MEMEBENCH: DIAGNOSING CULTURAL MEME INTERPRETATION WITH RUBRICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large vision-language models have improved at describing visual content, but accurate descriptions do not ensure interpretation when meaning depends on knowledge beyond the pixels. Memes expose this gap because they rely on cultural entities, background knowledge, and community conventions. Most meme benchmarks reduce interpretation to labels or holistic scores, obscuring where an explanation breaks down. We introduce **MemeBench**, a diagnostic benchmark of 1,253 Chinese and English memes with human-written references and quality-controlled VIKR annotations, centered on anime, comics, games, and adjacent online subcultures. Its **VIKR** schema decomposes explanations into Visual clues, Identity links, Knowledge units, and Reasoning mechanisms, each scored against binary rubric criteria. Across 26 LVLMs, every model covers visible content more reliably than the knowledge needed to interpret it, and even the strongest retains a 22.6% Visual–Knowledge gap. To test whether this diagnosis can guide improvement, we introduce **KAR**, an entity-guided retrieval baseline built on **CultureBase**. Across four controlled models, KAR raises VIKR Success by 3.6–7.4% and, compared with generic retrieval, repairs more answers and breaks fewer. Yet both retrieval conditions improve Identity and Knowledge while reducing Visual coverage in every comparison. MemeBench thus turns meme interpretation from a single score into a rubric-level account of what a model sees, what it knows, and what retrieval repairs or breaks.

1 INTRODUCTION

Large vision-language models (LVLMs) have made rapid progress in recognizing and describing visual content. Yet an accurate description does not guarantee a correct interpretation when meaning depends on knowledge beyond the pixels. In Figure 1, an LVLM correctly describes an ancient Chinese official holding an Oscar statuette, but fails to recognize Zhuge Liang mourning Zhou Yu and interprets the scene as political scheming rather than performative grief. As LVLMs are increasingly expected to interpret open-world visual content, measuring this separation between seeing and meaning becomes more important (Cohen et al., 2025; Yadav et al., 2025; Chiu et al., 2025; Myung et al., 2024).

Existing benchmarks provide limited resolution for studying this problem. General-purpose LVLM benchmarks assess broad visual perception and reasoning (Liu et al., 2024; Yue et al., 2024; Yu et al., 2023; Fu et al., 2023), while meme benchmarks have expanded from classification to captioning, question answering, and open-ended explanation (Nguyen et al., 2025; De Giorgis et al., 2026; Zhou et al., 2026; Zhao et al., 2025; Zhong & Baghel, 2024). Most evaluations nevertheless reduce an interpretation to a label, an answer, or a holistic correctness score. Such a score can indicate whether a response is acceptable, but cannot show whether it misdescribes the image, misses the referenced entity, lacks the relevant background knowledge, or fails to connect these elements into the intended meaning. Models with similar overall scores may omit different information, and the same augmentation may improve one part of an explanation while weakening another. Table 1 compares MemeBench with existing benchmarks along these dimensions.

We introduce **MemeBench**, a bilingual diagnostic benchmark for open-ended meme interpretation. It contains 1,253 Chinese and English memes collected from Bilibili, Reddit, and ImgFlip, centered on ACG (anime, comics, and games) and adjacent online subcultures. These communities provide a

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

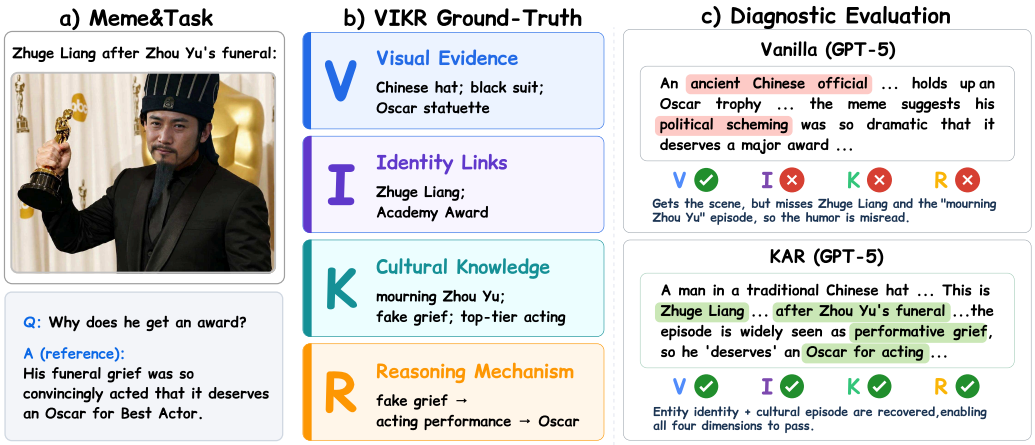


Figure 1: The task and evaluation of MemeBench. **a)** meme with reference answer. **b)** VIKR ground truth and per-dimension rubrics. **c)** In this paired example, the Vanilla response misses the entity and the cultural episode and fails I/K/R; KAR (§4) retrieves both and passes all four. More examples in Appendix E.

concentrated test bed for knowledge-dependent image interpretation because their memes often rely on long-tail characters, source works, community events, and conventions. MemeBench references 2,072 distinct entities, most of which occur in only one meme. Unlike recognition or multiple-choice evaluation, open-ended interpretation requires models to supply rather than select the relevant entities and cultural background. This exposes incomplete but plausible explanations and tests whether external evidence recovers missing content without displacing correct content.

At the core of MemeBench is **VIKR**, which organizes a complete explanation into four kinds of observable content. They cover **V**isual clues, **I**dentity links, **K**nowledge units, and **R**easoning mechanisms. Binary rubric criteria measure which requirements a response covers within each dimension. Passing every rubric criterion in all four dimensions marks a complete interpretation, while individual failures localize omitted content. VIKR characterizes the observable semantics of an explanation while leaving model-internal processing unspecified. It can identify what an explanation is missing and test whether an augmentation fills the diagnosed gap.

Benchmark	Expl.	Bi.	Dec.	Diag.	Aug.
Hateful Memes	✗	✗	✗	✗	✗
MemeCap	✓	✗	✓	✗	✗
D-HUMOR	✗	✗	✗	✗	✗
MemeIntent	✓	✗	✓	✗	✗
MUnd/MetaGPT	✓	✓	✓	✗	✗
MemeLens	✗	✓	✗	✗	✗
M-QUEST	✗	✗	✓	✓	✗
MemeQA	✗	✗	✓	✓	✗
CII-Bench	✗	✗	✗	✗	✗
MemeBench	✓	✓	✓	✓	✓

Table 1: Comparison with existing meme benchmarks: open-ended explanation rather than labels or QA (Expl.), Chinese and English coverage (Bi.), annotation of intermediate meaning components (Dec.), component-level scoring of responses (Diag.), and supplying targeted evidence to test component-level changes (Aug.).

Our evaluation of 26 commercial and open-source LVLMs reveals a consistent separation between visual description and culturally grounded interpretation. The strongest model reaches 60.3% SUCCESS, whereas blind human explanations on a stratified 150-item subset reach 83.0% under the same rubrics. Every model covers visible content more reliably than the background knowledge needed to explain it. This separation remains substantial even for the strongest model, whose Visual coverage exceeds its Knowledge coverage by 22.6%. VIKR also distinguishes models that a single SUCCESS score places together, showing that similar aggregate performance can conceal different missing content. The two language-associated online-cultural ecosystems exhibit the same broad structure, with their larger differences concentrated in Identity and Knowledge rather than Visual coverage.

We next ask whether this diagnosis can guide improvement. We compare reasoning guidance, generic web retrieval, and **KAR**, an entity-guided retrieval condition that grounds cultural entities through **CultureBase** before web search. Reasoning guidance alone does not produce a consistent improvement. Across the four controlled models, both retrieval conditions improve Identity and Knowledge, while Visual coverage declines in all eight retrieval comparisons. KAR raises SUCCESS

by 3.6–7.4% over Vanilla, and in paired comparison with generic web retrieval, it improves SUCCESS for every model, repairs more answers, and breaks fewer. A positive net score can conceal opposing changes within an explanation. By revealing what an explanation gets right, what it misses, and how retrieval changes both, MemeBench turns open-world meme interpretation from a single score into an actionable diagnosis for model improvement. Because the same rubrics score each explanation before and after augmentation, a method earns credit only when it fills the diagnosed gap while preserving content the model already had.

Our contributions are threefold.

1. We introduce **MemeBench**, a bilingual benchmark of 1,253 Chinese and English memes, with VIKR annotations and dimension-specific rubrics for diagnosing open-ended interpretations, scored by two LLM judges whose intersection agrees with human labels at Cohen's κ=0.81.
2. Through an evaluation of 26 LVLMs, we show that models cover visible content more reliably than the entities and background knowledge needed for interpretation, while similar overall scores can conceal different missing content.
3. We introduce **KAR** as an entity-guided retrieval baseline and show that targeted evidence changes explanation components selectively, demonstrating how MemeBench evaluates both component-level gains and preservation under model augmentation.

2 RELATED WORK

Meme understanding benchmarks. Meme benchmarks have progressed from harmful-content classification (Kiela et al., 2020; Pramanick et al., 2021; Bui et al., 2025; Lu et al., 2024) to captioning (Hwang & Shwartz, 2023) and higher-level semantics (Kasu et al., 2025; Nandy et al., 2024; Zhang et al., 2025; Xu et al., 2025). MemeIntent (Park et al., 2024) generates intent descriptions with annotated background knowledge. MetaGPT (Xu et al., 2026) targets metaphor detection, domain extraction, and interpretation, while MemeLens (Shahroor et al., 2026) specializes a multilingual multitask VLM for memes. Cross-cultural work studies transcreation and hateful-meme detection (Zhao et al., 2026; Wang et al., 2026), while MemeQA (Nguyen et al., 2025) and M-QUEST (De Giorgis et al., 2026) decompose understanding through QA. These tasks broaden coverage but do not diagnose free-form explanations through requirement-level Visual, Identity, Knowledge, and Reasoning coverage.

Knowledge-augmented multimodal models. CulturalVQA (Nayak et al., 2024) probes geographically grounded cultural knowledge through visual QA, while BLEnD-Vis (Tan et al., 2026) targets everyday cultural knowledge through visual QA and multiple-choice variants. AVMeme Exam (Jiang et al., 2026) extends cultural evaluation to audio-visual memes. Retrieval-augmented generation has been extended to visual QA (Lewis et al., 2020; Chen et al., 2022; Caffagni et al., 2024; Li et al., 2024), with entity-centric resources supporting factual QA and recognition (Marino et al., 2019; Chen et al., 2023; Mensink et al., 2023; Hu et al., 2023). These settings do not connect requirement-level diagnosis to retrieval: search agents learn *when* to search (Wu et al., 2025; Zhang et al., 2026), whereas KAR uses entity grounding to determine *what* cultural knowledge to query before issuing web searches.

3 MEMEBENCH

MemeBench evaluates open-ended meme interpretation. Given a meme image and a standard explanation prompt, a model produces a free-form explanation without candidate answers or multiple-choice options. Evaluation uses the per-dimension VIKR rubrics to measure coverage of visible content, invoked entities, relevant background, and the reasoning that connects them (Figure 1). Because an answer may satisfy only some requirements, these scores preserve both covered and omitted content instead of collapsing the explanation into a holistic judgment.

3.1 DATA CONSTRUCTION

MemeBench centers on ACG and adjacent online cultures as a concentrated stress test for interpretation that depends on long-tail entities and community-specific knowledge. It comprises **1,253 memes**

(768 Chinese, 485 English) collected between June and December 2025: Chinese samples come from Bilibili, a platform centered on ACG and internet culture, while English samples come from Reddit and ImgFlip.

Construction begins with manual screening for three properties: cultural knowledge beyond surface recognition, an explicitly explainable intended meaning, and sufficient image quality, yielding **1,500 candidate memes**. For each retained candidate, a bilingual annotator familiar with ACG and internet meme culture writes a reference explanation covering visual content, entity identity, cultural background, and humor mechanism, and assigns language and domain metadata. These human-written explanations serve as the semantic source for subsequent structured annotation. Gemini-3.1-Pro then converts each explanation into the VIKR schema (§3.2). Quality control combines automated schema and entity-link validation with targeted manual review of VIKR layer separation, rubric consistency, ambiguity, image quality, and safety; 247 items are removed, producing the final **1,253-item** evaluation set. A reference-independence study on a language–domain-stratified subset of 150 items asks two participants to re-explain each meme under blind conditions, with conflicts adjudicated independently. The study finds 98.0% of references fully or partially compatible and leads to 3 revisions and 0 removals; Appendix A.2 gives the protocol and Appendix A.1 the full selection breakdown.

3.2 VIKR Annotation Schema

VIKR represents a complete explanation through four observable response requirements: **V**isual clues, **I**dentity links, **K**nowledge units, and **R**easoning mechanisms. This representation turns free-form evaluation into a structured diagnosis that localizes missing content while keeping every requirement tied to the same reference interpretation. The schema operates entirely at the response level. It distinguishes whether a response describes the scene, grounds the referenced entities, supplies the required cultural background, and connects these elements to the intended meaning. These requirements may co-occur in an explanation, but each is evaluated independently rather than used as a proxy for another.

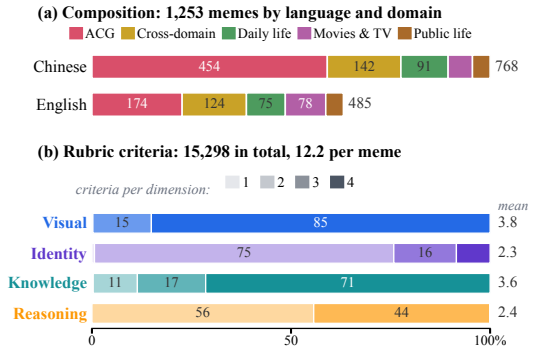


Figure 2: Benchmark scope and evaluation granularity. **(a)** Memes per language, segmented by domain; ACG is the largest domain in both languages (counts in Appendix A.1). **(b)** Share of memes (%) with one to four rubric criteria in each VIKR dimension (darker means more; hues follow Figure 1), with the mean at right; SUCCESS requires all criteria to pass.

The **Visual (V)** layer provides an entity-agnostic visual description, including OCR transcription and the appearance of salient entities, without naming them. The **Identity (I)** layer names entities and their source works, each linked by a shared ID to its appearance in the V layer. Separating Identity from Knowledge is diagnostically important because recognizing an entity does not determine which of its many cultural associations the scene invokes. The **Knowledge (K)** layer captures the background facts and cultural codes the meme requires, and the **Reasoning (R)** layer documents the humor trigger, mechanism, and core message. Each meme also carries language and domain metadata for stratified analysis; metadata does not participate in evaluation.

3.3 Evaluation Protocol

Each VIKR layer yields 1–4 binary rubric criteria, and a layer's score $d_i \in \{0, 1\}$ for $d \in \{V, I, K, R\}$ is set to 1 only if all its criteria are satisfied. The Identity rubric requires exact entity identification ("Kēnán" [Conan] rather than "a detective boy"); the Reasoning rubric accepts semantically equivalent formulations. This criterion-level representation supports both dimension-specific diagnosis and a strict all-dimension primary metric:

$$\text{Success}_i = V_i I_i K_i R_i, \tag{1}$$

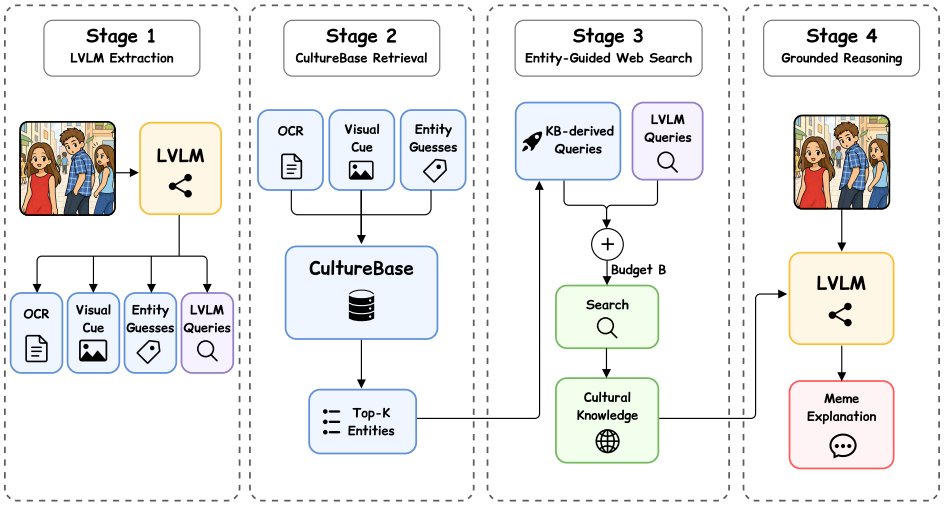


Figure 3: **KAR pipeline. Stage 1:** the LVLM extracts OCR text, a visual cue phrase, entity hypotheses, and suggested web queries. **Stage 2:** these cues retrieve the top-K CultureBase entities. **Stage 3:** entity-derived and LVLM-suggested queries share a budget B for web search. **Stage 4:** the LVLM explains the meme from the original image and the retrieved cultural knowledge. Search-CoT follows the same pipeline without Stage 2.

which is 1 exactly when the response satisfies every annotated requirement. The product preserves the four diagnostic scores while reserving overall success for explanations that cover the complete interpretation. Appendix B.2 reports a three-of-four sensitivity analysis.

Gemini-3.1-Pro and GPT-5.1 independently score each response against the same structured rubric (Appendix B.1), seeing the meme and response but not the model identity. Following the LLM-as-a-judge paradigm (Zheng et al., 2023; Lee et al., 2025), an item passes only when both judges agree. Dimension scores aggregate from these intersection decisions. On the independently human-labeled calibration subset, this intersection reaches Cohen's κ=0.81 on SUCCESS, above either judge alone (0.78 and 0.74) and within 0.03 of the 0.84 agreement between two human raters (Appendix A.2). Across all 26 models, the single-judge leaderboards also agree at Spearman ρ=0.99, showing that the judges differ mainly in strictness rather than model ranking (Appendix B.2).

3.4 BENCHMARK STATISTICS

MemeBench contains 628 ACG items; the remaining 625 span cross-domain references, daily life, movies and television, and public life, broadening the range of cultural references within this sampling frame (Figure 2(a); counts in Appendix A.1). Because the Chinese and English subsets come from different platforms and domain mixtures, cross-language results characterize two language-associated online-cultural ecosystems. The benchmark references 2,072 distinct entities, 86.9% of which appear in exactly one meme, while even the most frequent entity occurs in only 1.3% of items. A further 27.6% of memes cite at least two source works, so coverage of a few popular characters is insufficient. Its rubrics contain 15,298 binary criteria, 12.2 per meme on average (Figure 2(b)). Visual and Knowledge contain at least three criteria for 99.8% and 88.6% of memes, respectively, and SUCCESS requires every rubric criterion to pass. Together, the long-tailed reference space and dense rubric requirements test both the breadth of cultural grounding and the completeness of each explanation.

4 ENTITY-GUIDED RETRIEVAL

To test whether the diagnosed bottlenecks respond to targeted evidence, we build **KAR** (**K**nowledge-**A**ware **R**etrieval), an entity-guided retrieval condition that grounds cultural entities before searching the web. The design uses entity identity as the handle by which background knowledge becomes searchable, since a query anchored to the right entity reaches background that a generic visual description cannot.

CultureBase. The grounding step draws on **CultureBase**, a curated multilingual inventory of cultural entities. Each record stores the canonical name, aliases, source work, character traits, visual description, and cultural usage patterns of one entity, embedded with bge-m3 (Chen et al., 2024) for multilingual similarity retrieval. CultureBase supplies entity anchors and query context; the final explanation is generated jointly from the original image and retrieved web evidence.

Pipeline. KAR runs in four stages, shown in Figure 3. Stage 1 prompts the LVLM to extract OCR text o_i, a short visual cue phrase h_i, entity hypotheses $\mathcal{H}_i$ with confidence levels, and model-suggested web queries $\mathcal{Q}_i^{\text{vlm}}$. Stage 2 forms a three-route query set from the OCR text, cue phrase, and medium- or high-confidence entity names:

$$\mathcal{R}_i = \{o_i, h_i\} \cup \{n_\ell \mid c_\ell \in \{\text{medium}, \text{high}\}\}. \tag{2}$$

Each CultureBase entity is scored by its best match across these routes:

$$\text{score}_{ij} = \max_{r \in \mathcal{R}_i} g(r)^\top u_j, \tag{3}$$

where $g(\cdot)$ is the bge-m3 embedding function and u_j is the embedding of entity e_j. KAR retains the top five entities above the fixed similarity threshold τ=0.5, allowing textual and visual cues to recover an identity even when the model's initial name hypothesis is incomplete. Stage 3 converts each retained entity into a templated query containing its name, source, and meme origin, then merges these queries with $\mathcal{Q}_i^{\text{vlm}}$ under a shared budget. The merged set is issued to Tavily, and Stage 4 returns the retrieved evidence with the original image for the final explanation.

The generic condition, Search-CoT, shares the extractor, search interface, query budget, evidence formatting, and grounded reasoner, and differs only in omitting the CultureBase stage. If no CultureBase entity exceeds τ, no entity-grounded query is added and KAR follows the same query path as Search-CoT. Both retrieval conditions use two LVLM calls and at most five Tavily queries per item; KAR adds one local CultureBase lookup while keeping the web-query budget fixed. This isolates entity grounding under a fixed web backend and evidence budget; Appendix C gives prompts, hyperparameters, and coverage definitions.

5 Experiments

5.1 Setup

We evaluate MemeBench along two complementary axes. Closed-book evaluation measures image-only interpretation under the standard prompt, while paired augmentation settings test how explanations for the same items change when reasoning guidance or external evidence is added. All experiments use the full benchmark (n=1,253).

Metrics. The primary metric is Success (Eq. 1), which requires all four VIKR dimensions to pass. We also report per-dimension coverage as defined in §3.3.

Inference settings. The four settings are ordered so that each adds exactly one variable to the previous one. *Vanilla* gives only the meme image and the standard prompt; *CoT* adds step-by-step reasoning instructions without external evidence. *Search-CoT* adds generic web evidence through KAR's extractor, search interface, evidence formatting, and grounded reasoner while omitting CultureBase. *KAR* adds CultureBase entity grounding before web search (§4). Both retrieval settings use Tavily under the same query budget; adjacent settings thus isolate reasoning guidance, external evidence, and entity grounding in turn.

Models. We evaluate 26 models, including nine commercial systems (Google DeepMind, 2026; Singh et al., 2026) and open-source families spanning Qwen3-VL (Bai et al., 2025), InternVL3.5 (Wang et al., 2025), Kimi-K2.5 (Team et al., 2026), MiMo (Team et al., 2025), and LLaVA-OneVision (An et al., 2025). This main leaderboard compares general-purpose LVLMs under one image-to-free-form protocol; judge robustness and aggregation sensitivity are reported in Appendix B.2. Table 2 lists them in full. Figure 5 summarizes augmentation outcomes for all 26 models; Table 3 gives two complementary four-model comparisons. Gemini-3.1-Pro and Qwen3-VL-235B-A22B-Instruct span the leaderboard's capability range. GPT-5.6-Sol and Kimi-K2.5 form

Model	Overall (n=1,253)					Chinese (n=768)					English (n=485)				
	SUCC.	V	I	K	R	SUCC.	V	I	K	R	SUCC.	V	I	K	R
COMMERCIAL MODELS															
Gemini-3.1-Pro	**60.3**	**92.3**	**74.6**	**69.7**	**76.7**	**49.9**	**90.4**	**66.8**	**61.1**	**67.6**	**76.7**	**95.3**	**87.0**	**83.3**	**91.1**
Gemini-3-Flash	53.7	86.0	68.7	65.5	70.3	41.8	82.2	58.9	54.8	59.4	72.6	92.2	84.3	82.5	87.6
GPT-5.6-Sol	40.9	72.5	58.7	55.1	60.7	33.1	69.3	51.6	46.4	51.3	53.4	77.7	69.9	69.1	75.7
Claude-Opus-5	38.3	74.7	50.4	50.9	57.9	28.3	69.9	41.5	39.7	46.9	54.2	82.3	64.5	68.7	75.5
Qwen3.6-Plus	28.4	65.1	47.2	42.8	47.7	25.5	62.9	44.0	37.8	41.0	33.0	68.7	52.4	50.7	58.4
GPT-5.6-Terra	27.8	61.9	45.4	42.0	46.9	22.3	57.2	40.5	33.2	37.8	36.5	69.5	53.2	55.9	61.4
GPT-5.6-Luna	18.0	52.1	34.9	31.3	35.0	15.5	49.1	31.4	24.9	28.8	22.1	56.9	40.4	41.4	44.7
Claude-Sonnet-5	17.8	59.2	33.3	30.2	35.6	12.2	55.3	25.0	20.8	26.0	26.6	65.4	46.4	44.9	50.7
Qwen3.6-Flash	12.3	53.2	28.6	25.7	29.0	10.5	50.7	26.2	21.0	22.7	15.1	57.1	32.4	33.2	39.0
OPEN-SOURCE MODELS															
Kimi-K2.5	**37.9**	**71.1**	**59.2**	**54.5**	**58.7**	**32.4**	**68.1**	**56.1**	**48.2**	**50.5**	**46.6**	**75.9**	**64.1**	**64.5**	**71.5**
Qwen3-VL-235B-A22B-Instruct	14.5	53.2	30.6	26.1	27.3	10.2	49.9	26.0	18.9	20.3	21.4	58.4	37.9	37.5	38.4
MiMo-V2-Omni	14.3	48.1	29.7	26.4	28.8	10.8	44.3	24.9	19.4	21.2	19.8	54.2	37.3	37.5	40.8
MiMo-V2.5	13.7	47.2	30.5	26.9	28.3	9.9	43.6	25.7	19.5	19.4	19.8	52.8	38.1	38.6	42.5
Qwen3-VL-235B-A22B-Thinking	12.0	50.2	27.5	22.9	24.7	9.6	48.4	23.0	17.6	19.3	15.7	53.0	34.6	31.3	33.4
Qwen3-VL-32B-Instruct	8.4	47.6	20.1	19.1	20.0	5.9	44.4	16.5	13.9	13.2	12.4	52.8	25.8	27.2	30.9
Qwen3-VL-32B-Thinking	7.7	44.5	20.5	17.9	19.7	5.5	41.5	17.1	12.5	13.4	11.1	49.3	26.0	26.4	29.7
Qwen3-VL-30B-A3B-Thinking	5.8	41.6	18.3	14.7	15.8	3.6	38.0	14.1	10.5	10.7	9.3	47.2	24.9	21.2	23.9
Qwen3-VL-8B-Thinking	3.4	32.1	14.0	10.0	9.9	2.7	30.5	12.0	6.6	7.8	4.3	34.6	17.3	15.3	13.2
Qwen3-VL-30B-A3B-Instruct	3.3	39.7	14.8	12.2	11.3	2.7	37.9	10.9	8.3	7.7	4.1	42.7	21.0	18.4	17.1
Qwen3-VL-8B-Instruct	3.3	34.0	13.2	9.3	9.3	2.0	30.2	8.7	6.0	6.1	5.4	40.0	20.2	14.6	14.2
InternVL3.5-38B	2.6	35.6	12.5	7.8	8.2	1.4	33.5	9.5	4.3	4.4	4.5	39.0	17.3	13.4	14.2
InternVL3.5-30B-A3B	1.1	26.4	8.8	5.6	6.2	0.7	24.6	6.6	3.0	3.8	1.9	29.3	12.2	9.7	10.1
InternVL3.5-20B-A4B	1.1	22.0	8.5	5.4	5.3	1.2	20.1	6.8	2.9	2.9	1.0	25.2	11.3	9.5	9.1
InternVL3.5-14B	1.1	26.0	9.3	6.0	6.1	1.2	25.0	7.0	4.0	3.8	1.0	27.6	13.0	9.1	9.9
InternVL3.5-8B	1.0	21.1	8.3	5.2	4.9	0.7	19.3	6.1	2.9	2.5	1.6	23.9	11.8	8.9	8.9
LLaVA-OV-8B-RL	0.4	18.8	9.0	4.2	3.2	0.1	14.1	6.5	2.1	2.0	0.8	26.2	13.0	7.4	5.2

Table 2: Closed-book VIKR performance (%), scored under the dual-judge intersection (Gemini-3.1-Pro ∩ GPT-5.1). SUCC. requires $V \wedge I \wedge K \wedge R$ and is shaded by value; V/I/K/R are layer-wise coverage. Within each model group, **bold** marks the best and underline the second-best score per column. Unscored responses count as failures. Per-judge leaderboards are in Appendix B.2.

a matched commercial–open-source pair whose Vanilla profiles agree within 2.0% on every VIKR dimension, holding the initial diagnosis nearly fixed while model provenance varies.

5.2 CLOSED-BOOK RESULTS

Table 2 reports performance under a common image-only, closed-book protocol. On this measure, MemeBench remains far from saturated. Gemini-3.1-Pro leads with 60.3% SUCCESS, followed by Gemini-3-Flash (53.7%) and GPT-5.6-Sol (40.9%). Claude-Opus-5 follows at 38.3%, while Kimi-K2.5 is the strongest open-source model at 37.9%, only 3.0% behind GPT-5.6-Sol; every model outside the top three remains below 40%. Within the GPT-5.6 family, Sol (40.9%), Terra (27.8%), and Luna (18.0%) are consistently ordered across SUCCESS and all four dimensions, with the largest separations appearing in Identity, Knowledge, and Reasoning. Within Qwen3-VL, Instruct leads at 235B and 32B, while Thinking leads at 30B and ties at 8B; Thinking therefore does not automatically yield higher complete VIKR coverage. Closed-book performance also differs across the two sampled ecosystems. Gemini-3.1-Pro scores 76.7% on English and 49.9% on Chinese, while GPT-5.6-Terra scores 36.5% and 22.3%, respectively. The same direction holds for GPT-5.6-Sol and Kimi-K2.5, showing that the aggregate difference is not confined to one model family. The six models above 65% Visual coverage still span 28.4–60.3% SUCCESS, motivating the layer-wise diagnosis below.

5.3 VIKR DIAGNOSTICS

Every model sits on the same side of the Visual-Knowledge diagonal (Figure 4(a)). The gap spans 14.6 to 29.0% across the evaluated models and remains 22.6% for the strongest, showing that Knowledge remains a bottleneck across capability tiers. Layer-wise scoring also separates models a single score ranks together. The tightest SUCCESS tie in Table 2 is GPT-5.6-Luna and Claude-Sonnet-5, 0.2%

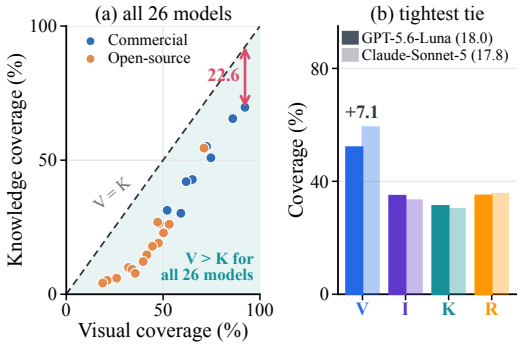


Figure 4: **VIKR diagnostics. (a)** Visual–Knowledge coverage of 26 models; the arrow marks the strongest model's 22.6% gap. **(b)** Closest SUCCESS pair (0.2%).

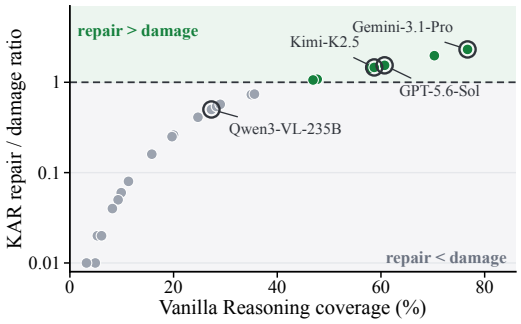


Figure 5: **KAR repair–damage balance.** Ratio of repair to damage rates under KAR versus Vanilla Reasoning coverage (26 models); dashed line: equal rates.

apart, yet their failures sit in different places (Figure 4(b)). Claude-Sonnet-5 covers 7.1% more visual content while grounding 1.6% fewer entities and 1.1% less background knowledge, so its bottleneck is cultural grounding, not perception. The same diagnostic separation appears near the top of the leaderboard: Claude-Opus-5 and Kimi-K2.5 are only 0.4% apart in SUCCESS, but Opus covers 3.6% more Visual content and 8.8% fewer Identity links, and their Visual–Knowledge gaps differ by 7.2%. Across Table 2, seven model pairs lie within 0.6% in SUCCESS while their Visual–Knowledge gaps differ by at least 4%.

The same resolution appears within individual models. Responses that identify an entity without supplying the required background (I=1, K=0) account for a median 13.0% of responses. The reverse pattern (I=0, K=1) accounts for 5.2%, while responses that satisfy both but miss the interpretation (I=1, K=1, R=0) account for 2.2%. Across the 26 Vanilla runs, the pooled responses realize all 16 VIKR states, including 0001 and 0010, with non-monotonic states accounting for 10.9% to 29.2% per model (Appendix B.2). These patterns preserve response-level distinctions that a single SUCCESS score collapses. Identity and Reasoning also track one another closely—within 4% for most models—while remaining separable in the observed state patterns.

Cross-language results further localize the gap. From English to Chinese, Gemini-3.1-Pro drops 4.9% on Visual but 20.2% on Identity and 22.2% on Knowledge; for GPT-5.6-Terra, the Knowledge gap is 22.7% versus 12.3% on Visual. Thus, the larger differences concentrate in the culturally loaded dimensions. Because the subsets differ in platform and domain composition, including ACG shares of 59% in Chinese and 36% in English, they characterize two sampled online-cultural ecosystems rather than a causal language effect.

This diagnostic structure yields a testable prediction. If VIKR captures distinguishable response requirements, targeted evidence should selectively change the diagnosed layers rather than move all four dimensions together.

5.4 RETRIEVAL ANALYSIS

Across the four models, KAR raises Knowledge by 4.1 to 5.0% and Identity by up to 6.4%, directly moving the diagnosed bottlenecks. CoT changes SUCCESS by at most 3.1% without a consistent direction, whereas both retrieval settings improve Identity and Knowledge.

Paired transitions expose what net SUCCESS conceals. Under KAR, Gemini-3.1-Pro repairs 26.5% of its failures and breaks 11.5% of its passes; Qwen3-VL-235B-A22B-Instruct repairs 9.2% and breaks 18.1%. Yet Qwen gains more net SUCCESS (+5.2% versus +3.6%), reversing the ranking by preservation quality. Across 26 models, ratio estimates are non-decreasing from CoT to Search-CoT to KAR, with 21 strict increases. Under KAR, the seven estimates above 1 are exactly the seven models with highest Vanilla Reasoning coverage (Figure 5). Because the rates condition on different Vanilla subsets, the ratio measures preservation balance, not net efficacy; every KAR ΔSUCCESS estimate remains positive. The comparison therefore emphasizes preservation quality rather than only average score changes.

Model / Setting	VIKR coverage (Vanilla / Δ) V	I	K	R	Success and preservation SUCC.↑	Rep.↑	Dmg.↓	Ratio↑
Gemini-3.1-Pro	92.3	74.6	69.7	76.7	60.3	–	–	–
+ CoT	−0.1	+1.0	+0.7	+0.8	60.4	16.3	**10.5**	1.55
+ Search-CoT	−2.9	+3.4	+2.4	+2.2	60.6	21.7	13.8	1.57
+ KAR	−2.5	+5.7	+4.1	+3.7	**63.9**	**26.5**	11.5	**2.31**
GPT-5.6-Sol	72.5	58.7	55.1	60.7	40.9	–	–	–
+ CoT	−1.0	+0.5	+0.3	+0.4	40.7	11.4	16.8	0.67
+ Search-CoT	−1.6	+3.9	+3.0	+2.5	44.1	17.6	17.7	0.99
+ KAR	−0.4	+6.4	+5.0	+4.2	**47.8**	**21.2**	**13.8**	**1.54**
Kimi-K2.5	71.1	59.2	54.5	58.7	37.9	–	–	–
+ CoT	−1.1	+0.4	+0.3	+0.3	37.8	10.7	17.7	0.60
+ Search-CoT	−1.5	+3.6	+3.0	+2.3	41.6	17.1	18.1	0.94
+ KAR	−0.2	+6.0	+4.9	+3.9	**45.3**	**20.6**	**14.1**	**1.46**
Qwen3-VL-235B-A22B	53.2	30.6	26.1	27.3	14.5	–	–	–
+ CoT	−2.9	−0.5	−0.4	−0.4	11.4	1.4	29.7	0.05
+ Search-CoT	−6.8	+1.2	+2.8	+0.8	17.7	8.0	25.3	0.32
+ KAR	−9.0	+2.1	+4.7	+1.4	**19.7**	**9.2**	**18.1**	**0.50**

Table 3: Controlled augmentation on four representative models, showing component-level changes and preservation of Vanilla performance. Shaded rows give absolute Vanilla VIKR and SUCCESS; augmented rows give VIKR changes (gains, losses) and absolute SUCCESS. All entries except **Ratio** are in %. **Rep./Dmg.** are shares of Vanilla failures repaired/passes broken; **Ratio** is their quotient. In each outcome column, **bold** marks the best augmented setting per model. Qwen3-VL-235B-A22B is Instruct.

The layer transitions explain why these rankings differ. Across all eight retrieval comparisons, Identity and Knowledge rise while Visual falls; Reasoning also rises, but by a smaller margin. These opposing movements provide direct response-level evidence that VIKR captures distinct changes that a single aggregate score cannot expose.

Because Search-CoT matches KAR except for CultureBase, their comparison tests entity grounding under controlled retrieval conditions. KAR repairs more items and breaks fewer than generic web search on every detailed model. For Gemini-3.1-Pro, Search-CoT leaves the ratio nearly unchanged (1.55 to 1.57), whereas adding CultureBase raises it to 2.31, showing the value of entity grounding in this controlled comparison. Appendix C.2 details CultureBase construction and the entity-grounding stage. The matched GPT-5.6-Sol–Kimi-K2.5 pair tests augmentation from near-identical initial VIKR profiles across model provenance; their final ratios differ by only 0.08.

Inspection of pass-to-fail cases suggests *evidence crowding*, where retrieved context displaces image-grounded or task-relevant content. Qwen3-VL-235B-A22B-Instruct provides the clearest example, losing 9.0% of Visual coverage under KAR. Thus, MemeBench tests whether evidence fills diagnosed gaps without displacing coverage the model already had.

6 CONCLUSION

We introduced MemeBench, a bilingual diagnostic benchmark of 1,253 Chinese and English memes centered on ACG and adjacent online cultures. VIKR evaluates open-ended explanations through Visual, Identity, Knowledge, and Reasoning requirements, reserving SUCCESS for complete interpretations. Across 26 LVLMs, every model covers visible content more reliably than required knowledge; even the strongest retains a 22.6% Visual–Knowledge gap, and its 60.3% SUCCESS leaves clear headroom relative to blind human explanations (83.0% on the study subset). Layer-wise profiles also distinguish models with similar aggregate scores. Paired augmentation shows why this diagnosis matters: retrieval improves Identity and Knowledge but can reduce Visual coverage, while KAR achieves a better repair–damage balance than generic retrieval across all four models. MemeBench thus gives culturally grounded interpretation a concrete, rubric-level target for improvement.

AI USE STATEMENT

In this work, we used generative AI tools for four purposes. First, for generating the structured dataset annotations: Gemini-3.1-Pro converts each human-written reference explanation into the VIKR schema and its rubric (§3.1); these outputs undergo automated schema and entity-link validation and targeted manual review, and a blind human study independently validates the references (Appendix A.2). Second, for retrieval and discovery, namely searching for and identifying related literature. Third, for research execution: AI coding assistants helped implement and debug the inference, evaluation, and analysis code; the authors reviewed this code and verified its outputs. Fourth, to aid and polish the writing, including editing for readability and formatting text, tables, figures, and references. We have not used generative AI tools for research ideation, including proposing hypotheses or designing the methodology and experiments; to interpret results; to draft sections of the paper; or to formulate or prove mathematical claims. Separately from this assistance, LLMs are components of the evaluated methodology: Gemini-3.1-Pro and GPT-5.1 serve as judges calibrated against human labels (§3.3), and LVLMs are the systems under evaluation and the extractor and reasoner in KAR (§4). We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

MemeBench is built from publicly accessible memes on Bilibili, Reddit, and ImgFlip and is intended for diagnostic evaluation, not for training. Memes frequently contain stereotypes, political commentary, and culturally sensitive content; the benchmark excludes explicitly hateful content, while some retained memes carry cultural stereotypes that circulate in their source communities (Appendix D). Participants in the human validation study only explained or assessed explanations of public memes (Appendix A.2). Annotations, metadata, evaluation code, and CultureBase are released under CC BY-NC-SA 4.0, and the dataset repository includes a takedown protocol (Appendix D.2).

REPRODUCIBILITY STATEMENT

Section 3 and Appendix A describe data construction, the VIKR rubric, and the human validation protocol. Appendix B.1 reproduces the standard, CoT, and judge prompts together with decoding settings, and Appendix C formalizes KAR and CultureBase, including thresholds and query budgets. Appendix D.2 provides the dataset card. We release the annotations, evaluation code, CultureBase, inference configurations, frozen model responses, and judge outputs, so that every reported result can be recomputed from fixed artifacts.

REFERENCES

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv:2509.23661*, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv:2511.21631*, 2025.

Minh Duc Bui, Katharina Von Der Wense, and Anne Lauscher. Multi3hate: Multimodal, multilingual, and multicultural hate speech detection with vision–language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 9714–9731, 2025.

Davide Caffagni, Federico Cocchi, Nicholas Moratelli, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Wiki-llava: Hierarchical retrieval-augmented generation for multimodal llms. In *Proc. of CVPR*, pp. 1818–1826, 2024.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. BGE M3-Embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv:2402.03216*, 4(5), 2024.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proc. of EMNLP*, pp. 5558–5570, 2022.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. Can pre-trained vision and language models answer visual information-seeking questions? In *Proc. of EMNLP*, pp. 14948–14968, 2023.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, et al. Culturalbench: A robust, diverse and challenging benchmark for measuring lms' cultural knowledge through human-ai red-teaming. In *Proc. of ACL*, pp. 25663–25701, 2025.

Ido Cohen, Daniela Gottesman, Mor Geva, and Raja Giryes. Performance gap in entity knowledge extraction across modalities in vision language models. In *Proc. of ACL*, pp. 29095–29108, 2025.

Stefano De Giorgis, Ting-Chih Chen, and Filip Ilievski. M-quest–meme question-understanding evaluation on semantics and toxicity. *arXiv:2603.03315*, 2026.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv:2306.13394*, 2023.

Google DeepMind. Gemini 3.1 pro model card. Technical report, Google DeepMind, 2026. `https://deepmind.google/models/model-cards/gemini-3-1-pro/`.

Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. In *Proc. of ICCV*, pp. 12065–12075, 2023.

EunJeong Hwang and Vered Shwartz. Memecap: A dataset for captioning and interpreting memes. In *Proc. of EMNLP*, pp. 1433–1445, 2023.

Xilin Jiang, Qiaolin Wang, Junkai Wu, Xiaomin He, Zhongweiyang Xu, Yinghao Ma, Minshuo Piao, Kaiyi Yang, Xiuwen Zheng, Riki Shimizu, et al. Avmeme exam: A multimodal multilingual multicultural benchmark for llms' contextual and cultural knowledge and thinking. *arXiv preprint arXiv:2601.17645*, 2026.

Sai Kartheek Reddy Kasu, Mohammad Zia Ur Rehman, Shahid Shafi Dar, Rishi Bharat Junghare, Dhanvin Sanjay Namboodiri, and Nagendra Kumar. D-humor: Dark humor understanding via multimodal open-ended reasoning. In *Proc. of ICDM*, pp. 377–386. IEEE, 2025.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *Proc. of NeurIPS*, 33:2611–2624, 2020.

Yukyung Lee, Joonghoon Kim, Jaehee Kim, Hyowon Cho, Jaewook Kang, Pilsung Kang, and Najoung Kim. Checkeval: A reliable llm-as-a-judge framework for evaluating text generation using checklists. In *Proc. of EMNLP*, pp. 15782–15809, 2025.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Proc. of NeurIPS*, 33:9459–9474, 2020.

Chuanhao Li, Zhen Li, Chenchen Jing, Shuo Liu, Wenqi Shao, Yuwei Wu, Ping Luo, Yu Qiao, and Kaipeng Zhang. Searchlvlms: A plug-and-play framework for augmenting large vision-language models by searching up-to-date internet knowledge. *Proc. of NeurIPS*, 37:64582–64603, 2024.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *Proc. of ECCV*, pp. 216–233. Springer, 2024.

Junyu Lu, Bo Xu, Xiaokun Zhang, Hongbo Wang, Haohao Zhu, Dongyu Zhang, Liang Yang, and Hongfei Lin. Towards comprehensive detection of chinese harmful memes. *Proc. of NeurIPS*, 37: 13302–13320, 2024.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proc. of CVPR*, pp. 3195–3204, 2019.

Thomas Mensink, Jasper Uijlings, Lluis Castrejon, Arushi Goel, Felipe Cadar, Howard Zhou, Fei Sha, André Araujo, and Vittorio Ferrari. Encyclopedic vqa: Visual questions about detailed properties of fine-grained categories. In *Proc. of ICCV*, pp. 3113–3124, 2023.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki A Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew A Ayele, et al. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *Proc. of NeurIPS*, 37:78104–78146, 2024.

Abhilash Nandy, Yash Agarwal, Ashish Patwa, Millon Madhur Das, Aman Bansal, Ankit Raj, Pawan Goyal, and Niloy Ganguly. *** yesbut***: A high-quality annotated multimodal dataset for evaluating satire comprehension capability of vision-language models. In *Proc. of EMNLP*, pp. 16878–16895, 2024.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Van Steenkiste, Lisa Anne Hendricks, Karolina Stańczak, and Aishwarya Agrawal. Benchmarking vision language models for cultural understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5769–5790, 2024.

Khoi PN Nguyen, Terrence Li, Derek Lou Zhou, Gabriel Xiong, Pranav Balu, Nandhan Alahari, Alan Huang, Tanush Chauhan, Harshavardhan Bala, Emre Guzelordu, et al. Memeqa: Holistic evaluation for meme understanding. In *Proc. of ACL*, pp. 18926–18946, 2025.

Jeongsik Park, Khoi PN Nguyen, Terrence Li, Suyesh Shrestha, Megan Kim Vu, Jerry Yining Wang, and Vincent Ng. Memeintent: Benchmarking intent description generation for memes. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 631–643, 2024.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. Momenta: A multimodal framework for detecting harmful memes and their targets. In *Findings of EMNLP*, pp. 4439–4455, 2021.

Ali Ezzat Shahroor, Mohamed Bayan Kmainasi, Abul Hasnat, Dimitar Dimitrov, Giovanni Da San Martino, Preslav Nakov, and Firoj Alam. Memelens: Multilingual multitask vlms for memes. *arXiv preprint arXiv:2601.12539*, 2026.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. OpenAI GPT-5 system card. *arXiv:2601.03267*, 2026.

Bryan Chen Zhengyu Tan, Weihua Zheng, Zhengyuan Liu, Nancy Chen, Hwaran Lee, Kenny Tsu Wei Choo, and Roy Ka-Wei Lee. Blend-vis: Benchmarking multimodal cultural understanding in vision language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4647–4669, 2026.

Core Team, Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, Kainan Bao, Hao Tian, Hailin Zhang, Gang Wang, Dawei Zhu, Cici, Chenhong He, Bowen Ye, Bowen Shen, Zihan Zhang, Zihan Jiang, Zhixian Zheng, Zhichao Song, Zhenbo Luo, Yue Yu, Yudong Wang, Yuanyuan Tian, Yu Tu, Yihan Yan, Yi Huang, Xu Wang, Xinzhe Xu, Xingchen Song, Xing Zhang, Xing Yong, Xin Zhang, Xiangwei Deng, Wenyu Yang, Wenhan Ma, Weiwei Lv, Weiji Zhuang, Wei Liu, Sirui Deng, Shuo Liu, Shimao Chen, Shihua Yu, Shaohui Liu, Shande Wang, Rui Ma, Qiantong Wang, Peng Wang, Nuo Chen, Menghang Zhu, Kangyang Zhou, Kang Zhou, Kai Fang, Jun Shi, Jinhao Dong, Jiebao Xiao, Jiaming Xu, Huaqiu Liu, Hongshen Xu, Heng Qu, Haochen Zhao, Hanglong Lv, Guoan Wang, Duo Zhang, Dong Zhang, Di Zhang, Chong Ma, Chang Liu, Can Cai, and Bingquan Xia. Mimo-vl technical report, 2025. URL https://arxiv.org/abs/2506.03569.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Y Charles, HS Che, Cheng Chen, Guanduo Chen, et al. Kimi K2.5: Visual agentic intelligence. *arXiv:2602.02276*, 2026.

Mo Wang, Kaixuan Ren, Pratik Jalan, Ahmed Ashraf, Tuong Vy Vu, Rahul Seetharaman, Shah Nawaz, and Usman Naseem. From native memes to global moderation: Cross-cultural evaluation of vision–language models for hateful meme detection. In *Proceedings of the ACM Web Conference 2026*, pp. 9788–9799, 2026.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv:2508.18265*, 2025.

Jinming Wu, Zihao Deng, Wei Li, Yiding Liu, Bo You, Bo Li, Zejun Ma, and Ziwei Liu. Mmsearch-r1: Incentivizing lmms to search. *arXiv:2506.20670*, 2025.

Bo Xu, Chenyuan Wang, Xinyu Chen, Hongfei Lin, and Feng Xia. Metagpt: A large vision-language model for meme metaphor understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 16040–16048, 2026.

Zhijun Xu, Siyu Yuan, Yiqiao Zhang, Jingyu Sun, Tong Zheng, and Deqing Yang. Punmemecn: A benchmark to explore vision-language models' understanding of chinese pun memes. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 18705–18721, 2025.

Srishti Yadav, Lauren Tilton, Maria Antoniak, Taylor Arnold, Jiaang Li, Siddhesh Milind Pawar, Antonia Karamolegkou, Stella Frank, Zhaochong An, Negar Rostamzadeh, et al. Evaluation of cultural competence of vision-language models. *arXiv:2505.22793*, 2025.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv:2308.02490*, 2023.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proc. of CVPR*, pp. 9556–9567, 2024.

Chenhao Zhang, Xi Feng, Yuelin Bai, Xeron Du, Jinchang Hou, Kaixin Deng, Guangzeng Han, Qinrui Li, Bingli Wang, Jiaheng Liu, et al. Can mllms understand the deep implication behind chinese images? In *Proc. of ACL*, pp. 14369–14402, 2025.

Ruiyang Zhang, Qianguo Sun, Chao Song, Yiyan Qi, and Zhedong Zheng. Vsearcher: Long-horizon multimodal search agent via reinforcement learning. *arXiv:2603.02795*, 2026.

Yuming Zhao, Peiyi Zhang, and Oana Ignat. Beyond translation: Cross-cultural meme transcreation with vision-language models. *arXiv:2602.02510*, 2026.

Zhengyi Zhao, Shubo Zhang, Yuxi Zhang, Yanxi Zhao, Yifan Zhang, Zezhong Wang, Huimin Wang, Yutian Zhao, Bin Liang, Yefeng Zheng, et al. Memereacon: Probing contextual meme understanding in large vision-language models. In *Proc. of EMNLP*, pp. 3559–3582, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Proc. of NeurIPS*, 36:46595–46623, 2023.

Yang Zhong and Bhiman Kumar Baghel. Multimodal understanding of memes with fair explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2007–2017, 2024.

Shijia Zhou, Saif M Mohammad, Barbara Plank, and Diego Frassinelli. I came, i saw, i explained: Benchmarking multimodal llms on figurative meaning in memes. *arXiv:2603.23229*, 2026.

Domain	ZH	EN	Total	%
ACG	454	174	628	50.1
Cross-Domain	142	124	266	21.2
Daily Life	91	75	166	13.2
Movies & TV	47	78	125	10.0
Public Life	34	34	68	5.4
Total	**768**	**485**	**1,253**	

Table 4: Joint domain-by-language distribution of the retained evaluation set (n=1,253); these are the counts plotted in Figure 2.

Compatibility outcome	Count	%
Fully compatible	124	82.7
Partially compatible	23	15.3
Materially conflicting	3	2.0
Indeterminate	0	0.0
Total	150	100.0

Table 5: Independent reference validation on the stratified human-study subset. Adjudication results in 3 revised references and 0 removed items.

A Benchmark Construction and Human Validation

A.1 Dataset Selection and Composition

Quality control removes 247 of the 1,500 candidates (§3.1). Table 4 shows the domain and language distributions of the retained set.

A.2 Human Study and Calibration

Design, roles, and blinding. We use one 150-item study to validate reference independence and calibrate the LLM judges. The subset is sampled before annotation review and stratified jointly by language and analysis domain, preserving the benchmark proportions up to integer rounding. The protocol assigns four participants to role-separated responsibilities:

1. **P1 (primary annotator)** authored the original explanations and metadata for the full candidate pool.
2. **P2 and P3 (blind respondents)** each answer all 150 memes using the same explanation prompt as evaluated models. They see only the meme and prompt—not the existing reference, VIKR fields, rubric, model outputs, or one another's answer.
3. **P4 (independent validator)** did not create the original references and did not contribute blind explanations. P4 compares the two blind explanations with the existing reference and adjudicates substantive semantic conflicts.

P2 and P3 are bilingual Chinese–English speakers familiar with online meme culture; language proficiency and self-reported ACG familiarity are recorded before the study. All 300 blind answers are frozen before any participant is shown a reference or rubric.

Reference validation and adjudication. P4 compares each pair of blind explanations with the existing reference on three content-bearing criteria: core entity identity, required cultural background, and humor or communicative mechanism. Each item is assigned one of four outcomes: *fully compatible* (same interpretation and required facts), *partially compatible* (same core reading with complementary or nonessential differences), *materially conflicting* (incompatible entity, background, or mechanism), or *indeterminate* (insufficient evidence to select a reading). For every material conflict, P4 reviews the image and external cultural sources, records a rationale, and chooses among retaining the existing reference, revising the reference/rubric to admit a supported reading, or removing an item that lacks a defensible evaluation target. References and rubrics are frozen only after these decisions.

Comparison	V	I	K	R	SUCCESS
Human–human	0.79	0.91	0.85	0.81	0.84
Gemini–human	0.72	0.86	0.79	0.74	0.78
GPT–human	0.68	0.83	0.75	0.70	0.74
Intersection–human	0.76	0.88	0.82	0.77	0.81

Table 6: Pre-adjudication human–human agreement and judge–human agreement (Cohen's κ) on 150 frozen model responses.

	V	I	K	R	SUCCESS
Human	97.0	94.0	90.0	91.0	83.0
Gemini-3.1-Pro	92.3	74.6	69.7	76.7	60.3
Gap	+4.7	+19.4	+20.3	+14.3	+22.7

Table 7: Human reference performance on the study subset (%), scored on the same rubrics as model responses, against the leading model from Table 8. Human SUCCESS has a 95% interval of [77.7, 87.7].

Human labels and judge calibration. To calibrate automatic evaluation on model-like outputs, we additionally select one frozen candidate response per study item (150 responses total), balanced across model families and performance bands before human labeling. Model identity and both LLM decisions are hidden. P1 and P4 independently label every rubric criterion; their pre-adjudication agreement is reported below, and disagreements are resolved by discussion to produce the human gold labels. Neither rater evaluates a response they authored. We then compare Gemini-3.1-Pro, GPT-5.1, and their intersection separately with these gold labels using Cohen's κ.

Human reference performance. Because the blind explanations of §A.2 answer the same prompt given to the evaluated models, scoring them against the same VIKR rubrics under the same conjunctive SUCCESS definition places human performance on the benchmark's own scale. Table 7 reports it alongside the strongest model.

Human performance establishes substantial headroom on the benchmark while localizing it to the culturally grounded dimensions. The leading model is within 4.7 points of human Visual coverage, whereas the gaps on Identity and Knowledge are 19.4 and 20.3 points, respectively. This contrast reinforces the central finding that progress in description has outpaced progress in culturally grounded interpretation.

B EVALUATION AND DIAGNOSTIC ANALYSES

B.1 INFERENCE AND JUDGING PROTOCOL

The standard prompt used for all models is reproduced below.

Decoding. For stochastic decoding, we set temperature to 1.0 and top-p to 0.95; locally served open-weight models use greedy decoding. Each model produces one response per item. We release the complete inference configurations together with the frozen responses and judge outputs, allowing every reported result to be recomputed from fixed artifacts.

Standard inference prompt

You are an expert at analyzing multimodal content and memes.
Analyze this image step-by-step:

1. **Visual Observation**: Describe the visual elements and text (OCR) you see in detail.
2. **Identity Recognition**: Identify the specific entities, characters, origin works, or public figures involved. Use proper nouns (character names, work titles), not generic descriptions.
3. **Knowledge Retrieval**: Explain the cultural or factual background knowledge needed to understand this meme. What would an outsider need to know?
4. **Cultural Decoding**: Explain the underlying joke, irony, or metaphor. How do the visual elements and cultural context combine to create humor?

For Chinese memes, an equivalent Chinese prompt is used.

The CoT condition replaces the standard prompt with the reasoning-oriented prompt shown next.

CoT inference prompt

You are an expert at analyzing multimodal content and memes.
Look at this meme image carefully. Think through your analysis step-by-step, showing your reasoning process as you go.
Start by describing what you see, then work through identifying who or what is depicted, recall relevant cultural background, and finally explain how the humor works.
Think aloud as you analyze: show your reasoning, not just your conclusions. If you're uncertain about something, say so and explain your reasoning for your best guess.
After your step-by-step analysis, provide a brief summary of:

- What entities are involved (use proper nouns)
- What cultural knowledge is needed
- Why it's funny

Dual-judge rubric. Two LLM judges (Gemini-3.1-Pro and GPT-5.1) independently evaluate each candidate response against the ground-truth annotation. Each judge receives the full GT JSON (`visual`, `identity`, `knowledge`, `reasoning`, and `eval_checklist`) along with the candidate response; the rubric is stored in the `eval_checklist` field. The core instructions are reproduced verbatim below.

Dual-judge rubric prompt

Evaluate the Candidate Response across 4 dimensions defined in the GT's `eval_checklist`. Each dimension contains an array of checklist items — you must verify every item in the array.
Aggregation Rule. A dimension scores 1 ONLY if ALL checklist items in that dimension are satisfied. If any single item fails, the entire dimension scores 0.
Visual Accuracy. Verify each item in `eval_checklist.visual`. Does the candidate correctly describe key visual elements and OCR text?
Identity Recognition (STRICT). Verify each item in `eval_checklist.identity`. The candidate MUST use the exact entity names (e.g., "Conan Edogawa" not "a boy with glasses", "Thanos" not "a purple titan"). Generic descriptions always fail.
Knowledge Grounding. Verify each item in `eval_checklist.knowledge`. Does the candidate demonstrate relevant encyclopedic knowledge? Semantic equivalence is sufficient; exact wording is not required.
Reasoning Logic. Verify each item in `eval_checklist.reasoning`. Does the candidate explain how trigger $\rightarrow$ reference $\rightarrow$ humor mechanism works? Semantic equivalence is sufficient.

Each judge outputs a binary vector $[V, I, K, R]$ and a rubric score $V+I+K+R$. A response counts as SUCCESS only when all four dimensions pass under both judges (dual-judge intersection). Both judges are queried with their provider's default decoding parameters and score each response exactly once.

Model	Gemini-3.1-Pro only					GPT-5.1 only					Intersection				
	SUCC.	V	I	K	R	SUCC.	V	I	K	R	SUCC.	V	I	K	R
Commercial Models															
Gemini-3.1-Pro	**62.8**	**93.5**	**78.5**	**70.3**	**79.1**	**72.7**	**95.6**	**79.1**	**85.7**	**85.7**	**60.3**	**92.3**	**74.6**	**69.7**	**76.7**
Gemini-3-Flash	57.9	89.6	74.5	66.4	74.5	63.0	90.5	71.8	79.2	78.5	53.7	86.0	68.7	65.5	70.3
Qwen3.6-Plus	31.0	70.2	52.3	44.0	50.9	40.4	78.1	52.8	61.9	63.0	28.4	65.1	47.2	42.8	47.7
Qwen3.6-Flash	15.5	63.0	35.9	26.6	32.9	20.7	66.3	33.8	45.7	45.1	12.3	53.2	28.6	25.7	29.0
GPT-5.6-Sol	42.5	75.7	61.9	55.6	63.0	55.1	83.9	63.8	75.7	74.8	40.9	72.5	58.7	55.1	60.7
GPT-5.6-Terra	29.8	65.7	49.4	42.4	48.2	40.1	76.9	51.6	65.7	67.4	27.8	61.9	45.4	42.0	46.9
GPT-5.6-Luna	19.3	56.1	38.5	31.6	36.3	30.0	69.8	42.6	55.9	57.5	18.0	52.1	34.9	31.3	35.0
Claude-Opus-5	40.8	77.2	54.7	51.2	60.0	50.0	87.8	56.5	75.3	73.4	38.3	74.7	50.4	50.9	57.9
Claude-Sonnet-5	19.0	65.4	37.7	30.7	38.5	28.1	73.1	40.1	52.1	52.1	17.8	59.2	33.3	30.2	35.6
Open-Source Models															
Kimi-K2.5	41.2	74.8	64.5	55.5	62.3	52.6	83.3	65.8	75.7	73.7	37.9	71.1	59.2	54.5	58.7
MiMo-V2.5	15.7	56.1	34.1	27.9	33.1	22.3	62.0	35.8	48.8	44.5	13.7	47.2	30.5	26.9	28.3
MiMo-V2-Omni	16.4	57.9	34.2	27.5	32.7	21.9	61.1	35.1	49.2	45.6	14.3	48.1	29.7	26.4	28.8
Qwen3-VL-235B-A22B-Instruct	16.8	63.0	37.1	26.7	33.1	22.7	64.3	35.7	48.3	45.1	14.5	53.2	30.6	26.1	27.3
Qwen3-VL-235B-A22B-Thinking	13.4	57.7	32.0	23.8	28.6	20.1	65.9	32.6	44.8	41.4	12.0	50.2	27.5	22.9	24.7
Qwen3-VL-32B-Instruct	10.3	59.7	26.8	19.8	24.7	14.0	60.0	23.6	38.5	34.7	8.4	47.6	20.1	19.1	20.0
Qwen3-VL-32B-Thinking	8.9	53.3	25.1	18.9	23.3	14.2	59.7	24.5	36.7	36.0	7.7	44.5	20.5	17.9	19.7
Qwen3-VL-30B-A3B-Instruct	5.8	52.9	23.2	14.0	15.5	7.3	51.2	17.4	26.8	23.5	3.3	39.7	14.8	12.2	11.3
Qwen3-VL-30B-A3B-Thinking	7.0	52.3	24.3	15.8	19.2	11.7	57.1	22.8	33.0	31.0	5.8	41.6	18.3	14.7	15.8
Qwen3-VL-8B-Instruct	4.2	50.0	19.9	10.9	12.6	6.5	44.5	15.6	23.8	20.8	3.3	34.0	13.2	9.3	9.3
Qwen3-VL-8B-Thinking	4.1	43.2	18.6	10.7	13.1	7.7	47.2	17.2	24.2	24.2	3.4	32.1	14.0	10.0	9.9
InternVL3.5-38B	2.8	46.8	17.3	8.6	12.1	6.5	52.9	15.1	22.3	22.3	2.6	35.6	12.5	7.8	8.2
InternVL3.5-30B-A3B	1.7	37.3	14.2	6.7	9.0	3.0	44.4	10.9	17.2	17.1	1.1	26.4	8.8	5.6	6.2
InternVL3.5-20B-A4B	1.6	33.8	13.2	6.3	7.9	2.6	37.2	10.1	15.4	13.5	1.1	22.0	8.5	5.4	5.3
InternVL3.5-14B	1.5	34.8	14.7	7.3	9.3	4.1	43.9	11.7	17.6	15.8	1.1	26.0	9.3	6.0	6.1
InternVL3.5-8B	1.4	29.1	10.8	5.7	6.5	3.4	37.3	11.3	15.9	15.3	1.0	21.1	8.3	5.2	4.9
LLaVA-OV-8B-RL	0.6	24.7	13.5	4.5	4.1	3.5	35.2	12.5	15.9	14.3	0.4	18.8	9.0	4.2	3.2

Table 8: Vanilla leaderboard under each judge alone and under their intersection (%, overall split, n=1,253). Bold marks the best value within each judge block. The two single-judge rankings correlate at Spearman ρ=0.99, and each correlates with the intersection at $\rho \geq 0.99$; Gemini-3.1-Pro ranks first under all three evaluation views.

B.2 Robustness, Sensitivity, and Diagnostic Breakdowns

Both judges scored all 26 vanilla runs independently over the full benchmark (n=1,253 each) using the same rubric version. Table 8 reports the resulting single-judge and intersection leaderboards.

The model ordering is highly stable across evaluation views: the two single-judge leaderboards correlate at Spearman ρ=0.99, and each correlates with the intersection at $\rho \geq 0.99$. Gemini-3.1-Pro ranks first under all three views. Together with the human calibration in Table 6, this agreement supports using the dual-judge intersection as a stable comparative measure.

VIKR diagnostic breakdowns. **Response-state distribution.** Table 9 reports the full distribution over the 16 VIKR states for each model with a complete judge pass.

Domain and language strata. Tables 10 and 11 report vanilla SUCCESS by domain and by language, the latter both overall and restricted to the ACG domain.

Metric sensitivity. The primary SUCCESS metric asks whether an explanation covers all four VIKR dimensions. We define a complementary item-level sensitivity metric

$$\text{Success}_i^{3/4} = \mathbf{1}\{V_i + I_i + K_i + R_i \geq 3\}\,, \tag{4}$$

which tests whether conclusions depend on requiring complete four-dimensional coverage.

Table 12 reports this sensitivity analysis for the 26-model Vanilla leaderboard under the same dual-judge and full-denominator protocol as the primary results. The relaxed criterion raises absolute rates by 1.9–15.8 percentage points while preserving the model ordering closely (Spearman ρ=0.99), so the leaderboard's capability gradient does not depend on the strict conjunction.

C KAR and CultureBase

This appendix gives the formal definition of the entity-guided retrieval condition shown in Figure 3 and the construction of CultureBase.

Model	0000	0001	0010	0011	0100	0101	0110	0111	1000	1001	1010	1011	1100	1101	1110	1111
Gemini-3.1-Pro	3.4	0.4	0.2	0.5	1.0	0.2	0.7	1.4	12.1	3.6	1.3	3.9	3.2	6.5	1.4	60.3
Gemini-3-Flash	7.1	0.0	0.2	0.7	1.7	0.8	1.4	2.0	13.4	3.2	1.4	5.1	3.4	4.8	1.0	53.8
GPT-5.6-Sol	15.7	1.4	1.4	1.0	2.1	0.3	1.6	3.8	12.9	3.7	1.0	4.2	3.3	5.3	1.2	40.9
Claude-Opus-5	16.5	1.5	1.2	1.0	1.0	0.4	0.8	2.7	17.8	5.5	0.8	5.2	3.0	3.3	0.9	38.4
Kimi-K2.5	14.2	1.1	1.8	1.0	3.8	1.3	1.3	4.2	12.6	3.6	1.1	5.2	4.3	4.4	2.0	38.0
Qwen3.6-Plus	22.3	1.1	1.0	1.4	2.7	1.3	1.2	3.3	16.1	4.5	1.7	4.3	5.2	3.6	1.6	28.7
GPT-5.6-Terra	24.6	1.8	1.2	1.6	3.5	0.7	1.6	3.1	15.9	4.2	1.1	4.3	3.9	3.5	1.3	27.8
GPT-5.6-Luna	33.8	2.2	1.8	1.2	3.4	0.8	1.6	3.0	18.0	3.9	1.1	3.1	4.0	2.6	1.4	18.0
Claude-Sonnet-5	30.3	1.3	1.1	1.0	2.6	0.9	1.4	2.1	21.9	5.3	1.3	4.5	4.8	2.7	0.9	17.8
Qwen3-VL-235B-A22B-Instruct	35.4	1.5	1.1	0.5	4.4	0.6	1.7	1.4	21.8	3.0	2.6	3.3	4.6	2.5	1.0	14.6
MiMo-V2-Omni	40.5	1.4	1.2	1.0	3.5	1.3	1.2	1.8	17.6	3.5	1.6	3.5	3.8	2.1	1.8	14.3
MiMo-V2.5	39.4	1.9	1.4	0.9	4.0	0.8	1.8	2.5	18.0	2.3	1.6	3.9	4.3	2.3	1.0	13.8
Qwen3.6-Flash	35.5	2.2	1.2	1.2	3.0	0.4	0.8	2.3	21.1	3.6	2.0	4.5	5.8	2.6	1.4	12.3
Qwen3-VL-235B-A22B-Thinking	38.2	1.8	1.7	0.8	4.0	0.5	0.9	1.7	22.0	3.3	1.8	2.7	5.1	2.1	1.4	12.0
Qwen3-VL-32B-Instruct	43.4	1.4	1.5	1.0	2.8	0.6	0.8	0.7	24.4	2.7	1.9	3.5	3.7	1.8	1.4	8.4
Qwen3-VL-32B-Thinking	44.2	1.8	1.3	0.8	3.8	0.6	0.7	1.9	23.9	2.6	1.9	2.8	3.3	1.8	1.0	7.7
Qwen3-VL-30B-A3B-Thinking	48.5	1.0	1.8	0.7	4.2	0.2	0.8	1.1	23.5	2.6	1.4	2.2	3.1	2.2	0.9	5.8
Qwen3-VL-8B-Thinking	57.8	1.3	1.9	0.8	3.8	0.2	0.7	0.5	20.0	1.6	1.1	1.1	3.9	1.1	0.7	3.4
Qwen3-VL-8B-Instruct	58.7	1.0	0.9	0.4	2.9	0.3	0.9	0.8	22.2	1.3	1.0	1.3	3.3	1.0	0.8	3.3
Qwen3-VL-30B-A3B-Instruct	52.4	0.7	1.4	0.8	2.7	0.6	0.5	1.0	24.4	1.7	1.8	1.8	3.8	1.4	1.6	3.3
InternVL3.5-38B	57.4	1.2	0.8	0.3	3.1	0.5	0.6	0.5	24.2	1.7	1.1	0.8	3.5	0.6	1.0	2.6
InternVL3.5-30B-A3B	66.6	1.4	1.1	0.5	2.7	0.7	0.2	0.3	18.9	1.0	1.0	0.8	2.7	0.4	0.6	1.1
InternVL3.5-20B-A4B	70.9	0.7	1.0	0.3	3.7	0.2	0.6	0.5	15.9	1.1	0.7	0.8	1.6	0.5	0.3	1.1
InternVL3.5-14B	66.6	1.2	0.8	0.6	3.0	0.4	0.9	0.5	18.8	0.9	0.9	0.9	2.4	0.6	0.4	1.1
InternVL3.5-8B	72.1	0.8	1.5	0.3	3.2	0.2	0.6	0.2	14.8	1.1	0.4	0.6	2.0	0.6	0.5	1.0
LLaVA-OV-8B-RL	74.8	0.7	1.4	0.1	3.5	0.0	0.6	0.2	12.4	0.6	0.6	0.4	3.0	0.8	0.5	0.4

Table 9: Distribution over the 16 VIKR pass/fail states (% of scored responses, digits ordered V I K R). Across the 26 Vanilla runs, 16 of 16 states occur in the pooled responses, including states such as `0001` and `0010` that a deterministic V–I–K–R chain would not produce. Percentages exclude items that either judge failed to score.

C.1 Retrieval Pipeline

Stage 1, LVLM extraction. Given a meme image x_i, an extractor prompt asks the LVLM to produce

$$z_i = (o_i, h_i, \mathcal{H}_i, \mathcal{Q}_i^{\text{vlm}}), \tag{5}$$

where o_i is OCR text, h_i is a short visual cue phrase, $\mathcal{H}_i = \{(n_\ell, s_\ell, c_\ell)\}_{\ell=1}^{L_i}$ is a set of entity hypotheses (name, source, confidence), and $\mathcal{Q}_i^{\text{vlm}}$ is a set of model-suggested web queries.

Stage 2, CultureBase retrieval. To ground the extracted entity hypotheses, KAR builds a three-route query set from OCR text, visual cue phrase, and entity names:

$$\begin{aligned}\mathcal{R}_i = &\{o_i\} \cup \{h_i\} \\ &\cup \{n_\ell : (n_\ell, s_\ell, c_\ell) \in \mathcal{H}_i,\ c_\ell \in \{\text{medium}, \text{high}\}\}.\end{aligned} \tag{6}$$

Each CultureBase entity is scored by its best match across routes:

$$\text{score}_{ij} = \max_{r \in \mathcal{R}_i} g(r)^\top u_j, \tag{7}$$

where $g(\cdot)$ is the BGE-M3 embedding function and u_j is the entity embedding. The top-K entities above a similarity threshold are retained:

$$\begin{aligned}\widehat{\mathcal{C}}_i = \{e_j : j \in &\\ &\text{TopK}_{j:\,\text{score}_{ij} \geq \tau}(\text{score}_{ij})\}, \\ &K{=}5,\ \tau{=}0.5.\end{aligned} \tag{8}$$

Stage 3, entity-guided web search. Retrieved entities are converted into search queries via a template $\psi(e_j)$ (e.g., "`<name> <source> meme origin`"). The final query set combines model-suggested and entity-grounded queries under a shared budget $B{=}5$:

$$\mathcal{Q}_i^{\text{kar}} = \text{Budget}_B\left(\mathcal{Q}_i^{\text{vlm}} \cup \{\psi(e_j) : e_j \in \widehat{\mathcal{C}}_i\}\right). \tag{9}$$

Each query is executed with a web-search backend (Tavily), yielding a combined evidence set $\mathcal{S}_i$.

Model	ACG (628)	Cross-Dom. (266)	Daily (166)	Movies (125)	Public (68)
Gemini-3.1-Pro	52.1	65.8	69.3	69.6	75.0
Gemini-3-Flash	44.0	60.5	62.6	66.4	72.1
GPT-5.6-Sol	32.2	42.9	51.2	54.4	64.7
Kimi-K2.5	28.0	41.0	42.8	59.2	66.2
Claude-Opus-5	23.4	44.0	56.6	59.2	70.6
Qwen3.6-Plus	18.6	31.2	37.4	42.4	60.3
GPT-5.6-Terra	17.5	32.3	38.0	38.4	60.3
GPT-5.6-Luna	10.0	18.4	24.1	34.4	45.6
Claude-Sonnet-5	7.2	19.9	30.7	32.8	48.5
MiMo-V2-Omni	6.7	16.9	19.9	24.0	42.6
Qwen3-VL-235B-A22B-Instruct	6.4	15.0	23.5	25.6	45.6
Qwen3.6-Flash	5.6	13.9	18.1	21.6	36.8
MiMo-V2.5	4.9	18.8	21.1	24.0	38.2
Qwen3-VL-235B-A22B-Thinking	4.1	11.7	24.1	24.0	33.8
Qwen3-VL-32B-Thinking	3.0	7.5	12.1	15.2	26.5
Qwen3-VL-32B-Instruct	1.9	8.3	18.1	17.6	27.9
Qwen3-VL-30B-A3B-Thinking	1.1	6.8	10.2	12.0	23.5
Qwen3-VL-8B-Thinking	0.6	2.3	8.4	8.0	11.8
InternVL3.5-38B	0.5	2.3	7.2	4.8	8.8
InternVL3.5-30B-A3B	0.5	0.4	1.8	4.0	2.9
Qwen3-VL-30B-A3B-Instruct	0.3	5.3	4.8	6.4	13.2
Qwen3-VL-8B-Instruct	0.3	2.6	8.4	8.0	11.8
InternVL3.5-20B-A4B	0.3	0.4	1.8	3.2	5.9
InternVL3.5-14B	0.3	1.1	3.0	1.6	2.9
InternVL3.5-8B	0.2	1.1	3.0	0.8	4.4
LLaVA-OV-8B-RL	0.0	0.4	0.6	1.6	1.5

Table 10: Dual-judge Vanilla SUCCESS (%) by domain, with item counts in parentheses. Unscored responses count as failures. Per-domain V/I/K/R and 95% bootstrap intervals are released with the analysis code.

	All domains			Within ACG		
Model	ZH	EN	Δ	ZH	EN	Δ
Gemini-3.1-Pro	49.9	76.7	+26.8	47.1	64.9	+17.8
Gemini-3-Flash	41.8	72.6	+30.8	37.9	59.8	+21.9
Claude-Opus-5	28.3	54.2	+26.0	18.5	36.2	+17.7
GPT-5.6-Sol	33.1	53.4	+20.3	30.4	36.8	+6.4
Kimi-K2.5	32.4	46.6	+14.2	26.4	32.2	+5.8
GPT-5.6-Terra	22.3	36.5	+14.2	16.7	19.5	+2.8
Qwen3.6-Plus	25.5	33.0	+7.5	17.8	20.7	+2.9
Claude-Sonnet-5	12.2	26.6	+14.4	5.7	10.9	+5.2
GPT-5.6-Luna	15.5	22.1	+6.6	10.6	8.6	-2.0
Qwen3-VL-235B-A22B-Instruct	10.2	21.4	+11.3	4.4	11.5	+7.1
MiMo-V2.5	9.9	19.8	+9.9	3.7	8.1	+4.3
MiMo-V2-Omni	10.8	19.8	+9.0	6.0	8.6	+2.7
Qwen3-VL-235B-A22B-Thinking	9.6	15.7	+6.0	3.1	6.9	+3.8
Qwen3.6-Flash	10.6	15.1	+4.5	4.8	7.5	+2.6
Qwen3-VL-32B-Instruct	5.9	12.4	+6.5	1.1	4.0	+2.9
Qwen3-VL-32B-Thinking	5.5	11.1	+5.7	1.8	6.3	+4.6
Qwen3-VL-30B-A3B-Thinking	3.6	9.3	+5.6	0.2	3.5	+3.2
Qwen3-VL-8B-Instruct	1.9	5.4	+3.4	0.2	0.6	+0.3
InternVL3.5-38B	1.4	4.5	+3.1	0.0	1.7	+1.7
Qwen3-VL-8B-Thinking	2.7	4.3	+1.6	0.2	1.7	+1.5
Qwen3-VL-30B-A3B-Instruct	2.7	4.1	+1.4	0.0	1.1	+1.1
InternVL3.5-30B-A3B	0.7	1.9	+1.2	0.2	1.1	+0.9
InternVL3.5-8B	0.7	1.6	+1.0	0.0	0.6	+0.6
InternVL3.5-20B-A4B	1.2	1.0	-0.1	0.2	0.6	+0.3
InternVL3.5-14B	1.2	1.0	-0.1	0.2	0.6	+0.3
LLaVA-OV-8B-RL	0.1	0.8	+0.7	0.0	0.0	+0.0

Table 11: Dual-judge Vanilla SUCCESS (%) by language, over all domains and restricted to ACG. Unscored responses count as failures. The within-ACG comparison controls domain composition more closely; both views characterize the two sampled language-associated online-cultural ecosystems.

Stage 4, grounded reasoning. The LVLM receives the original image together with the retrieved evidence $\mathcal{S}_i$ and produces the final explanation. In the Vanilla and CoT conditions this evidence set is empty, and in Search-CoT it is populated without the CultureBase stage.

C.2 CULTUREBASE

CultureBase is a curated knowledge base of ∼5,000 cultural entities crawled from Moegirl (Chinese ACG encyclopedia) and KnowYourMeme (English meme database). Each record stores the entity's name, aliases, source work, character traits, visual description, and cultural usage patterns. Entities are embedded with bge-m3 (Chen et al., 2024) for multilingual cosine-similarity retrieval.

Model	4/4	3/4	Δ
Gemini-3.1-Pro	60.3	73.5	+13.2
Gemini-3-Flash	53.7	66.6	+12.9
GPT-5.6-Sol	40.9	55.5	+14.5
Claude-Opus-5	38.3	50.4	+12.0
Kimi-K2.5	37.9	53.7	+15.8
Qwen3.6-Plus	28.4	41.2	+12.8
GPT-5.6-Terra	27.8	40.0	+12.2
GPT-5.6-Luna	18.0	28.2	+10.1
Claude-Sonnet-5	17.8	28.1	+10.3
Qwen3-VL-235B-A22B-Instruct	14.5	22.8	+8.2
MiMo-V2-Omni	14.3	23.5	+9.2
MiMo-V2.5	13.7	23.5	+9.7
Qwen3.6-Flash	12.3	23.1	+10.8
Qwen3-VL-235B-A22B-Thinking	12.0	19.8	+7.8
Qwen3-VL-32B-Instruct	8.4	15.7	+7.3
Qwen3-VL-32B-Thinking	7.7	15.0	+7.3
Qwen3-VL-30B-A3B-Thinking	5.8	12.1	+6.3
Qwen3-VL-8B-Thinking	3.4	6.7	+3.4
Qwen3-VL-30B-A3B-Instruct	3.3	9.1	+5.8
Qwen3-VL-8B-Instruct	3.3	7.1	+3.8
InternVL3.5-38B	2.6	5.6	+3.0
InternVL3.5-30B-A3B	1.1	3.2	+2.1
InternVL3.5-20B-A4B	1.1	3.2	+2.1
InternVL3.5-14B	1.1	3.5	+2.4
InternVL3.5-8B	1.0	3.0	+1.9
LLaVA-OV-8B-RL	0.4	2.3	+1.9

Table 12: Metric sensitivity for the 26-model Vanilla leaderboard under the dual-judge intersection (%). **4/4** is the primary SUCCESS criterion; **3/4** accepts any response covering at least three VIKR dimensions. Δ is the absolute increase under the relaxed aggregation. The denominator remains all 1,253 items, so unscored responses count as failures.

These records provide entity anchors for the KAR pipeline (§4); the retrieved names, aliases, and source works are then used to guide web searches for cultural background.

D RESPONSIBLE USE AND DATASET CARD

D.1 SCOPE AND RESPONSIBLE USE

Research scope. MemeBench is ACG-centered by design. ACG communities are where meme conventions are densest and where the knowledge needed to decode an image is most clearly external to the image itself, which is the regime this benchmark is built to measure. This concentrated setting makes long-tail entities, source works, and community conventions central to the task, while the non-ACG half of the benchmark tests the same evaluation framework across adjacent online-cultural domains. We release a frozen, dated snapshot so that all reported results remain reproducible as online culture evolves.

Data governance. Meme images are sourced from publicly accessible platforms (Bilibili, Reddit, ImgFlip). We release annotations, metadata, evaluation code, and CultureBase under CC BY-NC-SA 4.0. Meme images are hosted alongside annotations on the dataset repository for reviewer access and reproducibility.

Ethical considerations. Memes frequently contain stereotypes, political commentary, and culturally sensitive content. MemeBench excludes explicitly hateful content, and some retained memes carry cultural stereotypes that circulate in their source communities. Required cultural knowledge is defined relative to the source community of each meme, which is what makes the target decodable and auditable. The benchmark is designed for diagnostic evaluation.

D.2 DATASET CARD

Dataset name. MemeBench v1

Size. 1,253 memes (768 Chinese, 485 English) after quality filtering from an initial pool of 1,500. All experiments use the full set; null responses are scored as failures, not excluded.

Task. Open-ended meme explanation.

Annotation schema. Four-layer VIKR: Visual, Identity, Knowledge, Reasoning. Per-dimension binary evaluation rubrics.

Sources. Chinese: Bilibili. English: Reddit, ImgFlip. All from publicly accessible content.

Annotation process. Human-written explanations (`checked_gt`) $\rightarrow$ model-assisted structured annotation $\rightarrow$ automated validation and targeted review $\rightarrow$ two blind explanations on a stratified subset $\rightarrow$ independent adjudication and reference/rubric freeze.

Intended use. Diagnostic evaluation of LVLM cultural knowledge. Not intended for training.

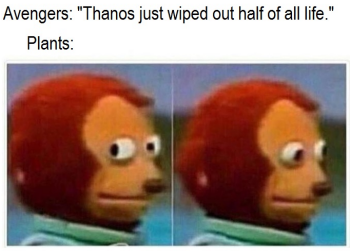
Scope. ACG-centered by design, with adjacent online-cultural domains represented in the same evaluation framework. The release is a dated snapshot, and automatic evaluation is calibrated against human labels and aggregated across two judges.

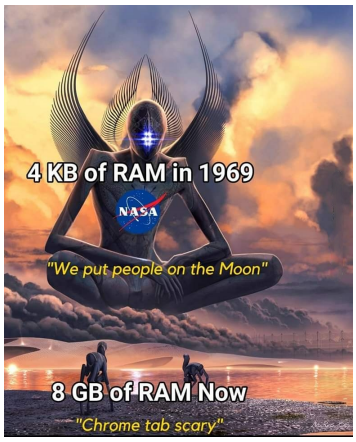
License. Annotations, metadata, evaluation code, and CultureBase are released under CC BY-NC-SA 4.0. Meme images are hosted on the dataset repository. A takedown protocol is included in the repository.

E QUALITATIVE VIKR CASES

Table 13 shows five representative cases in a compact table format. The top two rows are English memes, and the bottom three rows are Chinese memes. Each annotation cell expands the original VIKR fields from the MemeBench annotations.

Table 13: Raw qualitative VIKR table for two English and three Chinese meme cases. Each row presents the meme image and the original structured Visual, Identity, Knowledge, and Reasoning annotation fields. Annotation text is reproduced verbatim; the translation and romanization glosses under each Chinese image are added here for readability and are not part of the annotation.

Meme	Raw VIKR annotation
	{ "visual": { "description": "The image consists of black text at the top and a two-panel image below. The top text is a dialogue. The two-panel image shows a fuzzy puppet character with brown hair and a tan face. In the left panel, the puppet's large eyes are looking sideways towards the viewer. In the right panel, the puppet's eyes have shifted to look forward, avoiding eye contact. The puppet's expression is neutral but appears nervous or awkward due to the eye movement.", "ocr": "Avengers: Thanos just wiped out half of all life. / Plants:", "entities": ["e1: A fuzzy puppet character with brown hair and a tan face, looking sideways in the first panel and then looking away in the second panel, conveying an awkward or nervous expression."] }, "identity": { "entities": ["e1: name: Monkey Puppet / Pedro the Monkey; source: Okiku Naru Ko."] }, "knowledge": { "facts": ["In the Marvel Cinematic Universe, Thanos's snap is an event that eliminates half of all living beings in the universe.", "The films showing the aftermath of Thanos's snap focused on the disappearance of humans and animals, largely ignoring the effect on plant life.", "The Monkey Puppet reaction image, often called Awkward Look Monkey, is widely used in online culture to convey feelings of discomfort, guilt, or wanting to avoid attention."], "cultural_code": "Monkey Puppet meme (Awkward Look Monkey)" }, "reasoning": { "trigger": "The juxtaposition of the text stating half of all life being wiped out, followed by Plants: and the awkward monkey puppet looking away.", "reference": "The realization that plants are technically life and should have been affected by Thanos's snap, but the movies completely ignored this, so plants are humorously depicted as awkwardly staying quiet to avoid drawing attention to their survival.", "core_message": "The meme points out the logical flaw in Avengers: Infinity War where Thanos wiped out half of all life but plants seemingly survived unharmed, humorously personifying plants as awkwardly trying not to draw attention to this oversight.", "logic": "RV", "logic_detail": "Rule Violation (logical plot hole in the movie's premise) combined with Identity Collision (personifying plants as having awkward human reactions)." } }

Meme	Raw VIKR annotation
	{ "visual": { "description": "An illustration showing a giant, alien-like being sitting cross-legged in the sky, hovering over a landscape. Below, on the ground, two small figures look up at the giant being. Text is overlaid on the image in different positions. A circular blue logo with red and white elements is placed on the chest of the giant being.", "ocr": "4 KB of RAM in 1969 / NASA / We put people on the Moon / 8 GB of RAM Now / Chrome tab scary", "entities": ["e1: A giant, alien-like figure with glowing blue eyes and wings, sitting cross-legged in the sky.", "e2: Two small figures standing on the ground looking up.", "e3: A circular blue logo with a red chevron and white text NASA."] }, "identity": { "entities": ["e1: name: 4 KB of RAM in 1969; source: null.", "e2: name: 8 GB of RAM Now; source: null.", "e3: name: NASA; source: null."] }, "knowledge": { "facts": ["The Apollo Guidance Computer used during the Apollo missions had approximately 4 kilobytes of random-access memory (RAM).", "NASA's Apollo program successfully landed humans on the Moon for the first time in 1969 with the Apollo 11 mission.", "In contemporary consumer computing, 8 gigabytes of RAM is commonly regarded as an entry-level or baseline memory capacity for personal computers.", "Since the 2010s, Google Chrome has frequently been described by users and technology commentators as being resource-intensive, particularly in its use of system memory when multiple tabs or extensions are open."], "cultural_code": "Chrome RAM usage meme" }, "reasoning": { "trigger": "The contrast between the text 4 KB of RAM in 1969 associated with the giant, powerful figure and the text 8 GB of RAM Now associated with the small, fearful figures.", "reference": "The historical fact of the Apollo 11 moon landing versus the modern internet complaint about Google Chrome's memory consumption.", "core_message": "It is ironic that humanity accomplished the monumental feat of landing on the moon with only 4 KB of RAM, while modern computers with vastly more memory (8 GB) struggle to handle a single Google Chrome tab.", "logic": "IR", "logic_detail": null } }

Meme	Raw VIKR annotation


Translation: I am here to help you.
Romanized: Xiyouji = Journey to the West.
	{
	"visual": {
	"description": "A fluffy, anthropomorphic hamster standing on its hind legs in front of a small wooden log cabin. The hamster is pointing forward with its right paw and has its mouth open as if speaking or shouting. In its left paw, it holds a simple, poorly drawn green stick with an orange three-pronged fork top, resembling a trident. The text Wo Lai Zhu Ni is overlaid at the bottom.",
	"ocr": "Wo Lai Zhu Ni",
	"entities": ["e1: A standing hamster with an open mouth, pointing forward with one paw and holding a crudely drawn trident in the other."]
	},
	
	"identity": {
	"entities": ["e1: name: Huangfeng Dasheng (Yellow Wind Sage); source: Heishenhua: Wukong (Black Myth: Wukong)."]
	},
	
	"knowledge": {
	"facts": ["Heishenhua: Wukong is an action role-playing game developed by the Chinese game company Game Science and adapted from Xiyouji.", "Huangfeng Dasheng is a monster character in Heishenhua: Wukong. He resembles a rat-like demon and uses a trident as his weapon in battle.", "In Heishenhua: Wukong, Huangfeng Dasheng appears to assist the player during a fight against a giant beetle-like boss and shouts Wo Lai Zhu Ni.", "Wo Lai Zhu Ni is one of Huangfeng Dasheng's entrance lines in Heishenhua: Wukong and has been widely quoted among players because of its voice delivery and staging."],
	"cultural_code": "Wo Lai Zhu Ni (Black Myth: Wukong meme)"
	},
	
	"reasoning": {
	"trigger": "The combination of the cute hamster, the crudely drawn trident, and the exact quote Wo Lai Zhu Ni.",
	"reference": "The epic entrance of the powerful Huangfeng Dasheng from the game Black Myth: Wukong.",
	"core_message": "The humor comes from the absurd contrast between a majestic, powerful game character's iconic entrance line and a cute, harmless hamster holding a badly drawn trident acting out the scene.",
	"logic": "IC",
	"logic_detail": "Identity Collision (IC) mixed with Irony (IR). The meme maps a cute, harmless hamster holding a poorly drawn weapon to the powerful, fearsome rat demon from the game, creating a stark and humorous contrast between the grand, serious quote and the silly visual."
	}
	}


Translation: When you are on a bus but have nothing to grab onto.
Romanized: Huoyingrenzhe = Naruto; chakela = chakra.
	{
	"visual": {
	"description": "An image with text at the top and an illustration below. The text is in Chinese. The illustration shows a close-up of several pairs of feet wearing open-toed sandals and wrapped lower legs. The feet are glowing with a light blue aura against a dark, textured ground.",
	"ocr": "When you are on a bus but have nothing to grab onto:",
	"entities": ["e1: Feet wearing open-toed sandals and leg wraps, glowing with a blue aura."]
	},

Meme	Raw VIKR annotation

	"identity": {
	"entities": ["e1: name: Chakela Xifu; source: Huoyingrenzhe."]
	},
	
	"knowledge": {
	"facts": ["In the world setting of the Japanese manga and anime Huoyingrenzhe, chakela is a supernatural energy formed by combining spiritual energy and physical energy.", "Chakela Xifu is a technique in Huoyingrenzhe in which ninja concentrate chakela on the soles of their feet or palms, allowing them to attach to surfaces such as trees, walls, and ceilings.", "By precisely controlling the amount and flow of chakela, ninja in Huoyingrenzhe can perform movements ordinary people cannot, such as walking on water or walking upside down."],
	"cultural_code": null
	},
	
	"reasoning": {
	"trigger": "The image of glowing feet from Huoyingrenzhe combined with the text about having nothing to hold onto on a bus.",
	"reference": "The everyday struggle of keeping balance on a moving bus without a handrail, solved by the fictional ninja ability to magically stick feet to the floor.",
	"core_message": "It playfully suggests that one needs ninja-level magical abilities to maintain balance on a moving bus without holding a handrail.",
	"logic": "IR",
	"logic_detail": "Applying a highly specialized, magical combat skill (sticking to walls with chakela) to a mundane, relatable daily annoyance (balancing on a bus)."
	}
	}


Translation: Zhuge Liang after returning from Zhou Yu's funeral.
Romanized: Sanguo Yanyi = Romance of the Three Kingdoms; Diaoxiao = staged weeping.
	{
	"visual": {
	"description": "The image shows a man wearing a traditional Chinese ancient-style black hat with vertical stripes and a long black beard, grafted onto a body wearing a modern black suit. He is holding a golden statuette (resembling an Oscar) in his raised right hand. He has a subtle smirk. The background features white curtains and large golden silhouettes of the same statuette. There is black Chinese text at the top of the image.",
	"ocr": "Zhuge Liang after returning from Zhou Yu's funeral: / ab / bo / Xiaohongshu / Xiaohongshu ID: 507813434",
	"entities": ["e1: A man's face with a long black beard and a traditional ancient Chinese black hat with vertical stripes, superimposed on a body wearing a modern black suit, holding a golden statuette.", "e2: A golden statuette of a standing figure, held by the man, with larger versions in the background."]
	},
	
	"identity": {
	"entities": ["e1: name: Zhuge Liang; source: Sanguo Yanyi.", "e2: name: Oscar statuette (Academy Award); source: Academy Awards."]
	},

Meme	Raw VIKR annotation

	"knowledge": {
	"facts": ["In Sanguo Yanyi, Zhuge Liang and Zhou Yu have a complex relationship of rivalry and cooperation, and the two are portrayed as rivals who admire each other while competing against each other.", "Sanguo Yanyi includes a scene in which Zhuge Liang goes to mourn Zhou Yu, weeping bitterly and speaking with such sincerity that those present are deeply moved.", "Zhuge Liang Diaoxiao is often used in Chinese online culture and derivative works to discuss acting skills, disguised emotion, and similar topics, becoming a scene label with extended meaning.", "The trophy of the Academy Awards is commonly called the Oscar statuette, a gilded little golden figure and an important symbol of the award."],
	"cultural_code": "Zhuge Liang Diaoxiao / Oscar Yingdi"
	},
	
	"reasoning": {
	"trigger": "The text Zhuge Liang after returning from Zhou Yu's funeral combined with the image of him holding an Oscar statuette.",
	"reference": "The cultural reinterpretation that Zhuge Liang's intense weeping at Zhou Yu's funeral was actually just top-tier acting rather than genuine sorrow.",
	"core_message": "Zhuge Liang's grief at Zhou Yu's funeral was so fake yet convincingly performed that he deserves an Oscar for Best Actor.",
	"logic": "IC",
	"logic_detail": "Identity Collision: Treating historical/fictional political maneuvering as a modern acting performance worthy of a film award."
	}
	}