When 'YES' Meets 'BUT': Can Large Models Comprehend Contradictory Humor Through Comparative Reasoning?

Tuo Liang[1*], Zhe Hu[2*], Jing Li[2], Hao Zhang[1], Yiren Lu[1], Yunlai Zhou[1], Yiran Qiao[1], Disheng Liu[1], Jeirui Peng[1], Jing Ma[1], Yu Yin[1✉]

https://vulab-ai.com/projects/yesbut-v2/

arXiv:2503.23137v2 [cs.CV] 15 Apr 2026

Abstract—Understanding humor—particularly when it involves complex, contradictory narratives that require comparative reasoning—remains a significant challenge for large vision-language models (VLMs). This limitation hinders AI's ability to engage in human-like reasoning and cultural expression. In this paper, we investigate this challenge through an in-depth analysis of comics that juxtapose panels to create humor through contradictions. We introduce the YESBUT (V2), a novel benchmark with 1,262 comic images from diverse multilingual and multicultural contexts, featuring comprehensive annotations that capture various aspects of narrative understanding. Using this benchmark, we systematically evaluate a wide range of VLMs through four complementary tasks spanning from surface content comprehension to deep narrative reasoning, with particular emphasis on comparative reasoning between contradictory elements. Our extensive experiments reveal that even the most advanced models significantly underperform compared to humans, with common failures in visual perception, key element identification, comparative analysis and hallucinations. We further investigate text-based training strategies and social knowledge augmentation methods to enhance model performance. Our findings not only highlight critical weaknesses in VLMs' understanding of cultural and creative expressions but also provide pathways toward developing context-aware models capable of deeper narrative understanding though comparative reasoning.

Index Terms—Vision Language Models, Image Understanding, Visual Comparative Reasoning, Humor Understanding, Benchmark

1 INTRODUCTION

COMICS are more than just entertainment—they are intricate puzzles of visual narratives that merge images, text, and spatial relationships to convey layered meanings. Unlike novels, which rely solely on text, or videos, which typically follow a linear progression, comics often use fragmented yet interconnected panels to construct meaning through juxtaposition, requiring sophisticated comparative reasoning to interpret. While humans naturally interpret this nonlinear structure due to their cognitive abilities for comparison and cultural knowledge, current vision-language models (VLMs) struggle to replicate such nuanced understanding and comparative reasoning [1], [2], [3].

Evaluating how effectively VLMs perform comparative reasoning to interpret visual juxtaposition is essential for developing socially intelligent AI systems [4] Comics encapsulate complex human emotions, cultural nuances, and symbolic storytelling [5], making them a rigorous testbed for assessing AI's interpretive and comparative capabilities. A deeper understanding of visual juxtaposition can enhance scene interpretation, facilitate AI-driven storytelling that resonates with human audiences, and improve creative content generation in a more human-like manner. Addressing the challenges of visual juxtaposition allows VLMs to move beyond basic pattern recognition toward nuanced, context-aware reasoning, fostering deeper and more meaningful human–AI interactions.

While previous studies [2], [6] have explored humor understanding with VLMs, most studies have focused on single-panel comics, where the narrative is self-contained within a single image. In contrast, our work examines how VLMs handle juxtaposition—a technique that places contrasting elements side by side to create humor, irony, or thought-provoking narratives [7], [8]. This approach challenges VLMs to engage in *comparative reasoning and analysis* to decipher the intricate relationships between panels to capture the overall narrative [9], [10], [11].

In this work, we explore whether AI can recognize and interpret implicit contradictions in juxtaposed panels that contribute to meaning and humor. To illustrate, consider the comic in Fig. 1, which humorously illustrates the concept of a "learning curve." In the "Yes" panel, a man observes a steady upward curve labeled "LEARNING CURVE," symbolizing smooth progress. However, in the "But" panel, the curve twists into a closed loop, trapping him inside, humorously depicting the illusion of learning without real progress. This example highlights the need for deep contextual understanding and comparative reasoning—capabilities that remain limited in current models due to the constraints of the autoregressive paradigm, which hinders bidirectional reasoning [12], [13], [14].

- *Tuo Liang, Hao Zhang, Yiren Lu, Yunlai Zhou, Yiran Qiao, Disheng Liu, Jeirui Peng, Jing Ma, Yu Yin are with Computer and Data Sciences Department, Case Western Reserve University. Email: tuo.liang@case.edu*
- *Zhe Hu and Jing Li are with Department of Computing, The Hong Kong Polytechnic University. Email: zhe-derek.hu@connect.polyu.hk*

* *These authors contribute equally to this research.*

✉ *Correspondence to: Yu Yin (yu.yin@case.edu).*

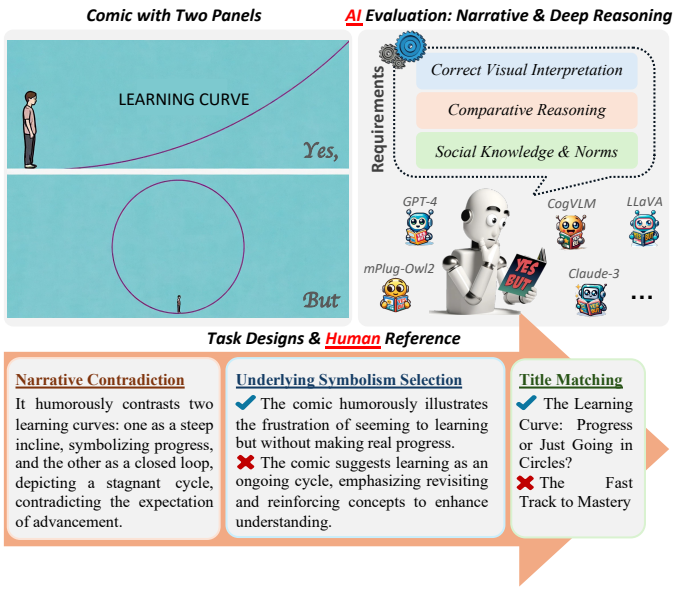

Fig. 1: We introduce the YESBUT (V2), a benchmark for assessing AI's ability to interpret juxtaposed comic panels with contradictory narratives. Unlike existing benchmarks, it emphasizes visual understanding, comparative reasoning, and social knowledge. To capture the layered reasoning required for interpreting these contradictions, we design multi-tiered tasks—ranging from basic content recognition to deep narrative comprehension—ensuring a comprehensive assessment of AI's interpretative abilities.

We identify three primary challenges in achieving robust juxtaposition understanding in comics (Figure 1). The first challenge is **accurate visual interpretation**, which involves decoding the visual elements within each panel. The second is **comparative reasoning**, which requires integrating and comparing key elements across multiple panels to detect contradictions that create humor or irony in the overall narrative. Finally, **social and cultural comprehension** is essential, as recognizing subtle social cues, conventions, and cultural contexts significantly influences the interpretation of comic humor and emotional responses. To this end, we introduce YESBUT, the first benchmark designed to assess VLMs' capacity for recognizing humor through juxtaposition and contradiction. Our dataset is uniquely annotated to capture multiple layers of narrative complexity. Each sample includes a literal description of the scene, an explicit contradiction statement that clarifies the humorous contrast, the underlying symbolism or message conveyed by the comic, a title summarizing the overall theme, and relevant background knowledge—including social norms, cultural references, and linguistic context—essential for full interpretation of the underlying humor.

Based on our annotations, we propose four complementary tasks that progressively assess comic understanding across different cognitive levels: (1) Literal Description Writing evaluates a model's perceptual ability by requiring to generate a surface-level description with an explicit depiction of the visual and textual content presented in the comic, (2) Contradiction Generation focuses on identifying and articulating the core contradiction in the narrative that often serves as the basis for humor; (3) Underlying Symbolism Selection measures deeper interpretative reasoning by challenging models to infer the abstract message or commentary embedded in the comic; and (4) Title Matching assesses holistic understanding and thematic summarization by requiring models to select a title that accurately encapsulates the comic's thematic essence. This hierarchical evaluation establishes a systematic approach for measuring progress in machine humor comprehension while highlighting specific reasoning capabilities needed for more sophisticated semantic understanding of visual narratives.

Building on our preliminary work [15], which introduced a small-scale dataset of 349 images and identified significant gaps in VLMs' comic narrative understanding, this paper presents a substantial expansion and refinement. We evaluate a wide range of models including both recent LLMs and VLMs using our extended benchmark. The experiments as well as extensive analysis reveal that current large models still face challenges for contradictory humor understanding with comparative reasoning, and tend to make common errors including visual perception, key element identification, and hallucination in narrative understanding. Our analysis further show that augmenting social knowledge and model training through textual data distillation can improve model performance. Our key contributions are:

- **Larger and More Diverse Dataset**: We expand YESBUT from 349 to 1,262 images, improving the benchmark's robustness and diversity to facilitate a more comprehensive evaluation of VLM capabilities. The comics encompass multi-cultural backgrounds, enabling richer assessment of linguistic and cultural influences on humor comprehension.
- **Comprehensive Model Evaluation**: We conduct a systematic evaluation of a diverse set of VLMs and LLMs, including general-purpose models, reasoning-enhanced models, and those supporting multi-image inputs, providing unprecedented comparative insights into their capabilities and limitations.
- **In-Depth and Fine-grained Analysis**: Through detailed statistical and ablation studies, we identify critical factors affecting model performance on juxtaposition-based humor and categorize specific failure patterns.
- **Practical Model Improvement Strategy**: We propose simple yet effective approaches to improve VLMs' understanding of juxtaposition-based humorous images, offering a practical direction for future research.

These contributions collectively establish a rigorous benchmark for evaluating machine humor comprehension while advancing the development of VLMs capable of deeper semantic reasoning in multimodal narratives.

2 Related Work

2.1 Vision-Language Model Evaluations

Recent advancements in large vision-language models (VLMs) have demonstrated remarkable capabilities, including following human instructions and performing tasks such as image captioning, visual question answering, and multimodal reasoning through zero-shot prompting [16],

[17], [18], [19], [20]. To systematically assess these models, numerous benchmarks have been developed for both language-only [21], [22], [23], [24] and vision-language [25], [26], [27], [28], [29] tasks. However, despite their effectiveness in measuring fundamental abilities such as linguistic comprehension [30], [31] and general problem-solving [32], existing evaluations often overlook deeper aspects of contextual reasoning and social intelligence. This limitation is critical, as VLMs still struggle to engage in nuanced social reasoning and accurately interpret human-centric contexts [33], [34]. Without rigorous assessments of these advanced capabilities, AI systems risk misinterpretations in real-world applications, where nonlinear inference and social considerations are crucial. Unlike existing works, we propose the development of targeted evaluation tasks to rigorously assess complex semantic reasoning and multimodal situational understanding. Such enhanced evaluations will not only diagnose the strengths and limitations of current models but also provide clear directions for advancing socially aware AI.

2.2 Computational Humor

Humor is a fundamental aspect of human communication, and its computational understanding has become a growing area of research [35], [36]. Early studies primarily focused on textual humor, leveraging linguistic features such as wordplay, incongruity, and sentiment shifts to detect [37], [38], [39] and generate humor [40]. While these approaches advanced humor detection, even state-of-the-art large language models (LLMs) like ChatGPT still struggle with nuanced humor comprehension [41], [42], particularly in cases requiring deep narrative understanding. Although LLMs can recognize humor to some extent, they often fail to grasp the intricate, context-dependent humor that arises in complex narratives.

More recently, research has expanded into multimodal humor analysis, integrating visual and textual elements to predict humor across different formats. Studies on humorous cartoon captions [2], [43], [44], visual humor prediction [45], [46], and humor detection in videos [47], [48] have shown promising advancements. Similar to our work, prior research [2], [6] has explored AI-driven comprehension of memes [49], cartoons [50], and comics [2], [6], [51], primarily focusing on humor within isolated images. However, a notable gap in the literature is the insufficient exploration of humor embedded within multi-panel narratives.

Unlike single-panel humor, which often relies on direct punchlines, multi-panel comics involve sequential storytelling, evolving contradictions, and juxtaposition to construct humor. These characteristics contribute to deeper and more complex humor structures. Our work leverages these features to investigate how comic juxtapositions—characterized by contradictory narratives—challenge existing computational models.

2.3 Visual Reasoning

Recent research has leveraged neurosymbolic methods [52], [53], [54] to enhance image understanding. These methods typically break down visual reasoning into explicit steps, using powerful vision models to improve object recognition. Another approach, such as Vision-Language Models (VLMs) [55], [56], [57], [58], [59], focuses on strengthening the vision encoder and improving alignment through visual prompt tuning. While these techniques have made significant progress in surface-level image understanding, our task requires a deeper level of visual reasoning. Specifically, our goal is to detect and comprehend contradictions between two comic panels—something that goes beyond suffacial image understanding.

Previous research has explored the visual reasoning capabilities of large models in various tasks, including commonsense reasoning [27], [60], [61], visual question answering [62], visio-linguistic compositionality [63], and science question answering [64]. However, our work shifts the focus to comparative reasoning with nonlinear narrative. Unlike linear narrative, which follows structured rules and is relatively straightforward for AI to process, comparative reasoning requires models to interpret ambiguous, multi-layered information without explicit guidance. This makes the task significantly more challenging, as it demands a deeper level of natural language understanding and cognitive modeling.

Our work is also relevant to comparative reasoning. Previous studies have primarily focused on textual scenarios [65], [66]. Recently, this task has been extended to multimodal contexts involving image pairs [67]. Our task, however, presents a unique challenge: the two panel images in comics collectively construct a contradictory narrative. This requires models to perform both local comparisons between panel elements and global narrative interpretation to grasp the comic's underlying message. Unlike previous approaches, our task demands not only comparative reasoning between visual elements but also an understanding of human behavior and social dynamics to accurately interpret the narrative's intended meaning.

3 The YesBut (V2) Dataset

Our benchmark consists of comics with juxtaposed panels that feature an inherent contradiction or unexpected narrative twist. Specifically, each sample includes (1) a two-panel comic that contains a contradictory narrative; (2) a literal description of the scene; (3) an explicit contradiction statement that clarifies the humorous contrast; (4) the underlying symbolism or message conveyed by the comic; (5) a title summarizing the overall theme; and (6) additional features, including social knowledge and linguistic context necessary for interpreting the comic.

Different from the previous work [15], we incorporate additional background features as a fundamental component of our dataset, as it plays a crucial role in comic interpretation. Social knowledge refer to human norms, cultural contexts, and awareness of social events and references that are necessary to interpret the comic's meaning. Linguistic context refers to the language and cultural background in which the comic was created, influencing how the narrative is perceived. Our dataset includes images in multiple languages (*e.g.*, English, Chinese, and Russian), with variations in linguistic and cultural environments affecting language comprehension.

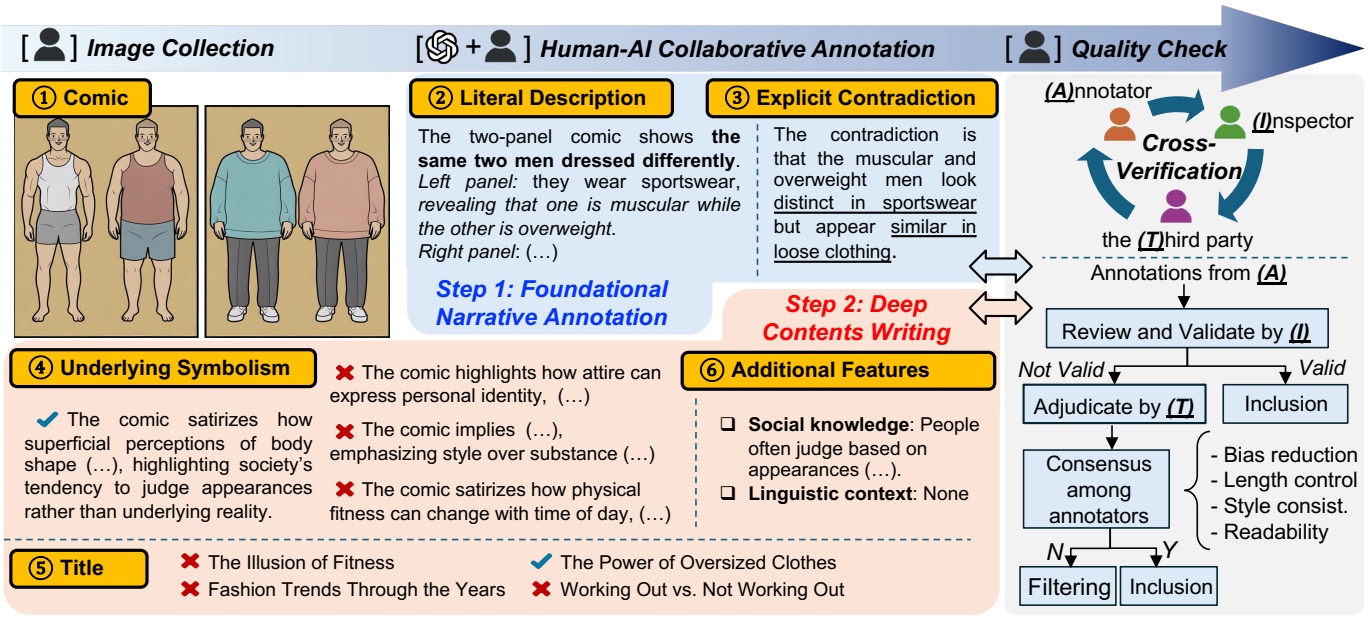

Fig. 2: Overview of the Data Construction Pipeline. The dataset construction begins with manually *collecting images* from social media platforms, verified by human reviewers to ensure authenticity and relevance. Next, a *progressive human-AI collaborative annotation* stage is employed to enhance labeling accuracy and efficiency. Finally, a rigorous *quality control and cross-verification* stage is conducted with multiple annotators to refine and validate the dataset.

The process of data construction is introduced in Section 3.1, followed by a summary of key statistics in Section 3.2. Based on annotated components, we designed multi-tiered tasks to systematically evaluate comic comprehension (Section 3.3).

3.1 Dataset Construction

To construct the dataset, we first introduce the *image collection* process, followed by a two-stage annotation process: a *progressive human-AI collaborative annotation* stage and a *quality check and cross-verification* stage. Fig. 2 provides an overview of the data construction pipeline.

3.1.1 Image Collection

Our dataset consists of captionless comics collected from social media platforms[1]. These comics depict conflicting narratives of everyday life, capturing humorous, ironic, or thought-provoking scenarios. To ensure data quality and relevance, we applied several preprocessing steps. First, we implemented deduplication techniques to remove identical or near-duplicate images, enhancing dataset diversity. Next, we filtered out images with ambiguous or unclear meanings to minimize noise and facilitate meaningful analysis. Additionally, we conducted content moderation to exclude comics containing inappropriate, offensive, or harmful material. After completing these preprocessing steps, our final curated dataset comprises 1,264 captionless comics, offering a valuable resource for studying visual storytelling, humor, and contrasting perspectives in everyday life.

1. https://twitter.com and https://www.pinterest.com/

3.1.2 Progressive Human-AI Collaborative Annotation

For each comic, we annotate the five key components: literal description, contradiction explanation, underlying symbolism, comic title, and relevant background knowledge (*i.e.*, social norms, linguistic context, and whether it contains text). To ensure high-quality annotations, we primarily rely on human annotators to provide gold-standard labels. The annotation process involved eleven human judges, who were selected based on linguistic and cultural proficiency. They underwent a brief training session to familiarize themselves with the annotation guidelines and ensure consistency across judgments. Throughout the process, annotators worked independently, and disagreements were resolved through discussion.

To minimize the cost and effort of manual annotation while ensuring high-quality data, we developed a progressive human-AI collaborative pipeline that leverages GPT-4 for structured data annotation. This pipeline functions as an interactive dialogue system, where AI and human annotators collaborate iteratively to refine and enhance annotations across multiple steps. The annotation process begins with AI-assisted generation: given a comic image, GPT-4 initially produces a narrative description along with an explanation of its contradictory logic (*i.e.*, **Step 1: Foundational Narrative Annotation**). These outputs then undergo a collaborative refinement phase, during which human annotators review, correct, and enhance them to establish the gold-standard descriptions and contradictions. Once validated, these refined annotations serve as the foundation for deeper annotations.

Building upon this initial step, GPT-4 is further prompted to generate deeper and progressively complex insights, including underlying symbolism, a compelling comic title, and relevant background knowledge (*i.e.*, **Step**

TABLE 1: Data Statistics. Avg. Len. is the average number of words.

YesBut Components		#Num	Ave. Len.
	Image	1,262	-
	Literal Description	1,262	134
	Explicit Contradiction	1,262	33
	Underlying Symbolism	5,048	26
	Title	5,048	6
Addition Features	Social Knowledge	3,407	97
	Linguistic Context	1,262	1

2: Deep Contents Writing). At each step, human annotators actively engage with AI-generated content, verifying its accuracy, contextual appropriateness, and interpretative depth. Through this recursive collaboration, the pipeline ensures that AI outputs are not only computationally efficient but also aligned with human-level understanding and nuance.

Structuring the annotation process into progressive steps allows AI to tackle increasingly complex tasks while reducing cognitive load and maintaining high interpretative accuracy. For instance, the pipeline first generates simpler, literal narratives, which serve as a foundation for interpreting the deeper symbolic meanings in comics. This approach enhances efficiency and reduces annotation costs. Additionally, for the underlying symbolism and title generation tasks, GPT-4 generates hard negative counterparts and distractions to construct multiple-choice questions for our experiments. Human annotators also annotate panel bounding boxes, which are used for panel separation in experiments and included in the dataset. Example prompts used for annotation are provided in Appendix B.1.

3.1.3 Quality Check with Cross-Verification

The annotation process incorporates multiple quality checks to ensure accuracy, consistency, and consensus among annotators. We implemented a cross-verification procedure in which each annotated comic undergoes review by a designated inspector, who verifies correctness and flags any ambiguities or quality concerns. If an annotation is unclear or inconsistent, a third annotator serves as an adjudicator to make the final decision. Comics with unresolved ambiguities, controversial content, or potential bias are filtered out. Finally, one of the authors conducts a comprehensive review to validate the annotations before their inclusion in the benchmark.

3.2 Dataset Statistics and Attribute Distribution

We analyze key attributes in our dataset that influence VLM comprehension of narrative contradictions and humorous contrasts in comics. Our dataset consists of 1,262 comics, each accompanied by high-quality annotations. A statistical breakdown of annotated components, including their quantity and length, is presented in Table 1.

Beyond basic statistics, we examine the distribution of comics across various categorical attributes to gain deeper insights into the dataset's structure. Fig. 3 illustrates these distributions, with representative examples from each category. A comprehensive breakdown of category-wise statistics is provided in Appendix B.2.

3.2.1 Presence of Text and Linguistic Context

Our dataset includes both text-embedded and purely visual comics, allowing us to investigate how textual elements influence model performance. To facilitate this analysis, each comic is annotated to indicate the presence or absence of embedded text. This distinction allows for an assessment of whether textual components enhance model comprehension or if models can primarily infer humor from visual cues. Notably, 58% of the comics in our dataset contain embedded text, while 42% rely solely on visual elements (Fig. 3, left). This balanced distribution enables a comprehensive evaluation of how models process different modalities of humor and the extent to which textual information contributes to their interpretative abilities.

For comics containing embedded text, we further categorized them based on linguistic dependency, recognizing that language variations influence humor perception. The two linguistic categories are defined as follows:

- **Linguistic-Dependent:** Comics that require knowledge of a specific language (e.g., English, Chinese, or Russian) to understand the humor. These typically involve puns, homophones, idioms, or culturally specific wordplay.
- **Transferable Text:** Comics containing text that can be directly translated into another language (*e.g.*,, Chinese to English) without altering its humor or meaning. This category also includes comics in which humor arises purely from visual cues, making them universally understandable regardless of language.

3.2.2 Social Knowledge

Many comics require social knowledge, including human norms, cultural contexts, or common sense beyond simple visual interpretation. To quantify the impact of social knowledge, we provide annotations that highlight comics requiring deeper contextual understanding beyond basic visual or texture elements. These annotations distinguish comics that are self-explanatory and those that demand external social knowledge or cultural norms for interpretation. As illustrated in Fig. 3 (middle), 86% images require social knowledge for proper interpretation. Only a small fraction can be readily understood without additional background knowledge.

3.2.3 Humor categories

To systematically analyze humor and encompass its diverse thematic dimensions, we categorize humorous content into 15 domains, each representing distinct contexts and topics frequently encountered by audiences (Fig. 3, right). The dataset predominantly features humor about Work Jokes & Complaints (18%) and Internet Culture & Technology (16%), highlighting workplace frustrations and digital-era satire. Mid-tier categories include Fashion Trends & Jokes (9%), Travel & Adventure (8%), Finance & Money Matters (8%), and Sports & Fitness (7%), emphasizing humor derived from personal experiences, social trends, and lifestyle habits. Additional everyday-life categories ($\leq$ 6%) encompass Food & Culinary Experiences, Daily Life & Routine Humor, Relationships & Social Life, Entertainment & Pop Culture, and Education & Student Life. Lastly, niche categories ($\leq$ 3%)

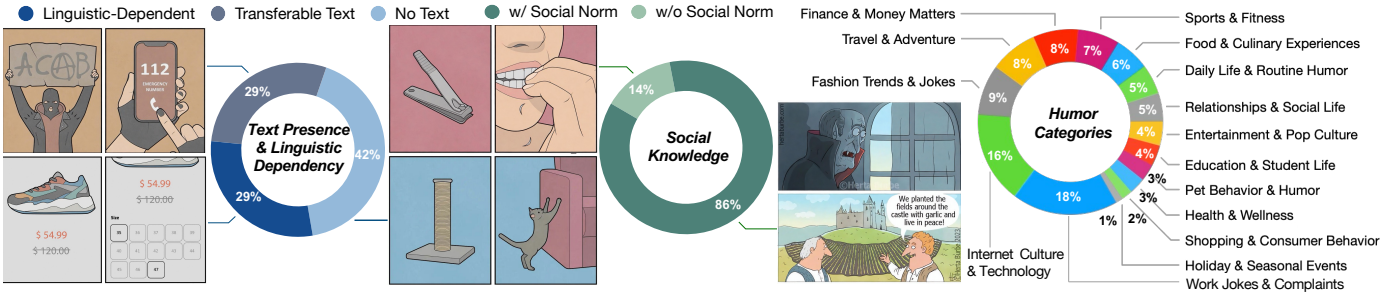

Fig. 3: Distribution of the original 1,264 comics downloaded from social media based on different aspects, including embedded text presence, reliance on social knowledge, and distinct humor categories. Overall, we show that our YESBUT exhibits balanced text presence, provides insights into social norms and cultural expectations, and captures a diverse thematic range of humor.

like Pet Behavior & Humor, Health & Wellness, Shopping & Consumer Behavior, and Holiday & Seasonal Events, highlight humor derived from specific cultural contexts and consumer behaviors. It can be seen that the collected comics cover a broad thematic scope and its reflection of social, cultural, and individual experiences.

3.3 Task Design: Do Large Models Understand Humor in Juxtaposition?

Our goal is to assess the ability of contemporary large-scale (visual) language models to recognize and interpret humor arising from narrative contradictions commonly found in comics. To achieve this, we have devised four targeted tasks that progressively assess the model's comprehension, reasoning abilities—particularly non-linear reasoning—and sensitivity to the nuanced, abstract, and symbolic content embedded in comic narratives:

Task 1: Literal Description Writing. This initial task measures the model's basic narrative comprehension. Given a two-panel comic as input, the model is tasked with generating a concise textual description capturing the overarching narrative depicted across the panels. Unlike standard image captioning tasks, which focus on isolated visual details, this exercise requires the model to synthesize the events into a coherent textual narrative.

Task 2: Explicit Contradiction Generation. Moving beyond literal interpretation, this task evaluates the model's ability to recognize the contradictions arising from the juxtaposed panel of the comics. Presented again with the comic as input, the model must generate a textual explanation clearly identifying and elaborating on the contradiction embedded within the narrative. Successful performance indicates the model's capability to reason logically and non-linearly about narrative inharmoniousness, which is crucial for humor understanding.

Tasks 3: Underlying Symbolism Selection. Humor often conveys deeper symbolic or conceptual messages beyond its surface narratives. In this task, we evaluate the model's sensitivity to these implicit symbolic meanings. The model is provided with a comic and four possible symbolic interpretations—one correct and three carefully designed distractors that appear plausible but are incorrect. By selecting the correct interpretation, the model demonstrates its ability for abstract reasoning and deep narrative comprehension.

Task 4: Title Matching. This final task evaluates whether the model can associate comics with an appropriate title. Given a comic strip and four possible titles (one correct and three plausible distractors), the model selects the most suitable title summarizing the underlying meaning and narrative abstraction of the comic. Because the title encapsulates the nuanced, abstract humor within the narrative, successful performance indicates a sophisticated grasp of both humor and high-level narrative understanding.

Collectively, these tasks provide a robust evaluation framework to explore multiple levels of humor comprehension, ranging from surface literal and logical interpretations to in-depth abstract symbolic reasoning, ultimately revealing whether large-scale language models genuinely understand humor in juxtaposition.

4 EXPERIMENTS

This section outlines the models, evaluation metrics, and settings used in our experiments. The goal is to evaluate model performance across different tasks, including text generation (literal description writing and contradiction generation) and multiple-choice questions (symbolism selection and title matching). We consider two types of models: Vision-Language Models (VLMs), which process both images and text, and Large Language Models (LLMs), which rely on image-to-text conversion before evaluation.

4.1 Models and Settings

We evaluate a diverse set of VLMs and LLMs in a zero-shot setting. For VLMs, we consider three categories: (1) General-purpose VLMs that support single-image input, including LLaVA-1.5 [68], CogVLM [69], and GPT-4v [58]; (2) Multi-image finetuned models that allow processing of multiple images as input, including LLaVA-OneVision [70], Qwen2-VL [71], and ChatGPT4-Vision-Turbo [58]; (3) Reasoning-Enhanced VLMs specifically designed to improve reasoning capabilities, such as LLaVA-Next [59] and GPT-4o [72].

For LLMs, we evaluate GPT-4 [73], Deepseek-r1 [74], Llama3 [17] and Qwen2 [75]. Since LLMs cannot process images directly, we first use LLaVA-1.6 13B [59] to generate captions that serve as literal descriptions for each comic [2].

2. We select LLaVA-1.6 specifically for its strong image captioning capabilities and because its parameter count is comparable to the LLMs.

TABLE 2: Model performance on tasks ranging from foundational narrative understanding to deep content reasoning. Our evaluation includes a diverse set of VLMs: general-purpose models, reasoning-enhanced models, and models capable of processing multi-image inputs. Additionally, we assess advanced LLMs, which use captions generated by LLaVa-Next-13B as input. For the tasks of literal description and contradiction generation, we report the BERT score (F1), ROUGE-2 (F1), and GPT evaluation scores. The best scores are highlighted in **bold**, while the second-best scores are underlined.

Type	Model	Literal Description			Contradiction			Symbolism	Title
		BERT	R-2	GPT	BERT	R-2	GPT	Accuracy	Accuracy
General-Purpose VLMs	LLaVA-1.5-7B	86.67	59.31	3.32	85.97	57.04	3.23	51.13	64.50
	LLaVA-1.5-13B	85.62	55.97	3.31	85.86	56.18	3.20	71.57	68.31
	CogVLM2	86.58	56.05	3.37	86.58	56.05	2.98	37.40	54.43
	GPT4-Vision-Turbo	87.41	60.02	3.65	87.27	47.98	3.55	74.05	71.36
Multi-Image VLMs	LLaVA-OneVision-0.5B	85.15	46.57	3.11	86.99	43.20	2.22	36.87	38.52
	LLaVA-OneVision-7B	85.26	42.13	3.55	87.63	46.84	3.02	67.64	70.80
	LLaVA-OneVision-72B	86.42	52.03	3.63	87.61	46.82	3.55	80.30	78.48
	Qwen2-VL-7B	87.16	57.10	3.55	86.83	55.70	3.09	74.48	74.72
	Qwen2-VL-72B	88.16	60.84	3.77	87.71	**60.07**	3.49	79.98	**81.25**
Reasoning-Enhanced VLMs	LLaVA-Next-7B	86.12	57.33	2.94	85.55	56.05	2.57	59.35	56.57
	LLaVA-Next-13B	86.49	56.92	3.05	85.57	55.83	3.12	70.36	66.88
	LLaVA-Next-72B	86.00	56.00	3.26	87.15	46.23	3.31	74.72	71.63
	GPT-4o	**88.96**	**63.13**	**3.96**	**87.85**	49.07	**3.72**	**80.38**	80.62
LLMs	GPT-4	-	-	-	86.57	46.32	2.86	61.85	59.89
	Deepseek-r1-70B	-	-	-	87.20	45.38	3.37	65.32	57.47
	Llama3-8B	-	-	-	86.62	43.95	3.24	60.62	55.71
	Llama3-70B	-	-	-	86.71	43.12	3.52	67.35	62.68
	Qwen2.5-7B	-	-	-	86.60	46.20	3.28	64.98	56.58
	Qwen2.5-72B	-	-	-	86.74	46.33	3.31	67.59	71.47

These captions replace the original images, allowing the LLMs to process them alongside the questions. As a result, LLMs are not evaluated on the literal description task.

For implementation, we set the temperature to 1 for GPT-4o and ChatGPT. For all other models, we maintain default parameter settings and prompt templates during inference. To mitigate prompt variance across different tasks, we create three distinct prompts per task and report the average score across these runs. Further details on model specifications, prompt templates, and experimental procedures are provided in the Appendix C.5.

4.2 Evaluation Metrics

To ensure a thorough assessment, we employ both automated metrics and human evaluations. The automated evaluation uses multiple metrics to capture different aspects of model performance, while human evaluations help assess qualitative aspects of text generation. For multiple-choice question tasks (*i.e.,* underlying symbolism selection and title matching), we adopt accuracy as the primary evaluation metric. For text generation tasks (*i.e.,* literal description writing and contradiction generation), we follow text generation research [76], [77] and employ a diverse set of metrics: word overlap-based metrics such as ROUGE-2 (R-2), word vector-based metrics including BERT Score, and GPT-based Score. Recent studies have demonstrated that GPT-based evaluation methods align closely with human judgment [78]. By incorporating multiple metrics, we aim to comprehensively assess the quality of model-generated descriptions and contradiction texts across different dimensions.

Recognizing the limitations of automatic evaluation for text generation tasks, human evaluations are conducted to provide additional insights. We establish a 5-point scoring system (1=worst, 5=best) and engage human judges to evaluate several aspects. For literal description, we evaluate: *Correctness*, which measures accuracy in conveying the comic's narrative; *Completeness*, which assesses coverage of all important narrative elements; and *Fidelity*, which examines the absence of hallucinations, ensuring all content is supportable by the comic images. For contradiction generation, we evaluate on aspects of Correctness and Fidelity using the same criteria. Please refer to Appendix C.3 for more details.

5 Main Results

In this section, we present a comprehensive analysis of model performance across the tasks. Table 2 shows the main results for all models tested.

5.1 Literal Description Writing

For literal description writing, the results show that commercial models generally outperform open source alternatives. Among all models, GPT-4o achieves the highest performance scores. For open source models with identical architectures, we observe a positive correlation between model size and performance, suggesting that larger models possess enhanced comic comprehension and description

generation capabilities. Notably, Qwen2-VL-72B ranks second, with performance metrics approaching those of GPT-4o, indicating a narrowing gap between commercial and open source VLMs.

5.2 Contradiction Generation

For the contradiction generation task, we include both VLMs and LLMs, with the latter utilizing image captions generated by LLaVA-1.6 13B as input. Our evaluation employing multiple metrics reveals that GPT-4o achieves superior performance in BERT and GPT scores, indicating high semantic similarity and language quality, while Qwen2-VL demonstrates the highest ROUGE-2 (R-2) score, suggesting better lexical overlap with reference contradictions. Interestingly, GPT-4o's ROUGE-2 score was surpassed by several 7B-parameter models despite its overall stronger performance. This discrepancy can be attributed to the inherent limitations of ROUGE-2 evaluation, which primarily measures bigram overlap. Upon manual analysis of the generated outputs, we observed that GPT-4o typically produces longer outputs with paraphrasing and expanded context, which results in reduced lexical overlap with reference contradictions.

Moreover, we can observe that the LLaVA-OneVision series models consistently outperforms both LLaVA-1.5 and LLaVA-Next series models in this task. We hypothesize that this superior performance stems from their multi-image training approach, which enhances their ability to detect and reason about relationship changes across multiple panels—a crucial skill for understanding comic narratives and accurately capturing inter-panel relationships.

5.3 Deep Reasoning Tasks

The underlying philosophy selection and title matching tasks require sophisticated reasoning based on comic narratives. Our results show that for the underlying symbolism selection task, GPT-4o achieves the highest accuracy (80.38%), while for the title matching task, Qwen2-VL-72B demonstrates the highest accuracy of 81.25%.

A key finding is that larger models generally demonstrate better comic comprehension capabilities, aligning with previous research indicating enhanced reasoning abilities in models with higher parameter counts. Comparing across the LLaVA model series with equivalent parameter counts, LLaVA-Next models consistently outperform LLaVA-1.5 models, while LLaVA-OneVision models generally surpass LLaVA-Next models. This performance hierarchy can be attributed to the progressive improvements in each model series: LLaVA-Next enhances reasoning abilities and world knowledge [59], while LLaVA-OneVision further incorporates supervised fine-tuning on multi-image and video inputs, strengthening its capacity to understand relationships and changes across multiple images. These findings suggest a promising direction for improving VLMs' understanding of juxtaposed humorous comics: enhancing models' reasoning and multi-image relationship comprehension through targeted fine-tuning.

Additionally, we observe that LLMs consistently underperform compared to VLMs of equivalent scale. This performance gap can be attributed to the LLMs receiving literal descriptions generated by LLaVA-Next-13B as input, thus inheriting any errors, information loss, or hallucinations present in the VLM-generated descriptions. We provide a more detailed analysis of description quality's impact on LLM performance in Section 6.1.2.

Another notable observation is that most models perform worse on title matching than on underlying symbolism selection, with Qwen2-VL-72B being the sole exception. This discrepancy likely stems from titles being shorter and more abstract representations of narratives that do not explicitly convey the underlying comic concepts. Consequently, distinguishing between correct titles and distractors requires deeper, more rigorous understanding and reasoning capabilities, presenting a more significant challenge for most models.

5.4 Comparison with Human Performance

To establish a human performance baseline for better understanding model performance, we conduct a controlled evaluation with three human participants on a sample of 50 randomly selected comics. Participants are asked to complete the same Underlying Symbolism and Title Matching tasks as our evaluated models.

As shown in Fig. 4, human evaluators substantially outperform even the strongest AI models on both deep reasoning tasks, especially on the more complex title matching task. This performance gap highlights the significant challenges AI systems face when tasks require complex non-linear reasoning, abstract concept interpretation, and cultural context understanding. These results confirm that despite recent advances in multimodal capabilities, substantial improvements are still needed for VLMs to achieve human-like comprehension of juxtaposition-based humor in comics.

5.5 Human Evaluation

To complement our automatic evaluation metrics, we conduct human evaluations on a randomly selected subset of 30 samples to assess the quality of both literal descriptions and contradiction generations. The results, presented in Fig. 5, align with the trends observed in automatic evaluations: commercial models consistently outperform open-source alternatives across both tasks, with GPT-4o achieving the highest overall scores. Among open-source models, Qwen2-VL-72B stands out, demonstrating performance levels approaching those of GPT-4o.

Additionally, we observe that models generally achieve higher scores for literal descriptions than for contradiction generation. This performance gap highlights the inherently greater complexity of the contradiction generation task, which requires sophisticated comparative reasoning across multiple visual elements in paired panels. In contrast, generating literal descriptions is a more straightforward process, primarily relying on visual perception of the comic image.

Furthermore, the results indicate a correlation between model scale and performance quality, suggesting that larger parameter counts contribute to improved image comprehension. Notably, models specifically designed for multi-image processing and enhanced reasoning—such as those in

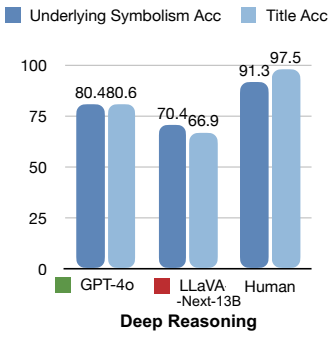

Fig. 4: Human performance on deep reasoning tasks.

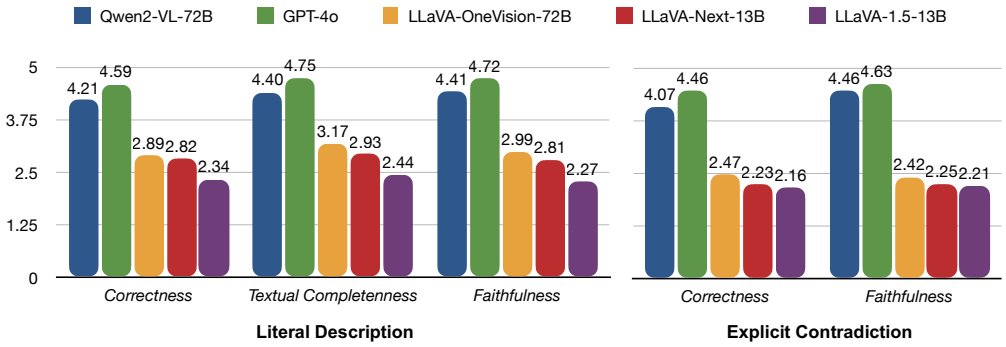

Fig. 5: Human Evaluation of Literal Description and Contradiction Generation Tasks.

the LLaVA-OneVision series—outperform general-purpose VLMs of similar scale, reinforcing our earlier findings.

Lastly, we identify a strong positive correlation between performance on literal description and contradiction generation tasks. Models that excel in one task tend to perform well in the other, indicating that these abilities are fundamental to visual narrative comprehension. Unlike superficial content understanding, interpreting juxtaposed comic panels requires an integrated set of advanced capabilities, including precise image perception, contextual narrative interpretation, extensive world knowledge, and sophisticated reasoning. This interdependence underscores the importance of these core competencies in effective visual understanding.

6 Analysis and Discussion

In this section, we conduct a series of experiments to investigate various factors influencing the deep reasoning performance of VLMs on comic understanding. Our investigation follows a structured progression: **(*A*) Basic Factors Affecting Model Performance**, examining elements such as embedded text and surface descriptions; **(*B*) Methodological Enhancements**, incorporating techniques like task decomposition, panel splitting, and model fine-tuning for deep reasoning tasks; **(*C*) Broader Enhancements**, assessing the impact of external social knowledge; and **(*D*) Case Study and Error Analysis**, conducting detailed examinations of specific instances and identifying common errors.

6.1 Basic Factors Affecting Model Performance

6.1.1 The Role of Embedded Text in Comic Images

In our dataset, 58% of the comics contain embedded text, while 42% rely solely on visual elements (as depicted in Fig. 3 left). Analyzing the impact of this embedded text on model performance reveals notable trends, as the results detailed in Table 3. For the title matching task, all models exhibit improved performance on images with embedded text, especially on comics with transferable text. This improvement suggests that embedded text provides direct contextual information, aiding VLMs in understanding the general content and augmenting the reasoning about the relationships of the comic panels.

Contrary to expectations, in the underlying symbolism deep reasoning task, models other than the Qwen2 series perform worse on images containing embedded text compared to those without. Given that a portion of the comics contain Chinese text (257 images with embedded Chinese text and 113 with English text) a plausible explanation is that Qwen2 models' proficiency in Chinese contributes to their superior performance [79]. In contrast, other models may lack robust multilingual capabilities, hindering their performance in deep reasoning tasks. Further discussions on language influence are provided in Appendix D.2.

These observations underscore the importance of integrating advanced multilingual understanding capabilities into VLMs to enhance their performance in tasks involving complex visual and textual information.

TABLE 3: Impact of embedded text in comic images on model performance. "w/o text" refers to comic images without embedded text. Images containing embedded text, which can assist in comic comprehension, are further categorized into two groups: ling. dept. (Linguistic-Dependent) and trans. txt. (Transferable Text).

Model	Emb. Text	Symbolism Acc.	Title Acc.
GPT-4o	*w/o text*	**82.36**	77.42
	ling. dept.	77.85	77.26
	trans. txt.	80.06	**88.73**
LLaVA-OneVision-7B	*w/o text*	**71.11**	69.61
	ling. dept.	61.02	63.80
	trans. txt.	69.32	**82.03**
LLaVA-OneVision-72B	*w/o text*	**82.05**	77.42
	ling. dept.	77.09	72.83
	trans. txt.	81.02	**85.81**
Qwen2-VL-7B	*w/o text*	70.11	69.17
	ling. dept.	76.97	77.47
	trans. txt.	**78.33**	**77.97**
Qwen2-VL-72B	*w/o text*	77.67	76.74
	ling. dept.	**83.37**	82.49
	trans. txt.	79.89	**86.58**

6.1.2 Influence of Surface Descriptions on Deep Reasoning

Effective deep reasoning about comic narratives fundamentally depends on accurate comprehension of surface content within the images. To investigate this relationship, we examine how the quality of comic descriptions influences subsequent deep reasoning performance across the evaluated models.

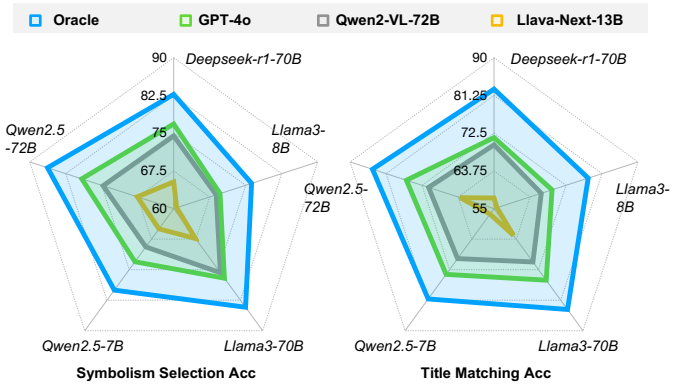

Fig. 6: LLMs' performance on deep reasoning tasks using different image descriptions as inputs. These image descriptions are from our annotations (*i.e.*, Oracle) or generated from VLMs (*e.g.*, GPT-4o, Qwen2-VL-72B, and Llava-Next-13B).

For LLMs, which rely entirely on textual descriptions as input for reasoning tasks, we conduct comparative experiments using descriptions generated by different VLMs as well as human-authored oracle descriptions. As shown in Fig. 6, descriptions generated by more advanced VLMs (specifically GPT-4o and Qwen2-VL-72B) lead to substantially improved reasoning performance across all evaluated LLMs compared to descriptions from LLaVA-Next-13B. This performance differential is consistent across LLM scales and architectures. Most notably, when provided with human-authored oracle descriptions, all LLMs achieve their highest performance scores, establishing an upper bound on potential performance.

These findings demonstrate a strong positive correlation between literal description quality and deep reasoning capabilities. The substantial performance variations observed when using identical LLMs with different input descriptions highlight the critical importance of accurate surface-level understanding as a foundation for higher-order reasoning about comic narratives. This suggests that improvements in VLMs' descriptive capabilities could yield cascading benefits for downstream reasoning tasks, even without architectural changes to reasoning components.

6.2 Methodological Enhancements

6.2.1 Does Task Decomposition Lead to Better Deep Reasoning

VLMs typically approach deep reasoning tasks in an end-to-end manner, simultaneously performing image captioning, narrative understanding, and deep reasoning when presented with comic images. This integrated approach requires models to manage multiple cognitive tasks of varying complexity concurrently. In this section, we investigate whether decomposing the complex reasoning tasks into sequential stages improves performance outcomes.

In particular, we implement a two-stage decomposition approach: first prompting the VLM to generate a literal description of the comic for surface understanding, then directing it to predict the result based on both the original comic image and the generated description. Fig. 7

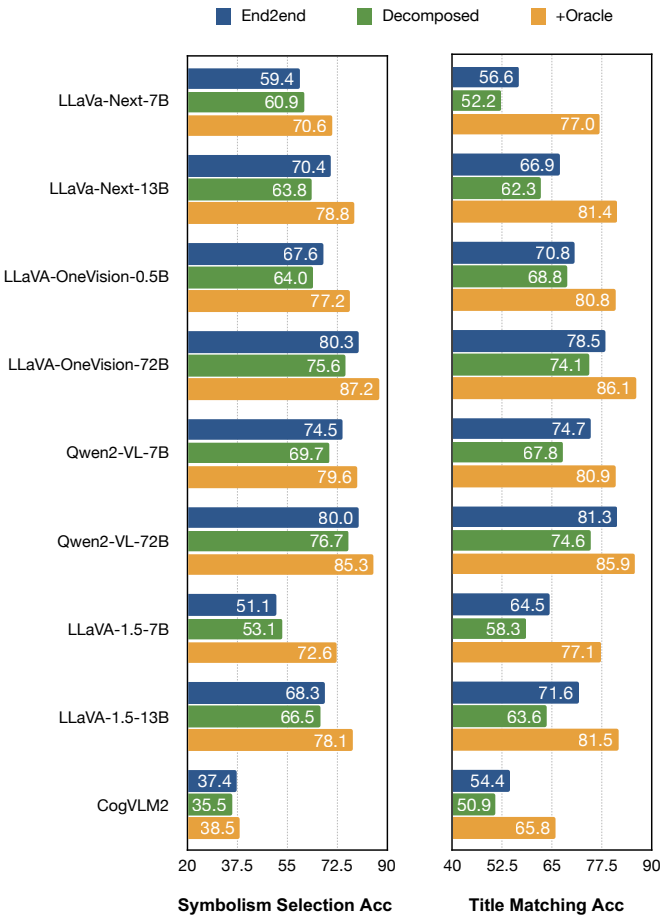

Fig. 7: Comparison of VLM deep reasoning performance in end-to-end and decomposed settings. In the decomposed setting, we additionally incorporate oracle literal descriptions provided by humans as an upper bound.

presents the comparative results between this decomposed approach and the standard end-to-end reasoning. Contrary to intuitive expectations, our findings indicate that task decomposition with model-generated descriptions does not consistently improve performance. In fact, we observe a performance decline across all models in the title selection task when augmented with their own generated descriptions. However, when we substitute model-generated descriptions with human-authored oracle descriptions in the decomposed setting, all VLMs demonstrate significant performance improvements.

These results suggest several important insights about current VLM capabilities. First, models continue to struggle with accurately recognizing critical visual elements in comic images that are essential for deep narrative reasoning. Second, the introduction of potentially flawed self-generated descriptions can propagate and amplify errors through the reasoning process. Our manual analysis reveals that models frequently generate descriptions containing hallucinations or omit crucial narrative elements (with further discussions in Section 6.4), subsequently misleading their own reasoning processes. This creates a compounding error effect that undermines the potential benefits of task decomposition. These findings highlight the importance of addressing

TABLE 4: Comparison of model performance using a *Single* comic image versus *Multi*-images split by panels. For each setting, we present results of both end-to-end and decomposed predictions as in Section 6.2.1.

Model	Setting		Symbolism Acc.	Title Acc.
LLaVA-OneVision-0.5B	*Single*	End2end	36.87	38.53
		Decomposed	**39.16**	**46.44**
	Multi	End2end	38.59	38.51
		Decomposed	35.82	40.17
LLaVA-OneVision-7B	*Single*	End2end	**67.64**	**70.83**
		Decomposed	64.00	68.83
	Multi	End2end	65.45	70.21
		Decomposed	55.55	57.05
LLaVA-OneVision-72B	*Single*	End2end	**80.67**	77.10
		Decomposed	67.59	58.56
	Multi	End2end	80.30	**78.48**
		Decomposed	75.55	74.13
Qwen2-VL-7B	*Single*	End2end	**74.48**	74.72
		Decomposed	69.73	67.75
	Multi	End2end	72.27	**75.42**
		Decomposed	64.74	62.20
Qwen2-VL-72B	*Single*	End2end	**79.98**	**81.25**
		Decomposed	76.66	74.60
	Multi	End2end	78.29	79.71
		Decomposed	71.32	63.63

fundamental visual perception and description accuracy as prerequisites for improving complex reasoning capabilities in future VLM development.

6.2.2 Does Splitting Comic Panels into Separate Images Enhance Performance

In our standard experimental setup, we evaluate VLMs by providing a single composite comic image containing both panels and requiring models to perform deep reasoning across the entire visual narrative. This approach, however, introduces potential challenges in panel disambiguation, as models must correctly identify and distinguish information from left and right panels to accurately interpret their relationship.

Given that some recent VLM architectures, such as the LLaVA-OneVision series, have been specifically trained on multiple image inputs and video sequences, we conduct analysis to see whether separating comic panels into distinct image inputs can improve performance by better aligning with their training paradigm. This approach theoretically enables models to process each panel individually before integrating information across panels, potentially facilitating more precise relationship modeling.

The results are presented in Table 4. Contrary to expectations, the results show that performance consistently decreases when using split panel inputs compared to single composite images across all evaluated models. These results suggest that the sequential processing of separated panels appears to disrupt rather than enhance models' ability to capture cross-panel relationships essential for deep reasoning. Moreover, despite training on multi-image inputs, current models still struggle with the particular cognitive challenge of identifying and reasoning about subtle narrative relationships, contradictions, and thematic connections across sequential panels. This finding highlights a critical gap in current VLMs' cross-image reasoning capabilities, particularly for narratively linked visual content that requires integrative understanding rather than independent processing of each visual element.

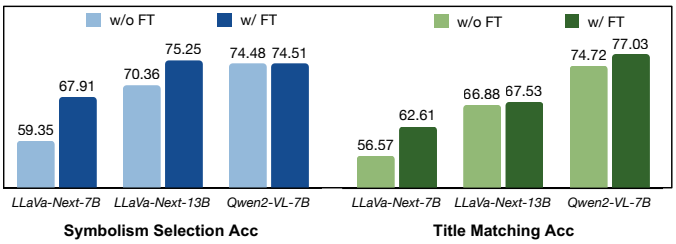

Fig. 8: Model performance on deep reasoning tasks with and without fine-tuning.

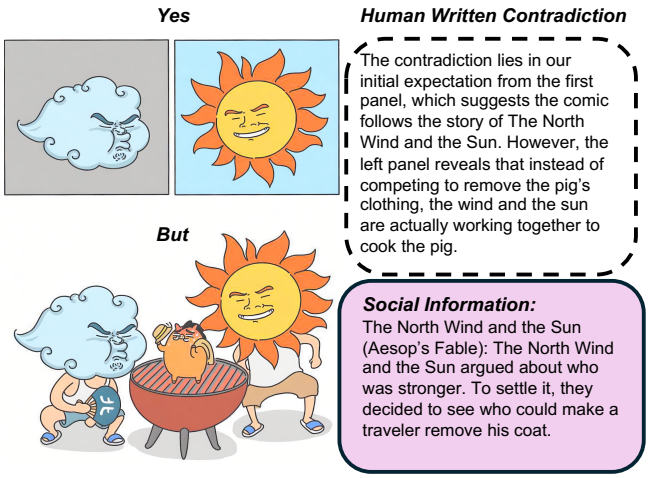

Fig. 9: A sample comic that requires additional social knowledge for full comprehension.

6.2.3 Model Finetuning for Deep Reasoning Tasks

In this section, we explore whether finetuning can enhance model performance on deep reasoning tasks for comic understanding. However, a significant challenge is the scarcity of large-scale comic datasets suitable for training. To address this limitation, we develop a text-only training methodology that circumvents the need for extensive image resources.

Our approach employs a weakly supervised textual data synthesis pipeline utilizing powerful LLMs such as GPT-4o as a data generator. Rather than requiring paired image-text data, we substitute comic images with textual descriptions of narratives, accompanied by corresponding reasoning questions. This method leverages the sophisticated text generation capabilities of LLMs while eliminating the dependency on visual data for training.

Specifically, for data generation, we carefully select 10 diverse examples from our labeled dataset to serve as few-shot prompts. Using these exemplars, we prompt GPT-4o to synthesize 20,000 contradictory comic scene descriptions. For each synthesized scene description, we further employ GPT-4o to generate corresponding questions targeting underlying symbolism selection and title matching tasks.

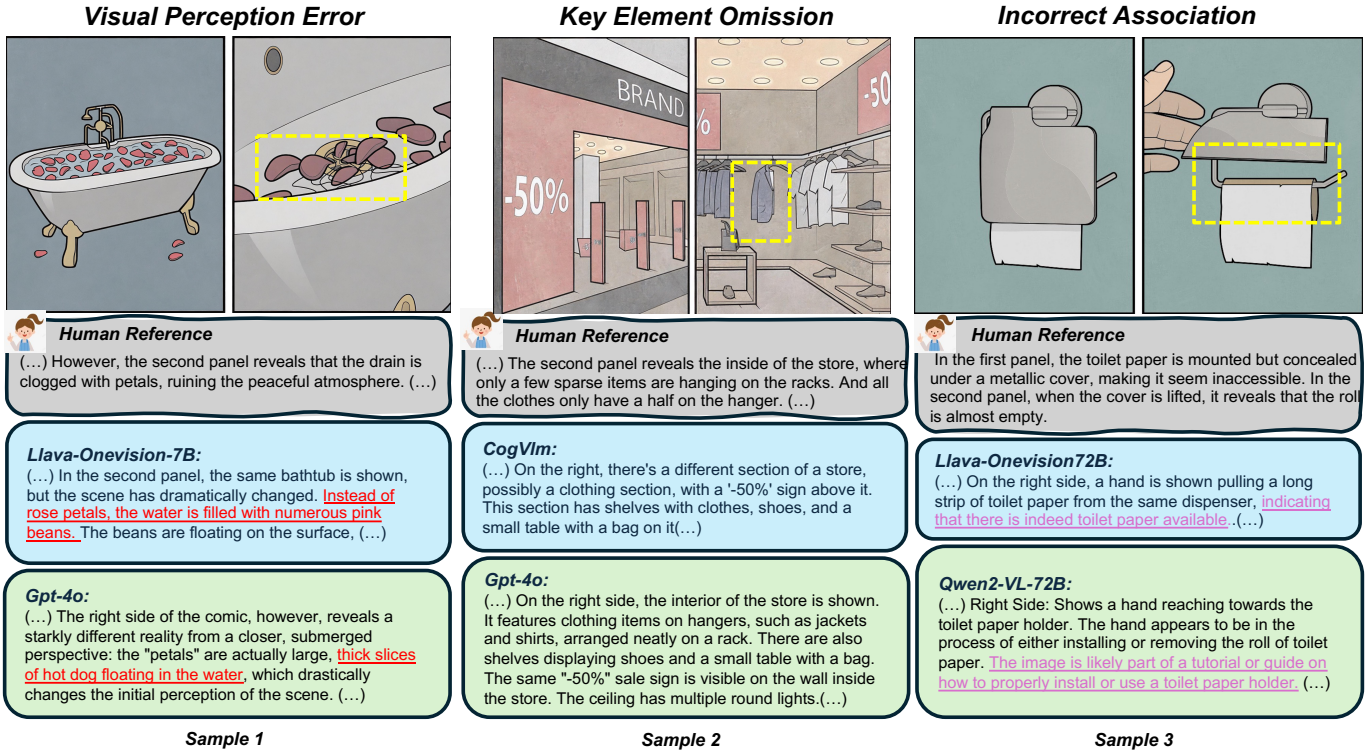

Fig. 10: Sample outputs of model-generated literal descriptions with highlighted errors of different types.

This synthetic dataset is then used to finetune only the language components of the VLMs, leaving their visual perception modules unchanged. This targeted approach allows us to enhance the models' comic understanding and reasoning capabilities while maintaining their original visual processing architecture. More detailed information regarding the model training process and data generation methodology is provided in Appendix C.4.

Fig. 8 presents the comparative results before and after finetuning. The results demonstrate consistent performance improvements across all evaluated models after finetuning, with particularly notable gains in symbolism accuracy for LLaVA-Next-7B (8.56 percentage points) and LLaVA-Next-13B (4.89 percentage points). These results validate the effectiveness of our text-only training approach for enhancing deep reasoning capabilities in VLMs, even without modifying their visual components.

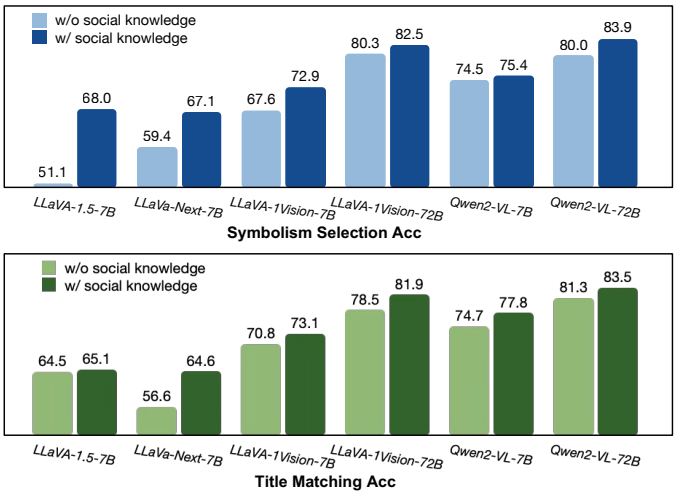

Fig. 11: The impact of external social knowledge.

6.3 Broader Enhancements with Social Knowledge Augmentation

Understanding humorous comics fundamentally requires comprehensive knowledge of social events and human behavioral norms (Fig. 9). Prior research has shown that large models continue to exhibit limitations in capturing human intent and social nuances [1]. This raises an important question: *Can explicitly integrating social knowledge enhance model comprehension of humor?* To answer this, we designed experiments to test the role of social knowledge in deep reasoning tasks related to comic understanding.

We conduct experiments by supplementing model prompts with our annotated social knowledge specific to each comic's context. As shown in Fig. 11, all evaluated VLMs achieve either improved or comparable performance when augmented with explicit social knowledge. This consistent pattern of improvement suggests that current models may possess insufficient or inadequately activated social knowledge for effectively interpreting juxtaposed humorous content. The performance gains observed through knowledge augmentation indicate that the models can effectively utilize such information when explicitly provided.

These findings suggest that enhancing models' inherent social knowledge representation or leveraging retrieval mechanisms for relevant knowledge could yield substantial improvements in human-centered reasoning tasks. This points toward promising avenues for developing more

socially-aware VLMs through targeted knowledge integration and reasoning enhancement techniques.

6.4 Case Study and Error Analysis

To provide qualitative insights into model limitations, we analyze model outputs and present common errors of VLM-generated descriptions in Fig. 10. Our analysis reveals several recurring error patterns. We categorize these errors into three types:

6.4.1 Visual Perception Error

Visual perception error occurs when models incorrectly identify visual elements, attributing erroneous identities or characteristics to objects present in the image. This error type represents a fundamental failure in visual perception. In Sample 1 (Fig. 10), we observe that LLaVA-OneVision-7B misidentifies flower petals as "pink beans," while even the more sophisticated GPT-4o incorrectly labels these same petals as "ham slices." These perceptual errors cascade into subsequent reasoning processes, establishing flawed premises that undermine higher-level understanding.

6.4.2 Key Element Omission

Key element omission errors occur when models fail to recognize or acknowledge significant visual elements present in the comic panels. This issue is particularly common in comics, where the model needs to identify the visual cues to correctly understand the overall narrative. In Sample 2 (Fig. 10), which depicts a "50% off" sale showing only half of each clothing item available for purchase (the visual punchline), both CogVLM2 and GPT-4o completely omit this critical visual element in their descriptions. Such omissions eliminate essential information required for understanding the comic's humor.

6.4.3 Incorrect Association

Incorrect association errors occur when models make up non-existent information or hallucinations for the visual content. This error type often manifests as hallucinated details or narratives that extend beyond what is present in the comic. Sample 3 (Fig. 10) demonstrates this error pattern: the comic depicts a nearly-empty paper roll that appears to have more paper than it actually contains. LLaVA-OneVision-72B incorrectly associates this image with an imagined narrative suggesting "there is still some paper left," while Qwen2-VL-72B fabricates a non-existent tutorial context. These hallucinated associations impose incorrect interpretive frameworks that fundamentally alter the comic's intended meaning.

In summary, these error patterns highlight the complex challenges in comic understanding that extend beyond simple visual recognition. Each error type disrupts a different aspect of the interpretive process, from basic perception to contextual framing, collectively undermining models' ability to grasp the nuanced meanings encoded in comic narratives.

7 Conclusion

This paper investigates this limitation through the analysis of comics that use juxtaposed panels to create humorous contradictions. We introduce YesBut, a novel, multi-tiered benchmark designed to evaluate the layered reasoning necessary for humor comprehension, ranging from basic content recognition to deep narrative inference. Our extensive experiments demonstrate that even state-of-the-art VLMs struggle to match human performance, exposing critical gaps in their ability to grasp nuanced contextual relationships. To address these shortcomings, we propose a novel text-only training strategy that synthesizes textual data to strengthen VLMs' language processing capabilities—eliminating the need for costly image-text paired training data without sacrificing performance. Beyond identifying key weaknesses in VLMs' understanding of cultural and creative expressions, our findings chart a promising path toward more robust, context-aware AI models capable of deeper narrative reasoning. This work lays the foundation for improving AI's ability to process humor, contradictions, and complex multimodal narratives, ultimately advancing human-AI interaction in creative domains.

Acknowledgments

All data samples collected are sourced from publicly available content on social media platforms. To maintain content integrity, we carefully review and filter out any samples that may contain offensive or harmful material. The Large Vision-Language Models (VLMs) used in our experiments are pretrained on diverse web corpora, which may inherently introduce biases into their outputs. We encourage users to critically assess the ethical considerations of generated outputs when applying them in future research. Our annotation process involves a team of ten human judges, each compensated with an average hourly wage of $11, ensuring fair and ethical remuneration for their contributions.

This work made use of the High Performance Computing Resource in the Core Facility for Advanced Research Computing at Case Western Reserve University, which is supported by NSF award NSF-2117439. We also thank the support from OpenAI Researcher Access grants #0000007745.

References

[1] Z. Hu and T. Shu, "Language models, agent models, and world models: The law for machine reasoning and planning," *arXiv preprint arXiv:2312.05230*, 2023. 1, 12

[2] J. Hessel, A. Marasovic, J. D. Hwang, L. Lee, J. Da, R. Zellers, R. Mankoff, and Y. Choi, "Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 688–714. [Online]. Available: https://aclanthology.org/2023.acl-long.41 1, 3

[3] A. Rayhan, R. Rayhan, and S. Rayhan, "Artificial general intelligence: Roadmap to achieving human-level capabilities," 2023. 1

[4] M. Koivisto and S. Grassini, "Best humans still outperform artificial intelligence in a creative divergent thinking task," *Scientific reports*, vol. 13, no. 1, p. 13601, 2023. 1

[5] R. Duncan and M. J. Smith, *The power of comics: History, form and culture*. A&C Black, 2009. 1

[6] Y. Yang, Z. Li, Q. Dong, H. Xia, and Z. Sui, "Can large multimodal models uncover deep semantics behind images?" *arXiv preprint arXiv:2402.11281*, 2024. 1, 3

[7] J. O. Young, *Art and knowledge*. Routledge, 2003. 1

[8] T. Groensteen, *Comics and narration*. Univ. Press of Mississippi, 2013. 1

[9] E. Bearne, "Rethinking literacy: Communication, representation and text," *Reading*, vol. 37, no. 3, pp. 98–103, 2003. 1

[10] J. Dittmer, "Comic book visualities: a methodological manifesto on geography, montage and narration," *Transactions of the Institute of British Geographers*, vol. 35, no. 2, pp. 222–236, 2010. 1

[11] J. Schechter, "Juxtaposition: A new way to combine logics," *The Review of Symbolic Logic*, vol. 4, no. 4, pp. 560–606, 2011. 1

[12] P. J. Kuttner, M. B. Weaver-Hightower, and N. Sousanis, "Comics-based research: The affordances of comics for research across disciplines," *Qualitative Research*, vol. 21, no. 2, pp. 195–214, 2021. 1

[13] Y. Tong, Y. Wang, D. Li, S. Wang, Z. Lin, S. Han, and J. Shang, "Eliminating reasoning via inferring with planning: A new framework to guide llms' non-linear thinking," *arXiv preprint arXiv:2310.12342*, 2023. 1

[14] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg *et al.*, "Sparks of artificial general intelligence: Early experiments with gpt-4," *arXiv preprint arXiv:2303.12712*, 2023. 1

[15] Z. Hu, T. Liang, J. Li, Y. Lu, Y. Zhou, Y. Qiao, J. Ma, and Y. Yin, "Cracking the code of juxtaposition: Can ai models understand the humorous contradictions," in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds., vol. 37. Curran Associates, Inc., 2024, pp. 47 166–47 188. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2024/file/540a6eefb60428c8547a27253f9a2a59-Paper-Conference.pdf 2, 3

[16] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *Advances in neural information processing systems*, vol. 35, pp. 27 730–27 744, 2022. 2

[17] AI@Meta, "Llama 3 model card," 2024. [Online]. Available: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md 2, 6, 17

[18] S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, and J. Gao, "Large language models: A survey," *arXiv preprint arXiv:2402.06196*, 2024. 2

[19] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen, "A survey on multimodal large language models," *arXiv preprint arXiv:2306.13549*, 2023. 2

[20] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763. 2

[21] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing *et al.*, "Judging llm-as-a-judge with mt-bench and chatbot arena," *Advances in Neural Information Processing Systems*, vol. 36, 2024. 3

[22] Y. Dubois, C. X. Li, R. Taori, T. Zhang, I. Gulrajani, J. Ba, C. Guestrin, P. S. Liang, and T. B. Hashimoto, "Alpacafarm: A simulation framework for methods that learn from human feedback," *Advances in Neural Information Processing Systems*, vol. 36, 2024. 3

[23] Y. Wang, Z. Yu, Z. Zeng, L. Yang, C. Wang, H. Chen, C. Jiang, R. Xie, J. Wang, X. Xie *et al.*, "Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization," *arXiv preprint arXiv:2306.05087*, 2023. 3

[24] Y. Huang, Y. Bai, Z. Zhu, J. Zhang, J. Zhang, T. Su, J. Liu, C. Lv, Y. Zhang, Y. Fu *et al.*, "C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models," *Advances in Neural Information Processing Systems*, vol. 36, 2024. 3

[25] K. Ying, F. Meng, J. Wang, Z. Li, H. Lin, Y. Yang, H. Zhang, W. Zhang, Y. Lin, S. Liu *et al.*, "Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi," *arXiv preprint arXiv:2404.16006*, 2024. 3

[26] Y. Bitton, H. Bansal, J. Hessel, R. Shao, W. Zhu, A. Awadalla, J. Gardner, R. Taori, and L. Schimdt, "Visit-bench: A benchmark for vision-language instruction following inspired by real-world use," *arXiv preprint arXiv:2308.06595*, 2023. 3

[27] N. Bitton-Guetta, Y. Bitton, J. Hessel, L. Schmidt, Y. Elovici, G. Stanovsky, and R. Schwartz, "Breaking common sense: Whoops! a vision-and-language benchmark of synthetic and compositional images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2616–2627. 3

[28] B. Li, R. Wang, G. Wang, Y. Ge, Y. Ge, and Y. Shan, "Seed-bench: Benchmarking multimodal llms with generative comprehension," *arXiv preprint arXiv:2307.16125*, 2023. 3

[29] B. Li, Y. Ge, Y. Ge, G. Wang, R. Wang, R. Zhang, and Y. Shan, "Seed-bench-2: Benchmarking multimodal large language models," *arXiv preprint arXiv:2311.17092*, 2023. 3

[30] L. Parcalabescu, M. Cafagna, L. Muradjan, A. Frank, I. Calixto, and A. Gatt, "Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena," *arXiv preprint arXiv:2112.07566*, 2021. 3

[31] Y. Oh, P. Ahn, J. Kim, G. Song, S. Lee, I. S. Kweon, and J. Kim, "Exploring the spectrum of visio-linguistic compositionality and recognition," *arXiv preprint arXiv:2406.09388*, 2024. 3

[32] D. Roberts and L. Roberts, "Smart vision-language reasoners," *arXiv preprint arXiv:2407.04212*, 2024. 3

[33] Z. Hu, Y. Ren, J. Li, and Y. Yin, "VIVA: A benchmark for vision-grounded decision-making with human values," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 2294–2311. [Online]. Available: https://aclanthology.org/2024.emnlp-main.137/ 3

[34] R. Li, S. Sun, M. Elhoseiny, and P. Torr, "Oxfordtvg-hic: Can machine make humorous captions from images?" in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 20 293–20 303. 3

[35] J. Palmer, *Taking humour seriously*. Routledge, 2003. 3

[36] N. De Pisapia, F. Bacci, D. Parrott, and D. Melcher, "Brain networks for visual creativity: a functional connectivity study of planning a visual artwork," *Scientific reports*, vol. 6, no. 1, p. 39185, 2016. 3

[37] L. Chen and C. M. Lee, "Predicting audience's laughter using convolutional neural network," *arXiv preprint arXiv:1702.02584*, 2017. 3

[38] A. Cattle and X. Ma, "Recognizing humour using word associations and humour anchor extraction," in *Proceedings of the 27th international conference on computational linguistics*, 2018, pp. 1849–1858. 3

[39] D. Yang, A. Lavie, C. Dyer, and E. Hovy, "Humor recognition and humor anchor extraction," in *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2015, pp. 2367–2376. 3

[40] M. Amin and M. Burghardt, "A survey on approaches to computational humor generation," in *Proceedings of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, S. DeGaetano, A. Kazantseva, N. Reiter, and S. Szpakowicz, Eds. Online: International Committee on Computational Linguistics, Dec. 2020, pp. 29–41. [Online]. Available: https://aclanthology.org/2020.latechclfl-1.4 3

[41] S. Jentzsch and K. Kersting, "Chatgpt is fun, but it is not funny! humor is still challenging large language models," *arXiv preprint arXiv:2306.04563*, 2023. 3

[42] S. Zhong, Z. Huang, S. Gao, W. Wen, L. Lin, M. Zitnik, and P. Zhou, "Let's think outside the box: Exploring leap-of-thought in large language models with creative humor generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 246–13 257. 3

[43] D. Shahaf, E. Horvitz, and R. Mankoff, "Inside jokes: Identifying humorous cartoon captions," in *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 1065–1074. 3

[44] D. Radev, A. Stent, J. Tetreault, A. Pappu, A. Iliakopoulou, A. Chanfreau, P. de Juan, J. Vallmitjana, A. Jaimes, R. Jha *et al.*, "Humor in collective discourse: Unsupervised funniness detection in the new yorker cartoon caption contest," *arXiv preprint arXiv:1506.08126*, 2015. 3

[45] V. Jain, F. d. S. A. Feitosa, and G. Kreiman, "Is ai fun? humordb: a curated dataset and benchmark to investigate graphical humor," *arXiv preprint arXiv:2406.13564*, 2024. 3

[46] A. Chandrasekaran, A. K. Vijayakumar, S. Antol, M. Bansal, D. Batra, C. L. Zitnick, and D. Parikh, "We are humor beings: Understanding and predicting visual humor," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4603–4612. 3

[47] Y. Kayatani, Z. Yang, M. Otani, N. Garcia, C. Chu, Y. Nakashima, and H. Takemura, "The laughing machine: Predicting humor in video," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 2073–2082. 3

[48] Y. Liu, T. Shen, D. Zhang, Q. Sun, S. Li, and G. Zhou, "Comment-aided video-language alignment via contrastive pre-training for short-form video humor detection," *arXiv preprint arXiv:2402.09055*, 2024. 3

[49] E. Hwang and V. Shwartz, "MemeCap: A dataset for captioning and interpreting memes," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 1433–1445. [Online]. Available: https://aclanthology.org/2023.emnlp-main.89 3, 18

[50] D. Radev, A. Stent, J. Tetreault, A. Pappu, A. Iliakopoulou, A. Chanfreau, P. de Juan, J. Vallmitjana, A. Jaimes, R. Jha, and R. Mankoff, "Humor in collective discourse: Unsupervised funniness detection in the new yorker cartoon caption contest," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds. Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 475–479. [Online]. Available: https://aclanthology.org/L16-1076 3

[51] X. Wang, H. Xia, J. Song, L. Guan, Y. Yang, Q. Dong, W. Luo, Y. Pu, Y. Wang, X. Meng *et al.*, "Beyond single frames: Can lmms comprehend temporal and contextual narratives in image sequences?" *arXiv preprint arXiv:2502.13925*, 2025. 3

[52] Z. Chen, Q. Zhou, Y. Shen, Y. Hong, H. Zhang, and C. Gan, "See, think, confirm: Interactive prompting between vision and language models for knowledge-based visual reasoning," *arXiv preprint arXiv:2301.05226*, 2023. 3

[53] F. Ke, Z. Cai, S. Jahangard, W. Wang, P. D. Haghighi, and H. Rezatofighi, "Hydra: A hyper agent for dynamic compositional visual reasoning," in *European Conference on Computer Vision*. Springer, 2024, pp. 132–149. 3

[54] T. Gupta and A. Kembhavi, "Visual programming: Compositional visual reasoning without training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 953–14 962. 3

[55] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-vl: A frontier large vision-language model with versatile abilities," *arXiv preprint arXiv:2308.12966*, 2023. 3, 17

[56] W. Wang, Q. Lv, W. Yu, W. Hong, J. Qi, Y. Wang, J. Ji, Z. Yang, L. Zhao, X. Song *et al.*, "Cogvlm: Visual expert for pretrained language models," *arXiv preprint arXiv:2311.03079*, 2023. 3

[57] Q. Ye, H. Xu, J. Ye, M. Yan, A. Hu, H. Liu, Q. Qian, J. Zhang, F. Huang, and J. Zhou, "mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration," 2023. 3

[58] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023. 3, 6

[59] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee, "Llava-next: Improved reasoning, ocr, and world knowledge," January 2024. [Online]. Available: https://llava-vl.github.io/blog/2024-01-30-llava-next/ 3, 6, 8, 17

[60] Y. Wang and Y. Zhao, "Gemini in reasoning: Unveiling commonsense in multimodal large language models," *arXiv preprint arXiv:2312.17661*, 2023. 3

[61] R. Zellers, Y. Bisk, A. Farhadi, and Y. Choi, "From recognition to cognition: Visual commonsense reasoning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6720–6731. 3

[62] D. A. Hudson and C. D. Manning, "Gqa: A new dataset for real-world visual reasoning and compositional question answering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6700–6709. 3

[63] T. Thrush, R. Jiang, M. Bartolo, A. Singh, A. Williams, D. Kiela, and C. Ross, "Winoground: Probing vision and language models for visio-linguistic compositionality," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5238–5248. 3

[64] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, and A. Kalyan, "Learn to explain: Multimodal reasoning via thought chains for science question answering," *Advances in Neural Information Processing Systems*, vol. 35, pp. 2507–2521, 2022. 3

[65] M. Yu, Z. Zhang, W. Yu, and M. Jiang, "Pre-training language models for comparative reasoning," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 12 421–12 433. [Online]. Available: https://aclanthology.org/2023.emnlp-main.763/ 3

[66] N. Jindal and B. Liu, "Identifying comparative sentences in text documents," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, 2006, pp. 244–251. 3

[67] J. Kil, Z. Mai, J. Lee, A. Chowdhury, Z. Wang, K. Cheng, L. Wang, Y. Liu, and W.-L. H. Chao, "Mllm-compbench: A comparative reasoning benchmark for multimodal llms," *Advances in Neural Information Processing Systems*, vol. 37, pp. 28 798–28 827, 2024. 3

[68] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," *arXiv preprint arXiv:2310.03744*, 2023. 6, 17

[69] W. Hong, W. Wang, M. Ding, W. Yu, Q. Lv, Y. Wang, Y. Cheng, S. Huang, J. Ji, Z. Xue *et al.*, "Cogvlm2: Visual language models for image and video understanding," *arXiv preprint arXiv:2408.16500*, 2024. 6, 17

[70] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu *et al.*, "Llava-onevision: Easy visual task transfer," *arXiv preprint arXiv:2408.03326*, 2024. 6, 17

[71] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge *et al.*, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *arXiv preprint arXiv:2409.12191*, 2024. 6

[72] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford *et al.*, "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024. 6

[73] D. Zhu, J. Chen, X. Shen, xiang Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," 2023. 6

[74] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi *et al.*, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025. 6

[75] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei *et al.*, "Qwen2. 5 technical report," *arXiv preprint arXiv:2412.15115*, 2024. 6

[76] A. Celikyilmaz, E. Clark, and J. Gao, "Evaluation of text generation: A survey," *arXiv preprint arXiv:2006.14799*, 2020. 7

[77] M. Gao, X. Hu, J. Ruan, X. Pu, and X. Wan, "Llm-based nlg evaluation: Current status and challenges," *arXiv preprint arXiv:2402.01383*, 2024. 7

[78] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, "Judging llm-as-a-judge with mt-bench and chatbot arena," in *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. [Online]. Available: https://openreview.net/forum?id=uccHPGDlao 7

[79] A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang, G. Dong, H. Wei, H. Lin, J. Tang, J. Wang, J. Yang, J. Tu, J. Zhang, J. Ma, J. Yang, J. Xu, J. Zhou, J. Bai, J. He, J. Lin, K. Dang, K. Lu, K. Chen, K. Yang, M. Li, M. Xue, N. Ni, P. Zhang, P. Wang, R. Peng, R. Men, R. Gao, R. Lin, S. Wang, S. Bai, S. Tan, T. Zhu, T. Li, T. Liu, W. Ge, X. Deng, X. Zhou, X. Ren, X. Zhang, X. Wei, X. Ren, X. Liu, Y. Fan, Y. Yao, Y. Zhang, Y. Wan, Y. Chu, Y. Liu, Z. Cui, Z. Zhang, Z. Guo, and Z. Fan, "Qwen2 technical report," 2024. [Online]. Available: https://arxiv.org/abs/2407.10671 9

Appendix A
Ethics Statement

A.1 Copyright and License

All data samples collected are sourced from publicly available content on social media platforms. We ensure compliance with copyright by utilizing original links to comics without infringement. In addition, we obtained permission from the author artist (e.g., Anton Gudim, Liz Climo) to conduct our benchmark using these public images. Additionally, we commit to open-sourcing our annotated benchmark, providing corresponding links to each comic image. We diligently review samples, filtering out potentially offensive or harmful content.

A.2 Human Annotation

Ten human annotators participate in our labeling process, receiving an average hourly wage of $11 to ensure fair compensation. We take steps to mitigate biases by maintaining a diverse group of annotators and providing clear annotation guidelines.

A.3 Model Bias & Ethical Considerations

The large vision-language models (VLMs) used in our experiments are pretrained on diverse web corpora, which may introduce biases into their outputs. We encourage users to critically assess potential ethical implications when applying these models in future research.

Appendix B
Data Annotation Details

B.1 Annotation Prompts

To balance efficiency and accuracy, we employ an AI-human collaborative pipeline for annotation. The AI assists in generating initial components, while human annotators refine outputs to ensure quality. The specific prompts used for AI generation are detailed in Fig. 12. We present a sample comic with all tasks in Fig. 13.

B.2 Detailed Data Distribution

Table 5 provides a detailed breakdown of the dataset distribution based on different attributes, including Presence of Text, Social Norms, and Humor Categories. To cluster images under humor categories, a compact and informative descriptive sentence is generated for each image based on its oracle description. These sentences are then processed using GPT-4 to identify thematic patterns, grouping them into distinct humor-related categories. This process results in 15 well-defined classes, each representing a unique type of humor. The diverse range of everyday life scenarios captured in our dataset provides a strong foundation for evaluating humor understanding in various contexts.

TABLE 5: Dataset Statistics Overview

Attributes	Subcategory		# of image
Presence of Text		No Text	530
	With Text	Linguistic-Dependent	370
		Transferable Text	362
Social Norms		Yes	1091
		No	171
Humor Categories		Work Jokes & Complaints	233
		Internet Culture & Technology	202
		Fashion Trends & Jokes	112
		Travel & Adventure	106
		Finance & Money Matters	100
		Sports & Fitness	92
		Food & Culinary Experiences	77
		Daily Life & Routine Humor	66
		Relationships & Social Life	62
		Entertainment & Pop Culture	53
		Education & Student Life	46
		Pet Behavior & Humor	37
		Health & Wellness	36
		Shopping & Consumer Behavior	24
		Holiday & Seasonal Events	16

Tasks	Prompts
Literal Description & Contradiction	*The given comic with two panels shows the same situation from two opposite sides with contradictions. You need to first read and understand the comic. Generate a detailed description to illustrate the narrative of the comic and explain the contradiction of what makes the comic interesting or sarcastic.*
Symbolism	*Write a brief description of the underlying moral of the narrative in one sentence, and include what phenomenon is it satirizing and what we can learn from the comic.*
Title	*Produce a short eye-catching title reflecting the narrative.*
Negative Symbolism	*Generate five contextualized, plausible, but ultimately incorrect criticisms and moral lessons we can learn from the image, each in one sentence as distracters. Keep the length and style the same as the correct one.*
Negative Title	*Provide five seemingly reasonable, eye-catching but incorrect titles.*
Social Knowledge/Norms	*Based on the image, provide me the necessary social knowledge of the comic.*

Fig. 12: Prompts for Data Annotation

Literal Description:
In the first panel, titled "Beginning of Dating," the couple shares a sweet, romantic moment while eating together. The woman, smiling lovingly, holds up a fork and says, "Say ahhh!" The man, cheerful and delighted, eagerly opens his mouth and responds, "AHHH~" The scene is filled with affection, care, and mutual admiration as they politely share food. In the second panel, titled "After a Few Years," the same couple is shown in a dramatically different and chaotic scenario. The woman is sprawled across the couch with a bag of snacks, greedily stuffing food into her mouth while crumbs fly everywhere. Meanwhile, the man desperately reaches for the snacks, shouting "GIMME SOME!!!" She nonchalantly pushes his face away with her hand, illustrating how their dynamic has shifted from loving gestures to playful, messy food fights over time.

Contradiction:
The contradiction in this comic lies in the shift from polite, romantic gestures early in a relationship to unfiltered, competitive behavior once the couple becomes more comfortable with each other.

Underlying Philosophy Selection:
A. The comic implies that early playful, light-hearted interactions fade in established relationships, wrongly contrasting how dynamics mature over time to maintain humor and warmth naturally.
B. The comic highlights food sharing humor as central to evolving relationships, suggesting continuity from romantic to relational depth, which mistakenly implies a lack of deeper bonding over time.
C. The comic illustrates how relationships evolve from idealized beginnings to more humorous and relaxed dynamics as partners grow closer and drop pretenses.
D. The comic promotes playful competition as common in relationships from the start, mistakenly suggesting that such dynamics are immediate and not the product of evolving comfort over time.
Answer: C

Title Matching:
A. Dining Manners: Everlasting Elegance
B. Perfect Meals: Romantic Illusion
C. Eating Together: From Sweet to Chaotic
D. Date Nights: Posing Perfection
Answer: A

Social Knowledge:
1. Relationship Dynamics: Early in relationships, people often display more romantic and considerate behavior. Over time, as they become more comfortable, their interactions can become more casual and playful.

Linguistic context: English

Contain_text: yes

Fig. 13: Sample Comic with All Annotated Tasks.

Appendix C
Experiments Details

C.1 Model Details

Our experiments include both cutting-edge proprietary and open-source VLMs and LLMs, enabling a comprehensive evaluation across diverse architectures. For commercial VLMs, we use GPT-4o (*gpt-4o-2024-08-06*) and GPT-4-Vision-turbo (*gpt-4-turbo-2024-04-09*) [3].

Among open-source VLMs, our selection includes LLaVA-Next, covering 7B, 13B, and 72B parameter sizes [59], as well as LLaVA-1.5 in 7B and 13B variants [68]. We also incorporate CogVLM2 [69], Qwen2-VL with 7B and 72B versions [55], and LLaVA-OneVision, which is available in 7B and 72B configurations [70].

For LLMs, we use the Llama 3 instruction variant in both 8B and 70B sizes [17], GPT-4 (*gpt-4-0613*), DeepSeek-R1-70B (*DeepSeek-R1-Distill-Llama-70B*), and Qwen2.5, available in 7B and 72B versions.

C.2 Implementation Details

All commercial models are accessed through their official API, while open-sourced models are implemented using Hugging Face Transformers [4]. Inference for GPT-3, GPT-4, GPT-4o, and GPT-4-Vision-Turbo is performed with a temperature of 1.0, while other models follow their default parameter settings or use greedy decoding. Experiments are conducted on A100 (80GB) and A6000 GPUs.

For multiple-choice question (MCQ) evaluation, the models are explicitly instructed to directly output an option in the prompt. Answers are parsed using hard rules, and if no valid option is detected, a random choice is assigned. For generation task evaluation, we apply rouge-score [5] to

3. https://platform.openai.com/docs/models/

4. https://huggingface.co/docs/transformers/en/index
5. https://pypi.org/project/rouge-score/

Prompts for Literal Description:

- Candidate literal description: gen
- Reference literal description: ref

Task: You need to determine how accurately the above candidate literal description matches the given reference literal description of a comic narrative.

Using a scale from 1 to 5, rate the accuracy with which the candidate description matches the reference description, with 1 being the least accurate and 5 being the most accurate.

Please directly output a score by strictly following this format: [[score]], for example: Rating: [[3]].

Prompts for Contradiction:

Background: You are an impartial judge. You will be given a literal description of a comic that presents the same situation from two opposing perspectives, highlighting contradictions. You will also be provided with a gold-standard illustration as reference that effectively demonstrates these narrative contradictions.

Your task is to evaluate the quality of a generated illustration and determine whether it accurately depicts the narrative contradictions in the comic. Then, assign a score on a scale of 1 to 5, where 1 is the lowest and 5 is the highest, based on its quality.

- The literal description of the comic: description
- The reference contradiction illustration: ref
- The generated contradiction illustration: gen

Please directly output a score by strictly following this format: [[score]], for example: Rating: [[3]].

Fig. 14: Prompts for GPT-based Evaluations

compute ROUGE score, and calculate the BERT score using the official implementation [6]. For GPT based evaluations for literal description and contradiction, we use *gpt-3.5-turbo-0125* version. The prompts we used are shown in Fig. 14.

C.3 Human Evaluations

For literal description writing and contradiction generation tasks, we randomly select 40 samples from each task for human evaluation. Model outputs are anonymized and shuffled before being presented to the reviewer. Following [49], the evaluation considers three key aspects:

- **Correctness**: The model's output accurately conveys the narrative of the comic.
- **Completeness**: The model's output covers all important elements of the comic's narrative.
- **Faithfulness**: All content in the model's output is supported by the comic image, with no hallucinated information.

6. https://github.com/Tiiiger/bert_score

Data Generation Prompt:

You are a creative comic writer and a question-generation expert. Your task is to craft engaging narratives for ten two-panel comics. Each comic should depict the same situation from two contrasting perspectives, emphasizing contradictions or opposing viewpoints. The narratives should be thought-provoking, emotionally engaging, and capable of sparking discussions about some popular topics and events of daily life.
After crafting the comic narrative, generate two challenging multiple-choice questions: - Moral question: Explore the underlying moral or philosophical lesson presented in the comic. - Title question: Encourage thoughtful selection of a title that best captures the essence of the comic.
The questions should inspire deep reflection and provide meaningful answer choices that encourage nuanced thinking.
Below are several examples:

Example_1

Example_2

...

Example_10
Now generate five comic and questions. The output should strictly be a jsonlist, with output presented as a JSON object:

Fig. 15: Prompts used for Data Generation.

For literal description writing, we evaluate all three aspects, while for contradiction generation, only correctness and faithfulness are assessed.

C.4 Model Finetuning Details for Deep Reasoning Tasks

Our approach employs a weakly supervised textual data synthesis pipeline using powerful LLMs, such as GPT-4o, as a data generator. Instead of relying on paired image-text data, we replace comic images with textual descriptions of narratives, accompanied by corresponding reasoning questions. This approach leverages the advanced text generation capabilities of LLMs while eliminating the need for visual data during training.

Data Generation. To construct the dataset, we manually select 10 diverse comic samples and utilize the few-shot learning ability of GPT-4o to generate 20,000 scene descriptions using the prompt shown in Fig. 15. A temperature of 1.0 is set for GPT-4o to encourage diversity in scene generation.

Next, based on these initial exemplars, GPT-4o synthesizes 20,000 contradictory comic scene descriptions. For each generated scene, additional prompts are used to create reasoning questions targeting symbolism selection and title matching tasks.

Finetuning Process. The generated dataset is then used to finetune *the language components* of VLMs while keeping their visual perception modules unchanged. This targeted finetuning enhances the models' comic understanding and reasoning abilities while preserving their original visual

TABLE 6: Comparison of Training Hyperparameters for Qwen2-VL-7B, LLaVA-13B, and LLaVA-7B.

Hyperparameter	Qwen2-VL-7B	LLaVA-Next-13B	LLaVA-Next-7B
Per-device Train Batch Size	4	4	4
Gradient Accumulation Steps	4	4	4
Floating Point Precision	bf16	bf16	bf16
Logging Steps	25	25	25
Evaluation Strategy	Steps	Steps	Steps
Evaluation Steps	500	500	500
Save Steps	500	500	500
LoRA Rank (r)	256	128	32
LoRA Alpha	1536	512	64
Number of Training Epochs	5	5	5
LoRA Target Modules	q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj	q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj	q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj

TABLE 7: Example evaluation prompts used for Literal Description, Contradiction Generation, Symbolism Selection, and Title Matching. Three distinct prompts are designed for each task to minimize bias and ensure robust model evaluation.

Task	Prompts
LiteralDescription	**Prompt 1:** The given comic shows the same situation from two opposite sides with contradictions. Write a one-paragraph literal description to describe the narrative of the comic. **Prompt 2:** Please literally describe the context of the image in detail. **Prompt 3:** Give me a detailed literal description of the image.
Contradiction Generation	**Prompt 1:** The given comic shows the same situation from two opposite sides with contradictions. Write a short explanation to illustrate the contradiction of the two sides. **Prompt 2:** Analyze the provided image, which is divided into two or more panels, each illustrating contrasting views of the same scenario. Describe the elements visible in each panel. Then concisely interpret how these elements convey contrasting perspectives in one or two sentences. Focus and only output the contradiction. **Prompt 3:** Given an image with two or more panels showing a contrast relationship, describe the elements visible in each panel and concisely interpret the contradiction in one or two sentences.
Underlying Symbolism Selection	**Prompt 1:** The given comic shows the same situation from two opposite sides with contradictions. Which of the following options best represents the underlying Symbolism of the comic? {MCQ Options} Just output the choice. **Prompt 2:** You are presented with an image divided into panels, each illustrating contrasting views of the same scenario. Which of the following options best represents the Symbolism of the image provided? {MCQ Options} Select the correct option by typing the corresponding letter (A, B, C, or D). **Prompt 3:** Given an image with two or more panels showing contrast, select the best option representing the deep semantic of the image. {MCQ Options} Just output the correct option as (A, B, C, or D), no more explanation.
Title Matching	**Prompt 1:** The given comic presents the same situation from two opposing perspectives, highlighting contradictions. Which of the following titles is most suitable for the comic? {MCQ Options} Output only the selected choice. **Prompt 2:** You are presented with an image divided into two or more panels, each depicting contrasting perspectives of the same scenario. Which of the following title options best represents the given image? {MCQ Options} Select the correct option by typing the corresponding letter (A, B, C, or D). **Prompt 3:** Given an image divided into two or more panels, a contrast relationship exists between the panels. Identify the best title from the following options that represents the image. {MCQ Options} Output only the corresponding letter (A, B, C, or D) without any additional explanation.

processing architecture. Table 6 presents the key training hyperparameters used for fine-tuning Qwen2, LLaVA-Next-7B, and LLaVA-Next-13B via the LoRA method.

C.5 Evaluation Prompts

To ensure a fair evaluation, we design three distinct prompts, each independently crafted by different individuals to minimize biases introduced by prompt variations. Model performance is assessed using these prompts, and

results are averaged across all tasks.

The prompts are designed to assess diverse aspects of model reasoning and understanding, covering Literal Description, Contradiction Generation, Symbolism Selection, and Title Matching are provided in Table 7.

Appendix D
More Results

D.1 More Sample Outputs

We present more randomly picked sample outputs on literal description and contradiction generation in Fig. 16 and Fig. 17.

D.2 Impact of Language Context on Deep Reasoning Tasks

We analyze how different language backgrounds influence a model's ability to understand juxtaposition humor in deep reasoning tasks. Based on our statistical results in Table 8, we compare model performance across three settings: language-independent, Chinese-context, and English-context.

In the English-context setting, LLaVA-OneVision-72B and GPT-4o achieved the highest accuracy in deep reasoning tasks. Specifically, LLaVA-OneVision-72B attained 86.98% accuracy in symbolism selection, while GPT-4o achieved 81.90% in title matching.

TABLE 8: Comparison of model performance on deep reasoning tasks across different language contexts. Accuracy results for Symbolism Selection and Title Matching are reported for Language-independent, English-context, Chinese-context, and all images.

Model	Setting	Symbolism Acc.	Title Acc.
GPT-4o	Language-Free	81.12	81.65
	Chinese-Context	75.23	75.23
	English-Context	83.81	81.90
	All	80.38	80.62
LLaVA-OneVision-0.5B	Language-Free	40.62	42.21
	Chinese-Context	24.30	20.56
	English-Context	27.62	41.90
	All	36.87	38.53
LLaVA-OneVision-7B	Language-Free	69.67	72.96
	Chinese-Context	55.61	57.94
	English-Context	73.33	77.14
	All	67.64	70.81
LLaVA-OneVision-72B	Language-Free	81.23	79.92
	Chinese-Context	72.74	70.56
	English-Context	86.98	81.27
	All	80.30	78.48
Qwen2-VL-7B	Language-Free	73.59	73.81
	Chinese-Context	75.23	78.04
	English-Context	80.95	76.19
	All	74.48	74.72
Qwen2-VL-72B	Language-Free	78.79	80.81
	Chinese-Context	83.18	83.18
	English-Context	83.81	80.95
	All	79.98	81.25

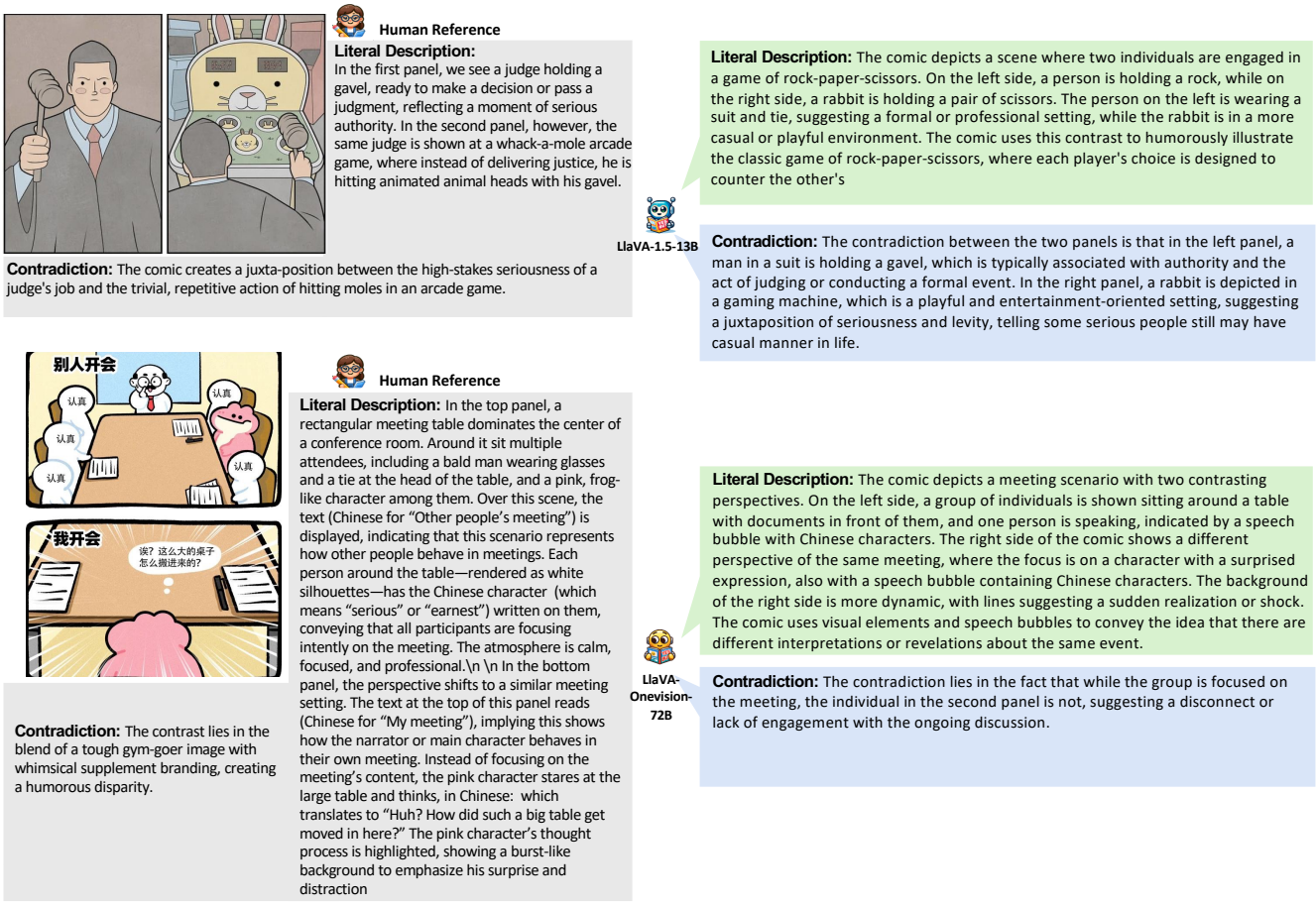

Fig. 16: Sample outputs of model generated literal description and contradiction.

Human Reference

Literal Description: In the first two panels, a happy woman lovingly holds her adorable dog close to her face, exclaiming, “Awwww!” and “Who’s a good boy?” as the dog excitedly licks her cheek. The moment is filled with warmth and affection, showcasing the deep connection between pet and owner. The bottom panel, labeled “5 MINUTES EARLIER”, hilariously reveals what the dog was doing just moments before. The same dog is sprawled out on the floor in an undignified position, enthusiastically licking its own butt. In the background, a purple dog toy and food bowls labeled “BOEDY” add to the humor, emphasizing the all-too-relatable reality of pet ownership.

Contradiction: The contradiction in this comic lies in the woman’s affectionate reaction to the dog’s kisses, completely unaware that just moments earlier, the same dog had been licking itself in a rather unsanitary manner.

LlaVA-Next-7B

Literal Description: In the first panel, a woman lovingly holds her small dog in her arms, smiling and saying, "Awwww! Who's a good boy?" The dog looks up at her with a content and innocent expression, reinforcing the warmth and affection between pet and owner. The second panel, labeled "5 minutes earlier!", delivers the punchline. The same dog is now standing on its hind legs with its front paws raised in the air, as if it was dancing or performing an unusual trick. This unexpected contrast adds humor by suggesting that the dog’s adorable appearance was preceded by a moment of goofy or possibly embarrassing behavior. The comic is drawn in a playful cartoon style with a limited color palette, emphasizing its lighthearted nature.

Contradiction: The contradiction in this comic lies in the shift between the perceived meaning of praise and the actual behavior of the dog moments earlier. In the first panel, the woman praises the dog, saying, "Aww! Who's a good boy?", which implies that the dog is well-behaved and deserving of affection. However, in the second panel, labeled "5 minutes earlier!", the dog is seen in an entirely different pose, raising its paw in a playful or possibly chaotic manner. This contrast humorously suggests that the dog’s behavior may not have been as "good" as the praise implies, creating a comedic disconnect between expectation and reality.

Human Reference

Literal Description: This comic humorously contrasts human perceptions of danger and wildlife conservation over time. Panel One (Before): In the top panel, two men are hiding behind a tree, cautiously observing a tiger walking by. They appear startled and fearful, reflecting the common perception of wild animals as dangerous threats.
Panel Two (Now): In the bottom panel, the narrative is flipped. Instead of being a feared predator, the tiger is now portrayed as vulnerable and endangered. This shift highlights the irony of how human attitudes have changed, emphasizing the real danger that wildlife faces today—extinction and habitat loss.

Contradiction: This contradiction serves as a sarcastic commentary on how society perceives wild animals as threats while, in reality, humans are the true danger to their existence. The comic cleverly highlights this irony, emphasizing the shift from fearing wildlife to recognizing the urgent need to protect it from extinction and habitat destruction.

GPT-4o

Literal Description: The image is a comic strip divided into three panels, each depicting a different scene under the labels "BEFORE", "NOW", and "TIGER IN DANGER". In the "BEFORE" panel, two men appear alarmed and are gesturing to each other to be quiet, as they sneak past a tree. Behind the tree, a cheetah is peeking out, seemingly unnoticed by the men. The "NOW" panel shows the roles reversed; two cheetahs are behind the same tree, appearing shocked and gesturing to each other to be quiet. A man walks past the tree, oblivious to the cheetahs. He looks confident and is carrying a camera and a bag, suggesting he might be a tourist or photographer. The "TIGER IN DANGER" panel shows the man from the "NOW" panel walking away cheerfully with his camera, completely unaware of his surroundings. The title suggests a commentary on wildlife endangerment or human encroachment into animal habitats.

Contradiction: In the first panel, two people are hiding from a tiger, indicating fear or caution. In the second panel, the roles are reversed, with the tigers hiding from a man carrying a gun, suggesting the tigers are now in danger.

Fig. 17: Sample outputs of model generated literal description and contradiction.

In contrast, models performed differently in the Chinese-context setting. Qwen2-VL-72B consistently achieved the highest accuracy in both deep reasoning tasks. This result reinforces the superior understanding and reasoning ability of the Qwen2 series in Chinese, further validating our previous findings on text-included comics and Qwen’s specialized performance in this context.

The language-independent setting yielded the most stable performance across all models. Since this setting primarily involves common scene humor, it does not require comprehension within a specific language. In this scenario, LLaVA-OneVision-72B achieved 81.23% in the symbolism selection task, while GPT-4o attained 81.56% in title matching.