Cracking the Code of Juxtaposition: Can AI Models Understand the Humorous Contradictions

Zhe Hu[1]*, **Tuo Liang**[2]*, **Jing Li**[1], **Yiren Lu**[2], **Yunlai Zhou**[2], **Yiran Qiao**[2], **Jing Ma**[2], **Yu Yin**[2]
[1]Department of Computing, The Hong Kong Polytechnic University
[2]Department of Computer and Data Sciences, Case Western Reserve University
[1]zhe-derek.hu@connect.polyu.hk, jing-amelia.li@polyu.edu.hk
[2]{txl859,yxl3538,yxz3057,yxq350,jxm1384,yu.yin}@case.edu
https://vulab-ai.github.io/YESBUT_Homepage/

Abstract

Recent advancements in large multimodal language models have demonstrated remarkable proficiency across a wide range of tasks. Yet, these models still struggle with understanding the nuances of human humor through juxtaposition, particularly when it involves nonlinear narratives that underpin many jokes and humor cues. This paper investigates this challenge by focusing on comics with contradictory narratives, where each comic consists of two panels that create a humorous contradiction. We introduce the YESBUT benchmark, which comprises tasks of varying difficulty aimed at assessing AI's capabilities in recognizing and interpreting these comics, ranging from literal content comprehension to deep narrative reasoning. Through extensive experimentation and analysis of recent commercial or open-sourced large (vision) language models, we assess their capability to comprehend the complex interplay of the narrative humor inherent in these comics. Our results show that even state-of-the-art models still lag behind human performance on this task. Our findings offer insights into the current limitations and potential improvements for AI in understanding human creative expressions.

1 Introduction

"The world is indeed comic, but the joke is on mankind."

—H. P. Lovecraft

Comics are a unique blend of visual art and narrative that encapsulate a wide range of human experiences and emotions. Understanding comics often requires significant social reasoning skills and cultural knowledge, as they heavily rely on context, cultural references, and visual metaphors. Furthermore, comics frequently employ nonlinear narratives [1, 2], demanding rigorous reasoning to grasp underlying ideas. Recent large (vision) language models have achieved impressive performance on various tasks [3–5], yet their ability to comprehend these complex human expressions remains insufficiently explored [6–8].

Examining AI models' ability to understand comics is essential for advancing their social and semantic comprehension. As a significant part of human creative expression, comic offers valuable insights into human emotions and cultural contexts [9]. This understanding is crucial for developing socially intelligent systems and enhancing AI-related creativity, thereby improving user experience in applications such as recommendation systems and automated content creation tools.

Previous studies have applied vision language models (VLMs) to understand humor and deep semantics [7, 10]. However, these studies often focus on single-panel comics and do not investigate

* These authors contribute equally.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

the more complex case of nonlinear narratives created through juxtaposition, a fundamental element in comics. Juxtaposition involves placing two contrasting elements together to provoke thought or evoke humor [11, 12]. This technique requires readers to pause and reassess the meaning, engaging in *nonlinear thinking* to reason about the relationships between panels for overall idea. [13–15].

In this work, we examine VLMs' ability to understand comics, specifically focusing on humor derived from juxtaposition. Our goal is to determine if large models can accurately comprehend the complex and contradictory narratives present in comics. Such contradictions challenge conventional semantic interpretations and demand deeper analysis. For example, Figure 1 shows a comic with two panels: in the first, a driver stops for ducks to cross the road ("Yes"), and in the second, the driver enters a "Peking Duck" restaurant ("But"), highlighting the contradiction in human-animal relationships through juxtaposition.

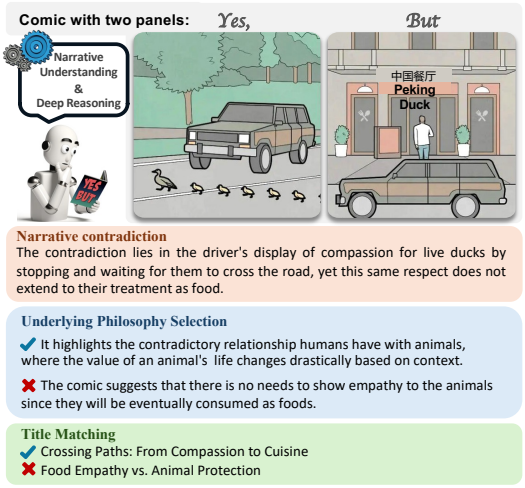


Figure 1: We introduce YESBUT dataset for comic understanding of juxtaposed comic panels. Given a two-panel comic with a contradictory narrative, we propose several tasks including narrative understanding, underlying philosophy selection and title matching, tackling different levels of comic understanding. (Comic by Anton Gudim).

Understanding such juxtaposition in comics significant challenges for the models. First, **it requires deep comprehension of human norms**, recognizing that people often have conflicting feelings, and identifying subtle social cues and contexts tied to cultural backgrounds. Additionally, **it demands nonlinear reasoning to grasp the overall narrative**, as the story is conveyed through the interplay of two panel elements, forming the core of the narrative beyond the literal meaning of each single panel. This type of juxtaposition necessitates critical thinking about similarities and differences, requiring in-depth reasoning. However, current models lack the ability to process information through nonlinear and deep thinking effectively, as the autoregressive paradigm of LMs limits their bidirectional reasoning capabilities [16–18]. By emphasizing these contradictions, we aim to push AI models to develop more sophisticated semantic understanding, enriching their interpretative capabilities.

To this end, we collected and annotated a new benchmark, YESBUT, for understanding comics with juxtaposition, focusing on *contradictory* narratives. Each comic is annotated with a literal description, a contradiction illustration, the underlying philosophy it reveals or satirizes, and a title that summarizes the overall narrative, as shown in Figure 1. We then propose four tasks: (1) literal description writing, to produce a surface description of the comic narrative; (2) contradiction generation, where the model illustrates the narrative contradiction; (3) underlying philosophy selection, which targets at selecting the correct philosophy the comic reflects; and (4) title matching, where the model matches the comic with a proper title. These tasks jointly cover different levels of comic understanding, from literal content comprehension to more in-depth narrative reasoning, providing a thorough evaluation of comic understanding capabilities.

We conducted comprehensive experiments on the YESBUT dataset, evaluating both commercial and open-sourced large (visual) language models. Both automatic and human evaluation results indicate that commercial VLMs outperform their open-sourced counterparts on most tasks. However, even the highest scores are far from perfect (e.g., 84.1% accuracy for underlying philosophy selection and 63.3% for title matching), underscoring the need for further advancements in this area. Additionally, our analysis reveals that augmenting models with oracle comic descriptions can significantly enhance performance, highlighting the considerable gap in current models' understanding of comic narratives. We release our annotations, code, and model results, aiming to provide valuable insights for future AI research on understanding human creative expression.

2 Related Work

Large Models and Evaluations. Recent large (vision) language models have demonstrated remarkable performance in following human instructions and performing various downstream tasks through zero-shot prompting [19–21, 4]. Various benchmarks have been proposed to evaluate their performance, encompassing both language-only tasks [22–25] and vision-language tasks [26–30]. These tasks primarily focus on assessing the fundamental capabilities of large models. However, the ability of large models to perform in-depth social reasoning and accurately understand human contexts remains underexplored [6].

Computational Humor. Humor is a vital component of human communication [31]. Our research is closely related to the computational understanding of humor. Previous studies have addressed humor recognition [32–34] and generation [35], with recent work expanding to multimodal data, such as visual humor prediction [36], humorous cartoon caption identification [37, 7, 38], and humor prediction in videos [39, 40]. Despite advancements, recent work shows that LLMs such as ChatGPT has not fully solved computational humor yet [41]. In this work, we design tasks to evaluate large vision-language models on their ability to understand humor through comic juxtapositions with contradictory narratives, requiring deep narrative comprehension. Through this study, we aim to provide insights into the capabilities of AI in processing and appreciating humor.

Interpretation of Human Creative Expressions. Visual artwork, encompassing mediums such as drawings, paintings, and sculptures, has been a profound aspect of human culture and cognition. These creative expressions are not merely decorative; they are deeply entwined with the ways humans perceive, interpret, and communicate their experiences and emotions [42]. Understanding these human creative expressions necessities valuable insights of human emotions, societal values, and cultural contexts, which is crucial for developing socially intelligent systems and enhancing AI-related creativity [43]. Previous research has explored AI interpretation of visual human creative expressions in tasks such as meme [44] and cartoon [45] understanding. Similar to our work, studies like [7] and [10] apply AI models to comprehend comics. However, these studies primarily focus on single-panel comics, emphasizing humor and deep semantics. In contrast, our work aims to investigate the significant feature of juxtaposition for understanding contradictory narratives.

Visual Reasoning. Our task is also related to the visual reasoning ability, where the model requires in-depth reasoning to comprehend contradictions between two comic panels. Previous research has examined visual reasoning capabilities of large models in tasks involving commonsense reasoning [46, 28, 47], visual question answering [48], visio-linguistic compositionality [49], and science question answering [50]. Unlike these studies, our task involves nonlinear reasoning, which necessitates AI to navigate multi-dimensional and complex information layers, often without explicit directives. While linear reasoning present their challenges, they usually exhibit clearer rules and structures, making them more accessible for AI to process with existing algorithms and models. Consequently, nonlinear reasoning represents a more intricate task, demanding higher natural language processing and cognitive modeling capabilities from AI systems.

3 The YESBUT Dataset

Our benchmark consists of YESBUT comics featuring contradictory narratives. Specifically, each sample includes: (1) a two-panel comic that forms a narrative with inherent contradictions; (2) a literal description of the comic narratives; (3) an explanation that illustrates the contradiction within the narrative; (4) the deep philosophy or underlying message the comic aims to convey; and (5) a title of the comic. Based on these components, we construct various tasks for comic understanding.

3.1 Image Collection

Our dataset consists of captionless comics, primarily from Anton Gudim's "YES, BUT" series [51], each featuring two panels depicting contradictory everyday scenarios. We scraped the images from social media [2] and conducted preprocessing, including deduplication, filtering out comics with more than two panels, and removing any inappropriate or offensive content. This process resulted in a final dataset of 348 comics.

[2] https://twitter.com and https://www.pinterest.com/

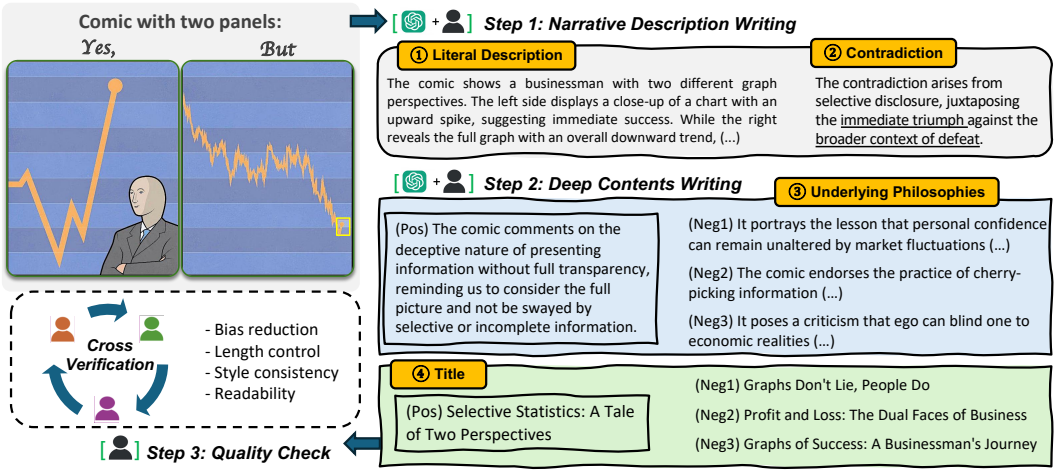


Figure 2: Overview of the data construction pipeline. Pos represents the positive options, and Neg stands for the negative options.

3.2 Data Annotation

For each comic, we annotate the corresponding literal description, contradiction explanation, underlying philosophy and comic title [7, 10]. We primarily rely on human annotators to obtain gold-standard annotations. Eight human judges participated in the annotation process, all of whom are proficient English speakers based in English-speaking countries and have at least a Bachelor's degree. Our annotation process included two stages: the progressive human-AI collaborative annotation stage and the quality check and cross-verification stage. The pipeline is illustrated in Figure 2.

Progressive Human-AI Collaborative Annotation. In this stage, we randomly assigned comic samples to each annotator, instructing them to first exclude any comics that may contain offensive, hateful, or sexual material before beginning the annotation process. To reduce human effort and costs associated with data annotation from scratch, we designed a human-AI collaboration pipeline utilizing GPT-4 [52] for data annotation and component writing.

The pipeline operates through dialogue interactions with the GPT-4 model. Given a comic image, we first prompt GPT-4 to generate narrative descriptions, illustrating the comic's narrative and explaining the contradictory logic between the two panels. Human annotators then modify and annotate the contents to obtain a literal description and contradiction explanation.

After obtaining the gold-standard description, both the comic and the description are used as input to prompt GPT-4 deep content writing, including the underlying philosophy and an eye-catching comic title. The underlying philosophy aims to foster a deep understanding of the comic and reveal the phenomenon it satirizes or the lesson it conveys; and the title is a more abstractive expression that reflects the overall narrative. Both components will be further checked by human annotators. Additionally, for the underlying philosophy and title understanding tasks, GPT-4 generates hard negative counterparts and distractions to design multiple-choice questions for our experiments. The prompts we used are provided in the Appendix A.

Our human-AI collaborative annotation pipeline is effective as it leverages a progressive prompting strategy, annotating each component from easy to difficult. Understanding the underlying philosophy of the comic, for example, requires first understanding the literal narratives and contradictions. This approach reduces annotation costs and improves overall efficiency.

Quality Check with Cross Verification. To ensure the quality and accuracy of the components and reduce objective bias from different human annotators, we introduced a cross verification stage. In this process, one annotator is assigned as an inspector for each comic. The inspector checks the annotated results to ensure all components are correct, unbiased, and appropriate. If any content is found to be of low quality or ambiguous, a third annotator is brought in as a judge to determine the final version. We exclude the comics with ambiguous or controversial narratives. This process ensures the quality of the annotated components for benchmark construction.

After the cross-verification process, one of the authors reviews each sample to verify the validity and appropriateness of the components. Finally, we obtain a dataset of 348 comics, each accompanied by high-quality components. The statistics of the components are shown in Table 1.

Components	#Num	Avg. Len.
Image	348	-
Literal Description	348	80
Contradiction	348	31
Philosophy	1,392	24
Title	1,392	6

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

Mitigating Annotation Bias. Our benchmark focuses on *common interpretation* of humor. However, the subjectivity of this task may introduce bias. To mitigate this issue, we have taken several steps in our annotation process: (1) Our annotators come from different genders and diverse cultural backgrounds, providing a range of perspectives; (2) Multiple quality checks and verifications are incorporated to ensure consensus among different annotators, with controversial or potentially biased comics being filtered out; (3) Annotations are further validated by cross-referencing social media comments for each comic to ensure alignment with widely accepted interpretations; (4) Recognizing that tasks such as generating titles and philosophical contents are inherently open-ended and involve subjective data annotation, we frame them as selection tasks, and ensure that the correct option is clearly and objectively superior than the negative options to mitigate subjectivity.

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

We aim to evaluate the capabilities of recent large (visual) language models in understanding humor through contradictions. This is challenging because it requires both social reasoning about human events and nonlinear logical reasoning about the narratives, going beyond the literal understanding of the comic. We design a series of tasks that require different levels of narrative understanding and reasoning abilities to evaluate the models' performance in reading comics.

1. Literal Description Writing. The first task is to generate the literal description of the comic narrative. We formulated this task as a text generation task: given an input comic, the model is required to generate a short description illustrating the narrative from the two panels of the comic. This task is different from the traditional image captioning, which requires the model to illustrate the comic narrative instead of solely focusing on image description.

2. Contradiction Generation evaluates whether the model can understand the contradiction within the narrative juxtaposition. Similarly, it is formulated as a text generation task.

3. Underlying Philosophy Selection. Understanding comics requires grasping not only the surface meaning of the images but also the underlying ideas the authors aim to convey. This task evaluates the model's ability to recognize the comic's underlying philosophy. It is formulated as a multiple-choice question answering (MCQ) task: given an input comic and four candidates of its underlying philosophy, the model must predict the correct option. The negative choices are crafted by annotators to be relevant to the comic, requiring reasoning to make the correct prediction.

4. Title Matching evaluates whether the models can identify the corresponding title, which is challenging because the title acts as an abstraction of the narrative and requires a proper understanding of the comic's content. Similar to the underlying philosophy task, it is formulated as a multiple-choice question answering task, where the most is asked to select the correct title from four options.

4 Experiments

4.1 Models and Settings

We evaluate the models' performance in a zero-shot manner using both recent VLMs and LLMs. For VLMs, the comic image and questions are provided as inputs for output prediction. We include both commercial models such as GPT-4 [52] and Claude-3[53], as well as open-sourced models including LLaVa [3, 54], CogVLM [55], Qwen-VL [56], mPLUG-Owl2 [57], and InstructBLIP [58].

For LLMs, since they cannot directly process images, we use the LLaVa-1.6 13B model generated literal descriptions as inputs due to its strong performance. We include ChatGPT, the Llama3 instruction model [20], and the Mistral 7B instruction model [59]. More details of the models are included in Appendix B.

Setting	Model	Literal Description			Contradiction			Philosophy	Title
		BERT	R-2	GPT	BERT	R-2	GPT	Accuracy	Accuracy
VLMs	GPT-4	**88.32**	**87.46**	**3.76**	87.64	**83.21**	**4.03**	82.76	60.25
	Claude-3	87.68	80.30	3.28	86.93	80.63	3.79	**84.10**	56.42
	LLaVA-1.6-34B	86.45	67.67	2.86	86.04	75.95	3.51	78.83	**63.31**
	LLaVA-1.6-13B	81.34	75.95	2.96	86.48	80.96	3.36	69.16	55.08
	LLaVA-1.5-13B	78.77	58.21	2.51	86.48	67.67	3.36	69.73	48.75
	InstructBlip-13B	85.20	35.28	2.69	85.54	51.15	2.54	30.75	22.70
	CogVLM	80.80	55.51	2.65	87.07	69.96	3.76	61.30	49.52
	Qwen-VL-Chat	79.03	51.58	2.76	86.41	59.77	3.25	59.10	42.05
	mPlug-Owl2	78.26	47.38	2.57	86.20	48.05	2.59	62.17	43.10
	LLaVA-1.6-7B	80.71	70.36	2.79	86.58	75.36	3.24	47.41	37.07
	InstructBlip-7B	76.02	38.02	2.60	86.32	66.29	2.85	25.86	26.44
LLMs	ChatGPT	-	-	-	**87.78**	67.42	3.54	75.86	49.52
	Llama-3-8B-Instruct	-	-	-	87.41	70.52	3.59	72.13	49.71
	Mistral-7B-Instruct	-	-	-	87.01	67.70	3.64	66.00	45.98

Table 2: Main results. For literal description and contradiction, we report BERT score (recall), BLEURT (BLT), and GPT evaluation score. For philosophy and title, we report accuracy (%). Best scores are **bold** and the second best ones are marked with underline.

Implementation Details. For GPT-4 and ChatGPT, we set the temperature as 1. For other models, we use the default parameter settings during inference. To reduce variance across different task prompts, we create three distinct prompts for each task and report the average scores from three runs with each prompt. The specific prompts and additional details are in Appendix B.

4.2 Evaluation Metrics

For the philosophy and title understanding tasks, which are formulated as multiple-choice question answering, we use accuracy as the evaluation metric. For generation tasks including literal description and contradiction, we apply reference-based evaluation metrics commonly used in text generation studies [60], and report ROUGE-2 (recall) [61] and BERT Score (recall) [62]. [3] Recent work shows GPT-based evaluation aligns well with human judgements [63–65], and we also apply ChatGPT for evaluation [4]. The prompts for GPT-based evaluation are provided in Appendix B.

Due to the limitation of the automatic evaluations for text generation, we also include human evaluation to assess the quality of the outputs for the literal description and contradiction generation tasks. We hire three human judges to rate each aspect on a scale of 1 (worst) to 5 (best). For literal description, following [44], we evaluate: (1) **Correctness**: Does the model output correctly convey the narrative of the comic? (2) **Completeness**: Does the model output cover all the important elements of the comic narrative? (3) **Faithfulness**: Can all contents from the model output be supported by the comic image (i.e., there are no hallucinations)? For contradiction generation, we evaluate **Correctness** and **Faithfulness**. More details are provided in Appendix B.4.

5 Main Results

The main experimental results are shown in Table 2. For VLMs, the original image is directly used as input, while for LLMs, the generated comic description is used as input.

[3] We report recall scores considering the open-ended nature of the outputs, as there can be multiple valid expressions. Our focus is on evaluating whether the key points are covered by the model outputs, ensuring a more precise assessment of content coverage.

[4] We utilize different GPT variants for specific purposes: *gpt4-turbo* for data annotation, *gpt-4-vision-preview* for experiments, and *gpt-3.5-turbo-0125* for GPT-based evaluation in text generation tasks. This helps reduce potential evaluation bias toward GPT-4's own generation. Further details are in Section B.

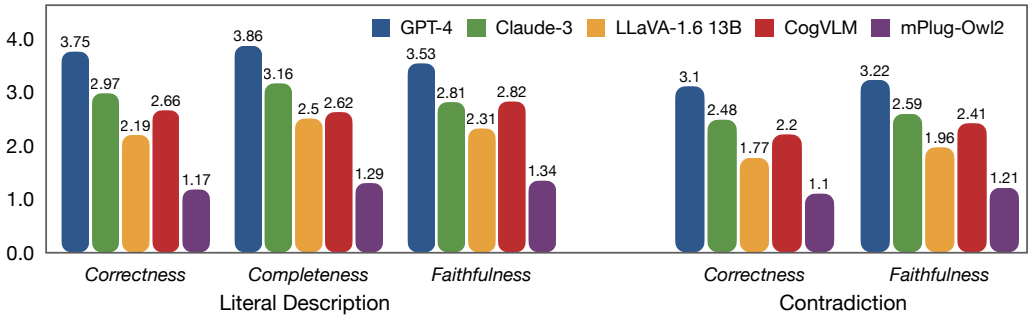


Figure 3: Human Evaluation on literal description and contradiction generation.

5.1 Narrative Understanding Tasks

Literal Description: We evaluate the results of VLMs only for this task. We observe that the two commercial models generally outperform the smaller open-sourced models. Among these models, GPT-4 achieves the highest scores. For the open-sourced models, the larger model variants (13B) consistently achieve better scores than their 7B counterparts, indicating that larger models have a superior ability to understand the image and produce higher-quality literal descriptions.

Contradiction Generation: A similar trend is observed where GPT-4 and Claude-3 achieve better results than other VLM models. Notably, LLaVA-1.6 variants outperform their counterparts in generating contradiction descriptions. This is likely due to their improved reasoning ability and world knowledge [54], which are essential for understanding comic narratives and accurately capturing the relationship between the two panels. For LLMs, unlike VLMs, the Llama-3 and Mistral models achieve results comparable to ChatGPT. Another interesting observation is that Llama-3 and Mistral obtain similar or better results for contradiction generation compared to open-sourced VLMs, despite not having access to the original comic images.

5.2 Deep Reasoning Tasks

The Underlying Philosophy Selection and Title Matching tasks require in-depth reasoning based on the comic narratives. As seen in Table 2, for philosophy selection, Claude-3 achieves the best accuracy with 84.10%, while for title matching, the LLaVA-1.6 34B variant ranks the highest with 63.31% accuracy. One key observation is that larger models usually perform better in-depth understanding of the comics, aligning with the findings that larger models typically exhibit superior reasoning abilities [66, 67].

Additionally, LLMs achieve performance comparable to open-sourced VL models. This can be attributed to the strong reasoning abilities of models like Llama-3 and Mistral [59, 20], which are crucial for understanding narratives and performing nonlinear reasoning to grasp deep semantics. Further analysis on the influence of descriptions for LLMs is provided in Section 6.1.

Another observation is that model performance on title matching is consistently lower than on underlying philosophy selection. Titles are shorter and more abstract versions of the narrative and do not explicitly convey the underlying idea of the comic. Therefore, distinguishing the correct title from distractions requires a deeper rigorous understanding and reasoning abilities, making it more challenging for models. Notably, the human evaluation results show a similar trend of our proposed GPT-based evaluation, demonstraining its effectiveness.

5.3 Human Evaluations

We conduct human evaluations on 30 randomly selected samples to assess the output quality of literal descriptions and contradiction generation, as shown in Figure 3. Similar trends are observed in both human and automatic evaluations: commercial models generally outperform open-source models in producing both literal descriptions and contradictions, with GPT-4 achieving the highest scores in both tasks. Additionally, the scores for literal descriptions are consistently higher than those for contradictions across all models, suggesting that understanding narrative contradictions is

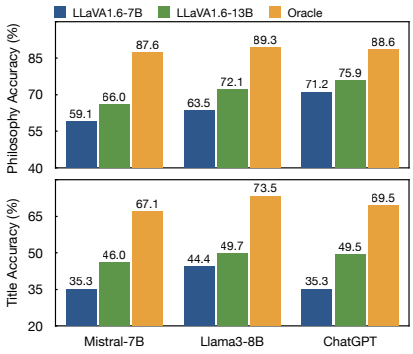


Figure 4: LLMs using different image description as input.

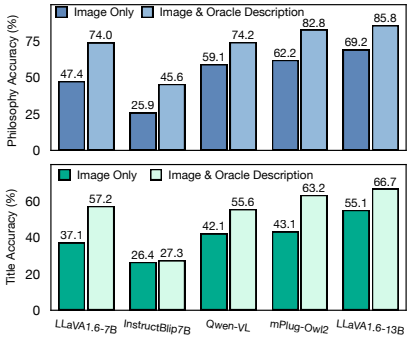


Figure 5: VLMs with image only input and image + oracle description as inputs.

more challenging than generating literal descriptions, which requires in-depth reasoning to compare the various aspects of both panels.

A comparison of the scores for literal description and contradiction reveals a strong correlation between the two tasks: models that perform well on literal descriptions also tend to achieve good results on contradictions. This indicates that understanding comic juxtaposition requires a diverse set of skills, including image understanding, narrative comprehension, and reasoning abilities.

6 Analysis and Discussion

6.1 How Does Literal Understanding of the Comic Influence Deep Reasoning?

We investigate whether the quality of surface-level literal descriptions influences subsequent deep reasoning tasks. For LLMs, we provide different literal descriptions generated by LLaVA-1.6 7B and 13B variants, as well as oracle descriptions written by humans, as model inputs. The results are shown in Figure 4. As the quality of literal descriptions improves, the prediction accuracy for both underlying philosophy and title selection also improves. This demonstrates a strong correlation between deep reasoning and literal narrative understanding. However, a significant performance gap remains compared to when oracle descriptions are used.

We further examine the performance of VLMs by providing them with additional oracle descriptions. The results are shown in Figure 5. Compared to using only the comic image as input, augmenting with human-written literal descriptions significantly improves the deep reasoning results for all VLMs. This confirms that correctly reasoning about the underlying semantics of a comic requires first accurately understanding its surface narrative. However, the performance gap indicates that current VLMs still lag in narrative understanding.

Additionally, an interesting observation from Figures 4 and 5 is that when the oracle literal description is provided as (partial) input, LLMs tend to outperform their VLM counterparts in both philosophy and title selection. For example, LLaVA-1.6-7B employs Mistral-7B as the language model backbone, yet its performance under oracle description is significantly worse than that of Mistral-7B . One possible reason is that incorporating oracle descriptions makes VLM input much longer, thus making prediction more challenging. We provide further discussions in Section6.2.

6.2 Is Decomposing Literal Description Helpful for Deep Reasoning of VLMs?

VLMs typically predict results in an end-to-end fashion, requiring the model to perform image captioning, narrative understanding, and deep reasoning all at once. Here, we investigate whether decomposing the task into separate stages of narrative understanding and in-depth reasoning can improve model performance. Specifically, we first prompt the VLM to produce a literal description of a comic; then the VLM predicts results based on both the comic image and the description. The results are shown in Table 3.

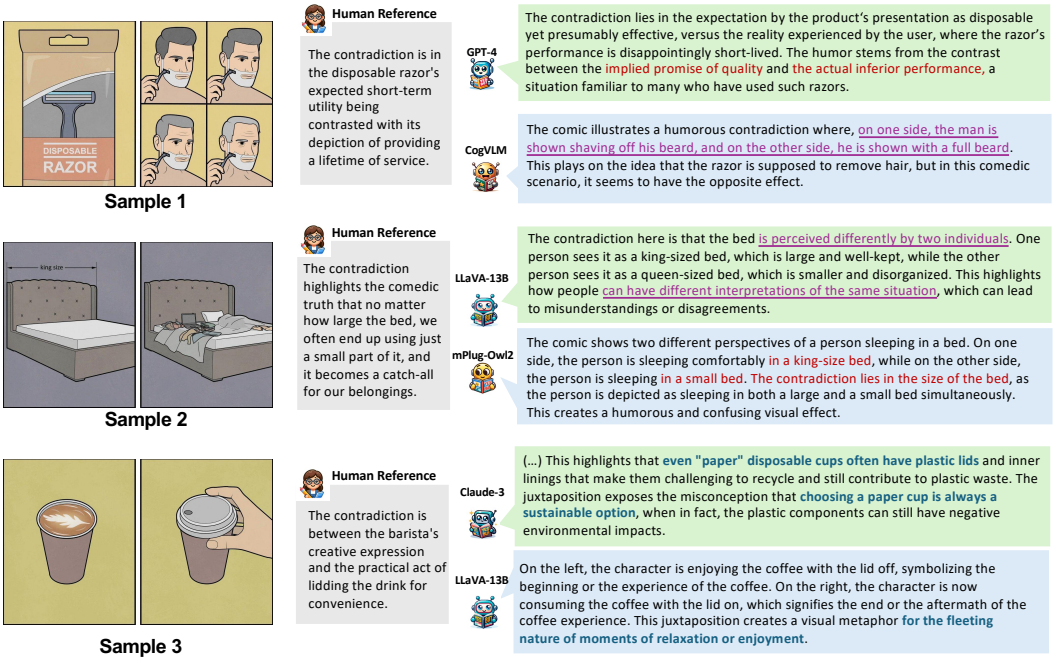


Figure 6: Sample outputs of contradiction explanations generated by different vision language models, along with human written references. We highlight different types of errors in model outputs.

As observed, decomposing the task and augmenting it with a literal description does not necessarily improve performance. In fact, when descriptions are incorporated, performance across all models declines on the title selection task, which contrasts with previous findings [10]. One possible explanation for this drop in performance is that the generated descriptions may contain errors, negatively impacting the model's deep understanding. Another explanation could be the length of the generated descriptions (e.g., the LLaVA-1.6 13B model's descriptions average around 170 words), leading to longer and more complex prompts that make prediction more challenging. We leave a more detailed investigation of this issue for future work.

Models	Philosophy	Title
LLaVA-1.6-13B	69.16	55.08
↪ w/ desp.	68.68	48.76
Qwen-VL-Chat	59.10	42.05
↪ w/ desp.	59.58	37.55
mPlug-Owl2	62.17	43.10
↪ w/ desp.	60.25	37.84
LLaVA-1.6-7B	47.41	37.07
↪ w/ desp.	53.07	34.96

Table 3: Decomposition model results augmenting the predicted description.

6.3 Error Analysis and Future Directions

We present sample outputs of contradictions generated by vision language models (VLMs) in Figure 6. VLMs can make various errors in contradiction understanding.

One type of error is **visual misinterpretation**, where the model incorrectly interprets the image contents. For example, in sample 1, CogVLM misinterprets the image by recognizing a person "shown with a full beard." Similarly, in sample 2, LLaVA-1.6 13B misunderstands the image contents and generates incorrect content about "two individuals," which is inconsistent with the comic. Such misinterpretations can lead to incorrect understanding of the narrative. These observations align with our previous findings in Section 6.1 and Section 6.2. *This highlights the need for future research to improve models' visual interpretation capabilities.*

Models also struggle to conduct **in-depth reasoning of the relationship** between two panels by recognizing their differences and similarities. In sample 1, while the comic implies a comparison between the expected disposable razor and its actual longevity, GPT-4 incorrectly explains the contradiction as being about the razor's quality. A similar error occurs with mPlug-Owl2 in sample 2, where it incorrectly thinks the bed sizes are different in the two panels, leading to a wrong illustration focusing on the bed size. *Future work might incorporate recent advanced reasoning*

approaches (e.g., multi-agent debate [68], test-time compute scaling [69]) to further improve model performance.

Another common error is **hallucination and incorrect association**. This is evident in sample 3. The original comic contrasts latte art before and after lidding the drink, but Claude-3 incorrectly associates the narrative with environmental protection, focusing on the plastic lid. Meanwhile, LLaVA-1.6 13B model suffers from hallucinations by interpreting the narrative as being about relaxation and enjoyment, which is unsupported by the original comic. *This suggests the need for improving world knowledge and social understanding abilities to enhance model performance on this task.* More sample outputs are in Appendix C.

7 Conclusion

In this work, we present YESBUT, the first benchmark dedicated to studying comic understanding through juxtaposition. YESBUT encompasses a variety of tasks that address both narrative comprehension and deep reasoning. The results indicate that state-of-the-art vision and language models still struggle with these tasks. We also offer a comprehensive analysis and discussion of errors to evaluate model performance. Current models still struggle to accurately interpret the visual contents and conduct in-depth reasoning of the underlying narratives. Through this study, we aim to provide insights for future research and advance the capabilities of AI models in understanding human context, ultimately contributing to more effective and culturally aware AI applications.

8 Limitations

We propose a comprehensive data annotation process to annotate each component. However, due to the subjectivity of comic interpretation, especially regarding the underlying ideas, there might be potential ambiguity. While we acknowledge the relatively small size of images, we rigorously collect comics and annotate each component, ensuring their high-quality and reliability. We plan to expand the dataset with the inclusion of different types of narratives in future work.

Our proposed benchmark focuses predominantly on recognizing and interpreting visual humor via juxtaposition, and may not cover all aspects of visual understanding required for more generalized AI applications. In the future, we intend to explore more deeply how AI can not only interpret but also creatively engage with content. This includes generating pivotal turning points from one perspective and creating counterpoints to given scenarios, like generating a "YES" image's counterpart.

9 Ethics Statement

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.

The Large Vision Language Models utilized in our experiments are pretrained using diverse web corpora, which may introduce biases in their outputs. We advise users to conscientiously evaluate the ethical implications of generated outputs when employing them in future research endeavors.

Data Annotation. Eight human judges are engaged in our annotation process. We compensate these judges with an average hourly wage of $11, ensuring fair remuneration for their contributions.

Acknowledgements

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] Jessica Pressman. *Digital modernism: Making it new in new media*. Oxford University Press, USA, 2014.

[2] Alan D Manning. Understanding comics: The invisible art. 1998.

[3] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.

[4] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023.

[5] Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and S Yu Philip. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2247–2256. IEEE, 2023.

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

[7] Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 688–714, Toronto, Canada, July 2023. Association for Computational Linguistics.

[8] Abu Rayhan, Rajan Rayhan, and Swajan Rayhan. Artificial general intelligence: Roadmap to achieving human-level capabilities, 2023.

[9] Randy Duncan and Matthew J Smith. *The power of comics: History, form and culture*. A&C Black, 2009.

[10] Yixin Yang, Zheng Li, Qingxiu Dong, Heming Xia, and Zhifang Sui. Can large multimodal models uncover deep semantics behind images? *arXiv preprint arXiv:2402.11281*, 2024.

[11] James O Young. *Art and knowledge*. Routledge, 2003.

[12] Thierry Groensteen. *Comics and narration*. Univ. Press of Mississippi, 2013.

[13] Eve Bearne. Rethinking literacy: Communication, representation and text. *Reading*, 37(3):98–103, 2003.

[14] Jason Dittmer. Comic book visualities: a methodological manifesto on geography, montage and narration. *Transactions of the Institute of British Geographers*, 35(2):222–236, 2010.

[15] Joshua Schechter. Juxtaposition: A new way to combine logics. *The Review of Symbolic Logic*, 4(4):560–606, 2011.

[16] Paul J Kuttner, Marcus B Weaver-Hightower, and Nick Sousanis. Comics-based research: The affordances of comics for research across disciplines. *Qualitative Research*, 21(2):195–214, 2021.

[17] Yongqi Tong, Yifan Wang, Dawei Li, Sizhe Wang, Zi Lin, Simeng Han, and Jingbo Shang. Eliminating reasoning via inferring with planning: A new framework to guide llms' non-linear thinking. *arXiv preprint arXiv:2310.12342*, 2023.

[18] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[19] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[20] AI@Meta. Llama 3 model card. 2024.

[21] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024.

[22] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2024.

[23] Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36, 2024.

[24] Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, et al. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. *arXiv preprint arXiv:2306.05087*, 2023.

[25] Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, et al. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36, 2024.

[26] Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, et al. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006*, 2024.

[27] Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schimdt. Visit-bench: A benchmark for vision-language instruction following inspired by real-world use. *arXiv preprint arXiv:2308.06595*, 2023.

[28] Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, and Roy 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*, pages 2616–2627, 2023.

[29] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.

[30] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench-2: Benchmarking multimodal large language models. *arXiv preprint arXiv:2311.17092*, 2023.

[31] Jerry Palmer. *Taking humour seriously*. Routledge, 2003.

[32] Lei Chen and Chong MIn Lee. Predicting audience's laughter using convolutional neural network. *arXiv preprint arXiv:1702.02584*, 2017.

[33] Andrew Cattle and Xiaojuan Ma. Recognizing humour using word associations and humour anchor extraction. In *Proceedings of the 27th international conference on computational linguistics*, pages 1849–1858, 2018.

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

[35] Miriam Amin and Manuel Burghardt. A survey on approaches to computational humor generation. In Stefania DeGaetano, Anna Kazantseva, Nils Reiter, and Stan Szpakowicz, editors, *Proceedings of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 29–41, Online, December 2020. International Committee on Computational Linguistics.

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

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

[38] Dragomir Radev, Amanda Stent, Joel Tetreault, Aasish Pappu, Aikaterini Iliakopoulou, Agustin Chanfreau, Paloma de Juan, Jordi Vallmitjana, Alejandro Jaimes, Rahul Jha, et al. Humor in collective discourse: Unsupervised funniness detection in the new yorker cartoon caption contest. *arXiv preprint arXiv:1506.08126*, 2015.

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

[40] Yang Liu, Tongfei Shen, Dong Zhang, Qingying Sun, Shoushan Li, and Guodong Zhou. Comment-aided video-language alignment via contrastive pre-training for short-form video humor detection. *arXiv preprint arXiv:2402.09055*, 2024.

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

[42] Nicola De Pisapia, Francesca Bacci, Danielle Parrott, and David Melcher. Brain networks for visual creativity: a functional connectivity study of planning a visual artwork. *Scientific reports*, 6(1):39185, 2016.

[43] Mika Koivisto and Simone Grassini. Best humans still outperform artificial intelligence in a creative divergent thinking task. *Scientific reports*, 13(1):13601, 2023.

[44] EunJeong Hwang and Vered Shwartz. MemeCap: A dataset for captioning and interpreting memes. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1433–1445, Singapore, December 2023. Association for Computational Linguistics.

[45] Dragomir Radev, Amanda Stent, Joel Tetreault, Aasish Pappu, Aikaterini Iliakopoulou, Agustin Chanfreau, Paloma de Juan, Jordi Vallmitjana, Alejandro Jaimes, Rahul Jha, and Robert Mankoff. Humor in collective discourse: Unsupervised funniness detection in the new yorker cartoon caption contest. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 475–479, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).

[46] Yuqing Wang and Yun Zhao. Gemini in reasoning: Unveiling commonsense in multimodal large language models. *arXiv preprint arXiv:2312.17661*, 2023.

[47] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731, 2019.

[48] Drew A Hudson and Christopher 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*, pages 6700–6709, 2019.

[49] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022.

[50] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

[51] "yes, but" series created by anton gudim. https://twitter.com/_yesbut_. Accessed: 2024.

[52] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[53] AI Anthropic. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 2024.

[54] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024.

[55] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023.

[56] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

[57] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration, 2023.

[58] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024.

[59] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[60] Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*, 2020.

[61] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[62] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020.

[63] David Chan, Suzanne Petryk, Joseph Gonzalez, Trevor Darrell, and John Canny. CLAIR: Evaluating image captions with large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13638–13646, Singapore, December 2023. Association for Computational Linguistics.

[64] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore, December 2023. Association for Computational Linguistics.

[65] Zhe Hu, Hou Pong Chan, and Yu Yin. Americano: Argument generation with discourse-driven decomposition and agent interaction. *arXiv preprint arXiv:2310.20352*, 2023.

[66] Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. Reasoning with language model prompting: A survey. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5368–5393, Toronto, Canada, July 2023. Association for Computational Linguistics.

[67] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

[68] Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023.

[69] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

A Data Annotation Details

Considering the workload of manually writing all components from scratch, we leverage a AI-human collaborative pipeline for annotation. The prompts for generating each component are listed in Table 4. After producing each component, human annotators will verify and modify the outputs for the final components. We present a sample comic with all tasks in Figure 7.

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.*
Underlying Philosophy	*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 Philosophy	*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.*

Table 4: Prompts for data annotation

Literal Description:

The comic is divided into two panels, each presenting a contradictory perspective of the same object-a mug. In the first panel, the mug is illustrated as an adorable fox with closed eyes, giving off a serene and cute vibe. It's an object that one would admire or find endearing. However, the second panel reveals a practical issue: a person attempts to drink from the fox-shaped mug, but its design-featuring protruding ears and head- awkwardly interferes, complicating the act of sipping comfortably.

Contradiction:
The comic illustrates a contradiction where a mug designed as an adorable fox is charming to look at but proves impractical to use due to its awkwardly protruding ears and head that hinder drinking.

Underlying Philosophy Selection:
A. The illustration critiques the clash between aesthetics and usability, emphasizing the need for a balanced consideration of both to ensure a harmonious and practical experience in any aspect of life.
B. The comic shows that adding more decorative elements to an object will enhance its value and enjoyment, when in fact, the opposite is true in this case.
C. The illustration suggests that the initial charming appearance of an item will always lead to a positive overall experience, disregarding any practical complications that arise later.
D. The image shows enduring inconvenience is a worthwhile sacrifice for the sake of owning something that looks unique or cute.
Answer: A

Title Matching:
A. Charming Design, Prickly Reality: The Fox Mug's Surprise
B. A Toast to Vulpine Grace
C. Harmony in a Sip
D. Enchanting Elixir: The Fox's Secret Brew
Answer: A

Figure 7: Sample comic with all annotated tasks.

Analysis on Data Diversity. In order to show the diversity of our benchmark, we prompt ChatGPT to generate topical keywords for each comic based on its description, and then cluster these keywords. All these scenarios are presented in Figure 8. As we can see, the comics in our benchmark encompass a diverse range of everyday life scenarios.

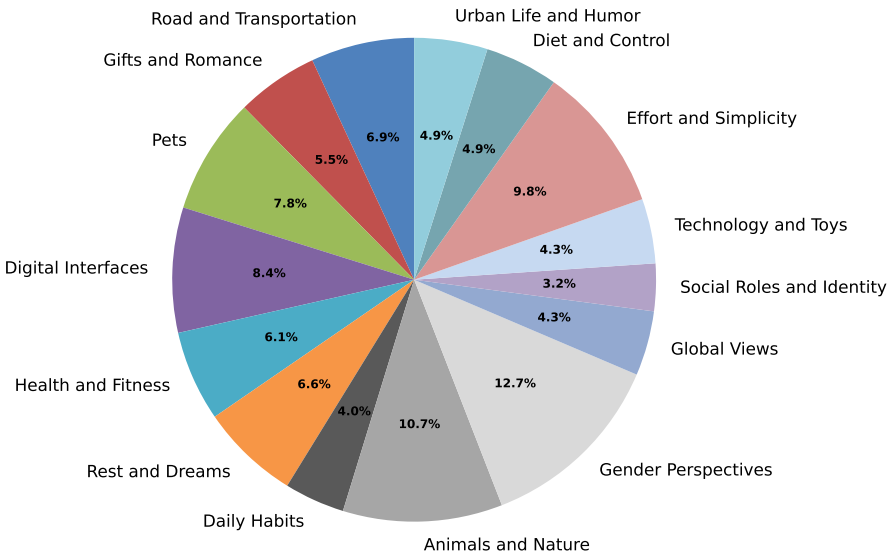


Figure 8: The clusters of comic topics covered by our benchmark.

B Experimental Details

B.1 Model Details

We include both commercial and open-sourced VLMs and LLMs in our experiments. For GPT-4, we use *gpt-4-vision-preview* version, and for ChatGPT we employ the *gpt-3.5-turbo-0125* model variant [5]. For Claude-3, we leverage the Claude 3 Ops model updated on 29th Feb, 2024 [6]. For open-sourced models, we include LLaVa1.6 (34B, 13B and 7B) [54], LLaVa1.5 13B [3], InstructBlip 13B and 7B variants [58], CogVLM [55], Qwen-VL [56], and mPLUG-Owl2 [57]. For LLMs, we use the Llama3 instruction variant [20], and the Mistral 7B instruction model [59].

B.2 Implementation Details

All commercial models are accessed through their official API. For open-sourced models, we implement the experiments using Hugging Face Transformers [7]. For GPT-4, Claude-3 and ChatGPT, we setting temperature as 1.0. For other models, we apply the default parameter setting or greedy decoding during inference. The experiments are conducted on NVIDIA 4090 and A6000 GPUs.

For MCQ evaluation, we explicitly instruct the model to directly output the option in prompts, and use hard rules to parse the answer. If none of the options can be parsed, we will assign it a random option. For generation task evaluation, we apply rouge-score [8] to compute ROUGE score, and calculate the BERT score using the official implementation [9]. For GPT based evaluations for literal description and contradiction, we use gpt-3.5-turbo-0125 version. The prompts we used are shown in Figure 9.

B.3 Experiment Prompts

To reduce the biases from different prompts, we design three different prompts by different people to evaluate models and report the average results for all tasks. We present the prompts used for Literal Description (Figure 13), Contradiction Generation (Figure 14), Underlying Philosophy Selection (Figure 15), and Title Matching (Figure 16).

[5] https://platform.openai.com/docs/models/
[6] https://www.anthropic.com/api
[7] https://huggingface.co/docs/transformers/en/index
[8] https://pypi.org/project/rouge-score/
[9] https://github.com/Tiiiger/bert_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]].

Figure 9: Prompts for GPT based evaluations.

B.4 Human Evaluation Details

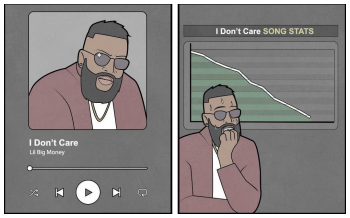
We present 30 random samples on each task for human evaluation. We anonymize the models and shuffle the outputs to the annotators. Following [44], we include the following aspects:

- **Correctness**: Does the model output correctly convey the narrative of the comic?
- **Completeness:** Does the model output cover all the important elements of the comic narrative?
- **Faithfulness:** Can all contents from the model output be supported by the comic image (i.e., there are no hallucinations)?

For Literal description, we evaluate on all three aspects. For contradiction, we evaluate on Correctness and Faithfulness.

C More Sample Outputs

Here, we present more randomly picked sample outputs on literal description and contradiction generation in Figure 10, Figure 11, and Figure 12.



Human Reference

Literal Description:
The left panel shows a music player interface with a song titled "I Don't Care" by Lil Big Money. The player has the typical features of a music app with play, pause, next, and previous buttons, and a progress bar that shows the song being played. The right panel reveals the musician, presumably Lil Big Money, looking stressed and worried as he observes the song stats on a chart that is plummeting. The chart is labeled "I Don't Care SONG STATS," and the descending line suggests that the song is not performing well.

Contradiction:
The contradiction is that while the song and the artist's persona radiate a carefree attitude, the reality is quite the opposite. Lil Big Money does care deeply about the performance of his song, as shown by the anxious demeanor and the dismal statistics on the graph.

Claude-3 Generated Description
The comic depicts a man with facial hair and sunglasses in two contrasting panels. In the first panel, titled "I Don't Care" with the subtitle "Lil Big Money", the man appears disinterested while music controls are shown below. In the second panel, titled "I Don't Care SONG STATS", a graph shows a sharp decline, implying the song's popularity is quickly fading. The juxtaposition suggests that despite the man's supposedly carefree attitude portrayed in the first panel, he is actually quite concerned about the performance and reception of his music as evidenced by the second panel showing the song's diminishing stats.

Claude-3 Generated Contradiction
The comic strip illustrates a satirical contradiction between the music artist's public persona and private attitude. In the first panel, the artist promotes their new song "I Don't Care" with a confident, indifferent expression. However, the second panel shows song stats plummeting, causing the artist to appear distressed and concerned, directly contradicting the song's "I Don't Care" message. This juxtaposition humorously highlights the discrepancy between the artist's projected apathetic image and their genuine emotional investment in the song's success or failure. The comic cleverly exposes how public figures may present one facade while privately experiencing the opposite sentiment.

CogVLM Generated Description
The comic portrays a man who is initially indifferent to a song titled 'I Don't Care' by Lil Big Money. However, upon seeing the song's statistics, he becomes deeply engrossed and contemplates its popularity, even though he initially claimed not to care about it.

CogVLM Generated Contradiction
The comic illustrates a humorous contradiction where the character claims not to care about song statistics, yet is deeply engrossed in observing them. On the left side, the character is shown confidently listening to a song, while on the right, he is seen contemplating the declining song statistics, indicating a sudden change in his attitude.

Figure 10: Sample outputs of model generated literal description and contradiction.

Human Reference

Literal Description:
This comic portrays the humorous contrast between expectations and reality within an artist's studio. In the first panel, the scene is set in a serene art studio, filled with classic art supplies: sculptures, a vase, brushes, and paints—all indicative of traditional fine arts. The empty canvas on the easel awaits the artist's touch. However, the second panel reveals the punchline: the artist has chosen to create a piece of anime or manga-style art, which stands in stark contrast to the classical environment and materials surrounding them.

Contradiction:
The contradiction of the comic lies in a traditional art setting prepared for classical art creation and the artist's unexpected choice to draw contemporary anime-style art instead.

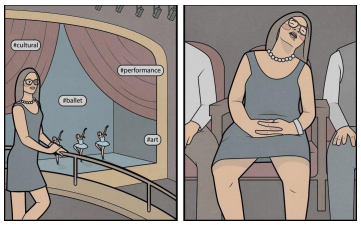
LLaVA-1.6 13B Generated Description
The comic depicts a scene of an artist's studio, with two panels showing the same space from opposite perspectives. In the left panel, the artist is seen standing in front of an easel with a blank canvas, holding a paintbrush, and surrounded by various art supplies. The studio is well-lit and appears to be a professional workspace. In the right panel, the same scene is shown from the opposite side, but the artist is now sitting at the easel, holding a smartphone, and the studio is now dimly lit, suggesting a different time of day or a change in the artist's mood or activity. The contrast between the two panels highlights the versatility of the artist's workspace and the different states of the artist's creative process.

LLaVA-1.6 13B Generated Contradiction
(...) The contradiction lies in the fact that the two sides represent two different stages of the creative process: one side is the beginning, where ideas are yet to be formed, and the other side is the end, where the artist has completed their work and is likely to be evaluating or contemplating the final product. This juxtaposition highlights the contrast between the initial stages of creation and the final stages of reflection and completion.

mPlug-Owl2 Generated Description
The comic depicts a man painting a picture of a woman in a white dress. However, the image is shown from two different perspectives, one where the woman is wearing a black dress and the other where she is wearing a white dress. This creates a contradiction in the visual representation of the woman's attire, making the comic visually intriguing and thought-provoking.

mPlug-Owl2 Generated Contradiction
In the comic, there is a man painting a picture of a woman in a studio. On one side, the man is painting a woman with a smile, while on the other side, the man is painting a woman with a frown. This contradiction highlights the difference in the emotions or expressions the artist is trying to convey in the two paintings.

Figure 11: Sample outputs of model generated literal description and contradiction.



Human Reference

Literal Description:
The first panel depicts a woman seemingly engrossed in a ballet, an epitome of high culture, poised to share her experience on social media with tags indicating her cultured taste and appreciation for the arts. In contrast, the second panel strips away the pretense to reveal the woman fast asleep during the performance.

Contradiction:
The contradiction lies between the cultured persona she broadcasts online and her true disinterest.

GPT-4 Generated Description
The comic is split into two panels, presenting a sequence of events with a humorous twist. In the first panel, a woman is seen standing in what appears to be an upscale theater, smiling and holding binoculars as she looks at the stage where a ballet performance is taking place. The ambiance is depicted as refined and high-culture, emphasized by hashtags such as "#cultural," "#performance," "#ballet," and "#art" scattered around the image, suggesting a sense of sophistication and appreciation for the fine arts. The second panel switches perspective to show the same woman sitting among an audience, seemingly relaxed with her head tilted back and eyes closed. The elegance implied in the first frame is juxtaposed with a comical hashtag "#fart," indicating that the woman is experiencing a less-than-dignified moment, thereby creating a comic disparity between the perception of cultured grace and the reality of an awkward bodily function.

GPT-4 Generated Contradiction
(...) This contradiction highlights the often humorous disparity between our idealized public image and our private, less glamorous realities. The comic plays on the juxtaposition of the refined setting of a ballet performance with the mundane and natural bodily function of passing gas, which stands in stark contrast to the cultural and artistic hashtags from the first image.

Claude-3 Generated Description
The comic strip presents a contradictory perspective on the performing arts. On the left side, a ballet performance is categorized under "cultural" and "art", with elegant ballerinas gracefully dancing on stage. In contrast, the right side depicts a woman slouching in a chair, suggesting a less refined or sophisticated setting. The hashtags imply this may represent some form of "performance", but the woman's casual, unrefined posture and appearance stand in stark juxtaposition to the poise and refinement of the ballet dancers. The comic highlights the subjectivity in labeling and perceiving different forms of performance and art, calling into question what truly constitutes "culture".

Claude-3 Generated Contradiction
(...) The juxtaposition highlights how the same artistic expression can be seen as either high-brow, cultured entertainment or as dull and unremarkable, depending on the viewer's perspective and appreciation for the medium. It points out the subjectivity in valuing and interpreting various art forms.

Figure 12: Sample outputs of model generated literal description and contradiction.

Prompt1:

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.

Prompt2:

Please literally describe the context of the image in detail.

Prompt3:

Give me a detailed literal description of the image.

Figure 13: Prompts for Literal Description Generation in experiments.

Prompt1:

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.

Prompt2:

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.

Prompt3:

Given an image, the image is divided into two or more panels. There is the contrast relationship in the image through panels. Describe the elements visible in each panel. Give me the concise interpretation how these panels convey contrasting perspectives, which you only need to output the contradiction in one or two sentences.

Figure 14: Prompts for Contradiction Generation in experiments.

Prompt1:

The given comic shows the same situation from two opposite sides with contradictions. Which of the following options best represents the underlying philosophy of the comic?
{MCQ Options}
Just output the choice:

Prompt2:

You are presented with an image, which is divided into two or more panels, each illustrating contrasting views of the same scenario.
Which of the following options best represents the philosophy of the image provided?
{MCQ Options}
Select the correct option by typing the corresponding letter (A, B, C, or D).

Prompt3:

Given an image, which has two or more panels. There is contrast in these panels.
Tell me the best option in the following options who represents the deep semantic of the image?
{MCQ Options}
Just tell me the correct option by outputing corresponding letter (A, B, C, or D), no more explanation.

Figure 15: Prompts for Underlying Selection Task in experiments.

Prompt1:

The given comic shows the same situation from two opposite sides with contradictions. Which of the following titles are the most suitable for the comic?
{MCQ Options}
Just output the choice:

Prompt2:

You are presented with an image, which is divided into two or more panels, each illustrating contrasting views of the same scenario. Which of the following title options best represents the image provided?
{MCQ Options}
Select the correct option by typing the corresponding letter (A, B, C, or D).

Prompt3:

Given an image, the image is divided into two or more panels. There is the contrast relationship in the image through panels.
Tell me the best title in the following title options who represents the image?
{MCQ Options}
Just tell me the correct option by outputing corresponding letter (A, B, C, or D), no more explanation.

Figure 16: Prompts for Title Matching Task in experiments.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of the work in Section 8.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The experimental settings are thoroughly explained in Section 4 and further detailed in Appendix Section B. Additionally, the code is released, ensuring that the experiments can be replicated as described, supporting the main claims and conclusions of the paper.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide the corresponding benchmark.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental settings are thoroughly explained in Section 4 and further detailed in Appendix Section B.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All relevant details, such as the type of compute workers for open-sourced models and the API specifications for commercial models, are provided in Section B.

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: We have ensured that all ethical guidelines were followed throughout the study, with careful consideration of fairness, transparency, and the responsible use of data and models. An Ethics statement is provided in Section 9.

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Please refer to the Section 1, Section 2, Section 8, and Section 9.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The creators and original owners of all assets used in the paper, including code, data, and models, are properly credited. Additionally, the licenses and terms of use for the comic data are explicitly stated and fully respected, as outlined in Section 9.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Section 3, Section 4 and Section B for newly proposed GPT based evaluation.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Section A

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]