ReMMD: Realistic Multilingual Multi-Image Agentic Verification for Multimodal Misinformation Detection

Chenhao Dang[1,2] **Dantong Zhu**[4] **Jun Yang**[5] **Conghui He**[2] **Weijia Li**[2,3†]
[1]Shanghai Jiaotong University [2]Shanghai Artificial Intelligence Laboratory
[3]Tsinghua University [4]Central South University
[5]China Electronics Technology Group Corporation 15th Research Institute
dangchenhao@pjlab.org.cn zhudantong@csu.edu.cn yangjun15s@cetc.com.cn
heconghui@pjlab.org.cn liweijia@sz.tsinghua.edu.cn
[†]Corresponding author.

Abstract

Multimodal misinformation detection is increasingly important because viral posts now combine long multilingual narratives, several images, mixed provenance, and subtle text–image framing errors. Existing benchmarks and methods remain poorly matched to this setting: they usually isolate short captions, single images, binary labels, or one manipulation source, while agentic verification remains costly under realistic evidence search. We present ReMMD, a realistic multilingual multi-image agentic verification framework for multimodal misinformation detection. ReMMD includes ReMMDBench, a real-world multimodal misinformation detection benchmark with 500 samples, 2,756 images, five monolingual languages, two cross-lingual settings, three text-length tiers, multi-image posts, five-way veracity labels, eight distortion labels, evidence provenance, and rationales. It also includes ReMMD-Agent, a persistent-memory verifier that decomposes posts into atomic points, builds a reusable evidence set, and predicts structured L1/L2/L3 outputs. Across proprietary systems, open LVLMs, MMD-Agent, and T^2-Agent, ReMMD-Agent obtains the best five-way veracity performance, with 41.80% accuracy and 39.12% macro-F1 using GPT-5.2, while reducing cost by 17.5% relative to MMD-Agent and 79.9% relative to T^2-Agent. The project is available at https://dang-ai.github.io/ReMMD.

1 Introduction

Real-world multimodal rumors and misinformation pervade news and social media, where text, images, screenshots, and generated or edited media jointly amplify false claims, threatening social trust, political processes, public figures, crisis response, and national security (Vosoughi et al., 2018; Lv et al., 2025). This risk has shaped a progression of evaluation resources, from large-scale

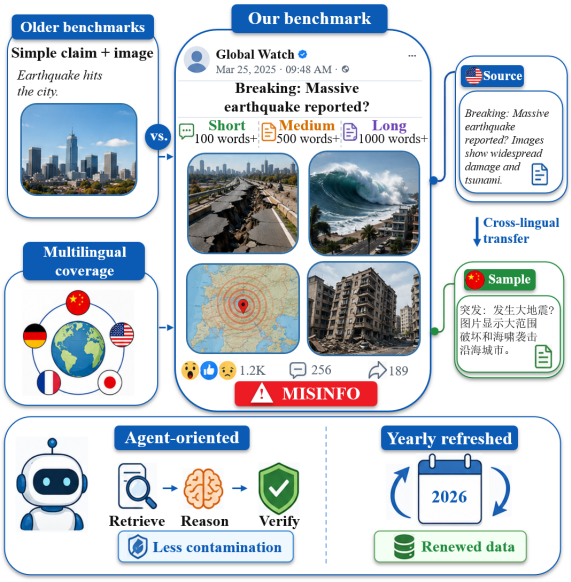


Figure 1: ReMMDBench is an agent-oriented benchmark for realistic multimodal misinformation detection, with yearly refreshed data to reduce contamination.

textual verification in FEVER (Thorne et al., 2018) to image–text mismatch evaluation in NewsCLIPpings (Luo et al., 2021), and more recently to mixed-source and dynamically refreshed settings in MMFakeBench (Liu et al., 2025) and VeriTaS (Rothermel et al., 2026). Methods follow suit, with VLM/LVLM-based multimodal misinformation detection (MMD) systems improving visual perception and retrieval (Wang et al., 2025; Liu et al., 2025), while T^2-Agent (Cui et al., 2026) extends tool-augmented verification through search-based reasoning.

Nevertheless, benchmark-driven progress still leaves a gap between existing MMD evaluations and operational deployment. Existing evaluations often simplify verification to isolated claims, single image–text pairs, coarse verdicts, or one manipulation source, while real deployments must handle long multilingual posts with many images, mixed visual provenance, partial truth, evolving evidence, and textual, visual, and cross-modal distortion at-

Benchmark	Dyn.	Agent	Lang.	Cross	Length	Images	Labels	Ratl.	Auto
NewsCLIPpings (Luo et al., 2021)	✗	✗	✗	✗	short	✗	2	part.	✗
MuMiN (Nielsen and McConville, 2022)	✗	✗	✓	part.	short	✗	2	✗	✗
COSMOS (Aneja et al., 2021)	✗	✗	✗	✗	short	✗	2	✗	✗
VERITE (Papadopoulos et al., 2024)	✗	✗	✗	✗	short	✗	3	part.	part.
AVeriTeC (Schlichtkrull et al., 2023)	part.	✗	✗	✗	med.	✗	4	✓	part.
MFC-Bench (Wang et al., 2025)	✗	✗	✗	✗	med.	part.	3	✓	part.
GroundMM (Yang et al., 2025a)	✗	✗	✗	✗	med.	part.	2+g	✓	part.
MMFakeBench (Liu et al., 2025)	✗	✓	✗	✗	short	✗	2+3	✓	part.
XFacta (Xiao et al., 2025)	part.	part.	✗	✗	med.	part.	2	✓	part.
VeriTaS (Rothermel et al., 2026)	✓	part.	part.	part.	med.	part.	4	✓	✓
M4FC (Geng et al., 2025)	part.	✗	✓	✓	med.	part.	multi	✓	part.
ReMMDBench (Ours)	✓	✓	✓	✓	all	✓	5+8	✓	✓

Table 1: Comparison with representative fact-checking and multimodal misinformation benchmarks. **Dyn.** indicates dynamically refreshed or update-aware data; **Agent** indicates whether designed for agentic verification; **Lang.** indicates multilingual coverage; **Cross** indicates cross-lingual evaluation; **Length** summarizes typical text length (short, med., long, or all tiers); **Images** indicates multi-image or image-aware samples; **Labels** gives the number or type of veracity and distortion labels, where "2+g" includes grounding supervision and "5+8" denotes five-way veracity plus eight distortion labels; **Ratl.** indicates rationale or explanation supervision; **Auto** indicates automated or dynamically assisted construction/evaluation. ✓, ✗, and part.denote yes, no, and partial support.

tribution (Giachanou et al., 2020; Müller-Budack et al., 2020). Addressing this gap requires systems that can decompose central claims, select evidential images, track provenance, reuse evidence, and attribute distortions across modalities, and recent tool-augmented MMD and production generalist agents make such agentic verification increasingly feasible (Cui et al., 2026; Shlomov et al., 2026).

This gap motivates the Realistic Multimodal Misinformation Detection Benchmark (ReMMDBench), a real-world, agent-oriented benchmark for evaluating systems under operational verification conditions, as illustrated in Figure 1 and positioned against prior resources in Table 1. ReMMDBench consists of single-text, multi-image samples spanning three length tiers, five monolingual languages, and two cross-lingual transfer settings, as summarized in Table 2. Since real-world fact-checking often requires graded verdicts rather than binary truth labels (Wang, 2017; Lee et al., 2023), each sample is annotated with a five-class L1 veracity label, L2 distortion labels selected from eight categories, and an L3 natural-language rationale.

Bridging this deployment gap under the same operational setting also calls for an agent that manages evidence before judgment. We introduce ReMMD-Agent, a real-world MMD verifier that decomposes posts into atomic claims and image bindings, retrieves web, image, and social evidence, and incrementally updates a persistent memory bank with reusable evidence. A structured judge then predicts L1 veracity, L2 distortion labels, and an L3 rationale from this evidence state. Like an experienced fact-checker, this workflow supports multidimensional judgment while remaining cost-efficient.

Together, ReMMDBench and ReMMD-Agent form ReMMD, a realistic multilingual multi-image agentic verification framework for MMD. On ReMMDBench, we evaluate two general-purpose closed-source agents and three open-source MMD agents, including ReMMD-Agent, using base models drawn from three backbone families and five total model sizes. ReMMD-Agent with GPT-5.2 sets the current best ReMMDBench result and reduces GPT-5.2 cost by 17.5% relative to MMD-Agent and 79.9% relative to T^2-Agent. Qwen3.5-9B also outperforms the closed-source agents on ReMMDBench and remains competitive on MMFakeBench.

Main Contributions of This Work

- **ReMMDBench.** A real-world, agent-oriented MMD benchmark with multilingual, multi-image, varied-length samples, five-way veracity, eight distortion labels, and rationales.
- **ReMMD-Agent.** A real-world MMD verifier that organizes claims, image bindings, and persistent evidence memory for low-cost structured judgment.
- **Evaluation.** Broad comparisons across commercial agents and open agentic baselines validate ReMMDBench and show ReMMD-Agent's high performance, analyzing language, length, distortion, transfer, and cost.

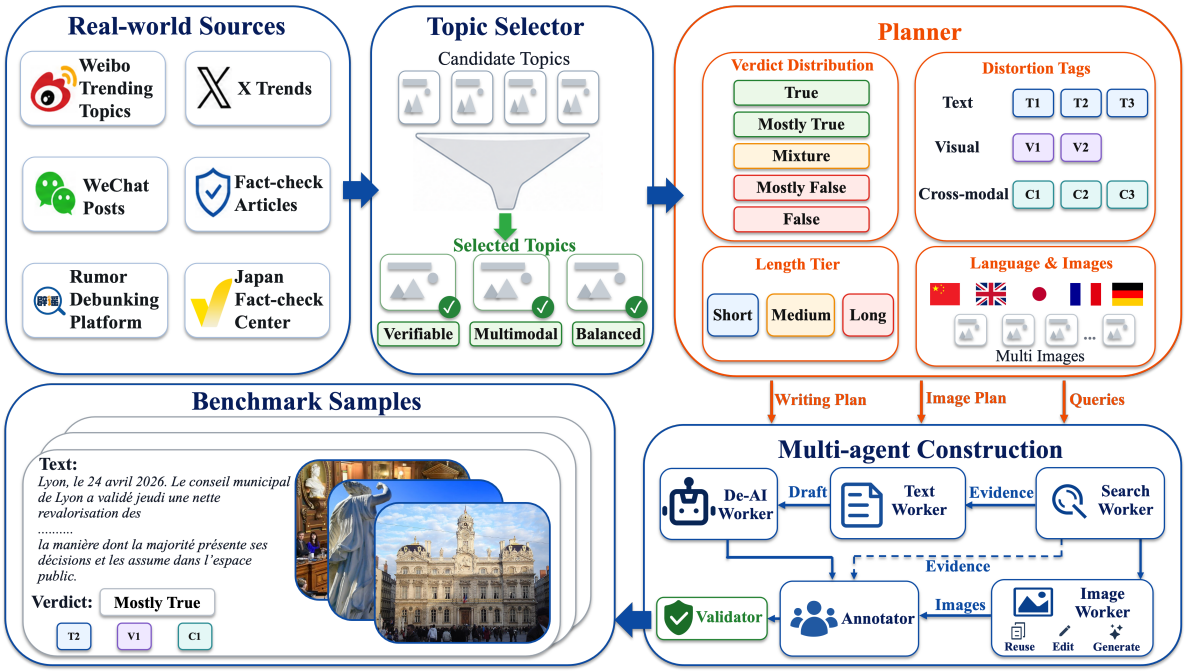


Figure 2: ReMMDBench turns real-world misinformation topics into controlled multilingual multi-image samples by planning language, length, visual provenance, and label conditions, then validating each instance against evidence, distortion annotations, text–image consistency, and image provenance before inclusion.

2 Related Work

Datasets and benchmarks. Textual verification benchmarks established evidence-grounded claim checking, from FEVER (Thorne et al., 2018) and MultiFC (Hanselowski et al., 2019) to news and social-media datasets such as CHEF (Hu et al., 2022), MDFEND (Nan et al., 2021), and FakeNewsNet (Shu et al., 2020). MM-COVID (Li et al., 2020) and MuMiN (Nielsen and McConville, 2022) broaden multilingual and social-media coverage, while multimodal resources study image repurposing, out-of-context use, unimodal bias, web evidence, localization, intent, and attribution (Sabir et al., 2018; Aneja et al., 2021; Luo et al., 2021; Papadopoulos et al., 2024; Yao et al., 2023; Schlichtkrull et al., 2023; Shao et al., 2023; Da et al., 2021; Guo et al., 2025). Recent benchmarks further cover AI-generated or edited news, mixed-source distortion, grounding, realism, multilinguality, and dynamic evaluation (Huang et al., 2024; Xu et al., 2024; Chen and Shu, 2024; Li et al., 2026; Xu et al., 2025; Liu et al., 2025; Yang et al., 2025a; Zhu et al., 2025; Xiao et al., 2025; Geng et al., 2025; Rothermel et al., 2026). Yet a large gap remains between these benchmark settings and real-world MMD deployment: existing data often isolates short captions, single images, limited languages, coarse labels, static evidence, or one manipulation source, whereas operational verification must handle long multilingual posts, many images, mixed provenance, graded veracity, and fine-grained text–visual distortion under changing evidence conditions.

Statistic	Count
Samples	500
Images	2,756
Avg. images	5.51
Monolingual	423
Cross-lingual	77
AI-generated image	237
AI-edited image	246
Any AI-touched image	384
Short / medium / long	173 / 159 / 168
English / Chinese	111 / 112
German / Japanese / French	67 / 68 / 65

Table 2: Core statistics of ReMMDBench. Length tiers are balanced, and most samples contain multiple images and at least one AI-touched visual item.

LVLMs and agentic verification. Large vision-language models are natural multimodal verifiers, but recent studies show that perception alone remains vulnerable to grounding errors, stale or adversarial evidence, and temporal contamination (Wang et al., 2025; Yang et al., 2025b; Chen et al., 2025; Xu et al., 2026; Xu and Yan, 2025). Agentic verification improves robustness by decomposing claims, asking targeted questions, and invoking retrieval or visual tools (Beigi et al., 2025; Liu et al., 2025; Cui et al., 2026); T^2-Agent fur-

ther expands tool-augmented reasoning with Monte Carlo Tree Search, but the added search increases cost. Thus, the core method challenge is not only perception or tool access, but evidence management at deployment scale. Realistic MMD requires long-horizon memory over many claims, images, sources, timestamps, provenance cues, and contradictions to support accurate classification, while high-concurrency applications also demand strict control of repeated retrieval and inference cost.

3 ReMMDBench

3.1 Benchmark Design and Construction

ReMMDBench is designed around controlled realism: the goal is not only to increase scale, but to make the verification pressures of real multimodal misinformation observable and measurable. Each sample is instantiated from a topic, language condition, text-length tier, image budget, visual provenance, and target label configuration. These factors jointly expose conditions that often co-occur in social media, including long narrative text, multiple evidential or decorative images, reused real media, and AI-generated or edited visuals.

Table 2 reports the main statistics. The benchmark is deliberately image-dense: only one sample has a single image, while 168 samples contain ten or eleven images. Text length is balanced across short, medium, and long tiers, with average length rising from 168.1 to 2,316.4 units and average image count from 2.35 to 10.05. Short posts therefore test whether a model avoids over-reading compact claims, whereas long posts require tracking entities, dates, quotations, and image order across a larger visual context. Additional distributional analysis is provided in Appendix G.

3.2 Annotation Schema

Each sample receives a hierarchical annotation consisting of an L1 veracity label, L2 distortion labels, and an L3 natural-language rationale. The L1 labels are ordered by severity: True, Mostly True, Mixture, Mostly False, and False. They distinguish fully supported claims, minor local errors, mixed true and false evidence, dominantly false conclusions with residual true details, and unsupported or contradicted core propositions. The middle labels are important because they encode whether an error changes the main conclusion or merely qualifies it. The distribution is near-balanced, and the average number of L2 labels increases from 0.00 for True to 4.41 for False. Detailed label frequencies are provided in Appendix B, and boundary notes are provided in Appendix C.

Topic category	Count	Percent
Entertainment and sports	107	21.4
International conflict	85	17.0
Public safety and disaster	70	14.0
Science, technology, and AI	66	13.2
Politics and public affairs	58	11.6
Society and culture	49	9.8
Finance and markets	35	7.0
Health and medicine	22	4.4
Other	8	1.6

Table 3: Topic distribution of ReMMDBench. The benchmark avoids concentrating on a single rumor domain, which helps distinguish general verification ability from topic memorization.

The L2 taxonomy separates textual, visual, and cross-modal distortions. Textual labels cover fabrication, distortion of a real factual basis, and misleading context; visual labels cover synthetic content and editing; cross-modal labels cover semantic, contextual, and pragmatic inconsistency. The labels are multi-label because a single post may distort text, manipulate images, and bind authentic visuals to the wrong context. We deliberately separate visual provenance from evidential force: an AI-touched image does not by itself make a post false, and an authentic image can still be misleading when attached to the wrong event.

3.3 Quality Control

ReMMDBench uses three-stage quality control. First, each candidate must contain a verifiable claim, at least one relevant image, and a gold label supported by evidence. Validators then reject cases driven by private context, satire, or normative disagreement, and audit whether the L1 verdict follows from the central claim and whether each L2 label is grounded in a concrete textual, visual, or cross-modal mismatch. A final pass aligns rationales with labels and verifies that image provenance is not conflated with veracity.

We keep the taxonomy compact to preserve reliability. Finer categories can separate manipulation subtypes, but they make annotation less stable and evaluation harder to interpret. The eight labels retain the distinctions most useful for fact-checking, namely whether the error lies in text, visual evidence, or their relation, while exact-match L2 remains a strict diagnosis metric.

Table 3 reports the topic mix. The benchmark is not concentrated in a single rumor domain: en-

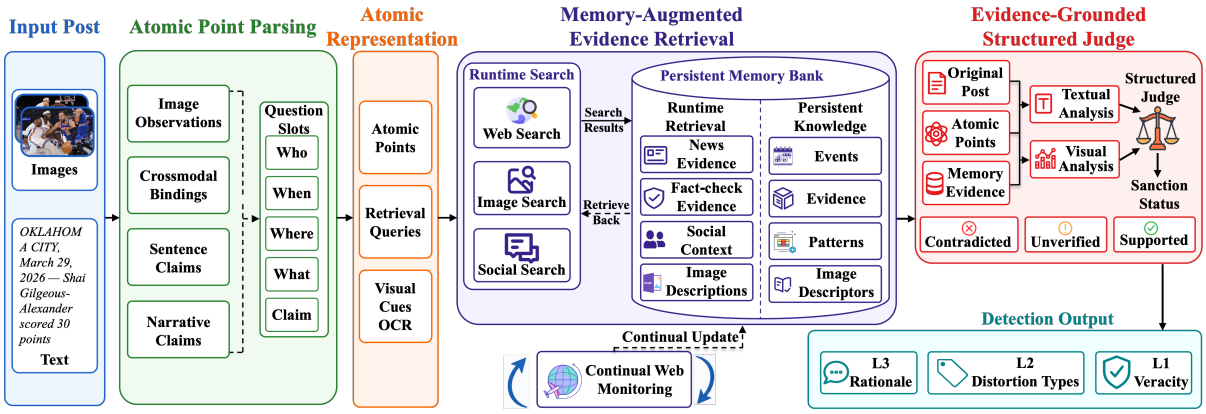


Figure 3: ReMMD-Agent verifies a multimodal post by first decomposing text and images into atomic claims, observations, and cross-modal bindings, then retrieving and reusing evidence in a persistent memory bank before a structured judge integrates textual, visual, and provenance cues to produce the L1 veracity label, L2 distortion diagnosis, and L3 rationale.

tertainment, conflict, public safety, science, politics, health, and finance require different evidence sources, testing both perceptual grounding and domain-sensitive retrieval. This breadth limits shortcut learning, since the same L1 verdict can arise from different combinations of textual distortion, image reuse, and cross-modal mismatch.

4 ReMMD-Agent

Figure 3 gives the computation graph of ReMMD-Agent. Given a post $s = (x, I)$, where x is the textual content and $I = \{i_m\}_{m=1}^{M}$ is the image set, the agent predicts (y, z, r): a five-way veracity label y, an eight-dimensional distortion vector $z \in \{0, 1\}^8$, and a concise rationale r. Rather than judging the full text–image bundle directly, ReMMD-Agent first compresses it into verifiable atomic units, retrieves and reuses evidence around those units, and performs judgment over an explicit evidence state. This design reduces information noise from long real-world narratives, such as background exposition, repeated assertions, and weakly relevant details, before expensive retrieval and makes the final decision traceable to claims, images, and sources.

4.1 Atomic Representation

The first stage maps the post into atomic points

$$A = \{a_j = (c_j, q_j, v_j, \tau_j)\}_{j=1}^{n},$$

where c_j is a minimal claim or visual observation, q_j is a retrieval query, v_j contains visual cues, and τ_j denotes the point type. Atomic points cover image observations, cross-modal bindings, sentence-level claims, and narrative-level claims. They retain only information that can affect verification, such as visible scenes, OCR, entities, overlays, and the way an image is used to support a specific event, location, person, time, number, or conclusion. This representation separates checkable content from long-form narrative noise and localizes retrieval to entities, dates, quantities, attributions, and image–text bindings. Near-duplicates are merged, and at most twelve points are retained per sample, reducing redundant searches and giving the judge a compact evidence state while preserving the central evidence needed for classification.

4.2 Memory-Augmented Retrieval

The second stage retrieves evidence for the atomic points and stores it in a sample-level memory bank $M_s = \{e_k\}_{k=1}^{K}$. Each record stores a type, source descriptor, optional timestamp, reliability note, and links to the points it may support or contradict. For each a_j, the system uses q_j and v_j to call web, image, and social search tools, yielding

$$R_j = \mathrm{TopK}_{e \in M_s} \, \mathrm{sim}(\phi(a_j), \phi(e)),$$

where $\phi(\cdot)$ is the text or multimodal representation for matching. The memory bank stores news reports, fact-checks, social context, image descriptions, event records, and reference descriptors. Crucially, M_s persists across atomic points: evidence retrieved for one textual claim can later support an image binding, resolve a temporal mismatch, or contradict a narrative-level conclusion. The memory bank therefore functions as an auditable evidence state rather than a transient prompt context, enabling reuse of high-value evidence and reducing repeated retrieval over overlapping claims.

Agent	Backbone	L1 Veracity				L2 Distortion			
		Accuracy	Precision	Recall	Macro-F1	Precision	Recall	Macro-F1	Exact Match
Manus	proprietary	33.00	33.58	33.00	33.13	43.69	42.32	42.75	7.60
ChatGPT	proprietary	30.20	32.65	30.15	28.24	41.71	47.56	43.63	3.00
MMD-Agent	GPT-5.2	26.40	23.77	26.38	23.42	42.24	48.83	41.98	2.00
	Gemma4-31B	25.60	26.83	25.60	25.86	40.94	42.11	40.82	5.20
	Qwen3.6-27B	25.40	26.30	25.44	25.36	41.59	38.61	37.94	4.80
	Qwen3.5-9B	26.20	26.16	26.24	26.00	40.10	40.31	38.15	5.80
	Qwen3.5-4B	25.00	25.42	25.04	24.66	43.08	32.70	34.45	7.20
T^2-Agent	GPT-5.2	28.20	29.91	28.14	26.00	42.15	47.15	42.68	2.60
	Gemma4-31B	24.40	24.97	24.41	24.60	41.16	38.73	38.73	7.60
	Qwen3.6-27B	26.00	24.07	26.13	23.40	41.63	27.09	27.74	4.20
	Qwen3.5-9B	25.60	23.98	25.67	23.31	38.83	27.14	28.58	3.20
	Qwen3.5-4B	21.20	20.41	21.20	19.92	37.95	27.26	28.50	1.80
ReMMD-Agent	GPT-5.2	**41.80**	**43.98**	**41.71**	**39.12**	44.31	47.01	45.15	5.00
	Gemma4-31B	33.60	34.21	33.59	33.76	43.67	41.31	42.27	7.80
	Qwen3.6-27B	30.40	32.85	30.36	30.07	**44.73**	39.20	41.10	**10.80**
	Qwen3.5-9B	37.20	39.72	37.11	37.18	44.58	**50.88**	**46.97**	10.00
	Qwen3.5-4B	29.20	32.07	29.12	28.58	42.59	42.89	41.94	6.00

Table 4: Full ReMMDBench results on 500 samples. Values are percentages. The grey rows mark general-purpose assistant baselines; each agent block merges five backbone rows.

4.3 Structured Evidence Judgment

The final stage receives (x, I, A, M_s) and auxiliary textual and visual analyses. The judge first assigns each atomic point a state $\sigma_j \in \{\text{supported}, \text{contradicted}, \text{unverified}\}$, then infers y from the evidence pattern over central claims and cross-modal bindings. This step is not a vote over atomic points: a contradicted peripheral number may shift a post from True to Mostly True, whereas a contradicted event attribution can determine the verdict even if many surface details are real. The L2 vector is assigned after L1 so that visual provenance is not treated as a shortcut for falsehood. The judge considers textual evidence, visual provenance, and image–text relations separately before selecting any distortion label, then outputs the veracity label y, distortion diagnosis z, and rationale r.

4.4 Implementation Details

Queries are issued in the original language. Cross-lingual samples additionally use an English or Chinese bridge query. Visual retrieval uses captions, OCR, named entities, and reverse-search descriptions when available. Auxiliary textual analysis flags fabrication, distortion, and misleading context, while visual analysis focuses on synthetic content, editing traces, source mismatch, and cross-modal consistency. These analyses are treated as soft evidence rather than hard rules. The resulting decomposition-and-memory pipeline keeps retrieval targeted, limits repeated tool use, and produces a compact evidence state that supports cost-efficient and auditable verification.

5 Experiments

5.1 Experimental Setup

We evaluate Manus (Manus, 2026), ChatGPT (OpenAI, 2026), MMD-Agent (Liu et al., 2025), T^2-Agent (Cui et al., 2026), and ReMMD-Agent on the full 500-sample ReMMDBench split. Manus uses Manus 1.6, and ChatGPT is evaluated through the OpenAI web interface. Model-backed agents use GPT-5.2, Gemma4-31B, Qwen3.6-27B, Qwen3.5-9B, and Qwen3.5-4B, with non-GPT open backbones deployed locally on H200 GPUs. All web retrieval uses the Serper API (Serper, 2026), and model-backed agents share the same evidence retriever and image-processing pipeline. We adapt MMD-Agent and T^2-Agent to multi-image samples while preserving their original label-selection rules, as detailed in Appendix D. Each system predicts the L1 five-way veracity label and L2 eight-label distortion vector. We report exact L1 accuracy and macro metrics, L2 macro metrics, and L2 exact match. GPT-5.2 cost is measured on the full benchmark under the same endpoint and tool-call budgets for all model-backed agents.

5.2 Overall Results

Table 4 shows that ReMMDBench remains difficult for all evaluated systems, which confirms that five-way, multi-image verification is substantially harder than detecting local suspicious cues. General-purpose assistants are competitive on some L2 metrics, but their weaker L1 results

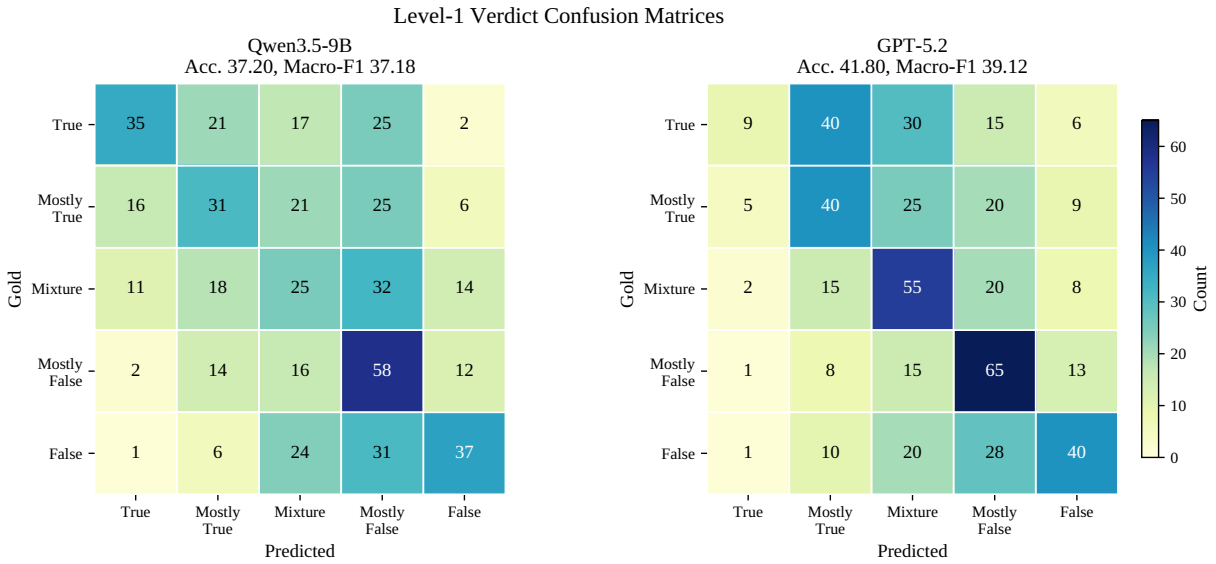


Figure 4: Count heatmaps for ReMMD-Agent L1 predictions. Both backbones recover substantial diagonal mass, but errors concentrate around adjacent middle labels where partial evidence must be calibrated rather than merely detected.

Variant	L1 Macro-F1	L2 Macro-F1
Full ReMMD-Agent	39.12	45.15
No memory bank	35.84	41.77
No atomic parsing	34.96	39.88
No visual auxiliary analysis	37.21	40.46
Single-pass LVLM judge	31.42	37.09

Table 5: Ablation on the GPT-5.2 ReMMD-Agent. Atomic parsing and memory reuse both contribute, and visual auxiliary analysis is especially important for L2 labels.

indicate that graded veracity depends on how evidence changes the central claim. ReMMD-Agent improves this calibration across backbone families. GPT-5.2 gives the best L1 performance, and Qwen3.5-9B gives the strongest L2 macro-F1 among comparable open-backbone runs.

The comparison with MMD-Agent and T^2-Agent shows that additional search is not sufficient unless evidence is organized around the right claims. MMD-Agent remains useful for distortion-oriented comparison, but struggles with partial-truth labels in long multi-image narratives. T^2-Agent explores more reasoning paths, yet the extra search does not consistently improve veracity. Figure 4 further shows that the remaining errors concentrate among neighboring middle labels, where models must judge the centrality of contradicted evidence rather than merely detect a suspicious cue. Appendix G reports additional GPT-backed confusion matrices.

Table 5 identifies the mechanism behind these gains. Atomic parsing reduces long-form information noise and supplies checkable units for retrieval and diagnosis, while memory supports provenance aggregation and cross-image evidence reuse. Removing either component hurts both L1 and L2, and the single-pass judge is weakest. Visual auxiliary analysis is especially important for L2 because visual edits and cross-modal mismatches can be diagnostic before they determine the final veracity label.

System	Total	Per sample	vs. ReMMD
MMD-Agent	$126.32	$0.2526	1.21×
T^2-Agent	$517.91	$1.0358	4.97×
ReMMD-Agent	$104.16	$0.2083	1.00×

Table 6: Full-benchmark GPT-5.2 cost audit. ReMMD-Agent reduces per-sample cost by 17.5% relative to MMD-Agent and 79.9% relative to T^2-Agent.

5.3 Fine-Grained Behavior

Figure 5 tests whether the gains persist under the main pressures built into ReMMDBench. Across text-length tiers, ReMMD-Agent is more stable than the baselines. This is most informative for long posts, where additional context also introduces more entities, dates, quotations, and image references. Atomic parsing turns this noisy context into checkable units, and memory reuse reduces retrieval of superficially related but temporally or geographically mismatched events. Language slices show that multilingual verification is not only a translation problem, since regional source availability, entity grounding, and cross-script nam-

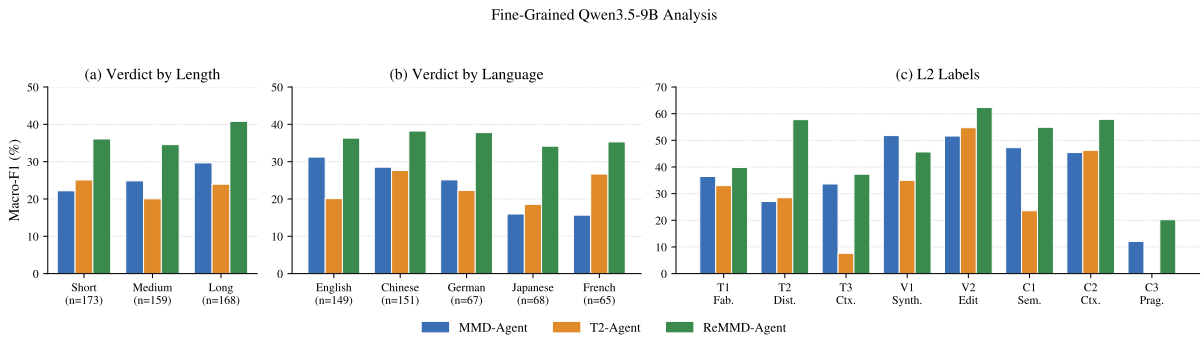


Figure 5: Fine-grained Qwen3.5-9B analysis across text length, language, and L2 labels.

System	Accuracy	Fake F1
MMD-Agent	0.592	0.673
T^2-Agent	0.639	0.715
ReMMD-Agent	**0.824**	**0.871**

Table 7: Transfer to the official MMFakeBench test split with Qwen3.5-9B and the same retrieval backend.

ing variation matter, especially for Japanese and French. Label slices show the clearest gains on distortion, editing, and cross-modal inconsistency labels, where evidence alignment is essential. The weaker advantage on synthetic visual content and pragmatic inconsistency suggests that low-level forensics and discourse-level support remain complementary challenges. Full numerical slices are reported in Appendix F.

5.4 Cost and Transfer

Table 6 shows that the gains do not come from greater spending. ReMMD-Agent is cheaper than MMD-Agent because evidence is reused across atomic points, and it is far cheaper than T^2-Agent because it avoids repeated expansion of tool-augmented reasoning paths. This matters for dynamic benchmarks and real deployments, where the same verifier may need to run repeatedly under high concurrency. Table 7 further shows that the policy is not specific to ReMMDBench. With the same Qwen3.5-9B backbone and retrieval backend, ReMMD-Agent transfers strongly to the large binary MMFakeBench test set. Appendix E gives the transfer setting in detail.

6 Discussion

The main lesson is that realistic MMD is an evidence-selection problem. A post may use real evidence to support a wrong conclusion, so fine-grained labels are necessary. Retrieval helps only when each source is tied to the claim or image it verifies. Visual authenticity alone is not enough, because real images can be misused and synthetic images do not automatically falsify the text.

The Qwen results support this view. Under the same ReMMD-Agent pipeline, Qwen3.5-9B outperforms Qwen3.6-27B on several metrics. This is not a general reversal of model scale. After retrieval and memory provide evidence, the backbone mainly needs to follow the schema, calibrate partial evidence, and avoid over-interpreting uncertainty. Larger models can be less stable on adjacent partial-truth labels.

The benchmark also clarifies future directions. Rationales should identify the claim, evidence, and image-text relation. Multilingual cases require local entity and source grounding, not only translation. Future systems should improve source-aware memory, temporal retrieval, multilingual entity linking, and metrics that separately evaluate visual edits, verdicts, and misleading mechanisms.

7 Conclusion

We introduced ReMMDBench and ReMMD-Agent to study multimodal misinformation under realistic verification conditions. ReMMDBench moves evaluation beyond short binary image-text cases by combining multilingual posts, multiple images, graded veracity, distortion labels, and rationales. ReMMD-Agent shows that this setting is best handled as evidence management. It decomposes posts into checkable units, reuses retrieved evidence through memory, and judges veracity and distortion from an explicit evidence state. Experiments show that this design improves calibration, supports fine-grained distortion diagnosis, reduces retrieval cost, and transfers beyond ReMMDBench. Taken together, ReMMD reframes realistic multimodal misinformation detection around evidence selection, grounding, and explanation across modalities.

Limitations

ReMMDBench contains 500 carefully constructed samples, which enables controlled analysis but is smaller than web-scale social-media corpora. The benchmark covers five languages and two cross-lingual directions, but it does not cover all linguistic communities, regional rumor ecosystems, or low-resource languages. Some generated or edited images may reflect the tools used during construction, so future releases should include a wider range of generators, editors, and real-world media sources. ReMMD-Agent also depends on external retrieval, and its results may vary with search-engine coverage, regional access, and temporal changes in online evidence. Finally, L3 rationales are audited qualitatively in this version; automatic rationale faithfulness evaluation remains future work.

Ethical Considerations

The benchmark is intended to support research on detecting and explaining multimodal misinformation, not to facilitate its creation or dissemination. Samples are constructed and annotated for evaluation, and potentially sensitive topics are handled through evidence-based labeling rather than persuasive rewriting. Because misinformation datasets may contain harmful claims, benchmark items should not be republished as standalone social content or used to amplify false narratives. Any release should include clear usage terms, provenance documentation, and contextual warnings for misleading material. ReMMD-Agent should be treated as decision support for trained fact-checkers or researchers, not as an automatic moderation authority or a substitute for human judgment. The released benchmark and code will be distributed for research use only under the license and usage terms specified in the release repositories.

References

Shivangi Aneja, Chris Bregler, and Matthias Nießner. 2021. Cosmos: Catching out-of-context misinformation with self-supervised learning. *arXiv preprint arXiv:2101.06278*.

Alimohammad Beigi, Bohan Jiang, Dawei Li, Zhen Tan, Pouya Shaeri, Tharindu Kumarage, Amrita Bhattacharjee, and Huan Liu. 2025. Can llms improve multimodal fact-checking by asking relevant questions? In *2025 IEEE International Conference on Big Data (BigData)*, pages 2732–2741. IEEE.

Canyu Chen and Kai Shu. 2024. Can llm-generated misinformation be detected? In *International Conference on Learning Representations*, volume 2024, pages 34687–34726.

Sanxing Chen, Yukun Huang, and Bhuwan Dhingra. 2025. Real-time factuality assessment from adversarial feedback. In *Proceedings of ACL*.

Xing Cui, Yueying Zou, Zekun Li, Peipei Li, Xinyuan Xu, Xuannan Liu, and Huaibo Huang. 2026. T2agent: A tool-augmented multimodal misinformation detection agent with monte carlo tree search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 175–183.

Jeff Da, Maxwell Forbes, Rowan Zellers, Anthony Zheng, Jena D. Hwang, Antoine Bosselut, and Yejin Choi. 2021. Edited media understanding frames: Reasoning about the intent and implications of visual misinformation. In *Proceedings of ACL*.

Jiahui Geng, Jonathan Tonglet, and Iryna Gurevych. 2025. M4fc: A multimodal, multilingual, multicultural, multitask real-world fact-checking dataset. *arXiv preprint arXiv:2510.23508*.

Anastasia Giachanou, Guobiao Zhang, and Paolo Rosso. 2020. Multimodal multi-image fake news detection. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 647–654.

Hao Guo, Zihan Ma, Zhi Zeng, Minnan Luo, Weixin Zeng, Jiuyang Tang, and Xiang Zhao. 2025. Each fake news is fake in its own way: An attribution multi-granularity benchmark for multimodal fake news detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pages 228–236.

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A richly annotated corpus for different tasks in automated fact-checking. In *Proceedings of the 23rd conference on computational natural language learning (CoNLL)*, pages 493–503.

Xuming Hu, Zhijiang Guo, GuanYu Wu, Aiwei Liu, Lijie Wen, and Philip S Yu. 2022. Chef: A pilot chinese dataset for evidence-based fact-checking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3362–3376.

Runsheng Huang, Liam Dugan, Yue Yang, and Chris Callison-Burch. 2024. Miragenews: Multimodal realistic ai-generated news detection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16436–16448.

Sian Lee, Aiping Xiong, Haeseung Seo, and Dongwon Lee. 2023. "fact-checking" fact checkers: A data-driven approach. *Harvard Kennedy School (HKS) Misinformation Review*.

Fanxiao Li, Jiaying Wu, Tingchao Fu, Yunyun Dong, Bingbing Song, and Wei Zhou. 2026. Drifting away from truth: Genai-driven news diversity challenges lvlm-based misinformation detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 588–596.

Yichuan Li, Bohan Jiang, Kai Shu, and Huan Liu. 2020. Mm-covid: A multilingual and multimodal data repository for combating covid-19 disinformation. *arXiv preprint arXiv:2011.04088*.

Xuannan Liu, Zekun Li, Pei Li, Huaibo Huang, Shuhan Xia, Xing Cui, Linzhi Huang, Weihong Deng, and Zhaofeng He. 2025. Mmfakebench: A mixed-source multimodal misinformation detection benchmark for lvlms. In *International Conference on Learning Representations*, volume 2025, pages 86327–86352.

Grace Luo, Trevor Darrell, and Anna Rohrbach. 2021. Newsclippings: Automatic generation of out-of-context multimodal media. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6801–6817.

Jinna Lv, Yuan Gao, Li Li, Lei Shi, and Siyu Li. 2025. Multi-modal fake news detection: A comprehensive survey on deep learning technology, advances, and challenges. *Journal of King Saud University Computer and Information Sciences*, 37(9):306.

Manus. 2026. Manus. https://manus.im/. Version 1.6. Accessed: 2026-05-26.

Eric Müller-Budack, Jonas Theiner, Sebastian Diering, Maximilian Idahl, and Ralph Ewerth. 2020. Multimodal analytics for real-world news using measures of cross-modal entity consistency. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, ICMR '20, pages 16–25, New York, NY, USA. Association for Computing Machinery.

Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. 2021. Mdfend: Multi-domain fake news detection. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 3343–3347.

Dan S Nielsen and Ryan McConville. 2022. Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 3141–3153.

OpenAI. 2026. ChatGPT. https://chatgpt.com/. Accessed: 2026-05-26.

Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis C Petrantonakis. 2024. Verite: a robust benchmark for multimodal misinformation detection accounting for unimodal bias. *International Journal of Multimedia Information Retrieval*, 13(1):4.

Mark Rothermel, Marcus Kornmann, Marcus Rohrbach, and Anna Rohrbach. 2026. Veritas: The first dynamic benchmark for multimodal automated fact-checking. *arXiv preprint arXiv:2601.08611*.

Ekraam Sabir, Wael AbdAlmageed, Yue Wu, and Prem Natarajan. 2018. Deep multimodal image-repurposing detection. *arXiv preprint arXiv:1808.06686*.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. Averitec: A dataset for real-world claim verification with evidence from the web. *Advances in Neural Information Processing Systems*, 36:65128–65167.

Serper. 2026. Serper API. https://serper.dev/. Accessed: 2026-05-26.

Rui Shao, Tianxing Wu, and Ziwei Liu. 2023. Detecting and grounding multi-modal media manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6904–6913.

Segev Shlomov, Alon Oved, Sami Marreed, Ido Levy, Offer Akrabi, Avi Yaeli, Łukasz Strąk, Elizabeth Koumpan, Yinon Goldshtein, Eilam Shapira, Nir Mashkif, and Asaf Adi. 2026. From benchmarks to business impact: Deploying IBM generalist agent in enterprise production. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 40423–40431.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Shengkang Wang, Hongzhan Lin, Ziyang Luo, Zhen Ye, Guang Chen, and Jing Ma. 2025. Mfc-bench: Benchmarking multimodal fact-checking with large vision-language models. In *ICLR Workshop on Data Problems for Foundation Models*.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Yuzhuo Xiao, Zeyu Han, Yuhan Wang, and Huaizu Jiang. 2025. Xfacta: Contemporary, real-world dataset and evaluation for multimodal misinformation detection with multimodal llms. *arXiv preprint arXiv:2508.09999*.

Cheng Xu, Changhong Jin, Yingjie Niu, Nan Yan, Yuke Mei, Shuhao Guan, Liming Chen, and M.-Tahar Kechadi. 2026. Livefact: A dynamic, time-aware benchmark for llm-driven fake news detection. *arXiv preprint arXiv:2604.04815*.

Cheng Xu and Nan Yan. 2025. Triplefact: Defending data contamination in the evaluation of llm-driven fake news detection. In *Proceedings of ACL*.

Qingzheng Xu, Huiqiang Chen, Heming Du, Hu Zhang, Szymon Łukasik, Tianqing Zhu, and Xin Yu. 2024. M3a: A multimodal misinformation dataset for media authenticity analysis. *Computer Vision and Image Understanding*, 249:104205.

Qingzheng Xu, Heming Du, Szymon Łukasik, Tianqing Zhu, Sen Wang, and Xin Yu. 2025. MDAM3: A misinformation detection and analysis framework for multitype multimodal media. In *Proceedings of the ACM Web Conference (WWW)*.

Bingjian Yang, Danni Xu, Kaipeng Niu, Wenxuan Liu, Zheng Wang, and Mohan Kankanhalli. 2025a. A new dataset and benchmark for grounding multimodal misinformation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 12571–12577.

Shuo Yang, Yuqin Dai, Guoqing Wang, Xinran Zheng, Jinfeng Xu, Jinze Li, Zhenzhe Ying, Weiqiang Wang, and Edith CH Ngai. 2025b. Realfactbench: A benchmark for evaluating large language models in real-world fact-checking. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 13435–13441.

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2023. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2733–2743.

Ye Zhu, Yunan Wang, and Zitong Yu. 2025. Multimodal fake news detection: Mfnd dataset and shallow-deep multitask learning. In *Proceedings of IJCAI*.

A Benchmark Examples

Figure 6 shows two non-sensitive ReMMDBench samples with multilingual text, multiple images, hierarchical labels, rationales, and evidence-centered analysis.

B Additional Benchmark Statistics

Analysis. Tables 8 and 9 separate the two annotation views that define ReMMDBench. The five L1 classes are close to balanced, which makes macro-F1 meaningful and prevents systems from succeeding by favoring a dominant verdict. The average number of L2 labels rises monotonically from True to False, showing that severe misinformation usually accumulates multiple forms of distortion rather than a single isolated cue. The distortion table further shows that visual editing, textual distortion, and cross-modal inconsistency all occur frequently. This distribution supports evaluating L1 veracity and L2 diagnosis together, since the same final verdict can arise from different combinations of textual, visual, and pragmatic evidence.

C Label Boundary Notes

T2 Distortion is assigned when a textual claim has a real factual basis but changes scope, intensity, attribution, relation, or conclusion. T3 Misleading Context is preferred when the content itself may be real but is placed in the wrong time, location, source, or event frame. V1 and V2 can co-occur when a real image is edited by inserting generated content. C1 concerns factual semantic conflict between text and image, C2 concerns context-frame mismatch, and C3 concerns stance, sentiment, or evidential-support mismatch.

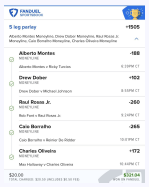
D Agent Adaptation Details

MMD-Agent and T^2-Agent were originally designed for single-image multimodal misinformation inputs. To evaluate them on ReMMDBench, we keep their original prompt structure and label-selection rules, and change only the input packing and evidence interface needed for multi-image samples. The full post text is passed unchanged, while images are serialized as ordered image slots with captions, OCR, named entities, and available provenance descriptors. Retrieval calls use the same query budget, image descriptors, Serper backend, and image-processing pipeline as ReMMD-Agent, and no ReMMDBench gold labels or rationales are exposed during inference. This adaptation makes the baselines executable on multi-image posts without giving them additional supervision or changing their decision taxonomy.

English short sample, ID 106. Entertainment, Sports & Celebrity. L1 verdict: **Mixture**. L2 labels: T3 Misleading Context, V2 Visual Editing, C2 Contextual Inconsistency, and C1 Semantic Inconsistency.

Ringside post

Edited card



Betting slip

Post excerpt. UFC326 was still trending on X after the Las Vegas card, with Oliveira celebration posts, betting screenshots, and recap threads circulating together. A graphic showing winners was treated as a same-night results board from the venue, and the post framed the ringside image, betting slip, and recap graphic as mutually reinforcing evidence.

Gold rationale. The sample combines real event-night posts and betting material with an edited fight-card image. The misleading step is not a single false caption, but the way the visual bundle makes the edited card appear to document confirmed outcomes and validated picks.

Analysis. The sample is compact, yet the evidence state includes an event photograph, a betting slip, and a stylized fight-card graphic. The veracity decision depends on whether the images document the same event in the way the text implies, making the case a test of cross-modal binding rather than surface suspicion.

Chinese medium sample, ID 026. Society. L1 verdict: **Mixture**. L2 labels: T3 Misleading Context, V2 Visual Editing, and C2 Contextual Inconsistency.

Edited poster

Staged visual

Ski-area context

Scenic context

Post excerpt. 近日，“四川绵阳九皇山景区月薪5万+招聘185cm以上、腹肌陪滑官/陪滑岗位”相关信息在社交平台走红。样本围绕招聘海报、景区回应和网友咨询展开，强调“月薪5万+”“长期招募”“岗位已招满”等传播要点，并将海报截图与滑雪场现场图组合为同一事件的视觉证据。

Gold rationale. 样本基于真实热点和媒体报道，但弱化了海报底部限制条款，并将阶段性或条件性收入叙述为长期稳定月薪。配图还混入归属不明或生成痕迹较强的现场图，使受众更容易形成“长期固定高薪且现场真实火爆”的偏差理解。

Analysis. 该样本的难点不在于话题是否存在，而在于视觉证据和文本强调是否保留了该说法成立的条件。验证器需要追踪薪资限定、区分已证实的招聘事实与宣传化表述，并避免把吸引眼球的场景图直接当作具体岗位条件的证据。

Figure 6: Benchmark examples from ReMMDBench: an English short sample and a Chinese medium sample with multimodal evidence and distortion annotations.

Verdict	Count	Avg. L2
True	100	0.00
Mostly True	99	2.03
Mixture	100	3.52
Mostly False	102	3.93
False	99	4.41

Table 8: Verdict distribution and average number of L2 labels.

E MMFakeBench Transfer Setting

The transfer experiment uses the official 10,000-instance MMFakeBench test split, whose class distribution is 70% fake and 30% true. All compared agents use Qwen3.5-9B and the same retrieval backend. ReMMD-Agent obtains 0.824 accuracy and 0.871 fake-class F1. MMD-Agent obtains 0.592 accuracy and 0.673 fake-class F1, while T^2-Agent obtains 0.639 accuracy and 0.715 fake-class F1.

Analysis. The transfer setting reduces the output space from five-way veracity and eight distortion labels to binary fake detection. ReMMD-Agent still keeps a large advantage, which suggests that its benefit is not limited to ReMMDBench-specific label definitions. The result is also informative for smaller open-source backbones: T^2-Agent performs more search, but the additional reasoning loop does not compensate for weaker evidence routing when the backbone capacity is limited. ReMMD-Agent's decomposition and memory reuse appear to supply a more stable control policy under the same retriever.

F Additional Fine-Grained Results

The following tables report the complete fine-grained slices used for Figure 5. Tables 10, 11, and 12 use the same grouped-agent layout as Table 4. Grey rows denote general-purpose assistant baselines, and agent rows are grouped by system family.

Distortion label	Count	Percent
T1 Fabrication	99	19.8
T2 Distortion	222	44.4
T3 Misleading Context	164	32.8
V1 Synthetic Visual Content	145	29.0
V2 Visual Editing	272	54.4
C1 Semantic Inconsistency	212	42.4
C2 Contextual Inconsistency	210	42.0
C3 Pragmatic Inconsistency	67	13.4

Table 9: Distortion-label frequency in ReMMDBench.

Short-text analysis. Short posts contain fewer claims and fewer images, but they provide less context for disambiguating entities and events. ReMMD-Agent still leads the strongest L1 results, especially with GPT-5.2, indicating that atomic decomposition is useful even when the textual input is compact. The L2 gap is smaller than in longer tiers because many distortions are visible from local cues, which lets assistant-style baselines remain competitive. Even so, exact match remains low across systems. This indicates that short posts often compress several cues into a small space, so a system must still decide whether a visual cue changes the central claim or only adds suspicious context.

Medium-text analysis. The medium tier is where simple scaling of context begins to fail. Several baselines improve L2 recall, but their L1 macro-F1 remains unstable because partial evidence must be assigned to the correct severity class. ReMMD-Agent/GPT-5.2 gives the strongest verdict performance, while ReMMD-Agent/Qwen3.5-9B gives the best L2 macro-F1. This split suggests that larger proprietary backbones help with calibrated verdict assignment, whereas the decomposition policy can still help a smaller open-source model detect distortion mechanisms. The tier is therefore diagnostic of the benchmark's main difficulty: additional narrative context creates more opportunities for evidence retrieval, but also increases the risk of treating peripheral contradictions as central.

Long-text analysis. The long tier is the most realistic stress test because the average sample contains about ten images and a much longer narrative. ReMMD-Agent/Qwen3.5-9B reaches the highest L1 macro-F1 and L2 macro-F1 in this slice, while ReMMD-Agent/GPT-5.2 has the highest accuracy. This pattern indicates that long posts reward evidence organization: additional context helps only when the agent can bind claims, images, and retrieved sources. The weaker T^2-Agent results show that expanding the reasoning search space is not enough if the retrieved evidence is not tied back to stable atomic units. Long posts also magnify the difference between retrieval volume and retrieval usefulness, since many plausible sources may describe neighboring events, reused images, or partially matching entities.

Language-slice analysis. Table 13 shows that ReMMD-Agent improves both verdict and distortion performance in every language. The gains

Agent	Backbone	L1 Veracity				L2 Distortion			
		Accuracy	Precision	Recall	Macro-F1	Precision	Recall	Macro-F1	Exact Match
Manus	proprietary	30.06	30.31	29.89	29.84	40.06	41.08	40.23	6.36
ChatGPT	proprietary	31.21	25.23	30.67	27.21	37.73	45.99	40.74	1.73
MMD-Agent	GPT-5.2	26.59	26.28	26.63	24.10	43.41	42.25	38.86	2.89
	Gemma4-31B	24.86	26.31	24.98	24.67	38.59	42.21	39.66	5.20
	Qwen3.6-27B	23.12	24.22	23.43	23.10	41.83	35.07	36.20	6.94
	Qwen3.5-9B	23.12	22.16	22.97	22.18	40.02	40.27	37.77	5.20
	Qwen3.5-4B	28.90	30.89	28.80	28.17	40.09	34.00	34.05	8.67
T^2-Agent	GPT-5.2	26.59	34.55	26.40	24.73	40.51	45.47	40.74	2.31
	Gemma4-31B	25.43	26.24	25.54	25.83	41.55	40.98	40.15	9.83
	Qwen3.6-27B	30.06	27.91	29.29	26.13	41.63	27.70	29.75	5.78
	Qwen3.5-9B	28.90	25.08	27.96	25.09	37.63	26.36	27.59	2.89
	Qwen3.5-4B	26.01	23.35	25.05	23.16	37.06	28.29	28.49	2.31
ReMMD-Agent	GPT-5.2	**41.04**	**40.11**	**40.51**	**37.91**	41.74	48.36	**44.45**	5.78
	Gemma4-31B	35.26	35.17	35.07	34.85	41.10	41.32	40.91	6.94
	Qwen3.6-27B	31.79	33.76	31.88	31.14	43.05	36.77	39.01	**9.83**
	Qwen3.5-9B	36.42	38.96	36.29	36.07	40.86	43.43	41.49	**9.83**
	Qwen3.5-4B	32.95	35.18	32.81	32.62	39.95	37.66	37.95	6.36

Table 10: Short-text subset results on 173 samples. Values are percentages and the table follows the same layout as Table 4.

Agent	Backbone	L1 Veracity				L2 Distortion			
		Accuracy	Precision	Recall	Macro-F1	Precision	Recall	Macro-F1	Exact Match
Manus	proprietary	29.56	30.38	28.98	29.34	44.57	46.41	44.83	8.18
ChatGPT	proprietary	25.16	32.05	25.97	24.48	40.21	46.86	41.97	4.40
MMD-Agent	GPT-5.2	23.27	18.02	23.21	19.13	38.56	49.06	39.43	2.52
	Gemma4-31B	23.90	24.53	23.26	23.61	40.72	41.24	40.08	5.66
	Qwen3.6-27B	30.82	32.68	30.56	30.98	39.41	38.43	37.59	6.29
	Qwen3.5-9B	26.42	25.28	25.23	24.84	38.49	36.58	35.93	9.43
	Qwen3.5-4B	25.79	24.49	25.30	24.69	41.24	31.35	33.41	6.92
T^2-Agent	GPT-5.2	28.30	23.57	29.86	25.92	37.94	46.03	39.82	1.89
	Gemma4-31B	25.16	26.33	24.97	25.08	35.52	35.35	33.96	8.18
	Qwen3.6-27B	25.79	24.01	25.31	23.54	40.00	26.93	28.10	3.14
	Qwen3.5-9B	22.64	20.49	22.70	20.03	34.08	26.81	27.65	3.77
	Qwen3.5-4B	19.50	19.91	20.12	19.04	36.60	25.66	27.71	1.26
ReMMD-Agent	GPT-5.2	**40.25**	**48.31**	**41.31**	**39.01**	45.87	49.55	47.00	5.66
	Gemma4-31B	33.33	33.09	32.76	32.10	41.44	40.50	40.61	10.69
	Qwen3.6-27B	30.19	33.28	29.62	29.59	**46.03**	38.70	41.29	**12.58**
	Qwen3.5-9B	33.96	36.08	34.79	34.55	45.13	**52.62**	**47.70**	11.32
	Qwen3.5-4B	22.64	26.79	22.95	22.50	40.56	43.98	41.39	6.92

Table 11: Medium-text subset results on 159 samples. Values are percentages and the table follows the same layout as Table 4.

Agent	Backbone	L1 Veracity				L2 Distortion			
		Accuracy	Precision	Recall	Macro-F1	Precision	Recall	Macro-F1	Exact Match
Manus	proprietary	39.29	39.02	38.91	38.84	46.92	40.52	43.23	8.33
ChatGPT	proprietary	33.93	34.35	32.99	31.90	47.74	49.76	47.83	2.98
MMD-Agent	GPT-5.2	29.17	26.42	28.53	25.82	44.28	54.67	46.18	0.60
	Gemma4-31B	27.98	28.41	27.92	27.65	43.30	42.95	42.01	4.76
	Qwen3.6-27B	22.62	23.11	22.73	22.29	44.04	41.75	38.99	1.19
	Qwen3.5-9B	29.17	30.36	29.33	29.66	41.79	43.80	39.82	2.98
	Qwen3.5-4B	20.24	21.46	20.91	20.43	48.59	32.89	35.70	5.95
T^2-Agent	GPT-5.2	29.76	31.51	29.14	27.65	47.89	49.98	46.90	3.57
	Gemma4-31B	22.62	22.39	22.40	22.31	45.90	39.20	40.95	4.76
	Qwen3.6-27B	22.02	19.35	23.74	19.20	38.81	26.54	25.18	3.57
	Qwen3.5-9B	25.00	24.96	25.90	23.95	43.80	28.18	29.88	2.98
	Qwen3.5-4B	17.86	17.09	18.34	16.98	40.67	27.51	28.89	1.79
ReMMD-Agent	GPT-5.2	**44.05**	41.18	**42.78**	39.41	45.36	42.71	43.37	3.57
	Gemma4-31B	32.14	32.19	32.33	31.75	49.36	41.78	44.89	5.95
	Qwen3.6-27B	29.17	30.86	29.09	28.46	45.00	41.85	42.59	**10.12**
	Qwen3.5-9B	41.07	**46.94**	40.24	**40.79**	46.88	**55.97**	**50.43**	8.93
	Qwen3.5-4B	31.55	34.43	29.82	29.00	46.42	46.44	45.51	4.76

Table 12: Long-text subset results on 168 samples. Values are percentages and the table follows the same layout as Table 4.

Language	System	L1 F1	L2 F1	ΔL1 vs. MMD	ΔL2 vs. MMD
English	MMD-Agent/Qwen3.5-9B	31.22	34.27	ref.	ref.
English	T^2-Agent/Qwen3.5-9B	20.09	27.45	-11.13	-6.82
English	ReMMD-Agent/Qwen3.5-9B	**36.30**	**47.76**	**+5.08**	**+13.49**
Chinese	MMD-Agent/Qwen3.5-9B	28.51	44.30	ref.	ref.
Chinese	T^2-Agent/Qwen3.5-9B	27.62	33.22	-0.90	-11.07
Chinese	ReMMD-Agent/Qwen3.5-9B	**38.20**	**48.81**	**+9.68**	**+4.51**
German	MMD-Agent/Qwen3.5-9B	25.11	39.23	ref.	ref.
German	T^2-Agent/Qwen3.5-9B	22.29	30.35	-2.82	-8.88
German	ReMMD-Agent/Qwen3.5-9B	**37.78**	**48.41**	**+12.67**	**+9.18**
Japanese	MMD-Agent/Qwen3.5-9B	15.97	30.58	ref.	ref.
Japanese	T^2-Agent/Qwen3.5-9B	18.56	21.02	+2.60	-9.57
Japanese	ReMMD-Agent/Qwen3.5-9B	**34.14**	**44.41**	**+18.18**	**+13.82**
French	MMD-Agent/Qwen3.5-9B	15.66	33.47	ref.	ref.
French	T^2-Agent/Qwen3.5-9B	26.68	25.27	+11.02	-8.20
French	ReMMD-Agent/Qwen3.5-9B	**35.28**	**38.28**	**+19.62**	**+4.81**

Table 13: Language-slice results for the Qwen3.5-9B backbone. The table reports verdict macro-F1 and distortion macro-F1, together with absolute gains over MMD-Agent.

are largest for Japanese and French on L1, where MMD-Agent is weakest, suggesting that multilingual verification is constrained by entity anchoring and regional evidence access rather than translation alone. T^2-Agent occasionally improves L1 over MMD-Agent, as in French, but its L2 performance drops sharply. This indicates that broader search may find enough evidence for a coarse verdict while still failing to diagnose the distortion mechanism. The consistent L2 gains are especially important because distortion labels require matching local expressions, named entities, and media provenance across languages, not merely translating the post into English.

Distortion-label analysis. Table 14 confirms that ReMMD-Agent is strongest on labels that require evidence alignment, especially T2 Distortion, V2 Visual Editing, C1 Semantic Inconsistency, and C2 Contextual Inconsistency. These labels depend on comparing the post with external evidence or with the intended image-text binding. The exception is V1 Synthetic Visual Content, where MMD-Agent performs best, suggesting that low-level generation artifacts and forensic cues remain useful even when retrieval is strong. C3 Pragmatic Inconsistency remains difficult for all systems because it depends on the rhetorical use of evidence rather than a single factual contradiction. This pattern supports the paper's central design choice: retrieval memory and atomic parsing are most valuable when the task is to decide how an otherwise plausible source is being used.

G Additional Benchmark Distributions and Confusion Matrices

Benchmark-distribution analysis. Figure 7 summarizes the design pressures behind ReMMD-Bench. The language panel shows that the benchmark is not English-centric and includes cross-lingual cases as a distinct condition. The distortion panel confirms that textual, visual, and cross-modal labels all occur frequently, so systems cannot optimize for a single manipulation family. The provenance panel shows that the dataset mixes reused source images, web-downloaded evidence images, generated images, and edited images. The length panel verifies that short, medium, and long posts are balanced, which makes the length-tier analysis in Figure 5 meaningful. Together, these distributions make the benchmark resistant to narrow shortcuts: a system must handle language variation, visual provenance, and text-image binding at the same time.

Image-count analysis. Figure 8 shows a long tail toward ten and eleven images. This shape is intentional rather than incidental. Many real social-media posts use carousel-style evidence, where some images are central and others are decorative, repeated, or weakly related. A verifier must therefore identify which images actually support the claim and which only add persuasive context. This is one reason ReMMDBench is difficult for agents that treat the image set as an undifferentiated visual bundle. The distribution also explains why memory reuse matters: once evidence is retrieved for one image or claim, it can often resolve later bindings without repeating the same search.

Confusion-matrix analysis. Figure 9 compares GPT-backed systems under the five-way verdict scale. Direct prompting and T^2-Agent show a visible tendency to avoid confident True predictions and to concentrate mass around middle labels. This suggests a conservative model bias: when the task involves misinformation, models often treat uncertainty itself as evidence of partial falsehood. ReMMD-Agent reduces this drift by forcing the judge to keep supported, contradicted, and unverified atomic points separate. The remaining confusion around Mostly True, Mixture, and Mostly False is expected, because these labels depend on the centrality of the disputed evidence rather than the mere presence of an error. The matrix therefore provides a qualitative explanation for the macro-F1 gains: the agent improves not by eliminating ambiguity, but by reducing systematic drift caused by unmanaged uncertainty.

Label	System	F1	Δ vs. MMD	Δ vs. T^2	Best
T1 Fabrication	MMD-Agent/Qwen3.5-9B	36.43	ref.	+3.43	
T1 Fabrication	T^2-Agent/Qwen3.5-9B	33.00	-3.43	ref.	
T1 Fabrication	ReMMD-Agent/Qwen3.5-9B	**39.80**	**+3.37**	**+6.80**	✓
T2 Distortion	MMD-Agent/Qwen3.5-9B	27.04	ref.	-1.43	
T2 Distortion	T^2-Agent/Qwen3.5-9B	28.47	+1.43	ref.	
T2 Distortion	ReMMD-Agent/Qwen3.5-9B	**57.75**	**+30.71**	**+29.27**	✓
T3 Misleading Context	MMD-Agent/Qwen3.5-9B	33.62	ref.	+26.01	
T3 Misleading Context	T^2-Agent/Qwen3.5-9B	7.61	-26.01	ref.	
T3 Misleading Context	ReMMD-Agent/Qwen3.5-9B	**37.29**	**+3.66**	**+29.68**	✓
V1 Synthetic Visual Content	MMD-Agent/Qwen3.5-9B	**51.78**	ref.	**+16.80**	✓
V1 Synthetic Visual Content	T^2-Agent/Qwen3.5-9B	34.98	-16.80	ref.	
V1 Synthetic Visual Content	ReMMD-Agent/Qwen3.5-9B	45.63	-6.14	+10.65	
V2 Visual Editing	MMD-Agent/Qwen3.5-9B	51.60	ref.	-3.13	
V2 Visual Editing	T^2-Agent/Qwen3.5-9B	54.73	+3.13	ref.	
V2 Visual Editing	ReMMD-Agent/Qwen3.5-9B	**62.29**	**+10.69**	**+7.56**	✓
C1 Semantic Inconsistency	MMD-Agent/Qwen3.5-9B	47.24	ref.	+23.66	
C1 Semantic Inconsistency	T^2-Agent/Qwen3.5-9B	23.57	-23.66	ref.	
C1 Semantic Inconsistency	ReMMD-Agent/Qwen3.5-9B	**54.88**	**+7.65**	**+31.31**	✓
C2 Contextual Inconsistency	MMD-Agent/Qwen3.5-9B	45.41	ref.	-0.85	
C2 Contextual Inconsistency	T^2-Agent/Qwen3.5-9B	46.26	+0.85	ref.	
C2 Contextual Inconsistency	ReMMD-Agent/Qwen3.5-9B	**57.86**	**+12.45**	**+11.60**	✓
C3 Pragmatic Inconsistency	MMD-Agent/Qwen3.5-9B	12.07	ref.	+12.07	
C3 Pragmatic Inconsistency	T^2-Agent/Qwen3.5-9B	0.00	-12.07	ref.	
C3 Pragmatic Inconsistency	ReMMD-Agent/Qwen3.5-9B	**20.21**	**+8.14**	**+20.21**	✓

Table 14: Per-label L2 F1 for the Qwen3.5-9B backbone. The only label where ReMMD-Agent is not best is V1, indicating that low-level synthetic-image cues remain complementary to evidence retrieval.

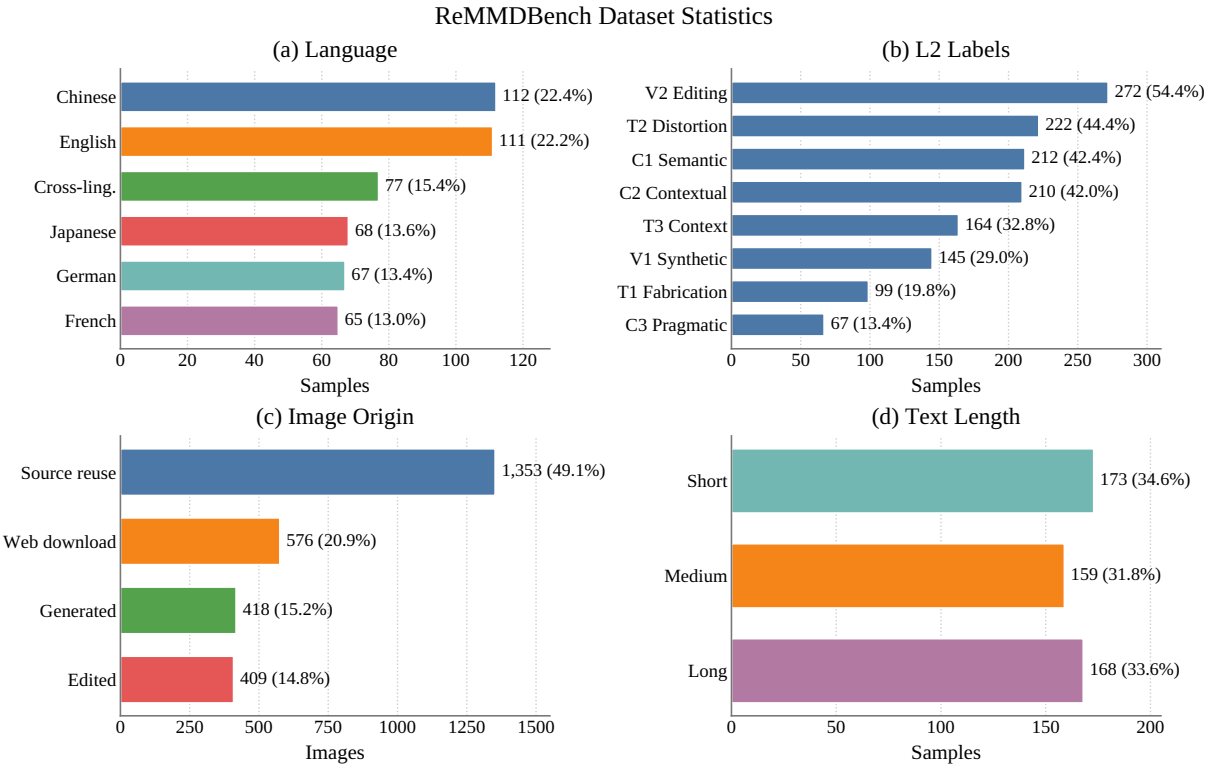


Figure 7: ReMMDBench statistics over language, L2 distortion labels, image provenance, and text-length tiers. Each panel reports counts with percentages in the corresponding sample or image population.

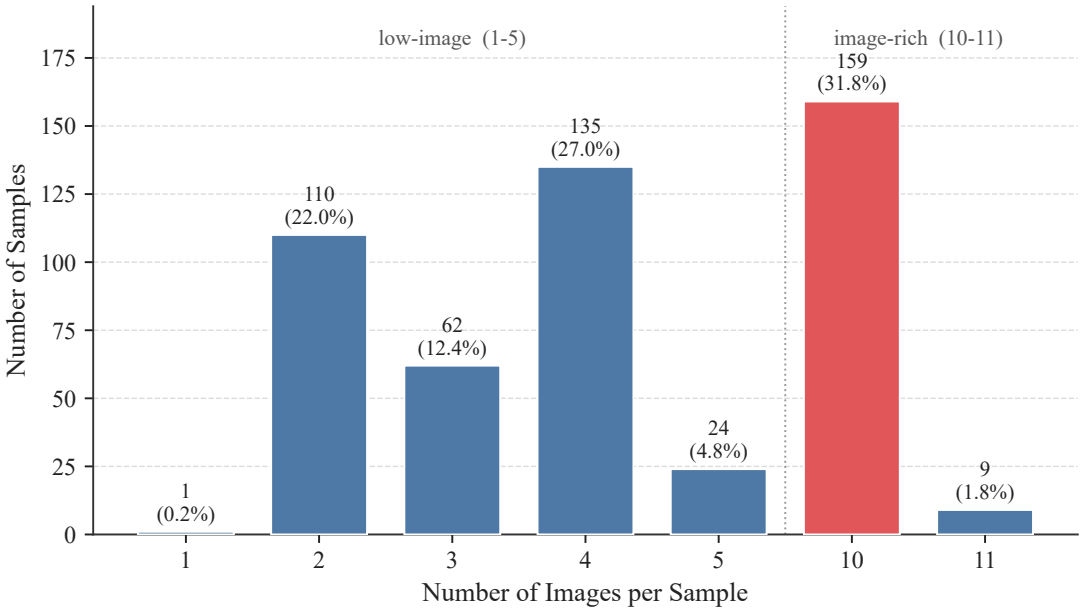


Figure 8: Distribution of images per sample in ReMMDBench. The long tail toward ten or eleven images is intentional and tests whether agents can aggregate evidence across carousel-style posts.

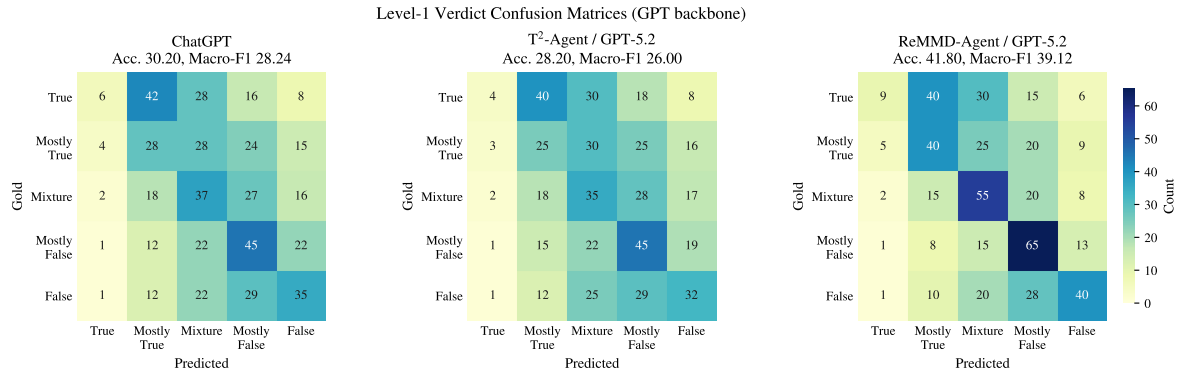


Figure 9: Appendix L1 count heatmaps for GPT-backed systems. Direct prompting and T^2-Agent often drift toward neighboring middle classes, while ReMMD-Agent recovers more diagonal mass without eliminating the intrinsic ambiguity of partial-truth cases.