PRISM: A CATEGORY-THEORETIC FRAMEWORK FOR MEASURING AND REFINING MULTIMODAL ANALOGIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Analogical reasoning involves identifying and preserving relational structures across domains. However, existing approaches to AI-driven multimodal analogy generation lack an interpretable measure of whether this structure is understood and maintained in the generated output. We address this gap with Pullback Refinement via Interpretable Structural Mapping (PRISM), a modality-agnostic framework for measuring and improving relational alignment in multimodal analogies, evaluated on visual metaphor generation. PRISM represents analogies as explicit relational mappings grounded in category theory and uses VLMs to instantiate these structures across modalities. Its first component, the pullback score, quantifies relational alignment from the resulting graph representation. On the AnaloBench benchmark, selecting the correct analogy purely by pullback score achieves 82.5% accuracy, demonstrating that the score captures meaningful relational information. PRISM's second component is an iterative refinement loop that uses the pullback score as an in-context feedback signal to iteratively revise the generated image towards greater relational depth. VLM-as-a-judge and human evaluations show that PRISM consistently improves metaphor consistency and analogy appropriateness over zero-shot generation, with human participants preferring the refined output in 57.65% of pairwise comparisons. However, a qualitative analysis reveals that refinement can favour visually crowded compositions rather than genuinely deeper relational correspondences. Our code is placed at `https://anonymous.4open.science/r/multimodal-analogy-review-6C64`.

1 INTRODUCTION

Human cognition is deeply rooted in analogical reasoning, which is the ability to identify and transfer relational structures across domains. Already present in young infants, it is central to learning, creativity and generalisation (Gentner & Hoyos, 2017). An analogy operates not by matching surface-level features, but instead by aligning the underlying relational structure between a source and a target domain (Gentner, 1983). This relational abstraction enables humans to apply knowledge across vastly different contexts, from scientific reasoning to creative expression. As AI systems are increasingly deployed in open-ended, multi-domain settings, the ability to reason analogically in a similarly generalisable manner becomes a critical capability.

Advances in large language model (LLMs) and vision-language model (VLMs) capabilities have produced systems that perform strongly on a wide range of tasks, including several benchmarks designed to probe analogical reasoning (Qin et al., 2025; Musker et al., 2025). However, strong benchmark performance does not necessarily imply genuine relational abstraction. There is growing evidence that models exploit statistical regularities in training data rather than performing true structural inference (Mitchell, 2021; Lewis & Mitchell, 2024). This distinction is especially pronounced in the multimodal setting, where recent work shows that current VLMs still struggle to generalise relational rules across visual transformations (Yiu et al., 2024), a difficulty echoed by recent diagnostic benchmarks for visual analogical mapping (Yilmaz et al., 2025), compositional analogies (Du et al., 2026), and visual metaphor understanding (Kundu et al., 2025). Furthermore, even when models produce correct outputs or generate explanations, their internal reasoning processes remain opaque and do not reliably correspond to interpretable analogical mappings (Chen et al., 2025; Opiełka et al., 2025).

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.

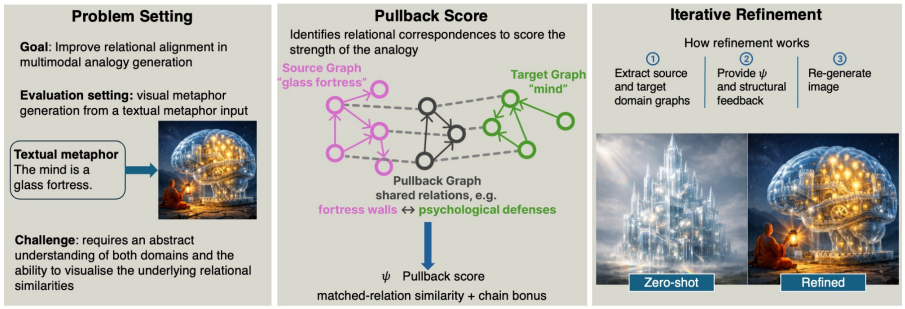


Figure 1: Overview of **PRISM**, a category-theoretic framework for structural alignment within multimodal analogies. The pullback score quantifies relational correspondences between source and target domain graphs and provides structural feedback for iterative refinement. We evaluate PRISM on visual metaphor generation, where refinement aims to produce analogies with stronger relational correspondences.

Existing approaches to analogical reasoning in AI tend to be either powerful but opaque, or interpretable but narrow. Structured prompting strategies and schema-based methods have demonstrated progress in specific task settings — such as scientific analogy generation (Yuan et al., 2023) or visual metaphor transfer (Xu et al., 2026) — but these approaches do not generalise across task formulations. What remains missing is a unified, human-interpretable representation of analogical structure that can operate across modalities and task types while still leveraging the broad knowledge encoded in state-of-the-art foundation models.

In this paper, we address this gap by proposing Pullback Refinement via Interpretable Structural Mapping (PRISM), a modality-agnostic structural framework for analogical reasoning that is both human-interpretable and generalisable across diverse reasoning tasks. PRISM represents analogies as explicit relational mappings grounded in category-theoretic constraints, and uses VLMs to populate and employ these representations across multimodal inputs. We evaluate PRISM using visual metaphor generation as the testbed, since this task requires an abstract understanding of both domains, and generative tasks are less prone to shortcut learning (Mitchell, 2021; Pekar et al., 2020).

Our contributions are two-fold: (1) We **translate a category-theoretic definition of analogical structure into an implementable representational schema**. Building on the formal account of analogies in Ott (2025), we instantiate a domain graph representation that realises category-theoretic analogies within a VLM-based pipeline. Rather than relying on task-specific heuristics, the representation captures the relational structure between a source and target domain in a unified, modality-agnostic form while remaining faithful to the underlying theory. From this representation, we derive the **pullback score**, a category-theoretic measure of relational alignment that quantifies the extent to which structure is preserved across a mapping. (2) We **design and evaluate PRISM's iterative refinement loop, which improves the relational quality of generated analogies using feedback on structural alignment**. This loop makes use of the pullback score as a feedback signal to encourage models to generate analogies that preserve deeper relational correspondences rather than superficial similarities, providing a practical step towards more transparent, generalisable and structurally grounded analogical reasoning in multimodal AI systems.

2 Related Work

Analogies in AI. Structure-Mapping Theory characterises analogies as alignments between the relational structures of a base and target domain rather than their surface attributes (Gentner, 1983; Gentner & Hoyos, 2017). This idea has been explored through symbolic systems based on hand-crafted representations, neural approaches that learn analogies from data, and neuro-symbolic methods that combine learned representations with explicit structure (Hofstadter, 1984; Evans, 1964; Marcus, 2003; Mikolov et al., 2013; Sadeghi et al., 2015; Mitchell, 2021; Lake et al., 2017; Evans & Grefenstette, 2018; Yi et al., 2018). More recent work has equipped language and vision-language models with symbolic scaffolds, including tool use and structured reasoning procedures (Schick et al., 2023; Gao et al., 2023; Olausson et al., 2023; Pan et al., 2023; Gupta & Kembhavi, 2023). Graph-based

approaches make entities, relations, and higher-order correspondences explicit, extending from the Structure-Mapping Engine and visual relational reasoning to learned graph alignment, multimodal knowledge graphs, and structured visual metaphor understanding (Falkenhainer et al., 1989; Lovett & Forbus, 2017; Ling et al., 2022; Zhang et al., 2023; Xu et al., 2026). We extend these directions by introducing a formal graph-theoretic scaffold that VLMs populate and reason over, using soft relational similarity and pullback operations to construct and quantify cross-domain correspondences.

Visual Metaphor Generation. Text-to-image models produce realistic images from prompts (Rombach et al., 2022), but are optimised for literal content rather than the relational structure underlying metaphors, and both VLMs and text-to-image generation models struggle with the abstraction visual metaphors require (Akula et al., 2023). Prior work narrows this gap by adding visual detail to prompts (Chakrabarty et al., 2023), grounding generation in explicit structures such as attribute-object binding (Zhang et al., 2024), or introducing iterative feedback, either through structured metaphor decomposition and multi-faceted rewards (Koushik et al., 2025) or VLM-generated critiques of the visual elaboration (Kundu et al., 2026). Xu et al. (2026) instead learns from visual references rather than text, transferring metaphors between images. Unlike this prior work, we ground refinement in an explicit, category-theoretic measure of relational alignment rather than a learned or heuristic reward.

3 BACKGROUND

Category theory studies structure-preserving mappings between collections of objects (see Awodey (2010) and Jencel (2021) for a comprehensive introduction). Ott (2025) applies this to analogies by representing each domain as a coloured multigraph, where nodes are the entities of the domain and edges are the relations between them, coloured such that two edges share a colour exactly when their relations are similar enough to be treated as analogous. The product of two domain graphs represents every possible pairing between their elements, and the pullback is a restricted product that additionally requires paired edges to share a colour, so that only structure-preserving correspondences survive. Individual analogies are then subgraphs of this pullback, and the preferred mapping is the one that preserves the largest number of relations.

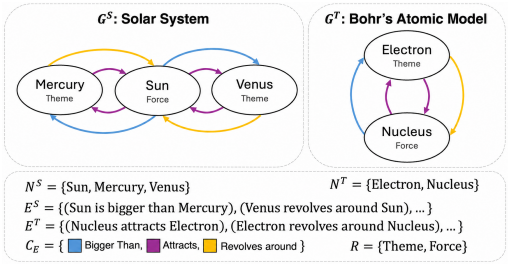


Figure 2: Coloured domain graphs for the solar system (G^S) and Bohr's atomic model (G^T). Shared node roles and edge colours indicate analogous entities and relations across the two domains.

We adopt this representation directly: a domain pair (D^S, D^T) is represented as a coloured graph pair

$$G = (G^S, G^T, c_E, c_N), \qquad G^S = (N^S, E^S), \qquad G^T = (N^T, E^T), \tag{1}$$

where $N = N^S \cup N^T$ is a set of nodes (entities) and $E = E^S \cup E^T$ is a set of edges (relations), each represented as a triple (u, r, v) with r the relation label. Edges in E^S are constrained to $u, v \in N^S$ and edges in E^T to $u, v \in N^T$, so that G^S and G^T remain structurally independent prior to the pullback operation. Every node is assigned a role by $c_N : N^S \cup N^T \to R$, drawn from a semantic-role vocabulary R, and every edge is assigned a colour by $c_E : E^S \cup E^T \to C_E$, drawn from a colour set C_E.

Formally, sharing c_E across the two domain graphs gives a cospan of edge-colouring maps into the common codomain C_E,

$$E^S \xrightarrow{c_E} C_E \xleftarrow{c_E} E^T, \tag{2}$$

and the pullback invoked is the pullback of Equation (2) in the category of sets

$$E^S \times_{C_E} E^T \;=\; \{(e, e') \in E^S \times E^T \mid c_E(e) = c_E(e')\}, \tag{3}$$

together with its projections onto E^S and E^T. An individual analogy is then a subset $\mathcal{P} \subseteq E^S \times_{C_E} E^T$ whose endpoints determine a single, consistent element of the node correspondence, computed by the matching procedure in Algorithm 2. Section 4 develops this pullback into a computable score.

4 METHODOLOGY

We consider a visual metaphor as a source and a target domain (D^S, D^T) related by a shared relational structure. Section 4.1 constructs a procedure that extracts the coloured graph pair (G^S, G^T) from a visual metaphor and defines a scalar pullback score $\psi(G^S, G^T) \in \mathbb{R}_{\geq 0}$ measuring how much of the target domain's relational structure is preserved in the source domain's depiction. Section 4.2 then composes this extraction procedure and score with an image-generation model and a refinement model into PRISM's iterative refinement loop, which uses ψ as feedback to improve subsequently generated visual metaphors.

4.1 THE PULLBACK FRAMEWORK FOR STRUCTURAL ANALOGY EXTRACTION AND EVALUATION

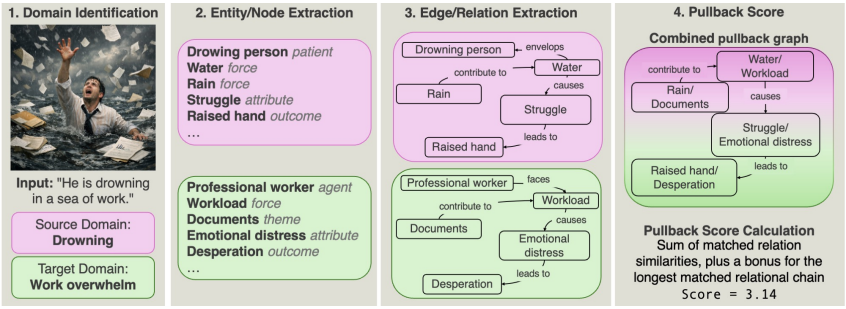


Figure 3: PRISM's four-stage domain-graph extraction pipeline for computing the pullback score from a visual metaphor: (1) source and target domain identification, (2) entity extraction, (3) relation extraction into coloured domain graphs, and (4) pullback score computation to quantify preserved relational structure.

Constructing coloured domain graphs. We formalise the extraction procedure as a map

$$\mathrm{ext} : \mathcal{X} \cup \mathcal{M} \longrightarrow \mathcal{G} \times \mathcal{G}, \qquad \mathrm{ext}(x) = (G^S, G^T) \tag{4}$$

where $\mathcal{X}$ is the space of visual metaphors, $\mathcal{M}$ is the space of textual metaphors, and $\mathcal{G}$ is the space of coloured graphs. Given a visual metaphor $x \in \mathcal{X}$ or a textual metaphor $m \in \mathcal{M}$, ext returns the coloured graph pair of its two domains. We instantiate ext as a four-stage VLM pipeline that converts a visual metaphor x into G^S and G^T through incremental abstraction, where each stage produces structured output that conditions the next. As illustrated in Figure 3, the pipeline identifies the source and target domains D^S, D^T, extracts their entities into node sets N^S, N^T, and extracts the relations between those entities into edge sets E^S, E^T of (u, r, v) triples.

Each extracted entity is also assigned a semantic role (e.g., "agent", "instrument", or "outcome"), instantiating the role map c_N. The role inventory is derived from case grammar (Fillmore, 1968) and semantic role labelling resources (Kipper et al., 2008). Replacing entity labels with semantic roles reduces the label scatter produced by unconstrained LLM extraction (Khojasteh et al., 2026). Relation predicates are likewise drawn from a shared closed vocabulary during extraction (Section A) for the same reason, to keep the extracted triples well-formed for the sentence encoder. This constrains only how individual edges are labelled, not how they are compared. Cross-domain correspondence does not require these predicate labels to match, since colour-matching is instead computed from the embedding similarity of the full role-substituted triple, so PRISM avoids the fixed, exact-match relation taxonomy used for structural alignment in prior graph-based methods such as SME.

Edge colours are computed by embedding role-substituted relation triples. Specifically, each edge (u, r, v) is rendered as the string "$c_N(u)\ r\ c_N(v)$", replacing its head and tail entities with their semantic roles (e.g., (*electron*, *revolves around*, *nucleus*) becomes (*theme*, *revolves around*, *force*)), ensuring that similarity depends on relational structure rather than lexical overlap. Using `all-mpnet-base-v2` (Reimers & Gurevych, 2019) as the sentence encoder $\mathcal{F}$, we represent the embedded source and target triples as $\hat{E}^S = \mathcal{F}(E^S) \in \mathbb{R}^{|E^S| \times d}$ and $\hat{E}^T = \mathcal{F}(E^T) \in \mathbb{R}^{|E^T| \times d}$. We then compute the cosine similarity matrix $\Phi = \hat{E}^S \cdot (\hat{E}^T)^\top \in \mathbb{R}^{|E^S| \times |E^T|}$, where Φ_{ij} denotes

the similarity between source edge i and target edge j. Edges with $\Phi_{ij} \geq \theta$ are assigned the same colour, instantiating c_E and replacing SME's exact relation matching with soft semantic equivalence (Mikolov et al., 2013). Algorithm 2 therefore matches edges directly on the thresholded pairwise similarities Φ_{ij} rather than on precomputed colour classes, so the pullback score is best understood as a practical relaxation of the categorical construction in Equation (3) rather than an exact instantiation of it. The complete extraction and graph construction procedure is summarised in Algorithm 1. We validate the extraction pipeline on the SCAR dataset (Yuan et al., 2023), with prompts and evaluation details provided in Sections A and B.

Algorithm 1 Pullback Score Computation

Require: Source domain graph $G^S = (N^S, E^S)$, target domain graph $G^T = (N^T, E^T)$, role map c_N, sentence encoder $\mathcal{F}$, matching threshold θ
Ensure: Pullback score ψ, matched pairs $\mathcal{P}$, coverage ρ
1: $E^S \leftarrow \{(u, r, v) \in E^S \mid u, r, v \text{ all defined}\}$, $E^T \leftarrow \{(u, r, v) \in E^T \mid u, r, v \text{ all defined}\}$
2: **if** $E^S = \emptyset$ **or** $E^T = \emptyset$ **then**
3: **return** $\psi = 0, \mathcal{P} = \emptyset, \rho = 0$
$\hat{E}^S \leftarrow \mathcal{F}\big(\{\text{“}c_N(u)\ r\ c_N(v)\text{”} \mid (u, r, v) \in E^S\}\big) \in \mathbb{R}^{|E^S| \times d}$ ▷ Embed role-substituted edges
$\hat{E}^T \leftarrow \mathcal{F}\big(\{\text{“}c_N(u)\ r\ c_N(v)\text{”} \mid (u, r, v) \in E^T\}\big) \in \mathbb{R}^{|E^T| \times d}$
4: $(\psi, \mathcal{P}, \rho) \leftarrow \textsc{GreedyPullback}(E^S, E^T, \hat{E}^S, \hat{E}^T, \theta)$ ▷ Algorithm 2 in Section D
5: **return** $\psi, \mathcal{P}, \rho$

The pullback score. Given the coloured graphs G^S and G^T produced above, the pullback graph contains all pairs of nodes and edges whose corresponding edges share a colour, thereby filtering the Cartesian product to structurally consistent correspondences. The resulting score is defined as

$$\psi(G^S, G^T) = \sigma + 0.5\,\ell, \qquad \sigma = \sum_{(i,j) \in \mathcal{P}} \Phi_{ij}, \tag{5}$$

where $\mathcal{P} \subseteq E^S \times E^T$ is the set of source-target edge pairs retained by the pullback, so pairs (i, j) with colour-matching similarity $\Phi_{ij} \geq \theta$, accepted greedily in order of decreasing similarity subject to every accepted pair preserving a single consistent node mapping across the whole set. Here, ℓ is the length, measured as the number of edges, of the longest cycle-free directed path in the target graph induced by those matches. The similarity term σ measures relational correspondence, while the chain bonus rewards metaphors whose correspondences form coherent relational chains rather than isolated matches. Neither component is arbitrary: we validate both, along with the role substitution introduced above, on an independent analogy benchmark below.

Validating the pullback score. Before using the pullback score as a feedback signal, we validate that it carries meaningful relational signal on an independent benchmark, separate from the visual metaphor dataset used throughout the rest of this paper. We use the AnaloBench dataset (Ye et al., 2024), a collection of stories of varying length, some of which are analogous. Given a narrative, the task is to identify the analogous story among four candidates (task T1 by Ye et al. (2024)).

We compare a zero-shot baseline, in which GPT-4.1 Mini directly selects the analogous story, against extracting domain graphs for the narrative and each candidate and selecting the candidate with the highest pullback score. Zero-shot prompting achieves 88.5% accuracy. Selecting purely by pullback score achieves 82.5% accuracy, with a mean reciprocal rank of 0.895 and a top-2 accuracy of 91.5%. The correct answer never received a pullback score of zero, indicating that the score always identifies at least some structural correspondence with the true analogy. Thus, although the pullback score is less accurate than zero-shot prompting as a stand-alone decision rule, it captures the structural signal needed for refinement.

The two approaches also make different errors. Of 200 examples, both methods agree on 159 correct and 17 incorrect cases, while 6 are solved only by the pullback score and 18 only by zero-shot prompting (McNemar $p = 0.0227$), indicating meaningfully different error distributions. Ablating the score confirms both of its components are necessary, as removing the chain bonus lowers accuracy to 76%, and removing the semantic-role abstraction lowers it to 24%, close to the 25% chance rate on this four-way task. Extended analysis is provided in Section C.

4.2 Feedback-Driven Refinement of Analogies

We propose PRISM's refinement loop, which combines an image-generation model gen, the extraction procedure ext, the pullback score ψ, and a refinement model ref into the feedback loop shown in Figure 4. Given a textual metaphor m and visual elaboration prompt p, gen produces a visual metaphor x. From x, we extract $(G^S, G^T) = \mathrm{ext}(x)$ and compute the pullback score $\psi(G^S, G^T)$. ref then revises the prompt using feedback derived from these signals. Using the pullback score as feedback steers generation towards analogies with stronger relational correspondences between domains.

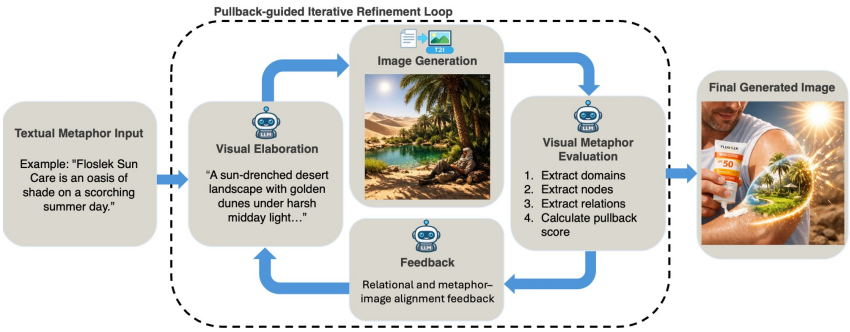


Figure 4: PRISM's refinement loop: a textual metaphor is visualised, scored via the pullback score, and revised in-context using feedback from the score and extracted domain graphs, repeating for up to k rounds.

Let p_0 denote the initial visual elaboration prompt of the textual metaphor m, produced by a separate elaboration model. At each refinement round $t \in \{1, \ldots, k\}$,

$$x_t = \mathrm{gen}(m, p_{t-1}), \qquad \text{(image generation)} \tag{6}$$

$$(G_t^S, G_t^T) = \mathrm{ext}(x_t), \qquad \psi_t = \psi(G_t^S, G_t^T), \qquad \text{(structural evaluation)} \tag{7}$$

$$p_t = \mathrm{ref}(m, p_{t-1}, f_t), \qquad \text{(prompt refinement)}, \tag{8}$$

where the refinement feedback at round t is $f_t = \{\psi_t, \rho_t, \mathcal{P}_t, G_t^S, G_t^T, a_t\}$. Here, ψ_t is the pullback score, $\rho_t = |\mathcal{P}_t|/|E_{T,t}|$ is the edge coverage, $\mathcal{P}_t$ is the set of matched edge pairs, G_t^S and G_t^T are the extracted source and target graphs, and a_t is the CLIPScore (Hessel et al., 2021) image–text alignment between x_t and m.

Together with the original metaphor m and previous elaboration p_{t-1}, these signals provide both global and local feedback. The scalar scores ψ_t, ρ_t, and a_t summarise structural quality, relational coverage, and semantic alignment, while $\mathcal{P}_t$ and the extracted graphs identify which target-domain relations remain unmatched. The refinement model is instructed to preserve correctly expressed structure and revise only the missing or inconsistent relations. The process terminates when either $t = k$ or the pullback score converges, i.e. $|\psi_t - \psi_{t-1}| < \varepsilon$, at which point the final image x_t is returned.

5 Experimental Setup

Baselines. We evaluate PRISM's generated analogies against two classes of baselines. First, we compare the refined images against the zero-shot outputs produced by Flux.2 Klein 4B (Black Forest Labs, 2026), Stable Image Core v1.1 (Stability AI, 2024), GPT Image 1.5 (OpenAI, 2025a) and Ideogram 4 (Ideogram, 2025). Second, we compare PRISM's refinement procedure against GEPA (Agrawal et al., 2025), a reflective prompt optimiser that iteratively proposes, evaluates, and selects prompt variants using natural-language feedback and Pareto-based evolutionary search. We evaluate GEPA with two different optimisation objectives, one that uses the pullback score, while the other uses CLIPScore (Hessel et al., 2021).

Dataset. The dataset contains two subsets, the first consists of 125 human-created visual metaphors collected from online sources and existing datasets (Hussain et al., 2017; Hwang & Shwartz, 2023),

including advertisements (83), cartoons and artworks (18), awareness campaigns (20), and memes (4). For each image, we manually transcribe the underlying metaphorical intent into a textual metaphor and annotate the corresponding source and target domains. The second subset contains 125 textual metaphors from the HAIVMet dataset (Chakrabarty et al., 2023). We publicly release the collected and annotated human-created metaphors, together with the visual metaphors generated for the full set of textual metaphors used in our experiments, at `https://doi.org/10.5281/zenodo.21891311`.

Metrics. Evaluating visual metaphors requires both understanding the intended metaphorical meaning and recognising the abstract source and target domains depicted in the image. We therefore rely on both automated and human judgements. We evaluate outputs using two VLM judges, Claude Opus 4.8 (Anthropic, 2026b) and GPT-5.2 (OpenAI, 2025b), and corroborate these scores with human ratings. Including GPT-5.2 provides an independent model family because the refinement loop uses Claude Sonnet 4.6. Each model is prompted to score each visual metaphor on three 10-point scales: *metaphor consistency (MC)*, measuring whether the image preserves the core logic of the textual metaphor, *analogy appropriateness (AA)*, assessing the validity of the functional and formal correspondences between the depicted domains, and *conceptual integration (CI)*, evaluating whether the two domains are fused naturally and coherently (Xu et al., 2026). The evaluation prompts adapt Xu et al. (2026) to assess generated images against textual metaphors rather than source images. We further conduct a human evaluation with 80 participants, in which each metaphor is independently reviewed by two participants. Each participant scored five pairs of zero-shot and refined metaphors on analogy appropriateness and metaphor consistency, and selects which image better captures the underlying relational structure. The full survey is provided in Section E.

Implementation. After comparing three state-of-the-art image generator models in a zero-shot setting (see Sections F and G), we selected GPT Image 1.5 (OpenAI, 2025a) as the primary image generator. We conduct hyperparameter tuning to select the base VLM used for visual elaboration and pullback extraction, to define the edge similarity matching threshold θ, and to calibrate the refinement-loop parameters k and ε (Section H). This procedure selects Claude Sonnet 4.6 (Anthropic, 2026b) as the feedback model, with $k = 10$, $\varepsilon = 0.05$, and a matching threshold of $\theta = 0.40$. To test the applicability of the methodology with different models, we later re-ran the experiments with GPT-6 Astra (OpenAI, 2026).

6 Results

Qualitative comparisons show that PRISM improves the recognisability of both domains and strengthens the relational correspondences between them. As shown in Figure 5, state-of-the-art image-generation models often fail to depict both domains of a metaphor clearly in zero-shot generation, particularly when one domain involves an abstract concept such as "mind" or "thought". PRISM produces more interpretable visual metaphors by making both domains recognisable and rendering their relationship more explicit. In example (a) from Figure 5, for instance, the refined image introduces smokestacks alongside the melting ice, visually linking industrial emissions to global ice loss and thereby clarifying the intended causal correspondence. This qualitative improvement is also reflected in the human evaluations, in which participants preferred the PRISM-refined image in 57.65% of forced-choice comparisons. Examples across all zero-shot generations and PRISM with different refinement models can be found in Appendix I.



Figure 5: Qualitative comparison of generated visual metaphors. (a) "Global ice melting is as dangerous as a missile." (b) "Thought is a vulture" (c) "My mind is a desert and I am an explorer"

PRISM significantly improves metaphor consistency and analogy appropriateness compared to zero-shot image generation. As Table 1 shows, PRISM outperforms zero-shot visual metaphor generation across all axes in both human and VLM evaluations. Notably, analogy appropriateness and metaphor consistency achieve significantly larger gains than conceptual integration. This indicates that refinement acts as a semantic corrector rather than an aesthetic

one, strengthening the metaphor's conceptual relations without necessarily improving overall image quality or composition. We also tested GPT-6 Astra as an alternative extraction and refinement model. Metaphor consistency, analogy appropriateness, and conceptual integration all improve on the Astra runs, but by smaller margins than with Claude Sonnet 4.6. This may be because the Astra zero-shot scores are already higher, so there is less room for improvement.

Figure 6 depicts how the scalar ratings between the zero-shot and refined images compare across the evaluation axes and shows that refined images are rated higher more often than zero-shot images. Across both axes, the combined proportion of 'refined better' and 'equally good' outcomes exceeds 50%. Ties in quality between zero-shot and refined are mostly favourable, but there is also a substantial proportion of metaphors in which zero-shot generation outperforms the refined metaphors.

Table 1: Zero-shot (ZS) versus PRISM-refined (Ref.) outputs, scored 0 to 10. VLM scores average all judges; human raters did not assess CI. Δ is the mean improvement.

Model / *Judge*		**MC**	**AA**	**CI**
Sonnet 4.6 *VLM*	ZS	6.42	6.36	7.85
	Ref.	7.65	7.44	8.15
	Δ	↑1.24	↑1.08	↑0.30
Sonnet 4.6 *Human*	ZS	5.48	5.27	–
	Ref.	5.93	5.84	–
	Δ	↑0.45	↑0.57	–
GPT-6 Astra *VLM*	ZS	7.65	7.33	7.27
	Ref.	8.18	7.81	7.32
	Δ	↑0.53	↑0.49	↑0.05

Table 2: Comparisons between PRISM, zero-shot (ZS) image-generation baselines, GEPA (Agrawal et al., 2025), and ablations of PRISM. "Overall" averages scores across all axes and both judges. Best result per column is shown in bold.

	Opus 4.8			**GPT-5.2**			**Overall**
Condition	**MC**	**AA**	**CI**	**MC**	**AA**	**CI**	
Stable Core (ZS)	3.75	3.70	4.90	4.30	4.95	6.40	4.67
Flux Klein (ZS)	5.05	4.90	6.06	5.16	5.63	7.32	5.69
GPT-Image (ZS)	6.25	6.15	7.50	6.50	6.70	7.85	6.83
Ideogram 4 (ZS)	5.26	5.37	6.84	5.37	6.05	7.24	6.02
GEPA (pullback)	6.57	6.29	6.34	7.02	6.62	6.88	6.62
GEPA (CLIPScore)	6.78	6.35	6.46	7.23	6.78	6.93	6.76
No domain guidance	4.32	4.45	4.92	4.88	5.15	5.87	4.93
One-shot extraction	6.24	6.23	6.03	6.75	6.78	6.80	6.47
PRISM (Ours)	**7.60**	**7.30**	**8.35**	**7.85**	**7.70**	**8.43**	**7.87**

Evaluating visual metaphors is inherently difficult and subjective. Agreement is limited both between human and automated evaluators and among human raters themselves. Human and VLM scores are significantly positively correlated, but only moderately for zero-shot images (Pearson's $r = 0.439$–0.446) and more weakly for refined images ($r = 0.270$–0.303). Similarly, agreement between human and VLM forced-choice preferences is only fair (Cohen's $\kappa = 0.270$, 66.2% raw agreement), with the VLM identifying 81.4% of human preferences for refined images but only 44.4% of preferences for zero-shot images. Human inter-rater agreement is also low for absolute scores, with Krippendorff's $\alpha = 0.082$ for analogy appropriateness and $\alpha = 0.111$ for metaphor consistency, but increases for forced-choice judgements ($\alpha = 0.233$). Nevertheless, raters agree on the direction of improvement for 71.5% of metaphors in analogy appropriateness and 74.7% in metaphor consistency, suggesting that relative comparisons are more reliable than absolute quality ratings. Together, these results show that visual metaphor evaluation depends on subjective interpretations of whether relational structure has been preserved, motivating the pullback score as a directly computable and less rater-dependent measure of relational alignment.

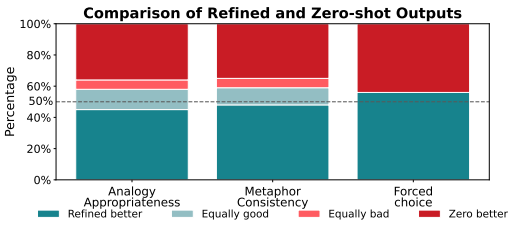


Figure 6: Human pairwise comparison of refined and zero-shot outputs. Tied ratings are classified as equally good for scores of at least 5 and equally bad otherwise.

Explicit domain guidance and structured refinement drive PRISM's performance. We conduct baseline and ablation comparisons. As shown in Table 2, PRISM achieves the highest overall score of 7.87, outperforming all zero-shot baselines, GEPA variants, and ablations. Among the GEPA baselines, optimisation with CLIPScore performs best, reaching an overall score of 6.76, compared

with 6.62 when GEPA is optimised directly with the pullback score. Both remain below the GPT-Image zero-shot baseline at 6.83 and substantially below PRISM. Replacing PRISM's multi-stage extraction with a single extraction step reduces performance to 6.47, while removing domain guidance from the refinement feedback, including both the input metaphor and CLIPScore, produces the largest degradation, to 4.93 overall. These ablations show that PRISM benefits from explicit source–target grounding and iterative structured refinement. GEPA performs better with CLIPScore than with the pullback score, suggesting that image–text alignment is a more effective prompt-level objective in this setting. However, neither variant matches the full pipeline, and GEPA with CLIPScore lacks PRISM 's modality-agnostic structural representation.



Input: The pharmaceutical industry is a money machine. **Input**: Failing to social distance is an explosion waiting to happen. **Input**: The sky is the limit.

Figure 7: Representative failure modes. **(a) Maximalism:** whereas the human-designed metaphor (Schmitz, 2025) communicates pharmaceutical corruption through a single dollar-shaped microscope, the refined output introduces multiple redundant monetary cues. **(b) Text as a shortcut:** generated images use written labels to identify domains such as "social distancing" or "limit", improving domain extraction without strengthening the underlying relational correspondence.

Limitations: maximalism and text as a shortcut. Qualitative inspection suggests two recurring failure modes. First, refinement often produces increasingly cluttered compositions by adding rather than removing visual elements. Second, generated images may rely on literal text to establish the intended domains; because the VLM-based extractor can read this text directly, such images may receive a high pullback score without expressing a clearer visual correspondence, as illustrated in Figure 7.

7 Conclusion

This paper introduced PRISM, comprising the pullback score as a category-theoretic measure of relational alignment computed from VLM-extracted domain graphs, and a feedback-driven refinement loop that uses it to guide iterative, test-time image generation. We validated that the pullback score carries meaningful analogical signal on the AnaloBench benchmark, and showed that VLMs struggle to depict the domains and relational correspondences of a textual metaphor zero-shot, a dissociation PRISM is designed to close. Across 200 metaphors, refinement yields statistically significant improvements in metaphor consistency and analogy appropriateness according to VLM judges and human raters. The direction of the improvement is replicated when running the loop using GPT-6 Astra, but smaller gains highlight the dependence of the pipeline's performance on the models used for image generation, extraction, and refinement. At the same time, our evaluations show that judging analogical quality, whether by VLM or by human, is itself difficult and only moderately consistent, highlighting the need for interpretable, structurally grounded metrics like the pullback score.

Beyond visual metaphor generation, we see this work as a step towards making analogical reasoning in multimodal AI systems more transparent, measurable, and improvable, rather than treating relational alignment as an unverifiable byproduct of generation. Applying PRISM beyond visual metaphor generation in future research, to other multimodal or purely textual analogical reasoning tasks, would test how far this structural approach to relational alignment generalises.

AI use statement

In this work, we used generative AI tools for helping develop theoretical models or conceptual frameworks, formulating mathematical claims, designing or providing feedback on research methodology or experiments, implementing methods and interpreting results. We have not used generative AI

tools for proposing or refining hypotheses, generating synthetic data sets, cleaning and reformatting datasets, or supporting qualitative and thematic data analysis. Providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, and assisting with translation are not applicable to this work. Additionally, we used generative AI tools for modifying scientific figures, creating or editing software code, editing the research paper to improve readability, summarizing existing literature and proposing a title for the research paper. We have reviewed all AI-assisted work.

The conceptual framework was grounded in the category-theoretic understanding of analogies in Ott (2025) and generative AI assisted in translating that formal account into an implementable graph schema. The pullback score in Equation (5) and the matching procedure in Algorithm 2 were formulated with AI assistance and verified by the authors with hand-crafted examples adapted from the SCAR dataset (Yuan et al., 2023). AI-generated code was reviewed by the authors and tested against expected behaviour, e.g. the extraction pipeline was independently verified on the SCAR dataset (Yuan et al., 2023) and the AnaloBench dataset (Ye et al., 2024).

The figures in the paper are hand-crafted by the authors, and only modified using generative AI. Two such examples are Figure 2, where the curved arrows were added using AI tools, and Figure 4, which contains AI-generated icons for LLMs and T2I generation models. For the literature review, AI was used to collect and summarize relevant literature, but each cited work was consulted by at least one author to confirm that they exist and to validate the claims attributed to them. All report sections were initially drafted by the authors, and refined for readability and clarity using generative AI tools. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Ethics statement

Human subjects. The human evaluations reported in Section 5 was reviewed and approved by our institute's Human Research Ethics Committee before approaching participants, and reviewed by the faculty's data management steward. Participants were recruited through a third-party survey exchange platform and compensated in platform credit. Furthermore, participation was voluntary with the option to withdraw at any time. Beyond the visual metaphor evaluations, we collected only age, gender and education levels, and participants also had the option to skip these questions. All images shown to participants were screened beforehand and visual metaphors that were deemed violent or disturbing were removed. However, we also acknowledge the limitations of the resulting sample. Recruitment through a survey exchange platform leads to a skewed pool of participants that is largely focused on students and researchers. Because metaphor interpretation is subjective and culturally situated, this sample should not be taken as a representation of the general audience.

Data provenance and licensing. In our evaluation, we draw our dataset from two main sources, as described in Section 5. 125 textual metaphors are collected from the HAIVMet dataset by Chakrabarty et al. (2023) and the 125 human-created visual metaphors used to inform the remaining textual metaphors were collected from online sources and existing research datasets (Hussain et al., 2017; Hwang & Shwartz, 2023). As to not violate these third-party copyrighted works, these collected visual metaphors are not redistributed in the online dataset found at `https://doi.org/10.5281/zenodo.21891311`. Instead, this dataset contains only the transcribed textual metaphors, the annotated source and target domains, and the source URLs of the images. Researchers wishing to reproduce our results on the human-generated subset can obtain the source images from their original providers under those providers' terms.

Reproducibility statement

We have aimed to make PRISM reproducible from the paper by including the algorithm for the pullback score computation in Algorithm 1 and Algorithm 2 and the prompts for the domain graph extraction pipeline in Section A. Section 4 furthermore specifies the sentence encoder used for the edge colouring and the used hyperparameter values are defined in Section 5 and Section H. Our code is published at `https://github.com/doubleblind-frog/iclr-2027` and the dataset of textual metaphors with manual annotations can be found at `https://doi.org/10.5281/zenodo.21891311`.

However, there are also a few factors that limit the reproducibility of the results. First of all, every stage of our pipeline is probabilistic, because image generation offers no seed control and domain-graph extraction varies across repeated calls on the same image. Additionally, most of our experiments use API-based proprietary models, which might be deprecated in favour of newer versions.

REFERENCES

Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alexandros G Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. GEPA: Reflective Prompt Evolution Can Outperform Reinforcement Learning. In *First Workshop on Foundations of Reasoning in Language Models*, 2025.

Arjun R. Akula, Brendan Driscoll, Pradyumna Narayana, Soravit Changpinyo, Zhiwei Jia, Suyash Damle, Garima Pruthi, Sugato Basu, Leonidas Guibas, William T. Freeman, Yuanzhen Li, and Varun Jampani. MetaCLUE: Towards Comprehensive Visual Metaphors Research. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23201–23211, Vancouver, BC, Canada, June 2023. IEEE. ISBN 979-8-3503-0129-8. doi: 10.1109/CVPR52729.2023.02222. URL https://ieeexplore.ieee.org/document/10204033/.

Anthropic. Claude Haiku 4.5, 2025. URL https://www.anthropic.com/claude/haiku.

Anthropic. Claude Opus 4.6, 2026a. URL https://www.anthropic.com/news/claude-opus-4-6.

Anthropic. Introducing Sonnet 4.6, 2026b. URL https://www.anthropic.com/news/claude-sonnet-4-6.

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. HealthBench: Evaluating Large Language Models Towards Improved Human Health, May 2025. URL http://arxiv.org/abs/2505.08775. arXiv:2505.08775 [cs].

Steve Awodey. *Category Theory*. OUP Oxford, June 2010. ISBN 978-0-19-161255-8. Google-Books-ID: zLs8BAAAQBAJ.

Black Forest Labs. black-forest-labs/FLUX.2-klein-4B · Hugging Face, April 2026. URL https://huggingface.co/black-forest-labs/FLUX.2-klein-4B.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3):904–911, September 2014. ISSN 1554-3528. doi: 10.3758/s13428-013-0403-5. URL https://doi.org/10.3758/s13428-013-0403-5.

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 Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 7370–7388, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.465. URL https://aclanthology.org/2023.findings-acl.465/.

Zheng Chen, Zhaoxin Feng, Jianfei Ma, Jiexi Xu, and Bo Li. Can LLMs Recognize Their Own Analogical Hallucinations? Evaluating Uncertainty Estimation for Analogical Reasoning. In Yuji Zhang, Canyu Chen, Sha Li, Mor Geva, Chi Han, Xiaozhi Wang, Shangbin Feng, Silin Gao, Isabelle Augenstein, Mohit Bansal, Manling Li, and Heng Ji (eds.), *Proceedings of the 3rd Workshop on Towards Knowledgeable Foundation Models (KnowFM)*, pp. 84–93, Vienna, Austria, August 2025. Association for Computational Linguistics. ISBN 979-8-89176-283-1. doi: 10.18653/v1/2025.knowllm-1.8. URL https://aclanthology.org/2025.knowllm-1.8/.

Jaemin Cho, Abhay Zala, and Mohit Bansal. DALL-EVAL: Probing the Reasoning Skills and Social Biases of Text-to-Image Generation Models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3020–3031, Paris, France, October 2023. IEEE. ISBN 979-8-3503-0718-4. doi: 10.1109/ICCV51070.2023.00283. URL https://ieeexplore.ieee.org/document/10378621/.

Yongkang Du, Xiaohan Zou, Minhao Cheng, and Lu Lin. CARV: A Diagnostic Benchmark for Compositional Analogical Reasoning in Multimodal LLMs, March 2026. URL https://arxiv.org/abs/2603.27958. arXiv:2603.27958 [cs]. Preprint, under review.

Richard Evans and Edward Grefenstette. Learning Explanatory Rules from Noisy Data. *Journal of Artificial Intelligence Research*, 61:1–64, January 2018. ISSN 1076-9757. doi: 10.1613/jair.5714. URL https://www.jair.org/index.php/jair/article/view/11172.

Thomas G. Evans. *A Program for the Solution of a Class of Geometric-analogy Intelligence-test Questions*. Air Force Cambridge Research Laboratories, Office of Aerospace Research, United States Air Force, 1964. Google-Books-ID: yhxaKcSU2HIC.

Brian Falkenhainer, Kenneth D. Forbus, and Dedre Gentner. The structure-mapping engine: Algorithm and examples. *Artificial Intelligence*, 41(1):1–63, November 1989. ISSN 0004-3702. doi: 10.1016/0004-3702(89)90077-5. URL https://www.sciencedirect.com/science/article/pii/0004370289900775.

Charles J. Fillmore. The case for case. *Universals in linguistic theory*, pp. 1–88, 1968. URL https://www.scirp.org/reference/referencespapers?referenceid=918107.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. PAL: Program-aided Language Models. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 10764–10799. PMLR, July 2023. URL https://proceedings.mlr.press/v202/gao23f.html.

Gemma. Gemma 4, 2026. URL https://deepmind.google/models/gemma/gemma-4/.

Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2): 155–170, April 1983. ISSN 0364-0213. doi: 10.1016/S0364-0213(83)80009-3. URL https://www.sciencedirect.com/science/article/pii/S0364021383800093.

Dedre Gentner and Christian Hoyos. Analogy and Abstraction. *Topics in Cognitive Science*, 9(3):672–693, 2017. ISSN 1756-8765. doi: 10.1111/tops.12278. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/tops.12278. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/tops.12278.

Tanmay Gupta and Aniruddha Kembhavi. Visual Programming: Compositional visual reasoning without training. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14953–14962, June 2023. doi: 10.1109/CVPR52729.2023.01436. URL https://ieeexplore.ieee.org/abstract/document/10204174. ISSN: 2575-7075.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.595. URL https://aclanthology.org/2021.emnlp-main.595/.

Douglas R. Hofstadter. The Copycat Project: An Experiment in Nondeterminism and Creative Analogies. Technical report, Massachusetts Institute of Technology The Artificial Intelligence Laboratory, January 1984. URL https://apps.dtic.mil/sti/html/tr/ADA142744/. Number: AIM755.

Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. Automatic Understanding of Image and Video Advertisements. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1100–1110, Honolulu, HI, July 2017. IEEE. ISBN 978-1-5386-0457-1. doi: 10.1109/CVPR.2017.123. URL http://ieeexplore.ieee.org/document/8099606/.

EunJeong Hwang and Vered Shwartz. MemeCap: A Dataset for Captioning and Interpreting Memes. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 1433–1445, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.89. URL https://aclanthology.org/2023.emnlp-main.89/.

Ideogram. Ideogram.ai, 2025. URL https://ideogram.ai.

Panic Jencel. Category Theory Illustrated, 2021. URL https://abuseofnotation.github.io/category-theory-illustrated/.

Mohammadhossein Khojasteh, Yifan Jiang, Stefano De Giorgis, Frank van Harmelen, and Filip Ilievski. Enhancing Structural Mapping with LLM-derived Abstractions for Analogical Reasoning in Narratives, March 2026. URL http://arxiv.org/abs/2603.29997. arXiv:2603.29997 [cs].

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. A large-scale classification of English verbs. *Language Resources and Evaluation*, 42(1):21–40, March 2008. ISSN 1572-8412. doi: 10.1007/s10579-007-9048-2. URL https://doi.org/10.1007/s10579-007-9048-2.

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. Poster presented at the NeurIPS 2025 Workshop on Generative and Protective AI for Content Creation (GenProCC), August 2025. URL http://arxiv.org/abs/2508.18569. arXiv:2508.18569 [cs].

Manishit Kundu, Sumit Shekhar, and Pushpak Bhattacharyya. Looking Beyond the Pixels: Evaluating Visual Metaphor Understanding in VLMs. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 23137–23158, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.1257. URL https://aclanthology.org/2025.findings-emnlp.1257/.

Manishit Kundu, Tejomay Kishor Padole, Sumit Shekhar, Biplab Banerjee, and Pushpak Bhattacharyya. CaRVE: Critiquing and Refining Visual Elaborations for Figurative Language Illustrations. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 42760–42777, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.2122. URL https://aclanthology.org/2026.findings-acl.2122/.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253, 2017. ISSN 0140-525X, 1469-1825. doi: 10.1017/S0140525X16001837. URL https://www.cambridge.org/core/product/identifier/S0140525X16001837/type/journal_article.

Martha Lewis and Melanie Mitchell. Evaluating the Robustness of Analogical Reasoning in Large Language Models, November 2024. URL http://arxiv.org/abs/2411.14215. arXiv:2411.14215 [cs].

Chen Ling, Tanmoy Chowdhury, Junji Jiang, Junxiang Wang, Xuchao Zhang, Haifeng Chen, and Liang Zhao. DeepGAR: Deep Graph Learning for Analogical Reasoning. In *2022 IEEE International Conference on Data Mining (ICDM)*, pp. 1065–1070, November 2022. doi: 10.1109/ICDM54844.2022.00132. URL https://ieeexplore.ieee.org/abstract/document/10027755. ISSN: 2374-8486.

Andrew Lovett and Kenneth Forbus. Modeling visual problem solving as analogical reasoning. *Psychological Review*, 124(1):60–90, 2017. ISSN 1939-1471, 0033-295X. doi: 10.1037/rev0000039. URL https://doi.apa.org/doi/10.1037/rev0000039.

Gary F. Marcus. *The Algebraic Mind: Integrating Connectionism and Cognitive Science*. MIT Press, January 2003. ISBN 978-0-262-63268-3. Google-Books-ID: 7YpuRUlFLm8C.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic Regularities in Continuous Space Word Representations. In Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff (eds.), *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 746–751, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL https://aclanthology.org/N13-1090/.

Melanie Mitchell. Abstraction and analogy-making in artificial intelligence. *Annals of the New York Academy of Sciences*, 1505(1):79–101, 2021. ISSN 1749-6632. doi: 10.1111/nyas.14619. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/nyas.14619. _eprint: https://nyaspubs.onlinelibrary.wiley.com/doi/pdf/10.1111/nyas.14619.

Sam Musker, Alex Duchnowski, Raphaël Millière, and Ellie Pavlick. LLMs as models for analogical reasoning. *Journal of Memory and Language*, 145:104676, December 2025. ISSN 0749596X. doi: 10.1016/j.jml.2025.104676. URL https://linkinghub.elsevier.com/retrieve/pii/S0749596X25000695.

Theo Olausson, Alex Gu, Ben Lipkin, Cedegao Zhang, Armando Solar-Lezama, Joshua Tenenbaum, and Roger Levy. LINC: A Neurosymbolic Approach for Logical Reasoning by Combining Language Models with First-Order Logic Provers. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5153–5176, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.313. URL https://aclanthology.org/2023.emnlp-main.313/.

OpenAI. GPT Image 1.5 Model | OpenAI API, 2025a. URL https://developers.openai.com/api/docs/models/gpt-image-1.5.

OpenAI. Introducing GPT-5.2, 2025b. URL https://openai.com/index/introducing-gpt-5-2/.

OpenAI. GPT-6 Astra. Large language model, 2026. URL https://openai.com/index/gpt-6-astra/. Accessed: 24 September 2026.

Gustaw Opiełka, Hannes Rosenbusch, and Claire E. Stevenson. Analogical Reasoning Inside Large Language Models: Concept Vectors and the Limits of Abstraction, March 2025. URL http://arxiv.org/abs/2503.03666. arXiv:2503.03666 [cs].

Claire Ott. Structural Similarity: Formalizing Analogies Using Category Theory. *Logics*, 3(4):12, December 2025. ISSN 2813-0405. doi: 10.3390/logics3040012. URL https://www.mdpi.com/2813-0405/3/4/12.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Wang. Logic-LM: Empowering Large Language Models with Symbolic Solvers for Faithful Logical Reasoning. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 3806–3824, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.248. URL https://aclanthology.org/2023.findings-emnlp.248/.

Niv Pekar, Yaniv Benny, and Lior Wolf. Generating Correct Answers for Progressive Matrices Intelligence Tests. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7390–7400. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/52cf49fea5ff66588408852f65cf8272-Abstract.html.

Chengwei Qin, Wenhan Xia, Tan Wang, Fangkai Jiao, Yuchen Hu, Bosheng Ding, Ruirui Chen, and Shafiq Joty. Relevant or Random: Can LLMs Truly Perform Analogical Reasoning? In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 23993–24010, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1230. URL https://aclanthology.org/2025.findings-acl.1230/.

Qwen. Qwen Studio, February 2026. URL https://qwenlm.github.io/blog/qwen3.5/. Section: blog.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763. PMLR, July 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

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. URL https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html.

Fereshteh Sadeghi, C. Lawrence Zitnick, and Ali Farhadi. Visalogy: Answering Visual Analogy Questions. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/hash/45f31d16b1058d586fc3be7207b58053-Abstract.html.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language Models Can Teach Themselves to Use Tools. *Advances in Neural Information Processing Systems*, 36:68539–68551, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/d842425e4bf79ba039352da0f658a906-Abstract-Conference.html.

Stephan Schmitz. NIH promises to create conflict-of-interest database for scientists, but offers few details, 2025. URL https://www.science.org/content/article/nih-promises-create-conflict-interest-database-scientists-offers-few-details.

Stability AI. Stability AI Core Models, 2024. URL https://stability.ai/core-models.

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, February 2026. URL http://arxiv.org/abs/2602.01335. arXiv:2602.01335 [cs].

Xiao Ye, Andrew Wang, Jacob Choi, Yining Lu, Shreya Sharma, Lingfeng Shen, Vijay Murari Tiyyala, Nicholas Andrews, and Daniel Khashabi. AnaloBench: Benchmarking the Identification of Abstract and Long-context Analogies. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 13060–13082, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.725. URL https://aclanthology.org/2024.emnlp-main.725/.

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/hash/5e388103a391daabe3de1d76a6739ccd-Abstract.html.

Nilay Yilmaz, Maitreya Patel, Yiran Lawrence Luo, Tejas Gokhale, Chitta Baral, Suren Jayasuriya, and Yezhou Yang. VOILA: Evaluation of MLLMs For Perceptual Understanding and Analogical Reasoning. In *Proceedings of the International Conference on Learning Representations (ICLR 2025)*, April 2025. URL https://openreview.net/forum?id=q5MUMlHxpd.

Eunice Yiu, Maan Qraitem, Anisa N. Majhi, Charlie Wong, Yutong Bai, Shiry Ginosar, Alison Gopnik, and Kate Saenko. KiVA: Kid-inspired Visual Analogies for Testing Large Multimodal Models. In *International Conference on Learning Representations*, pp. 26900–26919, October 2024. URL https://openreview.net/forum?id=vNATZfmY6R.

Siyu Yuan, Jiangjie Chen, Xuyang Ge, Yanghua Xiao, and Deqing Yang. Beneath Surface Similarity: Large Language Models Make Reasonable Scientific Analogies after Structure Abduction. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 2446–2460, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.160. URL https://aclanthology.org/2023.findings-emnlp.160/.

Linhao Zhang, Jintao Liu, Li Jin, Hao Wang, Kaiwen Wei, and Guangluan Xu. GOME: Grounding-based Metaphor Binding With Conceptual Elaboration For Figurative Language Illustration. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18500–18510, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1028. URL https://aclanthology.org/2024.emnlp-main.1028/.

Ningyu Zhang, Lei Li, Xiang Chen, Xiaozhuan Liang, Shumin Deng, and Huajun Chen. Multimodal Analogical Reasoning over Knowledge Graphs, March 2023. URL http://arxiv.org/abs/2210.00312. arXiv:2210.00312 [cs].

A PROMPTS

A.1 EXTRACTION PIPELINE PROMPTS

Prompt A.1: Stage 1: Domain Identification

STAGE 1 of dual-domain graph extraction from a visual metaphor.
YOU ARE GIVEN: a single image that depicts a visual metaphor. A visual metaphor combines two distinct conceptual domains:

- the TARGET domain (the abstract concept the metaphor is about) and
- the SOURCE domain (the concrete concept the target is being compared to).

For example, an image of a single man wreathed in fog is a visual metaphor where the TARGET is “loneliness" (abstract, what the image is about) and the SOURCE is “fog" (concrete, what loneliness is being compared to).
YOUR TASK: identify the two domains depicted.
YOUR OUTPUT WILL BE USED BY: a downstream pipeline that extracts a graph of nodes (entities) and edges (relations) for each domain independently, then computes a categorical pullback between the two graphs to score the structural strength of the metaphor. So the domain names you return must be:

- single canonical noun phrases, lowercase, no articles
- abstract enough to admit multiple instantiations (prefer “loneliness" over “the lonely old man’s feelings"; prefer “fog" over “the rolling grey morning fog"); however, if the domain clearly refers to a specific entity (e.g. “Eiffel Tower" or “Mercedes Benz"), return that as the domain name.
- No longer than 2-3 words (e.g. “social isolation" is fine, but “the feeling of being socially isolated in a crowd" is not)
- distinct from each other (target != source)

If the image does not clearly depict a visual metaphor (e.g. it is a literal photograph with no second domain implied), still return your best guess for what the dominant depicted concept is as the target, and “none" for the source. Be critical here: if you cannot clearly identify the two domains that the metaphor is trying to depict, return “none" for the source domain. The downstream pipeline will detect this.

Prompt A.2: Stage 2: Node Extraction

STAGE 2 of dual-domain graph extraction from a visual metaphor.
YOU ARE GIVEN:

- the same image as Stage 1
- the target domain and source domain identified in Stage 1

YOUR TASK: list the key entities (nodes) for each domain. Each node has two parts:

1. A ‘name‘: a concrete, image-grounded label for what this entity actually is in the depicted scene or in the abstract domain. Names should be specific to the domain and faithful to what is visible or genuinely implied by the image. Examples:
   - source domain “construction site": construction_worker, hammer, nail, concrete_block, steel_beam
   - target domain “creative work": writer, idea, manuscript, struggle, vision

   Names may be multi-word (snake_case). They should NOT be abstract role labels like “agent" or “instrument" — that’s what the ‘role‘ field is for.
2. A ‘role‘: a single label from this closed vocabulary, describing the structural function the entity plays in its own domain:{ROLE_VOCAB}. Pick the closest fit. Do not invent new role labels.

CRITICAL: role labels must describe what the entity does in ITS OWN DOMAIN. Do NOT choose role labels to make the two domains’ role lists look the same. If only the target domain has a clear ‘agent‘ and the source domain doesn’t, that’s fine — the asymmetry is informative. Forcing parallel role structures across domains defeats the purpose of the downstream structural-alignment scoring. Treat the two domains as independent role-typing exercises.
YOUR OUTPUT WILL BE USED BY: Stage 3 will see only the ‘name‘ fields (not the roles) when extracting relations, so that its reasoning is grounded in concrete image content rather than abstract structure.
Stage 4 will use the ‘role‘ fields to compute structural similarity between edges across domains.
NODE RULES:

- 4–7 nodes per domain.
- Names: concrete, image-grounded, snake_case, multi-word allowed. Be specific. Prefer “construction_worker" over “agent_figure"; prefer “rusty_nail" over “metal_object".
- Roles: from the closed vocabulary above, one per node.
- Names should be distinct within a domain. Roles MAY repeat within a domain (two ‘theme‘ nodes is fine if both genuinely play that structural part).

If a domain was identified as “none" in Stage 1, return an empty list for that domain.

Prompt A.3: Stage 3: Relation Extraction

STAGE 3 of dual-domain graph extraction from a visual metaphor.
YOU ARE GIVEN:

- the same image as Stage 1 and Stage 2
- the target and source domain names from Stage 1
- the per-domain node names from Stage 2 (CONCRETE names only, with role labels intentionally hidden)

YOUR TASK: for each domain, list the directed relations between its nodes. A relation is a triple (head, predicate, tail) where:

- head and tail are concrete node names from the Stage 2 list for that domain
- predicate captures a structural relationship that holds between the nodes within that domain

YOUR OUTPUT WILL BE USED BY: The categorical pullback algorithm, which compares the two graphs by finding a structure-preserving correspondence (between target_nodes and source_nodes) such that target_edges align with source_edges.
Important: when you write a relation, focus on what is actually happening in the depicted image (for the source domain) or what genuinely holds in the abstract concept (for the target domain). Do NOT try to make the two domains' relations mirror each other — the pullback step will discover any genuine parallels on its own.
RELATION RULES:

- Closed vocabulary ONLY: {RELATION_VOCAB}. Pick the closest match if nothing is exact; do not invent new predicates.
- 3–7 edges per domain. More than 7 risks adding unsupported edges.
- Both head and tail MUST be exact matches to names in the node list for that domain.
- Multiple edges between the same pair are allowed (different predicates).
- Do NOT add edges between target nodes and source nodes — each graph is intra-domain only. Cross-domain mapping happens in the pullback step, not here.
- Assess each domain independently based on the image and your reasoning, and define the relations between the nodes that are most salient for each independent domain. Do not coordinate across domains.

If a domain has an empty node list, return an empty edge list for it.

A.2 Iterative Refinement Prompt

Prompt A.4: Refinement with Feedback Prompt

You are refining the prompt given to a text-to-image model so that the image it produces is a STRONGER visual metaphor for a given idea.
A visual metaphor depicts an abstract TARGET concept by portraying it through a concrete SOURCE concept, such that the relationships among the source elements mirror the relationships in the target idea.
Here is the previous prompt you wrote and a diagnosis of how the image it produced fell short:
PREVIOUS PROMPT: {previous_elaboration}
DIAGNOSIS: {feedback}
Rewrite the prompt to fix the problems above. Specifically:

- Make BOTH the intended source and target concepts unmistakably present and recognisable in a single scene.
- Depict the relationships listed as missing, so the structure of the metaphor is visible, not just its surface objects.
- Keep the parts of the previous prompt that already worked; change only what the diagnosis flags.

Strict output rules:

- Output ONLY the new visual prompt itself.
- No preamble, no analysis, no headers, no markdown, no surrounding quotes, no commentary.
- Keep it under {MAX_WORDS} words (75 CLIP tokens).
- Write it as a single paragraph in natural prose.

B Can LLMs Extract and Populate Domain Graphs?

To ensure that PRISM's pullback score calculation pipeline (see Figure 3) is feasible, we first validated whether LLMs are able to complete steps (2) and (3), the actual extraction of a domain graph, for a specific domain. To test this behaviour, the SCAR benchmark (Yuan et al., 2023) was used, which contains analogous domains, their entities, and explanations of the relations between them.

Evaluating whether a given domain graph is valid or not is a difficult task. Therefore, we decided to follow a meta-evaluation similar to the one presented by Arora et al. (2025), in which a human-annotated binary response is used to test the accuracy of a given LLM-as-a-judge. We replicated this by selecting a random sample of 100 domains from the SCAR dataset and manually annotating these. Then, we altered half of the samples to produce incorrect domain graphs by altering or removing edges. Using this dataset, we assessed whether an LLM can act as a judge that can distinguish between valid and invalid domain graphs. Specifically, Claude Opus 4.6 (Anthropic, 2026a) was used as the judge LLM and achieved an accuracy of 88%, with an F1 score of 0.867. Furthermore, out of the 50 invalid domain graphs, only one was found by Claude Opus to be valid. Overall, Claude Opus 4.6 significantly outperforms the random baseline and errs conservatively — misclassifying valid graphs as invalid rather than the reverse — making it a suitable judge for this evaluation context.

Having established Claude Opus 4.6 as the judge, we then prompted Qwen 3.5 (Qwen, 2026), Gemma 4 (Gemma, 2026), and Claude Haiku 4.5 (Anthropic, 2025) to extract domain graphs for 100 randomly selected domains from the SCAR dataset, each with and without reasoning enabled. The results are shown in Table 3. Gemma 4 with reasoning enabled and Claude Haiku 4.5 with extended thinking performed the best with high accuracies above 80%. The results indicate that the quality of domain graph extraction does differ significantly between different models, but that there is the capability to extract valid domain graphs for various models, especially with reasoning modules enabled.

Table 3: Results for the domain graph extraction task on 100 samples of the SCAR dataset (Yuan et al., 2023), as judged by Claude Opus 4.6.

Model	Correctly generated graphs
Qwen 3.5	40%
Qwen 3.5 with reasoning	76%
Gemma 4	78%
Gemma 4 with reasoning	86%
Claude Haiku 4.5	56%
Claude Haiku 4.5 with extended thinking	84%

C Can the Pullback Score Provide Useful Signals in Analogical Reasoning?

This appendix provides extended analysis supporting the AnaloBench validation of the pullback score summarised in Section 4.1. AnaloBench (Ye et al., 2024) is a dataset of stories of various lengths, some of which are analogous. Given a narrative, the task (T1) is to correctly identify the analogous story from four given options (Ye et al., 2024). After determining that LLMs are able to extract valid domain graphs (see Section B), we used this task to configure the pullback score algorithm and its extraction prompts, and to validate whether the resulting score provides a useful signal for analogical reasoning, comparing a zero-shot baseline (GPT-4.1 Mini directly selecting the analogous story) against selecting the candidate with the highest pullback score.

We analysed the confidence of the pullback approach by examining the score margins between candidate answers. Figure 8 shows the distribution of the difference between the highest-ranked and second-highest-ranked pullback scores for correctly answered examples, alongside the difference between the highest-ranked score and the score assigned to the true answer for incorrectly answered examples. The margins for correct predictions were generally larger than those for incorrect predictions, suggesting that the pullback score expresses greater certainty when it identifies the correct analogy.

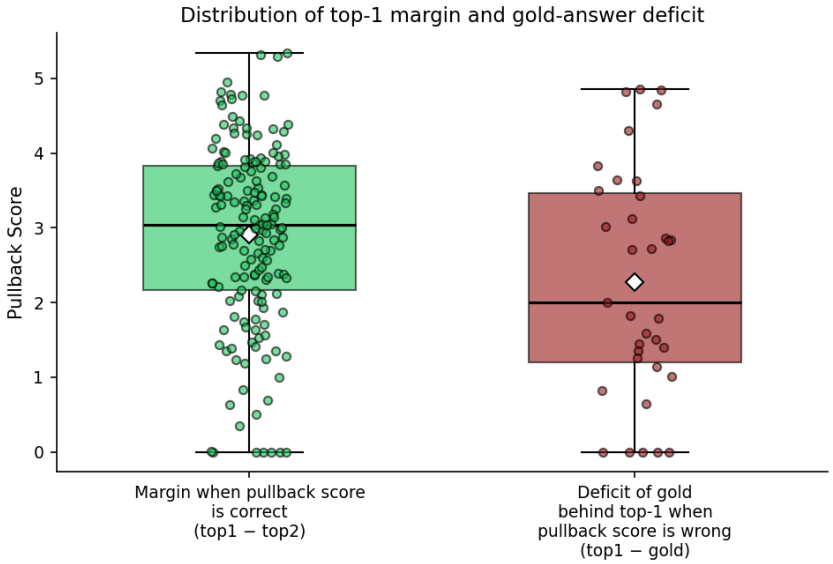


Figure 8: Confidence margins of the pullback score on the AnaloBench task: for correct predictions, the gap between the highest- and second-highest-scoring candidates; for incorrect predictions, the gap between the highest-scoring (wrong) candidate and the score of the true answer. Correct predictions show consistently larger margins than incorrect ones, indicating that the pullback score expresses greater certainty precisely when it identifies the correct analogy.

Qualitative inspection of the cases where the zero-shot approach succeeded while the pullback score approach failed suggested a potential bias towards denser and more highly connected graphs. In several of these cases, the pullback score assigned the highest score to an incorrect candidate whose extracted graph contained a larger number of (unmatched) relations than the graph corresponding to the correct answer. This observation highlights a potential limitation of the current scoring mechanism and suggests a direction for future refinement.

D Algorithm Description

This subroutine implements the matching step invoked by the pullback score computation introduced in Section 4 (Algorithm 1). GreedyPullback is a single-pass greedy heuristic: candidate edge pairs are sorted by similarity and accepted in order subject to the node-consistency check.

Algorithm 2 Greedy Semantic Pullback Matching

Require: Source edge set $E^S = \{(u_i, r_i, v_i)\}_{i=1}^m$, target edge set $E^T = \{(u'_j, r'_j, v'_j)\}_{j=1}^n$, precomputed embeddings $\hat{E}^S \in \mathbb{R}^{m\times d}$, $\hat{E}^T \in \mathbb{R}^{n\times d}$, matching threshold θ
Ensure: Pullback score ψ, matched pairs $\mathcal{P}$, coverage ρ

procedure GreedyPullback($E^S, E^T, \hat{E}^S, \hat{E}^T, \theta$)
  $\Phi \leftarrow \hat{E}^S(\hat{E}^T)^\top$ ▷ Cosine similarity matrix $\in \mathbb{R}^{m\times n}$
  $M \leftarrow \emptyset$ ▷ Node mapping: source node $\rightarrow$ target node
  $U_S \leftarrow \emptyset, U_T \leftarrow \emptyset$ ▷ Sets of already-matched edge indices
  $\mathcal{P} \leftarrow \emptyset$ ▷ Accepted edge pairs
  $C \leftarrow \{(\Phi_{ij}, i, j) \mid \Phi_{ij} \geq \theta\}$ sorted in descending order of Φ_{ij}
  for each $(\varphi, i, j) \in C$ **do**
    if $i \in U_S$ **or** $j \in U_T$ **then**
      continue ▷ Each edge may be matched at most once
    Let $e_i^S = (u_i, r_i, v_i)$ and $e_j^T = (u'_j, r'_j, v'_j)$
    $\Delta \leftarrow \{u_i \mapsto u'_j,\ v_i \mapsto v'_j\}$ ▷ Proposed node assignments for this pair
    if $\exists\, (a \mapsto b) \in M$ such that $a \in \mathrm{dom}(\Delta)$ and $\Delta(a) \neq b$ **then**
      continue ▷ Reject: violates global node consistency (pullback condition)
    $M \leftarrow M \cup \Delta$
    $U_S \leftarrow U_S \cup \{i\}, \quad U_T \leftarrow U_T \cup \{j\}$
    $\mathcal{P} \leftarrow \mathcal{P} \cup \{(i, j, \varphi)\}$
  end for
  $\sigma \leftarrow \sum_{(i,j,\varphi)\in\mathcal{P}} \varphi$ ▷ Total matched similarity, Equation (5)
  $\rho \leftarrow |\mathcal{P}| \,/\, |E^T|$ ▷ Coverage: fraction of target edges matched
  $\ell \leftarrow$ LongestSimplePath$\big(G^T[\mathcal{P}]\big)$ ▷ $G^T[\mathcal{P}]$: subgraph of G^T induced by matched edges; ℓ counted in edges
  $\psi \leftarrow \sigma + 0.5\,\ell$ ▷ Pullback score, Equation (5)
  return $\psi, \mathcal{P}, \rho$
end procedure

E User Study Instructions

Instructions: Evaluating AI-generated Visual Metaphors

You will be evaluating 5 pairs of images that have been generated for metaphorical sentences, e.g. "I am drowning in a sea of work". The images are supposed to be visual representations of these metaphors, with a focus on them depicting the underlying relations that give meaning to a metaphor. For each sentence, you will see two images.

For every image, you will need to evaluate the image on a scale from 0 to 10 for two questions, and then pick your preferred image of the two shown to you.

The scales that you need to evaluate the image on are metaphor consistency and analogy appropriateness.
Please base your ratings on the figurative meaning of the metaphors, not literal words, AI hallucinations or visual style.

Q1. Metaphor consistency: Does the given image maintain the fundamental creative logic and metaphorical structure from the metaphorical sentence it is based on? Does the image convey what the metaphor means, not just what it says?
Scoring Scale (0-10):
10: Perfect preservation - the core creative intent is identically transferred with the same logical structure
7-9: Strong preservation - clear continuity of creative logic with minor adaptations
4-6: Moderate preservation - some elements of the original intent remain but significant divergence
1-3: Weak preservation - only superficial similarities, core logic is different
0: No preservation - completely different creative approach

Figure 9: Part 1 of instructions provided to participants in the human evaluation.

Q2. Analogy appropriateness: Assess the quality and appropriateness of the analogy between the carrier (the metaphorical object used, e.g., "drowning") and the target (the actual subject, e.g., "work stress"). Are both of these domains clearly visible in the image? Is the relationship between them effectively communicated in the image? Is the analogical relationship intuitive to viewers without additional explanation?
Scoring Scale (0-10):
9-10: Exceptional - the analogy is highly appropriate with strong, multi-faceted correspondences that are both creative and well-justified
7-8: Strong - clear and meaningful analogy with solid grounding, minor room for improvement
5-6: Adequate - reasonable analogy that works but may lack depth or have some awkward aspects
3-4: Weak - forced or superficial analogy with limited meaningful correspondences
0-2: Inappropriate - arbitrary or nonsensical pairing with little justification

Q3. Pick the best image. From the two images shown, pick the image that better depicts the analogical mapping from the metaphorical sentence. If you find it hard to decide between the two images, pick the one that depicts the target and carrier domain more explicitly.

General guidelines:
Focus on meaning, not style. Ignore realism, lighting or beauty unless they prevent understanding. Don't penalise AI hallucinations like too many fingers on a hand unless it prevents understanding.
Text in images. Do not reward images simply because they repeat the sentence or domains within the sentence as text on the images. Imagery should carry the meaning.
Rate each image independently. Do not compare images across sentences.
If you do not understand the metaphor, you can search its meaning online.

Thank you for your careful ratings!

Figure 10: Part 2 of instructions provided to participants in the human evaluation.

F ZERO-SHOT GENERATION EXAMPLES

This appendix provides qualitative examples of the visual elaborations and generated images.

Table 4: Selected zero-shot generated visual metaphors. Each row shows the source metaphor, the model's visual elaboration, and the resulting generated image.

Metaphor	**Visual Elaboration**	**Generated Image (GPT Image 1.5)**
Love is a double edged sword.	A gleaming ornate sword suspended in mid-air, its blade splitting into two distinct halves—one radiating warm golden light and blooming roses, the other crackling with cold blue electricity and thorns. The background fades between serene twilight and stormy darkness. Fine mist swirls around the weapon, capturing the duality of beauty and pain.	
A sweet tooth is a predator chasing down your smile.	A sleek, shadowy predator with candy-colored fur stalks through a dreamlike landscape toward a luminous, golden smile floating in the distance. The creature's eyes glow with hunger as it prowls closer, leaving a trail of melting sweets and broken teeth. Soft, surreal lighting contrasts the predator's dark silhouette against pastel clouds and candy-striped terrain.	
The planet is a sinking ship.	A massive Earth sphere tilts precariously in turbulent waters, its continents cracking and flooding. Desperate figures cling to the edges as waves crash over the surface. Smoke rises from fissures. The sky darkens ominously. Lifeboats drift empty nearby, unreachable. The horizon swallows everything in murky depths, evoking apocalyptic urgency and collective doom.	
Social media is a hamster wheel.	An exhausted figure endlessly running inside a massive transparent hamster wheel, surrounded by glowing smartphone screens and notification badges. The wheel spins relentlessly in a dimly lit room, casting repetitive shadows. Despite constant motion, the scenery never changes. Scattered digital clutter accumulates around the base as the person runs faster, trapped in an endless cycle of movement without progress.	

G Extended Zero-Shot Results

This appendix reports extended results for zero-shot visual metaphors generated with Flux.2 Klein (Black Forest Labs, 2026), GPT-Image 1.5 (OpenAI, 2025a), and Stable Image Core v1.1 (Stability AI, 2024).

Metrics. We evaluate the generated images using domain fidelity, CLIP retrieval accuracy, and the pullback score. Domain fidelity measures the cosine similarity between the gold-standard source and target domains and the domains extracted from the generated image. CLIP retrieval accuracy (Radford et al., 2021; Hessel et al., 2021) measures consistency with the input metaphor by reporting how often the corresponding text or image is retrieved at rank 1 or within the top 5, following retrieval-based text-to-image evaluation protocols (Cho et al., 2023). The pullback score measures relational alignment between the domain graphs extracted from each image.

Generated images rarely depict both intended domains recognisably. As shown in Table 5, fewer than 10% of images recognisably depict both input domains under every condition, with GPT-Image zero-shot performing best at 9.4%. Either the source or target domain is recognisable more often, reaching 59.2% in the same condition. Although domain extraction admits multiple valid interpretations, these results indicate that simultaneously communicating both sides of a metaphor remains difficult for current image generators.

Table 5: Domain fidelity: percentage of metaphors for which the generated image recognisably depicts both or either input domain. Fidelity to both domains remains below 10% across all conditions.

Model	Both Domains Match	Either Domain Matches
GPT-Image Zero-Shot	**9.4%**	**59.2%**
GPT-Image Chain-of-Thought	6.5%	55.7%
Flux-Klein Zero-Shot	5.9%	54.2%
Stable-Core Chain-of-Thought	4.8%	44.0%
Stable-Core Zero-Shot	4.4%	52.0%
Flux-Klein Chain-of-Thought	4.2%	52.9%

GPT-Image zero-shot best preserves the input metaphor. CLIP retrieval produces a similar ranking. GPT-Image zero-shot achieves the highest retrieval accuracy in both directions, including 84% image-to-text Top-1 accuracy, compared with 59% for Stable-Core under chain-of-thought prompting (Table 6). Zero-shot prompting outperforms chain-of-thought prompting for all three generators, suggesting that the additional elaboration introduced by chain-of-thought prompting often causes the image to drift from the original metaphor. We therefore use GPT-Image zero-shot as the initial generation condition in the refinement experiments.

Table 6: CLIP retrieval accuracy between generated visual metaphors and their input texts. Higher Top-1 and Top-5 accuracy and lower mean rank indicate better retrieval.

	Image → Text			Text → Image		
Model	**Top-1 (%)**	**Top-5 (%)**	**Mean Rank**	**Top-1 (%)**	**Top-5 (%)**	**Mean Rank**
GPT-Image Zero	**84**	**96**	**1.57**	**82**	**96**	**2.54**
GPT-Image CoT	76	91	3.19	72	90	3.57
Flux-Klein Zero	76	92	2.60	67	87	4.50
Stable-Core Zero	69	87	4.22	61	82	6.35
Flux-Klein CoT	65	86	4.27	59	81	5.33
Stable-Core CoT	59	81	7.08	54	77	7.84

Relational structure can remain coherent even when the intended domains are not. Pullback scores exhibit less variation across conditions than either domain fidelity or CLIP retrieval. Scores are concentrated between 1 and 4, and comparisons against the human condition find no significant difference for any generator except Stable-Core zero-shot ($p_{\text{bonf}} = 0.040$). Human-created metaphors

nevertheless produce non-zero scores most consistently, at 94.7%, compared with 77.0–93.8% across the generated conditions.

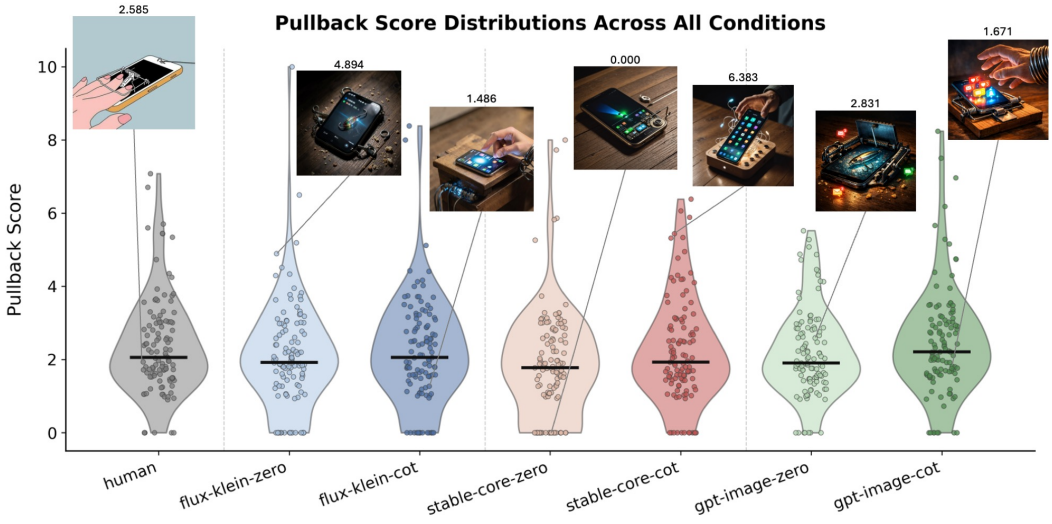


Figure 11: Distribution of pullback scores across six generation conditions and the human baseline. Annotated points show results for the metaphor "Your phone is a mouse trap".[2]

This result does not imply that generated images preserve the intended analogy as faithfully as human-created metaphors. The pullback score evaluates alignment between the domains extracted from an image, regardless of whether these are the intended domains. Taken together, the metrics therefore reveal a distinction between structural coherence and conceptual fidelity: generated images often contain a coherent relation between two depicted domains, but those domains frequently differ from the source and target specified by the input metaphor. PRISM addresses this gap by steering generation towards the particular relational structure intended by the input.

Table 7: Correlations between metaphor properties and CLIP retrieval accuracy. Sample sizes (n) are below the full 250 metaphors in some conditions because a subset of generations was blocked by the image generator's content moderation filters and could therefore not be scored. Significance is assessed using a Bonferroni-corrected threshold of $\alpha = 0.05/6 \approx 0.0083$ within each analysis. Significant p-values are shown in bold.

Analysis	**Condition**	n	**Pearson** r	p	**Spearman** ρ	p
Concreteness	Flux-Klein Zero	231	$+0.230$	**0.0004**	$+0.217$	**0.0009**
	Stable-Core Zero	243	$+0.186$	**0.0035**	$+0.171$	**0.0074**
	GPT-Image Zero	238	$+0.231$	**0.0003**	$+0.243$	**0.0002**
	Flux-Klein CoT	231	$+0.276$	$<$ **0.0001**	$+0.278$	$<$ **0.0001**
	Stable-Core CoT	243	$+0.208$	**0.0011**	$+0.191$	**0.0029**
	GPT-Image CoT	239	$+0.238$	**0.0002**	$+0.259$	**0.0001**
Nearness	Flux-Klein Zero	238	-0.064	0.3225	-0.066	0.3136
	Stable-Core Zero	250	-0.032	0.6145	-0.051	0.4242
	GPT-Image Zero	245	-0.131	0.0407	-0.136	0.0333
	Flux-Klein CoT	238	-0.065	0.3149	-0.074	0.2551
	Stable-Core CoT	250	-0.002	0.9805	-0.020	0.7528
	GPT-Image CoT	246	-0.042	0.5157	-0.050	0.4367

[2]Source of the human-created image for the metaphor "Your phone is a mouse trap": Pinterest (original creator unknown), `https://www.pinterest.com/pin/788833691001985793/`. Accessed 6 July 2026.

What makes a metaphor harder to depict? We additionally examine whether CLIP retrieval accuracy varies with two properties of the input metaphor: between-domain nearness and domain concreteness. Nearness is measured as the cosine similarity between the gold-standard source and target domains, with higher values indicating more similar domains. This analysis tests whether the advantage of near over far analogies previously observed in language-model reasoning (Khojasteh et al., 2026) also appears in visual metaphor generation.

Concreteness is additionally measured using the ratings of Brysbaert et al. (2014). Because a metaphor is constrained by its harder-to-visualise domain, we define its concreteness as the lower rating of its source and target domains, where ratings range from 1 for highly abstract concepts to 5 for highly concrete concepts.

As shown in Table 7, concreteness is positively associated with retrieval accuracy in every condition, with small but consistent correlations ($r = 0.186$–0.276). Metaphors are therefore easier to depict faithfully when both domains refer to readily visualisable concepts. By contrast, nearness shows no significant relationship with retrieval accuracy after correction for multiple comparisons. The difficulty of visual metaphor generation thus appears to depend more on whether the constituent domains can be rendered concretely than on their semantic similarity. Unlike language-based analogy selection, image generation does not receive a clear advantage from greater surface overlap between the domains.

H Model Selection & Hyperparameter Tuning

Refinement model selection. We conducted hyperparameter tuning experiments to select the most suitable model for generating the visual elaborations and the feedback metrics based on the pullback score extraction. In particular, we evaluated three state-of-the-art models: Gemma 4 31B, Claude Sonnet 4.6, and GPT 5.5.

Stopping criteria. Beyond model selection, we also wanted to investigate appropriate values for the stopping criteria (i.e. the maximum number of refinement rounds, k, and the convergence threshold, ε). In principle, the refinement loop could continue until a target pullback score is reached. In practice, however, computational and financial constraints make an unbounded number of iterations infeasible. Consequently, we treated both k and ε as tunable hyperparameters. We considered $k \in \{5, 10, 15\}$ and $\varepsilon \in \{0.05, 0.1, 0.2\}$.

Table 8: Hyperparameter tuning results for different visual elaboration models and stopping criteria. Scores correspond to the average ratings assigned by the three independent VLM judges.

Model	k	ε	Mean Pullback	Metaphor Consistency	Analogy Appropriateness	Conceptual Integration
Gemma 4 31B	5	0.05	2.8874	8.1534	8.6399	8.6667
Gemma 4 31B	5	0.10	2.8100	8.0467	8.5933	8.6000
Gemma 4 31B	5	0.20	2.8064	8.0467	8.5933	8.6134
Gemma 4 31B	10	0.05	3.1357	7.9800	8.4800	8.6200
Gemma 4 31B	10	0.10	3.0188	7.9000	8.4400	8.5667
Gemma 4 31B	10	0.20	2.9070	7.9400	8.4533	8.5867
Gemma 4 31B	15	0.05	3.1851	7.9800	8.5266	8.6334
Gemma 4 31B	15	0.10	3.0365	7.9000	8.4866	8.5934
Gemma 4 31B	15	0.20	2.9202	7.9267	8.5066	8.6267
Claude Sonnet 4.6	5	0.05	3.6607	8.1333	8.7266	8.7201
Claude Sonnet 4.6	5	0.10	3.6128	8.1866	8.7266	**8.7334**
Claude Sonnet 4.6	5	0.20	3.5388	8.1199	8.6866	8.6934
Claude Sonnet 4.6	10	0.05	4.1000	8.1599	8.7266	8.7201
Claude Sonnet 4.6	10	0.10	3.9788	**8.2133**	**8.7333**	8.7001
Claude Sonnet 4.6	10	0.20	3.8795	8.1599	8.6866	8.6734
Claude Sonnet 4.6	15	0.05	**4.1079**	8.1599	8.7000	8.7134
Claude Sonnet 4.6	15	0.10	3.9867	**8.2133**	8.7067	8.6934
Claude Sonnet 4.6	15	0.20	3.8874	8.1599	8.6600	8.6668
GPT-5.5	5	0.05	2.7944	7.7332	8.3916	8.0832
GPT-5.5	5	0.10	2.7231	7.6666	8.3083	8.1499
GPT-5.5	5	0.20	2.3086	7.6666	8.2250	8.1666
GPT-5.5	10	0.05	3.2636	7.1998	7.9416	7.5667
GPT-5.5	10	0.10	3.0258	7.1665	7.8916	7.6334
GPT-5.5	10	0.20	2.4495	7.6999	8.2750	8.1834
GPT-5.5	15	0.05	3.2636	7.1998	7.9416	7.5667
GPT-5.5	15	0.10	3.0258	7.1665	7.8916	7.6334
GPT-5.5	15	0.20	2.4495	7.6999	8.2750	8.1834

Calibration evaluation. Performing human evaluations for every configuration was impractical, so a panel of VLMs were used as judges instead. The VLM judge models used were Gemini 3.5 Flash, Claude Opus 4.8 and Llama-4 Maverick.[3]

This analysis was performed on a randomly selected holdout set of 25 visual metaphors, which is disjoint from the 200-metaphor evaluation set reported in Section 6. The results are shown in Table 8. Overall, Claude Sonnet 4.6 consistently outperformed Gemma 4 31B and GPT-5.5 across all evaluation settings and was therefore selected as the visual elaboration generator and feedback model.

The highest pullback score was achieved by Claude Sonnet 4.6 with $k = 15$ and $\varepsilon = 0.05$. However, this configuration only marginally outperformed the $k = 10$, $\varepsilon = 0.05$ setting (4.1079 vs. 4.1000), while requiring 50% more refinement iterations. Since the additional gains in both pullback score and VLM-based evaluation metrics were negligible across all configurations with $k \geq 10$, we selected $k = 10$ as a more computationally efficient operating point.

Although $\varepsilon = 0.10$ obtains marginally higher Metaphor Consistency and Analogy Appropriateness than $\varepsilon = 0.05$ at $k = 10$ (8.2133 vs. 8.1599, and 8.7333 vs. 8.7266), these differences are small relative to the noise expected from scoring only 25 holdout metaphors with the preliminary three-judge panel, whereas $\varepsilon = 0.05$ achieves a larger, more decisive margin on Conceptual Integration (8.7201 vs. 8.7001) and on the mean pullback score (4.1000 vs. 3.9788). Since ε is defined as the convergence threshold on the pullback score itself rather than on the downstream judge scores, we prioritised this larger, directly targeted margin over marginal, likely noise-level differences on axes ε was not selected to optimise.

Pullback matching threshold. The pullback score uses a cosine-similarity threshold, θ, to determine whether source- and target-domain relation edges are sufficiently similar to be matched. We calibrated θ on the pullback scores of the 125 original, human-authored visual metaphor images themselves by varying θ from 0.35 to 0.55 and measuring the resulting mean pullback score, the proportion of zero scores, and correlation with CLIP image-text similarity as an independent measure of image–metaphor alignment. This calibration therefore did not score any zero-shot or refined generated image: although the same underlying metaphors also appear in the 200-metaphor evaluation set of Section 6, no generated image scored there was used to select θ.

Table 9: Calibration of the pullback matching threshold θ on 125 human-authored visual metaphors. % Zero denotes the proportion of images for which no structural correspondence is recovered.

θ	Mean Pullback	% Zero	Spearman vs. CLIP
0.35	2.3144	2.4%	0.0753
0.40	**2.2305**	**3.2%**	0.0578
0.45	2.0538	7.2%	0.0690
0.55	1.4883	26.4%	-0.0242

As θ increases, matching becomes more restrictive: mean pullback scores decrease and zero scores become substantially more frequent, reaching 26.4% at $\theta = 0.55$. This is especially informative as these results are measures on a dataset of 125 real visual metaphors, so there are relational correspondences that should be identified in all provided images. Correlation with CLIP is weak across all settings, consistent with CLIP measuring holistic image–text similarity rather than relational correspondence. We therefore select $\theta = 0.40$ as a conservative operating point that retains almost all non-zero structural matches while avoiding the more permissive matching induced by $\theta = 0.35$.

[3]This calibration run used the three-judge panel in place at the time. The judging panel used for all subsequent evaluations reported in this paper is the two-judge panel described in Section 5.

I FULL QUALITATIVE COMPARISONS

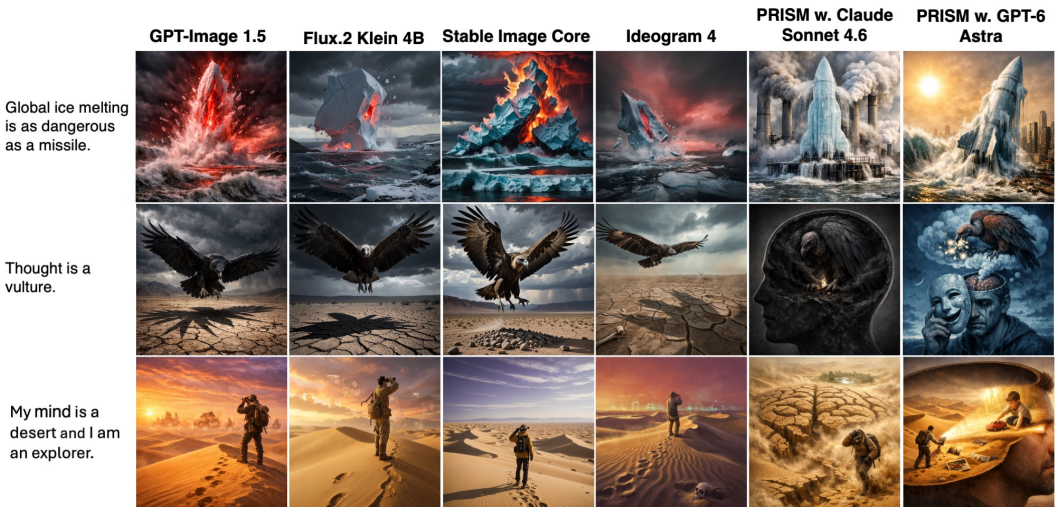


Figure 12: Comparison of four zero-shot generations pipelines with GPT-Image 1.5, Flux.2 Klein 4B, Stable Image Core and Ideogram 4 versus two versions of PRISM with Claude Sonnet 4.6 and GPT-6 Astra as the extraction and refinement models.