Computational Humor with Multimodal LLMs: Methods, Datasets, Evaluation, and Challenges

Tuo Liang[1,*], **Zhe Hu**[2,*], **Disheng Liu**[1], **Jing Li**[2], **Yu Yin**[1,†]

[1]Case Western Reserve University, [2]The Hong Kong Polytechnic University

Multimodal humor in memes, cartoons, and comics remains difficult for AI systems because intended meaning depends on non-literal mechanisms, shared cultural knowledge, and communicative intent rather than literal scene description. This survey focuses on visual humor understanding in single-image and multi-panel artifacts, while treating humor generation as an emerging downstream frontier. We position the literature against prior humor, sarcasm, and general MLLM surveys and organize it using a capability-centric hierarchy spanning recognition, interpretation and reasoning, and generation. Under this lens, we synthesize benchmark design, evaluation protocols, and modeling paradigms, tracing the field's shift from task-specific fusion models to large-model approaches based on multimodal alignment, evidence-grounded reasoning, and controlled generation. We conclude by highlighting the main barriers to progress: shortcut-prone evaluation, limited cultural and narrative coverage, weak evidence grounding, and unresolved safety and ownership concerns.

1 Introduction

Recent advances in AI have enabled models to jointly process text and images with unprecedented scale and performance. Multimodal Large Language Models (MLLMs) now achieve strong results on tasks such as image captioning, visual question answering, and cross-modal retrieval Yin et al. (2024); Comanici et al. (2025); Bai et al. (2025), which primarily emphasize the recognition, grounding, and reasoning of explicit visual and textual content. However, this paradigm remains limited when handling content whose meaning extends beyond literal perception Hwang & Shwartz (2023); Chen et al. (2024a); Nayak et al. (2024); Xu et al. (2025). Human communication frequently relies on expressive artifacts such as memes, comics, cartoons, and satirical images, where meaning arises from humor, cultural reference, symbolic association, or intentional incongruity between modalities. Interpreting such media requires reasoning about implicit meaning and communicative intent rather than merely recognizing observable content.

We refer to this problem setting as **multimodal humor understanding**: interpreting image-based artifacts whose humorous, satirical, or ironic meaning emerges from the interaction of visual content and associated text. We focus on single-image and multi-panel image artifacts, and we treat humor generation as an emerging downstream frontier that is useful for probing whether models have internalized the mechanisms of humorous interpretation, rather than as an independent capability to be evaluated in isolation.

Why this gap matters. The gap between perceptual recognition and communicative interpretation is not a minor residual error; it is a structural blind spot in how MLLMs are built and evaluated. A model can correctly identify a burning house, a courtroom, or a smiling face, and still fail to recover the metaphor, the social target, or the punchline that those elements jointly encode. This is because the core difficulty is not multimodal alignment in the usual sense, but reasoning about why a juxtaposition is funny, what background knowledge it presupposes, and what stance it communicates to an audience Saakyan et al. (2024); Ryan et al. (2025). As MLLMs are increasingly deployed in applications such as content moderation, social media analysis,

*These authors contributed equally.
†Corresponding author.

and creative assistance, systematically understanding their capabilities, limitations, and failure modes in these tasks becomes increasingly urgent.

The gap in existing surveys. Despite growing interest, existing surveys only cover this space partially. Surveys of text-based computational humor Amin & Burghardt (2020); Kalloniatis & Adamidis (2024); Loakman et al. (2025); Lemmens & De Marez (2026) center on linguistic phenomena such as puns, wordplay, and verbal irony, but do not address visual grounding or cross-modal incongruity. Multimodal surveys exist, but each narrows to a single task: sarcasm detection Farabi et al. (2024); Gao et al. (2025a) or meme classification Afridi et al. (2020); Ren et al. (2026). Broader MLLM surveys Yin et al. (2024), meanwhile, typically subsume creative and figurative understanding as one capability under general multimodal reasoning, without dedicated systematic treatment.

This survey is designed to fill the space. Rather than grouping work by task name or model architecture, we organize the literature around three progressively demanding capabilities: *recognition*, where systems detect or classify humorous phenomena; *interpretation and reasoning*, where systems explain the mechanism, target, or implicit meaning; and *generation*, where models must produce humor-consistent outputs grounded in the input. This framing clarifies not only what different benchmarks measure, but also which modeling assumptions are needed to move from label prediction to evidence-grounded interpretation.

Concretely, this survey makes the following contributions:

- We propose a capability-centric organization that aligns background concepts, benchmark design, and modeling paradigms.
- We synthesize datasets and evaluation protocols with particular attention to what current benchmarks do and do not measure about interpretive understanding.
- We identify the technical and socio-technical bottlenecks that currently limit progress, including shortcut-prone evaluation, sparse mechanism-level annotation, weak cultural grounding, and safety and ownership concerns.

Survey scope. We include Multimodal Visual Humor, such as *multimodal humor*, *meme understanding*, *visual sarcasm*, *satire detection*, *comic understanding*, *humorous captioning*, and *visual figurative language*. We focus on contemporary MLLMs work, and briefly introduce task-specific multimodal models before the MLLM era in Appendix C We include papers that study image-based artifacts and where success depends on reasoning beyond literal grounding, typically through visual–text interaction, non-literal mechanisms, or communicative intent. We exclude audio- and video-centric humor, generic aesthetic modeling, and persuasive media unless they are directly used as evaluation resources for image-based humor understanding. We additionally used backward and forward snowballing from benchmark and survey papers to recover closely related work. Our goal is not an exhaustive catalog of every humor-adjacent dataset, but a principled synthesis of the resources and modeling strategies that most directly test multimodal visual humor understanding.

2 Background

2.1 Definition of Multimodal Humor

In this survey, we use **multimodal humor** to refer to communicative artifacts that intentionally convey meaning beyond literal depiction through the coordinated use of visual content and associated text, such as internet memes, editorial cartoons, comic strips, and satirical images.

Unlike natural images or descriptive text, human creative media are designed to communicate expressive, humorous, narrative, or satirical meaning, requiring interpretation of what is implied rather than what is directly shown. Their interpretation relies not only on perceptual recognition but also on inference of implicit intent and shared social knowledge. We focus on single-image and multi-panel image artifacts where meaning emerges from visual–text interaction and from non-literal reasoning.

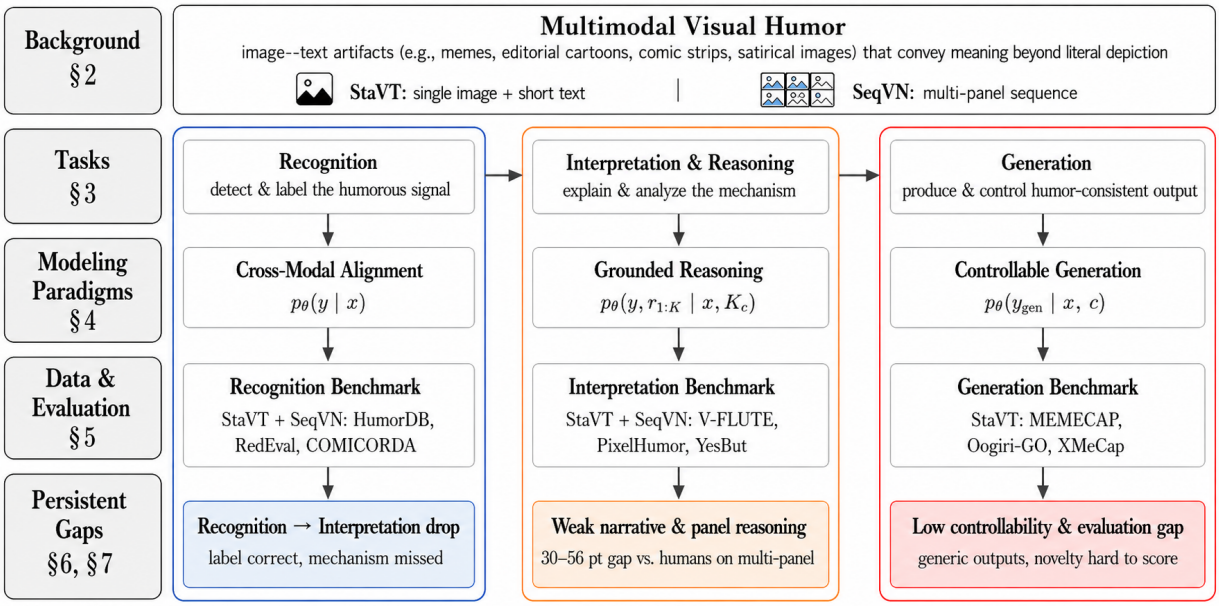

Figure 1 Overview of the survey. We define *multimodal humor* and its two *representation forms*: StaVT (single image + short text) and SeqVN (multi-panel sequence). Three progressively demanding *capabilities*— Recognition, Interpretation & Reasoning, and Generation—form the columns (§3). Row 3 gives the dominant *modeling paradigm* for each capability: cross-modal alignment $p_\theta(y \mid x)$, grounded reasoning $p_\theta(y, r_{1:K} \mid x, \mathcal{K}_x)$, and controllable generation $p_\theta(y_{\text{gen}} \mid x, c)$ (§4, Table 1). Row 4 lists the *benchmark family and evaluation protocol*, tagged by representation form (§5, Table 2); Row 5 (accented) highlights the *persistent gaps* exposed by our cross-benchmark analysis (§6, §7). Horizontal arrows mark capability progression; vertical arrows trace each column top-to-bottom from task to open problem.

2.2 Representation Forms

Multimodal humor spans a wide range of image–text data forms, including single images with captions and multi-panel sequences with dialogue. Modeling multimodal creative understanding can be formulated as learning a conditional mapping: $p(y|x)$, where x denotes multimodal inputs and y represents a task-specific output, such as a humor label, explanation, or generated continuation. The structure of x determines the perceptual encoding, alignment mechanisms, and architectural components required to model this conditional distribution.

Unlike conventional vision or language tasks, creative media often distribute meaning across modalities and (for comics) across panels. Therefore, representation form defines not only the input space but also the model's required integration capacity. We categorize representation forms into two modeling-oriented classes:

(1) Static Visual–Textual Artifacts (StaVT) include memes, editorial cartoons, and captioned images, where $x = \{x^{img}, x^{text}\}$ consists of a single image optionally paired with short text. Meaning frequently arises from cross-modal incongruity or implicit symbolic alignment. Modeling requires joint embedding spaces, multimodal alignment, and sensitivity to social knowledge priors Shifman (2013); Sharma et al. (2020). Architectures typically rely on vision–language encoders or MLLMs with cross-attention mechanisms.

(2) Sequential Visual Narratives (SeqVN) encompass comic strips and multi-panel memes, where meaning emerges from progression across panels, with $x = \{x_1, x_2, \ldots, x_T\}$ representing an ordered visual sequence. Understanding these forms requires tracking entities and events, inferring causal relations, and integrating information across a visual sequence Paval et al. (2025); Wang et al. (2025).

Across all categories, data form shapes not only perceptual requirements but also the types of reasoning and alignment needed for understanding and generating creative meaning.

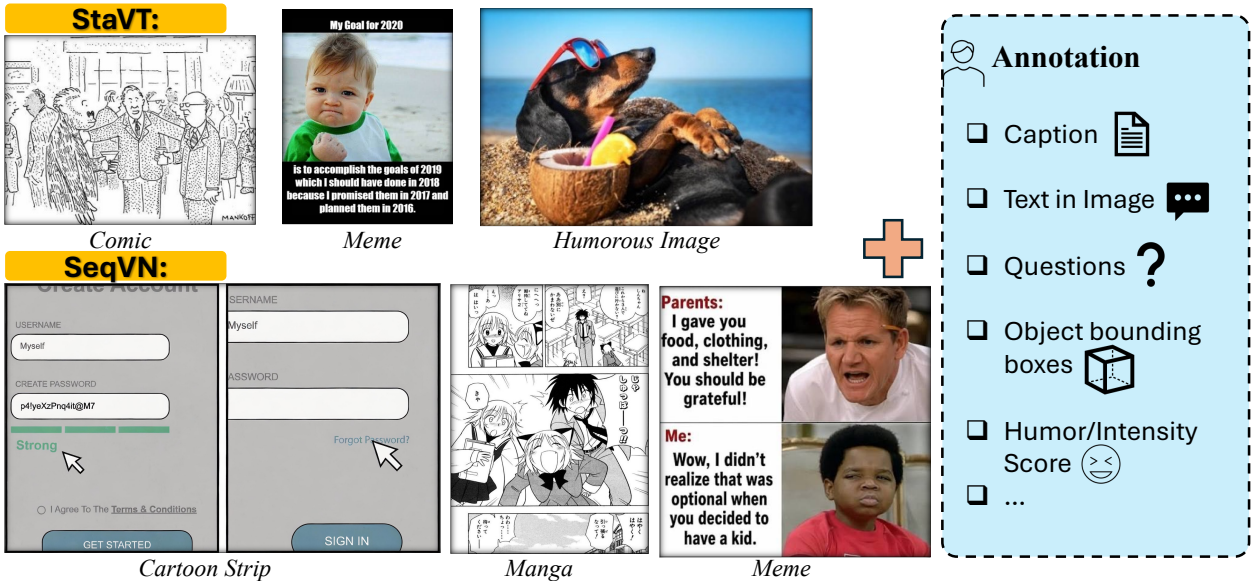

Figure 2 We categorize visual humor data into Static Visual–Textual Artifacts (StaVT), such as comics Hessel et al. (2023b), memes, and humorous images Jain et al. (2025), and Sequential Visual Narratives (SeqVN), such as cartoon strips Liang et al. (2025), manga Ikuta et al. (2025), and multi-panel memes. Common annotations include captions, image text, questions, object bounding boxes, humor or intensity scores, and task-specific labels.

2.3 Creative Meaning Construction

Beyond representation form, creative artifacts convey meaning through non-literal mechanisms. For AI systems, interpretation therefore requires more than recognizing visual and textual patterns: models must also infer why those patterns are used Hu & Shu (2023). In this survey, we treat creative meaning construction as the interaction between recurrent expressive mechanisms.

Across data forms, four mechanisms recur. **Incongruity** creates meaning through a mismatch between expectation and observation, often through cross-modal conflict in multimodal humor Forabosco (1992); Veale (2004); Schifanella et al. (2016); Farabi et al. (2024). **Analogy and Conceptual Mapping** project structure from a familiar source domain onto a target concept, as in metaphor and symbolic representation Lakoff & Johnson (2024); Refaie (2003); Foss (2004). **Exaggeration and Hyperbole** amplify attributes against implicit norms for emphasis or affect Kreuz et al. (1996); Zhang & Wan (2024), while **Narrative Structure** organizes setup, payoff, and causal progression over time Genette (1980); Bruner (1991); Paval et al. (2025). Together, these mechanisms explain how creative artifacts encode meaning beyond surface semantics and why they remain challenging for literalist AI systems.

2.4 Why Multimodal Humor Is Hard for AI

Multimodal humor is fundamentally challenging for AI models because the intended meaning often cannot be directly inferred from observable inputs. Unlike literal multimodal tasks where answers are grounded in explicit visual or textual evidence, creative understanding requires reasoning over latent variables such as implied metaphors, violated expectations, cultural references, and communicative intent. Models must detect incongruity, infer implicit norms, integrate external socio-cultural knowledge, and reason about why an artifact was produced and how it is meant to be interpreted. As a result, creative understanding goes beyond multimodal feature fusion, demanding the integration of perceptual alignment, structured reasoning, and pragmatic inference within a unified modeling framework.

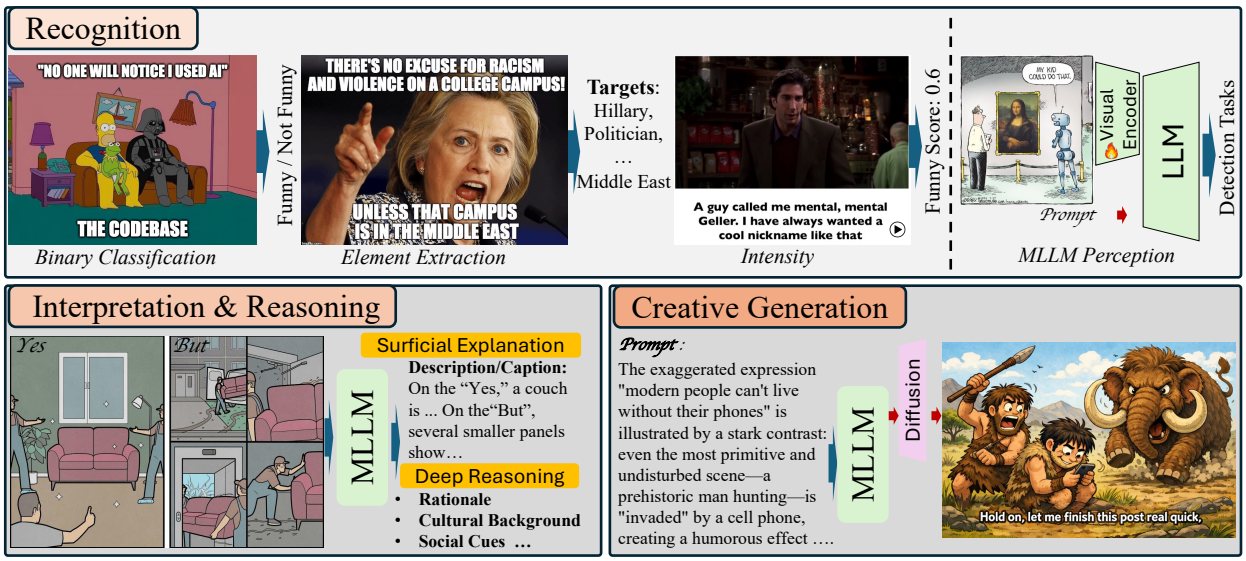

Figure 3 Capability-centric task hierarchy for multimodal humor. We organize tasks into three levels—Recognition, Interpretation and Reasoning, and Generation—reflecting increasing requirements for non-literal grounding, mechanism-aware inference, and evaluation beyond discriminative labels.

3 Task Hierarchy: Recognition, Interpretation, and Generation

We organize existing work into three capability levels and pair each with the evaluation signal it most directly demands (Figure 3). The hierarchy is not purely chronological: recent benchmarks often mix levels, but the distinction is useful because gains in label prediction do not automatically transfer to explanation or generation.

3.1 Level 1: Recognition

Recognition tasks ask whether a multimodal input contains a humorous, sarcastic effect and, in more fine-grained settings, which components instantiate it. Typical problems include binary or multi-class classification, target or role extraction, and intensity estimation. These tasks dominate early work on multimodal sarcasm, meme classification, and visual humor detection Cai et al. (2019); Hasan et al. (2021); Zhang & Wan (2024). They require reliable visual-text alignment and sensitivity to local cues, but they do not by themselves verify whether a model has captured the mechanism that makes an artifact funny or critical.

Recognition tasks are usually evaluated with discriminative metrics such as accuracy, F1, AUROC, or correlation with human ratings. These metrics are appropriate for testing cue sensitivity and class separation, yet they remain weak proxies for genuine understanding: a model can predict a correct label by exploiting recurrent surface patterns without being able to explain the intended meaning.

3.2 Level 2: Interpretation and Reasoning

Interpretation tasks ask models to explain *why* an artifact is humorous, satirical, or ironic. A useful distinction is between **descriptive explanation**, which verbalizes salient cues or paraphrases the joke, and **mechanism-grounded reasoning**, which identifies the specific conflict, analogy, target, or narrative step that produces the effect Hu et al. (2024); Saakyan et al. (2024); Wang et al. (2024b). These tasks require explicit cross-modal grounding, abstraction over non-literal meaning, and often external knowledge or context to resolve implicit references. For multi-panel inputs, they also require temporal and narrative reasoning across panels rather than within a single image Paval et al. (2025); Wang et al. (2025).

Evaluation at this level must combine language quality with interpretive faithfulness. Automatic metrics such as BLEU or BERTScore Papineni et al. (2002); Zhang et al. (2019) are useful for checking lexical or semantic overlap, but they remain insufficient when several explanations are plausible. Stronger protocols ask whether

Table 1 Modeling paradigms for multimodal humor in the large-model era. $x = (x^v, x^t)$ denotes the image–text input; y denotes a recognition or interpretation output; $r_{1:K}$ denotes an intermediate rationale chain; $\mathcal{K}_x$ denotes input-specific external knowledge retrieved or selected for x; and c denotes a creative control signal.

Paradigm	Core Technique	Objectives	Strength	Limitation	Representative
Cross-Modal Alignment (§4.1)	Instruction tuning; contrastive VL pre-training; modular expert routing	$p_\theta(y \mid x)$: label prediction from a joint visual–textual representation	Scalable; strong on detection & classification; transfers across domains	Mechanism stays implicit; prone to shortcut learning & hallucinated intent	SoMeLVLM Zhang et al. (2024b), MMoE Yu et al. (2024), YesBut-v2 Liang et al. (2025)
Grounded Reasoning (§4.2)	CoT prompting; rationale supervision; theory-guided decomposition; external retrieval	$p_\theta(y, r_{1:K} \mid x, \mathcal{K}_x)$: jointly infer the interpretation and rationale chain, optionally grounded in retrieved knowledge	Transparent; mechanism-aware explanations; supports cultural, temporal, and social grounding	Computationally expensive; prompt-sensitive; retrieval noise and reasoning drift can mislead the model	HumorChain Zhang et al. (2025), MemeMind Gu et al. (2025), MemeX Sharma et al. (2023c)
Controllable Generation (§4.3)	Prompt-based generation; instruction tuning; control signals; style/content constraints	$p_\theta(y_{\text{gen}} \mid x, c)$: generate humor-consistent output conditioned on the input and creative control signal	Enables targeted humor captioning, rewriting, and style-controlled output	Low controllability; generic outputs; novelty and humor quality hard to evaluate	MemeCap Hwang & Shwartz (2023), ViPE Shahmohammadi et al. (2023), MemeCraft Wang & Lee (2024)

a model identifies the relevant cues, names the right mechanism, and stays consistent with available evidence, often through rubric-based human judgment or LLM-assisted evaluation Liu et al. (2023); Hu et al. (2024).

3.3 Level 3: Generation

Generation tasks require models to produce humor-consistent outputs such as captions, punchlines, explanations, or comic continuations conditioned on an input artifact. In this survey, we treat generation as an emerging downstream frontier rather than the core of the field: successful generation presupposes at least partial understanding of incongruity, target selection, tone, and narrative setup Hwang & Shwartz (2023); Li et al. (2023); Tanaka et al. (2024). The challenge is therefore not only to produce fluent text, but also to maintain faithfulness to the source image and control over the mechanism being realized.

Because valid outputs are diverse, evaluation at this level relies heavily on human preference judgments or rubric-based assessment. Reference-based metrics remain weak proxies for humor quality and novelty, so the most informative benchmarks combine generation quality with tests of faithfulness to the source image, intended target, and rhetorical device.

4 Modeling Paradigms in the Large-Model Era

With the rise of MLLMs, multimodal humor modeling has shifted from handcrafted fusion toward alignment-driven representation learning, explicit reasoning, and evidence-grounded generation. We organize the modeling literature by the **capability** it primarily supports rather than by model family, because the same backbone behaves very differently when optimized for *recognition*, *interpretation*, or *controlled generation*. Table 1 summarizes the three resulting paradigms along their input formulation, technical core, and characteristic failure modes; the strongest recent systems combine them rather than treating them as substitutes. We detail each paradigm below and close with a cross-paradigm analysis.

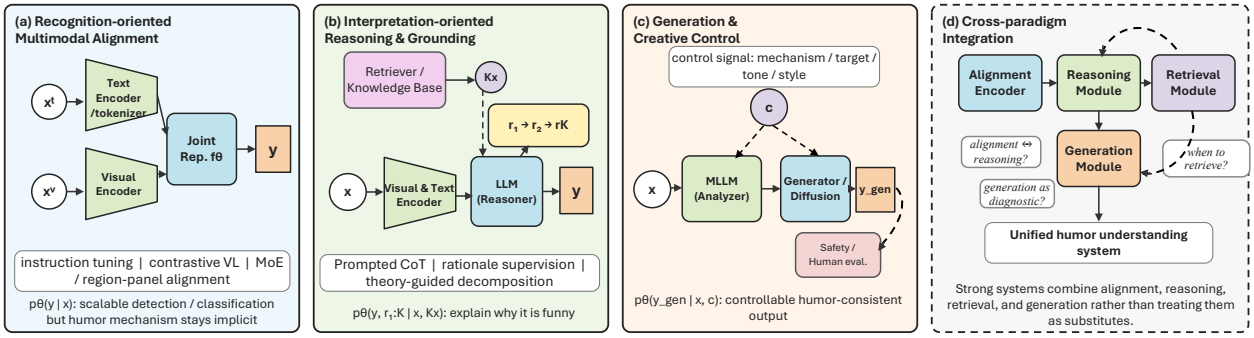

Figure 4 (a) Recognition-oriented multimodal alignment maps visual and textual inputs into a shared representation for scalable humor detection and classification, but leaves the underlying humor mechanism implicit. (b) Interpretation-oriented reasoning and grounding introduces intermediate rationales and optional external knowledge to explain why an artifact is humorous. (c) Generation and creative control conditions humorous output generation on control signals such as mechanism, target, tone, and style, with evaluation and safety constraints. (d) Cross-paradigm integration combines alignment, reasoning, retrieval, and generation into a unified framework, highlighting open questions about alignment–reasoning interaction, retrieval timing, and generation as a diagnostic of understanding.

4.1 Recognition-Oriented Multimodal Alignment

Recognition-oriented systems learn a joint encoder f_θ that maps visual and textual inputs into a shared semantic space and minimizes a task loss $\mathcal{L} = \mathbb{E}[\ell(g_\phi(f_\theta(x^v, x^t)),\ y)]$. The critical design choices are (i) *how* f_θ is trained—via contrastive pre-training, instruction tuning, or modular routing—and (ii) *at what granularity* visual features are extracted. Three strategies span this design space.

Creativity-oriented instruction tuning adapts a pre-trained MLLM with creativity-specific $(x^v, x^t, \text{instruction}, y)$ supervision. The principle is that humor, sarcasm, and meme communication have statistical regularities absent from generic VL corpora, so domain-specific fine-tuning consistently outperforms general-purpose checkpoints on recognition benchmarks Zhang et al. (2024b). A contrastive alternative optimizes an InfoNCE loss over image–text embeddings, exposing cross-modal correspondences for multi-task meme classification Shah et al. (2024). Across both strategies, gains are most pronounced on surface-level labels; performance on nuanced intent lags behind, suggesting that alignment alone captures *what co-occurs* but not *why*.

Modular and text-centric designs decouple perception from reasoning. Mixture-of-expert routing assigns vision and language tokens to specialized sub-networks, improving robustness when humor depends on only one modality Yu et al. (2024). A more radical approach converts all non-textual modalities into text via captioning, reducing multimodal fusion to unified self-attention Hasan et al. (2023); Baluja (2025). This text-centric strategy is surprisingly effective for humor understanding—even without architectural changes to the LLM—but incurs information loss on visually dense inputs where spatial layout or fine-grained detail carries the joke.

Multi-panel and region-aware alignment addresses sequential visual narratives $(x = \{x_1^v, \ldots, x_T^v\})$, where meaning emerges across panels. The key technical challenge is to capture both intra-panel content and inter-panel relations (entity co-reference, causal flow, setup–punchline structure). Current approaches range from multi-panel instruction tuning Liang et al. (2025), to panel-selection tasks that test narrative coherence Vivoli et al. (2025), to RL-trained region-level encoders that attend to character expressions and speech bubbles within each panel Chen et al. (2025b). These extensions are necessary because entity continuity and visual salience across panels are prerequisites for downstream interpretation.

Collectively, alignment methods are scalable and strong on classification, but the humor mechanism remains implicit: a model can predict the correct label while failing to recover the violated expectation or social target that drives the humor—a gap confirmed by the recognition-to-interpretation drop in Section 6.

4.2 Interpretation-Oriented Reasoning and Grounding

Once tasks demand explanation, alignment alone is insufficient because it captures *what co-occurs* but not *why it is funny*. Interpretation-oriented methods address this by conditioning the prediction on a chain of intermediate reasoning steps: $p(y \mid x) \approx \prod_{k=1}^{K} p(r_k \mid r_{<k}, x) \cdot p(y \mid r_{1:K}, x)$, where each r_k is a natural-language rationale that exposes part of the humor mechanism. The design space varies along two axes: (i) *how the reasoning trace is obtained*—prompted at inference, learned from human annotations, or imposed by theory—and (ii) *where missing knowledge comes from*—the model's own parameters or an external retriever.

Prompted vs. supervised reasoning. The cheapest approach elicits chain-of-thought reasoning at inference time via carefully constructed prompts Wei et al. (2022). Prompting models to first describe each panel and then articulate the conflict yields clear gains on multi-panel humor, where the contradiction is cross-panel rather than within a single frame Hu et al. (2024); similar staged prompts decompose conversational jokes into setup, incongruity, and resolution Chen et al. (2024c). However, prompted reasoning is inherently unstable: output quality varies with prompt phrasing, and models can hallucinate plausible-sounding but factually wrong rationales. Training-time rationale supervision is more robust. When models are jointly trained to generate the reasoning chain *and* the label—i.e., $p(y, r_{1:K} \mid x)$—both accuracy and interpretability improve substantially over pattern-based baselines, as demonstrated at scale on harmful-meme datasets Gu et al. (2025). An alternative is to apply an information bottleneck objective that compresses joke representations to retain only mechanism-relevant features before explanation Hwang et al. (2025). The practical trade-off is clear: prompted CoT is zero-cost but fragile; rationale supervision is strong but requires expensive human annotation and risks overfitting to annotator phrasing.

Theory-guided decomposition. A deeper commitment to structure anchors the reasoning pipeline in established humor theories, imposing fixed stages rather than free-form chains. The dominant template follows the incongruity-resolution model: (i) *setup extraction*—identify the expected scenario; (ii) *conflict detection*—localize the violated expectation; (iii) *resolution*—explain how the conflict produces humor Tikhonov & Shtykovskiy (2024); Zhang et al. (2025). This staged design generalizes better than single-pass prediction because each stage is independently evaluable, and theory labels (incongruity, superiority, relief) can steer the decomposition. The same principle of explicit intermediate targets appears in multimodal QA frameworks that chain evidence retrieval with step-by-step reasoning Agarwal et al. (2024), in figurative-language benchmarks that frame understanding as natural-language inference over non-literal hypotheses Saakyan et al. (2024, 2025), and in probes showing that visual metaphor and symbolic-graphics comprehension collapses without intermediate decomposition Kundu et al. (2025); Qiu et al. (2024). The limitation is that theory-guided pipelines can over-analyze simple humor or introduce reasoning drift when the mechanism does not neatly fit the assumed template.

External knowledge retrieval and grounding. When the missing information is social, cultural, or temporal rather than perceptual, reasoning over the input alone is insufficient. Knowledge-grounded systems augment the prediction with a retrieval step: $p(y \mid x, \mathcal{K}_x)$, where $\mathcal{K}_x = \mathcal{R}(x; \mathcal{D})$ is evidence selected from an external corpus $\mathcal{D}$. Three integration strategies have emerged. *Dense passage retrieval* encodes the meme as a query and retrieves background documents that explain implicit cultural references Sharma et al. (2023c). *Knowledge-graph and commonsense injection* augments feature representations with structured world knowledge, which is especially valuable for detecting offensiveness that cannot be inferred from surface cues Garg et al. (2025); Kumari et al. (2025). *Feature-space adaptation* fuses domain-specific attributes (e.g., brand, persuasion technique) directly into the visual encoder via lightweight adapters, avoiding the latency of an explicit retriever Jia et al. (2023). A related technique retrieves similar labeled examples at inference time for in-context learning, enabling domain-shift robustness without retraining Tang et al. (2024); reasoning-knowledge distillation from large LLMs into smaller students offers another route to grounding Lin et al. (2023). Retrieval is critical for satire, time-sensitive memes, and culturally loaded references, but it introduces noise: irrelevant evidence can mislead the model and obscure the actual creative mechanism.

4.3 Generation and Creative Control

Generation can be framed as conditional decoding $p(y_{\text{gen}} \mid x, c)$, where x is the source artifact and c encodes a creative control signal (intended mechanism, target concept, tone). The central difficulty is that effective c

presupposes interpretive competence: generating a funny caption for an image requires understanding *what is funny about* that image. Current work separates along three design principles.

Mechanism-explicit captioning adopts a two-stage *analyze-then-generate* pipeline. The analysis stage identifies the incongruity, salient objects, or violated expectation; the generation stage decodes a caption conditioned on that analysis. This architecture—instantiated through incongruity-resolution CoT prompting Tanaka et al. (2024), grounded humor captioning benchmarks Li et al. (2023), and cascaded describe-explain-caption pipelines Hwang & Shwartz (2023)—consistently outperforms direct prompting on image-specificity. A practical finding is that decoding-time interventions such as logit bias and negative sampling further suppress generic outputs, confirming that the bottleneck is not fluency but groundedness. The cascaded design has its own cost: errors in early stages propagate, making the overall quality sensitive to the weakest link.

Creative artifact synthesis extends generation beyond text to novel images or multimodal artifacts, typically by separating semantic planning (LLM) from visual rendering (diffusion model). Visual metaphor generation pipelines first extract source and target domains, then compose a concrete scene description, and finally render it—allowing each module to be evaluated and improved independently Chakrabarty et al. (2023); Shahmohammadi et al. (2023). Meme generation adds a stance or safety dimension: unconstrained generation produces more novel outputs but also higher rates of harmful content Wang & Lee (2024). A distinct thread explores "creative leaps"—humor requiring associative jumps beyond logical entailment. The emerging finding is that moderate-distance leaps between setup and resolution are rated funniest by human judges; too close is predictable, too far is nonsensical Zhong et al. (2024); Wang et al. (2024a).

Human-AI co-creation addresses a key limitation of fully autonomous systems: the absence of audience feedback during creative production. Interactive tools that suggest incongruity-based options for human refinement show that even novice users produce higher-quality humor when the system handles mechanism generation while the human curates tone and relevance Kariyawasam et al. (2024). Extending generation to temporal media such as short-video commentary further raises the bar, requiring both cross-modal alignment and temporal reasoning Ouyang et al. (2025).

Generation remains bottlenecked by controllability and evaluation. Models often regress toward generic captions or reuse familiar templates. Current metrics—both reference-based (BLEU, BERTScore) and human preference—cannot reliably distinguish genuine novelty from paraphrasing. Generation is therefore best viewed as a stress test of interpretive competence: without understanding the mechanism, faithful creative output is unlikely.

4.4 Cross-Paradigm Analysis

The three paradigms are complementary: alignment provides scalable perception, reasoning adds interpretive depth, and retrieval supplies missing context. The strongest recent systems already combine elements—e.g., theory-guided reasoning over alignment-tuned features Zhang et al. (2025), CoT supervision with commonsense grounding Gu et al. (2025), or reasoning decomposition feeding into generation Tanaka et al. (2024)—but integration remains ad hoc rather than principled.

Three open questions emerge from this landscape. First, *alignment–reasoning interaction*: does finer-grained visual alignment (region-level, panel-level) reduce the burden on downstream reasoning, or does it introduce redundant detail that increases reasoning drift? Second, *retrieval timing*: should evidence be injected before reasoning (to inform the trace) or after an initial reasoning pass (to fill identified gaps)? Current systems use fixed timing, but adaptive strategies could improve efficiency. Third, *generation as a diagnostic for understanding*: the field treats generation as a downstream application, yet generation quality could serve as a stronger probe of interpretive competence than MCQ accuracy—if appropriate evaluation protocols existed. Addressing these questions will require joint benchmarks that evaluate alignment, reasoning, and generation on the same artifacts, a direction we revisit in Section 6.

Table 2 Capability-aligned summary of benchmark resources. Detailed dataset inventories appear in Appendix Tables 4 and 5.

Level	Representative Resources	Typical Supervision / Target	Typical Evaluation	Recurring Limitation
Recognition	RedEval Tang et al. (2024), HumorDB Jain et al. (2025), COMICORDA Martínek et al. (2024), Inside Jokes Shahaf et al. (2015)	Binary labels, role labels, dialogue acts, or intensity scores	Accuracy, F1, AUROC, correlation	Shortcut learning remains easy; labels reveal little about interpretive depth.
Interpretation & reasoning	Do Androids Laugh at Electric Sheep? Hessel et al. (2023b), V-FLUTE Saakyan et al. (2025), PixelHumor Ryan et al. (2025), YesBut Hu et al. (2024)	Explanations, rationales, answer selection, cross-panel reasoning targets	MCQ accuracy, BERTScore, rubric-based human or LLM-assisted judgment	Limited cultural coverage and weak evidence annotation make faithful evaluation difficult.
Generation	MEMECAP Hwang & Shwartz (2023), OxfordTVG-HIC Li et al. (2023), Oogiri-GO Zhong et al. (2024), XMeCap Chen et al. (2024b)	Captions, punchlines, continuations, or humorous rewrites	Reference overlap, LLM-as-judge, human preference	Novelty, faithfulness, and safety are hard to score automatically; dedicated benchmarks remain sparse.

5 Datasets and Benchmarks

In this section, we align benchmarks with the same capability hierarchy used for tasks and models. This makes it easier to distinguish resources that only test label prediction from those that require explanation, cross-panel inference, or image-grounded generation. Detailed inventories are provided in Appendix Tables 4 and 5.

Recognition Resources. Recognition benchmarks remain the most abundant and the most standardized. In the image-based setting, they cover humorous cartoons, sarcastic image–text pairs, and multi-panel comics through labels for humor presence, target type, dialogue act, or intensity Shahaf et al. (2015); Tang et al. (2024); Jain et al. (2025); Martínek et al. (2024). These resources are useful for training perception and alignment modules, but they remain classification-oriented and therefore provide only indirect evidence about whether a model understands the underlying joke, target, or narrative conflict.

Interpretation and Reasoning Resources. More recent benchmarks explicitly probe interpretive depth by pairing image-based inputs with explanations, rationales, multiple-choice reasoning questions, or cross-panel inference targets. Datasets such as *Do Androids Laugh at Electric Sheep?*, V-FLUTE, PixelHumor, and the YesBut series move beyond surface labels and ask whether models can identify why a caption is funny, what contradiction drives a panel sequence, or how an implicit target should be resolved Hessel et al. (2023b); Saakyan et al. (2025); Hu et al. (2024); Ryan et al. (2025). These resources are far more diagnostic than recognition datasets, but they remain smaller, culturally narrower, and more expensive to annotate.

Generation Resources. Dedicated generation benchmarks are fewer and often repurpose understanding data as supervised targets. MEMECAP and OxfordTVG-HIC center humorous caption generation from static images, while Oogiri-GO and XMeCap extend the space toward joke completion, meme rewriting, or continuation over structured inputs Hwang & Shwartz (2023); Li et al. (2023); Zhong et al. (2024); Chen et al. (2024b). Their main value is diagnostic: they reveal whether a model can transform an interpretation of the source image into a controlled, novel output. At present, however, generation benchmarks remain sparse, and their evaluation protocols are still less mature than those used for recognition or explanation.

Table 3 Results on humor understanding benchmarks. All numbers are accuracy (%). Model results are obtained by evaluating a broader and more recent set of MLLMs using the benchmark-specific prompts and task definitions reported in the original papers. Human performance figures, where available, are taken from the corresponding benchmark papers. Task abbreviations: for **YesBut-v2** Liang et al. (2025), *Moral* and *Title* denote choosing the correct moral or title from multiple choices; **NYCC** Hessel et al. (2023b) denotes selecting the most suitable *Caption* for a New Yorker cartoon from options A–E; **MemeQA** Nguyen et al. (2025) denotes multiple-choice *Fill*-in-the-blank question answering on memes; **ExHVV** Sharma et al. (2023a) denotes *Role* selection for entities from {hero, villain, victim} in memes; **DarkHumor** Kasu et al. (2025) denotes binary *Detect*ion of dark humor (Yes/No); **HumorDB** Jain et al. (2025) denotes binary *Detect*ion of humor (Yes/No). For **MangaUB** Ikuta et al. (2025), *RecBg*, *CharCnt*, *PanelLoc*, *NextInf*, and *Onom* denote *recognition_background*, *character_count*, *panel_localization*, *next_panel_inference*, and *onomatopoeia_scene*, respectively.

Models	Recognition								Interpretation & Reasoning			
	ExHVV	DarkHumor	HumorDB	MangaUB					YesBut-v2		NYCC	MemeQA
	Class.	*Detect*	*Detect*	*RecBg*	*CharCnt*	*PanelLoc*	*NextInf*	*Onom*	*Moral*	*Title*	*Caption*	*Fill*
Qwen2.5-VL-7B	71.65	47.37	65.90	91.69	85.20	79.19	35.01	87.21	67.33	76.94	47.34	39.61
LLaVA-OneVision-7B	73.48	48.06	64.06	97.57	93.79	67.59	33.37	83.66	67.64	70.80	58.14	40.96
Qwen3-VL-8B	76.24	49.43	71.02	95.10	92.02	83.64	48.55	89.03	74.69	80.51	50.94	49.84
Qwen3-VL-8B-Thinking	78.64	**66.43**	70.59	94.35	90.07	82.26	32.76	91.34	70.67	78.82	46.49	37.66
InternVL3.5-8B	77.64	49.77	72.58	94.97	90.87	78.58	46.68	88.45	69.30	76.10	48.86	49.50
InternVL3.5-30B-A3B	79.44	54.76	70.45	96.59	89.98	82.06	48.22	90.10	75.46	–	51.89	49.40
Qwen3.5-35B-A3B	**80.03**	66.37	68.60	92.90	91.22	87.27	33.46	84.57	76.06	78.40	60.25	**60.41**
Gemma-4-31B-it	69.35	57.68	51.10	39.20	12.85	10.01	23.51	63.20	52.71	35.65	22.31	40.16
Qwen3.5-27B	77.24	65.84	**74.71**	94.04	**95.21**	90.95	54.08	83.25	**84.72**	**83.28**	61.79	54.52
InternVL3.5-38B	70.06	61.47	51.98	69.16	12.94	41.72	45.43	69.64	64.58	69.65	34.81	51.39
GPT-4o	74.80	61.50	59.95	**97.60**	94.50	**99.21**	**64.30**	**95.80**	80.38	80.62	**82.30**	59.60
Human	81.00	–	85.00			–			91.30	97.50	94.00	81.90

6 Cross-Benchmark Empirical Analysis

To move beyond qualitative synthesis, we extend the original evaluation protocols of the surveyed benchmarks to a broader and more recent set of MLLMs, while retaining the benchmark-specific prompts reported in the corresponding papers. We organize the resulting scores according to the capability hierarchy introduced in Section 3. This cross-benchmark empirical analysis serves two purposes: (1) it provides an updated snapshot of current MLLM performance across different humor-understanding capabilities, and (2) it reveals recurring patterns that are difficult to observe from individual benchmarks alone, particularly the contrast between recognition-oriented tasks and interpretation- or reasoning-oriented tasks. Human performance figures, where available, are taken from the original benchmark papers.

Evaluation protocol. For each benchmark, we follow the prompt format and task definition reported in the original paper. We use the official evaluation split whenever available and evaluate a broader set of recent MLLMs under the same benchmark-specific protocol. Because prompts differ across benchmarks by design, the results support within-benchmark model comparison and descriptive cross-benchmark analysis, rather than a strictly controlled comparison of task difficulty.

6.1 Cross-Level Performance Landscape

Table 3 summarizes MLLM results across seven humor-understanding benchmarks and twelve task settings. Following our capability hierarchy, we group ExHVV, DarkHumor, HumorDB, and MangaUB under *recognition*, since these tasks primarily require models to identify roles, detect humor categories, or recognize visual and structural elements. We group YesBut-v2, NYCC, and MemeQA under *interpretation & reasoning*, since these tasks require models to infer morals, select appropriate titles or captions, and fill in missing semantic content.

Several patterns emerge from this cross-benchmark view.

Recognition-oriented tasks generally yield higher reported accuracies, but remain far from uniformly solved. Models generally perform strongly on visually grounded recognition tasks, especially the MangaUB subtasks. GPT-4o achieves 97.60% on background recognition, 99.21% on panel localization, 64.30% on next-panel inference, and 95.80% on onomatopoeia-scene recognition, obtaining the best score on four of the

five MangaUB subtasks. Open-source models are also competitive on several recognition tasks: Qwen3.5-27B reaches 95.21% on character counting and 74.71% on HumorDB, while Qwen3.5-35B-A3B achieves the best ExHVV role-classification accuracy at 80.03%. However, recognition is not uniformly easy. DarkHumor remains difficult for most models, with the best score reaching only 66.43% with Qwen3-VL-8B-Thinking. This suggests that recognition remains challenging when it depends on implicit social norms, taboo framing, or pragmatic cues.

Interpretation and reasoning remain the central bottleneck. Compared with recognition-oriented tasks, interpretation-oriented benchmarks show larger and more consistent gaps from human performance. On YesBut-v2, the best model score is 84.72% for moral selection and 83.28% for title selection, both achieved by Qwen3.5-27B, while human performance reaches 91.30% and 97.50%, respectively. On NYCC caption selection, GPT-4o substantially outperforms all other models with 82.30%, but still trails the human score of 94.00%. MemeQA shows a similar gap: the best model, Qwen3.5-35B-A3B, reaches 60.41%, compared with 81.90% for humans. These results indicate that current MLLMs can often recognize salient entities or visual structures, but still struggle to infer the intended humorous mechanism, implicit punchline, or culturally appropriate interpretation.

Model ranking varies substantially across humor capabilities. No single model dominates all task types. GPT-4o is strongest on NYCC and most MangaUB subtasks, suggesting strong visual recognition and caption-matching ability. Qwen3.5-27B performs best on both YesBut-v2 subtasks and HumorDB, indicating stronger performance on moral and title inference as well as general humor detection. Qwen3.5-35B-A3B achieves the highest ExHVV and MemeQA scores, while Qwen3-VL-8B-Thinking performs best on DarkHumor. This fragmented ranking suggests that humor understanding is not a single monolithic ability. Different benchmarks emphasize different combinations of visual recognition, social knowledge, narrative inference, and pragmatic reasoning.

Reasoning-oriented variants do not consistently improve humor understanding. The comparison between Qwen3-VL-8B and Qwen3-VL-8B-Thinking is particularly revealing. The Thinking variant substantially improves DarkHumor detection from 49.43% to 66.43% and improves ExHVV from 76.24% to 78.64%, suggesting that deliberative reasoning can help when the task requires social or normative judgment. However, it decreases performance on several interpretation tasks, including YesBut-v2 Moral, NYCC, and MemeQA, and also drops sharply on MangaUB next-panel inference. This mixed pattern indicates that explicit reasoning does not automatically translate into better humor understanding. In some settings, extended reasoning may help identify implicit intent, while in others it may divert the model from direct visual–semantic matching or introduce unnecessary intermediate steps.

Human performance remains substantially higher on interpretation-heavy tasks. Where human results are available, the largest gaps appear in tasks requiring semantic or pragmatic interpretation. Humans outperform the best model by 6.58 points on YesBut-v2 Moral, 14.22 points on YesBut-v2 Title, 11.70 points on NYCC, and 21.49 points on MemeQA. The gap is also large on HumorDB, where the best model reaches 74.71% compared with 85.00% for humans. These gaps reinforce the main finding of this section: current MLLMs have made considerable progress on visual and categorical recognition, but still fall short in recovering the intended meaning, communicative function, and incongruity structure that make visual humor understandable to humans.

7 Challenges and Future Directions

Despite rapid progress in MLLMs on literal scene perception, a substantial gap remains between recognizing *what is depicted* and interpreting *what is meant* in multimodal visual humor. Closing this gap requires moving beyond physical description toward socio-cultural, rhetorical, and value-laden interpretation. We synthesize this gap into interconnected challenges and outline concrete directions for the community.

7.1 The Evaluation Crisis

Most benchmarks reduce humor understanding to multiple-choice questions (MCQs) or binary classification Hessel et al. (2023a); Hu et al. (2024); Yang et al. (2024). This is scalable but poorly matched to the phenomenon: discriminative accuracy is inflated by shortcut learning, and creative interpretation is inherently open-ended.

Future directions. A promising direction is *rubric-based generative evaluation*, where models produce free-form interpretations that are assessed along disentangled dimensions such as incongruity detection, target identification, and cultural grounding, rather than a single aggregate label Gunjal et al. (2025); Liang et al. (2025). This can be coupled with LLM-as-judge frameworks to operationalize humor-specific rubrics, rewarding plausible novel interpretations while penalizing unsupported or hallucinated reasoning Liu et al. (2025, 2023). Such a paradigm better aligns evaluation with the inherently open-ended and interpretive nature of humor understanding.

7.2 Social and Cultural Grounding

Creative artifacts depend on shared background knowledge, implicit norms, and audience beliefs that current MLLMs reason about poorly Jiang et al. (2025); Hu & Shu (2023); Chiu et al. (2025). First, current large models still lack theory-of-mind–like reasoning to infer whose perspective is expressed, what belief is being challenged, how an audience is meant to react, and often default to flat literal description instead Chen et al. (2025a). In addition, parametric knowledge has a fixed cutoff, so memes tied to breaking news or ephemeral trends become opaque Kasai et al. (2024), while training data remains predominantly English and Western, producing systematic blind spots on non-Western visual symbols and humor conventions Yu et al. (2025); Park et al. (2025).

Early efforts such as CHumor for Chinese humor He et al. (2024), Chinese multimodal sarcasm Gao et al. (2025b), and StandUp4AI for multilingual stand-up Barriere et al. (2025) broaden coverage, but still capture only a small slice of global humor traditions. Retrieval-augmented generation (RAG) Lewis et al. (2020) offers a promising path: retrieval-guided learning improves hateful-meme detection Mei et al. (2024) and contextualized meme explanation Sharma et al. (2023c), yet humor demands retrieval of not only topical knowledge but also the social norms and cultural contexts that make jokes intelligible.

Future directions. (i) *Multicultural, per-culture benchmarks* with annotations of the background knowledge each item requires, rather than a single aggregate accuracy; (ii) *humor-aware RAG*: retrieval pipelines that surface community norms, trending discourse, and event timelines alongside factual context, paired with periodic knowledge refresh to mitigate temporal decay; (iii) *explicit intent and stance modeling* (critique, mockery, self-deprecation) as a structured, trainable proxy for theory-of-mind reasoning.

7.3 Narrative Reasoning in Sequential Humor

Our cross-benchmark analysis (Section 6) shows that sequential, multi-panel humor exposes the widest model–human gap on panel-sequencing and temporal reordering Ryan et al. (2025); Wang et al. (2025). Unlike single-image humor, where the incongruity sits within one frame, multi-panel humor requires maintaining entity identity across panels, building an expectation from setup panels, and localizing the exact point where the narrative violates it. Current architectures largely process panels in isolation, missing these temporal and causal dependencies.

Future directions. (i) *Panel-aware architectures* that explicitly encode panel order and cross-frame entity co-reference, borrowing temporal-attention and state-tracking inductive biases from video understanding; (ii) *setup–punchline decomposition* as an explicit training objective rather than holistic pattern matching; (iii) *diagnostic benchmarks* isolating individual narrative skills (next-panel prediction, swapped-panel detection) for targeted evaluation beyond aggregate accuracy.

7.4 Toward Unified, Safe Systems

Recognition, interpretation, and generation are currently treated as independent tasks with separate pipelines, despite being deeply intertwined. A system that can explain why a meme is funny should, in principle, be better at generating one. This fragmentation is compounded by a safety problem specific to humor: the same non-literal mechanisms that make humor effective (incongruity, exaggeration, irony) are also what makes harmful content hard to detect. The Hateful Memes Challenge showed that neither unimodal classifier can reliably catch cross-modally-constructed hate Kiela et al. (2020), and follow-up work shows detecting *who* is targeted is as important as detecting *that* harm occurs Sharma et al. (2022a); Pramanick et al. (2021). As generation models improve, this becomes dual-use: the same system that writes witty captions can be steered toward harassment or stereotype-reinforcing content Weidinger et al. (2022); Wang & Lee (2024), and current VLMs remain vulnerable to adversarial meme-based attacks Lee et al. (2025). Overly conservative filters, in turn, risk censoring legitimate satire and self-deprecating humor — an unresolved safety–creativity trade-off Jha et al. (2024a). Separately, creative datasets are frequently scraped without creator consent, raising unresolved data-provenance and "right to style" concerns as models move toward style-imitative generation.

Future directions. (i) *Joint training across the capability hierarchy* (recognition, interpretation, generation) to test whether gains at one level transfer to the others, and *self-consistency checks* (e.g., a model that rates a meme funny but explains it blandly) as a diagnostic for genuine understanding; (ii) *context-sensitive, rhetoric-aware moderation*: safety classifiers conditioned on whether an artifact critiques versus promotes a harmful stance, paired with humor-specific red-teaming rather than keyword filtering; (iii) *provenance-aware training*: datasets and models that record content origin, licensing, and consent, auditable back to their training sources.

8 Conclusion

Multimodal visual humor challenges AI systems as its meaning extends beyond surface perception. Achieving robust understanding and generation therefore requires moving toward socially and rhetorically grounded reasoning. This survey introduces a capability-centric task hierarchy that clarifies how creative understanding progresses from recognition to interpretation and generation. We further outline future directions for meaningful, reliable, and responsible engagement with human-created media.

Limitations

Despite growing interest in multimodal humor, this survey has several limitations. First, most existing research and datasets focus on Western, internet-centric visual humor forms (e.g., memes, cartoons), leaving many cultural traditions and non-mainstream media underexplored. Second, as multimodal models evolve rapidly, some observations may not fully generalize to future architectures or training paradigms.

Ethical considerations

Understanding creative expression poses distinct ethical challenges. Creative media may potentially convey harmful or sensitive content implicitly through humor, irony, or symbolism, increasing the risk of misinterpretation, bias amplification, or over-censorship. Dataset bias and cultural imbalance further threaten fairness and robustness, particularly for marginalized communities. In addition, many creative datasets raise unresolved copyright and ownership concerns, especially as models increasingly transition from understanding to generation and style imitation. Addressing these issues requires context-aware evaluation, transparent dataset practices, and greater emphasis on interpretability and human oversight when deploying such systems.

References

Tariq Habib Afridi, Aftab Alam, Muhammad Numan Khan, Jawad Khan, and Young-Koo Lee. A multimodal memes classification: A survey and open research issues. In *The Proceedings of the Third International Conference on Smart City Applications*, pp. 1451–1466. Springer, 2020.

Siddhant Agarwal, Shivam Sharma, Preslav Nakov, and Tanmoy Chakraborty. Mememqa: multimodal question answering for memes via rationale-based inferencing. *arXiv preprint arXiv:2405.11215*, 2024.

Miriam Amin and Manuel Burghardt. A survey on approaches to computational humor generation. In *Proceedings of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pp. 29–41, 2020.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Ashwin Baluja. Text is not all you need: Multimodal prompting helps llms understand humor. In *Proceedings of the 1st Workshop on Computational Humor (CHum)*, pp. 9–17, 2025.

Kate Barnes, Tiernon R. Riesenmy, Minh Duc Trinh, Eli Lleshi, Nóra Balogh, and Roland Molontay. Dank or not? analyzing and predicting the popularity of memes on reddit. *Applied Network Science*, 6, 2020. URL https://api.semanticscholar.org/CorpusID:227227964.

Valentin Barriere, Nahuel Gomez, Leo Hemamou, Sofia Callejas, and Brian Ravenet. Standup4ai: A new multilingual dataset for humor detection in stand-up comedy videos. *arXiv preprint arXiv:2505.18903*, 2025.

Jerome Bruner. The narrative construction of reality. *Critical inquiry*, 18(1):1–21, 1991.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 2506–2515, 2019.

Tuhin Chakrabarty, Arkadiy Saakyan, Olivia Winn, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki, and Smaranda Muresan. I spy a metaphor: Large language models and diffusion models co-create visual metaphors. *arXiv preprint arXiv:2305.14724*, 2023.

Arjun Chandrasekaran, Ashwin K Vijayakumar, Stanislaw Antol, Mohit Bansal, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. We are humor beings: Understanding and predicting visual humor. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4603–4612, 2016.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024a.

Ruirui Chen, Weifeng Jiang, Chengwei Qin, and Cheston Tan. Theory of mind in large language models: Assessment and enhancement. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31539–31558, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1522. URL https://aclanthology.org/2025.acl-long.1522/.

Yule Chen, Yufan Ren, and Sabine Süsstrunk. Zooming into comics: Region-aware rl improves fine-grained comic understanding in vision-language models. *arXiv preprint arXiv:2511.06490*, 2025b.

Yuyan Chen, Songzhou Yan, Zhihong Zhu, Zhixu Li, and Yanghua Xiao. Xmecap: Meme caption generation with sub-image adaptability. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 3352–3361, 2024b.

Yuyan Chen, Yichen Yuan, Panjun Liu, Dayiheng Liu, Qinghao Guan, Mengfei Guo, Haiming Peng, Bang Liu, Zhixu Li, and Yanghua Xiao. Talk funny! a large-scale humor response dataset with chain-of-humor interpretation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17826–17834, 2024c.

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 *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25663–25701, 2025.

Jiwan Chung, Seungwon Lim, Jaehyun Jeon, Seungbeen Lee, and Youngjae Yu. Can visual language models resolve textual ambiguity with visual cues? let visual puns tell you! *arXiv preprint arXiv:2410.01023*, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Shafkat Farabi, Tharindu Ranasinghe, Diptesh Kanojia, Yu Kong, and Marcos Zampieri. A survey of multimodal sarcasm detection. *arXiv preprint arXiv:2410.18882*, 2024.

Giovannantonio Forabosco. Cognitive aspects of the humor process: The concept of incongruity. 1992.

Sonja K Foss. Theory of visual rhetoric. In *Handbook of visual communication*, pp. 163–174. Routledge, 2004.

Xiyuan Gao, Shekhar Nayak, and Matt Coler. Spoken in jest, detected in earnest: A systematic review of sarcasm recognition-multimodal fusion, challenges, and future prospects. *IEEE Transactions on Affective Computing*, 2025a.

Xiyuan Gao, Bruce Xiao Wang, Meiling Zhang, Shuming Huang, Zhu Li, Shekhar Nayak, and Matt Coler. A multimodal chinese dataset for cross-lingual sarcasm detection. In *Proc. Interspeech 2025*, pp. 3968–3972, 2025b.

Rahul Garg, Trilok Padhi, Hemang Jain, Ugur Kursuncu, and Ponnurangam Kumaraguru. Just kiddin': Knowledge infusion and distillation for detection of indecent memes. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 23067–23086, 2025.

Gérard Genette. *Narrative discourse: An essay in method*, volume 3. Cornell University Press, 1980.

Hexiang Gu, Qifan Yu, Saihui Hou, Zhiqin Fang, Huijia Wu, and Zhaofeng He. Mememind: A large-scale multimodal dataset with chain-of-thought reasoning for harmful meme detection. *arXiv preprint arXiv:2506.18919*, 2025.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains. *arXiv preprint arXiv:2507.17746*, 2025.

Diandian Guo, Cong Cao, Fangfang Yuan, Yanbing Liu, Guangjie Zeng, Xiaoyan Yu, Hao Peng, and Philip S Yu. Multi-view incongruity learning for multimodal sarcasm detection. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 1754–1766, 2025.

Md Kamrul Hasan, Sangwu Lee, Wasifur Rahman, Amir Zadeh, Rada Mihalcea, Louis-Philippe Morency, and Ehsan Hoque. Humor knowledge enriched transformer for understanding multimodal humor. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 12972–12980, 2021.

Md Kamrul Hasan, Md Saiful Islam, Sangwu Lee, Wasifur Rahman, Iftekhar Naim, Mohammed Ibrahim Khan, and Ehsan Hoque. Textmi: Textualize multimodal information for integrating non-verbal cues in pre-trained language models. *arXiv preprint arXiv:2303.15430*, 2023.

Ruiqi He, Yushu He, Longju Bai, Jiarui Liu, Zhenjie Sun, Zenghao Tang, He Wang, Hanchen Xia, and Naihao Deng. Chumor 1.0: A truly funny and challenging chinese humor understanding dataset from ruo zhi ba. *arXiv preprint arXiv:2406.12754*, 2024.

Ming Shan Hee, Wen-Haw Chong, and Ka-Wei Roy Lee. Decoding the underlying meaning of multimodal hateful memes. In *32nd International Joint Conference on Artificial Intelligence (IJCAI 2023)*. International Joint Conferences on Artifical Intelligence (IJCAI), 2023.

Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 688–714, Toronto, Canada, July 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.41. URL https://aclanthology.org/2023.acl-long.41/.

Jack Hessel, Ana Marasović, Jena D Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 688–714, 2023b.

Eftekhar Hossain, Omar Sharif, Mohammed Moshiul Hoque, and Sarah Masud Preum. Deciphering hate: identifying hateful memes and their targets. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8347–8359, 2024.

Zhe Hu, Tuo Liang, Jing Li, Yiren Lu, Yunlai Zhou, Yiran Qiao, Jing Ma, and Yu Yin. Cracking the code of juxtaposition: Can ai models understand the humorous contradictions. *Advances in Neural Information Processing Systems*, 37:47166–47188, 2024.

Zhiting Hu and Tianmin Shu. Language models, agent models, and world models: The law for machine reasoning and planning. *arXiv preprint arXiv:2312.05230*, 2023.

EunJeong Hwang and Vered Shwartz. Memecap: A dataset for captioning and interpreting memes. *arXiv preprint arXiv:2305.13703*, 2023.

EunJeong Hwang, Peter West, and Vered Shwartz. Bottlehumor: Self-informed humor explanation using the information bottleneck principle. *arXiv preprint arXiv:2502.18331*, 2025.

Hikaru Ikuta, Leslie Wohler, and Kiyoharu Aizawa. Mangaub: A manga understanding benchmark for large multimodal models. *IEEE MultiMedia*, 2025.

Vedaant V Jain, Gabriel Kreiman, and Felipe dos Santos Alves Feitosa. Humordb: Can ai understand graphical humor? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 604–613, 2025.

Prince Jha, Raghav Jain, Konika Mandal, Aman Chadha, Sriparna Saha, and Pushpak Bhattacharyya. Memeguard: An llm and vlm-based framework for advancing content moderation via meme intervention. *arXiv preprint arXiv:2406.05344*, 2024a.

Prince Jha, Krishanu Maity, Raghav Jain, Apoorv Verma, Sriparna Saha, and Pushpak Bhattacharyya. Meme-ingful analysis: Enhanced understanding of cyberbullying in memes through multimodal explanations. *arXiv preprint arXiv:2401.09899*, 2024b.

Zhiwei Jia, Pradyumna Narayana, Arjun Akula, Garima Pruthi, Hao Su, Sugato Basu, and Varun Jampani. Kafa: Rethinking image ad understanding with knowledge-augmented feature adaptation of vision-language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pp. 772–785, 2023.

Liwei Jiang, Taylor Sorensen, Sydney Levine, and Yejin Choi. Can language models reason about individualistic human values and preferences? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6757–6794, 2025.

Antonios Kalloniatis and Panagiotis Adamidis. Computational humor recognition: a systematic literature review. *Artificial Intelligence Review*, 58(2):43, 2024.

Hasindu Kariyawasam, Amashi Niwarthana, Alister Palmer, Judy Kay, and Anusha Withana. Appropriate incongruity driven human-ai collaborative tool to assist novices in humorous content generation. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, pp. 650–659, 2024.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asli, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, and Kentaro Inui. Realtime qa: What's the answer right now? *Advances in Neural Information Processing Systems*, 36, 2024.

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–a benchmark dataset and method. *arXiv preprint arXiv:2509.06771*, 2025.

Anas Anwarul Haq Khan, Tanik Saikh, Arpan Phukan, and Asif Ekbal. Hope 'the paragraph guy'explains the rest: Introducing mesum, the meme summarizer. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 6654–6668, 2024.

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. *Advances in neural information processing systems*, 33:2611–2624, 2020.

Roger J Kreuz, Richard M Roberts, Brenda K Johnson, and Eugenie L Bertus. Figurative language occurence and co-occurence in contemporary literature. *Advances in Discourse Processes*, 52:83–98, 1996.

Gitanjali Kumari, Jitendra Solanki, and Asif Ekbal. Memedetoxnet: Balancing toxicity reduction and context preservation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 25076–25098, 2025.

Manishit Kundu, Sumit Shekhar, and Pushpak Bhattacharyya. Looking beyond the pixels: Evaluating visual metaphor understanding in vlms. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 23137–23158, 2025.

George Lakoff and Mark Johnson. *Metaphors we live by*. University of Chicago press, 2024.

DongGeon Lee, Joonwon Jang, Jihae Jeong, and Hwanjo Yu. Are vision-language models safe in the wild? a meme-based benchmark study. *arXiv preprint arXiv:2505.15389*, 2025.

Roy Ka-Wei Lee, Rui Cao, Ziqing Fan, Jing Jiang, and Wen-Haw Chong. Disentangling hate in online memes. In *Proceedings of the 29th ACM international conference on multimedia*, pp. 5138–5147, 2021.

Jens Lemmens and Victor De Marez. Computational humor modeling: A survey on the state of the art. *ACM Computing Surveys*, 58(7):1–37, 2026.

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. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

Runjia Li, Shuyang Sun, Mohamed Elhoseiny, and Philip Torr. Oxfordtvg-hic: Can machine make humorous captions from images? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20293–20303, 2023.

Bin Liang, Chenwei Lou, Xiang Li, Min Yang, Lin Gui, Yulan He, Wenjie Pei, and Ruifeng Xu. Multi-modal sarcasm detection via cross-modal graph convolutional network. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pp. 1767–1777. Association for Computational Linguistics, 2022.

Tuo Liang, Zhe Hu, Jing Li, Hao Zhang, Yiren Lu, Yunlai Zhou, Yiran Qiao, Disheng Liu, Jeirui Peng, Jing Ma, et al. When'yes'meets' but': Can large models comprehend contradictory humor through comparative reasoning? *arXiv preprint arXiv:2503.23137*, 2025.

Hongzhan Lin, Ziyang Luo, Jing Ma, and Long Chen. Beneath the surface: Unveiling harmful memes with multimodal reasoning distilled from large language models. *arXiv preprint arXiv:2312.05434*, 2023.

Hui Liu, Wenya Wang, and Haoliang Li. Towards multi-modal sarcasm detection via hierarchical congruity modeling with knowledge enhancement. *arXiv preprint arXiv:2210.03501*, 2022.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*, 2023.

Zijun Liu, Peiyi Wang, Runxin Xu, Shirong Ma, Chong Ruan, Peng Li, Yang Liu, and Yu Wu. Inference-time scaling for generalist reward modeling. *arXiv preprint arXiv:2504.02495*, 2025.

Tyler Loakman, William Thorne, and Chenghua Lin. Who's laughing now? an overview of computational humour generation and explanation. In *Proceedings of the 18th International Natural Language Generation Conference*, pp. 780–794, 2025.

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

Jiří Martínek, Pavel Král, Ladislav Lenc, and Josef Baloun. Comicorda: Dialogue act recognition in comic books. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 3566–3578, 2024.

Jingbiao Mei, Jinghong Chen, Weizhe Lin, Bill Byrne, and Marcus Tomalin. Improving hateful meme detection through retrieval-guided contrastive learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5333–5347, 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. *arXiv preprint arXiv:2409.13592*, 2024.

Reuben Narad, Siddharth Suresh, Jiayi Chen, Pine SL Dysart-Bricken, Bob Mankoff, Robert Nowak, Jifan Zhang, and Lalit Jain. Which llms get the joke? probing non-stem reasoning abilities with humorbench. *arXiv preprint arXiv:2507.21476*, 2025.

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 *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18926–18946, 2025.

Xuan Ouyang, Senan Wang, Bouzhou Wang, Siyuan Xiahou, Jinrong Zhou, and Yuekang Li. Laugh, relate, engage: Stylized comment generation for short videos. *arXiv preprint arXiv:2511.03757*, 2025.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

ChaeHun Park, Yujin Baek, Jaeseok Kim, Yu-Jung Heo, Du-Seong Chang, and Jaegul Choo. Evaluating visual and cultural interpretation: The k-viscuit benchmark with human-vlm collaboration. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21960–21974, 2025.

Sandro Paval, Pascal Meißner, and Ivan P. Yamshchikov. ComicScene154: A scene dataset for comic analysis. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 31562–31568, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1608. URL https://aclanthology.org/2025.emnlp-main.1608/.

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. *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 4439–4455, 2021.

Zeju Qiu, Weiyang Liu, Haiwen Feng, Zhen Liu, Tim Z Xiao, Katherine M Collins, Joshua B Tenenbaum, Adrian Weller, Michael J Black, and Bernhard Schölkopf. Can large language models understand symbolic graphics programs? *arXiv preprint arXiv:2408.08313*, 2024.

Elisabeth El Refaie. Understanding visual metaphor: The example of newspaper cartoons. *Visual communication*, 2 (1):75–95, 2003.

Chengjuan Ren, Dongwon Jeong, Ming Wu, Yi Huang, Yuhan Gao, and Yuejia Li. A survey of multimodal hate meme detection. *Expert Systems with Applications*, pp. 132507, 2026.

Yuriel Ryan, Rui Yang Tan, Kenny Tsu Wei Choo, and Roy Ka-Wei Lee. Humor in pixels: Benchmarking large multimodal models understanding of online comics. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 14024–14050, 2025.

Arkadiy Saakyan, Shreyas Kulkarni, Tuhin Chakrabarty, and Smaranda Muresan. V-flute: Visual figurative language understanding with textual explanations. *CoRR*, 2024.

Arkadiy Saakyan, Shreyas Kulkarni, Tuhin Chakrabarty, and Smaranda Muresan. Understanding figurative meaning through explainable visual entailment. 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. 1–23, 2025.

Rossano Schifanella, Paloma De Juan, Joel Tetreault, and Liangliang Cao. Detecting sarcasm in multimodal social platforms. In *Proceedings of the 24th ACM international conference on Multimedia*, pp. 1136–1145, 2016.

Siddhant Bikram Shah, Shuvam Shiwakoti, Maheep Chaudhary, and Haohan Wang. Memeclip: Leveraging clip representations for multimodal meme classification. *arXiv preprint arXiv:2409.14703*, 2024.

Dafna Shahaf, Eric Horvitz, and Robert Mankoff. Inside jokes: Identifying humorous cartoon captions. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1065–1074, 2015.

Hassan Shahmohammadi, Adhiraj Ghosh, and Hendrik Lensch. Vipe: Visualise pretty-much everything. *arXiv preprint arXiv:2310.10543*, 2023.

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas Pykl, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Bjorn Gamback. Semeval-2020 task 8: Memotion analysis–the visuo-lingual metaphor! *arXiv preprint arXiv:2008.03781*, 2020.

Shivam Sharma, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. Disarm: Detecting the victims targeted by harmful memes. *arXiv preprint arXiv:2205.05738*, 2022a.

Shivam Sharma, Firoj Alam, Md Shad Akhtar, Dimitar Dimitrov, Giovanni Da San Martino, Hamed Firooz, Alon Halevy, Fabrizio Silvestri, Preslav Nakov, and Tanmoy Chakraborty. Detecting and understanding harmful memes: A survey. *arXiv preprint arXiv:2205.04274*, 2022b.

Shivam Sharma, Siddhant Agarwal, Tharun Suresh, Preslav Nakov, Md Shad Akhtar, and Tanmoy Chakraborty. What do you meme? generating explanations for visual semantic role labelling in memes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 9763–9771, 2023a.

Shivam Sharma, Atharva Kulkarni, Tharun Suresh, Himanshi Mathur, Preslav Nakov, Md Shad Akhtar, and Tanmoy Chakraborty. Characterizing the entities in harmful memes: Who is the hero, the villain, the victim? *arXiv preprint arXiv:2301.11219*, 2023b.

Shivam Sharma, S Ramaneswaran, Udit Arora, Md Shad Akhtar, and Tanmoy Chakraborty. Memex: Detecting explanatory evidence for memes via knowledge-enriched contextualization. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 5272–5290, 2023c.

Limor Shifman. *Memes in digital culture*. MIT press, 2013.

Kohtaro Tanaka, Kohei Uehara, Lin Gu, Yusuke Mukuta, and Tatsuya Harada. Content-specific humorous image captioning using incongruity resolution chain-of-thought. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2348–2367, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.152. URL https://aclanthology.org/2024.findings-naacl.152/.

Binghao Tang, Boda Lin, Haolong Yan, and Si Li. Leveraging generative large language models with visual instruction and demonstration retrieval for multimodal sarcasm detection. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1732–1742, 2024.

Alexey Tikhonov and Pavel Shtykovskiy. Humor mechanics: Advancing humor generation with multistep reasoning. *arXiv preprint arXiv:2405.07280*, 2024.

Tony Veale. Incongruity in humor: Root cause or epiphenomenon? 2004.

Emanuele Vivoli, Artemis Llabrés, Mohamed Ali Souibgui, Marco Bertini, Ernest Valveny Llobet, and Dimosthenis Karatzas. Comicspap: understanding comic strips by picking the correct panel. In *International Conference on Document Analysis and Recognition*, pp. 337–350. Springer, 2025.

Han Wang and Roy Ka-Wei Lee. Memecraft: Contextual and stance-driven multimodal meme generation. In *Proceedings of the ACM Web Conference 2024*, pp. 4642–4652, 2024.

Han Wang, Yilin Zhao, Dian Li, Xiaohan Wang, Gang Liu, Xuguang Lan, and Hui Wang. Innovative thinking, infinite humor: Humor research of large language models through structured thought leaps. *arXiv preprint arXiv:2410.10370*, 2024a.

Jiquan Wang, Lin Sun, Yi Liu, Meizhi Shao, and Zengwei Zheng. Multimodal sarcasm target identification in tweets. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8164–8175, 2022.

Xiaochen Wang, Heming Xia, Jialin Song, Longyu Guan, Qingxiu Dong, Rui Li, Yixin Yang, Yifan Pu, Weiyao Luo, Yiru Wang, Xiangdi Meng, Wenjie Li, and Zhifang Sui. Beyond single frames: Can LMMs comprehend implicit narratives in comic strip? In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 6436–6452, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.342. URL https://aclanthology.org/2025.findings-emnlp.342/.

Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Fuxiao Liu, Gedas Bertasius, et al. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 416–442, 2024b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 214–229, 2022.

Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, et al. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models. *arXiv preprint arXiv:2504.15279*, 2025.

Shweta Yadav, Cornelia Caragea, Chenye Zhao, Naincy Kumari, Marvin Solberg, and Tanmay Sharma. Towards identifying fine-grained depression symptoms from memes. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8890–8905, 2023.

Yixin Yang, Zheng Li, Qingxiu Dong, Heming Xia, and Zhifang Sui. Can large multimodal models uncover deep semantics behind images? In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 1898–1912, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.113. URL https://aclanthology.org/2024.findings-acl.113/.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 2024.

Haofei Yu, Zhengyang Qi, Lawrence Keunho Jang, Russ Salakhutdinov, Louis-Philippe Morency, and Paul Pu Liang. Mmoe: Enhancing multimodal models with mixtures of multimodal interaction experts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10006–10030, 2024.

Haorui Yu, Yang Zhao, Yijia Chu, and Qiufeng Yi. Seeing symbols, missing cultures: Probing vision-language models' reasoning on fire imagery and cultural meaning. In *Proceedings of the 9th Widening NLP Workshop*, pp. 1–8, 2025.

Huixuan Zhang and Xiaojun Wan. Image matters: A new dataset and empirical study for multimodal hyperbole detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 8652–8661, 2024.

Jiajun Zhang, Shijia Luo, Ruikang Zhang, and Qi Su. Humorchain: Theory-guided multi-stage reasoning for interpretable multimodal humor generation. *arXiv preprint arXiv:2511.21732*, 2025.

Jifan Zhang, Lalit Jain, Yang Guo, Jiayi Chen, Kuan Zhou, Siddharth Suresh, Andrew Wagenmaker, Scott Sievert, Timothy T Rogers, Kevin G Jamieson, et al. Humor in ai: Massive scale crowd-sourced preferences and benchmarks for cartoon captioning. *Advances in Neural Information Processing Systems*, 37:125264–125286, 2024a.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

Xiaoqiang Zhang, Ying Chen, and Guangyuan Li. Multi-modal sarcasm detection based on contrastive attention mechanism. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pp. 822–833. Springer, 2021.

Xinnong Zhang, Haoyu Kuang, Xinyi Mou, Hanjia Lyu, Kun Wu, Siming Chen, Jiebo Luo, Xuan-Jing Huang, and Zhongyu Wei. Somelvlm: A large vision language model for social media processing. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2366–2389, 2024b.

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. *arXiv preprint arXiv:2505.17433*, 2025.

Shanshan Zhong, Zhongzhan Huang, Shanghua Gao, Wushao Wen, Liang Lin, Marinka Zitnik, and Pan Zhou. Let's think outside the box: Exploring leap-of-thought in large language models with creative humor generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13246–13257, 2024.

Naitian Zhou, David Jurgens, and David Bamman. Social meme-ing: Measuring linguistic variation in memes. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3005–3024, 2024.

A Evaluation Details

We evaluate all models in a zero-shot setting. For each benchmark, we retain the task definition, prompt format, answer format, evaluation split, and scoring procedure reported in the corresponding original paper. Except for GPT-4o, all evaluated models use publicly available checkpoints hosted on Hugging Face. For GPT-4o, we use the fixed API version `gpt-4o-2024-05-13`. All experiments were completed by May 29, 2026.

Each model is queried once per question. For open-source models, decoding is performed with `do_sample=true`; all remaining benchmark-specific generation and evaluation settings follow the corresponding original papers. When an official evaluation split is available, we use it directly; otherwise, we follow the split or sampling procedure described in the benchmark paper. Because prompts and evaluation protocols differ across benchmarks by design, the reported results are intended primarily for within-benchmark model comparison and descriptive cross-benchmark analysis, rather than for a strictly controlled comparison of task difficulty. Since each question is evaluated with a single sampled generation, small score differences may be affected by decoding stochasticity and should not be interpreted as statistically significant.

B Overview of Datasets and Benchmarks

Here we provide a comprehensive tabular overview of datasets and benchmarks studied in this survey, organized according to the capability hierarchy introduced in the main paper.

C Task Specific Models Before MLLM Era

Prior to MLLMs, multimodal visual humor understanding was dominated by task-specific discriminative architectures tightly coupled with individual tasks such as sarcasm, humor, or metaphor detection.

Early models focused on effective fusion mechanisms for heterogeneous features. Schifanella et al. (2016) first incorporated visual cues into sarcasm detection via separate encoders and concatenation, while Cai et al. (2019) showed hierarchical fusion of text, images, and attributes better captures cross-modal interactions. Zhang et al. (2021) introduced contrastive attention to explicitly model inter-modal incongruity for finer detection of cross-modal discrepancies. These approaches improved reasoning on specific tasks, but they also introduced stronger task assumptions and often depended on curated knowledge sources or structures.

Beyond fusion design, later work emphasized implicit knowledge and commonsense. HKT Hasan et al. (2021) injected humor-related knowledge into Transformer architecture for deeper incongruity modeling. Lee et al. (2021) and Liang et al. (2022) leveraged object-level visual representations to construct richer contextual embeddings and cross-modal graphs, facilitating localized reasoning over visual–textual conflicts. Liu et al. (2022) incorporated external commonsense and semantic knowledge into hierarchical congruity modeling, improving implicit intent interpretation.

Overall, non-MLLM approaches made important progress by improving fusion design, structured representations, and knowledge injection. However, their gains were typically task-dependent, and their scalability and cross-domain generalization remained limited, as also noted in prior surveys Sharma et al. (2022b); Farabi et al. (2024). This limitation motivates the later shift toward more general multimodal foundation-model paradigms.

Table 4 Overview of multimodal visual humor datasets focused on **Recognition** tasks. Recognition datasets typically use binary labels (e.g., humorous vs. non-humorous), element-level labels (e.g., punchlines or rhetorical roles), and intensity scores (e.g., degree of funniness or offensiveness). In Data Forms, StaVT denotes Static Visual–Textual Artifacts; SeqVN denotes Sequential Visual Narratives. In Mechanism, Multi denotes datasets that involve multiple mechanisms defined in Sec. 2. The Availability column provides links to publicly accessible datasets, and "N/A" indicates unpublished datasets.

Dataset	Venue	Data Forms	Mechanism	Size	Avail.
StaVT: Meme					
Goal: Humor & Entertainment, Satire & Social Critique					
D-HUMOR Kasu et al. (2025)	ICDM'25	StaVT: Meme	Multi	4,379	Link
MemeMind Gu et al. (2025)	Arxiv'25	StaVT: Meme	Multi	43,223	*N/A*
TOXICN MM Lu et al. (2024)	NeurIPS'2	StaVT: Meme	Multi	12K	Link
PrideMM Shah et al. (2024)	EMNLP'2	StaVT: Meme	Multi	5,063	Link
BHM Hossain et al. (2024)	ACL'24	StaVT: Meme	Multi	7,148	*N/A*
Ext-Harm-P Sharma et al. (2022a)	NAACL'2	StaVT: Meme	Multi	4,446	Link
HVVMemes Sharma et al. (2023b)	EACL'23	StaVT: Meme	Multi	6,933	Link
RESTORE Yadav et al. (2023)	ACL'23	StaVT: Meme	Multi	4,664	Link
Dank or not? Barnes et al. (2020)	App. Net. Sci.'21	StaVT: Meme	Multi	70K	*N/A*
The Hateful Memes Challenge Set Kiela et al. (2020)	NeurIPS'2	StaVT: Meme	Multi	10K	Link
StaVT: Sarcastic Image					
Goal: Satire & Social Critique					
RedEval Tang et al. (2024)	NAACL'24	StaVT: Sarcastic Image	Sarcasm	1,004	Link
SPMSD Guo et al. (2025)	COLING'25	StaVT: Sarcastic Image	Sarcasm	1K	*N/A*
MSTI dataset Wang et al. (2022)	ACL'22	StaVT: Sarcastic Image	Sarcasm	5,015	Link
Multi-Modal Sarcasm Detection in Twitter Cai et al. (2019)	ACL'19	StaVT: Sarcastic Image	Sarcasm	24,635	*N/A*
Sarcasm in Multimodal Social Platforms Schifanella et al. (2016)	ACM MM'16	StaVT: Sarcastic Image	Sarcasm	10K	*N/A*
StaVT: Humorous Image / Cartoon					
Goal: Humor & Entertainment					
HumorDB Jain et al. (2025)	ICCV'25	StaVT: Humorous Image	Multi	3,542	Link
Inside Jokes Shahaf et al. (2015)	KDD'15	StaVT: Cartoon	Multi	76,928	*N/A*
SeqVN: Comic Strip					
Goal: Humor & Entertainment					
COMICORDA Martínek et al. (2024)	COLING'25	SeqVN: Comic Strip	Narrative	1,438	*N/A*
AVH & FOR Chandrasekaran et al. (2016)	CVPR'16	SeqVN: Comic Strip	Narrative	7,150	Link

Table 5 Overview of multimodal visual humor datasets focused on **Understanding** and **Generation**. These datasets typically pair multimodal inputs with explanations or rationales—often augmented with contextual or external knowledge—to justify intended meaning (e.g., humor, irony, or satire), alongside target outputs that support coherent and controllable generation (e.g., memes or comics with captions/rationales). StaVT denotes Static Visual–Textual Artifacts; SeqVN denotes Sequential Visual Narratives. In Mechanism, Multi denotes datasets that involve multiple mechanisms defined in Sec. 2. The Availability column provides links to publicly accessible datasets, and "N/A" indicates unpublished datasets.

Dataset	Venue	Mechanism	Size	Avail.
StaVT: Meme / ***Goal:*** *Humor & Entertainment, Satire & Social Critique*				
MemeReaCon Zhao et al. (2025)	EMNLP'25	Multi	1,565	*N/A*
MEMESAFETY-BENCH Lee et al. (2025)	EMNLP'25	Multi	50,430	Link
MemeMind Gu et al. (2025)	Arxiv'25	Multi	43,223	*N/A*
MemeQA Nguyen et al. (2025)	ACL'25	Multi	9K	Link
SEMANTICMEMES Zhou et al. (2024)	NAACL'24	Multi	3.8M	Link
MMD Khan et al. (2024)	EMNLP-Fdgs'24	Multi	13,494	Link
MultiBully-Ex Jha et al. (2024b)	EACL'24	Multi	5,854	Link
Oogiri-GO Zhong et al. (2024)	CVPR'24	Multi	130K	Link
MemeMQACorpus Agarwal et al. (2024)	ACL-Fdgs'24	Multi	1,880	*N/A*
ICMM Jha et al. (2024a)	ACL'24	Multi	1K	Link
OxfordTVG-HIC Li et al. (2023)	ICCV'23	Multi	2.9M	Link
MEMECAP Hwang & Shwartz (2023)	EMNLP'23	Multi	6,300	Link
MCC Sharma et al. (2023c)	ACL'23	Multi	3.4K	Link
HatReD Hee et al. (2023)	IJCAI'24	Multi	3,228	Link
ExHVV Sharma et al. (2023a)	AAAI'22	Multi	3K	Link
StaVT: Cartoon / ***Goal:*** *Humor & Entertainment*				
HumorBench Narad et al. (2025)	Arxiv'25	Multi	300	*N/A*
Humor in AI Zhang et al. (2024a)	NeurIPS'24	Incongruity	2.2M	Link
Do Androids Laugh at Electric Sheep? Hessel et al. (2023b)	ACL'23	Multi	24,048	*N/A*
StaVT: Social Media Image / ***Goal:*** *Humor & Entertainment, Satire & Social Critique, Emotion & Aesthetic Experience*				
V-FLUTE Saakyan et al. (2025)	NAACL'25	Multi	6,027	Link
SoMeLVLM Zhang et al. (2024b)	ACL'24	Multi	653.8K	Link
StaVT: Humorous Image / ***Goal:*** *Humor & Entertainment*				
HumorDB Jain et al. (2025)	ICCV'25	Multi	3,542	Link
VisualPun_UNPIE Chung et al. (2024)	EMNLP'24	Incongruity	1K	Link
SeqVN: Comic Strip / ***Goal:*** *Humor & Entertainment*				
MangaUB Ikuta et al. (2025)	IEEE MM'25	Narrative	18,179	Link
AI4VA-FG Chen et al. (2025b)	Arxiv'25	Narrative	16,264	*N/A*
PixelHumor Ryan et al. (2025)	EMNLP-Fdgs'25	Multi	2.8K	Link
YesBut-v2 Liang et al. (2025)	Arxiv'25	Multi	1,262	Link
YesBut Hu et al. (2024)	NeurIPS'24	Multi	348	Link
YesBut (synth. 3D stick) Nandy et al. (2024)	EMNLP'24	Multi	2,547 (syn.)	Link
ComicsPAP Vivoli et al. (2025)	Arxiv'25	Narrative	103,933	Link
SeqVN: Meme / ***Goal:*** *Humor & Entertainment*				
XMeCap Chen et al. (2024b)	ACM MM'24	Multi	12,320	*N/A*