Do AI Systems Understand What We Mean? A Survey of Multimodal Human Creative Expression Understanding

Anonymous ACL submission

Abstract

Multimodal human creativity, such as memes, cartoons, comics, advertisements, and humorous or satirical videos, presents unique challenges for AI systems due to its non-literal, culturally grounded, and rhetorically structured nature. While Multimodal Large Language Models (MLLMs) excel at physical-world understanding, they continue to struggle with creative communication. This survey provides a systematic overview of multimodal human creativity research. We first introduce a unified taxonomy that characterizes creative content along data forms, meaning-making mechanisms, and communicative goals. We then organize existing work into a capability-oriented hierarchy, spanning recognition, interpretation, and generation. Within this framework, we audit the architectural shifts that have shaped the transition from task-specific models to MLLMs. We conclude by identifying critical benchmark trends and open socio-technical challenges, providing a roadmap toward AI systems capable of sophisticated creative understanding.

1 Introduction

Recent advances in Multimodal Large Language Models (MLLMs) have demonstrated strong capabilities in physical-world grounding, such as scene understanding, object recognition, and factual visual question answering (Yin et al., 2024; Comanici et al., 2025; Bai et al., 2025). These successes have motivated expectations that MLLMs can move beyond literal perception to comprehend more abstract and socially grounded forms of human communication. In particular, human creative expression, which dominates contemporary online communication, appears to be a natural next frontier for multimodal AI. However, growing evidence shows that even state-of-the-art MLLMs struggle to interpret such content reliably (Jia et al., 2024; Saakyan et al., 2024; Joshi, 2025).

We define **multimodal human creative expression** as the strategic synthesis of heterogeneous signals (e.g., linguistic, visual, auditory, and temporal cues) to convey meaning that is intentionally *non-literal* and *context-dependent* (Boden, 1998; Wiggins, 2006; Colton and Wiggins, 2012). Common examples include comics, memes, editorial cartoons, and satirical videos. Unlike factual communication, these artifacts use rhetorical mechanisms like metaphor, irony, and incongruity to create a "semantic gap" between the literal depiction and the intended message.

Decoding these artifacts is challenging for MLLMs because their intended meaning is rarely recoverable from surface-level perception alone. While a model may accurately recognize objects, actions, or scenes at the perceptual level, it often fails to infer their communicative function. For instance, an MLLM may accurately identify a "burning house" in an image (physical perception), yet fail to resolve its function as a satirical metaphor for "climate policy" (creative interpretation). Thus, achieving a deep understanding of such content requires more than multimodal alignment; it requires inferring implicit intent and grounding its interpretation in cultural norms and socio-historical context (Saakyan et al., 2024; Ryan et al., 2025).

Despite the increasing interest, existing surveys on human creativity understanding remain fragmented across data, task, and modeling dimensions. From a task perspective, prior reviews often focus on isolated objectives such as detection, understanding (Sharma et al., 2022b; Farabi et al., 2024), or generation (Amin and Burghardt, 2020; Loakman et al., 2023), without capturing the full capability progression from recognition to interpretive reasoning and creativity synthesis. From a data perspective, surveys that consider multimodal inputs typically concentrate on a single content type, such as memes (Farabi et al., 2024), humor (Ren et al., 2024; Kalloniatis and Adamidis, 2024), or

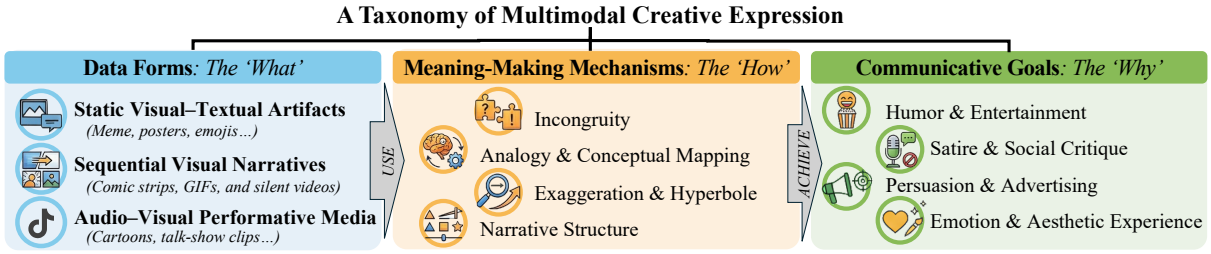

Figure 1: A unified data taxonomy of multimodal human creative expression.

text-only creativity (Amin and Burghardt, 2020; Su et al., 2025), overlooking the diversity of multimodal and **socially-grounded** inputs in real-world settings. From a modeling perspective, existing work (Nguyen and Ng, 2024; Loakman et al., 2025) tends to explore modeling techniques in a piecemeal manner, rather than providing a systematic and diverse discussion of multimodal alignment and reasoning mechanisms.

We argue that progress in multimodal creative AI is constrained not only by individual advances in data or models, but also by the lack of unified representations, capability-aware evaluation, principled synthesis across data, tasks, and models, and ethical grounding across the research pipeline. We address the limitation by systematically examining the problem from the following perspectives:

- **Unified Taxonomy:** We propose a unified taxonomy that organizes multimodal human creativity along *data forms*, *meaning-making mechanisms*, and *communicative goals*, connecting representation, interpretation, and creative intent beyond task- or model-centric views.
- **Capability-Centric Tasks:** We introduce a three-level task hierarchy—*Recognition*, *Interpretation and Reasoning*, and *Creative Generation*—that clarifies the capabilities required across datasets.
- **Capability-Guided Models:** We synthesize modeling paradigms according to the capabilities they support, highlighting how progress emerges from integrating perceptual alignment, reasoning, and grounded knowledge.
- **Benchmarks and Ethical Frontiers:** We bridge our taxonomy to existing benchmarks, and outline open socio-technical challenges to guide future creative AI development.

2 Background: A Taxonomy of Multimodal Creative Expression

To systematically characterize multimodal human creative expression, we propose a three-dimensional taxonomy organized around three fundamental questions: (1) **Data Forms**—*what* media the content is instantiated in; (2) **Meaning-Making Mechanisms**—*how* non-literal meaning is constructed; and (3) **Communicative Goals**—*why* the content is produced, as shown in Figure 1.

These dimensions capture complementary aspects of creative artifacts: their representational substrate, their underlying rhetorical and cognitive operations, and their pragmatic intent. Together, they provide a principled framework grounded in theories of multimodal communication and rhetoric (Foss, 2017; Lakoff and Johnson, 2024), while aligning naturally with contemporary AI research on multimodal understanding.

2.1 Data Forms: The "What"

Human creative expression spans a wide range of data forms, including text, images, audio, video, and their combinations. While early computational work focused primarily on textual humor and figurative language (Mihalcea and Strapparava, 2005), modern creative communication is overwhelmingly multimodal. Understanding such content requires models to integrate perceptual signals with cultural knowledge and rhetorical intent, going beyond literal semantic interpretation. We group creative data forms into three broad categories:

Static Visual–Textual Artifacts include memes, posters, and emojis. These artifacts typically consist of a single image, often paired with minimal text, to convey humor, stance, or affect. Interpretation depends on visual grounding, symbolic association, and shared socio-cultural context (Shifman, 2013; Sharma et al., 2020).

Sequential Visual Narratives encompass comic strips, GIFs, and silent videos, where meaning emerges from temporal progression across frames. 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., 2025b).

Audio–Visual Performative Media include cartoons, talk-show clips, and short-form videos on platforms such as TikTok or Reels. These formats

incorporate speech, prosody, facial expressions, music, and editing rhythm, making interpretation contingent on fine-grained temporal alignment and pragmatic inference (Marone, 2016).

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

2.2 Meaning-Making Mechanisms: The "How"

Creative meaning often arises from rhetorical and cognitive mechanisms that introduce ambiguity, surprise, or abstraction. We highlight four recurrent mechanisms that cut across data forms.

Incongruity refers to a mismatch between expectation and observation and underlies many forms of humor, irony, and sarcasm (Forabosco, 1992; Veale, 2004). In multimodal settings, incongruity often emerges from cross-modal clashes—e.g., a benign image paired with an unexpected caption—requiring models to reconcile conflicting cues (Schifanella et al., 2016; Farabi et al., 2024).

Analogy and Conceptual Mapping construct meaning by projecting structure from a familiar source domain onto a target concept. This mechanism unifies *metaphor* and *symbolism* (Lakoff and Johnson, 2024), where images, language, or other modalities jointly instantiate cross-domain correspondences that must be interpreted non-literally (Refaie, 2003; Foss, 2004).

Exaggeration and Hyperbole amplify attributes for comic or emphatic effect. In multimodal artifacts, exaggeration often arises through visual distortion, intensified motion, or prosodic emphasis, with meaning emerging from cross-modal amplification rather than any single signal (Kreuz et al., 1996; Zhang and Wan, 2024).

Narrative Structure operates primarily in sequential media, where meaning is shaped by setups, payoffs, and causal relations across time (Genette, 1980; Bruner, 1991). Understanding such content requires identifying how events and modalities jointly signal story progression (Paval et al., 2025).

These mechanisms explain how creative artifacts encode meaning beyond surface semantics, posing challenges distinct from literal understanding.

2.3 Communicative Goals: The "Why"

The third dimension concerns communicative intent—the purpose a creative artifact serves for its audience. This pragmatic layer shapes how both data forms and mechanisms should be interpreted. We identify four recurring goals.

Humor and Entertainment aim to amuse or delight. Beyond recognizing incongruity, models must infer humorous intent to distinguish jokes from errors or misinformation (Raskin, 1979; Hasan et al., 2019a; Hu et al., 2024).

Satire and Social Critique use irony or absurdity to comment on political or social issues. Correctly identifying satirical intent is crucial for avoiding literal misinterpretation (Burfoot and Baldwin, 2009; Li et al., 2020; Nandy et al., 2024).

Persuasion and Advertising seek to influence beliefs or behavior rather than entertain. Creative ads and propaganda memes employ metaphor, emotional framing, and stylization to persuade audiences (Forceville, 2002; Kumar et al., 2023; Wang et al., 2025a).

Emotion and Aesthetic Experience aim to evoke affective or aesthetic responses, such as nostalgia, empathy, or wonder. These goals require interpreting how multimodal cues jointly produce emotional resonance (Bruner, 1991; Christ et al., 2024; Shi et al., 2025a; Padó and Thomas, 2025).

Although communicative goals often overlap, identifying the primary intent provides essential context for interpretation (e.g., distinguishing satire from literal claims). In our taxonomy, communicative goals form the pragmatic dimension that complements data forms and meaning-making mechanisms, together offering a unified view of how creative meaning is constructed and interpreted.

3 Task Hierarchy: Recognition, Interpretation, and Generation

We summarize existing tasks into a three-level capability hierarchy to provide a structured lens on current models (Figure 2); importantly, each level aligns with distinct evaluation signals that reflect the core requirements of the tasks.

3.1 Level 1: Recognition

Understanding human creativity begins with recognition tasks, ranging from binary classification, multi-class classification and MLLMs perception. Early work focuses on **binary classification**, where systems determine whether an input exhibits a rhetorical or non-literal phenomenon. Prior to large language models, this paradigm dominats creativity research across humor (Yang et al., 2015; Chan-

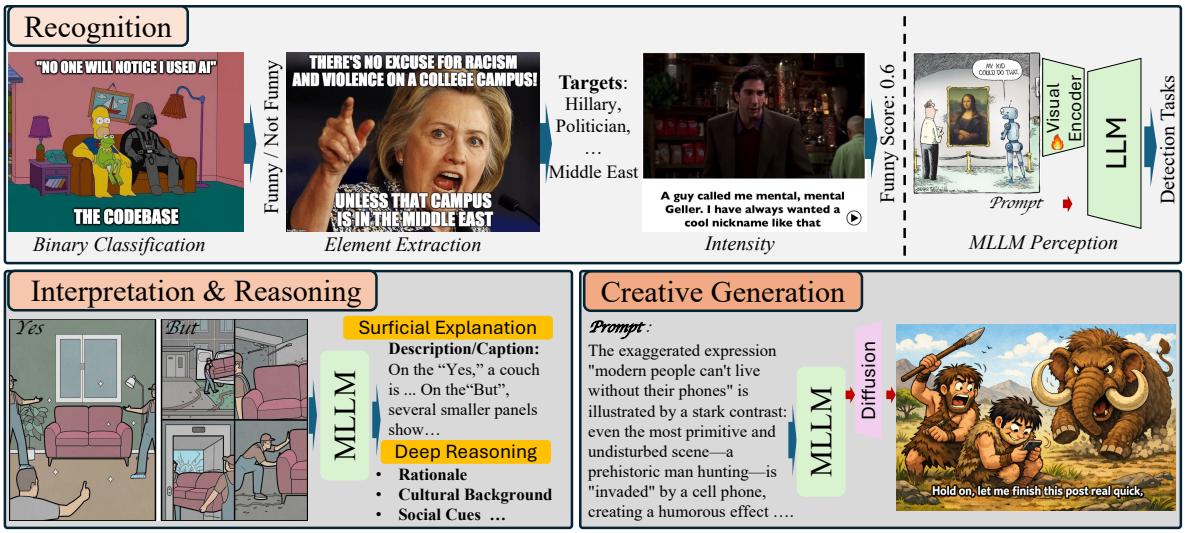

Figure 2: Capability-centric task hierarchy for multimodal human creativity.

drasekaran et al., 2016; Hasan et al., 2021), sarcasm (Schifanella et al., 2016; Liang et al., 2022; Liu et al., 2022a) etc., relying mainly on basic perceptual encoding, multimodal fusion, and surface-level cue discrimination.

Moving beyond binary decisions, recognition tasks extend to finer-grained settings (multi classification), such as **element extraction** and **intensity analysis**. Element extraction identifies constituent components of creative expressions, including persuasion techniques (Dimitrov et al., 2021b), entity roles (Sharma et al., 2022a, 2023b), sentiment and rhetorical categories in memes (Sharma et al., 2020), and punchlines or dialogue cues in multi-panel content (Kayatani et al., 2021; Liu et al., 2022b; Martínek et al., 2024). In contrast, intensity analysis quantifies the strength or perceived quality of creative effects, such as degrees of funniness (Alnajjar et al., 2022) or offensiveness (Kumari et al., 2025). These tasks require object-level localization, semantic disentanglement of overlapping signals, and judgement of creativity intensity.

MLLMs Perception exploits the language understanding and knowledge priors of LLMs to extend multimodal recognition, primarily by aligning semantic and visual representations through instruction tuning or finetune for detection-oriented tasks (Zhang and Wan, 2024; Huang et al., 2024; Gu et al., 2025). Compared to the others, these tasks still focus on recognition rather than deep reasoning, but they require stronger cross-modal semantic alignment, and language-grounded visual abstraction, enabling models to map perceptual cues to high-level labels using natural language supervision rather than handcrafted features.

For evaluation, recognition tasks are commonly assessed using discriminative metrics such as accuracy, F1, or correlation with human ratings. While suitable for measuring perceptual alignment and cue sensitivity, these metrics fail to capture interpretive depth or causal understanding required for higher-level creative reasoning.

3.2 Level 2: Interpretation and Reasoning

Recognition reflects surface-level understanding, where models map multimodal cues to labels, while this level advances to deeper reasoning, requiring models to explain why creative effects arise. **Surficial Explanation** produces descriptive or paraphrastic explanations (Hu et al., 2024) that verbalize creative content, but does not explicitly model underlying creative mechanisms or factors. This task requires natural language generation ability (Vaswani et al., 2017; Mann et al., 2020), cross-modal semantic alignment beyond label prediction (Liu et al., 2023a), and basic social and commonsense knowledge (Zhang et al., 2024b), enabling models to produce fluent, plausible descriptions by leveraging salient surface cues in a given instance without explicit multi-step reasoning.

Deep Interpretation and Reasoning requires models to explicitly reason about the mechanisms underlying creative effects through structured or theory-guided inference. It involves externalizing reasoning processes (e.g. CoT (Chen et al., 2024b), multi-stages (Tikhonov and Shtykovskiy, 2024)), abstracting non-literal concepts such as symbolism (Hussain et al., 2017; Yu et al., 2025) or metaphor (Akula et al., 2023; Saakyan et al., 2024), and integrating contextual (Wang et al., 2024b; Kumari et al., 2025) or external knowledge (Sharma et al., 2023c; Garg et al., 2025) to justify *why* a

creative effect arises rather than merely describing *what* is present.

At this level, Evaluation must go beyond fluency or surface plausibility. While surficial explanations can be assessed via semantic relevance and visual grounding (e.g., BERTScore (Zhang et al., 2019), BLEU (Papineni et al., 2002)), deep interpretation requires evaluating correct identification of rhetorical mechanisms and reasoning quality—such as correctness and consistency (Hu et al., 2024; Xiao et al., 2024)—motivating explanation-aware and evidence-based protocols (e.g., designed human evaluation and G-Eval (Liu et al., 2023b)).

3.3 Level 3: Creativity Generation

Creative expression generation spans diverse output forms—single-image (memes, humorous images, advertising posters), sequential (comics, visual narratives), and textual (captions, jokes)—with cross-modal consistency as a primary challenge: maintaining coherent creative intent and rhetorical structure across disparate representations. The field has evolved from **pattern-based approaches** that relied on predefined templates or rules, to **MLLM-guided generation** that produces flexible, context-aware content through learned multimodal representations. This shift requires models to move beyond pattern reuse toward deep interpretation and reasoning, including internalizing creative mechanisms (e.g., incongruity (Tanaka et al., 2024), metaphor (Chakrabarty et al., 2023), CoT (Zhong et al., 2024; Wang et al., 2024a)), aligning multimodal context with language generation (Shahmohammadi et al., 2023), and synthesizing novel content grounded in learned representations and knowledge priors.

Creative generation requires evaluating both output quality and consistency with the intended creative intent. Existing automatic metrics largely assess surface similarity or visual quality and struggle to capture the diversity and subjectivity of human creativity, making human evaluation still the most reliable—yet costly—option.

4 Modeling Paradigms

Our literature review covers the decade from 2015 to 2025, capturing the paradigm shift from early neural architectures to modern Foundation Models. To ensure technical rigor, we included peer-reviewed research and high-impact pre-prints focused on non-literal multimodal intent, novel benchmarks, or creative modeling paradigms. We excluded text-only studies and papers restricted to factual image captioning, as these lack the rhetorical or creative complexity central to this survey.

4.1 Task Specific Models: Non-MLLM Era

Prior to MLLMs, creativity 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.

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. (2022a) incorporated external commonsense and semantic knowledge into hierarchical congruity modeling, improving implicit intent interpretation.

Overall, non-MLLM approaches relied on carefully-engineered fusion strategies and explicit knowledge injection. While effective on specific tasks, these models remained limited in scalability and generalization, as summarized in related surveys (Sharma et al., 2022b; Farabi et al., 2024).

4.2 Large Models: Current Era

With the rise of MLLMs, modeling paradigms for human creativity have fundamentally shifted. Beyond task-specific discriminative architectures, LLM-based systems bring strong linguistic priors, world knowledge, and abstraction capabilities that enable not only recognition, but also explanation and interpretation of creative phenomena. This transition reframes multimodal creativity understanding from handcrafted feature fusion toward alignment-driven representation learning, multi-step reasoning, and knowledge-grounded interpretation, laying the foundation for the three emerging paradigms in the MLLM era: perceptual represen-

tation and multimodal alignment, reasoning mechanisms, and external knowledge integration.

Perceptual Representation and Multimodal Alignment. Modeling of human creativity with MLLMs is shifting from task-specific architectures toward large-scale multimodal alignment, where perception and semantics are jointly learned. Models such as SoMeLVLM (Zhang et al., 2024b) show that instruction tuning on creativity-oriented corpora enables handling humor, memes, and implicit intent without explicit symbolic pipelines. Complementary directions explore architectural modularity (Yu et al., 2024) to decouple perception and reasoning, and text-centric representation strategies (Hasan et al., 2023; Baluja, 2025) that reduce fusion complexity while improving interpretability.

Beyond single-panel inputs, recent work extends alignment to structured and long-form creativity settings. YesBut-v2 (Liang et al., 2025) scales instruction tuning to multi-panel memes requiring cross-panel reasoning; MeSum (Khan et al., 2024) frames meme understanding as multimodal summarization by fusing image, text, and audio with vision backbones and LLMs; FunnyNet-W (Liu et al., 2024b) adapts aligned representations to video-based humor; and ComicsPAP (Vivoli et al., 2025) targets multi-panel comic understanding. Closely related, AI4VA-FG (Chen et al., 2025) improves fine-grained comic understanding by learning region-aware visual representations via reinforcement learning, showing that more precise perceptual alignment at the region level further enhances multimodal understanding in complex visual narratives.

In general, these methods offer high scalability by using broad visual-linguistic priors to interpret diverse creative content. While effective for "surface-level" tasks like meme classification, it often suffers from semantic shallowness, relying on superficial correlations rather than deep rhetoric. This frequently leads to hallucinated intent, where models invent plausible but incorrect justifications for creative artifacts they cannot truly decode.

Reasoning Mechanics. Reasoning-based approaches emphasize *explicit* intermediate inference processes for transparent, controllable interpretation rather than implicit representation. A prominent line adopts *CoT* reasoning (Wei et al., 2022) to expose intermediate semantic steps. Tanaka et al. (2024) employs incongruity theory by prompting models to reason through expectation violation before generating humor, while Gu et al. (2025) extends this at scale with human-annotated multimodal CoT traces for harmful memes, demonstrating that explicit reasoning substantially improves both accuracy and interpretability over pattern-based baselines.

Beyond vanilla CoT, recent work proposes *multi-step and theory-guided reasoning*. Humor Mechanic (Tikhonov and Shtykovskiy, 2024) demonstrates that structured multi-step reasoning grounded in humor theory generalizes better than single-pass predictions, motivating multimodal extensions. Zhang et al. (2025) explicitly encodes humor theories into a staged framework, decomposing understanding into sequential conflict detection, resolution, and affective judgment.

Other studies highlight *abstract, symbolic, and cross-modal reasoning* as essential for creativity understanding. Loakman et al. (2024) shows MLLMs can perform explicit reasoning over sound symbolism by aligning abstract auditory–visual associations, while Liu et al. (2025) probes reasoning over non-literal, abstract visual concepts. Kundu et al. (2025) and Qiu et al. (2024) further demonstrate that interpreting metaphors, diagrams, or symbolic graphics requires explicit multi-step reasoning over latent structure rather than surface perception.

Collectively, these studies demonstrate that explicit reasoning enhances transparency by breaking down rhetorical conflicts into logical steps, enabling understanding that extends beyond simple pattern matching. However, this comes at the cost of increased computational latency and the risk of reasoning drift, where models over-analyze simple humor or become trapped in "logical loops" when faced with ambiguity.

External Knowledge Integration. Understanding multimodal creative expression often requires knowledge beyond what is explicitly present in the input, motivating the integration or retrieval of external knowledge. MemeX (Sharma et al., 2023c) explicitly formulates meme understanding as an explanatory evidence retrieval task, retrieving background documents to explain implicit cultural or contextual references. Similarly, Garg et al. (2025) and Kumari et al. (2025) incorporate external commonsense and cultural knowledge to better interpret indecent or harmful memes, showing that offensiveness and intent cannot be inferred from surface cues alone. In advertising, KAFA (Jia et al., 2023) augments vision–language models with structured

ad-related knowledge to adapt visual features toward persuasion-aware semantics. Beyond static knowledge injection, Tang et al. (2024) leverages retrieval-based external knowledge to improve robustness in sarcasm detection under domain shift.

These works demonstrate that external knowledge integration via retrieval or adaptation is crucial for grounding the understanding in cultural, social, and commonsense contexts. However, it might introduce retrieval noise. Irrelevant or distractive background data can interfere with the model's internal logic, often obscuring the core message or distracting the system from the specific visual cues that drive the creative intent.

5 Datasets and Benchmarks

In this section, we organize resources by task complexity, ranging from low-level recognition to high-level interpretation and creative generation. Detailed summaries of datasets and benchmarks are provided in Appendix Tables 1 and 2.

Recognition Resources. Early benchmarks for multimodal creativity primarily focus on existence-level recognition using simple textual or visual cues. For example, in the domain of humor, initial studies such as *Inside Jokes* (Shahaf et al., 2015) and early meme-based work ground humor detection in textual comparison and incongruity modeling (Tanaka et al., 2022). Subsequent benchmarks (e.g., UR-FUNNY (Hasan et al., 2019b), MUCH (Guo et al., 2024), and the Laughing Machine (Kayatani et al., 2021)) expand recognition to audiovisual cues, dialogue context, and temporal dynamics. More recent work further moves toward fine-grained and socially grounded signals, including laughter intensity and timing (Alnajjar et al., 2022), multimodal laughter detection (Kuznetsova and Strapparava, 2024), comic mischief (Baharlouei et al., 2024), and real-world reaction and stand-up comedy videos (Hyun et al., 2024; Barriere et al., 2025).

Overall, recognition datasets have shifted toward fine-grained, sequential annotations but remain classification-oriented and insufficient for evaluating deep reasoning or generation.

Understanding and Generation Resources. Understanding and generation are closely linked at the dataset level, as both are grounded in paired multimodal inputs (e.g., image/video–text). Thus, many datasets designed for understanding naturally support generation-oriented evaluation, since interpreting creativity often requires producing structured language grounded in multimodal context. In the MLLM era, datasets increasingly move beyond descriptive captioning toward reasoning-aware and knowledge-augmented evaluation. Early resources such as MemeCap (Hwang and Shwartz, 2023), EXCLAIM (Sharma et al., 2023a), and OxfordTVG-HIC (Li et al., 2023) extend captioning into explanatory understanding by requiring models to articulate intent or communicative goals, though they largely remain at a surface-level reasoning stage. More recent datasets, including Mementos (Wang et al., 2024b), MemeMind (Gu et al., 2025), and MemeGuard (Jha et al., 2024a), explicitly incorporate human-written rationales or CoT supervision to support deeper reasoning.

Complementarily, benchmarks such as the Yes-But series (Hu et al., 2024; Nandy et al., 2024; Liang et al., 2025), V-Flute (Saakyan et al., 2025), UnPIE (Chung et al., 2024), and Zhong et al. (2024) probe reasoning over incongruity, symbolism, and implicit knowledge, while video-focused datasets (e.g., Smile (Hyun et al., 2024), V-HUB (Shi et al., 2025b), EmoVid (Qiu et al., 2025)) further extend understanding and generation to temporal, affective, and social dynamics.

Overall, multimodal creativity datasets are moving beyond static recognition toward reasoning-oriented understanding across visual, temporal, and social cues, underscoring the need for fine-grained annotations and explicit reasoning supervision.

6 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 human creative expression. Understanding these contents requires moving beyond physical description toward socio-cultural, rhetorical, and value-laden interpretation. This shift introduces a distinct set of technical and conceptual challenges. In this section, we synthesize the core challenges and propose future directions for the community.

6.1 The Evaluation Crisis

A central limitation in current research lies in evaluation. Most benchmarks rely on multiple-choice questions (MCQs) or binary classification tasks (Hessel et al., 2023a; Hu et al., 2024; Yang et al., 2024). While these formats enable scalable comparison, they are poorly aligned with the nature of creative interpretation because:

First, such benchmarks often permit **surface-level shortcut learning** (Li et al., 2024). Models may exploit lexical cues, dataset artifacts, or answer priors without engaging in genuine rhetorical reasoning. Correctly selecting a label for a joke or meme does not guarantee that the model understands the underlying humor, irony, or critique.

Second, creativity interpretation is inherently **open-ended and subjective**. They admit multiple plausible readings depending on context and audience. Reducing these phenomena to a single "correct" option obscures interpretive depth and discourages explanation-driven reasoning.

Future evaluation must therefore move beyond discriminative setups toward *explainable and generative assessments*, where models are required to articulate why an artifact is humorous, ironic, or persuasive, and to justify their interpretations.

6.2 Data Granularity and Domain Diversity

To move beyond surface recognition, AI requires a new generation of datasets that goes beyond simple labeling. The first is need for granularity, where existing datasets often treat "creativity" as a monolithic attribute. There is a pressing need for fine-grained evaluation, separating the literal scene description from the figurative intent, the emotional subtext, and the specific rhetorical device.

Moreover, the diverse domains of data remain a problem. While social media memes are abundant, more specialized creative forms, such as political cartoons, visual satire, and culturally-specific rituals, remain underrepresented. Creating high-fidelity, expert-annotated datasets for these "long-tail" domains is essential for robust model training.

6.3 Social Knowledge

Creative artifacts are deeply embedded in social context. Their meaning often depends on shared background knowledge, implicit norms, and assumptions about audience beliefs. Current MLLMs, despite strong semantic modeling, remain limited in their ability to reason about such social pragmatics (Hu and Shu, 2023).

In particular, many failures stem from a lack of **theory-of-mind–like reasoning**: models struggle to infer whose perspective is being expressed, what beliefs are being challenged, or how an audience is expected to react. Moreover, creative expression is frequently tied to **temporally-evolving events**, trends, or controversies, requiring access to up-to-date external knowledge.

Future systems must therefore integrate social reasoning, explicit modeling of intent and audience, and retrieval mechanisms that ground interpretation in contemporary and community-specific context, rather than treating creative artifacts as isolated.

6.4 Safety, Ethics, and Ownership

Detecting Hidden Harm and Bias. Creative media often embeds harmful narratives through irony, sarcasm, and coded symbols, making subtle toxicity harder to flag than overt hate speech (Sharma et al., 2022b). Detection is further complicated by the multimodal nature of these formats—where meaning emerges only from the interplay of text and imagery—and cultural subjectivity, which often leads to inconsistent human annotations and false positives (Cao et al., 2024).

Ensuring Safe and Unbiased AI Outputs. Generative models frequently mirror demographic biases, reinforcing societal prejudices at scale. While alignment and filtering are standard mitigations, they present a "safety-utility" trade-off: overly conservative filters can stifle benign creative expression, yet remain vulnerable to brittle multimodal adversarial attacks (Lee et al., 2025). Ensuring safe, unbiased interaction with human-centered creative media remains an open challenge requiring layered alignment and robust, continuous monitoring.

Copyright and Intellectual Property. Creative datasets often scrape content from social media and art platforms without explicit consent from creators. This raises critical concerns regarding data provenance and the "Right to Style." As models move toward generation, the ability to mimic a specific artist's style or a writer's satirical tone without attribution poses a threat to the economic and intellectual rights of human creators.

6.5 Conclusion

Multimodal human creative expression 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 unified taxonomy and a capability-centric task hierarchy that clarify 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 creative expression, this survey has several limitations. First, most existing research and datasets focus on Western, internet-centric creative forms (e.g., memes, cartoons), leaving many cultural traditions and non-mainstream media underexplored. Moreover, while our taxonomy offers a unified organizational framework, real-world creative artifacts often span multiple data forms, mechanisms, and goals, making strict categorization imperfect. Finally, 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

Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas J Guibas. 2021. Artemis: Affective language for visual art. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11569–11579.

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

Arjun R Akula, Brendan Driscoll, Pradyumna Narayana, Soravit Changpinyo, Zhiwei Jia, Suyash Damle, Garima Pruthi, Sugato Basu, Leonidas Guibas, William T Freeman, and 1 others. 2023. Metaclue: Towards comprehensive visual metaphors research. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23201–23211.

Khalid Alnajjar, Mika Hämäläinen, Jörg Tiedemann, Jorma Laaksonen, and Mikko Kurimo. 2022. When to laugh and how hard? a multimodal approach to detecting humor and its intensity. *arXiv preprint arXiv:2211.01889*.

Miriam Amin and Manuel Burghardt. 2020. 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*, pages 29–41.

Elaheh Baharlouei, Mahsa Shafaei, Yigeng Zhang, Hugo Jair Escalante, and Thamar Solorio. 2024. Labeling comic mischief content in online videos with a multimodal hierarchical-cross-attention model. *arXiv preprint arXiv:2406.07841*.

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, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

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

Kate Barnes, Tiernon R. Riesenmy, Minh Duc Trinh, Eli Lleshi, Nóra Balogh, and Roland Molontay. 2020. Dank or not? analyzing and predicting the popularity of memes on reddit. *Applied Network Science*, 6.

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

Margaret A Boden. 1998. Creativity and artificial intelligence. *Artificial intelligence*, 103(1-2):347–356.

Digbalay Bose, Rajat Hebbar, Tiantian Feng, Krishna Somandepalli, Anfeng Xu, and Shrikanth Narayanan. 2023. Mm-au: Towards multimodal understanding of advertisement videos. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 86–95.

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

Clint Burfoot and Timothy Baldwin. 2009. Automatic satire detection: Are you having a laugh? In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 161–164, Suntec, Singapore. Association for Computational Linguistics.

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

Jingtao Cao, Zheng Zhang, Hongru Wang, Bin Liang, Hao Wang, and Kam-Fai Wong. 2024. Ospc: Detecting harmful memes with large language model as a catalyst. In *Companion Proceedings of the ACM Web Conference 2024*, pages 1892–1895.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _obviously_ perfect paper). *arXiv preprint arXiv:1906.01815*.

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

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

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

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

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

Lukas Christ, Shahin Amiriparian, Manuel Milling, Ilhan Aslan, and Björn Schuller. 2024. Modeling emotional trajectories in written stories utilizing transformers and weakly-supervised learning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7144–7159.

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

Simon Colton and Geraint A Wiggins. 2012. Computational creativity: The final frontier? In *ECAI 2012*, pages 21–26. IOS Press.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021a. Detecting propaganda techniques in memes. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pages 6603–6617.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021b. Semeval-2021 task 6: Detection of persuasion techniques in texts and images. In *Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021)*, pages 70–98.

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

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *arXiv preprint arXiv:1708.00524*.

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

Charles Forceville. 2002. *Pictorial metaphor in advertising*. Routledge.

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

Sonja K Foss. 2017. *Rhetorical criticism: Exploration and practice*. Waveland Press.

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

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

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

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

Hongyu Guo, Wenbo Shang, Xueyao Zhang, Shubo Zhang, Xu Han, and Binyang Li. 2024. Much: A multimodal corpus construction for conversational humor recognition based on chinese sitcom. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11692–11698.

SI Harini, Somesh Singh, Yaman K Singla, Aanisha Bhattacharyya, Veeky Baths, Changyou Chen, Rajiv Ratn Shah, and Balaji Krishnamurthy. 2025. Long-term ad memorability: Understanding & generating memorable ads. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5707–5718. IEEE.

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

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

Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. 2019a. UR-FUNNY: A multimodal language dataset for understanding humor. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2046–2056, Hong Kong, China. Association for Computational Linguistics.

Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed Ehsan Hoque. 2019b. Ur-funny: A multimodal language dataset for understanding humor. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 2046–2056.

Ming Shan Hee, Wen-Haw Chong, and Ka-Wei Roy Lee. 2023. 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).

Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023a. 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)*, pages 688–714, Toronto, Canada. Association for Computational Linguistics.

Jack Hessel, Ana Marasović, Jena D Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023b. 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)*, pages 688–714.

Eftekhar Hossain, Omar Sharif, Mohammed Moshiul Hoque, and Sarah Masud Preum. 2024. 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)*, pages 8347–8359.

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

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

Jianzhao Huang, Hongzhan Lin, Liu Ziyan, Ziyang Luo, Guang Chen, and Jing Ma. 2024. Towards low-resource harmful meme detection with lmm agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2269–2293.

Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. 2017. Automatic understanding of image and video advertisements. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1705–1715.

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

Lee Hyun, Kim Sung-Bin, Seungju Han, Youngjae Yu, and Tae-Hyun Oh. 2024. Smile: Multimodal dataset for understanding laughter in video with language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1149–1167.

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

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

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

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

Mengzhao Jia, Can Xie, and Liqiang Jing. 2024. Debiasing multimodal sarcasm detection with contrastive learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 18354–18362.

Zhiwei Jia, Pradyumna Narayana, Arjun Akula, Garima Pruthi, Hao Su, Sugato Basu, and Varun Jampani. 2023. 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)*, pages 772–785.

Narendra Nath Joshi. 2025. Evaluating human perception and bias in ai-generated humor. In *Proceedings of the 1st Workshop on Computational Humor (CHum)*, pages 79–87.

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

Sai Kartheek Reddy Kasu, Mohammad Zia Ur Rehman, Shahid Shafi Dar, Rishi Bharat Junghare, Dhanvin Sanjay Namboodiri, and Nagendra Kumar. 2025. D-humor: Dark humor understanding via multimodal open-ended reasoning–a benchmark dataset and method. *arXiv preprint arXiv:2509.06771*.

Yuta Kayatani, Zekun Yang, Mayu Otani, Noa Garcia, Chenhui Chu, Yuta Nakashima, and Haruo Takemura. 2021. The laughing machine: Predicting humor in video. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2073–2082.

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

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

Dayoon Ko, Sangho Lee, and Gunhee Kim. 2023. Can language models laugh at youtube short-form videos? *arXiv preprint arXiv:2310.14159*.

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

Yaman Kumar, Rajat Jha, Arunim Gupta, Milan Aggarwal, Aditya Garg, Tushar Malyan, Ayush Bhardwaj, Rajiv Ratn Shah, Balaji Krishnamurthy, and Changyou Chen. 2023. Persuasion strategies in advertisements. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 57–66.

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

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

Anna Kuznetsova and Carlo Strapparava. 2024. Multimodal and multilingual laughter detection in stand-up comedy videos. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11884–11889.

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

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

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

Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. 2024. Naturalbench: Evaluating vision-language models on natural adversarial samples. *Advances in Neural Information Processing Systems*, 37:17044–17068.

Lily Li, Or Levi, Pedram Hosseini, and David Broniatowski. 2020. A multi-modal method for satire detection using textual and visual cues. In *Proceedings of the 3rd NLP4IF Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 33–38, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

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

Bin Liang, Chenwei Lou, Xiang Li, Min Yang, Lin Gui, Yulan He, Wenjie Pei, and Ruifeng Xu. 2022. 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, pages 1767–1777. Association for Computational Linguistics.

Tuo Liang, Zhe Hu, Jing Li, Hao Zhang, Yiren Lu, Yunlai Zhou, Yiran Qiao, Disheng Liu, Jeirui Peng, Jing Ma, and 1 others. 2025. When'yes'meets' but': Can large models comprehend contradictory humor through comparative reasoning? *arXiv preprint arXiv:2503.23137*.

Dairu Liu, Ziyue Wang, Minyuan Ruan, Fuwen Luo, Chi Chen, Peng Li, and Yang Liu. 2025. Visual abstract thinking empowers multimodal reasoning. *arXiv preprint arXiv:2505.20164*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

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

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

Yang Liu, Tongfei Shen, Dong Zhang, Qingying Sun, Shoushan Li, and Guodong Zhou. 2024a. Comment-aided video-language alignment via contrastive pre-training for short-form video humor detection. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 442–450.

Zhi-Song Liu, Robin Courant, and Vicky Kalogeiton. 2024b. Funnynet-w: Multimodal learning of funny moments in videos in the wild. *International Journal of Computer Vision*, 132(8):2885–2906.

Zhisong Liu, Robin Courant, and Vicky Kalogeiton. 2022b. Funnynet: Audiovisual learning of funny moments in videos. In *Proceedings of the Asian Conference on Computer Vision*, pages 3308–3325.

Tyler Loakman, Yucheng Li, and Chenghua Lin. 2024. With ears to see and eyes to hear: Sound symbolism experiments with multimodal large language models. *arXiv preprint arXiv:2409.14917*.

Tyler Loakman, Aaron Maladry, and Chenghua Lin. 2023. The iron (ic) melting pot: Reviewing human evaluation in humour, irony and sarcasm generation. *arXiv preprint arXiv:2311.05552*.

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

Xinwei Long, Kai Tian, Peng Xu, Guoli Jia, Jingxuan Li, Sa Yang, Yihua Shao, Kaiyan Zhang, Che Jiang, Hao Xu, and 1 others. 2025. Adsqa: Towards advertisement video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23396–23407.

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

Ben Mann, Nick Ryder, Melanie Subbiah, J Kaplan, P Dhariwal, A Neelakantan, P Shyam, G Sastry, A Askell, S Agarwal, and 1 others. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 1(3):3.

Vittorio Marone. 2016. Looping out loud. *The European Journal of Humour Research*, 4(4):50–66.

Jiří Martínek, Pavel Král, Ladislav Lenc, and Josef Baloun. 2024. 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)*, pages 3566–3578.

Rada Mihalcea and Carlo Strapparava. 2005. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 531–538, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Youssef Mohamed, Runjia Li, Ibrahim Sa'id Ahmad, Kilichbek Haydarov, Philip Torr, Kenneth Church, and Mohamed Elhoseiny. 2024. No culture left behind: Artelingo-28, a benchmark of wikiart with captions in 28 languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20939–20962.

Saif M. Mohammad and Svetlana Kiritchenko. 2018. An annotated dataset of emotions evoked by art. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*, Miyazaki, Japan.

Naila Murray, Luca Marchesotti, and Florent Perronnin. 2012. Ava: A large-scale database for aesthetic visual analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2408–2415.

Abhilash Nandy, Yash Agarwal, Ashish Patwa, Millon Madhur Das, Aman Bansal, Ankit Raj, Pawan Goyal, and Niloy Ganguly. 2024. Yesbut: A high-quality annotated multimodal dataset for evaluating satire comprehension capability of vision-language models. *arXiv preprint arXiv:2409.13592*.

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

Khoi PN Nguyen, Terrence Li, Derek Lou Zhou, Gabriel Xiong, Pranav Balu, Nandhan Alahari, Alan Huang, Tanush Chauhan, Harshavardhan Bala, Emre Guzelordu, and 1 others. 2025. Memeqa: Holistic evaluation for meme understanding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18926–18946.

Khoi PN Nguyen and Vincent Ng. 2024. Computational meme understanding: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21251–21267.

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

Sebastian Padó and Kerstin Thomas. 2025. Artwork interpretation with vision language models: A case study on emotions and emotion symbols. *arXiv preprint arXiv:2511.22929*.

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

Sandro Paval, Pascal Meißner, and Ivan P. Yamshchikov. 2025. ComicScene154: A scene dataset for comic analysis. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31562–31568, Suzhou, China. Association for Computational Linguistics.

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. 2024. Can large language models understand symbolic graphics programs? *arXiv preprint arXiv:2408.08313*.

Zongyang Qiu, Bingyuan Wang, Xingbei Chen, Yingqing He, and Zeyu Wang. 2025. Emovid: A multimodal emotion video dataset for emotion-centric video understanding and generation. *arXiv preprint arXiv:2511.11002*.

Victor Raskin. 1979. Semantic mechanisms of humor. In *Annual Meeting of the Berkeley Linguistics Society*, pages 325–335.

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

Chengjuan Ren, Ziyu Guo, Ping Zhang, and Yuhan Gao. 2024. Humor detection using deep learning in 10 years: A survey. *Métodos numéricos para cálculo y diseño en ingeniería: Revista internacional*, 40(1):1–13.

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

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

Arkadiy Saakyan, Shreyas Kulkarni, Tuhin Chakrabarty, and Smaranda Muresan. 2025. 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)*, pages 1–23.

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

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, and 1 others. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294.

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

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

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

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

Shivam Sharma, Siddhant Agarwal, Tharun Suresh, Preslav Nakov, Md Shad Akhtar, and Tanmoy Chakraborty. 2023a. What do you meme? generating explanations for visual semantic role labelling

in memes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 9763–9771.

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

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

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

Shivam Sharma, S Ramaneswaran, Udit Arora, Md Shad Akhtar, and Tanmoy Chakraborty. 2023c. 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)*, pages 5272–5290.

Yi Shi, Wenlong Meng, Zhenyuan Guo, Chengkun Wei, and Wenzhi Chen. 2025a. Enhancing meme emotion understanding with multi-level modality enhancement and dual-stage modal fusion. *arXiv preprint arXiv:2511.11126*.

Zhengpeng Shi, Hengli Li, Yanpeng Zhao, Jianqun Zhou, Yuxuan Wang, Qinrong Cui, Wei Bi, Songchun Zhu, Bo Zhao, and Zilong Zheng. 2025b. V-hub: A visual-centric humor understanding benchmark for video llms. *arXiv preprint arXiv:2509.25773*.

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

Yuchen Su, Yonghua Zhu, Ruofan Wang, Zijian Huang, Diana Benavides-Prado, and Michael Witbrock. 2025. A survey of pun generation: Datasets, evaluations and methodologies. *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 7375–7395.

Kohtaro Tanaka, Kohei Uehara, Lin Gu, Yusuke Mukuta, and Tatsuya Harada. 2024. Content-specific humorous image captioning using incongruity resolution chain-of-thought. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2348–2367, Mexico City, Mexico. Association for Computational Linguistics.

Kohtaro Tanaka, Hiroaki Yamane, Yusuke Mori, Yusuke Mukuta, and Tatsuya Harada. 2022. Learning to evaluate humor in memes based on the incongruity theory. In *Proceedings of the Second Workshop on When Creative AI Meets Conversational AI*, pages 81–93.

Binghao Tang, Boda Lin, Haolong Yan, and Si Li. 2024. 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)*, pages 1732–1742.

Li Tao, Shunsuke Nakamura, Xueting Wang, Tatsuya Kawahara, Gen Tamura, and Toshihiko Yamasaki. 2024. A large-scale television advertising dataset for detailed impression analysis. *Multimedia Tools and Applications*, 83(7):18779–18802.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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

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

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

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

Wenshuo Wang, Ziyou Jiang, Junjie Wang, Mingyang Li, Jie Huang, Yuekai Huang, Zhiyuan Chang, Feiyan Duan, and Qing Wang. 2025a. Learning from mistakes: Enhancing harmful meme detection via misjudgment risk patterns. *arXiv preprint arXiv:2510.15946*.

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. 2025b. Beyond single frames: Can LMMs comprehend implicit narratives in comic strip? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6436–6452, Suzhou, China. Association for Computational Linguistics.

Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu,

Fuxiao Liu, Gedas Bertasius, and 1 others. 2024b. 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)*, pages 416–442.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Geraint A Wiggins. 2006. A preliminary framework for description, analysis and comparison of creative systems. *Knowledge-based systems*, 19(7):449–458.

Ruiyu Xiao, Lei Wu, Yuhang Gou, Weinan Zhang, and Ting Liu. 2024. Prove your point!: Bringing proof-enhancement principles to argumentative essay generation. *arXiv preprint arXiv:2410.22642*.

Shweta Yadav, Cornelia Caragea, Chenye Zhao, Naincy Kumari, Marvin Solberg, and Tanmay Sharma. 2023. 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)*, pages 8890–8905.

Diyi Yang, Alon Lavie, Chris Dyer, and Eduard Hovy. 2015. Humor recognition and humor anchor extraction. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2367–2376.

Yixin Yang, Zheng Li, Qingxiu Dong, Heming Xia, and Zhifang Sui. 2024. Can large multimodal models uncover deep semantics behind images? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1898–1912, Bangkok, Thailand. Association for Computational Linguistics.

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

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

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

Huixuan Zhang and Xiaojun Wan. 2024. 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)*, pages 8652–8661.

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

Jifan Zhang, Lalit Jain, Yang Guo, Jiayi Chen, Kuan Zhou, Siddharth Suresh, Andrew Wagenmaker, Scott Sievert, Timothy T Rogers, Kevin G Jamieson, and 1 others. 2024a. Humor in ai: Massive scale crowdsourced preferences and benchmarks for cartoon captioning. *Advances in Neural Information Processing Systems*, 37:125264–125286.

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

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

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

Zhengyi Zhao, Shubo Zhang, Yuxi Zhang, Yanxi Zhao, Yifan Zhang, Zezhong Wang, Huimin Wang, Yutian Zhao, Bin Liang, Yefeng Zheng, and 1 others. 2025. Memereacon: Probing contextual meme understanding in large vision-language models. *arXiv preprint arXiv:2505.17433*.

Shanshan Zhong, Zhongzhan Huang, Shanghua Gao, Wushao Wen, Liang Lin, Marinka Zitnik, and Pan Zhou. 2024. 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*, pages 13246–13257.

Naitian Zhou, David Jurgens, and David Bamman. 2024. 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)*, pages 3005–3024.

A Overview of Datasets and Benchmarks

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

Table 1: Overview of multimodal human creativity datasets focused on **Recognition** tasks. Recognition datasets typically use binary labels (e.g., humorous vs. non-humorous), element-level labels (e.g., persuasion techniques, punchlines, or rhetorical roles), and intensity scores (e.g., degree of funniness or offensiveness) to capture different granularities of creative signals. In Data Forms, StaVT denotes Static Visual–Textual Artifacts; SeqVN denotes Sequential Visual Narratives; PM denotes Audio–Visual Performative Media. In Mechanism, Multi denotes this dataset contain multi rhetorical mechanics we define in Sec. 2. The Availability column provides links to publicly accessible datasets and "N/A" indicate unpublished datasets.

Dataset	Venue	Data Forms	Mechanism	Communicative Goal	Size	Availability
D-HUMOR (Kasu et al., 2025)	ICDM'25	StaVT: *Meme*	Multi	Humor & Entertainment, Satire & Social Critique	4,379	Link
MemeMind (Gu et al., 2025)	Arxiv'25	StaVT: *Meme*	Multi	Humor & Entertainment, Satire & Social Critique	43,223	N/A
TOXICN MM (Lu et al., 2024)	NeurIPS'24	StaVT: *Meme*	Multi	Humor & Entertainment, Satire & Social Critique	12K	Link
PrideMM (Shah et al., 2024)	EMNLP'24	StaVT: *Meme*	Multi	Humor & Entertainment, Satire & Social Critique	5,063	Link
BHM (Hossain et al., 2024)	ACL'24	StaVT: *Meme*	Multi	Humor & Entertainment, Satire & Social Critique	7,148	N/A
Ext-Harm-P (Sharma et al., 2022a)	NAACL'23	StaVT: *Meme*	Multi	Humor & Entertainment, Satire & Social Critique	4,446	Link
HVVMemes (Sharma et al., 2023b)	EACL'23	StaVT: *Meme*	Multi	Humor & Entertainment, Satire & Social Critique	6,933	Link
RESTORE (Yadav et al., 2023)	ACL'23	StaVT: *Meme*	Multi	Humor & Entertainment, Satire & Social Critique	4,664	Link
Dank or not? (Barnes et al., 2020)	App. Net. Sci.'21	StaVT: *Meme*	Multi	Humor & Entertainment, Satire & Social Critique	70K	N/A
The Hateful Memes Challenge Set (Kiela et al., 2020)	NeurIPS'20	StaVT: *Meme*	Multi	Humor & Entertainment, Satire & Social Critique	10K	Link
Propaganda-techniques-in-memes (Dimitrov et al., 2021a)	ACL'21	StaVT: *Meme*	Multi	Persuasion & Advertising	950	Link
SemEval-2021 Task 6 (Dimitrov et al., 2021b)	ACL'21	StaVT: *Meme*	Multi	Persuasion & Advertising	950	Link
SemEval-2020 Task 8 (Sharma et al., 2020)	ACL'20	StaVT: *Meme*	Multi	Emotion & Aesthetic Experience	10K	N/A
RedEval (Tang et al., 2024)	NAACL'24	StaVT: *Sarcastic Image*	Sarcasm	Satire and Social Critique	1,004	Link
SPMSD (Guo et al., 2025)	COLING'24	StaVT: *Sarcastic Image*	Sarcasm	Satire and Social Critique	1K	N/A
MSTI dataset (Wang et al., 2022)	ACL'22	StaVT: *Sarcastic Image*	Sarcasm	Satire and Social Critique	5,015	Link
Multi-Modal Sarcasm Detection in Twitter (Cai et al., 2019)	ACL'19	StaVT: *Sarcastic Image*	Sarcasm	Satire and Social Critique	24,635	N/A
Sarcasm in Multimodal Social Platforms (Schifanella et al., 2016)	ACM MM'16	StaVT: *Sarcastic Image*	Sarcasm	Satire and Social Critique	10K	N/A
ArtELingo (Mohamed et al., 2024)	EMNLP'22	StaVT: *Artwork*	Multi	Emotion & Aesthetic Experience	80k	Link
WikiArt Emotions (Mohammad and Kiritchenko, 2018)	LREC'18	StaVT: *Artwork*	Multi	Emotion & Aesthetic Experience	4,105	Link
AVA (Murray et al., 2012)	CVPR'12	StaVT: *Artwork*	Multi	Emotion & Aesthetic Experience	250K	Link
HumorDB (Jain et al., 2025)	ICCV'25	StaVT: *Humorous Image*	Multi	Humor & Entertainment	3,542	Link
Image Matters (Zhang and Wan, 2024)	COLING'24	StaVT: *Hyperbole image*	Hyperbole	Emotion & Aesthetic Experience	2,160	Link
Persuasion Strategies in Advertisements (Kumar et al., 2023)	AAAI'23	StaVT: *Poster*	Multi	Persuasion & Advertising	64,832	Link
DeepMoji (Felbo et al., 2017)	EMNLP'17	StaVT: *Emoji*	Multi	Satire and Social Critique, Emotion & Aesthetic Experience	1246M	Link
Inside Jokes (Shahaf et al., 2015)	KDD'15	StaVT: *Cartoon*	Multi	Humor & Entertainment	76,928	N/A

(Continued with Table 1)

Dataset	Venue	Data Forms	Mechanism	Communicative Goal	Size	Availability
COMICORDA (Martínek et al., 2024)	COLING'24	SeqVN: *Comic Strip*	Narrative	Humor & Entertainment	1,438	N/A
AVH & FOR (Chandrasekaran et al., 2016)	CVPR'16	SeqVN: *Comic Strip*	Narrative	Humor & Entertainment	7,150	Link
StandUp4AI (Barriere et al., 2025)	EMNLP-Findings'25	PM: *Short-form Video*	Multi	Humor & Entertainment	330.00 hours	Link
Multimodal Comic Mischief Dataset (Baharlouei et al., 2024)	COLING'24	PM: *Short-form Video*	Multi	Humor & Entertainment	74.63 hours	Link
DY11k (Liu et al., 2024a)	ICMR'24	PM: *Short-form Video*	Multi	Humor & Entertainment	35.38 hours	Link
SMILE (Hyun et al., 2024)	NAACL-Findings'24	PM: *Short-form Video*	Multi	Humor & Entertainment	14.41 hours	Link
Friends (Alnajjar et al., 2022)	COLING'22	PM: *Short-form Video*	Multi	Humor & Entertainment	78.60 hours	N/A
MUCH (Guo et al., 2024)	COLING'24	PM: *Video*	Multi	Humor & Entertainment	62.00 hours	Link
Multimodal and Multilingual Laughter Detection (Kuznetsova and Strapparava, 2024)	COLING'24	PM: *Video*	Multi	Humor & Entertainment	17.00 hours	Link
The Laughing Machine (Kayatani et al., 2021)	WACV'21	PM: *Video*	Multi	Humor & Entertainment	77.42 hours	N/A
MUStARD (Castro et al., 2019)	ACL'20	PM: *Video*	Sarcasm	Satire and Social Critique	9.31 hours	Link
UR-FUNNY (Hasan et al., 2019b)	EMNLP'19	PM: *Video*	Multi	Humor & Entertainment	90.23 hours	Link

Table 2: A collection of datasets for multimodal creative **Understanding** and **Generation**. Understanding and generation datasets typically pair multimodal inputs with explanations or rationales—often augmented with contextual or external knowledge—to justify why creativity arises, alongside target creative outputs that support coherent and controllable generation across modalities (e.g., Memes, Comics or Videos + Caption/Rationales). In Data Forms, StaVT denotes Static Visual–Textual Artifacts; SeqVN denotes Sequential Visual Narratives; PM denotes Audio–Visual Performative Media. In Mechanism, Multi denotes this dataset contain multi rhetorical mechanics we define in Sec. 2. The Availability column provides links to publicly accessible datasets, and "N/A" indicates unpublished datasets.

Dataset	Venue	Data Forms	Mechanism	Communicative Goal	Size	Availability
MemeReaCon (Zhao et al., 2025)	EMNLP'25	StaVT: *Meme*	Multi	Satire & Social Critique	1,565	N/A
MEMESAFETY-BENCH (Lee et al., 2025)	EMNLP'25	StaVT: *Meme*	Multi	Humor & Entertainment	50,430	Link
MemeMind (Gu et al., 2025)	Arxiv'25	StaVT: *Meme*	Multi	Humor & Entertainment, Satire & Social Critique	43,223	N/A
MemeQA (Nguyen et al., 2025)	ACL'25	StaVT: *Meme*	Multi	Humor & Entertainment	9K	Link
SEMANTICMEMES (Zhou et al., 2024)	NAACL'24	StaVT: *Meme*	Multi	Humor & Entertainment, Satire & Social Critique	3.8M	Link
MMD (Khan et al., 2024)	EMNLP-Findings'24	StaVT: *Meme*	Multi	Humor & Entertainment	13,494	Link
MultiBully-Ex (Jha et al., 2024b)	EACL'24	StaVT: *Meme*	Multi	Humor & Entertainment	5,854	Link
Oogiri-GO (Zhong et al., 2024)	CVPR'24	StaVT: *Meme*	Multi	Humor & Entertainment	130K	Link
MemeMQACorpus (Agarwal et al., 2024)	ACL-Findings'24	StaVT: *Meme*	Multi	Humor & Entertainment	1,880	N/A
ICMM (Jha et al., 2024a)	ACL'24	StaVT: *Meme*	Multi	Humor & Entertainment	1K	Link
OxfordTVG-HIC (Li et al., 2023)	ICCV'23	StaVT: *Meme*	Multi	Humor & Entertainment	2.9M	Link
MEMECAP (Hwang and Shwartz, 2023)	EMNLP'23	StaVT: *Meme*	Multi	Humor & Entertainment	6,300	Link
MCC (Sharma et al., 2023c)	ACL'23	StaVT: *Meme*	Multi	Humor & Entertainment	3,4K	Link
HatReD (Hee et al., 2023)	IJCAI'24	StaVT: *Meme*	Multi	Satire & Social Critique	3,228	Link
ExHVV (Sharma et al., 2023a)	AAAI'22	StaVT: *Meme*	Multi	Humor & Entertainment	3K	Link
HAIVMet (Chakrabarty et al., 2023)	ACL'23	StaVT: *Artwork*	Metaphor	Humor & Entertainment, Emotion & Aesthetic Experience	6,476	Link
ArtELingo (Mohamed et al., 2024)	EMNLP'22	StaVT: *Artwork*	Multi	Emotion & Aesthetic Experience	80k	Link
LAION-Aesthetics (Schuhmann et al., 2022)	NeurIPS'22	StaVT: *Artwork*	Multi	Emotion & Aesthetic Experience	51.9M	Link
ArtEmis (Achlioptas et al., 2021)	CoRR'21	StaVT: *Artwork*	Multi	Emotion & Aesthetic Experience	81,446	Link
HumorBench (Narad et al., 2025)	Arxiv'25	StaVT: *Cartoon*	Multi	Humor & Entertainment	300	N/A
Humor in AI (Zhang et al., 2024a)	NeurIPS'24	StaVT: *Cartoon*	Incongruity	Humor & Entertainment	2.2M	Link
Do Androids Laugh at Electric Sheep? (Hessel et al., 2023b)	ACL'23	StaVT: *Cartoon*	Multi	Humor & Entertainment	24,048	N/A
V-FLUTE (Saakyan et al., 2025)	NAACL'25	StaVT: *Social Media Image*	Multi	Humor & Entertainment, Satire & Social Critique	6,027	Link
SoMeLVLM (Zhang et al., 2024b)	ACL'24	StaVT: *Social Media Image*	Multi	Humor & Entertainment, Satire & Social Critique, Emotion & Aesthetic Experience	653.8K	Link
HumorDB (Jain et al., 2025)	ICCV'25	StaVT: *Humorous Image*	Multi	Humor & Entertainment	3,542	Link
VisualPun_UNPIE (Chung et al., 2024)	EMNLP'24	StaVT: *Humorous Image*	Incongruity	Humor & Entertainment	1K	Link

(Continued with Table 2)

Dataset	Venue	Data Forms	Mechanism	Communicative Goal	Size	Availability
ImageMet (Kundu et al., 2025)	EMNLP-Findings'25	StaVT: *Metaphor Image*	Metaphor	Emotion & Aesthetic Experience	2,527	N/A
MangaUB (Ikuta et al., 2025)	IEEE MM'25	SeqVN: *Comic Strip*	Narrative	Humor & Entertainment	18,179	Link
AI4VA-FG (Chen et al., 2025)	Arxiv'25	SeqVN: *Comic Strip*	Narrative	Humor & Entertainment	16,264	N/A
PixelHumor (Ryan et al., 2025)	EMNLP-Findings'25	SeqVN: *Comic Strip*	Multi	Humor & Entertainment	2.8K	Link
YesBut-v2 (Liang et al., 2025)	Arxiv'25	SeqVN: *Comic Strip*	Multi	Humor & Entertainment	1,262	Link
YesBut (Hu et al., 2024)	NeurIPS'24	SeqVN: *Comic Strip*	Multi	Humor & Entertainment	348	Link
YesBut (synthetic 3D stick) (Nandy et al., 2024)	EMNLP'24	SeqVN: *Comic Strip*	Multi	Humor & Entertainment	2,547 (synthetic)	Link
ComicsPAP (Vivoli et al., 2025)	Arxiv'25	SeqVN: *Comic Strip*	Narrative	Humor & Entertainment	103,933	Link
XMeCap (Chen et al., 2024a)	ACM MM'24	SeqVN: *Meme*	Multi	Humor & Entertainment	12,320	N/A
Laugh, Relate, Engage (Ouyang et al., 2025)	Arxiv'25	PM: *Short-form Video*	Multi	Humor & Entertainment	1K	N/A
V-HUB (Shi et al., 2025b)	Arxiv'25	PM: *Short-form Video*	Multi	Humor & Entertainment	4.00 hours	N/A
EmoVid (Qiu et al., 2025)	Arxiv'25	PM: *Short-form Video*	Symbolism	Emotion & Aesthetic Experience	39.00 hours	N/A
ExFunTube (Ko et al., 2023)	EMNLP'23	PM: *Short-form Video*	Multi	Humor & Entertainment	84.00 hours	Link
MUStARD (Castro et al., 2019)	ACL'20	PM: *Short-form Video*	Sarcasm	Satire & Social Critique	9.31 hours	Link
AdsVQA (Long et al., 2025)	ICCV'25	PM: *Ads Video*	Multi	Persuation & Advertising	1,544 (22.70 hours)	Link
LAMDBA (Harini et al., 2025)	WACV'25	PM: *Ads Video*	Multi	Persuasion & Advertising	2,205	Link
TV Ads (Tao et al., 2024)	Multimedia Tools App.'24	PM: *Ads Video*	Multi	Persuasion & Advertising	14,490	N/A
MM-AU (Bose et al., 2023)	ACM MM'23	PM: *Ads Video*	Multi	Persuasion & Advertising	147.00 hours	N/A
Automatic Understanding of Image and Video Advertisements (Hussain et al., 2017)	CVPR'17	StaVT & PM: *Ads Poster & Video*	Multi	Persuation & Advertising	64,832 & 3,477	Link