




# Probing and Inducing Combinational Creativity in Vision-Language Models

Yongqian Peng<sup>1,2\*</sup>, Yuxi Ma<sup>1\*</sup>, Mengmeng Wang<sup>3</sup>, Yuxuan Wang<sup>3</sup>,  
Yizhou Wang<sup>4</sup>, Chi Zhang<sup>3</sup>, Yixin Zhu<sup>1</sup>, , Zilong Zheng<sup>3</sup>, 

<sup>1</sup> Institute for Artificial Intelligence, Peking University    <sup>2</sup> Yuanpei College, Peking University

<sup>3</sup> State Key Laboratory of General Artificial Intelligence, BIGAI

<sup>4</sup> Center on Frontiers of Computing Studies, School of Computer Science, Peking University

\*Equal contributors     yixin.zhu@pku.edu.cn, zlzheng@bigai.ai

## Abstract

The ability to combine existing concepts into novel ideas stands as a fundamental hallmark of human intelligence. Recent advances in Vision-Language Models (VLMs) like GPT-4V and DALL-E-3 have sparked debate about whether their outputs reflect combinational creativity—defined by M. A. Boden (1998) as synthesizing novel ideas through combining existing concepts—or sophisticated pattern matching of training data. Drawing inspiration from cognitive science, we investigate the combinational creativity of VLMs from the lens of concept blending. We propose the **Identification-Explanation-Implication (IEI) framework**, which decomposes creative processes into three levels: identifying input spaces, extracting shared attributes, and deriving novel semantic implications. To validate this framework, we curate CreativeMashup, a high-quality **dataset** of 666 artist-generated visual mashups annotated according to the IEI framework. Through extensive experiments, we demonstrate that in **comprehension** tasks, best VLMs have surpassed average human performance while falling short of expert-level understanding; in **generation** tasks, incorporating our IEI framework into the generation pipeline significantly enhances the creative quality of VLMs outputs. Our findings establish both a theoretical foundation for evaluating artificial creativity and practical guidelines for improving creative generation in VLMs. Project page: <https://ppyqq.github.io/aicc/>.

**Keywords:** creativity; combinational creativity; VLMs

## Introduction

*Creativity is just connecting things.*

— Steve Jobs

Creativity, a defining characteristic of human intelligence, enables the production of novel concepts, solutions, and artistic expressions (Mehrotra et al., 2024; M. A. Boden, 1998; Holyoak & Thagard, 1996). At its core, combinational creativity—the ability to generate new ideas by meaningfully combining familiar ones—represents one of the most fundamental creative processes (Han et al., 2019; M. Boden, 2009), influencing domains from art and design to scientific discovery (Guzdial & Riedl, 2018); see examples in Fig. 1. This uniquely human capacity allows us to actively shape our environment rather than merely respond to it (Gabora & Kaufman, 2010).

Recent advances in VLMs have demonstrated increasingly sophisticated generative capabilities across multiple modalities (Franceschelli & Musolesi, 2024; Chakrabarty et al., 2024; Bubeck et al., 2023; Ma et al., 2023), producing creative content that closely resembles human-generated work (Orwig et al., 2024; Koivisto & Grassini, 2023; Tian et al., 2024; Mehrotra et al., 2024). However, this apparent success raises fundamental questions about the underlying mechanisms: while existing research has examined VLMs’ creativity through specific lenses such as semantic association (Chen & Ding, 2023)

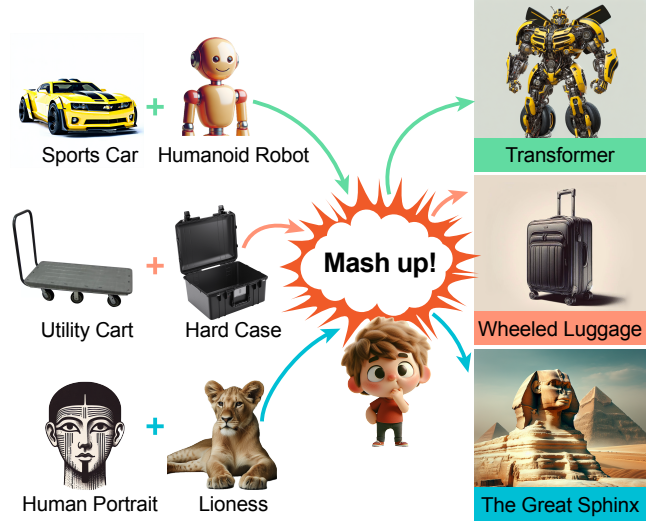


Figure 1: **Combinational creativity across domains.** Examples showing how combining two distinct elements creates novel concepts: sports car + humanoid robot → Transformer (entertainment), utility cart + hard case → wheeled luggage (industrial design), and human portrait + lioness → Great Sphinx (ancient architecture). Each combination demonstrates how merging basic elements generates innovative outcomes.

and divergent thinking (Bellemare-Pepin et al., 2024), we lack a systematic framework for evaluating whether these models implement genuine combinational creative processes or merely leverage statistical patterns in their training data.

To address this gap, we investigate VLMs through the theoretical framework of combinational creativity. Grounded in cognitive science, this framework provides explicit criteria for evaluating both creative processes and their outcomes. Specifically, we address two complementary research questions: ① How effectively can VLMs *understand* and *interpret* combinational creative processes and products? ② Can the explicit incorporation of the combinational creative thinking process enhance the *generative* capabilities of these systems?

Drawing inspiration from conceptual blending theory (Fauconnier & Turner, 2003), we introduce the Identification-Explanation-Implication (IEI) framework—a hierarchical decomposition of combinational creativity into three levels:

- **Identification:** Recognition of constituent elements (*What objects are combined?*)
- **Explanation:** Analysis of combinatorial mechanisms (*How are these objects combined?*)
- **Implication:** Interpretation of semantic meaning (*What message does this combination convey?*)

For comprehension tasks (Q①), this framework enables sys-

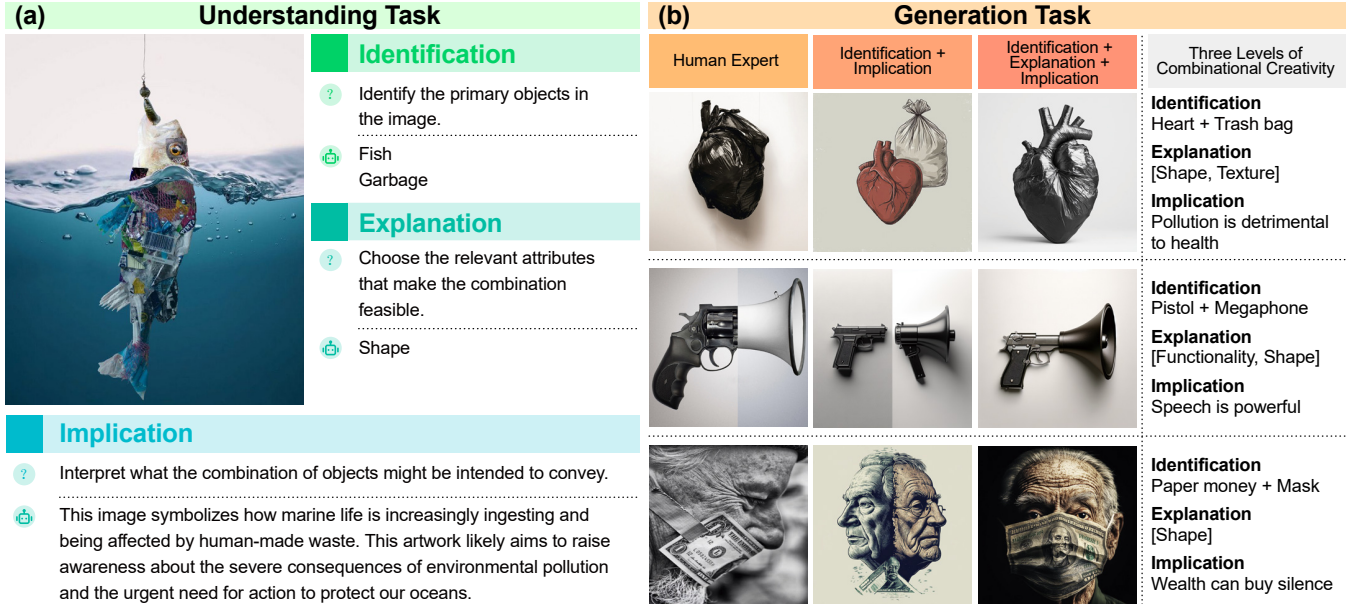


Figure 2: **Examples of the comprehension task and generation task.** (a) The understanding task demonstrates three evaluation components using a fish-garbage mashup image: human participants or VLMs identify primary objects, explain combination attributes, and interpret implications. (b) The generation task compares outputs from human experts and two model settings (Identification + Implication vs. Identification + Explanation + Implication) across three concept pairs (heart-trash, pistol-megaphone, paper money-mask).

tematic analysis through progressively deeper analytical levels. For generation tasks (Q2), we investigate two approaches while maintaining consistent input specifications: a method that explicitly incorporates conceptual attribute mapping based on our IEI framework, and a baseline approach using standard chain-of-thought prompting (Wei et al., 2022).

To validate this framework, we curate CreativeMashup, a novel benchmark dataset designed for combinational creativity. CreativeMashup features professional artists’ original visual mashups paired with expert annotations that decompose each creation according to the IEI framework’s three levels, providing ground-truth data for evaluating both comprehension and generation capabilities. Through extensive experiments on CreativeMashup, we demonstrate that state-of-the-art VLMs achieve above-average human performance in **comprehension** tasks while lagging behind the expert performance, and that the explicit incorporation of our framework significantly enhances their creative **generation** capabilities.

## Related Work

**Combinational Creativity** Margaret A. Boden’s seminal work (M. A. Boden, 1998, 2004; M. Boden, 2009) identifies three fundamental types of creativity, with combinational creativity—the generation of novel ideas through familiar combinations—being particularly relevant to Artificial Intelligence (AI) research. For AI systems to achieve this form of creativity, Boden argues they must develop three critical capabilities: rich associative memory, understanding of human values, and computational expression of these values (Peng et al., 2024). Despite recent advances in AI, these fundamental aspects remain largely unexplored in modern research (M. A. Boden, 1998, 2004; Ma et al., 2023).

**Conceptual Blending Theory** Conceptual blending theory, developed in cognitive science, provides a theoretical framework for understanding combinational creativity (Fauconnier & Turner, 2003). In its basic form, the theory describes four interconnected spaces: two input concept spaces, a generic space containing shared structures, and a blend space where new meanings emerge. The generic space serves as a crucial bridge, identifying common elements between input spaces that enable meaningful integration and the creation of novel insights (Fauconnier & Turner, 2003). Early computational implementations, such as Han et al. (2018)’s system, approached this challenge via hand-authored semantic networks, combining related concepts in structured ways.

**Creativity in VLMs** The emergence of pre-trained VLMs has demonstrated remarkable creative capabilities, raising fundamental questions about their relationship to human cognition and creative processes. Recent “probing” studies have employed both classical psychological measures and specialized methodologies to investigate VLMs’s creativity (Franceschelli & Musolesi, 2024; Tian et al., 2024; Koivisto & Grassini, 2023; Hessel et al., 2023). For instance, Koivisto & Grassini (2023)’s evaluated AI chatbots using the Alternative Uses Task and revealed that while AI systems generally outperformed average humans, the highest-quality human responses remained superior. Similarly, specialized evaluations (e.g., Akula et al. (2023)’s visual metaphor dataset and A. Ji et al. (2022)’s Tangram puzzle assessments) have provided insights into specific aspects of machine creativity and abstract reasoning.

However, existing research primarily focuses on performance metrics rather than *underlying creative mechanisms*. The proposed IEI framework addresses this limitation by providing a systematic approach to evaluating both the *comprehension* and *generation* aspects of combinational creativity in



VLMs, moving beyond simple performance measures.

## The IEI Framework

Drawing from conceptual blending theory (Fauconnier & Turner, 2003), we decompose combinational creativity into three hierarchical levels, each corresponding to progressively deeper analytical capabilities.

**Identification** This foundational level involves recognizing and isolating constituent *elements* that form potential combinations. Corresponding to input mental spaces in conceptual blending theory, this process requires the system not only to detect visual elements but also to understand their categorical identity and potential roles in combinations.

**Explanation** This intermediate level analyzes structural and semantic alignments between identified elements, mapping to the generic space in blending theory. This process encompasses (i) attribute extraction, identifying salient features of each element (*e.g.*, shape, functionality); (ii) cross-space mapping, finding corresponding features between elements; and (iii) compatibility assessment, evaluating which shared attributes could support meaningful combinations.

**Implication** The most sophisticated level involves deriving emergent meaning from the combinations, corresponding to the blend space in conceptual blending theory. It requires (i) pattern completion, inferring implied relationships beyond explicit combinations; (ii) semantic integration, synthesizing coherent meaning from combined elements; and (iii) cultural/contextual reasoning, understanding broader implications within social/cultural contexts.

The IEI framework systematically evaluates combinational creativity in comprehension and generation tasks. For comprehension, each level within the IEI framework represents progressively complex analytical capabilities, offering clear metrics to investigate comprehension capabilities. In generation tasks, the IEI framework guides the creative process systematically, facilitating more sophisticated and meaningful outputs compared to standard chain-of-thought approaches. This structured decomposition supports the design of experiments that explicitly integrate systematic creative thinking, enhancing VLMs’ capacity for contextually relevant creation.

## Experiment 1: Do VLMs Understand Combinational Creativity?

We design three comprehension tasks (see Fig. 2a) to investigate: To what extent can VLMs understand combinational creativity expressed in images?

### CreativeMashup Dataset Construction

To enable rigorous evaluation of combinational creativity comprehension, we developed CreativeMashup, a carefully curated benchmark dataset. Working with professional artists, we collected 666 visual mashups specifically designed to exemplify creative combinations. Each mashup features the deliberate blending of two ordinary physical objects against simple backgrounds, allowing a clear focus on the creative combinations.

The distinguishing feature of the dataset is its comprehensive annotation scheme aligned with our IEI framework. Expert annotators provided detailed labels across all three levels: (i) identification-level annotations of constituent objects, (ii) explanation-level documentation of combining attributes, and (iii) implication-level interpretations of semantic meaning. This rich annotation structure enables systematic evaluation of both human and machine understanding of combinational creativity at multiple cognitive levels.

### Task Setups

We design three tasks corresponding to the levels of our IEI framework, each probing progressively deeper aspects of combinational creativity understanding:

**Identification** This task probes the basic object recognitions in a visual mashup image. These basic objects serve as the source components that can be combined to create the final image. Given a visual mashup image, models must identify the two primary objects that have been combined, requiring both visual perception and conceptual understanding. We assess performance using precision and recall. Our system uses GPT-4o to perform object matching by incorporating the homogeneous meaning sets that were created during annotation, along with appropriate instructions/examples to guide the process.

**Explanation** This task examines the understanding of the combinatorial mechanisms underlying visual mashups. Models must identify specific attributes that allow for feasible combinations in a visual mashup image. Typically, these are attributes shared by the two basic objects, such as similar shapes, functions, or semantic properties that enable meaningful integration. When evaluating performance in recognizing the attributes, correctness is assessed by comparing the selected attributes with a predefined list of ground-truth attributes for each image. We use precision to measure this performance.

**Implication** This task assesses deeper semantic understanding by requiring models to interpret what the combination in a visual mashup image might be intended to convey and express. This includes identifying underlying themes, emotions, cultural references, or conceptual messages embedded in the creative combination. The task demands not just recognition of combined elements, but comprehension of emergent meaning that arises from their integration. To assess both models’ and humans’ capabilities in this task, we perform pairwise comparisons among the multiple implications associated with each image. To automate the evaluation pipeline, we use GPT-4o as an adjudicator. We validate the reliability of this method by having two human experts annotate preferences for 200 comparison results, achieving consistency rates of 85.1% and 81.6% with GPT-4o. This automated evaluation technique for open-ended responses has demonstrated robustness in previous research (J. Ji et al., 2024; Chiang et al., 2024).

### Experimental Setups

We establish human baselines and evaluate a comprehensive set of models to assess combinational creativity understanding.

Table 1: **Model performance on comprehension tasks.** Models are evaluated on identification (P/R), explanation (P), and implication WR metrics for combinational creativity understanding. **P** refers to precision, **R** to recall, and **WR** to winning rate. Human performance (gray row) serves as the baseline for VLMs.

| Models                                   | Identification |              | Explanation  | Implication |
|--|----------------|--------------|--------------|-------------|
|  | P↑             | R↑           | P↑           | WR↑         |
| Human Experts                            | -              | -            | -            | <b>78.3</b> |
| Average People                           | 53.42          | 70.33        | 69.89        | 51.0        |
| GPT-4o (OpenAI, 2024b)                   | <b>75.67</b>   | <b>85.00</b> | <b>74.19</b> | 73.5        |
| GPT-4V (OpenAI, 2023)                    | 60.83          | 75.00        | 63.44        | 71.9        |
| Gemini-1.5-Pro (Team et al., 2024)       | 73.67          | 81.33        | 54.34        | 71.7        |
| Claude-3.5-Sonnet (Anthropic, 2024b)     | 60.08          | 74.83        | <b>74.19</b> | 62.9        |
| Claude-3-Opus (Anthropic, 2024a)         | 63.17          | 72.50        | 65.59        | 39.2        |
| LLaVA-1.6-34B (Liu, Li, Li, & Lee, 2024) | 64.67          | 72.17        | 62.37        | 40.6        |
| LLaVA-1.6-13B (Liu, Li, Li, & Lee, 2024) | 60.33          | 67.33        | 40.86        | 34.3        |
| LLaVA-1.6-7B (Liu, Li, Li, & Lee, 2024)  | 50.33          | 57.83        | 48.39        | 20.8        |
| LLaVA-1.5-7B (Liu, Li, Wu, & Lee, 2024)  | 49.62          | 63.00        | 43.01        | 20.1        |
| MiniCPM (Hu et al., 2024)                | 64.40          | 72.33        | 50.54        | 41.7        |
| Qwen-VL-Chat (Bai et al., 2023)          | 55.50          | 62.50        | 65.59        | 41.9        |

**Human Baseline** To establish performance benchmarks for average human understanding of combinational creativity, we recruited participants through the Prolific platform, the protocol of which was approved by a university IRB. These participants followed identical annotation guidelines to those used in creating our expert-annotated dataset, ensuring direct comparability between expert and average human performance.

**Models** We evaluate a diverse array of 11 models (OpenAI, 2024b, 2023; Anthropic, 2024a,b; Liu, Li, Wu, & Lee, 2024; Liu, Li, Li, & Lee, 2024; Hu et al., 2024; Bai et al., 2023) spanning different capabilities and architectures. Our selection includes closed-source models like GPT-4o, Gemini-1.5-Pro, and Claude-3.5-Sonnet, as well as open-source alternatives such as LLaVA (Liu, Li, Wu, & Lee, 2024; Liu, Li, Li, & Lee, 2024) and Qwen (Bai et al., 2023). The models cover various parameter sizes (34B, 13B, 8B, and 7B) and different Large Language Model (LLM) architectures (Yi-34B, Llama3-8B, Vicuna-7B, and Qwen-7B), enabling comprehensive evaluation across model scales, architectures, and accessibility levels.

## Results

**Identification** As shown in Tab. 1, GPT-4o demonstrates superior performance in identifying the basic objects in visual mashups, achieving the highest precision and recall scores. Gemini-1.5-Pro follows as a strong performer with 73.67% precision and 81.33% recall. Average human participants show moderate competency in this task, achieving 53.42% precision and 70.33% recall, indicating that even basic object identification in creative combinations presents notable challenges.

**Explanation** In analyzing combinatorial mechanisms, GPT-4o maintains its leading position, while Claude-3.5-Sonnet shows particularly interesting results—despite moderate performance in identification, it matches GPT-4o’s top performance in explanation. A notable observation is Gemini-1.5-Pro’s performance disparity, showing strong identification capabilities but relatively weaker explanation abilities. Average humans demonstrate strong explanatory capabilities, achieving 69.89% precision and ranking third, suggesting humans’ natural ability to understand combinatorial mechanisms.

**Implication** The implication task reveals the most distinctive performance patterns. Human experts demonstrate

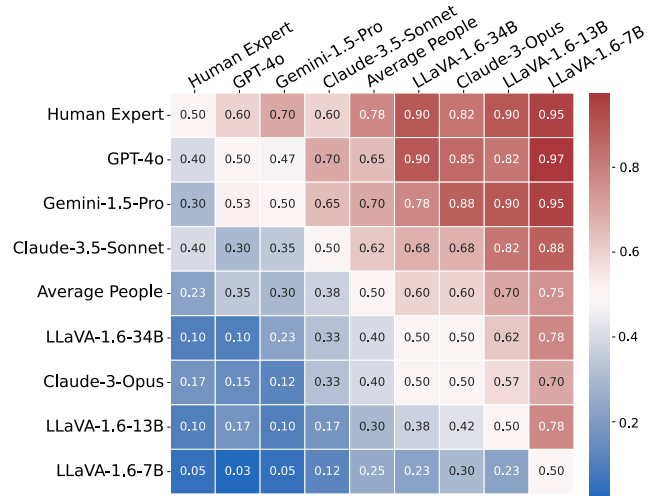


Figure 3: **Pairwise model comparison on implication task.** The heatmap displays winning probabilities, where each cell  $(i, j)$  shows the win rate of row model  $i$  vs. column model  $j$ . Darker red indicates higher win rates, while darker blue represents lower win rates.

superior understanding with a 78.3% winning rate, significantly outperforming all models. Average participants achieve a 51% winning rate, positioning in the middle range of performance. Among models, GPT-4o leads the artificial systems, followed by GPT-4V and Gemini-1.5-Pro, all achieving winning rates above 70%. These patterns are further confirmed through pair-wise comparisons (Fig. 3), where human experts consistently maintain the highest winning rate, while average participants typically fall within the middle performance range across model comparisons.

## Analysis and Discussions

Based on our comprehensive analysis of Tab. 1 and Fig. 3 and experimental data, we present three key findings.

**State-of-the-art VLMs have achieved above-average human performance in comprehension tasks** Our results demonstrate that leading models like GPT-4o and Claude-3.5-Sonnet have surpassed average human performance across all three tasks in understanding combinational creativity.

**Fusion combinational type shows higher cognition rates than replacement** Our visual mashups exhibit two distinct combination types: fusion and replacement. Fusion combinations merge characteristics from different basic objects, creating a blended object that preserves visual traits from both sources (see Fig. 4b). Replacement combinations involve one basic object substituting another within the structure (see Fig. 4a). Analysis of 11 models reveals higher precision metrics for fusion combinations in 9 models (see Fig. 5), mirroring typical human performance patterns. This disparity stems from the inherent nature of these combinations. Fusion combinations offer more readily identifiable characteristics from both basic objects, facilitating easier deconstruction. Conversely, replacement combinations demand a deeper understanding of contextual relationships or shared characteristics, as they predominantly display traits from one basic object while retaining only minimal characteristics from the other.



(a) The **replacement** of dynamite with soda cans. (b) The **fusion** between a fish and a toothpaste.

Figure 4: **Two types of combination in comprehension tasks.** (a) **Replacement** maintains functional or visual similarity while substituting for safer or more accessible alternatives. (b) **Fusion** merges two unrelated concepts to create a novel composite that inherits properties from both sources.

**Semantic attribute range and depth define implication quality** Our comparison of expert and average human implications reveals two critical factors determining interpretation quality: semantic attribute range and depth. For example, in Fig. 4a, an average person’s simple interpretation “A taste explosion” contrasts with the expert’s more nuanced analysis: “Drinking soda in excess is like detonating explosives within your body, slowly destroying your health.” The expert interpretation incorporates additional semantic attributes like “in excess” and “destroy health,” extending beyond the basic “taste” attribute. Similarly, in Fig. 4b, the profound metaphorical interpretation “Humanity relentlessly extracts resources” transcends the more literal observation “The fish is killed for its roe and packaged for human consumption.” While the average interpretation correctly identifies “human consumption,” it doesn’t explore deeper semantic connections between fish consumption and resource extraction. These examples illustrate how varying ranges and depths of semantic attributes lead to distinctly different levels of interpretation.

## Experiment 2: How to Induce VLMs’ Combinational Creativity?

Building on conceptual blending theory (Fauconnier & Turner, 2003), we hypothesize that explicitly incorporating a three-step thinking process would induce or enhance VLMs’ ability to generate products exhibiting combinational creativity.

### Task Setups

This task requires VLMs to generate creative images by combining provided objects to convey specific themes (see Fig. 2b). These thematic constraints are essential, as they enable systematic comparison of both novelty and usefulness in the generated images, two key dimensions in evaluating creative products.

Drawing from our annotated visual mashups dataset, we structured the input information across three levels: ① **Identification**: the input objects available for potential combination; ② **Explanation**: the shared attributes that may guide the combination process; ③ **Implication**: the semantic meaning of the creative output, which serves as a source for the theme.

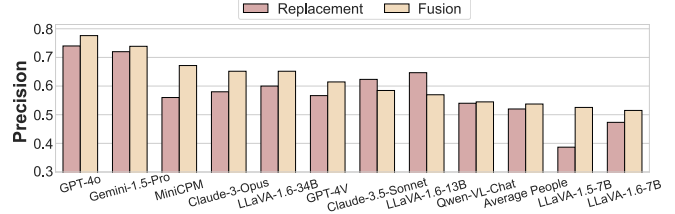


Figure 5: **Precision in identification task by combination type: replacement vs. fusion.** Precision metrics across combination categories show consistently higher precision for replacement-based combinations compared to fusion-based ones.

Our central research question examines whether explicitly incorporating a three-step IEI thinking process would enhance the combinational creativity manifested in the generated images. To investigate this, we compare three conditions:

1. **Identification+Implication (II)**: Models are instructed to creatively combine the input objects while considering the theme, using chain-of-thought prompting (Wei et al., 2022) without any external guidance in the combinational process.
2. **Identification+Explanation+Implication (IEI)**: Models are explicitly instructed to follow a three-step IEI thinking process, with additional guidance focusing on the potential shared attributes.
3. **Human Expert**: The original artist-generated visual mashups that serve as our reference point.

For both **II** and **IEI** conditions, we begin by providing the inputs as prompts to GPT-4o. This process involves combining basic objects at the conceptual level and generating appropriate prompts for text-to-image generative models. These generated prompts are then used to drive the image-generation process.

### Experimental Setups

**Human Evaluations** To assess the effectiveness of our approach, we recruited 50 participants through Prolific for preference evaluation. Participants were tasked with ranking image triplets (i.e., **Human Experts**, **II**, and **IEI**) from different experimental conditions based on two key criteria: image novelty and effectiveness of conveying the theme. To minimize potential biases, we randomized the presentation order of triplets across participants.

**Models** We evaluated our method across five state-of-the-art text-to-image generation models, encompassing both closed- and open-source models. The selected models included Midjourney (Midjourney, 2024), Flux-1.1-pro (Labs, 2024b), DALL-E-3 (OpenAI, 2024a), Flux-1.0-dev (Labs, 2024a), and Stable-Diffusion-3-medium (Esser et al., 2024). This diverse selection of models affords an examination of the generalizability of our approach across different generative models.

### Results

Fig. 6 presents human evaluation results of image triplets across different generation conditions. The Friedman test revealed significant differences in ranking scores among the three conditions for all models ( $p < 0.0001$ ). Post-hoc analysis revealed that **IEI**-generated images significantly outperformed **II**-generated images, while artist-generated images maintained superior performance overall.



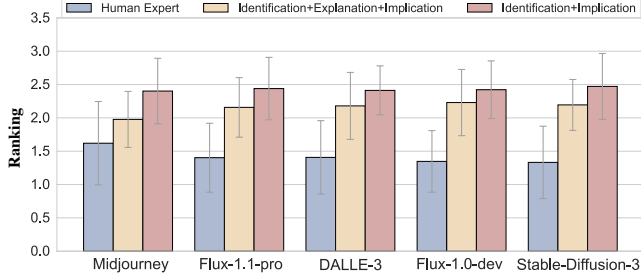


Figure 6: **Human preference rankings across in generation task.** We compare rankings from Human Expert, **IEI**, and **II** assessments. (Lower scores indicate better performance; error bars show variance.)

Among the evaluated models, Midjourney demonstrated the strongest performance. The **IEI** and **II** approaches achieved mean rankings of 1.98 and 2.40, respectively, leading ahead of other models’ performance. Notably, human-expert artworks showed the highest ranking variance compared to Midjourney-generated images, suggesting more diverse opinions on human-created art. For less sophisticated models, participants more consistently identified the superiority of human-created images, indicating that these models produce outputs more readily distinguishable from human artworks.

Fig. 7 illustrates the winning rates comparing **II** and **IEI** conditions across models. Midjourney exhibited superior performance in both conditions, with **IEI** enhancing its winning rate from 0.26 to 0.35. All models demonstrated improved winning rates under the **IEI** condition compared to their **II** performance, though improvement magnitudes varied. The performance differential between conditions diminished for lower-performing models, with Stable-Diffusion-3 showing minimal improvement from **IEI** enhancement.

## Analysis and Discussions

### IEI enhances creativity independent of prompt length

The primary distinction between **IEI** and **II** conditions lies in incorporating the explanation level of combinational creativity as a conceptual guiding mechanism. As evidenced in Figs. 6 and 7, **IEI** yielded significantly superior results compared to **II**. This improvement was particularly pronounced in more sophisticated models, which demonstrated greater benefits from the additional guidance. We attribute this pattern to these models’ enhanced capability to process and leverage the increased conceptual complexity in generating higher-quality images.

To verify that these improvements were not simply due to more verbose instructions, we conducted analyses of prompt lengths. An independent samples t-test comparing prompt lengths revealed no significant difference between **II** ( $M = 487.95$ ) and **IEI** ( $M = 514.53$ ) conditions ( $t = -0.84, p = 0.404$ ). To further investigate this relationship, we examined all instances where **IEI** outperformed **II**. In 46.0% of these successful cases, **IEI** prompts were shorter than their **II** counterparts, demonstrating that the explanation level’s effectiveness stems not from increased verbosity.

These findings suggest that introducing the three-step **IEI** thinking process, especially the explanation level of combinational creativity, provides essential conceptual guidance that

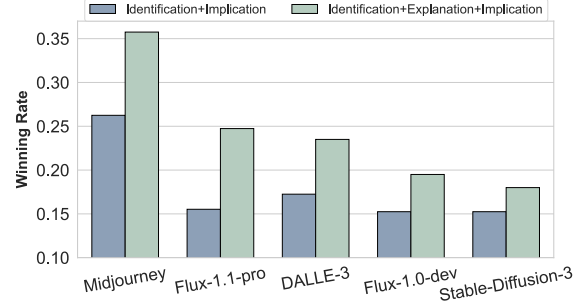


Figure 7: **Winning rates against human experts in generation task.** Win percentages comparing **II** and **IEI** settings against human expert assessment, demonstrating relative performance differences between the two approaches in human evaluation.

facilitates more meaningful concept integration. The improvement in creative output appears to derive from the structural and semantic guidance provided by the explanation level rather than from simply providing more detailed instructions to the image generation models.

**Text-to-image models are the bottleneck in generating visual combinational creativity images** Our analysis reveals a significant disparity between conceptual combination and visual execution capabilities in current AI systems. Using GPT-4o for concept-level object combination showed promising results, with 90% (36 out of 40) of DALL-E-3 generated examples successfully integrating basic objects and effectively conveying intended themes. However, the translation from conceptual description to visual output proved more challenging, with 27.8% (10 out of 36) of well-crafted descriptions failing to produce satisfactory visual results.

This performance gap illuminates a crucial challenge. While LLM demonstrate strong conceptual combination abilities, text-to-image models often struggle to execute these creative concepts faithfully. This observation aligns with our experimental findings, in which more sophisticated models showed greater improvement with enhanced instructions, suggesting that the ability to follow instructions significantly impacts the quality of visual generation. While these results highlight the potential of LLM as creative design tools for skilled human artists, they also indicate that advancing text-to-image models’ instruction-following capabilities remains a critical priority for achieving fully automated creative processes.

## Conclusion

This work advances the understanding of combinational creativity in VLMs through a three-level **IEI** framework. We contribute a framework-annotated dataset of visual mashups for probing understanding capabilities, and an **IEI**-based process that enhances generative performance in VLMs. Experiments reveal that while state-of-the-art models can surpass average human performance in recognizing combinational creativity, they still fall short of the expert level. We show that incorporating the combinational process into model operation enhances creative output, validating our framework’s utility. This work establishes a foundation for analyzing combinational creativity, with our dataset and evaluation methodologies serving as resources for future research in creative AI systems.

## Acknowledgments

We gratefully acknowledge Chengdong Ma and Qinghao Wang for their valuable assistance with data exploration, Yujia Peng for her suggestions on the human study, and Zhen Chen for her efforts in figure preparation. This work is supported in part by the National Science and Technology Major Project (2022ZD0114900), the National Natural Science Foundation of China (62376031), the Beijing Nova Program, the State Key Lab of General AI at Peking University, the PKU-Bingji Joint Laboratory for Artificial Intelligence, and the National Comprehensive Experimental Base for Governance of Intelligent Society, Wuhan East Lake High-Tech Development Zone.

## References

- Akula, A. R., Driscoll, B., Narayana, P., Changpinyo, S., Jia, Z., Damle, S., ... others (2023). Metaclue: Towards comprehensive visual metaphors research. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Anthropic. (2024a). *Claude 3*. <https://www.anthropic.com/news/claude-3-family>.
- Anthropic. (2024b). *Claude 3.5*. <https://www.anthropic.com/news/3-5-models-and-computer-use>.
- Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., ... Zhou, J. (2023). Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.
- Bellemare-Pepin, A., Lespinasse, F., Thölke, P., Harel, Y., Mathewson, K., Olson, J. A., ... Jerbi, K. (2024). Divergent creativity in humans and large language models. *arXiv preprint arXiv:2405.13012*.
- Boden, M. (2009). Creativity: How does it work. *The idea of creativity*, 28, 237–50.
- Boden, M. A. (1998). Creativity and artificial intelligence. *Artificial intelligence*, 103(1-2), 347–356.
- Boden, M. A. (2004). *The creative mind: Myths and mechanisms*. Routledge.
- Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... others (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.
- Chakrabarty, T., Laban, P., Agarwal, D., Muresan, S., & Wu, C.-S. (2024). Art or artifice? large language models and the false promise of creativity. In *ACM Conference on Human Factors in Computing Systems (CHI)*.
- Chen, H., & Ding, N. (2023). Probing the “creativity” of large language models: Can models produce divergent semantic association? In *Annual Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., ... others (2024). Chatbot arena: An open platform for evaluating llms by human preference. In *Proceedings of International Conference on Machine Learning (ICML)*.
- Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., ... others (2024). Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of International Conference on Machine Learning (ICML)*.
- Fauconnier, G., & Turner, M. (2003). Conceptual blending, form and meaning. *Recherches en Communication*, 19, 57–86.
- Franceschelli, G., & Musolesi, M. (2024). On the creativity of large language models. *AI & SOCIETY*, 1–11.
- Gabora, L., & Kaufman, S. B. (2010). Evolutionary approaches to creativity. In *The cambridge handbook of creativity* (pp. 279–300). Cambridge University Press.
- Guzdial, M. J., & Riedl, M. O. (2018). Combinatorial creativity for procedural content generation via machine learning. In *Aaai workshops*.
- Han, J., Park, D., Shi, F., Chen, L., Hua, M., & Childs, P. R. (2019). Three driven approaches to combinational creativity: Problem-, similarity- and inspiration-driven. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, 233(2), 373–384.
- Han, J., Shi, F., Chen, L., & Childs, P. R. (2018). The combinator-a computer-based tool for creative idea generation based on a simulation approach. *Design Science*, 4, e11.
- Hessel, J., Marasović, A., Hwang, J. D., Lee, L., Da, J., Zellers, R., ... Choi, Y. (2023). Do androids laugh at electric sheep? humor “understanding” benchmarks from the new yorker caption contest. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Holyoak, K. J., & Thagard, P. (1996). *Mental leaps: Analogy in creative thought*. MIT press.
- Hu, S., Tu, Y., Han, X., He, C., Cui, G., Long, X., ... others (2024). Minicpm: Unveiling the potential of small language models with scalable training strategies. In *Conference on Language Modeling (CoLM)*.
- Ji, A., Kojima, N., Rush, N., Suhr, A., Vong, W. K., Hawkins, R., & Artzi, Y. (2022). Abstract visual reasoning with tangram shapes. In *Annual Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Ji, J., Hong, D., Zhang, B., Chen, B., Dai, J., Zheng, B., ... Yang, Y. (2024). Pku-saferlhf: A safety alignment preference dataset for llama family models. *arXiv preprint arXiv:2406.15513*.
- Koivisto, M., & Grassini, S. (2023). Best humans still outperform artificial intelligence in a creative divergent thinking task. *Scientific reports*, 13(1), 13601.
- Labs, B. F. (2024a). *black-forest-labs/flux* (github repository). <https://github.com/black-forest-labs/flux>.
- Labs, B. F. (2024b). *Flux-1.1-pro*. <https://blackforestlabs.ai>.
- Liu, H., Li, C., Li, Y., & Lee, Y. J. (2024). Improved baselines with visual instruction tuning. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2024). Visual instruction tuning. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*.
- Ma, Y., Zhang, C., & Zhu, S.-C. (2023). Brain in a vat: On missing pieces towards artificial general intelligence in large language models. *arXiv preprint arXiv:2307.03762*.
- Mehrotra, P., Parab, A., & Gulwani, S. (2024). Enhancing creativity in large language models through associative thinking strategies. *arXiv preprint arXiv:2405.06715*.
- Midjourney. (2024). *Midjourney*. <https://www.midjourney.com/explore?tab=top>.
- OpenAI. (2023). *Gpt-4v(ision) system card*. [https://cdn.openai.com/papers/GPTV\\_System\\_Card.pdf](https://cdn.openai.com/papers/GPTV_System_Card.pdf).
- OpenAI. (2024a). *DALLE-3*. <https://openai.com/index/dall-e-3/>.
- OpenAI. (2024b). *GPT-4o*. <https://openai.com/index/hello-gpt-4o/>.
- Orwig, W., Edenbaum, E. R., Greene, J. D., & Schacter, D. L. (2024). The language of creativity: Evidence from humans and large language models. *The Journal of Creative Behavior*, 58(1), 128–136.
- Peng, Y., Han, J., Zhang, Z., Fan, L., Liu, T., Qi, S., ... Zhu, S.-C. (2024). The tong test: Evaluating artificial general intelligence through dynamic embodied physical and social interactions. *Engineering*, 34, 12–22.
- Team, G., Georgiev, P., Lei, V. I., Burnell, R., Bai, L., Gulati, A., ... others (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.
- Tian, P., Huang, T., Liu, M., Jiang, D., Spangher, A., Chen, M., ... Peng, N. (2024). Are large language models capable of generating human-level narratives? In *Annual Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... others (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*.