LET ANDROIDS DREAM: FROM VISUAL ELEMENTS TO EMERGENT IMPLICATIONS VIA CONTEXTUAL ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Image metaphor understanding differs from general VQA: the implied meaning is rarely explicit in pixels and instead emerges from the *compositional interaction* of visual elements together with social intent and culture-specific background. We identify *missing context* – the gap between surface perception and the external knowledge required to infer emergent implications – as a failure mode of current multimodal large language models (MLLMs). We formalize *Contextual Alignment* as inference over a latent context: selecting background knowledge that is relevant to the image's composition and useful for inferring its implication. We then propose ***Let Androids Dream (LAD)***, a training-free framework that operationalizes contextual alignment through test-time computation: (1) **Perception** extracts compositional visual structure, (2) **Search** acquires missing context via perception-guided retrieval, and (3) **Reasoning** aligns compositions with context to derive emergent implications. On S100, a benchmark of 100 high-level implication questions in English and Chinese, LAD raises the lightweight GPT-4o-mini from 44% to 74% (English) and from 42% to 52% (Chinese) Multiple-Choice Question (MCQ) accuracy, on par with the closed-book GPT-5.4 in English, and from 2.98 to 4.02 and from 3.36 to 3.66 on Open-Style Questions (OSQ); with GPT-4o, LAD obtains the highest OSQ scores among directly comparable systems (4.14 and 4.26). LAD improves all four backbones we test, including the open-source Qwen2.5-VL-7B (+18 MCQ and +1.30 OSQ points in English), and its gains hold on the full II-Bench and CII-Bench. Adding other search tools raises MCQ accuracy in five of six cases but changes OSQ scores by -0.48 to $+0.01$, whereas LAD's search stage improves both, suggesting that how context is acquired matters beyond access to search. Our OSQ judge is within one point of trimmed human means for 95.7% of answers, and LAD with web search also improves MMMU, SeedBench, and MMStar.

1 INTRODUCTION

> *Do androids dream of electric 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 electric sheep.*
>
> – Philip K. Dick (1968)

Metaphors pervade daily communication – from "time is money" to visual satire in editorial cartoons, serving as fundamental cognitive tools for conceptualizing abstract ideas (Lakoff & Johnson, 2008). Enabling AI to understand such metaphors, particularly in the visual modality, requires going far beyond literal object recognition, and it matters in practice for moderating harmful memes, analysing political communication, and interpreting images across cultures. Critically, the meaning of an image metaphor does not reside in individual visual entities themselves, but emerges from their *compositional interaction* under contextual and cultural priors – making this task qualitatively different from standard visual question answering.

We identify a systematic failure mode in current multimodal reasoning: *missing context*. When interpreting image metaphors, models must connect compositional visual structures to cultural, social,

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

or temporal meanings – emergent implications that frequently lie outside both the image content and the model's readily accessible parametric knowledge (Figure 1). Existing approaches either attempt explicit metaphor mapping (Zhong et al., 2024), which is brittle under many-to-many relations, or chain-of-thought prompting (Xu et al., 2024b), which reasons explicitly but only over knowledge already in the model's parameters. Scaling model parameters or task-specific training, including end-to-end reinforcement learning (Zhang et al., 2026), can partially internalize such knowledge through train-time investment (Liu et al., 2024; Zhang et al., 2024a), but these approaches are costly, opaque, and model-bound. A fundamental question remains: can we achieve effective contextual alignment at inference time, transparently and without any training?

We answer this affirmatively. We first formalize *Contextual Alignment* with a latent-context decomposition: the implication of an image depends on background knowledge that is not visible in it, and contextual alignment selects the context that is relevant to the image's composition and useful for inferring its implication (Section 3.1). We then propose ***Let Androids Dream (LAD)***, a training-free framework that operationalizes contextual alignment through three complementary stages: **Perception** to extract compositional visual structure, **Search** to retrieve missing contextual knowledge guided by perceived compositions, and **Reasoning** to align compositions with acquired context for implication inference. Unlike chain-of-thought prompting (which cannot access missing information), vanilla retrieval augmentation (which retrieves without perception guidance), or end-to-end training (which requires costly optimization), LAD operates at inference time on a frozen MLLM and is backbone-agnostic by design. Its technical novelty lies in three components that generic retrieval pipelines lack (Table 1): queries are planned from the *composition* of the image – keyword combinations naming its emotion, domain, and rhetoric – rather than from entity names; retrieved answers are ranked and refined by their relevance to the image's *implication* rather than to a question; and, unlike end-to-end and RL-trained methods, the acquired context is an explicit variable that can be inspected, removed, or replaced.

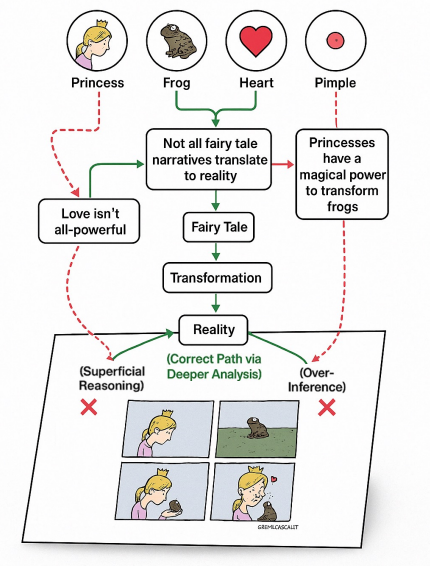

Figure 1: An image is worth a thousand words: For image metaphor understanding, different element combinations lead to different reasoning paths, while the correct path requires integrating all elements.

On S100, a set of 100 high-level implication questions from II-Bench and CII-Bench, LAD raises the multiple-choice (MCQ) accuracy of the lightweight GPT-4o-mini from 44% to 74% (English) and from 42% to 52% (Chinese), on par with the closed-book GPT-5.4 in English, and its open-style (OSQ) score from 2.98 to 4.02 and from 3.36 to 3.66; with GPT-4o, LAD obtains the highest OSQ scores among directly comparable systems. Because LAD uses search, we also compare it with search-augmented systems: other search tools raise MCQ accuracy in five of six cases but change OSQ by -0.48 to $+0.01$, whereas LAD improves both, and LAD improves all four backbones we test. For context-dependent reasoning, *how* inference-time computation acquires context matters, not only whether the model can search.

Our key contributions are as follows:

- **Problem Reformulation:** We reformulate image metaphor understanding as a context-dependent compositional reasoning problem and formalize it as inference over a latent context, which makes missing context explicit and maps perception, context acquisition, and reasoning to separable stages.
- **Failure Mode Identification:** We identify *missing context* as a failure mode of frozen MLLMs; ablations show that explicit chain-of-thought alone helps little, whereas structured perception with explicit reasoning (Stage I+III) gives most of the gain on three of four metrics and perception-guided context acquisition (Stage II) adds the rest.

- **Test-Time Contextual Alignment:** We propose LAD, a training-free, backbone-agnostic framework that achieves contextual alignment through test-time computation, and compare it with both closed-book and search-augmented systems. We also design an Open-Style Question (OSQ) evaluation protocol whose automatic scores are within one point of human ratings for 95.7% of answers.
- **Generalizability:** LAD improves four backbones (Qwen2.5-VL-7B/72B, GPT-4o-mini, GPT-4o) on both languages and formats, the GPT-4o-mini gain holds on the 2,199 full-benchmark items, and LAD with web search improves MMMU, SeedBench, and MMStar.

2 Related Work

2.1 Image Metaphor

Early work on image metaphor targeted specific phenomena such as humor (Hessel et al., 2023; Horvitz et al., 2024) and sarcasm (Desai et al., 2022). Explaining New Yorker cartoons (Hessel et al., 2023) and Chinese pun rebus art (Zhang et al., 2024b) requires knowledge that is not in the picture, but these studies do not supply it. Benchmarks for MLLMs followed: DeepEval (Yang et al., 2024b) proposes a taxonomy, and II-Bench (Liu et al., 2024) and CII-Bench (Zhang et al., 2024a) evaluate English and Chinese images. Methods fall into three groups: explicit metaphor mapping (CLOT (Zhong et al., 2024)), which scales poorly to many-to-many mappings and new cultural references; chain-of-thought prompting (C4MMD (Xu et al., 2024b)), which reasons only over parametric knowledge; and end-to-end reinforcement learning (MetaphorStar (Zhang et al., 2026)), which stores implication reasoning in the weights. LAD treats the task as *contextual alignment* at test time and keeps the acquired context as an explicit variable.

2.2 Vision-language Reasoning and Retrieval

Reasoning models such as o1 (OpenAI, 2024b), DeepSeek-R1 (DeepSeek-AI, 2025), and QwQ (Team, 2024b; Yang et al., 2024a) motivated vision-language reasoning, which converts images into textual grounding (Socratic Models (Zeng et al., 2022), LLaVA-CoT (Xu et al., 2024a), Mulberry (Yao et al., 2024)), elicits cues through multi-turn interaction (VoCoT (Li et al., 2024b), V* (Wu & Xie, 2023)), or reasons with visual intermediates (Visual Sketchpad (Hu et al., 2024), Whiteboard-of-Thought (Menon et al., 2024), CoVT (Qin et al., 2025)); these methods need task-specific training or stay within parametric knowledge. Retrieval supplies missing knowledge through adaptive retrieval (Self-RAG (Asai et al., 2024)), search agents (Chen et al., 2024), and trained multimodal search agents (MMSearch-R1 (Wu et al., 2025), WebWatcher (Geng et al., 2025); MMSearch-Plus (Tao et al., 2025)) that look up facts about depicted entities. The meaning of a metaphor, however, emerges from a *composition* of elements, so LAD plans queries from perceived compositions and selects context by implication (Table 1).

Table 1: Positioning of LAD. *Live retrieval*: web search at test time (Self-RAG retrieves from a static index); *composition-guided queries*: retrieval queries derived from the combination of visual elements and their emotion, domain, and rhetoric; *implication-aware selection*: retrieved content ranked and refined by relevance to the image's implied meaning; –: not applicable (no retrieval stage). LAD's WebSearch builds on MindSearch.

Method	Training-free	Multi-modal	Live retrieval	Adaptive retrieval	Composition-guided queries	Implication-aware selection
C4MMD (CoT distillation) (Xu et al., 2024b)	✗	✓	✗	–	–	–
Socratic Models (Zeng et al., 2022)	✓	✓	✗	–	–	–
Self-RAG (Asai et al., 2024)	✗	✗	✗	✓	✗	✗
MindSearch (Chen et al., 2024)	✓	✗	✓	✗	✗	✗
MMSearch-R1 (Wu et al., 2025)	✗	✓	✓	✓	✗	✗
MetaphorStar (Zhang et al., 2026)	✗	✓	✗	–	–	–
LAD (ours)	✓	✓	✓	✓	✓	✓

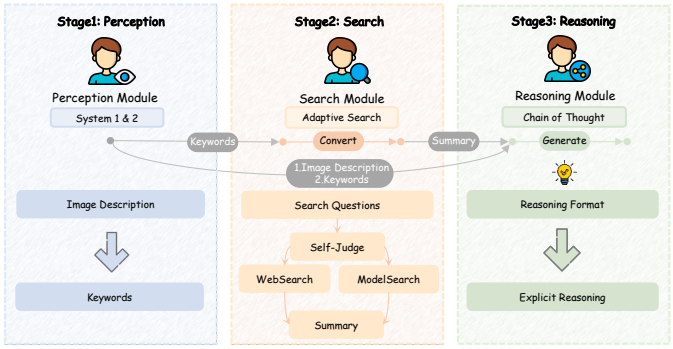

Figure 2: The general framework of Let Androids Dream (LAD), which includes three stages: (1) Perception: converting raw visual information into rich and multi-level textual representations (z), (2) Search: planning, routing, and condensing knowledge retrieval into a context summary ($\hat{c}$), and (3) Reasoning: generating context-aligned image metaphor interpretations via explicit reasoning.

3 METHOD

3.1 PROBLEM FORMULATION: CONTEXTUAL ALIGNMENT

Let x be an image and y its implication. The implication depends on background knowledge c – cultural conventions, social roles, events, idioms, puns – that is generally not visible in x:

$$p(y \mid x) = \textstyle\sum_c p(y \mid x, c)\, p(c \mid x). \tag{1}$$

An end-to-end MLLM must both read x and recall a suitable c from its parameters. We speak of *missing context* when the model's implicit estimate of $p(c \mid x)$ misses the context $c^\star$ that the implication requires, even though the model could identify y given $c^\star$. Because the relevant context depends on how elements are *composed*, we estimate it from a perceptual abstraction $z = \phi(x)$ and define *contextual alignment* as selecting a context $\hat{c}$ that is relevant to z and useful for inferring y:

$$\hat{c} = S\big(R(Q(z))\big), \qquad \hat{y} = \arg\max_y \, p_\theta(y \mid x, z, \hat{c}), \tag{2}$$

where Q plans questions from z, R answers them from parametric or external sources, and S selects and condenses the answers. Eq. 1 is a conceptual decomposition rather than an estimable model: LAD directly constructs $\hat{c}$ and does not explicitly estimate either probability. Since z is computed from x, its role is computational, and Eq. 2 exposes stages that our ablations remove or replace.

We propose Let Androids Dream (LAD), a training-free framework that implements Eq. 2 with prompted calls to one frozen MLLM in three stages (Figure 2; Algorithm 1). The same prompt templates (with Chinese versions for Chinese images) are reused for every backbone and benchmark; the English prompts are reproduced verbatim in Appendix L.

3.2 STAGE I: PERCEPTION

This stage addresses *perception failure* – the tendency of end-to-end models to process visual elements in isolation without explicitly capturing their compositional structure. The *Perception* stage transforms raw visual inputs into structured, hierarchical textual representations (z). First, the MLLM produces a one-paragraph description capturing coarse-grained visual information, including discernible text within the image, prominent colors, overall layout, and salient objects or entities. Following this, the MLLM condenses the description into seven keywords that encapsulate aspects relevant to metaphor understanding: the perceived emotion, the domain or context (e.g., political, social, cultural), and rhetorical devices, explicitly checking for puns and homophones. The keywords also re-emphasize key textual elements or entities and serve as the cues that guide knowledge retrieval in Stage II.

3.3 STAGE II: SEARCH

This stage addresses *missing context* – the external knowledge required for implication inference is absent from both the visual input and the model's readily accessible parametric knowledge. It performs perception-guided context acquisition ($\hat{c} = S(R(Q(z)))$) in three phases: Plan, Search, and Summary.

1. Plan (Q): Using the Stage-I keywords, the MLLM, guided by a prompt designed for image metaphor tasks, generates five search questions aimed at the latent meanings, cultural references, or background information of keyword combinations.

2. Search (R): Each question is routed by a Self-Judge mechanism to one of two knowledge sources.

(a) **Self-Judge**: The MLLM outputs a confidence score from 1 to 5 for answering the question from its own knowledge, together with a decision. The rubric lowers the score for internet or meme culture, highly popular knowledge, real-time content, niche entities, and questions involving more than two entities. Questions scored > 3 go to ModelSearch and questions scored ≤ 3 to WebSearch when the decision agrees; inconsistent or malformed outputs default to ModelSearch. This zero-shot self-assessment follows adaptive retrieval (Asai et al., 2024).

(b) **ModelSearch**: The MLLM answers the question from its own parametric knowledge with a prompt specialized for explaining memes and image metaphors.

(c) **WebSearch**: Following the planner–searcher design of MindSearch (Chen et al., 2024), a planner decomposes the question into sub-questions organized as a dependency graph, and a searcher issues each sub-question to the DuckDuckGo search engine, selects and reads pages, and answers it; the planner then composes the final answer.

Queries are text generated from Stage-I outputs: neither the image nor the benchmark question or options are sent to the search engine, and pages from the benchmark sources are excluded from the retrieved results. This reduces direct leakage, although pages that discuss the same images elsewhere cannot be ruled out.

3. Summary (S): **RankSummary** ranks the five question–answer pairs by their relevance to the metaphor implied by the image description and keeps the top three; **RefineSummary** rewrites them, guided by the ranking reasons, into one concise search summary by removing irrelevant or redundant information and reconciling the remaining pieces. This summary is the context $\hat{c}$ for Stage III.

3.4 STAGE III: REASONING

This stage addresses the *alignment gap* – how to derive emergent implications from the integration of compositional visual structures and acquired context. The MLLM receives the original image together with the description, the keywords, and the search summary. For MCQ, it articulates its reasoning inside `<think>...</think>` tags, connecting visual cues, keywords, and external knowledge, and gives the option inside `<answer>` tags; for OSQ, it states the implication concisely under the same instruction given to all baselines. The explicit trace exposes a readable rationale, which we do not assume faithfully reflects the model's internal computation.

Together, the three stages realize a complete test-time contextual alignment pipeline: from extracting compositional structure, through acquiring missing context, to aligning both for emergent implication inference – all without any parameter updates. Table 1 contrasts this design with prior decompositions.

4 EXPERIMENT

4.1 BASELINES

Models. We compare with open-source MLLMs from 7B to 78B parameters, closed-source models, and reasoning models. Because LAD uses search, we also evaluate *search-augmented* systems: GPT-4o-mini with OpenAI's native search as Stage II, and Grok-3 and GPT-4o with built-in or Perplexity search (Table 3(b)). Closed-book systems use the end-to-end prompts in Figures 6–7; model identifiers and decoding parameters are in Appendix C, and earlier baselines in Table 5.

Table 2: Overall results on S100: Multiple-Choice Question (MCQ, accuracy %) and Open-Style Question (OSQ, 1–5). Models are ordered as open-source (by size), closed-source, and reasoning models. **Bold**/underline: best/second best among the first five models and ours (the last seven were evaluated later; Appendix D); arrows: change over the base model (GPT-4o base: Table 4(a)).

Model	Multiple-Choice Question		Open-Style Question	
	en	zh	en	zh
Qwen2.5-VL-7B (Bai et al., 2025)	46	40	2.34	2.58
Qwen2.5-VL-72B (Bai et al., 2025)	72	56	1.56	3.12
InternVL3-78B (Zhu et al., 2025)	70	74	3.42	3.70
GPT-5.4 (OpenAI, 2026a)	74	72	3.64	3.82
Gemini-3.0-pro (Gemini Team, 2025)	76	**76**	3.82	3.96
Claude-Sonnet-5 (Anthropic, 2026d)	58	54	2.86	3.26
Qwen3.8-max (Qwen Team, 2026)	84	82	3.80	4.00
Gemini-3.1-pro (Google DeepMind, 2026a)	88	80	3.06	3.26
Grok-4.6 (xAI, 2026)	76	84	3.56	3.14
Claude-Opus-4.7 (Anthropic, 2026c)	64	54	3.02	3.14
Gemini-3.8-flash (Google DeepMind, 2026b)	78	84	3.46	4.00
GPT-5.6-Sol (OpenAI, 2026b)	78	58	3.12	2.94
Our Method				
GPT-4o-mini (OpenAI, 2024a)	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 ↑
Improvement	+30	+10	+1.04	+0.30
GPT-4o + LAD (Stage I + II + III)	**80** ↑	66 ↑	**4.14** ↑	**4.26** ↑

Evaluation. Our main benchmark is S100, a manually selected 100-image subset of II-Bench (Liu et al., 2024) and CII-Bench (Zhang et al., 2024a) (Table 8); the full test sets (1,399 and 800 items) are used in Section 4.4.4. To mitigate MCQ bias (Li et al., 2024a; Zheng et al., 2024), we also evaluate OSQ generation with GPT-4o rubric scoring, and report Wilson intervals and significance tests for MCQ (Appendix E).

4.2 Multiple-Choice Question

Implementation Details. Each S100 image (comics, posters, illustrations, memes, and traditional artworks) has one six-option question, mostly "What is the implication in this image?" (Figure 3).

Results and Analysis. Table 2 reports S100 MCQ results. With the lightweight GPT-4o-mini backbone, LAD reaches 74% on English MCQ – on par with the closed-book GPT-5.4 and close to Gemini-3.0-pro (76%) – and improves the base model by +30 (en) and +10 (zh) points; with GPT-4o, LAD ties GPT-4o with Perplexity search at 80%, the best English MCQ accuracy among directly comparable rows (Table 3(b)). The seven models evaluated later span 58–88% (en) and 54–84% (zh), above GPT-4o-mini + LAD in 12 of 14 model–language pairs; relay GPT-4o scores 52/52 vs. 74/58 originally. The English gain over GPT-4o-mini is significant ($p = 0.002$; Appendix E), and both gains hold on the full benchmarks (Section 4.4.4).

4.3 Open-Style Question

Evaluation Metric. Our metric probes both surface-level information and the emotion, domain, and rhetoric behind an image through five perspectives: ***Surface-level Information***, ***Emotional Expression***, ***Domain and Context***, ***Rhetorical Skills***, and ***Deep Implications*** (Figure 5, Appendix F).

MLLM-based Automatic Evaluation. GPT-4o (`gpt-4o-2024-11-20`, temperature 0) assigns each answer one holistic score from 1 to 5 based on the five perspectives, given the image and the human-written reference released with each benchmark (Appendix F). All systems, including LAD, receive the same OSQ instruction to state the implication concisely without describing the image.

Human Consistency Validation. Sixteen bilingual PhD students and researchers scored model answers for 50 random images (25 per language) with the same rubric; averaging human scores after removing the highest and lowest, the judge is within one point of this average for 96.5% of English and 94.8% of Chinese answers (**95.7%** overall). This criterion measures agreement at one-point resolution. Details are in Appendix G.

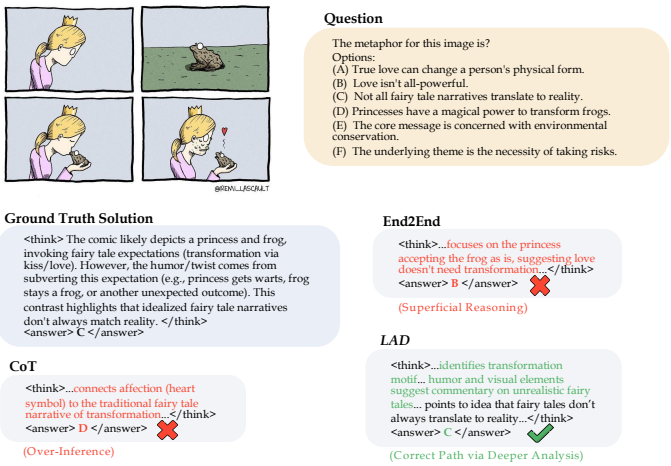

Figure 3: Case study of different methods on Multiple-Choice Question. The *End2End* remains superficial, *CoT* over-infers, while *LAD* gives the correct answer with a context-aligned rationale (outputs abbreviated; analysis in Appendix J).

Results and Analysis. Table 2 reports OSQ results. LAD improves GPT-4o-mini by +1.04 (en) and +0.30 (zh) and GPT-4o by +1.20 and +0.50; GPT-4o with LAD obtains the highest scores in both languages (4.14 and 4.26) and GPT-4o-mini with LAD the second-highest English score (4.02). In contrast, search-augmented baselines show no consistent OSQ improvement (e.g., Perplexity lowers GPT-4o's Chinese score from 3.76 to 3.28). These seven models span 2.86–3.80 in English, below GPT-4o-mini + LAD (4.02) and LAD-CoT (3.84), and reach 4.00 in Chinese vs. 3.66 for GPT-4o-mini + LAD; since the later setup also changes GPT-4o's scores, this comparison is indicative only. The Qwen backbones also gain in English (+1.30, +2.06), suggesting that the gains are not limited to the judge's model family; the native-search comparison, which shares Stages I and III with LAD, likewise suggests that the rubric does not explain LAD's edge over it.

4.4 Ablation Study

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

LAD-CoT (Stage I+III) improves GPT-4o-mini (MCQ en: 44%→68%, zh: 42%→44%; OSQ en: 2.98→3.84, zh: 3.36→3.58), whereas a standard chain-of-thought instruction (Figures 8–9), which reasons only over the model's parametric knowledge, adds +6 English MCQ points and leaves Chinese MCQ unchanged (Table 3(a)). LAD-CoT adds the perceptual abstraction and, for MCQ,

Table 3: Stage-wise ablation on S100: (a) reasoning strategies for Stage I+III; (b) search strategies for Stage II. Arrows mark changes relative to the "w/o" row of each block; LAD rows are shaded.

Setting	MCQ		OSQ	
	en	zh	en	zh
GPT-4o-mini				
w/o CoT	44	42	2.98	3.36
Std CoT	50 ↑	42	3.10 ↑	3.28 ↓
LAD-CoT	**68** ↑	**44** ↑	**3.84** ↑	**3.58** ↑

(a) Reasoning (Stage I+III).

Setting	MCQ		OSQ	
	en	zh	en	zh
Grok-3				
w/o search	66	64	3.24	**2.96**
Grok built-in search	**72** ↑	64	**3.25** ↑	2.92 ↓
GPT-4o				
w/o search	74	58	**2.94**	**3.76**
Perplexity search	**80** ↑	**66** ↑	2.88 ↓	3.28 ↓
GPT-4o-mini, with LAD Stages I and III				
w/o search (LAD-CoT)	68	44	3.84	3.58
OpenAI native search	72 ↑	48 ↑	3.62 ↓	3.34 ↓
LAD-Search	**74** ↑	**52** ↑	**4.02** ↑	**3.66** ↑

(b) Search (Stage II).

Table 4: Generalization results: (a) LAD with all three stages run by each backbone; (b) general VQA benchmarks. LAD uses web search; the other rows are closed-book. LAD rows are shaded.

Setting	MCQ (%)		OSQ	
	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** ↑
GPT-4o-mini				
w/o LAD	44	42	2.98	3.36
w/ LAD	**74** ↑	**52** ↑	**4.02** ↑	**3.66** ↑
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** ↑

(a) Different base models (S100).

Model	MMMU	SeedBench	MMStar
GPT-4o-mini	59.4	72.8	54.8
GPT-4o	**70.7**	76.7	**65.1**
GPT-4o-mini + LAD	67.9	**77.2**	60.3

(b) General VQA benchmarks (accuracy, %).

`<think>` reasoning: naming the domain, emotion, and rhetoric of a composition lets the model recall relevant conventions that end-to-end inference misses (Figure 3).

4.4.2 Stage II (Search)

LAD-Search improves all four metrics over LAD-CoT (MCQ en 68%→74%, zh 44%→52%; OSQ en 3.84→4.02, zh 3.58→3.66). In contrast, the other search tools usually raise MCQ accuracy but do not consistently raise OSQ (Table 3(b); Figure 4): with the backbone and Stages I and III fixed, replacing LAD's Stage II with OpenAI's native search lowers OSQ below the no-search setting (3.84→3.62, 3.58→3.34). LAD's search stage, which plans queries from the composition and ranks answers by implication, outperforms both alternatives on all four metrics.

4.4.3 Different Base Models

To test whether LAD is model-specific, we apply the full pipeline – all three stages, including Self-Judge, ModelSearch, and the WebSearch planner and searcher, run by the respective backbone with unchanged prompts – to the open-source Qwen2.5-VL-7B and 72B and to GPT-4o (Table 4(a)). LAD improves every backbone on every metric. English MCQ gains are larger for the two backbones with low closed-book accuracy (44–46%: +30 and +18) than for the two with high accuracy (72–74%: +4 and +6; Figure 4(c)), while Chinese gains (+6 to +10) are more uniform. OSQ im-

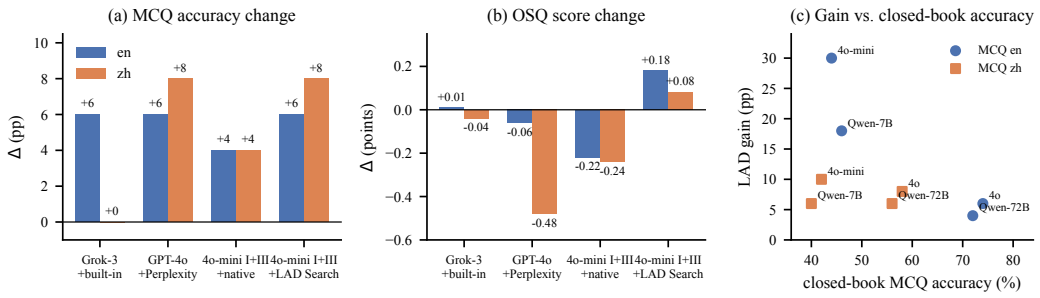

Figure 4: (a, b) Change in MCQ accuracy and OSQ score when search is added, relative to the same system without search (for GPT-4o-mini, relative to LAD-CoT). Other search tools raise MCQ accuracy in five of six cases but change OSQ by -0.48 to $+0.01$; LAD-Search raises both. (c) LAD gain versus closed-book MCQ accuracy of the backbone (Table 4(a)).

proves for all backbones (+0.30 to +2.06); the largest gain starts from Qwen2.5-VL-72B's closed-book English score of 1.56 (despite 72% MCQ accuracy); its cause was not identified.

4.4.4 GENERALIZATION EXPERIMENT

Experiments on Full Benchmarks. We evaluate LAD on II-Bench and CII-Bench at full scale (Table 6). LAD raises GPT-4o-mini from 63.5% to 81.2% and from 35.6% to 53.8% ($p < 10^{-12}$ for both; Table 7) and from 2.93 to 4.22 and 3.29 to 4.31 on OSQ, the highest OSQ scores among the compared systems. On II-Bench, it scores above GPT-4o (72.6%, $p < 10^{-7}$) and close to Claude-4.6-Sonnet (81.0%; difference not significant); on CII-Bench it is close to GPT-4o (54.1%; not significant) but below InternVL3-78B and Claude-4.6-Sonnet. LAD's gains thus hold at scale; the full-benchmark baselines are closed-book, and search-controlled comparisons exist only on S100 (Table 3(b)).

Experiments on General VQA Benchmarks. To examine whether LAD applies beyond metaphor understanding, we evaluate it on MMMU (expert-level reasoning), SeedBench (general understanding), and MMStar (general understanding) with GPT-4o-mini and unchanged prompts (Table 4(b)). LAD improves GPT-4o-mini on all three (+8.5 on MMMU, +4.4 on SeedBench, +5.5 on MMStar) and exceeds GPT-4o on SeedBench (77.2 vs. 76.7). With web search enabled, these are not controlled comparisons with closed-book models (Appendix I).

5 DISCUSSION

Perception, search, and reasoning. Stage I turns images into compositional abstractions of objects, emotion, domain, and rhetoric, and with Stage III accounts for most of the English MCQ gain; Stage II supplies cultural and social background, and its OSQ benefit disappears when it is replaced by generic search, consistent with retrieved context helping more when it is selected for the composition. Since entity facts can rule out MCQ distractors without explaining how entities combine, evaluations should also test generation (cognitive analogy: Appendix K).

Expected failure cases. Because queries are derived from the Stage I description, LAD cannot recover context for a decisive element that the description omits, such as small embedded text; off-topic or contradictory search results can still enter the context, since ranking only selects among the retrieved answers; and for abstract artworks whose meaning rests on style rather than recognizable entities, keyword queries retrieve little useful context.

Test-time vs. train-time contextual alignment. Contextual alignment can also be internalized by end-to-end reinforcement learning, as in MetaphorStar (Zhang et al., 2026). LAD instead needs no parameter updates, applies to frozen backbones, and exposes an inspectable context at the cost of 15–45 seconds of search per image (Appendix C); training a policy to plan questions from $\mathbf{z}$ could combine both.

6 CONCLUSION

We reformulate image metaphor understanding as a context-dependent compositional reasoning problem, in which meaning emerges from the interaction of visual elements under cultural and social context rather than from isolated recognition. We identify *missing context* as a failure mode of frozen MLLMs and formalize *Contextual Alignment* as inference over a latent context. Our framework, Let Androids Dream (LAD), realizes contextual alignment with training-free test-time computation through perception, search, and reasoning: it lifts a lightweight GPT-4o-mini from 44% to 74% English MCQ accuracy and from 2.98 to 4.02 English OSQ score on S100, improves all four backbones we test, and improves open-ended interpretation where generic search does not. More broadly, many multimodal reasoning tasks require aligning compositional visual structure with knowledge that is not present in the image. For such context-dependent reasoning, how inference-time computation acquires and selects context can matter as much as model scale. Combining test-time context acquisition with train-time alignment is a promising direction for future work.

Ethics Statement

All images come from the public releases of II-Bench (Liu et al., 2024) and CII-Bench (Zhang et al., 2024a); we release only item identifiers and code. The images are comics, illustrations, posters, memes, and artworks, not photographs of private individuals. The human evaluation (Appendix G) was carried out by 16 volunteer researchers from the authors' institutions, who rated anonymized model outputs; no personal data were collected, participants gave informed consent, and no institutional ethics review was sought for this minimal-risk rating task. Retrieved internet-culture context can carry stereotypes; LAD exposes this context as text, and its outputs should not be used for content moderation without human review.

Reproducibility Statement

Section 3 and Algorithm 1 specify the pipeline; Appendix L lists the English LAD and baseline prompts verbatim (the Chinese and WebSearch prompts are in the supplementary code); Appendix C gives model identifiers where available, decoding parameters, the Self-Judge rule, the search engine, and the exclusion of benchmark sources. The anonymized supplementary material contains the code for all three stages and the WebSearch planner–searcher, the evaluation and judging scripts, the list of S100 items, and the scripts for the statistics in Appendix E. All models were accessed through APIs or released weights.

AI Use Statement

We used LLM-based assistants (Claude and GPT models in the Cursor IDE) to polish the writing, convert the manuscript to the ICLR template, check references, plot Figure 4, compute the statistics in Appendix E from the reported results, and run the API evaluations of the later-evaluated baselines with the prompts of Appendix L. They were not used to design LAD or to generate data. The authors reviewed all AI-assisted content and take full responsibility for it.

References

Anthropic. System card: Claude Opus 4 & Claude Sonnet 4, 2025. URL https://www-cdn.anthropic.com/4263b940cabb546aa0e3283f35b686f4f3b2ff47.pdf.

Anthropic. Claude 4.6 Opus model card, 2026a. URL https://www.anthropic.com/research/claude-model-card.

Anthropic. Claude 4.6 Sonnet model card, 2026b. URL https://www.anthropic.com/research/claude-model-card.

Anthropic. Introducing Claude Opus 4.7, 2026c. URL https://www.anthropic.com/news/claude-opus-4-7.

Anthropic. Introducing Claude Sonnet 5, 2026d. URL https://www.anthropic.com/news/claude-sonnet-5.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *ICLR*, 2024.

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.

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.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

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

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

Gemini Team. Gemini 3 pro: Best for complex tasks and bringing creative concepts to life, 2025. URL https://deepmind.google/models/gemini/pro/.

Xinyu Geng, Peng Xia, Zhen Zhang, Xinyu Wang, Qiuchen Wang, Ruixue Ding, et al. WebWatcher: Breaking new frontier of vision-language deep research agent. *arXiv preprint arXiv:2508.05748*, 2025.

Google DeepMind. Gemini 3.1 Pro model card, 2026a. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/.

Google DeepMind. Gemini 3.8 Flash model card, 2026b. URL https://deepmind.google/models/model-cards/gemini-3-8-flash/.

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.

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.

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.

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

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.

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

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*, 2024a.

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*, 2024b.

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.

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

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.

OpenAI. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024a.

OpenAI. Learning to reason with LLMs, 2024b. URL https://openai.com/index/learning-to-reason-with-llms/.

OpenAI. GPT-5.4 system card, 2026a. URL https://openai.com/index/gpt-5-system-card.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition, 2026b. URL https://openai.com/index/gpt-5-6/.

Yiming Qin, Bomin Wei, Jiaxin Ge, Konstantinos Kallidromitis, Stephanie Fu, Trevor Darrell, et al. Chain-of-visual-thought: Teaching VLMs to see and think better with continuous visual tokens. *arXiv preprint arXiv:2511.19418*, 2025.

Qwen Team. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, 2026. URL https://qwen.ai/blog?id=qwen3.8.

ByteDance Seed. Doubao-1.5-thinking-vision-pro, 2025. URL https://console.volcengine.com/ark/region:ark+cn-beijing/model/detail?Id=doubao-1-5-thinking-vision-pro.

Xijia Tao, Yihua Teng, Xinxing Su, Xinyu Fu, Jihao Wu, Chaofan Tao, et al. MMSearch-Plus: Benchmarking provenance-aware search for multimodal browsing agents. *arXiv preprint arXiv:2508.21475*, 2025.

Qwen Team. QVQ: To see the world with wisdom, 2024a. URL https://qwenlm.github.io/blog/qvq-72b-preview/.

Qwen Team. QwQ: Reflect deeply on the boundaries of the unknown, 2024b. URL https://qwenlm.github.io/blog/qwq-32b-preview/.

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

Jinming Wu, Zihao Deng, Wei Li, Yiding Liu, Bo You, Bo Li, et al. MMSearch-R1: Incentivizing LMMs to search. *arXiv preprint arXiv:2506.20670*, 2025.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal LLMs. *arXiv preprint arXiv:2312.14135*, 2023.

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.

xAI. Grok 3 beta – the age of reasoning agents, 2025. URL https://x.ai/news/grok-3.

xAI. Introducing Grok 4.6, 2026. URL https://x.ai/news/grok-4-6.

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*, 2024a.

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

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*, 2024a.

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

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.

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.

Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, et al. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*, 2022.

Chenhao Zhang, Xi Feng, Yuelin Bai, Xinrun Du, et al. Can MLLMs understand the deep implication behind Chinese images? *arXiv preprint arXiv:2410.13854*, 2024a.

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

Tuo Zhang, Tiantian Feng, Yibin Ni, Mengqin Cao, Ruying Liu, Katharine Butler, et al. Creating a lens of Chinese culture: A multimodal dataset for Chinese pun rebus art understanding. *arXiv preprint arXiv:2406.10318*, 2024b.

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

Zhipu.ai. GLM-4V, 2024. URL https://www.bigmodel.cn/dev/api/normal-model/glm-4v.

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.

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.

APPENDIX OVERVIEW

- A: Extended limitations and expected failure cases.
- B: Full algorithm of LAD.
- C: Implementation details (models, decoding, routing, retrieval, baselines, cost).
- D: Additional baselines.
- E: Statistical analysis.
- F: Open-style evaluation protocol.
- G: Human evaluation study.
- H: S100 benchmark statistics.
- I: Search-comparison figure and general multimodal benchmarks.
- J: Qualitative analysis and case study.
- K: Cognitive motivation of the design.
- L: English prompts, verbatim.

A EXTENDED LIMITATIONS AND EXPECTED FAILURE CASES

Evaluation scale. S100 contains 50 items per language, so a single item changes accuracy by 2 points and the 95% interval half-width is 11–13 points. We use S100 for the broad model comparison, the ablations, and OSQ, where each system requires one judged answer per image, and we base significance claims on the full II-Bench and CII-Bench (Table 6, Appendix E). Ablations on the full benchmarks would give tighter estimates of each component's contribution.

Granularity of interventions. The component analysis adds or replaces whole stages. The formulation of Section 3.1 suggests finer interventions that we did not run: replacing $\hat{c}$ with the context of another image (to rule out that any additional text helps), supplying gold annotations of emotion, domain, and rhetoric as $\hat{c}$ (an upper reference; the II-Bench analysis (Liu et al., 2024) reports that such labels raise accuracy), removing the image from Stage III (to measure how much the answer depends on visual input once the description is available), and routing all questions to one source (to evaluate Self-Judge).

Prompt sensitivity. The prompt templates were written once and reused without change for Qwen2.5-VL-7B/72B, GPT-4o, and the general benchmarks (with Chinese versions for Chinese images); this reuse is not evidence of robustness, and we did not measure how results vary under paraphrased instructions.

Judge. OSQ uses one judge model (`gpt-4o-2024-11-20`; the `gpt-4o` alias for the later-evaluated models). The judge prompt was written for internet memes and is applied to all image types; its criteria on key elements, emotion, domain, rhetoric, and depth are general, but the persona may be less suited to traditional paintings. Human agreement (Appendix G) and same-backbone comparisons (Section 4.3) mitigate but do not remove judge-specific preferences.

Retrieval dependence and cost. Results depend on the search engine and its index at run time and can drift. The search stage adds 15–45 seconds and 860–1,070 tokens per image (Appendix C); caching repeated queries and early stopping when Self-Judge is confident would reduce this.

Expected failure cases. By construction, LAD cannot recover from a Stage-I description that omits the decisive element (e.g., small embedded text), because the keywords and queries are derived from it; Stage III still receives the image, which can partly compensate. When retrieved pages are off-topic or contradictory, the ranking step can only choose among the five answers, so irrelevant context may still enter $\hat{c}$. For highly abstract artworks whose implication rests on style rather than recognizable entities, keyword queries retrieve little useful context. Finally, retrieved internet-culture content can carry stereotypes that the model may reproduce.

B ALGORITHM

Algorithm 1: Let Androids Dream (LAD)

Input: Image x; task $T \in \{\text{MCQ}, \text{OSQ}\}$ (question and options for MCQ)
Output: Answer $\hat{y}$

// Stage I: Perception, $z = \phi(x)$
$d \leftarrow$ MLLM.Describe(x) /* one-paragraph description */
$k \leftarrow$ MLLM.ExtractKeywords(d) /* 7 keywords: emotion, domain, rhetoric */
// Stage II: Search, $\hat{c} = S(R(Q(z)))$
$\mathcal{Q} \leftarrow$ MLLM.PlanQuestions(k) /* 5 search questions */
$\mathcal{A} \leftarrow \emptyset$
for *each* $q \in \mathcal{Q}$ **do**
  $(s, dec) \leftarrow$ MLLM.SelfJudge(q) /* $s \in \{1, \ldots, 5\}$ */
  if $s \leq 3$ ***and*** $dec =$ *'WebSearch'* **then**
    $a \leftarrow$ WebSearch(q) /* external knowledge */
  else
    $a \leftarrow$ MLLM.ModelSearch(q) /* parametric knowledge */
  end
  $\mathcal{A} \leftarrow \mathcal{A} \cup \{(q, a)\}$
end
$(\mathcal{A}_3, r) \leftarrow$ MLLM.Rank($d, \mathcal{A}$) /* top-3 pairs and ranking reasons */
$\hat{c} \leftarrow$ MLLM.Refine($r, \mathcal{A}_3$) /* context summary */
// Stage III: Reasoning
$\hat{y} \leftarrow$ MLLM.Reason($x, d, k, \hat{c}, T$)
return $\hat{y}$

Function *WebSearch*(q)
  $G \leftarrow \{q\}$ /* planner graph, root = question */
  repeat
    MLLM.Planner adds sub-questions to G, each linked to the nodes it depends on
    for *each new node* u *(in parallel)* **do**
      $res \leftarrow$ DuckDuckGo(MLLM.RewriteQuery(u, answers of parents))
      $pages \leftarrow$ Fetch(MLLM.SelectPages(res, u)) /* benchmark sources excluded */
      answer(u) $\leftarrow$ MLLM.AnswerFromPages($pages, u$)
    end
  until *the planner decides to answer*
  return MLLM.Planner.Respond(G)

C IMPLEMENTATION DETAILS

Backbones. The default backbone is GPT-4o-mini (`gpt-4o-mini-2024-07-18`). In the backbone study (Table 4(a)), each backbone runs all three stages, including Self-Judge, ModelSearch, and the WebSearch planner and searcher, with identical prompts. API models are queried through their official APIs; open-source models are served locally.

Decoding. top_p is 0.9 for all calls. Temperatures: description 0.7, keyword extraction 0.7, question planning 0.7, Self-Judge 0.0, ModelSearch 0.5, ranking 0.0, refinement 0.5, OSQ reasoning 0.7, MCQ reasoning 0.5. Closed-book baselines use temperature 0.5 for MCQ and 0.7 for OSQ with the prompts in Appendix L. MCQ answers are parsed from the `<answer>` tag (LAD) or the "Answer: $LETTER" pattern (baselines).

Self-Judge. The judge outputs "[score, decision]" with an integer score from 1 to 5 (prompt in Appendix L). A question is sent to WebSearch if the score is at most 3 and the decision is WebSearch, and to ModelSearch if the score is above 3 and the decision is ModelSearch; inconsistent or malformed outputs default to ModelSearch.

WebSearch. The planner–searcher follows MindSearch (Chen et al., 2024), using the same backbone as LAD for planning, query rewriting, page selection, and answering, and DuckDuckGo as the search engine. Pages from the benchmark sources (the II-Bench and CII-Bench project, dataset, and paper pages) were excluded from the results in the search configuration of our runs; this filter is not

part of the released MindSearch client code. Queries are text only; the image and the benchmark question or options are never sent to the search engine.

Search-augmented baselines. "Grok-3 + built-in search" enables Grok-3's native web search; "GPT-4o + Perplexity" combines GPT-4o with Perplexity search; "GPT-4o-mini + native search" keeps LAD's Stage I and Stage III and replaces Stage II (planning, routing, retrieval, ranking, and condensing) with OpenAI's native web search, so the backbone and the Stage I/III prompts are shared with LAD. Because the search engine and the selection of retrieved content both differ, the comparison tests the acquisition pipeline as a whole rather than a single component.

OSQ references and judge. The reference answer is the human-written explanation released with II-Bench (English) and the concatenation of the human-written metaphorical meaning and explanation released with CII-Bench (Chinese). The judge is GPT-4o (`gpt-4o-2024-11-20`) at temperature 0; it receives the image, the reference, and the answer, and each answer is scored once.

Cost. Measured with the search questions executed concurrently, each search question takes 4–9 seconds and the whole search stage (including planning, ranking, and refinement) takes 15–45 seconds and consumes roughly 860–1,070 tokens per image. Stages I and III add three MLLM calls (description, keywords, reasoning).

D Additional Baselines

Table 5 lists the earlier S100 baselines omitted from Table 2 for space, including the search-augmented Grok-3 and GPT-4o systems also shown in Table 3(b); GPT-4o and Grok-3 are the base models of these systems and of the GPT-4o LAD rows. The later-evaluated models in Tables 2 and 5 (Gemini-3.8-flash, Qwen3.8-max, Claude-Sonnet-5, GPT-5.6-Sol, Grok-4.6, Gemini-3.1-pro, Qwen3-VL-Plus in thinking mode, Claude-Opus-4.7) were evaluated in September 2026 through an OpenAI-compatible API relay with the end-to-end prompts of Figures 6–7 (MCQ temperature 0.5, OSQ 0.7, one sample) and scored by the `gpt-4o` alias with the same judge prompt, because `gpt-4o-2024-11-20` was not available. Gemini-3.1-pro, Qwen3-VL-Plus, and Claude-Opus-4.7 are the `gemini-3.1-pro-preview`, `qwen3-vl-plus-think`, and `claude-opus-4-7` endpoints; Grok-4.6, Qwen3.8-max, and Qwen3-VL-Plus have 49 scored Chinese items (accuracy over 49) because one request failed repeatedly; refusals count as incorrect. We cannot verify which model versions the relay serves, and the cause of the GPT-4o differences was not identified. For reasoning endpoints, the relay returns only the final answer, occasionally preceded by a few words of reasoning residue, which we did not remove. Each endpoint passed a check in which it correctly described the text and costumes of an S100 comic. Under this setup, closed-book GPT-4o obtains 52/52 (MCQ) and 2.88/3.06 (OSQ), versus 74/58 and 2.94/3.76 in the original evaluation, so these rows are not directly comparable with the other rows.

Table 5: Additional baselines on S100 (MCQ accuracy %, OSQ 1–5; Qwen3-VL-Plus was evaluated in the later setup).

Model	MCQ en	MCQ zh	OSQ en	OSQ zh
GPT-4o (OpenAI, 2024a)	74	58	2.94	3.76
GLM-4.1V-8B (Zhipu.ai, 2024)	60	52	2.60	2.96
Grok-3 (xAI, 2025)	66	64	3.24	2.96
DeepSeek-VL2 (Wu et al., 2024)	46	36	2.82	2.86
GLM-4V-plus (Zhipu.ai, 2024)	64	64	3.01	3.12
Claude-4-Sonnet (Anthropic, 2025)	60	62	3.46	3.86
QVQ-72B (Team, 2024a)	62	56	3.10	3.42
Doubao-1.5-thinking-vision-pro (Seed, 2025)	66	66	3.16	3.90
Grok-3-reasoning (xAI, 2025)	74	64	3.06	2.92
Claude-4.6-Sonnet (Anthropic, 2026b)	64	62	3.56	3.72
Claude-4.6-Opus (Anthropic, 2026a)	70	82	3.60	3.94
Grok-3 + built-in search	72	64	3.25	2.92
GPT-4o + Perplexity search	80	66	2.88	3.28
Qwen3-VL-Plus (thinking) (Qwen Team, 2025)	88	76	3.08	4.02

E Statistical Analysis

Table 6: Full benchmarks (II-Bench: 1,399; CII-Bench: 800 items), with approximate 95% Wilson half-widths for MCQ. LAD uses web search; the other rows are closed-book.

	MCQ (%)		OSQ	
Model	II	CII	II	CII
GLM-4.1V-8B	$70.0_{\pm 2.4}$	$46.3_{\pm 3.4}$	2.83	3.06
GPT-4o	$72.6_{\pm 2.3}$	$54.1_{\pm 3.4}$	3.86	4.06
InternVL3-78B	$78.2_{\pm 2.2}$	$64.0_{\pm 3.3}$	3.68	4.06
Claude-4.6-Sonnet	$81.0_{\pm 2.1}$	$\mathbf{64.5}_{\pm 3.3}$	3.62	3.82
GPT-4o-mini	$63.5_{\pm 2.5}$	$35.6_{\pm 3.3}$	2.93	3.29
LAD (GPT-4o-mini)	$\mathbf{81.2}_{\pm 2.0}$	$53.8_{\pm 3.4}$	**4.22**	**4.31**

For an accuracy $\hat{p}$ on n items we report the 95% Wilson interval

$$\left(\hat{p} + \frac{z^2}{2n} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}\right) \Big/ \left(1 + \frac{z^2}{n}\right), \qquad z = 1.96.$$

Differences between two systems are approximated with an unpaired two-proportion z-test on correct-answer counts reconstructed from the reported accuracies; a paired McNemar test would require per-item predictions. The test ignores that both systems answer the same items; it is conservative when correctness is positively correlated across systems, which is typical but could not be verified without per-item predictions, so the p-values should be read as approximate. Table 7 lists the comparisons used in the paper.

Table 7: Significance of key comparisons (two-proportion z-test).

Benchmark	**Comparison**	**Accuracies (%)**	z	p
II-Bench (n=1,399)	GPT-4o-mini + LAD vs. GPT-4o-mini	81.2 vs. 63.5	10.48	1.1×10^{-25}
CII-Bench (n=800)	GPT-4o-mini + LAD vs. GPT-4o-mini	53.8 vs. 35.6	7.29	3.1×10^{-13}
II-Bench	GPT-4o-mini + LAD vs. GPT-4o	81.2 vs. 72.6	5.38	7.3×10^{-8}
II-Bench	GPT-4o-mini + LAD vs. Claude-4.6-Sonnet	81.2 vs. 81.0	0.14	0.89
CII-Bench	GPT-4o-mini + LAD vs. GPT-4o	53.8 vs. 54.1	−0.15	0.88
S100-en (n=50)	GPT-4o-mini + LAD vs. GPT-4o-mini	74 vs. 44	3.05	0.002
S100-zh (n=50)	GPT-4o-mini + LAD vs. GPT-4o-mini	52 vs. 42	1.00	0.32
S100-en	GPT-4o-mini + LAD vs. Gemini-3.0-pro	74 vs. 76	−0.23	0.82

F Open-Style Evaluation Protocol

Figure 5 shows the five aspects of the rubric: surface-level information (whether the key elements such as embedded text and important entities are captured), emotional expression, domain and context, rhetorical devices, and deep implications. The judge assigns a single holistic score from 1 to 5 following the level descriptions in the judge prompt (Appendix L). All systems receive the same OSQ instruction ("No need to describe images and text, only answer metaphors ... try to be concise"); the surface-level aspect concerns whether the answer is grounded in the correct elements, which interpretations typically reference, and it applies equally to all systems. OSQ complements MCQ because six-option questions can be answered by eliminating distractors and are sensitive to option order and format (Zheng et al., 2024; Li et al., 2024a; Myrzakhan et al., 2024).

G Human Evaluation Study

Participants. Sixteen PhD students and researchers, recruited as volunteers within the authors' institutions, took part. All are proficient in English and Chinese and familiar with comics, memes, and visual metaphors. Participation was voluntary and uncompensated, participants were informed

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 5: Aspects and level descriptions of the open-style evaluation.

about the purpose of the study and how their ratings would be used, and no personal data were collected. The study involved only rating model outputs for public images; no independent institutional ethics review or exemption determination was obtained.

Materials and instructions. We randomly sampled 25 English and 25 Chinese S100 images. For each image, participants saw the image, the human reference answer, and anonymized model answers in random order, and were instructed: "Score each answer from 1 to 5 using the scoring standard. Judge how well the answer captures the implied meaning of the image with respect to the key elements, emotion, domain, rhetorical devices, and depth of implication, compared with the reference answer. Do not let the length of the answer influence your score." The scoring standard was the same level description given to the judge (Appendix L).

Agreement measure. Let $h_{i,1}, \ldots, h_{i,m}$ be the human scores for answer i and $\tilde{h}_i$ their mean after removing the highest and the lowest score. With judge score g_i, agreement is the fraction of answers with $|\tilde{h}_i - g_i| \leq 1$. Trimming removes individual outliers before averaging; the criterion allows one point of disagreement on a 5-point scale and therefore measures coarse calibration. Agreement is 96.5% for English, 94.8% for Chinese, and 95.7% overall.

H S100 Benchmark Statistics

S100 was built by selecting 50 images from II-Bench and 50 from CII-Bench that are diverse in domain and image type and rich in visual and cultural content. Table 8 lists the distribution. Each image has one six-option MCQ; most questions ask for the overall implication, and some ask about specific details that are necessary for it.

I Additional Results: Search Comparison and General Benchmarks

Figure 4 (Section 4.4) visualizes the search ablation of Table 3(b) and the backbone study of Table 4(a).

Table 8: Distribution of domains and image types in S100.

English images (II-Bench)	
Society	21 (42%)
Life	16 (32%)
Art	6 (12%)
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%)

Chinese images (CII-Bench)	
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%)

To test whether the decomposition applies beyond implications, we ran LAD with GPT-4o-mini and unchanged prompts on MMMU (Yue et al., 2024) (expert-level multi-discipline reasoning), SeedBench (general understanding), and MMStar (general understanding). Table 4 (Section 4.4.4) reports accuracy. WebSearch was enabled with the same exclusion of benchmark sources as in the main experiments, so the comparison with closed-book models is not controlled for information access: part of the gain, especially on knowledge-intensive MMMU questions, may come from retrieved domain knowledge. We report these results as evidence that the pipeline transfers without task-specific changes, not as evidence of a reasoning improvement independent of retrieval.

J Qualitative Analysis

Figure 3 (Section 4) compares the end-to-end baseline, standard CoT, and LAD on a comic about a princess and a frog. The end-to-end model focuses on the princess accepting the frog and answers that love is not all-powerful (B), staying with the literal scene (*superficial reading*). Standard CoT links the heart symbol to the familiar transformation narrative and answers that princesses can transform frogs (D) (*over-inference*): it reasons explicitly, but with a context recalled from its parameters that does not fit this composition. LAD identifies the transformation motif, reads the humorous twist as a comment on fairy-tale expectations, and selects the correct option that fairy-tale narratives do not always translate to reality (C). The two failure patterns correspond to the perception failure and missing context that Stages I and II target (Sections 3.2–3.3). The superficial-reading failure motivates the emotion, domain, and rhetoric keywords of Stage I; the II-Bench analysis reports that supplying such labels raises accuracy (Liu et al., 2024).

K Cognitive Motivation

The title and epigraph pose two questions. *Do androids dream?*—can a model interpret implications at all—is addressed by Perception and Reasoning, which turn pixels into a structured reading and connect it to a conclusion. *Do they dream of electric sheep?*—are the interpretations specific and grounded in the right context—is addressed by Search, which supplies the cultural and social background the image presupposes.

Cognitive motivation. LAD's design is an engineering analogy, not a model of human cognition. Perception follows a coarse-to-fine order that mirrors how people first form a holistic impression of an image and then identify its key elements; in dual-process terms (Evans, 2003), the description corresponds to the fast holistic impression and the keyword extraction to deliberate analysis. Search follows accounts of active information seeking (Ikoja-Odongo & Mostert, 2006; Wilson, 2009): when people meet an unfamiliar reference, they look it up. Reasoning connects the elements with the retrieved background, which resembles the social inference involved in interpreting intent (Kim et al., 2025). We use these analogies only to motivate the decomposition; the experiments evaluate it on its own terms.

L PROMPTS

All prompts are shared by all backbones. We list the English LAD prompts verbatim; {<slot>} marks where inputs (description, keywords, questions, retrieved answers, or the MCQ) are inserted. The Chinese prompts are translations with the same structure and are included in the supplementary code, together with the WebSearch planner and searcher prompts from MindSearch.

Stage I: image description.

```
Please provide a description of the image in a paragraph. You should consider the role, text, color, and layout
 of the image. Focus on the text and important elements in the image. Try to be concise while ensuring the
correctness of the description.
```

Stage I: keyword extraction.

```
Please provide seven keywords most related to the metaphor based on the image description. You should consider
the emotion, domain, and rhetoric in the image description. Focus on the images text and important entities in
the image description. Note that there may be homophonic memes and puns, distinguish and provide them but do
not repeat the same element. Take a deep breath and think step by step, only output the keywords.
Image description: {<slot>}
```

Stage II: search-question planning.

```
# Role
You are a researcher familiar with internet culture and memes, skilled at uncovering and analyzing the deeper
meanings of internet memes.
## Attention
Your responsibility is to design precise search questions based on keywords that can accurately retrieve meme
culture or image metaphors.

## Skills
### Skill 1: Internet Search Skills
- Analyze the provided keywords and select the most relevant ones, understanding their potential meanings and
applications in metaphorical culture

### Skill 2: Cultural Analysis
- Deeply explore the background and history of keywords in internet culture, through information organization
and analysis, to uncover the deeper meanings behind the keywords

### Skill 3: Metaphor Comprehension
- Understand and analyze the metaphorical meanings of keywords in images, to better comprehend the cultural
background of meme images

## Workflow:
1. Analyze the provided keywords and select the most relevant ones, understanding their potential meanings and
applications in metaphorical culture.
2. Design specific and targeted search questions to improve the accuracy of the search.
3. Ensure that the search questions can guide users to find cultural backgrounds or metaphorical explanations
related to the meme images.

## Constraints
- Search questions should focus on meme culture or image metaphors, avoiding overly broad or irrelevant search
results
- Search questions should consider keyword combinations, avoiding single keyword searches as much as possible
- The total number of search questions should be 5, please design the most relevant questions

## Example:
Input:
"""
Keywords: irony, environmental protection, waste, contrast, empty slogans, lack of action, inconsistency
between words and deeds
"""
Output:
"""
1. "What are the specific manifestations of the inconsistency between environmental protection slogans and
actual actions? "

2. "What is the profound connotation of the popular Internet pictures comparing environmental protection and
waste? "

3. "Dual information pictures satirizing environmental protection slogans and lack of action? "

4. "Ironic contrast pictures of the contradiction between slogans and actual actions in environmental
protection themes? "
```

```
5. "How do Internet pictures show the ironic contrast between environmental protection slogans and wasteful
behavior? "
"""

## Solve:
keywords: {<slot>}
```

Stage II: Self-Judge routing.

```
# Task
Please evaluate your knowledge level, determine whether the question is suitable for direct answering or
requires external knowledge support, and provide a confidence score according to the Evaluation Standard.

## Evaluation Standard
- If the question contains Internet culture and meme culture, the confidence score is less than 3
- If the knowledge contained in the question is highly popular, the confidence score is less than 3
- If the real-time content contained in the question accounts for a large proportion, the confidence score is
less than 3
- If the entity contained in the question is relatively niche, the confidence score is less than 3
- If the question contains more than 2 entities, the confidence score is less than 3

## Workflow:
1. <Implicitly> Analyze the content of the question and identify whether it contains Internet culture and meme
culture elements.
2. <Implicitly> Evaluate the popularity of the question and the proportion of real-time content.
3. According to the scoring criteria, <Explicitly> provide a confidence score and decision.

## Constraints
- The confidence score must be between 1 and 5. If the score is bigger than 3, select ModelSearch. If the score
 is smaller than or equal to 3, select WebSearch.

## OutputFormat:
[Confidence score (1-5 points), Decision (ModelSearch/WebSearch)]

## Solve:
Question: {<slot>}
```

Stage II: ModelSearch.

```
# Role
You are a researcher familiar with internet culture and memes, skilled at uncovering and analyzing the deeper
meanings of internet memes.
## Attention
You are responsible for providing clear, professional and comprehensive explanations of Internet memes and
image metaphors based on Internet culture and meme culture, helping users understand the cultural meaning and
social impact behind them.

## Skills
### Skill 1: Metaphor Understanding
- Ability to accurately identify the memes or metaphors involved in the picture based on picture analysis, and
be able to extract the deep meaning

### Skill 2: Internet Culture Analysis
- Ability to analyze the origin, development and current popularity of memes or image metaphors, so that users
can better understand the cultural background of memes

### Skill 3: Providing Explanations
- Ability to provide users with clear and professional explanations of the cultural meaning and social impact
behind memes or image metaphors
- Ability to provide structured and easy-to-understand answers

## Workflow:
1. Identify the meme or picture metaphor involved in the question raised by the user.
2. Analyze the origin, development and current popularity of the meme or picture metaphor.
3. Explain the cultural meaning and social impact behind the meme or picture metaphor.
4. Provide a structured and easy-to-understand answer.

## Constraints
- The answer content needs to be logically clear and well-structured to ensure that readers can understand it
easily
- The answer content needs to focus on the meme culture or picture metaphor
- The answer part needs to be comprehensive and complete, and there should be no vague expressions such as "
based on the above content" to ensure the credibility of the information
- The language style needs to be professional and rigorous, and avoid colloquial expressions
- Maintain uniform grammar and vocabulary to ensure the consistency and coherence of the overall document
```

```
## Solve:
Question: {<slot>}
```

Stage II: relevance ranking.

```
# Role
You are a researcher familiar with Internet culture and memes, and are good at extracting deep meanings from
images, and matching and sorting these meanings with relevant questions and summaries.
## Attention
You are responsible for sorting the relevance of multiple question-summary pairs based on the image description,
 and selecting the three most relevant ones to help users understand the deep meaning of the image.

## Skills
### Skill 1: Image Analysis
- Ability to conduct a comprehensive analysis based on the user's description of the image

### Skill 2: Metaphor Understanding
- Ability to accurately identify metaphors in images based on image analysis and extract deep meaning

### Skill 3: Relevance Sorting
- Ability to sort the relevance of question-summary pairs based on the metaphorical understanding of the image,
 based on whether the content of the question-summary pair is related to the image metaphor
- Ability to provide reasons for the corresponding relevance sorting, and analyze the low-relevance content of
the question-summary pair

## Workflow:
1. Carefully analyze the image description and <implicitly> extract the metaphors and key elements in the image.

2. Analyze each question-summary pair and <implicitly> evaluate its relevance to the image metaphor.
3. According to the relevance, <implicitly> sort the question-summary pairs.
4. <display> Output the top 3 most relevant question-summary pairs after sorting and the reasons for sorting.

## Constraints
- Sorting should be based on the relevance of the image metaphor to ensure the accuracy and logic of the
sorting results
- If a part of the question-summary pair is not related to the image metaphor, please provide it in the
corresponding reason
- Only output the top 3 most relevant question-summary pairs after sorting and the reasons for sorting, avoid
using markdown format

## Examples:
Input:
"""
Image Description: "A person standing at a crossroads", Question-Summary Pair: "1. Question: "The importance of
 life choices"; 2. Question: "Decision-making process in the face of difficulties"; 3. Question: "Complexity of
 urban planning"; 4. Question: "Navigation skills in travel"; 5. Question: "Personal growth and self-discovery
""
"""
Output:
"""
1. Question: The importance of life choices
  - Sorting Reason: xxx
2. Question: Decision-making process in the face of difficulties
  - Sorting Reason: xxx
3. Question: Personal growth and self-discovery
  - Sorting Reason: xxx
"""

## Solve:
Image Description: {<slot>}
Question-Summary Pair: {<slot>}
```

Stage II: refinement into the context summary.

```
# Role
You are a researcher familiar with Internet culture and memes, and are good at rewriting the selected "question-
summary pairs" according to the relevance sorting reasons to improve the appeal of the content and the accuracy
 of the information.
## Attention
You are responsible for rewriting the selected "question-summary pairs" according to the relevance ranking
reasons, removing repeated semantics, parts with low relevance to Internet culture and meme culture, and
removing markdown and citation formatting symbols, and finally outputting a fused and rewritten content.
## Skills
### Skill 1: Internet Culture Insight
- Able to deeply understand classic film and television characters, classic film and television plots and meme
culture, and accurately grasp network hot spots and popular trends
```

```
### Skill 2: Content Analysis and Rewriting
- Able to conduct detailed analysis of "question-summary pairs" and identify the relevance to Internet culture
and meme culture
- Able to creatively rewrite content based on analysis results to improve the accuracy and attractiveness of
information

### Skill 3: Format Recognition and Processing
- Able to identify and remove markdown and reference format symbols to ensure the simplicity and readability of
 output content
- Able to integrate rewritten content and connect it with appropriate transition sentences to form a coherent
and fluent article

## Workflow:
1. Read and understand the relevance sorting results and all "question-summary pairs" provided by the user.
2. Based on the selection reasons, analyze the relevance of each "question-summary pair" to Internet culture
and meme culture.
3. Rewrite the selected 3 "question-summary pairs", remove repeated semantics, low relevance parts, and format
symbols.
4. Integrate the rewritten content, connect it with appropriate transition sentences to form a coherent and
fluent article.

## Constraints
- The rewritten content should maintain the original meaning and should not change the original information and
 viewpoints
- The language should be concise, clear, easy to understand, and related to the image metaphor
- Do not use any form of markdown and reference format symbols

## Solve:
Relevance Sorting Results: {<slot>}
Most Relevant Question-Summary Pair: {<slot>}
```

Stage III: reasoning for OSQ.

```
Image description: {<slot>};
Keywords: {<slot>};
Keywords description: {<slot>};
Please combine the image, keywords, and description information and try to understand the deep meaning of the
combination of the image and text. 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.
```

Stage III: reasoning for MCQ (the MCQ is inserted as "Question: ... Options: A. ... F. ...").

```
Image description: {<slot>};
Keywords: {<slot>};
Keywords description: {<slot>};
Please combine the image, keywords, and description information and try to understand the deep meaning of the
combination of the image and text. Answer the following multiple-choice questions with the correct option
$LETTER directly. Output the thinking process in <think> </think> and final correct answer in <answer> </answer>
 tags. The output format should be as follows: <think>...</think> <answer>...</answer>
Multiple-choice questions: {<slot>}
Answer:
```

OSQ judge (GPT-4o), reproduced as used, including the typo "Standrad".

```
# 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: {<slot>}

## 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 directly by strictly following this format: "[ratings]"

## Solve:
Model answer: {<slot>}
```

Closed-book baselines. Figures 6–7 show the end-to-end prompts used for all closed-book systems, and Figures 8–9 the standard chain-of-thought prompts of Table 3(a) (Chinese and English). In the Stage-III prompts, the slot labelled "Keywords description" receives the search summary $\hat{c}$; in the no-search setting (LAD-CoT) no search summary is provided.

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: End-to-end baseline prompt for 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: End-to-end baseline prompt for OSQ.

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: Standard chain-of-thought baseline prompt for MCQ (the "Std CoT" row of Table 3(a)).

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: Standard chain-of-thought baseline prompt for OSQ.