

Humor in Pixels: Benchmarking Large Multimodal Models Understanding of Online Comics

Yuriel Ryan*, Rui Yang Tan*, Kenny Tsu Wei Choo, Roy Ka-Wei Lee,
Singapore University of Technology and Design

{yurieljunlongryan_wang@mymail., ruiyang_tan@mymail., kenny_choo@, roy_lee@}sutd.edu.sg

Abstract

Understanding humor is a core aspect of social intelligence, yet it remains a significant challenge for Large Multimodal Models (LMMs). We introduce PixelHumor, a benchmark dataset of 2,800 annotated multi-panel comics designed to evaluate LMMs’ ability to interpret multimodal humor and recognize narrative sequences. Experiments with state-of-the-art LMMs reveal substantial gaps: for instance, top models achieve only 61% accuracy in panel sequencing, far below human performance. This underscores critical limitations in current models’ integration of visual and textual cues for coherent narrative and humor understanding. By providing a rigorous framework for evaluating multimodal contextual and narrative reasoning, PixelHumor aims to drive the development of LMMs that better engage in natural, socially aware interactions. PixelHumor is made available here: <https://github.com/Social-AI-Studio/PixelHumor>

1 Introduction

Motivation. Humor is a quintessential element of human communication and intelligence, reflecting our ability to perceive, interpret, and appreciate complex social and cultural nuances (Kuipers et al., 2008; Jiang et al., 2019). It plays a pivotal role in social interaction, creativity, and even cognitive development. The capacity to understand and generate humor is often considered a hallmark of advanced intelligence, involving sophisticated processes such as abstract thinking, contextual reasoning, and emotional perception.

In the realm of artificial intelligence, Large Language Models (LLMs) like GPT-4o (Achiam et al., 2023) have demonstrated remarkable proficiency in various natural language processing tasks, including text generation, translation, and question-answering (Chang et al., 2024; Laskar et al., 2023).



Figure 1: A XKCD (Munroe, 2024) comic¹ that depicts a launching rocket at risk of hitting Santa Claus.

However, their ability to understand or generate humor remains limited (Chang et al., 2024; Jentzsch and Kersting, 2023). Humor comprehension involves not only linguistic understanding (Attardo, 2009, 1997) but also the interpretation of subtle cues, double meanings, and cultural references, which are challenging for AI systems primarily trained on textual data.

The recent development of Large Multimodal Models (LMMs), such as GPT-4o (Achiam et al., 2023), Gemini-1.5 (Reid et al., 2024), Gemma 3 (Team et al., 2025), Qwen2-VL (Wang et al., 2024), and LLaVA-OneVision (Li et al., 2024), has extended AI capabilities to process and integrate information from multiple modalities, including text and visuals. This advancement opens new avenues for AI to tackle complex tasks that requires multimodal comprehension. Despite these advancements, there is a lack of systematic studies evaluating LMMs’ ability to understand visual humor, particularly in comics that combine images and text to convey contextual humorous narratives as exemplified in Figure 1. For instance, GPT-4o interpreted the Figure 1 as “*Santa Claus is hijacking the rocket*”. While the model is able to identify key objects in the comic (e.g., *Santa Claus* and *rocket*), it misinterpreted the humor, which suggests that the launching rocket is going to hit Santa Claus.

¹XKCD comics used in this work were for non-commercial purposes under the Creative Commons Attribution-NonCommercial 2.5 License. Access this comic at <https://xkcd.com/2559/>.

*Equal contribution

Research Objectives. To address this gap, we present PixelHumor, a benchmark dataset of 2,800 comics sourced from seven distinct creators. The dataset is carefully annotated for humor-related tasks, including classification, interpretation, and sequential recognition, offering a comprehensive framework to evaluate LMMs’ multimodal humor comprehension systematically. Through extensive experiments on state-of-the-art LMMs using the PixelHumor dataset, we demonstrate that these models struggle significantly with understanding humor, particularly when it involves complex compositional reasoning and multimodal integration.

Contributions. We summarize the main contributions of this work as follows: (i) We construct PixelHumor, a dataset of 2,800 comics with detailed annotations for humor understanding tasks, filling a critical gap in resources for evaluating multimodal humor comprehension in AI. (ii) We provide a comprehensive framework by outlining specific subtasks (identification, classification, interpretation, and sequential recognition) to evaluate the humor comprehension abilities of LMMs in a multimodal context. (iii) We benchmark state-of-the-art LMMs on the PixelHumor dataset and analyze their performance, highlighting the challenges these models face in understanding complex humorous content that requires advanced reasoning and contextual knowledge.

2 Related Works

Humor has been explored through multiple disciplinary lenses, including linguistic (Attardo, 2009, 1997), psychological (Martin and Ford, 2018), and philosophical perspective (Morreall, 1982, 1986), as well as its physiological benefits (Wilkins and Eisenbraun, 2009; McGraw and Warren, 2010). In this work, we focus on insights from the social science perspective to inform the design of benchmarks for LMMs. Humor, as a form of communication (Kuipers et al., 2008; Davis, 2008) and social intelligence (Yip and Martin, 2006), plays a pivotal role in various social and cultural contexts (Crawford, 2003; Jiang et al., 2019; Moody-Ramirez and Church, 2019), making its comprehension an essential aspect of human-centric AI development.

Despite advances in AI, LMMs struggle to replicate humor’s intricate interplay of linguistic, cultural, and contextual elements (Mirowski et al., 2024). This challenge is compounded by the overlap of online humor with potentially harmful con-

tent, prompting research on detecting (Ji et al., 2023; Cao et al., 2023b,a), clustering (Prakash et al., 2023b) and explaining (Lin et al., 2024; Lee et al., 2021; Hee et al., 2023) toxic memes (Hee et al., 2025; Hee and Lee, 2025), alongside efforts in meme generation (Wang and Lee, 2024; Sadasivam et al., 2020; Peirson V and Tolu-nay, 2018) and humorous image captioning (Zhang et al., 2024; Tanaka et al., 2024; Chandrasekaran et al., 2016; Hessel et al., 2023). Key humor understanding tasks—detection, classification, interpretation, and generation—have been identified, yet existing datasets address these in isolation. Text-based datasets like TalkFunny (Chen et al., 2024), One-liners (Mihalcea and Strapparava, 2005), Pun of the Day (Yang et al., 2015), and Ted-Laughter (Chen and Lee, 2017) focus on written jokes, while HumorDB (Jain et al., 2024) targets static visual humor classification and explanation. Multimodal datasets to address satire include YesBut (Nandy et al., 2024) and memes such as Memotion 3.0 (Mishra et al., 2023), MERMAID (Toh et al., 2023), and TotalDefMeme (Prakash et al., 2023a). Audio-text datasets like Big Bang Theory (Bertero and Fung, 2016) detect punchlines, whereas audiovisual datasets (MUSARD (Castro et al., 2019), UR-Funny (Hasan et al., 2019), MUMOR (Wu et al., 2021)) analyze dynamic humor. However, these datasets lack integrated evaluation of narrative sequencing and multi-panel humor, limiting their ability to assess LMMs’ temporal reasoning (see Appendix A for details).

To address these gaps, we introduce PixelHumor, a benchmark comprising 2,800 annotated multi-panel online comics. Unlike previous datasets, which rely on single panel images or static contexts, PixelHumor uniquely integrates humor detection, classification in eight styles, open-ended interpretation, and sequence recognition within a unified framework. This multi-panel focus enables evaluation of narrative and temporal reasoning, critical for humor delivery in visual storytelling, revealing LMM limitations in contextual integration. By holistically assessing these dimensions, PixelHumor provides a comprehensive tool to advance the research on LMM humor comprehension.

3 Dataset Construction

3.1 Data Collection

PixelHumor comprises comics from seven diverse sources: *Cyanide and Happiness* (ExplosmEnter-

Humor Style	Description
Comparison	This comic compares two or more objects/ideas to highlight similarities or differences, and the humor stems from this comparison.
Personification	This comic features an animal, creature, or plant with human-like qualities and the humor stems from this personification.
Exaggeration	This comic exaggerates actions, words, or situations to an absurd degree, and the humor lies in this overemphasis.
Pun	This comic utilizes wordplay or linguistic ambiguity to create humor.
Sarcasm	This comic expresses an idea that is the opposite of the speaker’s true intention, and the humor lies in this discrepancy.
Silliness	This comic contains absurd or foolish elements. The humor stems from the nonsensical or absurd nature of the situation or character.
Surprise	This comic features a twist or unexpected element. The humor stems from the subversion of expectations.
Dark	This comic incorporates dark, taboo, or potentially offensive ideas. The humor arises from a benign challenging or violation of conventional norms.

Table 1: Taxonomy of Humor Styles in PixelHumor tainment, 2024), *Peanuts* (Schulz, 2024), *Garfield* (Davis, 2024), *XKCD* (Munroe, 2024), *PhD Comics* (Cham, 2024), *They Can Talk* (Kelly, 2024), and *Saturday Morning Breakfast Cereal (SMBC)* (Weinersmith, 2024), totaling 2,800 comics (see AppendixD for source-specific statistics). These were selected to represent a broad spectrum of humor styles—*Comparison*, *Personification*, *Exaggeration*, *Pun*, *Sarcasm*, *Silliness*, *Surprise*, and *Dark*, adapted from (Taecharungroj and Nueangjamnong, 2015) and detailed in Table 1. Each of the sources exhibit a variety of humor styles, while some of them dominates in certain humor styles. For instance, *Cyanide and Happiness* emphasizes dark humor, *They Can Talk* focuses on personification, and *PhD Comics* highlights sarcasm, ensuring diverse comedic coverage.

Comics were collected for research purposes, using only publicly available content without modification. Compliance with intellectual property was ensured by adhering to robots.txt guidelines and fair use principles, with automated checks verifying access permissions. Data were used solely for annotation and analysis, not for model training, and will be deleted post-study. To respect copyright, PixelHumor will be released as URLs linking to original sources, preserving creator hosting.

3.2 Humor Annotation

Eight undergraduate students aged between 18 to 25 were recruited, and trained over two weeks, including annotating five trial comics and three practice sessions, using guidelines in Appendix B. They were instructed to identify the intended humor objectively, minimizing personal or demographic biases, and were briefed about sensitive content (e.g., dark humor), with the option to withdraw if they feel uncomfortable with any of the content.

The 2,800 comics were split into four sets of 700, each annotated by a pair of annotators to assess inter-annotator agreement. Tasks were released in 100-comic batches, with quality control checks on 10 random comics per batch, evaluating sound effect identification and panel number accuracy as objective metrics of attention and adherence to guidelines. Disagreements, occurring in 15% of cases, were resolved by a third annotator, with final labels determined by majority voting, ensuring consistency for ambiguous humor.

3.3 Dataset Analysis

Inter-Annotator Agreement. To assess annotation consistency and reliability, we computed agreement levels (exact matches for single-label annotations and overlaps for multi-label annotations) and Krippendorff’s Alpha (K-alpha) (Hayes and Krippendorff, 2007). The overall agreement level was 0.872, indicating strong alignment among annotators. However, the K-alpha score was lower at 0.556, suggesting limited reliability in certain annotations. Upon further analysis, we identified that the primary source of low reliability was the question: *Do you understand the humor in this comic?* The inherent humor-oriented nature of the selected comic sources resulted in a strong label imbalance, as most comics were perceived as humorous by annotators. This imbalance likely contributed to the lower reliability score².

Sound Effects and Humor. As shown in Figure 2(a), 85% of the 2,800 comics lack sound effects. Among the 15% that include them, approximately 70% feature onomatopoeic expressions (e.g., *BAM!*, *POW!*) that directly enhance humor, often tied to motion or action (assumed proportion; adjust with data). This suggests sound effects serve as a linguistic device amplifying comedic effect in

²To enhance transparency and facilitate further research, we will release both the individual annotations and the aggregated labels in the dataset.

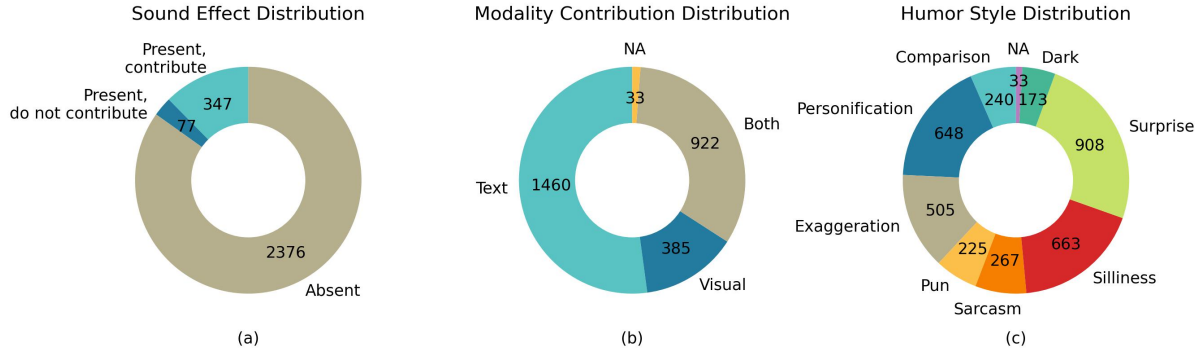


Figure 2: Distribution of sound effects, modality contribution, and humor styles in PixelHumor

visual storytelling, particularly in dynamic scenes. For LMMs, accurately interpreting these cues requires integrating auditory-like signals with visual context, a challenge given their text-heavy bias.

Modality Contribution to Humor. Figure 2(b) reveals that text is the primary humor driver in 52% of comics, underscoring the dominance of linguistic elements. Another 32% rely on a synergistic text-visual interplay, where images complement textual punchlines, while 16% are labeled non-humorous (NA) (assumed; adjust with data). These findings highlight that while text alone can convey humor, a significant portion of comics leverages multimodal cues for comedic impact. This poses a challenge for LMMs to model cross-modal dependencies, essential for capturing nuanced humor in narrative sequences.

Humor Style Distribution. Figure 2(c) shows *Surprise* as the most prevalent humor style (35%), reflecting its critical role in comedic timing and punchline delivery across sources. *Personification* follows closely (28%), particularly in comics featuring anthropomorphic characters, such as *Garfield* (Davis, 2024), *Peanuts* (Schulz, 2024), and *They Can Talk* (Kelly, 2024). Conversely, *Dark* humor is least common (5%), primarily concentrated in *Cyanide and Happiness* (ExplosmEntertainment, 2024), known for its provocative style. Non-humorous comics are labeled NA (assumed proportions; adjust with data). A detailed source-specific breakdown is provided in Appendix D. These patterns emphasize the need for LMMs to adapt to diverse humor styles, particularly those reliant on narrative context and cultural nuances.

4 Experiment Settings

In this section, we outline the experiment settings for evaluating LMMs’ ability to comprehend humor in comics using the PixelHumor dataset.

4.1 Task Definitions

We define four core tasks—*humor identification*, *humor classification*, *humor interpretation*, and *sequence recognition*—to comprehensively evaluate the multimodal humor comprehension capabilities of LMMs. These tasks target key dimensions of cognitive and narrative understanding, from basic cue detection to higher-order reasoning.

Humor Identification. This task tests whether a model can detect and localize humor within a comic. Subtasks include identifying the presence of humor, recognizing contributing factors such as sound effects, determining the most critical panel for humor delivery, and assessing the relative importance of textual versus visual elements. Success indicates basic multimodal perception skills, while failure highlights limitations in recognizing fundamental humor cues.

Humor Classification. Models must categorize comics into one or more humor styles based on a predefined taxonomy (Table 1). As a multi-class classification problem, this task challenges models to disentangle overlapping and subtle humor types. Accurate classification demonstrates an ability to generalize and apply nuanced conceptual understanding—critical given humor’s inherently subjective and context-dependent nature.

Humor Interpretation. This open-ended task assesses a model’s ability to explain why a comic is humorous. Models must generate natural language

explanations articulating how textual and visual elements interact to produce comedic effects. This task probes reasoning and abstraction capabilities beyond surface-level pattern recognition, aiming to reveal whether models can “*think*” about humor in human-like terms.

Sequence Recognition. Comics rely on a carefully structured narrative order to build context and deliver punchlines. In this task, models must correctly reconstruct the intended sequence of panels and associated textual elements. Success reflects an understanding of temporal dependencies and narrative flow across modalities—an essential aspect of coherent humor comprehension.

Task-specific prompts were designed to align closely with the objectives of each evaluation setting. Full details of the prompt construction are provided in Appendix E.

4.2 Benchmark Models

To evaluate LMMs’ ability to comprehend humor in comics, we benchmarked a diverse set of models spanning different architectures, parameter scales, and accessibility settings. All models were evaluated under a unified framework using standardized prompts (Appendix E) with a temperature of 0 to ensure consistency and deterministic outputs.

We included two closed-source models: GPT-4o³ (Achiam et al., 2023) and Gemini-1.5-Pro⁴ (Reid et al., 2024). GPT-4o represents the latest state-of-the-art in multimodal reasoning, while Gemini-1.5-Pro is a competitive model known for its strong performance in vision-language tasks.

For open-source evaluation, we selected four models to capture a range of capacities. Among larger models, we tested Qwen2-VL-72B (Wang et al., 2024), a high-capacity vision-language model with strong generalization across multimodal benchmarks, and Gemma3-27B (Team et al., 2025), noted for its advanced visual grounding and contextual reasoning abilities. To complement these, we included two smaller models: LLaVA-OneVision-7B-SI⁵ (Li et al., 2024), which is optimized for visual understanding at the single-image level, and Qwen2-VL-7B (Wang et al., 2024), a lightweight yet competitive model.

All models were evaluated in their publicly available pre-trained forms without task-specific fine-

³Evaluated using the GPT-4o 2024-08-06 model

⁴Evaluated using the gemini-1.5-pro-001 model

⁵The SI variant is pre-trained on single-image tasks, making it suitable for comic panel evaluation

Sub-Tasks	Model	F1	Prec.	Rec.
Humor Presence Identification	GPT-4o	0.983	0.983	0.988
	Gemini-1.5-Pro	0.984	0.982	0.988
	Qwen2-VL-72B	0.984	0.983	0.988
	Gemma3 27B	0.984	0.985	0.989
	LLaVA-OV 7B	0.982	0.977	0.988
	Qwen2-VL 7B	0.982	0.979	0.987
Sound Effect Identification	GPT-4o	0.821	0.821	0.820
	Gemini-1.5-Pro	0.790	0.845	0.742
	Qwen2-VL-72B	0.717	0.844	0.624
	Gemma3 27B	0.743	0.840	0.666
	LLaVA-OV 7B	0.703	0.833	0.609
	Qwen2-VL 7B	0.013	0.849	0.007
Panel Contribution	GPT-4o	0.765	0.788	0.762
	Gemini-1.5-Pro	0.717	0.746	0.713
	Qwen2-VL-72B	0.494	0.605	0.485
	Gemma3 27B	0.540	0.633	0.530
	LLaVA-OV 7B	0.501	0.618	0.507
	Qwen2-VL 7B	0.517	0.639	0.498
Modality Contribution	GPT-4o	0.626	0.699	0.656
	Gemini-1.5-Pro	0.613	0.661	0.631
	Qwen2-VL-72B	0.577	0.682	0.605
	Gemma3 27B	0.211	0.696	0.352
	LLaVA-OV 7B	0.562	0.597	0.586
	Qwen2-VL 7B	0.214	0.608	0.297

Table 2: Experiment result of various humor identification tasks. The best results are **bold**.

tuning. This zero-shot setting ensures a fair assessment of their generalization capabilities to novel, humor-centered multimodal reasoning tasks.

4.3 Evaluation Framework and Metrics

Automated metrics were employed for humor identification, humor classification, and sequence recognition, while human evaluations are conducted for humor interpretation.

For humor identification and humor classification, we used precision, recall, and weighted F1-score against human-annotated ground truth, capturing the models’ ability to detect humor and distinguish nuanced humor styles (Table 1).

Humor interpretation, being open-ended, was evaluated through human ratings on a 7-point Likert scale, assessing the relevance and coherence of the model-generated explanations.

Sequence recognition was evaluated by measuring models’ ability to reconstruct the correct order of panels and associated text. Visual sequencing was assessed using accuracy, while textual sequencing was further evaluated with Word Error Rate (WER) and Character Error Rate (CER) to capture fine-grained alignment.

5 Experiment Results

In this section, we present the experimental results across the four humor evaluation tasks: *humor identification*, *humor classification*, *humor interpreta-*

Model	F1.	Prec.	Rec.	Com.	Per.	Exa.	Pun.	Sar.	Sil.	Sur.	Dar.	N/A
GPT-4o	0.499	0.393	0.711	0.596	0.965	0.758	0.587	0.569	0.593	0.713	0.746	0.030
Gemini-1.5-Pro	0.480	0.393	0.673	0.421	0.819	0.481	0.471	0.854	0.655	0.789	0.584	0.091
Qwen2-VL 72B	0.375	0.455	0.382	0.304	0.840	0.521	0.409	0.251	0.267	0.128	0.358	0.182
Gemma3 27B	0.465	0.370	0.735	0.446	0.767	0.576	0.333	0.689	0.916	0.863	0.827	0.091
LLaVA-OV 7B	0.094	0.306	0.123	0.388	0.071	0.170	0.076	0.775	0.005	0.000	0.000	0.000
Qwen2-VL 7B	0.248	0.507	0.295	0.467	0.877	0.463	0.271	0.101	0.054	0.001	0.237	0.030

Table 3: Experiment result of humor classification task. We also report the Recall score of various humor style types. **Com.:** Comparison, **Per.:** Personification, **Exa.:** Exaggeration, **Pun.:** Pun, **Sar.:** Sarcasm, **Sil.:** Silliness, **Sur.:** Surprise, **Dar.:** Dark, **N/A:** Not Applicable. The best results are **bold**.

tion, and sequence recognition. Overall, closed-source models such as GPT-4o and Gemini-1.5-Pro consistently outperform open-source counterparts, particularly on tasks requiring fine-grained narrative understanding and multimodal reasoning. While all models excel at detecting the presence of humor, they struggle significantly with identifying key narrative moments, explaining humor coherently, and reconstructing comic sequences. These findings highlight fundamental challenges in current LMMs’ abilities to integrate visual and textual information over extended contexts. We report both quantitative metrics and qualitative analyses to provide a comprehensive assessment of model performance and limitations.

5.1 Humor Identification

Humor identification evaluates a model’s ability to detect humor cues, attribute their sources, and understand the relative contributions of textual and visual elements. Table 2 summarizes the results across four sub-tasks.

Humor Presence. All models achieved near-perfect performance, with F1-scores exceeding 0.98. Gemini-1.5-Pro, Qwen2-VL-72B, and Gemma3-27B recorded the highest F1-scores (0.984), each misclassifying only eight comics. However, given that the dataset is heavily skewed toward humorous comics (only 33 labeled as non-humorous), this high accuracy likely reflects label imbalance rather than deep humor understanding.

Sound Effect Identification. While GPT-4o achieved the highest F1-score (0.821) in recognizing the role of sound effects in humor, Qwen2-VL-7B performed poorly. The latter exhibited a strong bias toward predicting sound effects regardless of context, leading to extremely low recall (0.007). These results suggest that some models rely on superficial textual heuristics rather than genuine multimodal understanding.

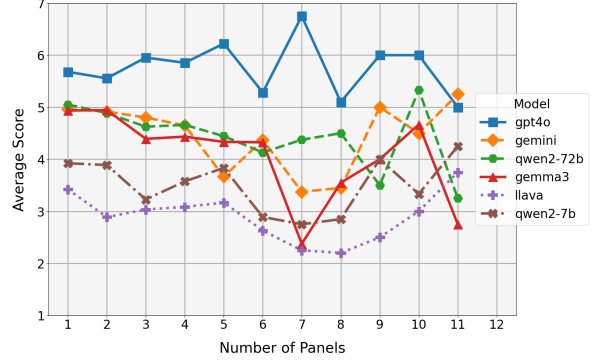


Figure 3: Average relevance scores for the comics (y-axis) plotted against the number of panels (x-axis). Comics with more than 8 panels make up a very small proportion (1.7%) of the rated comics.

Panel Contribution. Identifying the panel most responsible for humor posed a greater challenge. GPT-4o achieved the highest F1-score, while open-source models lagged significantly. Common errors included defaulting to reading order (e.g., left-to-right selection) rather than recognizing punch-line positioning. Notably, Qwen2-VL-72B frequently generated out-of-bounds panel indices, further reducing its performance. These patterns indicate that LMMs struggle to localize key narrative moments crucial for humor delivery.

Modality Contribution. This sub-task assessed whether humor was driven by text, visuals, or both. GPT-4o again outperformed others (F1 = 0.626), closely followed by Gemini-1.5-Pro (F1 = 0.613). In contrast, Qwen2-VL-7B and Gemma3-27B achieved low F1-scores, largely due to an over-reliance on predicting “Both” modalities regardless of context. These findings suggest that fine-grained attribution of humor sources remains a key weakness for many LMMs, particularly smaller or open-source models.

5.2 Humor Classification

Humor classification evaluates a model’s ability to distinguish between different humor styles within

Model	1 HS	2 HS	3 HS	4 HS	5+ HS
Ground Truth	2,052	643	96	9	0
GPT-4o	343	856	1,480	112	9
Gemini-1.5-Pro	39	1,227	1,239	276	19
Qwen2-VL 72B	2,420	252	119	8	1
Gemma3 27B	8	140	2,226	417	9
LLaVA-OV 7B	2,800	0	0	0	0
Qwen2-VL 7B	2,661	41	50	23	25

Table 4: Number of humor styles (HS) predicted by each model.

a comic. Table 3 presents the weighted average performance across models.

GPT-4o achieved the highest F1-score, followed closely by Gemini-1.5-Pro and Gemma3-27B. Larger open-source models, such as Qwen2-VL-72B, showed competitive performance but still trailed behind the closed-source models, while smaller models like Qwen2-VL-7B and LLaVA-OV-7B struggled significantly.

A key challenge for open-source models was a strong bias toward assigning a single humor style per comic, despite the multi-label nature of the task (Table 4). Qwen2-VL-7B, for instance, achieved relatively high precision but low recall, leading to lower F1-score. LLaVA-OV-7B consistently predicted only one humor style (primarily *Sarcasm*), severely limiting its ability to capture the interplay of multiple comedic elements. Similarly, Qwen2-VL-72B and Qwen2-VL-7B exhibited a bias toward predicting *Personification* and *Exaggeration*, reflecting a lack of nuanced humor modeling.

Across all models, *Personification* was classified most accurately, likely due to its clear visual and textual cues, particularly in comics featuring anthropomorphized characters (e.g., *Peanuts*, *Garfield*, *They Can Talk*). In contrast, styles requiring deeper contextual understanding, such as *Sarcasm* and *Dark* humor, were more frequently misclassified, highlighting limitations in models’ subtle inference abilities.

Another consistent pattern was poor recall for the “N/A” (non-humorous) category. This reflects an underlying bias toward humor predictions, likely influenced by the dataset’s skew toward humorous comics. As a result, models struggled to identify non-humorous instances, mirroring trends observed in the humor presence identification task.

5.3 Humor Interpretation

We selected 350 comics with high annotator agreement and rated model-generated explanations on a 7-point Likert scale to assess models’ ability to generate coherent explanations for comic-based humor.

Model	Mean	Median	STD
GPT-4o	5.801	6.500	1.371
Gemini-1.5-Pro	4.647	5.000	1.687
Qwen2-VL 72B	4.653	4.500	1.437
Gemma3 27B	4.439	4.500	1.796
LLaVA-OV 7B	3.039	3.000	1.595
Qwen2-VL 7B	3.477	3.500	1.712

Table 5: Statistics of relevance scores evaluated by human annotators on LMMs’ generated interpretation of the comics. The best results are **bold**.

Model	K-alpha	Exact Match	Mean	Median	STD
GPT-4o	0.252	0.368	1.334	1.000	1.534
Gemini-1.5-Pro	0.367	0.257	1.694	1.000	1.553
Qwen2-VL 72B	0.391	0.265	1.289	1.000	1.154
Gemma3 27B	0.476	0.243	1.251	1.000	1.069
LLaVA-OV 7B	0.456	0.237	1.551	1.000	1.394
Qwen2-VL 7B	0.444	0.274	1.491	1.000	1.379
Overall	0.523	0.274	1.435	1.000	1.368

Table 6: Inter-rater reliability and statistics of absolute rating difference by model.

Results are summarized in Table 5.

Quantitative Analysis. GPT-4o achieved the highest mean score for humor interpretation, substantially outperforming all open-source models. Gemini-1.5-Pro, Qwen2-VL-72B, and Gemma3-27B followed with comparable mean ratings, while smaller models such as Qwen2-VL-7B and LLaVA-OV-7B performed significantly worse. These results highlight a clear gap between proprietary and open-source models in humor reasoning, as well as the advantage of larger model sizes.

Figure 3 illustrates how comic length affects relevance scores. Performance remains relatively stable for comics with four or fewer panels, but begins to fluctuate as narratives grow longer, particularly beyond six panels. This trend suggests that while additional context can sometimes aid humor interpretation, longer sequential dependencies pose increasing challenges for LMMs.

Qualitative Analysis. Generated interpretations from smaller open-source models frequently hallucinated humor elements or produced generic, template-like explanations (See Appendix H for examples). While larger open-source models provided more plausible interpretations, they too struggled as the number of panels increased. Even top-performing closed-source models exhibited volatility in longer narratives, indicating persistent difficulties in maintaining multimodal coherence over extended sequences.

These findings point to long-context modeling as a critical bottleneck for humor comprehension.

Future improvements may benefit from hierarchical attention mechanisms, enhanced cross-modal alignment, or human-in-the-loop training strategies to better support narrative tracking and sequential humor reasoning.

Inter-evaluator Agreement. After evaluating the generated explanations for the humor interpretation task, we calculated the inter-rater reliability shown in Table 6. The Krippendorff’s alpha is 0.523 which reflects a moderate agreement between evaluators. This value is within the acceptable range for subjective tasks in NLP and HCI, such as emotion or empathy evaluation. Moreover, the low median rating difference (1.0) and standard deviation (~ 1.4) show that even when evaluators disagree, their ratings are generally close in magnitude. GPT-4o received the highest exact match agreement but a lower K-alpha as it’s ratings skew toward the upper end (scores 6–7); these higher scores reduces the score variance and suppresses alpha despite consistent judgments, a known behavior of this metric. These values indicate that the human evaluations of the LMM’s explanations remain largely consistent, despite the subjective nature of humor interpretation.

5.4 Human Preferences.

We further assessed the difficulty of humor interpretation for LMMs through a preference study. In this study, we selected 70 comics: 10 comics randomly chosen per source and a human was tasked to interpret the humor in these comics. After which, two participants were tasked to choose the best interpretation of the humor—one human and six generated—for the comics. Disagreements are resolved through a third annotator from the chosen explanations.

Human written explanations are mostly preferred over generated ones in interpreting humor. In 48 (68.6%) of the 70 samples (Table 7), the human written explanation was judged to be the best by human participants, despite the high mean scores achieved by GPT-4o in this task (Table 5). We attribute this outcome to two main factors: the superior reasoning abilities of humans and the repeated general reasoning patterns for humor by the models. Human-written explanations were able to better combine information from both modalities and describe how they synergize to amplify humor. In contrast, generated explanations often give general explanations that rely on absurdity

Explanation Source	Times Selected
Human	48
GPT-4o	14
Gemini-1.5-Pro	3
Qwen2-VL 72B	3
Gemma3 27B	1
LLaVA-OV 7B	0
Qwen2-VL 7B	1

Table 7: Results of the human preference study to select the best explanation out of the six generated and one human written explanation for 70 comics.

or unexpectedness, which might not be the main crux of the humor (see examples in Appendix H). Ultimately, these results indicate that humans are still far ahead of LMMs in more complex multi-modal humor tasks: fine-grained comprehension and interpretation.

5.5 Sequence Recognition

Sequence recognition—visual and text—evaluates a model’s ability to reconstruct the correct reading order of comic panels, a crucial component for maintaining narrative coherence and delivering humor. Each panel was randomly assigned a unique numerical label from 1 to N , displayed at the top-left corner.

Visual Sequence. Visual sequence recognition assesses a model’s ability to reconstruct the correct order of panels in the comics. Gemini-1.5-Pro achieved the highest panel sequencing accuracy (0.645), closely followed by GPT-4o (0.614), while open-source models lagged significantly (e.g., Qwen2-VL-72B at 0.343, Gemma3-27B at 0.310) (Table 8). LLaVA-OV-7B showed intermediate performance (0.422), outperforming other open models but still far behind the closed-source models.

Qualitative analysis for the visual sequencing task revealed several common failure patterns across all models. Many defaulted to conventional reading orders (e.g., left-to-right, top-to-bottom) without adapting to the randomized panel numbering. Others miscounted the number of panels or produced continuous sequences until reaching output limits, suggesting reliance on heuristic completions rather than true sequential reasoning. These errors underscore broader limitations in LMMs’ capacity for structured visual narrative processing, a key requirement for coherent humor understanding.

Text Sequence. Text sequence recognition assesses a model’s ability to reconstruct the correct order of dialogue or captions within comics. Ta-

Model	Panel Acc.	Text Acc.	WER	CER
GPT-4o	0.614	0.326	0.230	0.241
Gemini-1.5-Pro	0.645	0.258	0.328	0.339
Qwen2-VL 72B	0.343	0.176	0.443	0.496
Gemma3 27B	0.310	0.141	0.366	0.419
LLaVA-OV 7B	0.422	0.130	0.415	0.515
Qwen2-VL 7B	0.314	0.103	0.463	0.485

Table 8: Experiment results for sequence recognition tasks. The best results are **bold**.

ble 8 summarizes the results. GPT-4o achieved the highest accuracy (0.326) with the lowest error rates (WER = 0.230, CER = 0.241), demonstrating strong alignment between text elements and narrative structure. Gemini-1.5-Pro followed with slightly lower performance (Text Acc. = 0.258, WER = 0.328, CER = 0.339), while all open-source models lagged significantly. Qwen2-VL-7B performed worst (Text Acc. = 0.103, WER = 0.463, CER = 0.485).

Qualitative analysis for text sequencing revealed that larger open-source models, such as Qwen2-VL-72B and Gemma3-27B, occasionally produced repetitive or degenerate outputs (e.g., “*Ctrl-C, Ctrl-V...*”), highlighting weaknesses in maintaining coherent narrative flow despite increased model capacity. Smaller models frequently hallucinated or omitted key text elements, further disrupting generation.

These findings underscore that even advanced LMMs face persistent challenges in structured multimodal reasoning. Bridging the gap between panel and text sequencing performance will require training strategies that more effectively model sequential dependencies across modalities.

6 Discussion

This work introduced PixelHumor, a novel benchmark designed to evaluate LMMs ability to understand humor in online comics. Through comprehensive experiments across four core tasks, humor identification, humor classification, humor interpretation, and sequence recognition, we provided the first in-depth assessment of LMMs’ multimodal humor comprehension.

Our results reveal that while LMMs achieve near-perfect accuracy in detecting the presence of humor, they struggle with deeper aspects such as narrative sequencing, subtle humor style classification, and modality attribution. Sequence recognition results show that even top models often rely on conventional reading heuristics rather than accurately re-

constructing narrative flow. In humor classification, models consistently excel at explicit styles like *personification* but perform poorly on nuanced categories such as *sarcasm* and *dark humor*. Additionally, humor interpretation tasks highlight significant degradation in reasoning quality as narrative complexity increases.

These findings have important implications for the development of future multimodal AI systems. Our results suggest that current LMMs rely heavily on surface-level heuristics—such as favoring the textual modality—rather than engaging in true multimodal contextual and sequential reasoning. To genuinely comprehend humor, models must move beyond pattern recognition to understand causal, temporal, and multimodal relationships that underpin comic narratives. This will require advances such as hierarchical modeling of narrative structures, enabling models to track setups, twists, and punchlines across multiple panels and modalities. Moreover, improved cross-modal fusion mechanisms are needed to better integrate visual and textual cues, particularly for humor that emerges from their interaction rather than from either modality alone.

Future work could focus on developing models that better track long-range multimodal dependencies, distinguish finer-grained humor mechanisms, and adapt to diverse cultural humor norms. Extending PixelHumor to cover broader linguistic and stylistic variations could further stress-test LMMs’ reasoning limits. Future work could also cover web comics in other languages to assess an LMM’s ability to comprehend humor in multilingual settings.

7 Conclusion

PixelHumor is a challenging benchmark, comprised of web comics, to assess an LMM’s ability to comprehend and interpret humor. Our experiments highlight fundamental limitations in LMMs, and show that they are still far behind humans in complex humor-related tasks: in particular, humor interpretation where human written explanations are still largely preferred; humor style classification where models struggle to identify darker and sarcastic styles; and sequence recognition which is required in the build-up to a humorous punchline. This work lays the foundation for building AI systems that can engage more socio-intelligently with one of the most complex forms of human communication: humor.

Limitations

While PixelHumor marks a significant advancement in evaluating humor comprehension in LMMs, several limitations remain. Humor is inherently subjective, and despite rigorous annotation guidelines, screening, and training, individual biases may influence interpretations. Additionally, PixelHumor focuses on static visual-textual humor in comics, excluding temporal and dynamic elements present in videos and animations, which are critical to humor in other media formats. Furthermore, as the dataset is primarily in English and primarily based in Western media, it may over-represent Western humor conventions, limiting cross-cultural generalizability. Expanding PixelHumor to include multilingual and culturally diverse humor sources would be a valuable direction for future research.

Ethical Consideration

We collected the web comics from a diversified public online sources to capture a broad spectrum of humor styles. However, comics such as Cyanide and Happiness may contain violence, sexually explicit and dark humor that might be uncomfortable for some readers. We are including these comics as they are still valuable for future work in understanding dark humor, hate speech and their connections in intelligent multimodal systems. As a precaution, we will separate these comics in the benchmark and label them as “potentially harmful”.

All annotators were recruited through our university’s internal calls for part-time research assistantship. There were no academic incentives towards the annotators and paid for 20 dollars per hour, which exceeds the minimum wage in our local context (9 dollars per hour). The annotators were informed in advance that the dataset may contain examples of dark or sarcastic humor, and they could withdraw or skip any samples they found uncomfortable. Additionally, annotators were compensated fairly and were provided with clear annotation guidelines that focused on humor classification, not value judgments or endorsements of the content.

While releasing our dataset, we do not redistribute the images directly, but instead provide URLs and associated metadata. We also plan to include a detailed usage disclaimer and data documentation sheet (datasheet for datasets) that explicitly warns users of potential content sensitiv-

ities and encourages ethical use of the dataset in research contexts.

Acknowledgement

This research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG3-AMP-2024-08-001). We also sincerely thank Bryan Tan (Chen Zhengyu) and Jia Wang Peh for their additional annotations and evaluations, which substantially improved this work.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Salvatore Attardo. 1997. The semantic foundations of cognitive theories of humor.
- Salvatore Attardo. 2009. *Linguistic theories of humor*. Walter de Gruyter.
- Dario Bertero and Pascale Fung. 2016. [Deep learning of audio and language features for humor prediction](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 496–501, Portorož, Slovenia. European Language Resources Association (ELRA).
- Rui Cao, Ming Shan Hee, Adriel Kuek, Wen-Haw Chong, Roy Ka-Wei Lee, and Jing Jiang. 2023a. [Procap: Leveraging a frozen vision-language model for hateful meme detection](#). In *Proceedings of the 31st ACM International Conference on Multimedia, MM ’23*, page 5244–5252, New York, NY, USA. Association for Computing Machinery.
- Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. 2023b. [Prompting for multimodal hateful meme classification](#). *Preprint*, arXiv:2302.04156.
- Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. [Towards multimodal sarcasm detection \(an _Obviously_ perfect paper\)](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4619–4629, Florence, Italy. Association for Computational Linguistics.
- Jorge Cham. 2024. [Piled higher and deeper \(phd comics\)](#).
- 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.

- Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.
- Lei Chen and Chong Min Lee. 2017. [Predicting audience’s laughter during presentations using convolutional neural network](#). In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 86–90, Copenhagen, Denmark. Association for Computational Linguistics.
- Yuyan Chen, Yichen Yuan, Panjun Liu, Dayiheng Liu, Qinghao Guan, Mengfei Guo, Haiming Peng, Bang Liu, Zhixu Li, and Yanghua Xiao. 2024. 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.
- Mary Crawford. 2003. Gender and humor in social context. *Journal of pragmatics*, 35(9):1413–1430.
- Dineh Davis. 2008. *Communication and humor*. na.
- Jim Davis. 2024. [Garfield](#).
- ExplosmEntertainment. 2024. [Cyanide & happiness](#).
- Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. 2019. [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.
- Andrew F Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89.
- Ming Shan Hee, Wen-Haw Chong, and Roy Ka-Wei Lee. 2023. [Decoding the underlying meaning of multimodal hateful memes](#). *Preprint*, arXiv:2305.17678.
- Ming Shan Hee, Zihan Gao, Yinglong Wang, Xiangxiang Chu, Roy Ka-Wei Lee, and Zengchang Qin. 2025. [Contrastive instruction fine-tuning large multimodal model for hateful meme classification](#). *Proceedings of the International AAAI Conference on Web and Social Media*, 19(1):760–773.
- Ming Shan Hee and Roy Ka-Wei Lee. 2025. [Demystifying hateful content: Leveraging large multimodal models for hateful meme detection with explainable decisions](#). *Preprint*, arXiv:2502.11073.
- Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023. [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.
- Veedant Jain, Felipe dos Santos Alves Feitosa, and Gabriel Kreiman. 2024. Is ai fun? humordb: a curated dataset and benchmark to investigate graphical humor. *arXiv preprint arXiv:2406.13564*.
- Sophie Jentzsch and Kristian Kersting. 2023. Chatgpt is fun, but it is not funny! humor is still challenging large language models. *arXiv preprint arXiv:2306.04563*.
- Junhui Ji, Wei Ren, and Usman Naseem. 2023. [Identifying creative harmful memes via prompt based approach](#). In *Proceedings of the ACM Web Conference 2023, WWW ’23*, page 3868–3872, New York, NY, USA. Association for Computing Machinery.
- Tonglin Jiang, Hao Li, and Yubo Hou. 2019. Cultural differences in humor perception, usage, and implications. *Frontiers in psychology*, 10:123.
- James Kelly. 2024. [They can talk](#).
- Giselinde Kuipers et al. 2008. *The sociology of humor*. na.
- Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Xiangji Huang. 2023. A systematic study and comprehensive evaluation of chatgpt on benchmark datasets. *arXiv preprint arXiv:2305.18486*.
- 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, MM ’21*, page 5138–5147. ACM.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.
- Hongzhan Lin, Ziyang Luo, Wei Gao, Jing Ma, Bo Wang, and Ruichao Yang. 2024. [Towards explainable harmful meme detection through multimodal debate between large language models](#). In *Proceedings of the ACM Web Conference 2024, WWW ’24*, page 2359–2370, New York, NY, USA. Association for Computing Machinery.
- Rod A Martin and Thomas Ford. 2018. *The psychology of humor: An integrative approach*. Academic press.
- A Peter McGraw and Caleb Warren. 2010. Benign violations: Making immoral behavior funny. *Psychological science*, 21(8):1141–1149.

- 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.
- Piotr Mirowski, Juliette Love, Kory Mathewson, and Shakir Mohamed. 2024. A robot walks into a bar: Can language models serve as creativity supporttools for comedy? an evaluation of llms’ humour alignment with comedians. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1622–1636.
- Shreyash Mishra, S Suryavardan, Parth Patwa, Megha Chakraborty, Anku Rani, Aishwarya Reganti, Aman Chadha, Amitava Das, Amit Sheth, Manoj Chinakotla, et al. 2023. Memotion 3: Dataset on sentiment and emotion analysis of codemixed hindi-english memes. *arXiv preprint arXiv:2303.09892*.
- Mia Moody-Ramirez and Andrew B Church. 2019. Analysis of facebook meme groups used during the 2016 us presidential election. *Social Media+ Society*, 5(1):2056305118808799.
- John Morreall. 1982. A new theory of laughter. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 42(2):243–254.
- John Morreall. 1986. The philosophy of laughter and humor.
- Randall Munroe. 2024. [xkcd](#).
- Abhilash Nandy, Yash Agarwal, Ashish Patwa, Milon 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*.
- Abel L Peirson V and E Meltem Tolunay. 2018. Dank learning: Generating memes using deep neural networks. *arXiv preprint arXiv:1806.04510*.
- Nirmalendu Prakash, Ming Shan Hee, and Roy Ka-Wei Lee. 2023a. [Totaldefmeme: A multi-attribute meme dataset on total defence in singapore](#). In *Proceedings of the 14th ACM Multimedia Systems Conference, MMSys ’23*, page 369–375. ACM.
- Nirmalendu Prakash, Han Wang, Nguyen Khoi Hoang, Ming Shan Hee, and Roy Ka-Wei Lee. 2023b. [Prompttopic: Unsupervised multimodal topic modeling of memes using large language models](#). In *Proceedings of the 31st ACM International Conference on Multimedia, MM ’23*, page 621–631. ACM.
- Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.
- Aadhavan Sadasivam, Kausic Gunasekar, Hasan Davulcu, and Yezhou Yang. 2020. Memebot: Towards automatic image meme generation. *arXiv preprint arXiv:2004.14571*.
- Charles M. Schulz. 2024. [Peanuts](#).
- Viriya Taecharungroj and Pitchanut Nueangjamnong. 2015. Humour 2.0: Styles and types of humour and virality of memes on facebook. *Journal of Creative Communications*, 10(3):288–302.
- 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.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.
- Shaun Toh, Adriel Kuek, Wen-Haw Chong, and Roy Ka-Wei Lee. 2023. [Mermaid: A dataset and framework for multimodal meme semantic understanding](#). In *2023 IEEE International Conference on Big Data (BigData)*, pages 433–442.
- Han Wang and Roy Ka-Wei Lee. 2024. Memecraft: Contextual and stance-driven multimodal meme generation. In *Proceedings of the ACM on Web Conference 2024*, pages 4642–4652.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.
- Zach Weinersmith. 2024. [Saturday morning breakfast cereal \(smbc\)](#).
- Julia Wilkins and Amy Janel Eisenbraun. 2009. Humor theories and the physiological benefits of laughter. *Holistic nursing practice*, 23(6):349–354.
- Jiaming Wu, Hongfei Lin, Liang Yang, and Bo Xu. 2021. Mumor: A multimodal dataset for humor detection in conversations. In *Natural Language Processing and Chinese Computing: 10th CCF International Conference, NLPCC 2021, Qingdao, China, October 13–17, 2021, Proceedings, Part I 10*, pages 619–627. Springer.
- Di Yi 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, Lisbon, Portugal. Association for Computational Linguistics.

Jeremy A Yip and Rod A Martin. 2006. Sense of humor, emotional intelligence, and social competence. *Journal of research in personality*, 40(6):1202–1208.

Jifan Zhang, Lalit Jain, Yang Guo, Jiayi Chen, Kuan Lok Zhou, Siddharth Suresh, Andrew Wagenmaker, Scott Sievert, Timothy Rogers, Kevin Jamieson, et al. 2024. Humor in ai: Massive scale crowd-sourced preferences and benchmarks for cartoon captioning. *arXiv preprint arXiv:2406.10522*.

A Comparison with Existing Datasets

Table ?? shows the detailed difference between PixelHumor and other humor datasets. While we acknowledge the existence of prior humor-related datasets, these resources typically focus on single-panel, single-modality inputs and are limited to narrow task scopes, such as binary humor classification or caption ranking. In contrast, PixelHumor introduces a new paradigm for humor understanding that centers on multimodal, multi-panel narrative reasoning. Our dataset supports four distinct yet complementary tasks: humor detection, multi-label humor classification, explanation via panel contribution, and sequence reconstruction, offering a comprehensive and fine-grained benchmark for evaluating LMMs. To our knowledge, this is the first dataset to jointly evaluate LMMs’ capabilities across both temporal structure and modality attribution in humor, making PixelHumor a uniquely valuable resource for advancing this research area.

B Annotation Guidelines and Questions

Annotators answered the following five questions for each comic, with a clear purpose and rationale outlined for each task:

1. *Do you understand the humor in this comic? (Y/N)*

Purpose: To determine whether humor is present in the comic.

Rationale: Focuses on identifying the intended humor, independent of the annotator’s personal amusement, ensuring objectivity.

2. *Do the words with sound effects contribute to the humor in this comic? (MCQ: Present, contribute; Present, do not contribute; Absent) Choose Absent if there are no sound effects in this comic.*

Purpose: To assess the contribution of onomatopoeic sound effects to the comic’s humor.

Rationale: This objective measure serves as a control question for quality checks.

3. *Type the panel number that contributes the most to the humor. (NA if no humor is present)*

Purpose: To identify the punchline panel.

Rationale: Helps pinpoint the comic’s critical humorous element, even with randomized panel numbering.

Dataset	Modality	Single / Multi panel	Humor Styles	Task Coverage	Sequential Reasoning	Attribution Granularity
PixelHumor (ours)	I + T	Multi-panel	8 styles (multi-label)	Humor detection, humor interpretation, humor classification, sequence recognition	Yes	Panel + Modality
New Yorker Caption Contest (Hessel et al., 2023)	I + T	Single-panel	N/A	Caption evaluation (ranking, matching)	No	Whole image-text pair
Memotion 3.0 (Mishra et al., 2023)	I + T	Multi-panel	Sarcasm	Sentiment analysis, emotion classification, scales/intensity of emotion classes	No	Whole image-text pair
YesBut (Nandy et al., 2024)	I + T	Multi-panel	Satire	Satirical image detection, satirical image understanding, satirical image completion	No	Whole image-text pair
HumorDB (Jain et al., 2024)	I	Single-panel	N/A	Detection, interpretation, humor comparison	No	Image
TalkFunny (Chen et al., 2024)	T	N/A	Affiliative, self-enhancing, aggressive, self-defeating	Humor sentiment style classification, humor generation	No	Text-only
One-liners (Mihalcea and Strapparava, 2005)	T	N/A	Alliteration, antonymy, adult slang	Humor detection	No	Text-only
Pun of the day (Yang et al., 2015)	T	N/A	Pun	Humor detection, humor anchor extraction	No	Text-only
Ted-Laughter (Chen and Lee, 2017)	T	N/A	N/A	Humor detection	No	Text-only

Table 9: Summary of existing humor datasets. T: Text, I: Image, V: Video, A: Audio.

4. *Do the text or the visuals in this comic contribute more to the humor? (MCQ: Text, Visual, Both - Equal Contribution, NA) Indicate NA if this comic is not humorous to you.*

Purpose: To assess the relative importance of text and visuals in the comic’s humor.

Rationale: Evaluates multimodal reasoning, where humor may arise from the interaction of textual and visual elements.

5. *Which humor styles best describe the comic? (Select all that apply or NA)*

Purpose: To classify the comic into relevant humor styles from the defined taxonomy.

Rationale: Enables a nuanced categorization of humor styles, ensuring alignment with the dataset’s taxonomy.

C Humor Interpretation Guidelines

Evaluators are presented with an interpretation of the humor in the comic. They are to evaluate this explanation by giving it a score from 1 (strongly disagree) to 7 (strongly agree) and state their justifications for their score. The pointers below were provided to help the evaluators rate and write their justifications for the model’s explanation:

- Give an overall justification or reason for the rating of the comic.
- Reference exactly which part of the explanation is accurate or inaccurate with direct quotes.
- Elaborate on any points that the explanation may have missed out or are accurate in their inference (e.g. being able to infer the true intentions of the narrative which are commonly found in styles like sarcasm).
- Reference the taxonomy of humor styles (Table 1) as detailed in Question 5 as needed.

The participants **must fulfill at least two out of four of these pointers** in their justification for

their score to be accepted as a valid evaluation. Below, we provided examples of accepted or rejected justifications.

Example Justification for Humor Interpretations

Example Explanation: "The comic's humor stems from the unexpected juxtaposition of traditional religion and modern technology. We see a priest soliciting a donation, a common practice, but the punchline subverts our expectations. Instead of offering a typical excuse, the person being solicited implies they can donate digitally through an app called "Papal," a playful reference to the Pope. This unexpected twist, highlighting the increasing digitization of even traditionally cash-based interactions like charitable giving, creates the humor."

Good/Accepted Justification: The overall explanation is accurate, but the phrase "humor stems from the unexpected juxtaposition of traditional religion" is inaccurate as the main humor style is "surprise". The comic also uses a pun which was not reflected in the explanation. Rating: 5/7

D Additional Data Analysis

Data Source	Max	Min	Average
<i>Happiness and Cyanide</i>	17	1	4.34
<i>Peanuts</i>	11	2	4.53
<i>Garfield</i>	8	2	3.50
<i>XKCD</i>	10	1	2.57
<i>PhD Comics</i>	8	1	3.67
<i>They Can Talk</i>	7	1	2.87
<i>SMBC</i>	18	1	3.49

Table 10: Statistical distribution of panels in PixelHumor dataset breakdown by data source.

Data Source	Max	Min	Average
<i>Happiness and Cyanide</i>	161	0	24.87
<i>Peanuts</i>	152	0	38.10
<i>Garfield</i>	90	1	18.66
<i>XKCD</i>	253	0	39.29
<i>PhD Comics</i>	140	0	46.71
<i>They Can Talk</i>	45	0	15.27
<i>SMBC</i>	501	0	59.79

Table 11: Statistical distribution of words in PixelHumor dataset breakdown by data source.

Tables 10 and 11 present the statistical distributions of panels and words in the PixelHumor dataset, respectively. The dataset includes highly complex comics, with some featuring up to 18 panels and 501 words, as observed in *SMBC*. On average, comics from *Peanuts* and *Happiness and Cyanide* tend to have a higher number of panels, while those from *PhD Comics* and *SMBC* are notably more text-heavy.

Table 12 provides a breakdown of the PixelHumor dataset by data source and humor styles, showcasing the diversity of humor types represented across the various sources.

E Task-Specific Humor Evaluation Prompts

We designed the following prompts to evaluate LMMs on the four humor evaluation tasks outlined in Section 4.1. Questions 1–4 assess humor identification, testing the model’s ability to detect and analyze the presence of humor. Question 5 evaluates humor classification, supported by guidelines and definitions of the humor styles. Question 6 focuses on humor interpretation, requiring the model to articulate the reasoning behind the humor. Finally, Questions 7 and 8 address sequence recognition, evaluating the model’s ability to understand the intended order of textual and visual elements.

Prompts for Humor Identification Task

System prompt: *You are a humorous assistant that understands comics. You will be given comics and your task is to evaluate the comics.*

Question 1: *Do you understand the humor of this comics? Please output only a single word answer "Yes" or "No".*

Question 2: *Analyze the comic and respond based on the following criteria regarding text-based sound effects: (a) If there are no sound effects present in the comic, output "Absent". (b) If sound effects are present and contributing to the humor, output "Present, contribute". (c) If sound effects are present but do not contribute to the humor, output "Present, do not contribute".*

Question 3: *Which panel contributes the most to the humor of this comic? Please output only the labeled panel number.*

Question 4: *Is the text or the visual modality more important to the humor in this comic? Output "Both" if both modalities contribute humor to the comic. Please output only a single word answer "Text", "Visual", "Both".*

Data Source	Com.	Per.	Exa.	Pun.	Sar.	Sil.	Sur.	Dar.	N/A
Happiness and Cyanide	13	7	28	95	7	58	145	42	5
Peanuts	8	76	92	9	22	87	96	2	8
Garfield	7	131	43	5	59	85	69	1	0
XKCD	44	5	46	20	30	130	98	11	16
PhD Comics	59	4	104	20	39	56	116	1	1
They Can Talk	13	336	2	1	4	17	25	1	1
SMBC	47	14	81	14	13	97	106	26	2

Table 12: Breakdown of PixelHumor dataset by data source and humor types. **Com.:** Comparison, **Per.:** Personification, **Exa.:** Exaggeration, **Pun.:** Pun, **Sar.:** Sarcasm, **Sil.:** Silliness, **Sur.:** Surprise, **Dar.:** Dark, **N/A:** Not Applicable.

Prompts for Humor Classification Task

System prompt: You are a humorous assistant that understands comics. You will be given comics and your task is to evaluate the comics.

Question 5: Which humor styles best describe the comic? Here are some guidelines for each humor style.

Comparison: This comic compares two or more objects/ideas to reference the differences or similarities. This comic is funny because of this comparison.

Personification: This comic has at least one animal/creature/plant that acts like a human (talking, running on two legs etc.). This comic is funny because of this personified creature/plant.

Exaggeration: This comic attempts to exaggerate (overemphasize or magnify) something out of proportion. This comic is funny because of this exaggeration/absurdity.

Pun: This comic is funny because of the linguistic elements. Linguistic elements include: uncommon uses of language, double-meanings in phrases or words etc.

Sarcasm: This comic expresses an idea/thought that is not the real intention of the character/comic. This comic is funny because of the sarcasm present.

Silliness: There are elements in the comic which are absurd and/or ridiculous. The characters are or did something foolish. This comic is funny because of the silly elements.

Surprise: There was a twist in the narrative or an unexpected element in the comic. This comic is funny because of the twist or unexpected elements.

Dark: There are potentially sensitive, taboo or ideas that violate the norm in this comic where if taken out of context in this comic, might be offensive to others. This comic is because of these benign violations or the dark humor present.

You may select multiple humor styles but output only the humor styles "Comparison", "Personification", "Exaggeration", "Pun", "Sarcasm", "Silliness", "Surprise" or "Dark".

Prompts for Humor Interpretation Task

System prompt: You are a humorous assistant that understands comics. You will be given comics and your task is to evaluate the comics.

Question 6: Explain why this comic is funny or not funny in 3 sentences.

Prompts for Sequence Recognition Task

System prompt: You are a humorous assistant that understands comics. You will be given comics and your task is to evaluate the comics.

Question 7: In what order should the panels be read? Respond with the panel numbers only. Write the panel numbers followed by a comma. For example the answer "3,4,2,1" will mean that panels will be read in order of panel 3, then panel 4, then panel 2 and finally panel 1.

Question 8: For each panel, what are the text inside? Respond as {panel_number}: {text_within_panel}.

F Quantitative Error Analysis

We have also conducted a structured, quantitative error analysis by (i) providing confusion matrices for humor style classification across models, and (ii) presenting F1 score breakdowns across five reasoning tasks, segmented by humor style and number of comic panels. These allow us to identify patterns in model reasoning failures, for instance, sarcasm and surreal humor consistently present challenges across models, and performance tends to degrade as panel count increases. This deeper diagnostic view supports targeted directions for future work, such as better context modeling and culturally-informed reasoning modules.

Table 13 presents the confusion matrix for humor style classification task for each model. From the confusion matrix, we can see that models classified each of the comics to have multiple humor styles according to Table 4 predicted larger amount of False Positive classification, which lead to a lower precision score and hence low F1 score. On the

other hand, LLaVA-OV 7B which only predicted 1 humor style per comic has larger amount of False Negatives while most classifications on Sarcasm are actually False Positives.

Table 14 shows the F1 scores for each task broken down by humor styles for each model. In Task 3, panel contribution task, all tested models have better performance on comics with Personification as one of the humor styles as personification is more straightforward to identify than other humor styles. On the other hand, Exaggeration performs worse than most of the other humor styles. For Task 4, modality contribution task, most of the models (except Gemma3-27B and Qwen2-VL-7B) can identify the modality contributing to sarcastic comic better. While interestingly most of them failed to correctly identify the modality contribution for those comics with personification as humor style. (RY: I didn't include for Task 5, since it seems to be repeating Table 3, but that table is Recall instead of F1.)

Table 15 shows the F1 scores for each task broken down by number of panels for each model. For Task 3 to Task 5, we can see that when the number of panels increases, the model performance degrades, showing the model's capability in understanding long-form humor remains challenging.

G Case Studies

In all case studies, the XKCD comics are used for non-commercial purposes under the Creative Commons Attribution-NonCommercial 2.5 License.

Table 16 presents a case where models struggled with humor identification and classification but performed well in humor interpretation. While LLaVA-OV achieved the highest scores in both tasks, its explanation was widely disagreed upon by annotators, suggesting that high classification accuracy does not necessarily correlate with meaningful humor interpretation. In contrast, other models with lower classification scores provided interpretations that made more sense to human evaluators, highlighting the complexity of aligning AI-generated humor reasoning with human perception.

A common failure in sequence recognition is that models default to conventional reading orders, incorrectly associating speech bubbles with different panels. Table 17 illustrates how models fail to fully grasp context, even when they correctly recognize text. Both Gemini-1.5-Pro and LLaVA-OV produced hallucinated responses, introducing de-

tails absent from the comic, such as body language and facial expressions. Meanwhile, GPT-4o and Qwen2-VL failed to capture the punchline, missing the subtle wordplay on "time travel" and the gradual realization conveyed in the comic. Even when GPT-4o correctly identified the panel sequence and text, it struggled to connect them to the underlying humor, reinforcing that sequential reasoning alone is insufficient for humor comprehension.

Table 18 presents an intriguing contrast. GPT-4o misidentified the panel sequence but correctly mapped text to panels, producing an interpretation that annotators found reasonable. In contrast, Gemini-1.5-Pro correctly aligned the panels and text but failed to provide a humor explanation that resonated with human evaluators. This suggests that perfect sequence recognition may not be essential for humor comprehension, raising questions about the relative importance of narrative alignment versus inferential reasoning. Further studies are needed to explore this phenomenon and determine whether narrative structure alone is sufficient for AI to understand humor, or if deeper contextual reasoning is required.

Model	GPT-4o				Gemini-1.5-Pro				Qwen2-VL-72B			
Styles	TP	FP	TN	FN	TP	FP	TN	FN	TP	FP	TN	FN
Com.	143	364	2196	97	101	144	2416	139	73	69	2491	167
Per.	625	437	1715	23	531	305	1847	117	544	296	1856	104
Exa.	383	864	1431	122	243	437	1858	262	263	515	1780	242
Pun	132	220	2355	93	106	243	2332	119	92	221	2354	133
Sar.	152	571	1962	115	228	1578	955	39	67	314	2219	200
Sil.	393	793	1344	270	434	1049	1088	229	177	342	1795	486
Sur.	647	778	1114	261	716	978	914	192	116	88	1804	792
Dar.	129	356	2271	44	101	207	2420	72	62	61	2566	111
N/A	1	1	2766	32	3	5	2762	30	6	12	2755	27

Model	Gemma3-27B				LLaVA-OV 7B				Qwen2-VL 7B			
Styles	TP	FP	TN	FN	TP	FP	TN	FN	TP	FP	TN	FN
Com.	107	186	2374	133	93	302	2258	147	112	414	2146	128
Per.	497	293	1859	151	86	199	2096	419	234	624	1671	271
Exa.	291	580	1715	214	46	13	2139	602	568	347	1805	80
Pun	75	114	2461	150	17	63	2512	208	61	323	2252	164
Sar.	184	1120	1413	83	207	1765	768	60	27	185	2348	240
Sil.	607	1813	324	56	3	3	2134	660	36	79	2058	627
Sur.	784	1278	614	124	0	1	1891	908	1	0	1892	907
Dar.	143	605	2022	30	0	0	2627	173	41	53	2574	132
N/A	3	2	2765	30	0	2	2765	33	1	4	2763	32

Table 13: Confusion matrix of the model prediction on the Humor Classification task broken down by humor style. TP: True Positive, FP: False Positive, TN: True Negative, FN: False Negative.

Model	GPT-4o					Gemini-1.5-Pro				
Styles	Task 1	Task 2	Task 3	Task 4	Task 5	Task 1	Task 2	Task 3	Task 4	Task 5
Com.	1.000	0.911	0.760	0.690	0.594	0.998	0.878	0.683	0.649	0.477
Per.	1.000	0.928	0.835	0.512	0.763	1.000	0.879	0.772	0.525	0.690
Exa.	1.000	0.890	0.681	0.621	0.660	0.999	0.844	0.661	0.611	0.522
Pun	1.000	0.935	0.773	0.669	0.609	1.000	0.904	0.778	0.661	0.544
Sar.	1.000	0.910	0.777	0.764	0.586	0.998	0.883	0.741	0.749	0.708
Sil.	1.000	0.921	0.777	0.624	0.605	0.999	0.870	0.713	0.583	0.630
Sur.	0.999	0.901	0.782	0.641	0.664	0.999	0.870	0.728	0.637	0.692
Dar.	0.997	0.888	0.764	0.576	0.688	0.997	0.829	0.765	0.556	0.610
N/A	0.059	0.985	0.059	0.059	0.059	0.167	0.935	0.167	0.167	0.167

Model	Qwen2-VL-72B					Gemma3-27B				
Styles	Task 1	Task 2	Task 3	Task 4	Task 5	Task 1	Task 2	Task 3	Task 4	Task 5
Com.	1.000	0.822	0.528	0.647	0.415	1.000	0.836	0.547	0.229	0.504
Per.	1.000	0.784	0.531	0.456	0.678	1.000	0.812	0.597	0.311	0.665
Exa.	0.998	0.735	0.462	0.611	0.513	1.000	0.787	0.474	0.171	0.563
Pun	0.998	0.792	0.434	0.617	0.472	1.000	0.824	0.536	0.171	0.442
Sar.	0.998	0.811	0.474	0.651	0.353	1.000	0.832	0.531	0.118	0.622
Sil.	0.998	0.765	0.508	0.556	0.381	0.999	0.808	0.560	0.233	0.737
Sur.	0.997	0.751	0.443	0.582	0.253	0.999	0.790	0.502	0.186	0.716
Dar.	0.988	0.747	0.468	0.501	0.415	0.994	0.805	0.535	0.198	0.692
N/A	0.308	0.862	0.308	0.308	0.308	0.167	0.900	0.167	0.167	0.167

Model	LLaVA-OV-7B					Qwen2-VL-7B				
Styles	Task 1	Task 2	Task 3	Task 4	Task 5	Task 1	Task 2	Task 3	Task 4	Task 5
Com.	1.000	0.761	0.459	0.605	0.394	1.000	0.042	0.575	0.240	0.488
Per.	1.000	0.740	0.552	0.485	0.111	1.000	0.029	0.570	0.281	0.663
Exa.	0.999	0.719	0.479	0.589	0.200	0.999	0.029	0.474	0.207	0.439
Pun	1.000	0.805	0.463	0.563	0.098	1.000	0.017	0.450	0.232	0.340
Sar.	1.000	0.791	0.498	0.656	0.582	1.000	0.025	0.509	0.075	0.208
Sil.	0.999	0.754	0.512	0.583	0.030	0.998	0.048	0.547	0.257	0.140
Sur.	0.999	0.757	0.480	0.565	0.028	0.999	0.025	0.455	0.200	0.082
Dar.	1.000	0.761	0.523	0.453	0.008	0.994	0.026	0.431	0.336	0.253
N/A	0.000	0.862	0.000	0.000	0.000	0.059	0.000	0.059	0.059	0.059

Table 14: F1 score broken down by humor style for each task. Task 1: Humor Presence, Task 2: Sound Effect presence, Task 3: Panel Contribution, Task 4: Modality Contribution, Task 5: Humor Style Classification.

Model	GPT-4o					Gemini-1.5-Pro				
# Panels	Task 1	Task 2	Task 3	Task 4	Task 5	Task 1	Task 2	Task 3	Task 4	Task 5
1	0.959	0.976	0.870	0.718	0.495	0.961	0.945	0.785	0.672	0.456
2	0.978	0.933	0.887	0.542	0.594	0.978	0.925	0.807	0.540	0.564
3	0.993	0.919	0.827	0.584	0.479	0.993	0.881	0.810	0.562	0.458
4	0.987	0.913	0.757	0.641	0.524	0.987	0.871	0.721	0.636	0.504
5	0.967	0.829	0.538	0.567	0.528	0.967	0.788	0.556	0.575	0.511
6	1.000	0.886	0.728	0.598	0.493	1.000	0.817	0.665	0.572	0.538
7	1.000	0.928	0.658	0.598	0.470	1.000	0.818	0.681	0.616	0.494
8	0.972	0.792	0.595	0.486	0.477	0.972	0.766	0.603	0.581	0.481
9	0.926	1.000	0.567	0.486	0.470	0.926	0.952	0.242	0.617	0.451
10	1.000	0.855	0.355	0.609	0.576	1.000	0.757	0.414	0.639	0.427
11	1.000	0.904	0.133	0.630	0.399	1.000	0.836	0.207	0.778	0.487
12	1.000	1.000	0.667	0.556	0.875	1.000	0.667	0.667	0.556	0.625
13	1.000	0.333	0.000	0.333	0.667	1.000	0.333	0.500	0.333	0.667
14	1.000	1.000	0.000	0.000	1.000	1.000	1.000	0.000	0.000	1.000
17	1.000	1.000	0.000	0.000	1.000	1.000	0.000	1.000	0.000	1.000
18	1.000	1.000	1.000	1.000	0.000	1.000	1.000	0.000	1.000	0.000

Model	Qwen2-VL-72B					Gemma3-27B				
# Panels	Task 1	Task 2	Task 3	Task 4	Task 5	Task 1	Task 2	Task 3	Task 4	Task 5
1	0.969	0.900	0.844	0.675	0.372	0.966	0.907	0.876	0.345	0.461
2	0.974	0.820	0.600	0.549	0.450	0.978	0.840	0.529	0.337	0.590
3	0.992	0.735	0.407	0.473	0.347	0.993	0.786	0.533	0.185	0.433
4	0.987	0.779	0.364	0.615	0.373	0.987	0.823	0.416	0.160	0.492
5	0.967	0.631	0.386	0.586	0.396	0.967	0.652	0.306	0.197	0.478
6	0.994	0.705	0.302	0.425	0.387	1.000	0.735	0.375	0.172	0.476
7	1.000	0.687	0.474	0.579	0.340	1.000	0.748	0.410	0.160	0.461
8	0.962	0.708	0.466	0.450	0.273	0.972	0.694	0.332	0.172	0.460
9	0.926	0.863	0.383	0.664	0.356	0.926	0.952	0.197	0.181	0.465
10	1.000	0.563	0.268	0.641	0.455	1.000	0.633	0.205	0.236	0.487
11	1.000	0.728	0.133	0.519	0.462	1.000	0.904	0.211	0.274	0.398
12	1.000	0.667	0.444	0.556	0.650	1.000	1.000	0.000	0.167	0.400
13	1.000	0.333	0.500	0.333	0.250	1.000	0.333	0.000	0.333	0.500
14	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.000	1.000
17	1.000	0.000	0.000	0.000	1.000	1.000	0.000	0.000	0.000	1.000
18	1.000	1.000	0.000	1.000	0.000	1.000	1.000	0.000	1.000	0.000

Model	LLaVA-OV-7B					Qwen2-VL-7B				
# Panels	Task 1	Task 2	Task 3	Task 4	Task 5	Task 1	Task 2	Task 3	Task 4	Task 5
1	0.952	0.838	0.763	0.587	0.123	0.952	0.076	0.939	0.286	0.272
2	0.978	0.768	0.442	0.535	0.164	0.978	0.000	0.801	0.341	0.383
3	0.993	0.745	0.421	0.514	0.087	0.996	0.025	0.391	0.186	0.241
4	0.988	0.753	0.441	0.586	0.095	0.987	0.014	0.347	0.167	0.254
5	0.967	0.670	0.458	0.544	0.071	0.967	0.011	0.285	0.247	0.162
6	1.000	0.732	0.365	0.489	0.026	1.000	0.003	0.124	0.264	0.162
7	1.000	0.782	0.359	0.515	0.051	1.000	0.001	0.303	0.198	0.268
8	0.972	0.708	0.448	0.489	0.019	0.972	0.079	0.221	0.242	0.115
9	0.926	0.907	0.193	0.586	0.165	0.926	0.080	0.115	0.172	0.244
10	1.000	0.637	0.139	0.596	0.114	1.000	0.074	0.135	0.330	0.323
11	1.000	0.904	0.000	0.474	0.000	1.000	0.000	0.000	0.316	0.460
12	1.000	1.000	0.444	0.222	0.000	1.000	0.000	0.444	0.167	0.400
13	1.000	0.333	0.500	0.333	0.167	1.000	0.500	0.000	0.333	0.417
14	1.000	1.000	0.000	0.000	0.000	1.000	0.000	1.000	0.000	0.000
17	1.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	1.000	1.000
18	1.000	1.000	0.000	1.000	0.000	1.000	0.000	0.000	1.000	0.000

Table 15: F1 score broken down by number of panels for each task. Task 1: Humor Presence, Task 2: Sound Effect presence, Task 3: Panel Contribution, Task 4: Modality Contribution, Task 5: Humor Style Classification. Note that there are no comics with 15 and 16 panels.

Table 16: Case study 1 on an XKCD comic, access at <https://xkcd.com/234/>. Incorrect classifications are marked in red.

Model	Humor	Sound	Punchline Panel	Modality	Humor Type	Interpretation
Groundtruth	Yes	Yes	1	Text	Comparison	-
GPT-4o	Yes	NA	5	Text	Pun	This comic is funny because it cleverly combines the concept of lock-picking with programming, using technical jargon as a metaphor for escaping restraints. The punchline involves a play on words with "backslashes," which are used in coding, humorously suggesting they could help escape handcuffs. The humor lies in the absurdity of applying programming logic to a physical escape scenario. Score: [6,6]
Gemini-1.5-Pro	Yes	NA	2	Text	Pun, Silliness	This comic is mildly humorous due to the juxtaposition of romanticized lock-picking with the mundane reality of debugging code. The stick figure's script crashing is a relatable experience for programmers. The punchline about backslashes adds a layer of humor specific to coding, further enhancing the relatability for a niche audience. Score: [4,5]
Qwen2-VL 72B	Yes	NA	4	Text	Pun	This comic is funny because it cleverly juxtaposes the fascination with locks and the hacker instinct with the mundane frustration of debugging code. The humor lies in the contrast between the grandiosity of admiring Houdini's lock-picking skills and the petty annoyance of dealing with spaces in a Bash script. The punchline, where the character dismissively says Backslashes? Huh? Never mind, adds to the humor by highlighting the sudden realization that the problem was trivial and easily solved. Score: [2,3]

Continued on next page

Table 16: Case study 1 on an XKCD comic, access at <https://xkcd.com/234/>. Incorrect classifications are marked in red.

Model	Humor	Sound	Punchline Panel	Modality	Humor Type	Interpretation
Groundtruth	Yes	Yes	1	Text	Comparison	-
Gemma3 27B	Yes	NA	2	Both	Comparison, Exaggeration, Silliness	Ah, a classic XKCD! This comic is funny because it perfectly captures the frustrating experience of debugging code – starting with grand philosophical thoughts about locks and hacking, then spiraling down into the mundane reality of whitespace errors. The humor lies in the relatable contrast between the lofty aspirations of a programmer and the petty, infuriating bugs that plague their work, culminating in the realization that the problem was just a backslash. It's a joke that only someone who's stared at a computer screen for too long will *really* get, making it a delightfully nerdy chuckle. Score: [3,2]
LLaVA-OV 7B	Yes	Yes	5	Text	Comparison	The comic is funny because it plays on the idea of a person being fascinated by locks and trying to open them, only to realize that they are not as strong as they thought. The humor comes from the person's realization and the absurdity of the situation. Additionally, the comic includes a twist at the end where the person discovers that the lock was fake, adding to the comedic effect. Score: [1,1]
Qwen2-VL 7B	Yes	NA	2	Both	Pun	This comic is funny because it uses humor to highlight the challenges and frustrations of learning to code. The conversation between the two characters is relatable to anyone who has struggled with coding, and the punchline is unexpected and amusing. The comic also uses visual humor, with the character's expressions and body language adding to the humor. Score: [2,4]

Table 17: Case study 2 on an XKCD comic, access at <https://xkcd.com/209/>. Incorrect recognitions are marked in red.

Model	Sequence Identification	Text Identification	Interpretation
Groundtruth	1,3,2,4	1: COME EXPLORE THE FUTURE WITH ME! HUH? WHAT'S THAT YOU'RE IN? 3: TWO-SEAT KAYAK! I SEE, BUT WHY DO YOU HAVE IT? 2: WE'LL FIND OUT! THE FUTURE IS A BIG PLACE! SO THE KAYAK TRAVELS THROUGH TIME? 4: SURE! JUST LIKE EVERYTHING ELSE! IT ALSO GOES OVER WATER. COME ON!	-
GPT-4o	1,3,2,4	1: COME EXPLORE THE FUTURE WITH ME! 2: HUH? WHAT'S THAT YOU'RE IN? 3: TWO-SEAT KAYAK! 4: I SEE, BUT WHY DO YOU HAVE IT? 5: WE'LL FIND OUT! THE FUTURE IS A BIG PLACE! 6: SO THE KAYAK TRAVELS THROUGH TIME? 7: SURE! JUST LIKE EVERYTHING ELSE! IT ALSO GOES OVER WATER. COME ON!	This comic is funny because it plays on the literal interpretation of exploring the future with a mundane object, a two-seat kayak. The humor lies in the absurdity of using a kayak, which travels over water, as a metaphor for time travel. The punchline cleverly highlights the character's enthusiasm and the unexpected twist of exploring the future in such an ordinary way. Score: [3,1]
Gemini-1.5-Pro	4,2,3,1	1: COME EXPLORE THE FUTURE WITH ME! HUH? WHAT'S THAT YOU'RE IN? 2: A TWO-SEAT KAYAK! I SEE, BUT WHY DO YOU HAVE IT? 3: WE'LL FIND OUT! THE FUTURE IS A BIG PLACE! SO THE KAYAK TRAVELS THROUGH TIME? 4: SURE! JUST LIKE EVERYTHING ELSE! IT ALSO GOES OVER WATER. COME ON!	This comic is funny because of the absurd premise of using a kayak to travel through time. The straight-faced delivery of the first character, despite the obvious lack of logic, adds to the humor. Finally, the second character's acceptance of this flawed logic makes the situation even more ridiculous and funny. Score: [1,1]
Qwen2-VL 72B	1,2,3,4	1: COME EXPLORE THE FUTURE WITH ME! 2: HUH? WHAT'S THAT YOU'RE IN? 3: A TWO-SEAT KAYAK! 4: I SEE, BUT WHY DO YOU HAVE IT? 5: WE'LL FIND OUT! THE FUTURE IS A BIG PLACE! 6: SO THE KAYAK TRAVELS THROUGH TIME? 7: SURE! JUST LIKE EVERYTHING ELSE! IT ALSO GOES OVER WATER. COME ON!	This comic is funny because it plays on the idea of exploring the future, which is typically associated with time travel. The punchline reveals that the "future" the character is referring to is actually just a kayak trip, which is a clever and unexpected twist. The humor comes from the contrast between the grandiose idea of exploring the future and the mundane reality of a kayak ride. Score: [3,6]

Continued on next page

Table 17: Case study 2 on an XKCD comic, access at <https://xkcd.com/209/>. Incorrect recognitions are marked in red.

Model	Sequence Identification	Text Identification	Interpretation
Groundtruth	1,3,2,4	1: COME EXPLORE THE FUTURE WITH ME! HUH? WHAT'S THAT YOU'RE IN? 3: TWO-SEAT KAYAK! I SEE, BUT WHY DO YOU HAVE IT? 2: WE'LL FIND OUT! THE FUTURE IS A BIG PLACE! SO THE KAYAK TRAVELS THROUGH TIME? 4: SURE! JUST LIKE EVERYTHING ELSE! IT ALSO GOES OVER WATER. COME ON!	-
Gemma3 27B	1,2,3,4	1: COME EXPLORE HUH? WHAT'S THE FUTURE THAT YOU'RE IN WITH ME! 2: WOAHSEAT, I SEE, BUT WHY DO YOU HAVE IT? 3: WE'LL FIND OUT! THE FUTURE IS A BIG PLACE! 4: SO THE KAYAK TRAVELS THROUGH TIME! SURE. JUST LIKE EVERYTHING ELSE! IT ALSO GOES OVER WATER. COME ON!	Ah, a classic xkcd! This comic is funny because it takes a wildly complex concept - time travel - and applies the most mundane, literal solution: a kayak. The humor comes from the deadpan delivery and the complete disregard for any actual scientific principles of time travel, culminating in the realization that, of course, a kayak also floats on water. It's the kind of joke that makes you chuckle and then question your own logic. Score: [3,3]
LLaVA-OV 7B	1,2,3,4	1: Come explore the future with me! 2: Uh? What's that you're in? 3: I see, but why do you have it? 4: So the kayak travels through time? It also goes over water. Come on!	The comic is funny because it plays on the idea of a person trying to explain the concept of time travel to someone who is not familiar with it. The humor comes from the absurdity of the situation and the person's frustration in trying to convey a complex idea in simple terms. Additionally, the comic uses visual humor by showing the person's body language and facial expressions as they struggle to explain the concept. Score: [1,1]
Qwen2-VL 7B	1,2,3,4	1: COME EXPLORE THE FUTURE WITH ME! 2: HUH? WHAT'S THAT YOU'RE IN? 3: A TWO-SEAT KAYAK! 4: I SEE, BUT WHY DO YOU HAVE IT? 5: WE'LL FIND OUT! THE FUTURE IS A BIG PLACE! 6: SO THE KAYAK TRAVELS THROUGH TIME? 7: SURE! JUST LIKE EVERYTHING ELSE! IT ALSO GOES OVER WATER. COME ON!	This comic is funny because it plays on the idea of time travel and the absurdity of a two-seat kayak. The dialogue between the characters is also humorous, with the first character's offer to explore the future and the second character's confusion about the kayak's purpose. The final panel, with the second character's response, adds to the humor by emphasizing the absurdity of the situation. Score: [3,3]

Table 18: Case study 3 on an XKCD comic, access at <https://xkcd.com/473/>. Incorrect recognitions are marked in red.

Model	Sequence Identification	Text Identification	Interpretation
Groundtruth	2,1,4,3	2: SHE THREW ME OUT YELLING, "YOU DON'T SAY THOSE WORDS. "NOT IN THIS HOUSE." 1: IT'S BEEN TWO YEARS. I THOUGHT THE WOUNDS HAD HEALED. 4: BUT I STAND BY WHAT I SAID. 3: PLUTO NEVER SHOULD HAVE BEEN A PLANET.	-
GPT-4o	1,2,3,4	1: It's been two years. I thought the wounds had healed. 2: She threw me out yelling, "You don't say those words." "Not in this house." 3: Pluto never should have been a planet. 4: But I stand by what I said.	This comic is funny because it humorously exaggerates the emotional fallout from a scientific debate about Pluto's planetary status. The dramatic setup, akin to a personal or familial conflict, contrasts with the seemingly trivial nature of the argument over Pluto. The punchline reveals the absurdity of being thrown out of a house over an astronomical opinion, highlighting the humor in taking such a stance too seriously. Score: [6,6]
Gemini-1.5-Pro	2,1,4,3	1: IT'S BEEN TWO YEARS. I THOUGHT THE WOUNDS HAD HEALED. 2: HE THREW ME OUT YELLING, "YOU DON'T SAY THOSE WORDS. "NOT IN THIS HOUSE." 3: PLUTO NEVER SHOULD HAVE BEEN A PLANET. 4: BUT I STAND BY WHAT I SAID.	This comic is funny because it highlights the stubbornness of people and their unwillingness to admit they're wrong, even after a long time. The visual gag of the character remaining in the same position outside the house for two years adds to the humor. The absurdity of the situation, particularly the final panel's non-sequitur about Pluto, elevates the comic further. Score: [1,2]
Qwen2-VL 72B	1,2,3,4	1: IT'S BEEN TWO YEARS. 2: SHE THREW ME OUT, YELLING, "YOU DON'T SAY THOSE WORDS." "NOT IN THIS HOUSE." 3: BUT I STAND BY WHAT I SAID. 4: PLUTO NEVER SHOULD HAVE BEEN A PLANET.	This comic is funny because it plays on the absurdity of the situation. The first panel sets up a serious and dramatic tone, suggesting a deep familial conflict. However, the punchline in the last panel reveals that the argument was over the demotion of Pluto from planet status, which is a trivial and unexpected topic for such a heated argument. The contrast between the serious setup and the silly resolution creates a humorous effect. Score: [7,7]

Continued on next page

Table 18: Case study 3 on an XKCD comic, access at <https://xkcd.com/473/>. Incorrect recognitions are marked in red.

Model	Sequence Identification	Text Identification	Interpretation
Groundtruth	2,1,4,3	2: SHE THREW ME OUT YELLING, "YOU DON'T SAY THOSE WORDS. "NOT IN THIS HOUSE." 1: IT'S BEEN TWO YEARS. I THOUGHT THE WOUNDS HAD HEALED. 4: BUT I STAND BY WHAT I SAID. 3: PLUTO NEVER SHOULD HAVE BEEN A PLANET.	-
Gemma3 27B	1,2,3,4	1: IT'S BEEN TWO YEARS. I THOUGHT THE WOUNDS HAD HEALED. 2: SHE THREW ME OUT YELLING, "YOU DON'T SAY THOSE WORDS." "NOT IN THIS HOUSE." 3: AUTO NEVER SHOULD HAVE BEEN A PLANET. 4: BUT I STAND BY WHAT I SAID.	Alright, let's dissect this xkcd gem! This comic is funny because it takes a dramatic, relationship-ending argument (You don't say those words! Not in this house!) and then *completely* undercuts it with a bizarre, non-sequitur declaration about Pluto's planetary status. The escalation from domestic dispute to astronomical opinion is so unexpected and absurd that it lands as a delightfully nerdy punchline, and the visual of the aftermath just adds to the chaotic energy. It's a perfect example of xkcd's signature blend of relatable situations and delightfully strange tangents. Score: [7,7]
LLaVA-OV 7B	2,1,4,3	2: HE THREW ME OUT YELLING, "YOU DON'T SAY THOSE WORDS!" NOT IN THIS HOUSE. 3: BUT I STAND BY WHAT I SAID. 4: BUT I STAND BY WHAT I SAID.	The comic is funny because it plays on the idea of a person who has been hurt by someone else's words, but then realizes that they have been using those same words to hurt others. The humor comes from the irony of the situation and the unexpected twist at the end. However, some people may not find it funny if they do not understand the context or if they have experienced similar situations in their own lives. Score: [1,1]
Qwen2-VL 7B	2,1,4,3	1: It's been two years. 2: She threw me out yelling, "You don't say those words." "Not in this house." 3: Pluto never should have been a planet. 4: But I stand by what I said.	This comic is funny because it uses a series of panels to depict a humorous and relatable situation. The first panel sets up the premise, and the subsequent panels build on it with increasingly absurd and exaggerated actions. The final panel, which shows the character being thrown out of a house, is the punchline that ties everything together in a humorous way. The use of simple, stick-figure drawings and minimal text also adds to the humor by making the situation seem more absurd and exaggerated. Score: [4,4]

H Human Preference Examples

Below are samples of human-written interpretations chosen over GPT4o generated interpretations. In these samples, we highlight spans where a human incorporated multimodal reasoning that contributed to the humor; we posit that this (natural) multimodal reasoning ability is the most significant reason for human-written explanations being preferred in 48 of the 70 samples in this study (Table 7). In contrast, the generated explanations often use similar patterns of reasoning: descriptions of absurdity or unexpectedness. We showcase this discrepancy below where the human interpretation (that was preferred) did not explicitly use absurdity or unexpectedness in their reasoning.

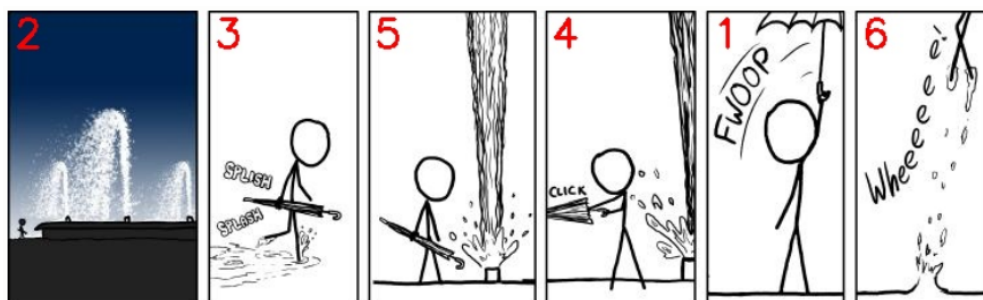


Figure 4: An xkcd comic that requires multimodal reasoning: pairing the sound effects with the comical actions of the stickman to understand how the humor is produced. Access at <https://xkcd.com/1038/>

Human interpretation: The humor in this comic lies in how the stickman "fly away" (in a cute and silly way) with his umbrella. We can observe this by following the narrative where the stickman encounters a powerful water jet at the water fountain. After which, he opened his umbrella to catch the strong water jet and be propelled upwards. The humor is amplified with the "wheeee" sound effect by the stickman.

GPT4o Interpretation: This comic is funny because it **subverts expectations** by showing a character using a fountain as a makeshift launch pad with an umbrella, which is an **absurd and unexpected use** of everyday objects. The sequence of panels builds up the anticipation, and the final "Wheeee" adds a playful tone to the character's unconventional flight. The humor lies in the combination of creativity and the sheer silliness of the situation



Figure 5: An xkcd comic that requires the user to have relevant world knowledge about the internet, as well as the pairing of expressive phrases such as "What are you doing!?" with the visual expression of the characters (raised hands depicting disbelief) to amplify humor. Access at <https://xkcd.com/262/>

Human interpretation: The humor in this comic plays with how the internet likes to put captions on cat related content to give them a more humorous/funny/human expression in the content (memes, short form videos etc). The humor is **further amplified by how the other man shouted "what are you doing" with raised hands, showing that he is shocked and/or confused and/or in disbelief that the man is doing something absurd to the cats.**

Gpt4o interpretation: This comic is funny because it plays on the internet meme culture of adding humorous captions to cat pictures, taking it to an **absurd literal level** by physically gluing captions to cats. The **unexpected and ridiculous action** contrasts with the normalcy of the conversation, enhancing the humor. Additionally, the cats' reactions add a layer of visual comedy.

Red text indicate a common reasoning pattern for humor in LMMs.

Green text indicate how the human's interpretation uses both modalities to explain the humor.