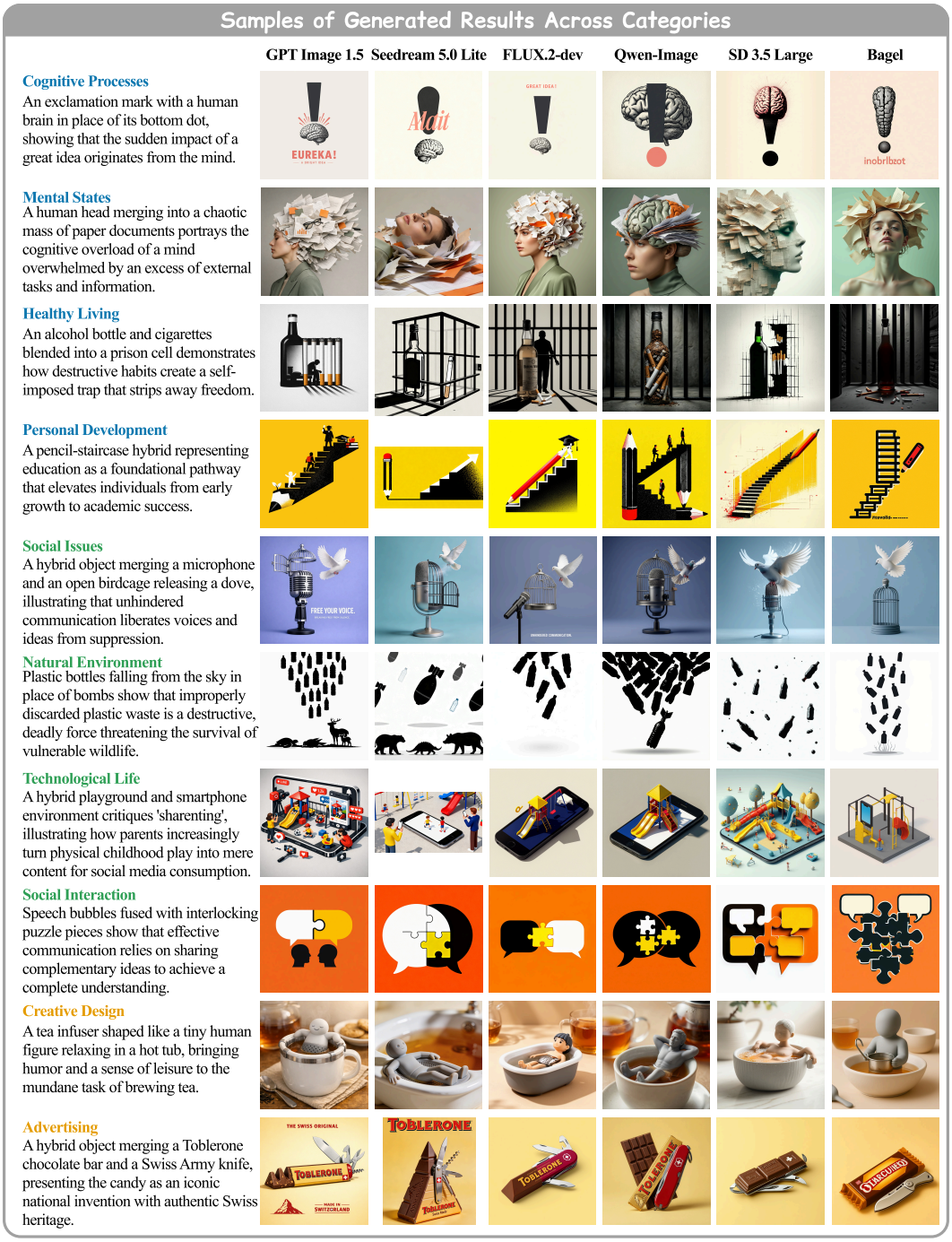


Figure 15: Generated visual metaphors across ten categories from VMetaphor-Bench.

### A.13 Broader Impacts

This work advances the evaluation of visual metaphor generation, a challenging aspect of text-to-image models that requires coherent cross-domain reasoning and expressive visual composition. By providing a structured benchmark with fine-grained evaluation criteria, it can support the development of more capable and interpretable generative models. Such progress may benefit real-world applications including creative design, education, and advertising, where visual metaphors are widely used to communicate messages effectively and engage audiences.

We also acknowledge potential negative implications. Improved visual metaphor generation could be misused to produce persuasive misleading imagery (e.g., manipulative advertising or political propaganda) that exploits the strong emotional and rhetorical effects of visual analogies. We encourage future work to consider safeguards against such misuse, and our benchmark itself is intended for evaluation rather than direct deployment.

### A.14 Data Licensing and Release

To respect the rights of original content creators on Pinterest, we do not redistribute the source images. Our public release consists of: (1) image URLs pointing to the original Pinterest pages, (2) all structured annotations (target/source domains, structure types, meaning, element mappings), (3) descriptive and conceptual prompts, and (4) the 9,594 multiple-choice questions. Users can retrieve images via the URLs for research purposes following Pinterest's terms of service. The annotations and prompts are released under CC-BY-4.0.