arXiv:2602.01335v2 [cs.CV] 27 Aug 2026

Beyond Pixels: Visual Metaphor Transfer via Schema-Driven Agentic Reasoning

YU XU*, YUXIN ZHANG*, and LIN GAO, University of Chinese Academy of Sciences, China
OLIVER DEUSSEN, University of Konstanz, Germany
TONG-YEE LEE, National Cheng Kung University, Taiwan
FAN TANG†, University of Science and Technology Beijing, China

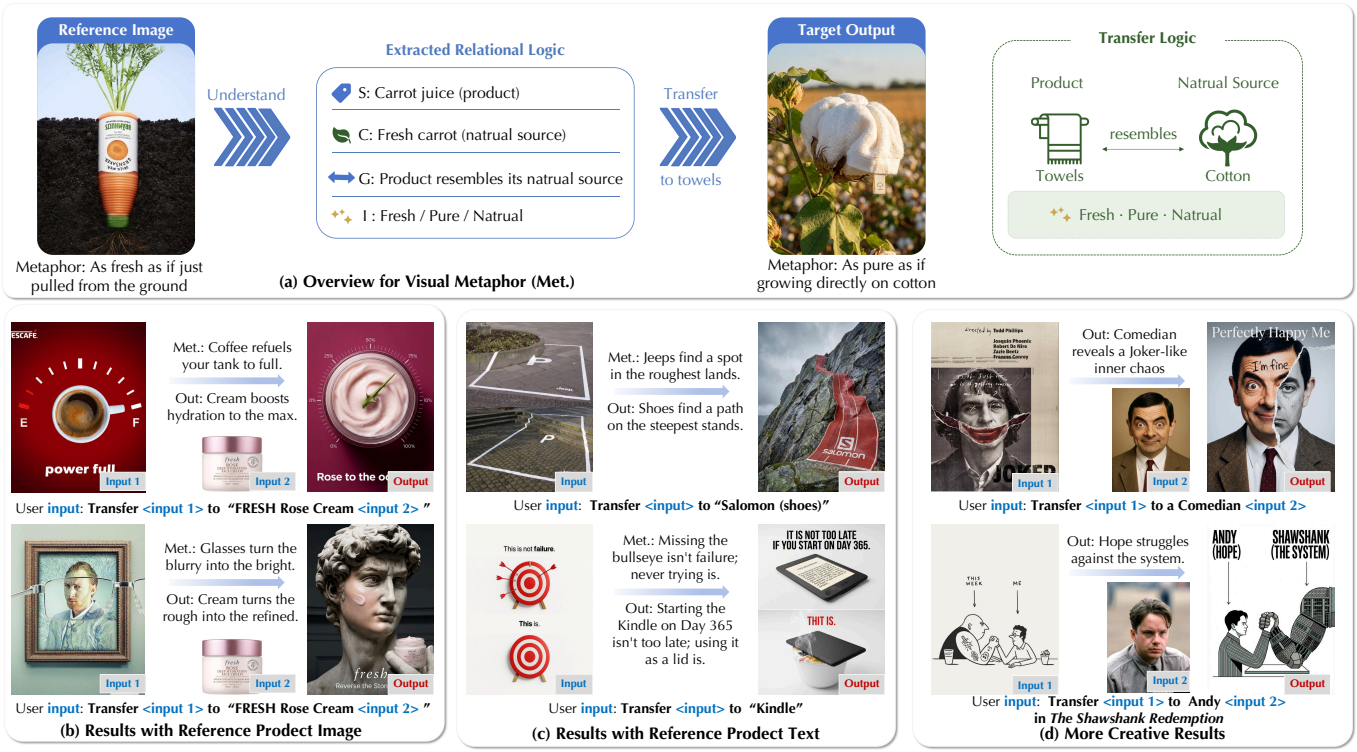

Fig. 1. Diverse image metaphor transfer results generated by our framework. For each pair, the left image serves as the *Reference* and the right is the *Generated Result*. Our model demonstrates robust capability across distinct cognitive levels.

A visual metaphor constitutes a high-order form of human creativity, employing cross-domain semantic fusion to transform abstract concepts into impactful visual rhetoric. Despite the remarkable progress of generative AI, existing models remain largely confined to pixel-level instruction alignment and surface-level appearance preservation, failing to capture the underlying abstract logic necessary for genuine metaphorical generation. To bridge this gap, we introduce the task of Visual Metaphor Transfer (VMT), which challenges models to autonomously decouple the "creative essence" from a reference image and re-materialize that abstract logic onto a user-specified target subject. We propose a cognitive-inspired, multi-agent framework that operationalizes Conceptual Blending Theory (CBT) through a novel Schema Grammar ($\mathcal{SG}$). This structured representation decouples relational invariants from specific visual entities, providing a rigorous foundation for cross-domain logic re-instantiation. Our pipeline executes VMT through a collaborative system of specialized agents: a perception agent that distills the reference into a schema, a transfer agent that maintains generic space invariance to discover apt carriers, a generation agent for high-fidelity synthesis and a hierarchical diagnostic agent that mimics a professional critic, performing closed-loop backtracking to identify and rectify errors across abstract logic, component selection, and prompt encoding. Extensive experiments and human evaluations demonstrate that our method significantly outperforms state-of-the-art baselines in metaphor consistency, analogy appropriateness, and visual creativity, paving the way for automated high-impact creative applications in advertising and media. Project page with source code and self-contained skills is https://yuci-gpt.github.io/Beyond-Pixels/.

CCS Concepts: • **Computing methodologies** → **Image manipulation**; *Image processing*.

Additional Key Words and Phrases: image generation, image reasoning, agent

*Equal contribution.
†Corresponding author: Fan Tang.

Authors' Contact Information: Yu Xu; Yuxin Zhang; Lin Gao, yu.xu915@gmail.com, yuxin.zhazel@gmail.com, gaolin@ict.ac.cn, University of Chinese Academy of Sciences, China; Oliver Deussen, oliver.deussen@uni-konstanz.de, University of Konstanz, Germany; Tong-Yee Lee, tonylee@mail.ncku.edu.tw, National Cheng Kung University, Taiwan; Fan Tang, tfan.108@gmail.com, University of Science and Technology Beijing, China.

1 Introduction

Visual metaphor operates at the upper limits of human creative cognition, where meanings are constructed through the integration of disparate semantic domains, enabling abstract ideas to be conveyed as visually articulated statements with layered, non-literal significance. Despite the remarkable progress of generative AI, existing text-to-image (T2I) [Rombach et al. 2022; Podell et al. 2024; Esser et al. 2024; blackforestlabs.ai 2024; Bao et al. 2025] and image-to-image [Mou et al. 2025; Wu et al. 2025; Wei et al. 2025; Liu and Chang 2025; Liu et al. 2025b; Dai et al. 2025; OpenAI 2025] models remain largely confined to pixel-level instruction alignment and the preservation of surface-level visual appearance, such as style, texture, or subjects, failing to capture the underlying abstract logic necessary for genuine metaphorical generation. Transitioning from pixel-level reconstruction to metaphorical synthesis requires identifying deep-seated relational invariants across disparate domains and executing creative conceptual blending to induce emergent meaning. Lacking an innate perception of such creative logic, current models cannot distill the metaphorical essence from a reference image and adapt it flexibly to novel contexts as humans do.

Research on visual metaphors has traditionally evolved along two parallel trajectories: interpretation and synthesis. Multimodal Large Language Models (MLLMs) have demonstrated fundamental cognitive abilities for metaphor interpretation, yet they struggle to parse non-literal semantics and deep symbolic relationships embedded in complex visual rhetoric without additional information or prompts [Akula et al. 2023; Kundu et al. 2025]. Simultaneously, synthesis methods remain predominantly text-driven, relying on mapping linguistic metaphors onto concrete objects through extensive textual prompts [Chakrabarty et al. 2023; Sun et al. 2025]. Despite their respective advancements, both paradigms converge on a shared limitation: an over-reliance on explicit, user-provided textual descriptions. This dependency creates a critical technical barrier to a more sophisticated creative capability—the ability for autonomously decoupling the underlying metaphorical logic from a visual reference and fluidly re-instantiating it within a novel context.

To bridge this gap, we introduce the task of Visual Metaphor Transfer (VMT). Unlike conventional subject customization or style transfer that focuses on visual appearance, VMT necessitates deconstructing the "creative essence" from a reference image and re-materializing that abstract logic onto a user-specified target subject. This paradigm shift presents two formidable challenges that push the boundaries of current generative AI: (1) Explicit metaphor modeling, which requires distilling domain-independent relational invariants from raw pixels into a structured representation; (2) Autonomous carrier adaptation, which demands the retrieval of a novel visual vehicle that not only aligns with the target subject's attributes but also preserves the original cognitive tension to induce a fresh emergence of meaning. Addressing these challenges requires a transition from passive pixel synthesis to active, agentic visual reasoning.

To address these challenges, we propose a multi-agent framework for VMT, operationalizing Conceptual Blending Theory (CBT) [Fauconnier and Turner 1998, 2003] from cognitive linguistics into an executable computational paradigm. Central to our approach is the Schema Grammar ($\mathcal{SG}$), a novel structural representation that decouples abstract relational invariants from specific visual entities. By encoding the intricate interplay between subjects, carriers, and semantic violations, $\mathcal{SG}$ provides a rigorous foundation for cross-domain logic re-instantiation. Specifically, our framework executes VMT through a collaborative pipeline of specialized agents: (1) a *perception agent* that distills the reference image into a structured schema; (2) a *transfer agent* that maintains generic space invariance to autonomously discover contextually apt carriers for new subjects; and (3) a *generation agent* that translates these logic blueprints into structured prompts for high-fidelity synthesis. Crucially, we introduce a hierarchical backtracing mechanism within a *diagnostic agent*, which mimics a professional "critic" by identifying the root causes of failure across abstract logic, component selection, or prompt encoding. This closed-loop refinement ensures that the final output transcends mere pixel-level consistency to achieve profound logical alignment. We frame our approach primarily as a system contribution. Because end-to-end SOTA models lack the structured reasoning required for metaphor transfer, our core contribution lies in operationalizing Conceptual Blending Theory into a multi-agent, schema-driven workflow, thereby unlocking zero-shot creative capabilities that are fundamentally absent in base generative models. Our main contributions are summarized as follows:

- **Cognitive-Logic Formalization**: We operationalize conceptual blending theory (CBT) by proposing the "Schema Grammar" representation, providing a rigorous cognitive science foundation for cross-domain carrier matching and metaphor synthesis.
- **Closed-Loop Multi-Agent Framework**: We develop a collaborative system encompassing perception, transfer, generation, and diagnostics. Notably, the proposed hierarchical backtracing mechanism significantly enhances generation reliability for complex metaphorical tasks.
- **Superior Experimental Performance**: Our method outperforms existing baselines in terms of metaphor consistency, analogy appropriateness, and visual creativity.

2 Related Work

2.1 Visual Metaphor Understanding and Generation

Visual metaphors serve as powerful rhetorical devices that convey abstract concepts through symbolic imagery. Unlike subject customization [Ruiz et al. 2023; Gal et al. 2023; Avrahami et al. 2023; Xu et al. 2025] or style transfer [Zhang et al. 2022; Hertz et al. 2024; Xu et al. 2026] approaches that focus on preserving visual appearance, metaphor understanding and generation requires capturing abstract symbolic relationships that convey meaning beyond literal visual similarity. Classic work on analogy and structure-mapping [Gentner 1983] and its computational realization in the Structure-Mapping Engine (SME) [Falkenhainer et al. 1989] establish that cross-domain transfer hinges on shared relational structure rather than surface attributes—a principle that motivates our schema-level Generic Space G and informs subsequent CBT-based blending accounts. MetaCLUE [Akula et al. 2023] shows that vision-language models struggle with metaphor understanding compared to literal images. Building on this foundation, metaphor understanding works primarily focus on multimodal contexts, employing techniques including

linking metaphor text to visual concepts with prompting [Xu et al. 2024], concept drift mechanisms [Qian et al. 2025], and prompt optimizer with reinforcement learning [Fan et al. 2024] to detect and interpret metaphorical content. Complementing understanding approaches, recent works explore generating visual metaphors through text-driven synthesis. I-spy-a-metaphor [Chakrabarty et al. 2023] propose a human–LLM–diffusion collaboration framework for generating visual metaphors from linguistic metaphors. Creative Blends [Sun et al. 2025] proposes an AI-assisted system that uses commonsense knowledge and LLMs to map abstract concepts to concrete objects and blend them via T2I models. TIAC [Liao et al. 2024] proposes a framework that maps abstract concepts to clear intents and semantic objects via LLMs to generate concept-aligned images. Mind's eye [Koushik et al. 2025] further explore self-evaluating visual metaphor generation framework with reinforcement learning. VisiBlends [Chilton et al. 2019], MetaMap [Kang et al. 2021], and PopBlends [Wang et al. 2023] focus on supporting human conception through rule extraction or design-aided tools. However, these text-driven methods generate metaphors from linguistic input rather than learning from visual examples, requiring explicit textual specification of metaphorical concepts, which limits their ability to extract and transfer reusable metaphorical representations across different visual contexts. In contrast, our method analyzes the core semantic meaning of metaphors and leverages the generic space from conceptual blending theory to transfer metaphorical representations to new subjects, achieving high-fidelity image-driven metaphor transfer.

2.2 Image Generation with Multimodal LLMs

Recent advances in multimodal large language models [Li et al. 2022; Seedream et al. 2025; Cao et al. 2025; DeepMind 2025; OpenAI 2025] demonstrate remarkable capabilities in understanding and reasoning across vision and language modalities. Building on these foundations, many image generation tasks leverage multimodal LLMs to decompose complex generation objectives into specialized subtasks and employ multi-agent frameworks to extend their functionality [Shaham et al. 2024; Sandoval-Castaneda et al. 2025]. For instance, SketchAgent [Vinker et al. 2025] utilizes LLMs with in-context learning to generate SVG strings that are subsequently rendered into sketches. MCCD [Li et al. 2025] proposes a multi-agent scene parsing and hierarchical compositional diffusion framework to achieve image generation for complex multi-object prompts. However, these methods primarily adopt a sequential execution paradigm where agents operate in a feed-forward manner without retrospective analysis. In contrast, our method introduces a critic module that traces back to evaluate the output of each agent in previous steps and performs targeted refinement, enabling iterative improvement and achieving more faithful metaphor transfer that aligns with the intended symbolic meaning.

3 Computational Modeling of Visual Metaphors

To bridge human cognitive creativity and metaphorical transfer, we formalize the task by operationalizing conceptual blending theory (CBT). This section provides the theoretical foundation and structural representation for agentic reasoning.

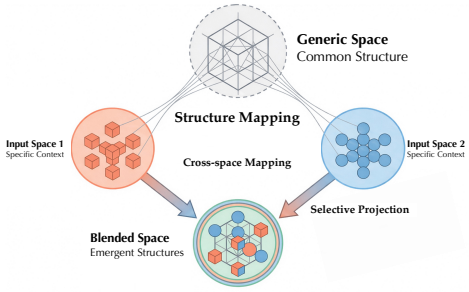

Fig. 2. Illustration for conceptual blending theory.

3.1 Conceptual Blending Spaces

Conceptual blending theory, proposed by Fauconnier and Turner [1998; 2003], posits that human creativity arises from integrating disparate mental spaces to generate novel meanings. A metaphor is not a simple linear mapping but a dynamic integration of four mental spaces, as illustrated in Fig. 2: (1) **Input Spaces** contain specific entities that provide raw content for the blend. Elements between these spaces are often linked by counterparts. (2) The **Generic Space** captures abstract, domain-independent relational invariants shared by both input spaces. It encodes the underlying logic (e.g., specific roles, frames, or schemata) that enables mapping between inputs. (3) The **Blended Space** is where elements from the inputs are selectively projected and integrated. Through composition, completion, and elaboration, this space gives rise to **emergent structures**–new relations and meanings absent from either input space alone.

3.2 Structured Representation of Visual Metaphoric

We bridge psychological operations and computational reasoning by mapping these spaces into a structured **schema grammar** ($\mathcal{SG}$). The 7-tuple is a direct operationalization of CBT [Fauconnier and Turner 1998, 2003]. We represent a visual metaphor as $\mathcal{SG} = \{S, C, A_S, A_{es}, G, V, I\}$, where each element maps onto a blending component (see Fig. 3 for a running example of each element):

- Entity instantiation ($\{S, C, A_S, A_{es}\}$): S and C instantiate the two Input Spaces—the subject (primary entity depicted) and the carrier (visual context or metaphorical vehicle). Visual metaphor embeds the subject into the carrier's domain. Inherent attributes A_S are the canonical properties of S in its original domain and provide the baseline for detecting violations. Expressive attributes A_{es} capture image-level rhetorical atmosphere (e.g., lighting, mood, composition); preserving A_{es} ensures the transferred image retains the reference's expressive tone.
- Relational bridging ($\{G\}$): The Generic Space G is the domain-independent relational invariant shared by S and C (functional, structural, relational, or emotional). Critically, G is a *constituent* of $\mathcal{SG}$, not a separate schema: $\mathcal{SG}_{\text{ref}}/\mathcal{SG}_{\text{tgt}}$ denote full schema grammars, while $G_{\text{ref}}/G_{\text{tgt}}$ denote only this relational bridging component.
- Synthesis operationalization ($\{V, I\}$): Selective projection and emergent structure in the Blended Space are realized through violation points V—semantic incongruities where S transgresses norms of C, derived by contrasting A_S with G—and Emergent

Meaning I, the high-level creative message induced by that cognitive tension.

Removing any component breaks this CBT mapping: without G there is no transferable invariant, without V no cognitive tension, and without I no target meaning. Our w/o CBT ablation (Sec. 5.5) confirms that this decomposition drives the observed gains. By decoupling relational logic (G, V, I) from visual entities (S, C) while retaining A_S and A_{es}, $\mathcal{SG}$ manipulates a metaphor's "creative essence" as a structured representation.

3.3 Task Formulation

Based on $\mathcal{SG}$, we formulate visual metaphor transfer as learning a mapping function $\mathcal{M}$ that migrates reference logic to a new subject. Given a reference schema $\mathcal{SG}_{\text{ref}}$ and target subject S_{tgt}, the framework must synthesize a target schema $\mathcal{SG}_{\text{tgt}}$ such that:

$$\begin{aligned} \mathcal{M}(\mathcal{SG}_{\text{ref}}, S_{\text{tgt}}) &\rightarrow \mathcal{SG}_{\text{tgt}}, \\ s.t., G_{\text{tgt}} &\equiv G_{\text{ref}}. \end{aligned} \tag{1}$$

A transfer succeeds if $\mathcal{SG}_{\text{tgt}}$ preserves the reference's abstract relational logic G while autonomously discovering a novel carrier C_{tgt} and violations V_{tgt} contextually appropriate for the new subject. This formalization transforms VMT from pixel-level reconstruction into structured search and instantiation within the space of Schema Grammars.

4 Method

In this section, we present our framework for visual metaphor transfer. Our method decomposes metaphor transfer into four executable stages: (1) The **Perception Agent** extracts an explicit reference schema $\mathcal{SG}_{\text{ref}}$ from the input image. (2) The **Transfer Agent** takes $\mathcal{SG}_{\text{ref}}$ and the target subject S_{tgt}, preserves the relational invariant G, and synthesizes $\mathcal{SG}_{\text{tgt}}$ by selecting a new carrier and redesigning the violation points. (3) The **Generation Agent** converts $\mathcal{SG}_{\text{tgt}}$ into a detailed T2I prompt that grounds the subject, carrier, violation, emergent meaning, and aesthetic attributes. (4) The **Diagnostic Agent** evaluates the generated image and performs prompt-, component-, or abstraction-level backtracking to correct failures in visual realization, carrier selection, or schema extraction. The overall architecture is illustrated in Fig. 3. p

4.1 Perception Agent

The Perception Agent only analyzes the reference image, without adapting it to the target subject. Given I_{ref}, it extracts a structured schema $\mathcal{SG}_{\text{ref}} = \{S, C, A_S, A_{es}, G, V, I\}$, making the implicit metaphor explicit. This stage separates surface entities from abstract relational logic, which direct prompting often fails to do. We employ a Vision-Language Model (VLM) guided by chain-of-thought reasoning in the sequence:$S/C \rightarrow A_S \rightarrow A_{es} \rightarrow G \rightarrow V \rightarrow I$. The model first identifies the subject S and carrier C, then extracts the subject attributes A_S and expressive attributes A_{es}, such as layout, lighting, mood, and rhetorical atmosphere. It next isolates the Generic Space G, the domain-independent relation connecting S and C. By contrasting A_S with the norms implied by G and C, it derives the Violation Points V that create cognitive tension, and finally infers the Emergent Meaning I. Formally, this extraction process is:

$$\mathcal{SG}_{\text{ref}} = \text{VLM}(I_{\text{ref}}, p_{\text{extract}}), \tag{2}$$

where p_{extract} denotes the extraction-specific prompt. This decomposition turns an implicit creative concept into a manipulable reference schema. The Perception Agent does not determine target adaptation; it only extracts reusable logic from the reference. By isolating G while retaining S, C, A_S, A_{es}, V, and I, it provides the necessary input for the Transfer Agent to re-instantiate the same abstract logic with a new subject and carrier. A complete schema example is provided in the supplementary material.

4.2 Transfer Agent

Given the reference schema $\mathcal{SG}_{\text{ref}}$ and target subject S_{tgt}, the Transfer Agent synthesizes a target schema $\mathcal{SG}_{\text{tgt}}$ by preserving the abstract relational logic while re-grounding it in a new domain. Unlike the Perception Agent, which analyzes only the reference image, this stage performs target-side schema synthesis: selecting a new carrier, redesigning violation points, and adapting the emergent meaning to the target subject. This transfer is guided by VLM reasoning that keeps the Generic Space G invariant across domains.

a) Transfer Objective: The goal is to generate $\mathcal{SG}_{\text{tgt}} = \{S^{\text{tgt}}, C^{\text{tgt}}, A_S^{\text{tgt}}, A_{es}^{\text{tgt}}, G, V^{\text{tgt}}, I^{\text{tgt}}\}$, where G is preserved from $\mathcal{SG}_{\text{ref}}$, while C^{tgt}, A_S^{tgt}, A_{es}^{tgt}, V^{tgt}, and I^{tgt} are re-instantiated for the target domain. This retains the reference metaphor's relational logic while realizing it through a target-specific visual configuration.

b) Reasoning process: We prompt the VLM to perform relational transfer through: $G \rightarrow A_S^{\text{tgt}} \rightarrow C^{\text{tgt}} \rightarrow V^{\text{tgt}} \rightarrow I^{\text{tgt}}$. It first analyzes the domain-independent nature of G, then profiles the inherent attributes and functional or symbolic roles of S_{tgt}. Next, it searches for a cross-domain carrier C^{tgt} sharing the same Generic Space G, designs violation points V^{tgt} that mirror the reference violation logic, and aligns I^{tgt} with the target context.

This transfer process is formalized as:

$$\mathcal{SG}_{\text{tgt}} = \text{VLM}(\mathcal{SG}_{\text{ref}}, S_{\text{tgt}}, p_{transfer}), \tag{3}$$

where $p_{transfer}$ specifies the relational reasoning chain and invariance constraint on G.

The resulting $\mathcal{SG}_{\text{tgt}}$ serves as the conceptual blueprint for visual synthesis. By keeping G invariant while re-instantiating C^{tgt}, V^{tgt}, and I^{tgt}, the Transfer Agent enables target-specific metaphor construction instead of directly copying the reference schema. This distinction is validated by the *w/o Transfer Agent* ablation, where directly feeding $\mathcal{SG}_{\text{ref}}$ to generation often preserves reference fragments but fails to form a coherent target-domain metaphor.

4.3 Generation Agent

Given the target schema $\mathcal{SG}_{\text{tgt}}$, the Generation Agent converts its structured components into a T2I prompt for visual realization. Unlike the Transfer Agent, which performs schema-level reasoning, this stage translates the subject, carrier, violation, emergent meaning, and expressive attributes into executable generative instructions:

$$P = \text{LLM}(\mathcal{SG}_{\text{tgt}}, p_{\text{generation}}), \tag{4}$$

where $p_{\text{generation}}$ denotes the generation-specific system prompt.

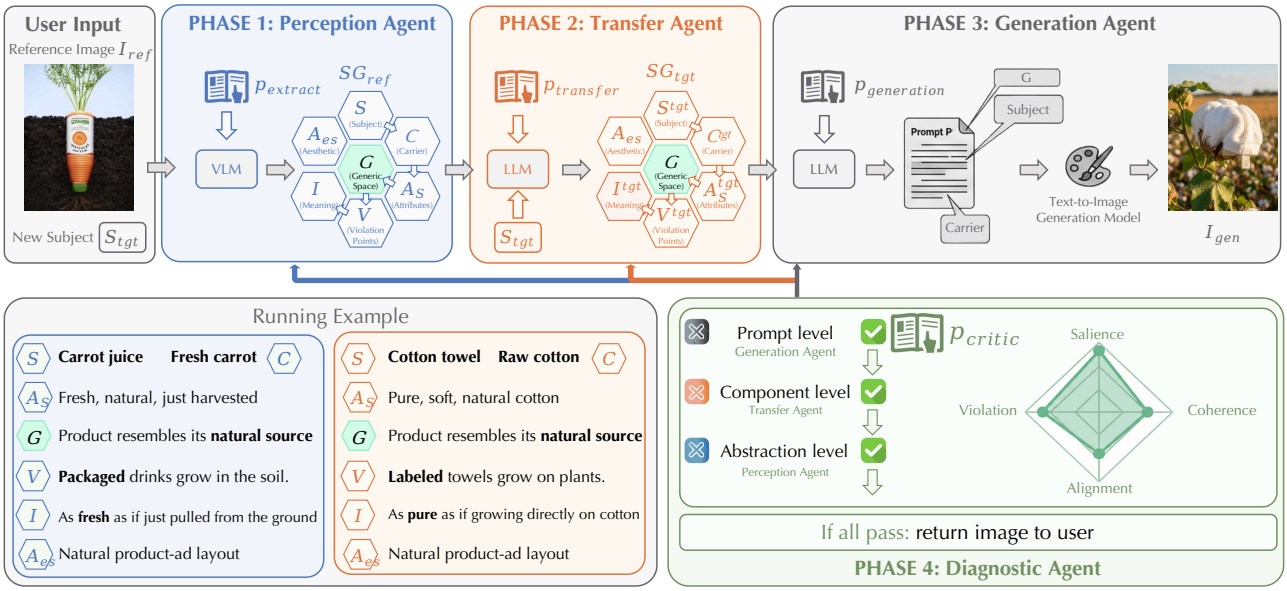

Fig. 3. Architecture of our Self-Reflective Agentic Framework for Visual Metaphor Transfer. The system consists of Perception, Transfer, Generation, and Diagnostic agents. It transforms a given visual metaphor (I_{ref}) into a new target context (I_{gen}) by extracting structured schema representations and mapping them from the source to the target domain ($\mathcal{SG}_{ref} \rightarrow \mathcal{SG}_{tgt}$). In the second column, the quoted input text is the user-specified target subject S_{tgt} (not a free-form metaphor description); the small explanatory text beneath is provided only to aid reader understanding and is *not* available to the model during experiments. A hierarchical feedback loop ensures the generated output faithfully preserves the metaphor logic while adapting to the new subject matter.

The prompt construction emphasizes C_{tgt} for scene layout, V_{tgt} for conceptual tension, and I_{tgt} with $A_{\text{es}}^{\text{tgt}}$ for style and atmosphere. Finally, the target metaphor image is synthesized by a pre-trained generator, which means $I_{\text{gen}} = \text{Gen}(P)$.

4.4 Diagnostic Agent

The generated image I_{gen} may still contain visual or conceptual errors caused by prompt ambiguity, unsuitable component choices, or incorrect schema abstraction. To address this, we introduce a VLM-based Diagnostic Agent that evaluates the result and performs hierarchical refinement. We further extend the Diagnostic Agent with a safety and alignment check: the LLM evaluates the outputs against safety guidelines and rejects or revises schemas that may induce malicious, offensive, or harmful alignments. Unlike the Generation Agent, which only realizes a fixed $\mathcal{SG}_{\text{tgt}}$, this stage decides whether errors should be corrected at the prompt, component, or abstraction level.

a) Diagnostic dimensions. The VLM examines I_{gen} across four complementary dimensions:

- **Subject Salience:** whether S^{tgt} is recognizable and preserves its core attributes;
- **Violation Realization:** whether V^{tgt} is visually explicit and structurally coherent;
- **Relational Coherence:** whether the Generic Space G is successfully instantiated and perceptible; and
- **Meaning Alignment:** whether the image conveys the intended I^{tgt} without negative ambiguity, malicious implication, or offensive association.

Rather than producing numerical scores, the VLM outputs qualitative issue descriptions and assigns each failure to a backtracking level. We provide the diagnostic decision table in the supplementary materials, mapping failure types to their corresponding revision levels and strategies.

b) Hierarchical backtracking and refinement. Based on the diagnostic output, we perform cascaded attribution through three levels governed by strict conditional failure triggers. *Prompt-level* backtracking is triggered when the VLM determines the conceptual structure ($\mathcal{SG}_{\text{tgt}}$) is sound, but visual details fail (e.g., P underspecifies spatial relations or atmosphere); we revise P and regenerate. *Component-level* backtracking is triggered when the VLM detects unbridgeable domain gaps that make the outputs unrecognizable or visually ambiguous, indicating that C^{tgt} or V^{tgt} is unsuitable; we return to the Transfer Agent to replace the visual carrier or redesign the violation. *Abstraction-level* backtracking is triggered when component modifications repeatedly fail (exceeding a local iteration threshold), suggesting that the underlying metaphorical logic (G) was extracted at an improper abstraction level; we revisit $\mathcal{SG}_{\text{ref}}$ and re-extract the relational logic. This refinement is formulated as:

$$I_{\text{final}} = \text{Refine}(I_{\text{gen}}, \mathcal{SG}_{\text{tgt}}, p_{\text{critic}}; \tau), \tag{5}$$

where p_{critic} denotes the diagnostic prompt and τ the iteration threshold. The loop stops when the diagnostic feedback indicates satisfactory quality or reaches the maximum iteration limit, yielding I_{final}.

5 Experiments

We first introduce the experiment settings in Sec. 5.1, and the present qualitative, quantitative, and human evaluation results in Secs. 5.2, 5.3 and 5.4. Finally, we conduct an ablation study and generalizability analysis in Secs. 5.5 and 5.6.

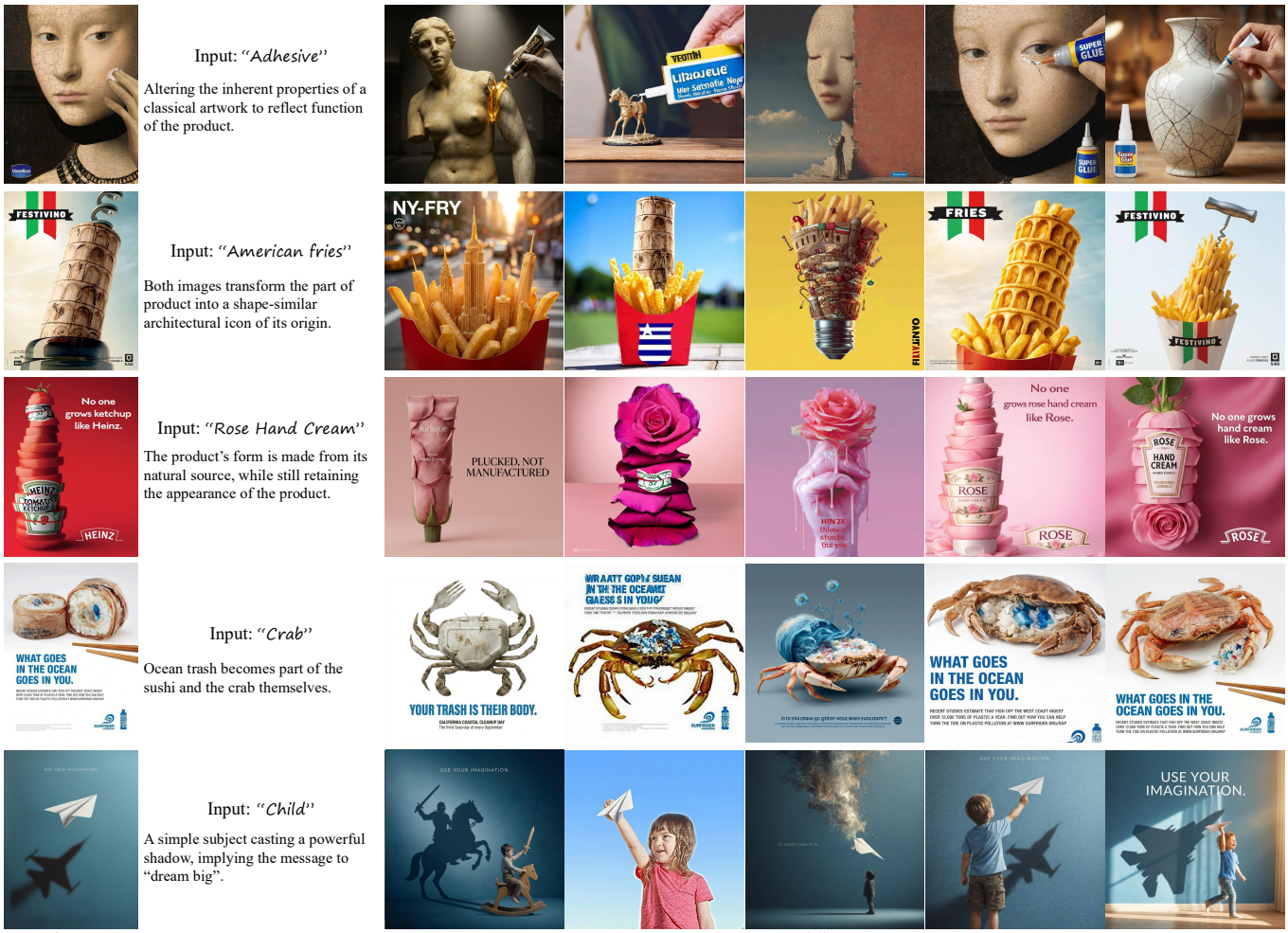

Fig. 4. Qualitative comparison with baseline methods.

5.1 Experimental settings

Baselines. We compare our approach against state-of-the-art multimodal image generation models with integrated visual understanding and reasoning capabilities, as metaphor transfer inherently requires analyzing the source image's creative concept before generating the target. We evaluate image generation baselines **BAGEL** [Deng et al. 2025], **Midjourney** [Midjourney 2026], **GPT-Image** [OpenAI 2025], and **Banana-pro** [DeepMind 2025], and use these names consistently throughout the paper. We further compare against I-spy-a-metaphor [Chakrabarty et al. 2023]. Because that method generates visual metaphors from linguistic metaphors, we prepend a linguistic-metaphor extraction stage over the reference image to align it with our VMT pipeline, then evaluate under the same metrics.

Datasets. We curated a diverse dataset of 126 visual metaphors from the internet, including product ads (32), memes (33), film posters (15), comics (10), and other creative works (36). This heterogeneous collection spans multiple domains to comprehensively test our framework's generalization across various metaphorical styles and compositions.

Metrics. Unlike conventional image evaluation methods such as CLIP [Ilharco et al. 2021] or DINO [Liu et al. 2025a] that primarily assess low-level visual features or semantic similarity, metaphor transfer requires evaluating high-level conceptual reasoning and abstract creative alignment, which necessitates the use of VLMs capable of understanding complex analogical relationships. We employ three frontier VLMs, Gemini-3-pro, GPT-5.2, and Claude-Sonnet-4.5 to assess generated images across multiple dimensions using 10-point scales : (1) Metaphor Consistency (MC), which measures whether the target metaphor preserves the core metaphor logic of the source; (2) Analogy Appropriateness (AA), which evaluates the validity of functional and formal correspondences between the carrier and target subject; and (3) Conceptual Integration (CI), which assesses whether the fusion between the subject and carrier appears natural and harmonious. We also evaluate image aesthetic quality with a SigLip-based predictor [Zhai et al. 2023] to ensure visual appeal. Note that these evaluation VLMs are distinct from the VLM

Fig. 5. **Qualitative comparison of different backbone combinations.** We validate the framework's generalizability by pairing different LLMs (Gemini, GPT) with various T2I models (Banana-pro, GPT-Image, FLUX). FLUX remains essentially competitive with commercial generators.

used in our iterative refinement process (Section 4.4), ensuring independent assessment of generation quality. Using three judges from different providers mitigates single-model bias; we additionally report Qwen-VL judgments in the supplementary material, where our method remains strongest. Per-metric human–VLM agreement is likewise reported in the supplementary material. The complete VLM evaluation prompts and the validation of VLM-as-judge reliability are provided in the supplementary material.

Implementation. We employ Gemini-3-pro as both the VLM and LLM in our pipeline, and utilize Banana-pro for image generation. The iteration threshold τ is set to 5 to balance refinement quality and computational efficiency. p_{extract}, p_{transfer}, $p_{\text{generation}}$ and p_{critic} are provided in the supplementary material. On the 126-metaphor set, the average number of calls per agent is Perception 1.14, Transfer 1.25, Generation 2.19, and Diagnostic 2.19 (ideal value 1); sequential dependence causes call counts to increase along the pipeline. Average wall-clock time is 4 min 11 s with about 8.3k tokens per sample (see supplementary Consumption). To assess reproducibility beyond commercial endpoints, we also re-run the full pipeline with open-source, version-pinned models—Qwen (LLM) and Qwen-VL (VLM) for Perception/Transfer/Diagnostic, and FLUX for generation—and report Strong Prompt / w/o CBT under the same open-source stack (supplementary Table); although absolute scores drop slightly, relative gains of our method persist, confirming that improvements stem from the framework rather than a black-box model.

5.2 Qualitative comparisons

As shown in Fig. 4, our method excels at decoupling abstract creative logic from source domains and re-materializing it within novel targets. While SOTA models like GPT-Image and Banana-pro are visually proficient, they rely on surface-level manipulation rather than decoding underlying metaphoric logic. For instance, in the "American Fries" task, these baselines merely substitute components without grasping the "regional architectural landmark with similar shape" metaphor, whereas in the "Rose Hand Cream" case, they erroneously preserve the "sliced" geometry from the reference, which is semantically incongruent with the new subject. This conceptual deficiency extends to structural integration: in the "Crab" example, baselines scatter trash as background clutter instead of merging it into the organism's anatomy, and in the "Child" scene, they fail to project a "powerful shadow" to convey the intended "dream big" message. Furthermore, models like BAGEL and Midjourney tend to generate results from scratch, leading to a loss of metaphoric alignment. In contrast, our method achieves superior semantic reasoning by successfully synthesizing New York landmarks that mirror the geometric form of fries while signifying their origin, blending organic rose textures with traditional packaging, and seamlessly embedding plastic waste into the crab's biological structure. By accurately mapping abstract relationships, our approach demonstrates a unique capacity to re-materialize abstract creative intents while ensuring high-level conceptual consistency.

Table 1. Quantitative evaluation results. Best results in **bold**.

Methods	Gemini-3-pro			GPT-5.2			Claude-4.5			Aes.↑
	MC↑	AA↑	CI↑	MC	AA	CI	MC	AA	CI	
BAGEL	5.17	4.55	5.05	6.21	5.83	6.07	6.05	5.58	5.95	4.77
Midjourney	5.33	5.57	6.09	6.33	6.46	6.24	6.51	5.94	6.06	5.22
GPT-Image	8.08	7.59	7.47	7.71	7.65	7.54	7.95	7.39	7.51	5.63
Banana-pro	8.75	7.68	7.33	7.95	7.77	7.37	8.08	7.42	7.74	5.57
I-spy-a-metaphor	8.86	8.02	7.49	8.14	7.98	7.55	8.52	7.83	7.94	5.48
Direct Prompt	8.79	8.03	7.63	8.13	7.96	7.69	8.44	7.85	7.92	5.59
Strong Prompt	9.02	8.13	7.68	8.38	8.04	7.72	8.59	7.95	8.02	5.61
w/o CBT	8.91	8.09	7.58	8.33	8.01	7.71	8.56	7.89	7.97	5.61
w/o Trans.	8.97	8.11	7.47	8.36	8.07	7.41	8.58	7.97	7.49	5.44
w/o Diag.	9.14	8.47	8.33	8.44	8.33	8.29	8.62	8.38	8.19	5.63
Ours	**9.31**	**8.97**	**8.76**	**8.62**	**8.51**	**8.58**	**8.73**	**8.61**	**8.36**	**5.68**

5.3 Quantitative comparisons

As Tab. 1 shows, our method consistently outperforms all baselines across three frontier VLMs and the aesthetic predictor. Notably,

we achieve the most significant improvement in the AA metric (a 16.8% increase over the runner-up) demonstrating that our proposed Metaphor Transfer Agent effectively identifies metaphorically consistent visual carriers that best match new subjects. Beyond AA, our approach maintains superior scores in MC and CI, while also securing the highest aesthetic score (5.68). This consensus among evaluators (Gemini, GPT, and Claude) underscores our framework's robustness in generating logically sound and visually harmonious metaphorical images without sacrificing artistic quality. Relative to the adapted I-spy-a-metaphor, our method improves MC/AA/CI under all three judges, indicating that schema-driven transfer and closed-loop diagnosis outperform linguistic-metaphor co-creation when aligned to the VMT setting.

5.4 Human evaluation study

To validate the perceptual quality and creative effectiveness between our method and the baselines, we conducted a comprehensive human evaluation study with 100 participants (age 14-55, 45 male, 55 female). The study comprises two tasks.

In Task 1, each participant evaluates 45 images generated by each method (ours and 4 baselines, totaling 225 images per participant) independently along five dimensions using 5-point Likert scales: (1) Metaphor Recognizability (MR); (2) Metaphor Ingenuity (MI); (3) Violation Appropriateness (VA); (4) Visual Integration (VI); and (5) Overall Visual Quality (VQ). The detail of the definitions are in the supplementary material. As shown in the left of Fig. 7, our method consistently outperforms all baselines across all five dimensions. Notably, it achieves a significant lead in MI (4.61) and VA (4.49), indicating that our framework produces more creative and purposeful metaphorical designs than SOTA models like Banana-pro and GPT-Image. Furthermore, our approach attains the highest scores in VI (4.77) and VQ (4.79), confirming that our focus on conceptual reasoning does not compromise aesthetic fidelity. These results collectively demonstrate our method's superior capability in synthesizing metaphorical images.

In Task 2, we conduct GSB (Good/Same/Bad) evaluation to assess user preference by asking participants which image in each pair delivered a more compelling metaphorical message. As shown in the right of Fig. 7, our method is consistently favored by participants, securing over 60% "Ours Better" ratings across all baseline comparisons. Notably, our approach outperforms strong commercial competitors such as GPT-Image and Banana-pro in 64.44% and 67.11% of cases, respectively, while being judged as inferior in fewer than 10% of pairs. This preference margin widens further against Midjourney (72.89%) and BAGEL (76.22%), demonstrating our framework's metaphorical syntheses are significantly more resonant and conceptually effective than existing SOTA models.

5.5 Ablation study

We conduct qualitative and quantitative comparisons across ablation variants, as shown in Fig. 6 and Tab. 1. To disentangle prompt strength, CBT-based schema description, staged transfer, and diagnostic refinement, all variants use the same reference image, target subject, generator, and evaluation protocol, differing only in the enabled components or provided descriptions.

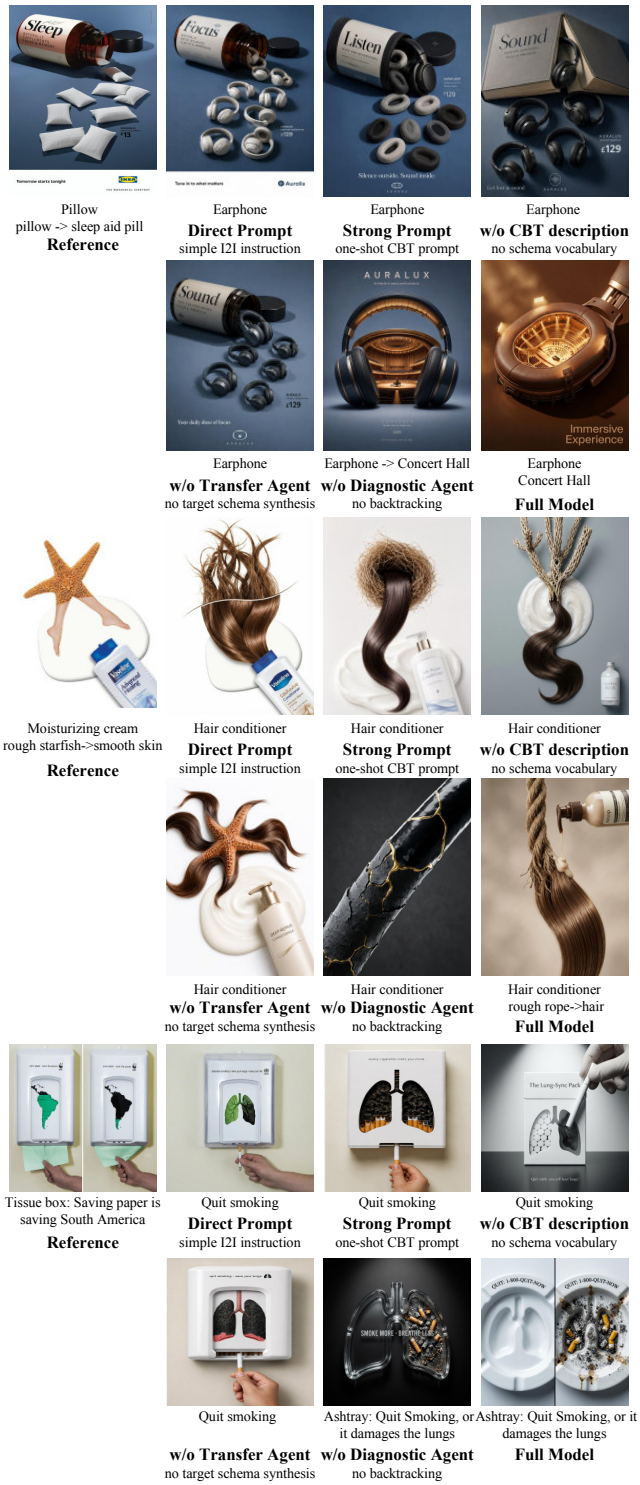

Fig. 6. **Qualitative comparison of ablation variants.**

Direct prompt. We first evaluate a direct I2I prompting baseline using Banana-pro [DeepMind 2025]. The model receives only a

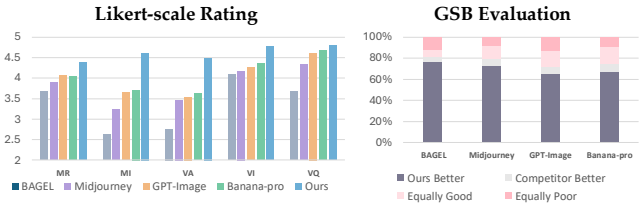

Fig. 7. Human evaluation study.

simple instruction: "*Understand this advertisement image, analyze its metaphors and creative ideas, and transfer this creative idea to the new product.*" Without structured guidance, the model often falls back to literal object substitution or salient-shape copying, while missing the underlying relational mechanism, as shown in Tab. 1. The drops in MC, AA, and CI indicate limited metaphor understanding under simple prompting.

Strong prompt with full CBT description. To test whether the gains mainly come from stronger prompt engineering, we introduce a manually designed strong-prompt baseline with a full CBT-style task description. The prompt explicitly instructs the LLM to analyze the reference image in terms of subject S, carrier C, relational invariant R, violation point V, and emergent meaning M, and to directly convert this one-shot analysis into a text-to-image prompt. Although the stronger prompt improves metaphor awareness compared to direct prompting, it remains a single-stage approach and therefore often yields unstable carrier selections or incomplete relational transfer. As illustrated in Tab. 1, it may capture partial visual or functional cues, yet still fails to consistently preserve the cross-domain metaphorical structure. These results indicate that prompt strength alone does not fully account for the observed improvements.

W/o CBT description . We then remove the CBT-based schema description while keeping the overall multi-stage prompting process (including iteration). Without explicit notions of subject, carrier, relational invariant, violation, and emergent meaning, the model lacks a structured vocabulary for decomposing and transferring the reference metaphor. As shown in Tab. 1, this variant can generate visually plausible results, but the outputs tend to rely on generic scene composition, superficial texture transfer, or loosely related causal cues rather than preserving the intended metaphorical logic. The performance drop relative to the full method shows that merely introducing iteration without the CBT schema is inferior, and that the CBT description is important for stabilizing metaphor decomposition and transfer.

W/o Transfer Agent . We next remove the Transfer Agent while retaining schema extraction. The Perception Agent extracts the reference schema $\mathcal{SG}_{\text{ref}}$, which is directly fed to an LLM to produce the T2I prompt, without synthesizing a target schema $\mathcal{SG}_{\text{tgt}}$. Thus, this variant has access to the reference schema but lacks explicit target-side carrier selection, violation redesign, and relational re-instantiation. Although it preserves more reference information than prompt-only variants, it adapts the schema poorly to the target domain, leading to unsuitable carriers, superficial attribute transfer, or incomplete instantiation of the reference mechanism, as shown in Tab. 1. The AA drop indicates that target-side schema synthesis is necessary for constructing coherent transferred metaphors.

Table 2. User study results for ablations. Best results in **bold**.

Methods	MR↑	MI↑	VA↑	VI↑	VQ↑
Direct Prompt	3.62	2.75	2.96	3.75	3.65
Strong Prompt	3.97	3.69	3.77	3.89	4.09
w/o CBT	3.88	3.66	3.68	4.04	4.12
w/o Trans.	4.01	3.71	3.83	4.33	4.25
w/o Diag.	4.26	4.05	3.98	4.61	4.48
Ours	**4.47**	**4.56**	**4.51**	**4.74**	**4.76**

W/o Diagnostic Agent. This variant keeps the Perception, Transfer, and Generation Agents but removes Phase 4, generating once from $\mathcal{SG}_{\text{tgt}}$ without prompt-, component-, or abstraction-level revision. Without hierarchical diagnosis, semantic and structural errors persist, including misleading source-context associations, ambiguous carrier realization, and failures to enforce the required visual layout, as shown in Tab. 1. The consistent score drops confirm the importance of closed-loop refinement.

Full model. The full model combines all components: Perception extracts $\mathcal{SG}_{\text{ref}}$, Transfer synthesizes $\mathcal{SG}_{\text{tgt}}$ by preserving the relational invariant while re-instantiating the carrier and violation, Generation converts $\mathcal{SG}_{\text{tgt}}$ into a T2I prompt, and Diagnosis performs hierarchical backtracking when needed. Across Rows 1–5 in Tab. 1, the full model better preserves the intended relational logic, selects suitable target-domain carriers, and enforces the required visual structures, demonstrating the necessity of each component for high-quality metaphor transfer.

User study on ablations. To verify that these component-wise gains are also perceptible to human viewers, we extend the human evaluation protocol to all ablation variants, participants rate the outputs of each ablation variant along the same five dimensions on 5-point Likert scales. As shown in Tab. 2, our full model receives the highest ratings on all dimensions. Direct Prompt is rated lowest, especially on MI (2.75) and VA (2.96), while Strong Prompt and w/o CBT plateau around 3.7 on these conceptual dimensions, confirming that neither stronger prompting nor iteration without the CBT schema suffices. Even against the strongest variant (w/o Diag.), the remaining gaps concentrate on MI (+0.51) and VA (+0.53), indicating that the perceived advantage stems from schema-driven reasoning and closed-loop diagnosis rather than rendering quality.

5.6 Generalizability analysis

To verify the robustness and model-agnostic nature of our framework, we evaluate it across different LLM and T2I generator combinations. As shown in Fig. 5, we test two reasoning backbones (Gemini, GPT-4) paired with three rendering engines (Banana-pro, GPT-Image, FLUX). In particular, FLUX results remain essentially competitive with commercial generators, supporting that the pipeline is not tied to a single proprietary image model.

Consistency in metaphorical mapping. The top row shows the "LEGO" scenario, where the core metaphor contrasts "disorganized effort" with "total inaction". Regardless of the T2I model, the framework translates the reference's "missed target" concept into a "chaotic LEGO pile", and "total inaction" into an "empty baseplate". This consistency shows that our method preserves the metaphor's logical structure across visual generators.

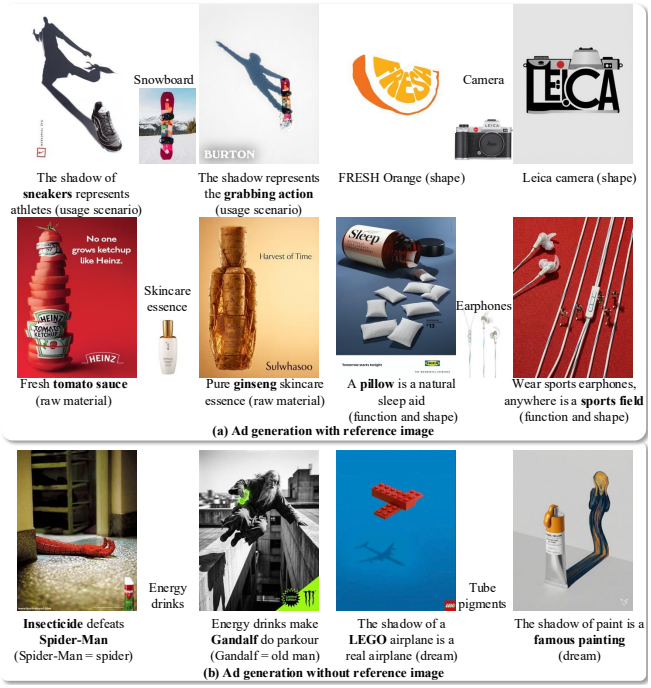

Fig. 8. **Application of product advertisements for multiple-reference-guided scenarios.** Our method can flexibly handle both visual-to-visual and text-to-visual creative workflows.

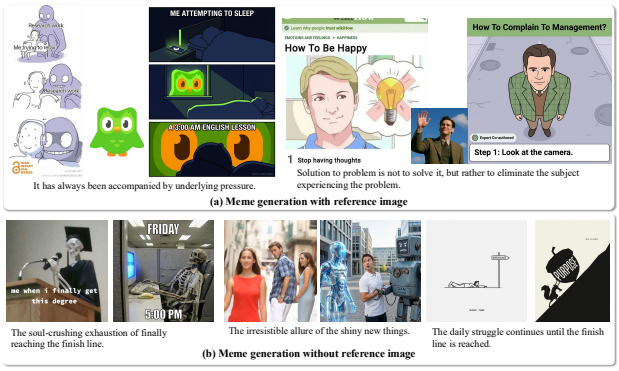

Fig. 9. **Application of meme image generation.**

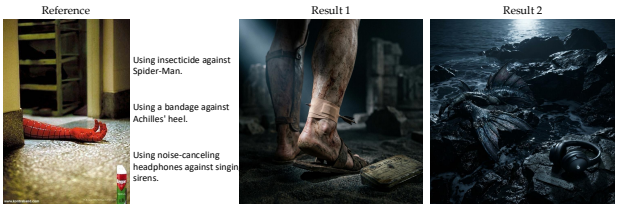

Fig. 10. Badcase. Cognitive overload and obscure symbolism.

Diversity in narrative interpretation. The bottom row ("Protecting Forests") shows how different LLMs shape the creative narrative while maintaining visual coherence. Gemini-driven variants interpret role reversal as *active retaliation*, prompting anthropomorphic trees to demolish urban buildings with axes. GPT-driven variants interpret it as *cultural satire*, depicting civilized trees displaying chainsaws as "hunting trophies." Despite these divergences, all T2I models faithfully render their prompts. This demonstrates that our framework supports creative flexibility in reasoning while preserving high-fidelity visual execution during generation.

We presented more applications for product advertisements and meme image generation in Figs. 8 and 9, bad cases in Fig. 10. For corresponding discussions, please refer to the supplementary materials.

6 Conclusions

We introduced visual metaphor transfer, a task that goes beyond pixel-level editing by extracting the underlying metaphor logic from a reference image and re-instantiating it on a user-specified new subject. To achieve this, we formalize metaphor structure with Schema Grammar $\mathcal{SG}$ and construct a closed-loop multi-agent pipeline comprising: a Perception Agent that extracts the schema, a Transfer Agent that preserves the Generic Space G while finding a new carrier, a Generation Agent that turns the schema into prompts, and a Diagnostic Agent that backtraces failures across prompt, component, and abstraction levels. Experiments show that this design improves metaphor consistency, analogy appropriateness, and conceptual integration compared with baselines.

Acknowledgments

We thank the reviewers for their insightful comments and suggestions. This work was partly supported by the National Natural Science Foundation of China under No. 62572458, National Science and Technology Council, Taiwan under Grant 113-2221-E006-161-MY3, and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy–EXC 2117–422037984.

References

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

Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. 2023. Break-a-scene: Extracting multiple concepts from a single image. In *SIGGRAPH Asia 2023 Conference Papers*. 1–12.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2025. Text to image generation with bidirectional multiway transformers. *Computational Visual Media* 11, 2 (2025), 405–422.

blackforestlabs.ai. 2024. FLUX, offering state-of-the-art performance image generation. https://blackforestlabs.ai/. Accessed: 2024-10-07.

Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xinchi Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu, et al. 2025. Hunyuanimage 3.0 technical report. *arXiv preprint arXiv:2509.23951* (2025).

Tuhin Chakrabarty, Arkadiy Saakyan, Olivia Winn, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki, and Smaranda Muresan. 2023. I spy a metaphor: Large language models and diffusion models co-create visual metaphors. *arXiv preprint arXiv:2305.14724* (2023).

Lydia B Chilton, Savvas Petridis, and Maneesh Agrawala. 2019. Visiblends: A flexible workflow for visual blends. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

Yusheng Dai, Chenxi Wang, Chang Li, Chen Wang, Kewei Li, Jun Du, Lei Sun, Jianqing Gao, Ruoyu Wang, and Jiefeng Ma. 2025. Latent swap joint diffusion for 2d long-form latent generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 11006–11015.

Google DeepMind. 2025. Gemini 3 Pro Image (Nano Banana Pro): advanced AI image generation and editing model. urlhttps://deepmind.google/models/gemini-image/pro/. [Online; accessed 2026-01].

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

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

Brian Falkenhainer, Kenneth D Forbus, and Dedre Gentner. 1989. The structure-mapping engine: Algorithm and examples. *Artificial intelligence* 41, 1 (1989), 1–63.

Zezhong Fan, Xiaohan Li, Kaushiki Nag, Chenhao Fang, Topojoy Biswas, Jianpeng Xu, and Kannan Achan. 2024. Prompt optimizer of text-to-image diffusion models for abstract concept understanding. In *Companion Proceedings of the ACM Web Conference 2024*. 1530–1537.

Gilles Fauconnier and Mark Turner. 1998. Conceptual integration networks. *Cognitive science* 22, 2 (1998), 133–187.

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

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. 2023. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *The Eleventh International Conference on Learning Representations*.

Dedre Gentner. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive science* 7, 2 (1983), 155–170.

Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. 2024. Style aligned image generation via shared attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4775–4785.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. *OpenCLIP*. doi:10.5281/zenodo.5143773

Youwen Kang, Zhida Sun, Sitong Wang, Zeyu Huang, Ziming Wu, and Xiaojuan Ma. 2021. MetaMap: Supporting visual metaphor ideation through multi-dimensional example-based exploration. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

Girish A Koushik, Fatemeh Nazarieh, Katherine Birch, Shenbin Qian, and Diptesh Kanojia. 2025. The Mind's Eye: A Multi-Faceted Reward Framework for Guiding Visual Metaphor Generation. *arXiv preprint arXiv:2508.18569* (2025).

Manishit Kundu, Sumit Shekhar, and Pushpak Bhattacharyya. 2025. Looking Beyond the Pixels: Evaluating Visual Metaphor Understanding in VLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*. 23137–23158.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*. PMLR, 12888–12900.

Mingcheng Li, Xiaolu Hou, Ziyang Liu, Dingkang Yang, Ziyun Qian, Jiawei Chen, Jinjie Wei, Yue Jiang, Qingyao Xu, and Lihua Zhang. 2025. MCCD: Multi-Agent Collaboration-based Compositional Diffusion for Complex Text-to-Image Generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 13263–13272.

Jiayi Liao, Xu Chen, Qiang Fu, Lun Du, Xiangnan He, Xiang Wang, Shi Han, and Dongmei Zhang. 2024. Text-to-image generation for abstract concepts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 3360–3368.

Feng Liu and Xiaobin Chang. 2025. IIDM: Image-to-image diffusion model for semantic image synthesis. *Computational Visual Media* 11, 2 (2025), 423–429.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. 2025a. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*. Springer, 38–55.

Yilong Liu, Houwen Zheng, and Shuojin Yang. 2025b. SAD: Style-aware diffusion adaptation for few-shot style transfer image generation. *Computational Visual Media* (2025).

Midjourney. 2026. Midjourney. https://www.midjourney.com.

Chong Mou, Yanze Wu, Wenxu Wu, Zinan Guo, Pengze Zhang, Yufeng Cheng, Yiming Luo, Fei Ding, Shiwen Zhang, Xinghui Li, et al. 2025. Dreamo: A unified framework for image customization. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*. 1–12.

OpenAI. 2025. GPT Image 1.5: The new ChatGPT Images is here. urlhttps://openai.com/zh-Hans-CN/index/new-chatgpt-images-is-here. [Online; accessed 2026-01].

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2024. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *The Twelfth International Conference on Learning Representations*.

Wenhao Qian, Zhenzhen Hu, Zijie Song, and Jia Li. 2025. Concept Drift Guided LayerNorm Tuning for Efficient Multimodal Metaphor Identification. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*. 1100–1108.

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

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 22500–22510.

Marcelo Sandoval-Castaneda, Bryan Russell, Josef Sivic, Gregory Shakhnarovich, and Fabian Caba Heilbron. 2025. EditDuet: A Multi-Agent System for Video Non-Linear Editing. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*. 1–11.

Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. 2025. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427* (2025).

Tamar Rott Shaham, Sarah Schwettmann, Franklin Wang, Achyuta Rajaram, Evan Hernandez, Jacob Andreas, and Antonio Torralba. 2024. A multimodal automated interpretability agent. In *Forty-first International Conference on Machine Learning*.

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

Yael Vinker, Tamar Rott Shaham, Kristine Zheng, Alex Zhao, Judith E Fan, and Antonio Torralba. 2025. Sketchagent: Language-driven sequential sketch generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 23355–23368.

Sitong Wang, Savvas Petridis, Taeahn Kwon, Xiaojuan Ma, and Lydia B Chilton. 2023. PopBlends: Strategies for conceptual blending with large language models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.

Yuxiang Wei, Yiheng Zheng, Yabo Zhang, Ming Liu, Zhilong Ji, Lei Zhang, and Wangmeng Zuo. 2025. Personalized image generation with deep generative models: A decade survey. *Computational Visual Media* (2025).

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. 2025. Qwen-image technical report. *arXiv preprint arXiv:2508.02324* (2025).

Bo Xu, Junzhe Zheng, Jiayuan He, Yuxuan Sun, Hongfei Lin, Liang Zhao, and Feng Xia. 2024. Generating multimodal metaphorical features for meme understanding. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 447–455.

Yu Xu, Fan Tang, Juan Cao, Xiaoyu Kong, Yuxin Zhang, Jintao Li, Oliver Deussen, and Tong-Yee Lee. 2026. Headrouter: A training-free image editing framework for mm-dits by adaptively routing attention heads. *ACM Transactions on Graphics* 45, 3 (2026), 1–14.

Yu Xu, Fan Tang, Juan Cao, Yuxin Zhang, Oliver Deussen, Weiming Dong, Jintao Li, and Tong-Yee Lee. 2025. B4M: Breaking Low-Rank Adapter for Making Content-Style Customization. *ACM Transactions on Graphics* 44, 2 (2025), 1–17. doi:10.1145/3728461

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*. 11975–11986.

Yuxin Zhang, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, Tong-Yee Lee, and Changsheng Xu. 2022. Domain enhanced arbitrary image style transfer via contrastive learning. In *ACM SIGGRAPH 2022 conference proceedings*. 1–8.