Let Androids Dream of Electric Sheep: A Human-Inspired Image Metaphor Understanding and Reasoning Framework

Chenhao Zhang[1,2] Yazhe Niu[1,3]
[1]Shanghai AI Laboratory [2]Huazhong University of Science and Technology
[3]The Chinese University of Hong Kong MMLab
zhangchenhao@pjlab.org.cn niuyazhe314@outlook.com

Abstract

Metaphorical comprehension in images remains a critical challenge for AI systems, as existing models struggle to grasp the nuanced cultural, emotional, and contextual implications embedded in visual content. From a cognitive perspective, image metaphors often require social reasoning signals (e.g., intention- and perspective-related cues) and benefit from a dual-process pipeline that combines fast perceptual abstraction with slower, evidence-grounded deliberation. While multimodal large language models (MLLMs) excel in general Visual Question Answer (VQA) tasks, they struggle with a fundamental limitation on image metaphor tasks: contextual gaps that obscure the relationships between different visual elements and their abstract meanings. Inspired by the human cognitive process, we propose ***Let Androids Dream (LAD)****, a novel framework for image metaphor understanding and reasoning. LAD addresses contextual missing through the three-stage framework: (1)* ***Perception****: converting visual information into rich and multi-level textual representations, (2)* ***Search****: iteratively searching and integrating cross-domain knowledge to resolve ambiguity, and (3)* ***Reasoning****: generating context-alignment image metaphor via explicit reasoning.*
Our framework with the lightweight GPT-4o-mini model achieves SOTA performance compared to 15+ MLLMs on English image metaphor benchmark and a huge improvement on Chinese benchmark, performing comparable with the Gemini-3.0-pro model on Multiple-Choice Question (MCQ) and outperforms the GPT-4o model 36.7% on Open-Style Question (OSQ). Generalization experiments also show that our framework can effectively benefit general VQA and visual reasoning tasks. Additionally, our work provides new insights into how AI can more effectively interpret image metaphors, advancing the field of vision-language reasoning and human-AI interaction. We fully open-scoured datasets, benchmarks, and method codes.

1. Introduction

> *Do androids dream of electronic sheep? The question actually has two levels: The first level is to ask if androids dream, and the second level is to ask if they dream of electronic sheep.*
>
> – Philip K. Dick (1968)

Metaphors are not just abstract concepts found in literature; they are also prevalent in our daily lives. For instance, when we say "time is money" or "life is a journey," we are using metaphors to convey complex ideas in a more contextual and understandable way. These metaphors highlight the integral role that metaphoric thinking plays in human communication and cognition. Just as we use metaphors to make sense of the world around us, we aim to enable AI to understand metaphors in a human-like manner. In linguistic terms, as George Lakoff and Mark Johnson elaborated in "Metaphors We Live By" [13], metaphors are not merely ornamental language devices but fundamental cognitive tools that allow us to conceptualize our surroundings. Metaphors possess characteristics such as systematicity, the creation of similarity, and imaginative rationality. Through cross-domain mapping, one concept can be used to comprehend another, allowing for a more insightful interpretation.

With the rapid advancement of large language models (LLMs), models such as OpenAI o1 [22], DeepSeek-R1 [4], and QwQ [28] have demonstrated remarkable text-reasoning capabilities. However, a significant amount of knowledge in the real world cannot be fully represented by text alone. Visual information, for instance, contains a wealth of knowledge that is not easily captured through text. As a result, there has been a growing interest in integrating visual information into text-reasoning tasks. Compared to language, vision is inherently complex, with its diverse representation of information, subjective understanding, and the difficulty in quantifying its information. In recent years, multimodal reasoning models such as QVQ [27] and K1.5 [26] have achieved outstand-

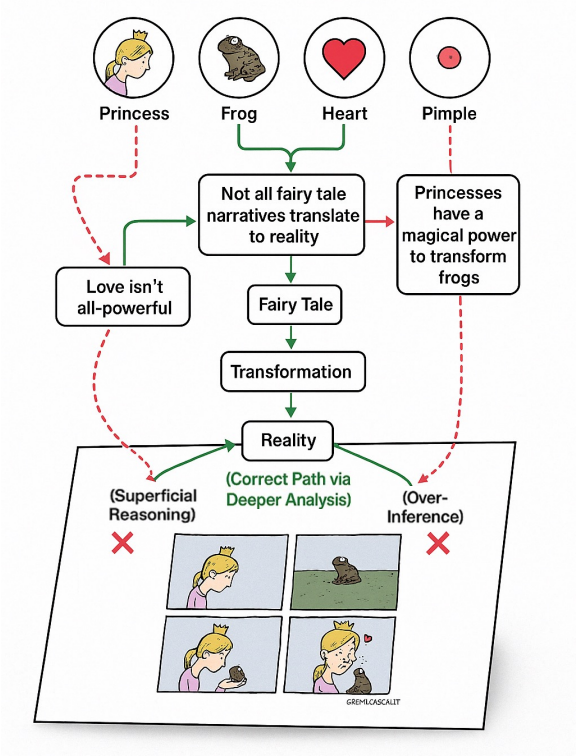


Figure 1. An image is worth a thousand words: For the image metaphor understanding task, different elements' combination leads to different thinking paths, but the correct path needs all elements with multiple reasoning thoughts.

ing performance. For example, K1.5 model has reached a high score on math, code and multimodal reasoning benchmarks [11, 16, 18, 29, 39]. However, these models still perform poorly on image metaphor questions [17, 40]. They tend to focus on the superficial elements of the image, neglecting the deeper connections and emotional expressions among these elements, as shown in Figure 1. It is important to note that these models excel at logical reasoning tasks, which are based on a different set of cognitive principles compared to image metaphor tasks. In contrast to the VQA task, which primarily centers on concrete image comprehension, the image metaphor entails a stronger emphasis on abstract meaning and higher-order reasoning capabilities. It is not a simple logical reasoning task and requires a different method to understand and generate implications. It requires the model to understand complex and abstract information, such as metaphors, symbols, and emotions in the image, rather than just the concrete contents.

image metaphor tasks consist of two main aspects: understanding and generation. Understanding image metaphor is a more complex and challenging task than understanding conventional images. It requires advanced cognitive abilities such as multi-hop reasoning and a sophisticated theory of mind (ToM), which are inherent to human cognition [17, 40]. In practice, this often manifests as social reasoning: models must infer the implied intent, perspective, or communicative goal behind a visual narrative, rather than only describing what is literally shown. Compared to understanding, generating implication is even more difficult. The fundamental challenge stems from the lack of contextual understanding of the key elements and internal relationships of the image. This lack of context hinders our ability to decipher the intended message or to create images that effectively convey specific meanings. Without the background of cultural, historical, or environmental context, the significance of key visual components remains elusive, impeding both interpretation and creative expression.

Existing approaches largely fall into (i) explicit metaphor mapping (e.g., CLOT [43]) and (ii) implicit reasoning (e.g., untrained chain-of-reasoning in C4MMD [35]). Mapping is brittle under many-to-many relations and shifting culture; implicit reasoning faces out-of-domain context gaps.

To address these issues, inspired by how humans resolve ambiguity, we argue the core difficulty is *contextual missing*. We propose a framework that converts visual cues into controllable textual abstractions and iteratively retrieves out-of-domain knowledge to support deeper reasoning. Experiments from both Multiple-Choice Question and Open-Style Question consistently verify the superiority of the proposed framework.

Our key contributions are listed as follows:

- We systematically analyze image metaphor tasks, highlighting their strong reliance on contextual and social reasoning (including ToM-related inference). We find the key difficulty lies in contextual missing, and propose a cognitively motivated direction for solving these tasks – Contextual Alignment.
- We propose a novel human-inspired three-stage framework Let Androids Dream (LAD) for image metaphor understanding and reasoning, including Perception, Search and Reasoning, which instantiates a dual-process pipeline (fast perceptual abstraction followed by slower, evidence-grounded deliberation). Our LAD implements the lightweight GPT-4o-mini model to achieve SOTA on English image metaphor benchmark (1300+ questions) and a huge improvement on Chinese image metaphor benchmark (800 questions), comparable with the Gemini-3.0-pro model and other top closed-source models on Multiple-Choice Question (MCQ). Generalization experiments also show that our framework can effectively benefit general VQA and visual reasoning tasks.
- We design the challenging Open-Style Question (OSQ) with comprehensive metrics to automatically evaluate the image metaphor tasks. This metric aligns 95.7% with human annotations, making it more suitable for diverse evaluation. Our LAD outperforms the GPT-4o model 36.7% on OSQ.

2. Related Work

2.1. Image Metaphor

Image metaphor involves humor, sarcasm, and broader metaphorical meaning. Early work often targeted specific aspects, such as humor recognition [7, 8] and sarcasm detection [5]. With the rise of LLMs/MLLMs, recent efforts emphasize more comprehensive evaluation: DeepEval [37] proposes a taxonomy, II-Bench [17] provides an English benchmark, and CII-Bench [40] extends it to Chinese. These benchmarks highlight that implication understanding requires multi-hop reasoning and theory of mind (ToM) style social/context reasoning [17, 40]. Existing solutions broadly fall into (i) explicit metaphor mapping (e.g., CLOT [43]), which is hard to scale to many-to-many mappings and evolving cultural references, and (ii) implicit reasoning via training-free CoT (e.g., C4MMD [35]), which can suffer from a large out-of-domain search space and context drift. Our LAD instead formulates implication as *contextual alignment*: it converts images into controllable textual abstractions and enriches them with targeted retrieval, aligning with human-inspired dual-process intuition-to-deliberation and ToM-relevant social reasoning.

2.2. Vision-language Reasoning

The rapid advancement of LLMs has demonstrated remarkable text reasoning capabilities, as evidenced by models such as o1 [22], DeepSeek-R1 [4], and QwQ [28, 36]. However, many reasoning signals are visual and contextual, motivating vision-language reasoning. Existing work mainly follows three directions: (1) converting vision into rich textual grounding (e.g., LLAVA-COT [34], Mulberry [38]); (2) multi-turn interaction to elicit fine-grained cues (e.g., VoCoT [15], V* [31]); and (3) tool-augmented reasoning that externalizes intermediate steps (e.g., Visual Sketchpad [9], Whiteboard-of-Thought [19]). In this landscape, LAD emphasizes a human-inspired separation of fast perception and slower deliberation, with a meta-cognitive routing step for evidence acquisition when context is missing.

3. Method

Inspired by the human cognitive process, we introduce a new paradigm for solving image metaphor tasks – Contextual Alignment. We have a detailed discussion for this point in Section 1 and Section 5. Therefore, we propose Let Androids Dream (LAD), a novel framework for image metaphor understanding and reasoning. This framework operates through the three-stage framework, as shown in Figure 2: (1) **Perception**: converting visual information into rich and multi-level texts, (2) **Search**: iteratively searching and integrating cross-domain knowledge to resolve ambiguity, and (3) **Reasoning**: generating context-alignment analysis via explicit reasoning.

3.1. Stage I: Perception

The initial stage, *Perception*, aims to transform raw visual inputs into structured, hierarchical textual representations, mirroring the human cognitive process of initial intuition-driven observation and subsequent identification of key elements. This stage operates in a manner analogous to human System 1 (intuitive, holistic processing) and System 2 (analytical, focused processing).

First, we utilize MLLM to process the input image and produce a detailed textual narrative. This description captures coarse-grained visual information, including discernible text within the image, prominent colors, overall layout, and salient objects or entities. This step provides a holistic foundational understanding of the content of the image. Following this, we derive a fine-grained keyword set. The MLLM condenses the above image description into a concise set of approximately 7 keywords. These keywords are specifically chosen to encapsulate critical aspects relevant to implication understanding, such as the perceived emotion, the domain or context (e.g., political, social, cultural) and any rhetorical devices that might be visually suggested. Keywords also re-emphasize crucial textual elements or entities identified in the description. This two-tiered representation, comprising a rich description and focused keywords, provides a robust foundation for the subsequent *Search* and *Reasoning* stages by converting unstructured visual data into actionable textual information. The keywords, in particular, serve as vital cues for guiding the knowledge retrieval in stage II.

3.2. Stage II: Search

The *Search* stage addresses semantic ambiguities and enhances contextual comprehension by iteratively retrieving and integrating cross-domain knowledge critical for interpreting image metaphors. This stage employs adaptive search, which dynamically selects the most appropriate search method. The process is systematically organized into three main phases: Plan, Search, and Summary.

1. Plan: The process begins by formulating targeted search queries. Using the keywords generated in Stage I, the MLLM, guided by a prompt specifically designed for image metaphor tasks, generates five different levels of search questions. These questions aim to uncover latent meanings, cultural references, or background information pertinent to the image metaphors.

2. Search: This phase executes the search based on the generated questions, employing the Self-Judge mechanism to determine the optimal search strategy for each question.

(a) **Self-Judge**: The MLLM acts as a judge, assigning a confidence score to each search question. This score reflects criteria such as the perceived popularity or commonness of the knowledge required, relevance to real-time or recent events, and whether the question in-

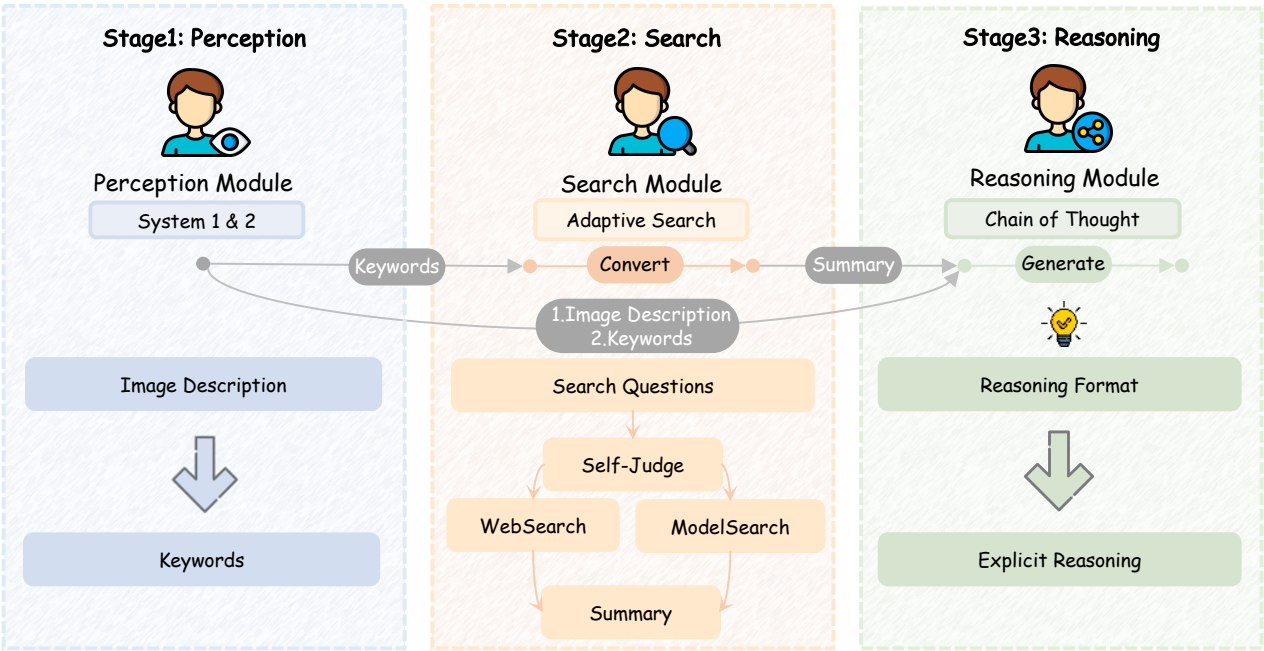


Figure 2. The general framework of Let Androids Dream (LAD), which includes three stages: (1) Perception: converting raw visual information into rich and multi-level textual representations, (2) Search: iteratively searching and integrating cross-domain knowledge to resolve ambiguity, and (3) Reasoning: generating context-alignment image metaphor interpretations via explicit reasoning.

volves contemporary internet slang or meme culture. Questions scoring high, indicating a need for up-to-date or niche information, are routed to WebSearch. Questions scoring low, suggesting the answer might reside within general world knowledge, are directed to ModelSearch. This adaptive routing optimizes for both knowledge coverage and inference efficiency.

(b) **ModelSearch**: For questions deemed suitable for internal knowledge retrieval, ModelSearch leverages the MLLM's own parametric memory. Using a specialized prompt, the model directly generates an answer based on its pre-trained knowledge base. This approach is efficient for recalling established facts or common concepts.

(c) **WebSearch**: For questions requiring external, dynamic, or highly specific information, WebSearch is invoked. Inspired by LLM search methods like MindSearch [3], but focusing on image metaphor tasks, our WebSearch component first employs the planner. The planner, acting as a high-level strategist, decomposes the initial search question into a series of more granular sub-questions. These sub-questions are structured into a directed acyclic graph (DAG), simulating a multi-step, exploratory information-seeking process. Subsequently, the searcher executes this plan. It performs hierarchical information retrieval for each sub-question from the internet, gathering relevant snippets and facts. This multi-agent method, with distinct planner and searcher modules, allows for parallel processing and dynamic refinement of the search strategy. The retrieved information for sub-questions is then synthesized to answer the original search question. This ensures access to recent developments and a broad spectrum of public knowledge, crucial for understanding contemporary image metaphors.

3. Summary: The raw outputs from the Search phase are refined into a concise search summary.

(a) **RankSummary:** The set of five question-answer pairs is evaluated. The MLLM ranks these pairs based on their relevance to understanding the core implication of the original image. The top three most relevant question-answer pairs are selected.

(b) **RefineSummary:** The selected pairs are further processed. The MLLM, guided by the ranking reason from the ranking step, rewrites and consolidates these pairs. This involves removing irrelevant or redundant information, reconciling diverse pieces of information, and potentially supplementing details to create a single, optimized, and concise search summary. This final summary serves as the enriched contextual input for Stage III.

3.3. Stage III: Reasoning

The final stage, *Reasoning*, performs explicit reasoning to derive contextually grounded interpretations of image metaphors. This stage synthesizes all previously gathered information — the hierarchical textual representations

from Stage I (descriptions and keywords) and the domain-enriched knowledge from Stage II — into a coherent implication framework.

For image metaphor tasks, we employ a specific reasoning format. The MLLM is prompted to articulate its reasoning trajectory using designated markers, such as "<think> ...</think>" special tokens. Within these markers, the model explicitly lays out its step-by-step reasoning process, connecting the visual cues, keywords, and external knowledge to arrive at the final image metaphor analysis and explanation. This domain-specific CoT method not only guides the model towards a more robust and grounded output, but also makes the inferential pathway transparent. The framework ultimately generates a contextually-aligned implication understanding that emerges from the integration of visual-semantic inputs and cross-domain knowledge, formalizing the LAD system's capacity for evidence-based visual reasoning.

3.4. LAD Pipeline

The Let Androids Dream (LAD) framework operates as a sequential pipeline, integrating the three distinct stages described in Figure 2 and Algorithm 1. Stage I (Perception) initiates the process. It takes an input image and employs the MLLM to generate a comprehensive image description. This description is then further processed to extract seven salient keywords. The outputs of this stage are the image description and the set of keywords. These keywords serve as the primary input for Stage II (Search). Here, the MLLM transforms the keywords into five targeted search questions. A self-judge mechanism then directs these questions to either ModelSearch (for internal knowledge retrieval) or WebSearch (for external, dynamic information). The resulting question-answer pairs are ranked for relevance, with the top three being selected and subsequently refined into a concise search summary. This search summary is the key output of Stage II. Finally, Stage III (Reasoning) receives the original image, the image description and keywords from Stage I, and the search summary from Stage II. The MLLM integrates these multi-modal inputs and, through an explicit reasoning process (guided by a structured CoT), generates the final image metaphor. This implication represents the culmination of the LAD pipeline's understanding and reasoning about the input image.

Algorithm 1 Let Androids Dream (LAD)

Input: Image IMG; Tasks T_{MCQ}, T_{OSQ}
Output: Answers A_{MCQ}, A_{OSQ}

```
// Stage I: Perception
img_dep ← MLLM.Perception(IMG)                    /* Gen. description. */
keywords ← MLLM.Perception(img_dep)               /* Gen. 7 keywords */
// Stage II: Search
search_qs ← MLLM.Plan(keywords)                   /* 5 questions for image metaphor */
all_qa ← ∅
for each q in search_qs do
    strategy ← MLLM.Self-Judge(q)
    if strategy = 'WebSearch' then
        answer ← WebSearch(q)                     /* External knowledge */
    end
    else if strategy = 'ModelSearch' then
        answer ← ModelSearch(q)                   /* Parametric knowledge */
    end
    all_qa.add((q, answer))
end
search_sum ← MLLM.Summary(img_dep, all_qa)        /* Rank top-3, refine */
// Stage III: Reasoning
A_MCQ ← MLLM.Reasoning(IMG, img_dep, keywords, search_sum, T_MCQ)
A_OSQ ← MLLM.Reasoning(IMG, img_dep, keywords, search_sum, T_OSQ)
return A_MCQ, A_OSQ

Function WebSearch(q)
    // Planner: Decompose query
    sub_qs ← MLLM.RewriteQuery(q)
    // Searcher: Hierarchical retrieval
    snippets ← SearchAPI.BatchQuery(sub_qs)       /* Titles, summaries, URLs */
    sel_urls ← MLLM.SelectPages(snippets, q)
    content ← PythonCrawler.FetchContent(sel_urls)
    // Summarizer: Generate answer
    summary ← MLLM.Summary(content, q)
    return summary
```

4. Experiment

4.1. Baselines

Models. To comprehensively compare with LAD, we carefully select a diverse range of MLLMs, encompassing both open-source and closed-source models, with the aim of covering a wide spectrum of model characteristics and scales. These models span parameter sizes from 7B to 300B, ensuring that models of varying complexity and capability are thoroughly assessed. In selecting the models, we focus on the following key aspects: 1) General and Reasoning models, 2) Open-Source and Closed-Source models, and 3) model parameter scaling law. The experiment setup is in Appendix A.

Evaluation. We report MCQ on II-Bench [17] and CII-Bench [40], and manually construct the 100-image high-level benchmark (Appendix C). To mitigate MCQ bias [14, 20, 41], we also evaluate generation with OSQ on the same images using fixed-question prompting and GPT-4o rubric scoring (human consistency in Appendix B). Further analyses are in Appendix D.

4.2. Multiple-Choice Question

4.2.1. Implementation Details

Our high-level benchmark includes diverse images such as comics, posters, illustrations, English and Chinese Internet memes, and Chinese traditional artworks, all rich in visual information and cultural significance. Each image is paired with one question, each offering six options with only one correct answer. The question is "What is the implication in

Model	Multiple-Choice Question		Open-Style Question	
	en	zh	en	zh
General Models				
Qwen2.5-VL-7B [2]	46%	40%	2.34	2.58
DeepSeek-VL2 [32]	46%	36%	2.82	2.86
GLM-4.1V-8B [42]	60%	52%	2.60	2.96
Gemini-2.0-flash [24]	70%	68%	1.60	3.12
Qwen2.5-VL-72B [2]	72%	56%	1.56	3.12
InternVL3-78B [44]	70%	74%	3.42	3.70
GLM-4V-plus [42]	64%	64%	3.01	3.12
Grok-3 [33]	66%	64%	3.24	2.96
Claude-4.0-Sonnet [1]	60%	62%	3.46	3.86
GPT-4o [21]	74%	58%	2.94	3.76
GPT-4.1 [21]	74%	62%	3.30	3.92
Vision-language Reasoning Models				
Gemini-2.0-flash-thinking [24]	64%	68%	1.66	2.84
QVQ-72B [27]	62%	56%	3.10	3.42
Doubao-1.5-thinking-vision-pro [23]	66%	66%	3.16	3.90
Grok-3-reasoning [33]	74%	64%	3.06	2.92
Gemini-3.0-pro [25]	**76%**	**76%**	3.82	**3.96**
Our Method				
GPT-4o-mini [21]	44%	42%	2.98	3.36
+ LAD (Stage I + III)	68% ↑	44% ↑	3.84 ↑	3.58 ↑
+ LAD (Stage I + II + III)	74% ↑	52% ↑	**4.02** ↑	3.66 ↑
Improv.	+30 (68.2%)	+10 (23.8%)	+1.04 (34.9%)	+0.3 (8.9%)

Table 1. Overall results of different models on Multiple-Choice Question and Open-Style Question. The best-performing model in each category is **in-bold**, and the second best is underlined. Performance differences relative to base models are shown as colodarkred subscripts: ↑ for improvements, ↓ for declines.

this image?" (mostly) or different levels of image understanding, such as overarching interpretation and nuanced details. A case study of different methods on MCQ is in Figure 3.

4.2.2. Results and Analysis

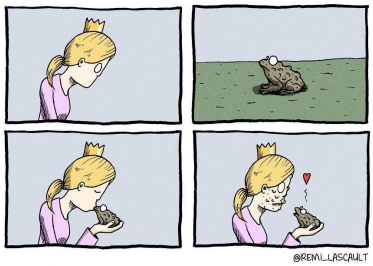
Table 1 reports MCQ results. LAD achieves SOTA with GPT-4o-mini: on English MCQ it matches Gemini-3.0-pro and exceeds Claude-4.0-Sonnet by 23.3%; on Chinese MCQ it is comparable to GPT-4o and surpasses DeepSeek-VL2 by 44.4%. Relative to base GPT-4o-mini, LAD yields +68.2% (en) and +23.8% (zh) improvements. We also observe limited advantage of current reasoning models over general models on implication, suggesting that RL-based reasoning gains may not directly transfer to this socially/contextually grounded task.

4.3. Open-Style Question

4.3.1. Implementation Details

Evaluation Metric. To comprehensively assess MLLMs' understanding of image metaphor, we develop a multi-faceted evaluation metric. This metric is designed to probe both the surface-level information readily apparent in the image and the deeper emotion, domain and rhetorical skills that inform its creation and interpretation. Our evaluation metric encompasses five key perspectives: ***Surface-level Information***, ***Emotional Expression***, ***Domain and Context***, ***Rhetorical Skills***, and ***Deep Implications***. For each perspective, we give its detailed description in Figure 4.

MLLM-based Automatic Evaluation. To evaluate image metaphor comprehension in MLLMs, we develop an MLLM-based evaluation standard based on evaluation metrics, as illustrated in Figure 4. Our experiment utilize the same dataset from MCQ experiment, comprising 50 English images and 50 Chinese images. We employ human-written descriptions and implication interpretations as ground truth. We choose the same MLLMs with MCQ experiment to generate image metaphors for these images, which are subsequently scored using GPT-4o and our evaluation standard. The evaluation prompt is in Appendix F. To validate the model's scoring efficacy, we enlist 16 PhD students and researchers well-versed in English and Chinese metaphorical



Question

The metaphor for this image is?
Options:
(A) True love can change a person's physical form.
(B) Love isn't all-powerful.
(C) Not all fairy tale narratives translate to reality.
(D) Princesses have a magical power to transform frogs.
(E) The core message is concerned with environmental conservation.
(F) The underlying theme is the necessity of taking risks.

Ground Truth Solution

<think> The comic likely depicts a princess and frog, invoking fairy tale expectations (transformation via kiss/love). However, the humor/twist comes from subverting this expectation (e.g., princess gets warts, frog stays a frog, or another unexpected outcome). This contrast highlights that idealized fairy tale narratives don't always match reality. </think>
<answer> C </answer>

End2End

<think>...focuses on the princess accepting the frog as is, suggesting love doesn't need transformation...</think>
<answer> B </answer> ✗

(Superficial Reasoning)

CoT

<think>...connects affection (heart symbol) to the traditional fairy tale narrative of transformation...</think>
<answer> D </answer> ✗

(Over-Inference)

LAD

<think>...identifies transformation motif... humor and visual elements suggest commentary on unrealistic fairy tales... points to idea that fairy tales don't always translate to reality...</think>
<answer> C </answer> ✓

(Correct Path via Deeper Analysis)

Figure 3. A case study of different methods on Multiple-Choice Question. The *End2End* method shows superficial reasoning and the *CoT* method shows over-inference, while our *LAD* framework shows the correct path via more contextual alignment analysis. The full prompt is listed in Appendix F.

Evaluation Metric	Evaluation Standard
1. Surface-level Information: • Identification of primary entities within the image • Analysis of color composition and application • Recognition of intricate details and their significance **2. Emotional Expression:** • Identification of conveyed emotions (e.g., tranquility, intensity, melancholy) • Depth of emotional resonance and its alignment with the image's theme • Consistency of emotional expression across the image's elements **3. Domain and Context:** • Recognition of the image's domain (e.g., art, commerce, social commentary) • Contextualization within its cultural, historical, or societal background • Evaluation of the image's innovation within its domain **4. Rhetorical Skills:** • Identification of rhetorical devices (e.g., symbolism, contrast, personification) • Analysis of how rhetorical techniques enhance the image's expression • Integration of rhetorical devices with metaphorical implications to create a cohesive interpretation **5. Deep Implications:** • Recognition of metaphorical elements and their layered meanings • Depth of interpretation of philosophical, cultural, or social values embedded in the image • Evaluation of the originality and creativity in metaphorical interpretation	**[1 point]:** Fails to capture key elements within the image (such as text, and important entities). Does not identify emotions, domain, or rhetorical devices. Only provides a superficial description of surface-level information, lacking depth and creativity, with a significant gap from the standard answer. **[2 points]:** Captures some key elements within the image, but the identification of emotions, domain, and rhetorical devices is vague. The description of surface-level information is relatively complete, but there is a clear deficiency in exploring deeper meanings, showing a noticeable gap from the standard answer. **[3 points]:** Effectively captures key elements within the image and initially identifies emotions, domain, and rhetorical devices. The description of surface-level information is relatively accurate, and there is some relevant expression of deep meanings. However, there is still room for improvement in depth and creativity, and it is generally close to the standard answer. **[4 points]:** Accurately captures key elements within the image and clearly identifies emotions, domain, and rhetorical devices. The description of surface-level information is detailed and precise, with a relatively deep exploration of deep meanings, demonstrating a certain level of creativity and depth. It is largely consistent with the standard answer but may have minor deficiencies in some details or depth. **[5 points]:** Accurately and precisely captures key elements within the image and profoundly identifies emotions, domain, and rhetorical devices. The description of surface-level information is comprehensive and precise, with unique insights into deep meanings, skillfully integrating image elements with metaphorical implications. It demonstrates exceptional creativity and depth, is highly consistent with the standard answer, and shows a profound grasp of metaphor creation and cultural understanding.

Figure 4. Evaluation metric and evaluation standard of Open-Style Question.

imagery to independently score the dataset. The human-model scoring consistency reached 95.7%, affirming the method's validity. The detailed human-model consistency study is in Appendix B.

4.3.2. Results and Analysis

Table 1 reports OSQ results. LAD achieves SOTA with GPT-4o-mini: on English OSQ it outperforms Gemini-3.0-pro by 5.2%, GPT-4o by 36.7%, and Claude-4.0-Sonnet by 16.2%; on Chinese OSQ it remains competitive with top closed-source models and surpasses Qwen2.5-VL-72B by 15.1% and DeepSeek-VL2 by 30%. Relative to base GPT-4o-mini, LAD improves by 34.9% (en) and 8.9% (zh). We observe larger gaps between reasoning and general models on OSQ than on MCQ, consistent with the higher difficulty of open-style implication generation and potential MCQ-format bias [41].

Model	Multiple-Choice Question		Open-Style Question	
	en	zh	en	zh
GPT-4o-mini				
w/o CoT	44%	42%	2.98	3.36
Standard CoT	50% ↑	42%	3.10 ↑	3.28 ↓
LAD-CoT	**68%** ↑	**44%** ↑	**3.84** ↑	**3.58** ↑

Table 2. Results of different CoT methods. Our LAD-CoT method achieves the best improvement.

4.4. Ablation Study

4.4.1. Stage I (Perception) and Stage III (Reasoning)

LAD-CoT (Stage I+III) substantially improves GPT-4o-mini (MCQ en: 44%→68%; OSQ en: 2.98→3.84; OSQ zh: 3.36→3.58), outperforming standard CoT (Table 2). Figure 3 provides a case study; prompts are in Appendix F.

4.4.2. Stage II (Search)

LAD-Search (Stage II) consistently improves both MCQ and OSQ (e.g., on GPT-4o-mini: MCQ en 68%→74%, zh 44%→52%; OSQ en 3.84→4.02, zh 3.58→3.66). In contrast, alternative search variants can improve MCQ but often degrade OSQ (Table 3), highlighting the importance of implication-aware retrieval and summarization.

Model	Multiple-Choice Question		Open-Style Question	
	en	zh	en	zh
Grok-3				
w/o search	66%	64%	3.24	2.96
Grok-Search	**72%** ↑	64%	3.25 ↑	2.92 ↓
GPT-4o				
w/o search	74%	58%	2.94	3.76
Perplexity (pro)	**80%** ↑	**66%** ↑	2.88 ↓	3.28 ↓
GPT-4o-mini				
w/o search	68%	44%	3.84	3.58
GPT-Search	72% ↑	48% ↑	3.62 ↓	3.34 ↓
LAD-Search	**74%** ↑	**52%** ↑	**4.02** ↑	**3.66** ↑

Table 3. Results of different search methods. Our LAD-Search method achieves the best improvement.

Model	Multiple-Choice Question		Open-Style Question	
	en	zh	en	zh
Qwen2.5-VL-7B				
w/o LAD	46%	40%	2.34	2.58
w/ LAD	**64%** ↑	**46%** ↑	**3.64** ↑	**3.36** ↑
Qwen2.5-VL-72B				
w/o LAD	72%	56%	1.56	3.12
w/ LAD	**76%** ↑	**62%** ↑	**3.62** ↑	**3.68** ↑
GPT-4o				
w/o LAD	74%	58%	2.94	3.76
w/ LAD	**80%** ↑	**66%** ↑	**4.14** ↑	**4.26** ↑
Gemini-3.0-pro				
w/o LAD	76%	76%	3.82	3.96
w/ LAD	**82%** ↑	**78%** ↑	**4.30** ↑	**4.46** ↑

Table 4. Results of different base models. Our LAD demonstrates the generalizability on different base models.

4.4.3. Different Base Models

To demonstrate the generalizability of our LAD framework beyond the GPT-4o-mini model, we conduct new experiments applying LAD to other base models, including the open-source Qwen2.5-VL series, the closed-source model GPT-4o and the latest closed-source model Gemini-3.0-pro. As the Table 4 shows, applying LAD framework significantly boosts the performance of all models across both MCQ and OSQ tasks, confirming that our framework is not model-specific and provides a robust and generalizable approach to enhancing image metaphor understanding.

4.4.4. Generalization Experiment

Model	MMMU_val	SeedBench	MMStar
GPT-4o-mini	59.4	72.8	54.8
GPT-4o	**70.7**	76.7	**65.1**
LAD (GPT-4o-mini)	67.9 % ↑	**77.2%** ↑	60.3 ↑

Table 5. Results of different models on general VQA benchmarks.

Experiments On Full Benchmarks. We further evaluate LAD on II-Bench and CII-Bench at full scale, and observe consistent gains over strong baselines; detailed setup, Table, and analysis are provided in Appendix E.

Experiments On General VQA Benchmarks. To further demonstrate that LAD is a generalizable reasoning framework, we evaluated it on three general multi-modal benchmarks: MMMU (Expert AGI and Visual Reasoning), SeedBench (General Understanding), and MMStar (General Understanding). We applied the LAD framework to GPT-4o-mini without modifying the core architecture. The results are presented in Table 5.

We find that the LAD framework provides huge improvements (e.g., +8.5% on MMMU). With LAD, the lightweight GPT-4o-mini surpasses the much larger GPT-4o on SeedBench (77.2 vs 76.7) and significantly closes the gap on others. These results confirm that our "Perception-Search-Reasoning" workflow addresses a fundamental cognitive gap in VLM reasoning, effectively handling tasks requiring visual commonsense and complex reasoning beyond just metaphor understanding.

5. Discussion

5.1. Human Cognitive Theory of Let Androids Dream

Our claim is that the LAD framework is analogous to human cognitive strategies, not a direct neuroscientific replica. Our goal is to create a system that reasons in a way that is transparent and aligns with how humans might tackle the same problem, not to simulate the human brain perfectly.

Our framework is directly inspired by established human cognitive science theories: (1) **Dual-Process Theory** [6]: The Perception stage mirrors the interplay between System 1 (the fast, intuitive, holistic impression of the image) and System 2 (the slower, analytical identification of key elements), and (2) **Active Information-Seeking Theory** [10, 30]: The Search stage is analogous to the human tendency to actively seek external information to resolve ambiguity. Humans do not reason in a vacuum; when we encounter an unfamiliar meme or cultural reference, a common cognitive act is to "Google it" to supplement our internal knowledge. Our WebSearch module directly simulates this deliberate information-foraging behavior.

5.2. How to Let Androids Dream? Perception and Reasoning

The question "How to Let Androids Dream?" metaphorically addresses the foundational challenge of enabling AI systems to interpret the nuanced implications embedded in images. Our framework tackles this by first emulating human-like perception (Stage I), converting raw visual input into rich, multi-level textual representations, including comprehensive descriptions and salient keywords. These

keywords are designed to capture not only objects and scenes but also potential emotional tones, relevant domains (e.g., cultural, social, political), and discernible rhetorical devices. Subsequently, LAD's Stage III employs an explicit, structured CoT process. This structured reasoning guides the model to systematically connect the perceived visual elements with retrieved contextual knowledge, thereby constructing a coherent understanding of implications. This method is vital because, as our experiments (Section 4) and recent work on social reasoning [12] show, comprehending implications extends beyond basic VQA tasks and classic logical reasoning; it inherently involves sophisticated social reasoning and the interpretation of contextual cues often missed by MLLMs.

5.3. How to Dream of Electric Sheep? Search

Building upon the capacity to analyze, "How to Dream of Electric Sheep?" delves into how AI can generate accurate and specific image metaphors—the metaphorical 'electric sheep'. LAD's Stage II (Search) is the key to achieving this goal. This stage acknowledges that the meaning of visual elements, particularly in metaphorical contexts, often relies on external information, such as cultural norms, historical events, or contemporary affairs, which may not be adequately represented in MLLMs' static pre-trained knowledge. LAD's adaptive search mechanism, which includes formulating targeted queries from keywords and dynamically selecting between internal ModelSearch and external WebSearch via Self-Judge, systematically enriches the initial perception with relevant cross-domain knowledge. This iterative retrieval and integration of contextual information, especially for popular metaphors or ambiguous visual cues, significantly broadens the model's interpretive horizon. By providing this essential external context, the Search stage empowers LAD to move beyond superficial interpretations and accurately capture the intended, often subtle, implications of an image, as demonstrated by its robust performance on Open-Style Question (OSQ).

6. Conclusion

Understanding image metaphors remains challenging for MLLMs, mainly due to contextual missing. Our work introduces Let Androids Dream (LAD), a novel three-stage framework: Perception, Search, and Reasoning. Inspired by human cognitive processes, this framework is designed to achieve contextual alignment by explicitly integrating visual interpretation with external knowledge retrieval. We conduct comprehensive experiments to demonstrate its effectiveness. Utilizing the lightweight GPT-4o-mini, LAD achieves top results on implication benchmarks, performing comparable or even surpassing Gemini-3.0-pro and other top closed-source models, particularly on challenging OSQ. In summary, LAD bridges the gap between superficial perception and reasoning in multimodal AI systems, offering a promising direction for contextual-alignment vision-language reasoning.

References

[1] Anthropic. Model card addendum: Claude 3.5 haiku and upgraded claude 3.5 sonnet, 2024. 6

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 6

[3] Zehui Chen, Kuikun Liu, Qiuchen Wang, Jiangning Liu, Wenwei Zhang, Kai Chen, and Feng Zhao. Mindsearch: Mimicking human minds elicits deep ai searcher. *arXiv preprint arXiv:2407.20183*, 2024. 4

[4] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 1, 3

[5] Poorav Desai, Tanmoy Chakraborty, and Md Shad Akhtar. Nice perfume. how long did you marinate in it? multimodal sarcasm explanation. In *AAAI*, 2022. 3

[6] Jonathan Evans. In two minds: dual-process accounts of reasoning. *Trends in Cognitive Sciences*, 2003. 8

[7] Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest. In *ACL*, 2023. 3

[8] Zachary Horvitz, Jingru Chen, Rahul Aditya, Harshvardhan Srivastava, Robert West, Zhou Yu, and Kathleen McKeown. Getting serious about humor: Crafting humor datasets with unfunny large language models. In *ACL*, 2024. 3

[9] Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *arXiv preprint arXiv:2406.09403*, 2024. 3

[10] Robert Ikoja-Odongo and Janneke Mostert. Information seeking behaviour: A conceptual framework. *South African Journal of Libraries and Information Science*, 2006. 8

[11] Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*, 2024. 2

[12] Hyunwoo Kim, Melanie Sclar, Tan Zhi-Xuan, Lance Ying, Sydney Levine, Yang Liu, Joshua B. Tenenbaum, and Yejin Choi. Hypothesis-driven theory-of-mind reasoning for large language models. *arXiv preprint arXiv:2502.11881*, 2025. 9

[13] George Lakoff and Mark Johnson. *Metaphors we live by*. University of Chicago press, 2008. 1

[14] Wangyue Li, Liangzhi Li, Tong Xiang, Xiao Liu, Wei Deng, and Noa Garcia. Can multiple-choice questions really be useful in detecting the abilities of llms? *arXiv preprint arXiv:2403.17752*, 2024. 5

[15] Zejun Li, Ruipu Luo, Jiwen Zhang, Minghui Qiu, and Zhongyu Wei. Vocot: Unleashing visually grounded multi-step reasoning in large multi-modal models. *arXiv preprint arXiv:2405.16919*, 2024. 3

[16] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023. 2

[17] Ziqiang Liu, Feiteng Fang, Xi Feng, Xinrun Du, Chenhao Zhang, et al. Ii-bench: An image implication understanding benchmark for multimodal large language models. In *NeurIPS*, 2024. 2, 3, 5, 11

[18] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *ICLR*, 2024. 2

[19] Sachit Menon, Richard Zemel, and Carl Vondrick. Whiteboard-of-thought: Thinking step-by-step across modalities. *arXiv*, 2024. 3

[20] Aidar Myrzakhan, Sondos Mahmoud Bsharat, and Zhiqiang Shen. Open-llm-leaderboard: From multi-choice to open-style questions for llms evaluation, benchmark, and arena. *arXiv preprint arXiv:2406.07545*, 2024. 5

[21] OpenAI. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 6, 11

[22] OpenAI. Learning to reason with llms, 2024. 1, 3

[23] ByteDance Seed. Doubao-1.5-thinking-vision-pro, 2025. 6

[24] Gemini Team. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 6

[25] Gemini Team. Gemini 3 pro: Best for complex tasks and bringing creative concepts to life, 2025. 6

[26] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1.5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025. 1

[27] Qwen Team. Qvq: To see the world with wisdom, 2024. 1, 6

[28] Qwen Team. Qwq: Reflect deeply on the boundaries of the unknown, 2024. 1, 3

[29] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. In *NeurIPS*, 2024. 2

[30] Thomas D. Wilson. Activity theory and information seeking. *Annu. Rev. Inf. Sci. Technol.*, 2009. 8

[31] Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. *arXiv preprint arXiv:2312.14135*, 2023. 3

[32] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024. 6

[33] xAI. Grok 3 beta — the age of reasoning agents, 2025. 6

[34] Guowei Xu, Peng Jin, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024. 3

[35] Yanzhi Xu, Yueying Hua, Shichen Li, and Zhongqing Wang. Exploring chain-of-thought for multi-modal metaphor detection. In *ACL*, 2024. 2, 3

[36] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 3

[37] Yixin Yang, Zheng Li, Qingxiu Dong, Heming Xia, and Zhifang Sui. Can large multimodal models uncover deep semantics behind images? In *ACL*, 2024. 3

[38] Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, and Dacheng Tao. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. *arXiv preprint arXiv:2412.18319*, 2024. 3

[39] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, 2024. 2

[40] Chenhao Zhang, Xi Feng, Yuelin Bai, Xinrun Du, et al. Can mllms understand the deep implication behind chinese images? *arXiv preprint arXiv:2410.13854*, 2024. 2, 3, 5, 11

[41] Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *ICLR*, 2024. 5, 7

[42] Zhipu.ai. Glm-4v, 2024. 6

[43] Shanshan Zhong, Zhongzhan Huang, Shanghua Gao, Wushao Wen, Liang Lin, Marinka Zitnik, and Pan Zhou. Let's think outside the box: Exploring leap-of-thought in large language models with creative humor generation. *arXiv preprint arXiv:2312.02439*, 2024. 2, 3

[44] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 6

A. Experiment Setup

We use the lightweight GPT-4o-mini-0718 [21] with LAD framework in experiments. We set the model temperature as 0.5 and top_p as 0.9 in MCQ experiments, and temperature as 0.7 and top_p as 0.9 in OSQ experiments. Additionally, we set the evaluation model GPT-4o temperature as 0 and evaluate more than three times to get the average score in OSQ experiments. All experiments are conducted on NVIDIA A800 GPUs.

B. Human-Model Consistency Study

To validate our automated OSQ evaluation based on the GPT-4o model, we conduct a human-model consistency study. We construct a dedicated dataset by randomly selecting 25 images with questions each from our English and Chinese OSQ. We recruit 16 PhD students and researchers, all proficient in both English and Chinese and experienced with metaphorical imagery, to independently score the model responses. Their evaluations are based on ground truth answers and the detailed scoring standard. We calculate human inter-annotator agreement by averaging the scores for each response after discarding the highest and lowest individual scores. This process yields the consistency of 94.8% for Chinese and 96.5% for English. The average human-model scoring consistency reached 95.7%, affirming the method's validity for assessing image metaphor comprehension.

C. Statistics

We manually construct the high-level benchmark by selecting 100 high-quality, diverse and representative images from II-Bench [17] and CII-Bench [40]. The general statistic is in Table 6.

D. Further Analysis on Method and Experiments

D.1. Analysis of Let Androids Dream Success

Our analysis points to two primary failure modes for baseline models, which Let Androids Dream (LAD) is designed to mitigate. These are illustrated in Figure 1 and the case study in Figure 3:

1. Superficial Reasoning: This occurs when a model only processes the literal, surface-level elements and misses the metaphorical meaning entirely. In Figure 3 the "End2End" baseline exemplifies this, failing to grasp the subversion of the fairy tale trope.

2. Over-Inference: This happens when a model incorrectly applies a known symbol or narrative without considering the full context. The "CoT" baseline in Figure 3 demonstrates this by connecting the heart symbol to a traditional fairy tale transformation without recognizing the comic's twist.

LAD succeeds by first creating a more structured understanding in the Perception stage and then grounding its reasoning with targeted external knowledge from the Search stage, which helps avoid both superficiality and incorrect inferences.

D.2. Analysis of Model Scaling and Image Metaphor Types

Our experiments have some insightful findings:

1. Model Scaling: By testing on QwenVL-2.5-7B and QwenVL-2.5-72B, we can analyze the effect of model scale. Our findings align with expectations: larger parameter models generally achieve better baseline performance, and both scales benefit from the LAD framework. This confirms that our method is effective across different model sizes.

2. Image Metaphor Types: Our benchmark was already designed to be diverse across various domains (e.g., life, society, art, psychology, Chinese traditional culture) and image types (e.g., comic, poster, meme). We find that models perform worse in domains containing abstract and complex information, like Art and Psychology. And models only observe the surface-level information and lack sufficient understanding of Chinese culture. In a further analysis using the annotations from the original II-Bench and CII-Bench, we observed that providing explicit labels for Emotion, Domain, and Rhetoric significantly enhances model accuracy, with Emotion labels providing the largest boost. This confirms that our framework's focus on identifying these elements in the Perception stage is well-founded.

E. Generalization on Full Benchmarks

We conduct the large-scale experiments with representative and top-performing models, including GPT-4o, Claude-3.5-Sonnet and GLM-4.1V-8B, on II-Bench (1,399 examples) and CII-Bench (800 examples) for both MCQ and OSQ tasks.

As shown in Table 7, LAD's gains remain consistent on larger-scale data. Applying LAD to GPT-4o-mini significantly narrows or surpasses larger baseline models. Relative to base GPT-4o-mini, LAD improves II-Bench from 63.5% to 81.2% on MCQ and 2.93 to 4.22 on OSQ; on CII-Bench it improves 35.6% to 53.8% (MCQ) and 3.29 to 4.31 (OSQ). This is consistent with the trend observed on our high-level benchmark and supports the robustness of LAD across evaluation scales.

Statistics of English Images	
Society	21 (42%)
Life	16 (32%)
Art	6 (2%)
Psychology	4 (8%)
Others	3 (6%)
Multi-panel Comic	16 (32%)
Single-panel Comic	9 (18%)
Illustration	5 (10%)
Meme	5 (10%)
Poster	5 (10%)
Painting	5 (10%)
Logo	5 (10%)

Statistics of Chinese Images	
Life	13 (26%)
Art	13 (26%)
Society	12 (24%)
Chinese Traditional Culture	6 (12%)
Environment	5 (10%)
Politics	1 (2%)
Illustration	15 (30%)
Single-panel Comic	10 (20%)
Poster	8 (16%)
Meme	8 (16%)
Painting	6 (12%)
Multi-panel Comic	3 (6%)

Table 6. General statistics of the high-level benchmark.

Model	Multiple-Choice Question		Open-Style Question	
	II-Bench (1399)	CII-Bench (800)	II-Bench (1399)	CII-Bench (800)
GLM-4.1V-8B	70.0%	46.3%	2.83	3.06
GPT-4o-mini	63.5%	35.6%	2.93	3.29
InternVL3-78B	78.2%	**64.0%**	3.68	4.06
GPT-4o	72.6%	54.1%	3.86	4.06
Claude-3.5-Sonnet	80.9%	54.1%	3.51	3.84
LAD (GPT-4o-mini)	**81.2%** ↑	53.8% ↑	**4.22** ↑	**4.31** ↑

Table 7. Results of different models on full benchmarks. The best-performing model in each category is **in-bold**. Performance differences relative to base models are shown as colodarkred subscripts: ↑ for improvements, ↓ for declines.

F. Prompts

In experiments, the prompts of different settings are as follows:

F.1. Evaluation

 # Role
You are an impartial judge who is familiar with Internet culture and memes, and is good at digging out and analyzing the deep meaning of Internet memes.

Attention
You are responsible for evaluating the quality of the answer provided by the model for Internet culture and memes. Your evaluation should refer to the human answer and image, and score based on the Evaluation Standard.

Evaluation Standard
- [1 point]:
Fails to capture key elements within the image (such as text, and important entities). Does not identify emotions, domain, or rhetorical devices. Only provides a superficial description of surface-level information, lacking depth and creativity, with a significant gap from the standard answer.
- [2 points]:
Captures some key elements within the image, but the identification of emotions, domain, and rhetorical devices is vague. The description of surface-level information is relatively complete, but there is a clear deficiency in exploring deeper meanings, showing a noticeable gap from the standard answer.
- [3 points]:
Effectively captures key elements within the image and initially identifies emotions, domain, and rhetorical devices. The description of surface-level information is relatively accurate, and there is some relevant expression of deep meanings. However, there is still room for improvement in depth and creativity, and it is generally close to the standard answer.
- [4 points]:
Accurately captures key elements within the image and clearly identifies emotions, domain, and rhetorical devices. The description of surface-level information is detailed and precise, with a relatively deep exploration of deep meanings, demonstrating a certain level of creativity and depth. It is largely consistent with the standard answer but may have minor deficiencies in some details or depth.
- [5 points]:
Accurately and precisely captures key elements within the image and profoundly identifies emotions, domain, and rhetorical devices. The description of surface-level information is comprehensive and precise, with unique insights into deep meanings, skillfully integrating image elements with metaphorical implications. It demonstrates exceptional creativity and depth, is highly consistent with the standard answer, and shows a profound grasp of metaphor creation and cultural understanding.

Standrad Answer:
Human answer: {}

Constraints
- Avoid any position biases and be as objective as possible
- Do not allow the length of the descriptions to influence your evaluation
- Output your final verdict by strictly following this format: "[ratings]"

Solve:
Model answer: {}

Figure 5. The evaluation prompt of Open-Style Question (OSQ).

F.2. End2End

Prompt in Chinese	Prompt in English
请根据提供的图片尝试回答以下单选题。直接回答正确选项，不要包含额外的解释。 请使用以下格式：“答案：$LETTER”，其中$LETTER是你认为正确答案的字母。 单选题：{} 答案：	Please try to answer the following multiple-choice questions based on the provided image. Answer the correct option directly without additional explanation. Please use the following format: "Answer: $LETTER", where $LETTER is the letter of the correct answer you think. Multiple-choice questions: {} Answer:

Figure 6. The end2end prompt of Multiple-Choice Question (MCQ).

Prompt in Chinese	Prompt in English
请结合以上图片，尽可能分析理解图片的深层含义。无需描述图片，仅回答图片隐喻。请保证回答的准确性并尽量简洁。	Please try to understand the deep meaning of the image. No need to describe images and text, only answer metaphors. Ensure the accuracy of the answer and try to be concise as much as possible.

Figure 7. The end2end prompt of Open-Style Question (OSQ).

F.3. CoT

Prompt in Chinese	Prompt in English
请根据提供的图片尝试回答以下单选题。逐步思考回答正确选项，不要包含额外的解释。 请使用以下格式：“答案：$LETTER”，其中$LETTER是你认为正确答案的字母。 单选题：{} 答案：	Please try to answer the following multiple-choice questions based on the provided image. Let's think step by step to answer the correct option directly without additional explanation. Please use the following format: "Answer: $LETTER", where $LETTER is the letter of the correct answer you think. Multiple-choice questions: {} Answer:

Figure 8. The CoT prompt of Multiple-Choice Question (MCQ).

Prompt in Chinese	Prompt in English
请结合以上图片，逐步思考尽可能分析理解图片的深层含义。无需描述图片，仅回答图片隐喻。请保证回答的准确性并尽量简洁。	Please try to think step by step to understand the deep meaning of the image. No need to describe images and text, only answer metaphors. Ensure the accuracy of the answer and try to be concise as much as possible.

Figure 9. The CoT prompt of Open-Style Question (OSQ).