

YesBut: A High-Quality Annotated Multimodal Dataset for evaluating Satire Comprehension capability of Vision-Language Models

Abhilash Nandy[♣] Yash Agarwal[♣] Ashish Patwa[♣] Millon Madhur Das[♣]
Aman Bansal[♣] Ankit Raj[◇] Pawan Goyal[♣] Niloy Ganguly[♣]
nandyabhilash@kgpian.iitkgp.ac.in

[♣]Indian Institute of Technology Kharagpur [♣]University of Massachusetts Amherst
[◇] Haldia Institute of Technology

Abstract

Understanding satire and humor is a challenging task for even current Vision-Language models. In this paper, we propose the challenging tasks of *Satirical Image Detection* (detecting whether an image is satirical), *Understanding* (generating the reason behind the image being satirical), and *Completion* (given one half of the image, selecting the other half from 2 given options, such that the complete image is satirical) and release a high-quality dataset **YesBut**, consisting of 2547 images, 1084 satirical and 1463 non-satirical, containing different artistic styles, to evaluate those tasks. Each satirical image in the dataset depicts a normal scenario, along with a conflicting scenario which is funny or ironic. Despite the success of current Vision-Language Models on multimodal tasks such as Visual QA and Image Captioning, our benchmarking experiments show that such models perform poorly on the proposed tasks on the **YesBut** Dataset in Zero-Shot Settings w.r.t both automated as well as human evaluation. Additionally, we release a dataset of 119 real, satirical photographs for further research¹.

1 Introduction

Satire is a form of humor that uses irony or exaggeration to criticize or mock people, politics, or society. It serves as a powerful tool to highlight issues, provoke thought, and often encourages a critical perspective on the subject matter. Satirical images posted on social media often consist of conflicting scenarios to convey irony and humor. Understanding such conflicting scenarios requires understanding interaction among entities and text (if any) within the image, along with common-sense knowledge and reasoning capabilities. Fig. 1 shows an example image conveying satire. The irony in the image is that the person is messaging

¹The dataset and code are available at https://github.com/abhi1nandy2/yesbut_dataset

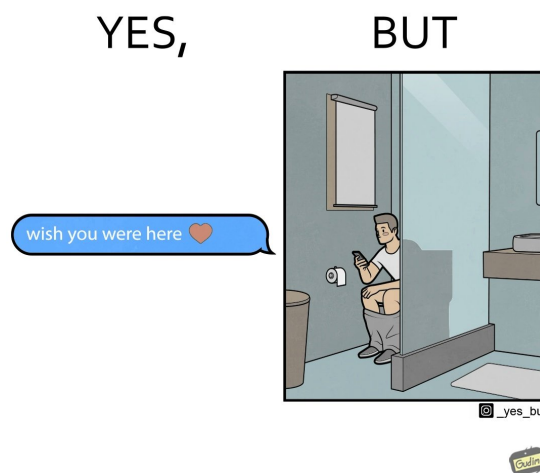


Figure 1: Satire conveyed through a social media image

someone a very heartfelt message on the mobile, while sitting on a toilet seat!

People convey humor on the internet and social media using images, GIFs, and videos. Previous studies have shown that memes (Buchel, 2012) and TV show Clips (Attardo et al., 2003) are prevalent means for expressing such humor. There have also been attempts at detecting (Hasan et al., 2019; Castro et al., 2019; Tanaka et al., 2022) and describing (Hwang and Schwartz, 2023) multimodal satire and humor. However, very few works have *simultaneously studied the detection, understanding, and comprehension of satirical situations in society in the multimodal setting*.

There has been a rise in the development of Vision-Language (VL) models (Liu et al., 2023; Huang et al., 2023; Peng et al., 2023; Zhu et al., 2023; OpenAI, 2023; Team, 2023). Such models have shown remarkable State-Of-The-Art (SOTA) performance on several downstream tasks such as Visual Question Answering and Image Captioning. Such models are pre-trained in a manner that im-

ages and text have shared embedding space, and that, images and their corresponding text descriptions have similar representations in that embedding space (Radford et al., 2021; Zhai et al., 2023).

In this paper, we investigate whether existing VL Models are able to decipher satire in images. To do so, we propose 3 benchmarking tasks - (1) *Satirical Image Detection* - Given an image, classify the image as being satirical or not (2) *Satirical Image Understanding* - Given a satirical image, describing in natural language why the image is satirical (3) *Satirical Image Completion* - Given a part of the image, correctly select the remaining part of the image from 2 options. These tasks go beyond image recognition and language understanding, and are challenging, as understanding satire usually involves understanding the punchline corresponding to a sudden twist or a funny quip in a given situation (Ramachandran, 1998). For example, in Fig. 1, the model needs to first comprehend the text "wish you were here", followed by understanding that the text was sent by a person sitting in the toilet from the image on the right, and then finally grasp the irony of the situation.

To evaluate the tasks, we collected a high-quality multimodal dataset *YesBut* consisting of 1,084 satirical and 1,463 non-satirical images, where each image contains 2 sub-images with the same/different artistic styles. In each *satirical image*, the left sub-image describes a scenario, and the right sub-image presents another scenario which either contradicts or pokes fun at the first scenario, creating an element of satire. Additionally, each such satirical sample is annotated to get the description of individual images inside the sample, as well as the overall description containing the punchline that conveys the satire.

We perform detailed evaluation on the satirical image detection, understanding, and completion tasks using recent VL models in zero-shot and zero-shot Chain-of-Thought (CoT) (Kojima et al., 2022) settings (as we want to observe how well the models can decipher satire without the support of additional training/in-context examples). We observe that the task of satirical image detection is especially difficult. Also, even though Gemini performs the best in Satirical Image Understanding and Completion tasks, there is a significant scope for improvement in SOTA VL Models in understanding and comprehending satire in images in zero-shot scenarios. Also, for further research, we release an additional set of 119 diverse, real, satiri-

cal photographs. We infer that SOTA VL Models fail to perform well even on real photographs (*see Section A in Appendix* for more details).

2 Background

2.1 Satirical and Humor Datasets

Previous works on satire and humor in NLP and Computer Vision mostly revolve around detecting satire in text (Rogoz et al., 2021) and multimodal scenarios (Li et al., 2020; Ionescu and Chifu, 2021), detecting humor in multimodal scenarios (Hasan et al., 2019; Castro et al., 2019), meme/joke captioning (Hwang and Schwartz, 2023; Hessel et al., 2023), etc. However, no such work performs a comprehensive and simultaneous evaluation of satire and humor detection, understanding, and comprehension capabilities of VL Models in Multimodal Scenarios.

2.2 Other Image Datasets

The WHOOPS benchmark, introduced by Bitton-Guetta et al. (2023), comprises unconventional images challenging commonsense expectations, both human-created and machine-generated, accompanied by corresponding textual descriptions. Specifically designed for tasks such as image captioning, image-text matching, visual question answering, and explanation generation, it provides a unique dataset for evaluating model performance in these domains. In contrast, our work performs a holistic evaluation of different SOTA VL Models on their ability to detect, understand, and comprehend satire in images.

3 Our Annotation Pipeline

The entire data collection and annotation pipeline is shown in Fig. 2. We curated a collection of annotated satirical and non-satirical images in this section in 4 stages.

3.1 Stage 1: Collecting Satirical Images from Social Media

We manually downloaded images from the posts in 'X' (erstwhile known as Twitter) handle @_yesbut_ (with proper consent). We manually filtered 283 images that are satirical, and annotated them in the next stage. Each image contains two sub-images (which are colorized sketches), one on the left showing a normal scenario, while one on the right is ironical/pokes fun at the left sub-image.

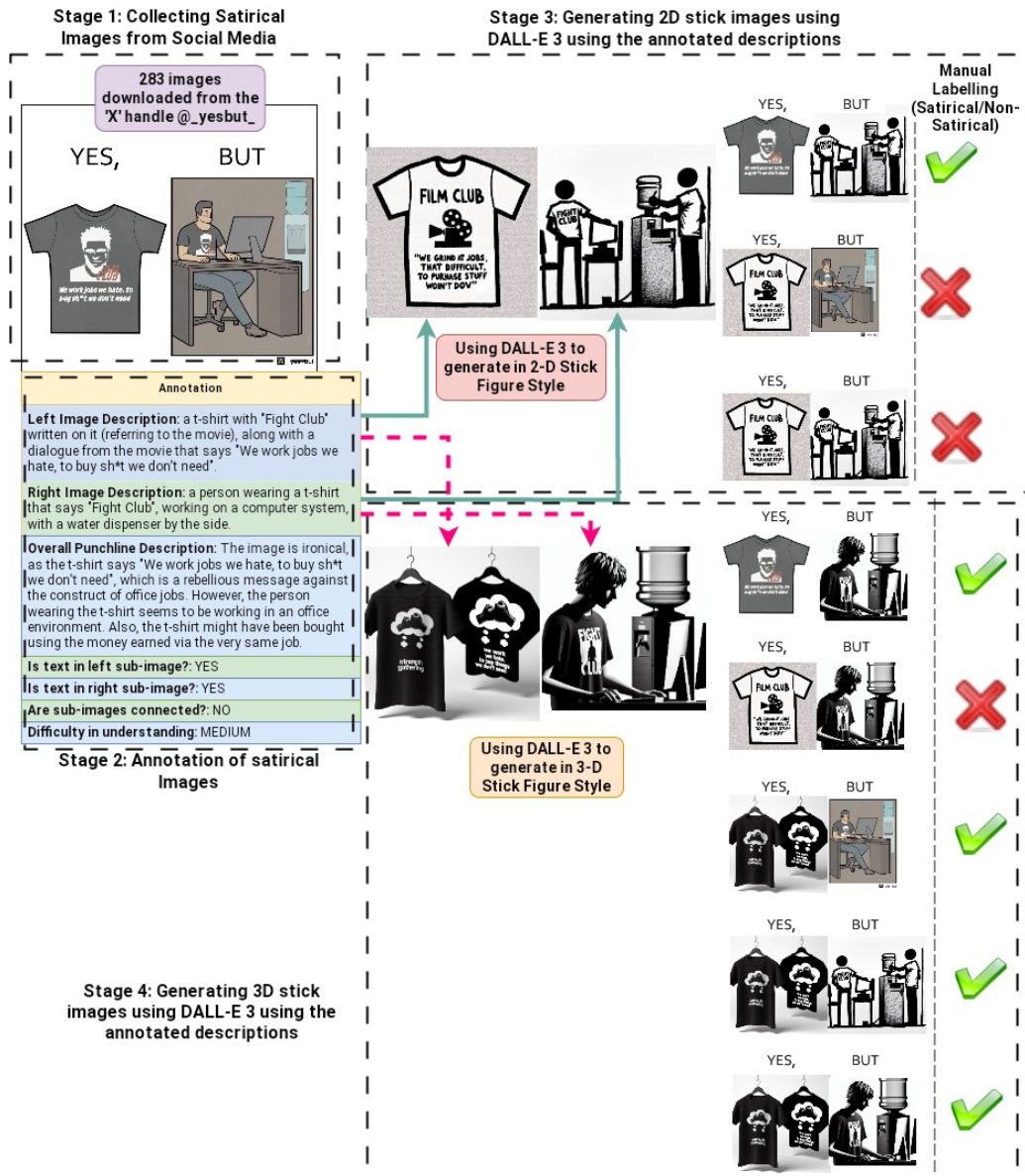


Figure 2: Our annotation Pipeline for *YesBut* in 4 Stages - (1) Collecting Satirical Images from Social Media (2) Human Annotation of satirical images (3) Generating 2D stick images using DALL-E 3 and annotated descriptions (4) Generating 3D stick images using DALL-E 3 and annotated descriptions

3.2 Stage 2: Annotation of satirical images

Textual descriptions and certain categorical features of *satirical* images were annotated using 5 annotators, all of whom met the qualification criteria of being undergraduate sophomore students or above, enrolled in English-medium colleges. Specifically, we collected the following features (these were given as annotator instructions) for every image - (1) Textual Description of the Left Sub-Image (2) Textual Description of the Right Sub-Image (3) Overall Textual Description which contains the punchline (4) A binary feature on whether the Left sub-Image contains any text (5) A binary feature on whether the Right sub-Image contains

any text (6) A binary feature on whether the sub-images can be created by dividing a larger image using a vertical line as a separator (this is later referred to as the) (7) A categorical feature on how difficult the annotation was. This can have 3 possible values - 'EASY' when the annotator does not need any additional help from the internet, 'MEDIUM' when the annotator needs additional help from the internet to understand the overall description, and 'HARD' when additional help from the internet is needed to write all the 3 aforementioned textual descriptions.

Fig. 3 shows distribution of the 283 images based on different aspects of image content and

annotated descriptions. We can see that - (1) from Fig. 3a, more than half of the images have no text, which would make it difficult for the VL Models to understand those images due to absence of a text modality; (2) from Fig. 3b, more than 94% of the images do not have connected sub-images, requiring the VL Models to understand the connection between the objects in the two sub-images; (3) from Fig. 3c, a significant 13.5% of the (MEDIUM and HARD) images required annotators to refer to the internet to annotate the images, which makes the dataset challenging; (4) from Fig. 3d, the overall descriptions of the images containing the punchline were classified into 4 different types of satire by passing the following prompt to ChatGPT (OpenAI, 2021) as input - "What are the different types of satire?", and then asking ChatGPT to further classify overall descriptions of images into 1 of the 4 obtained categories. Most of the images show Social Satire (it focuses on cultural trends, social conventions, and the absurdities of everyday life) and Horatian Satire (it aims to amuse rather than enrage, often using wit, irony, exaggeration to poke fun at societal norms and human folly).

3.3 Stage 3: Generating 2D stick images using DALL-E 3 on the annotated descriptions

To increase the size and the diversity of the dataset, we use the DALL-E 3 (Betker et al., 2023) image generation model to generate synthetic sub-images using the annotated left and right sub-image descriptions (obtained in Stage 2). We use the following prompt - "Draw using stick figures (black silhouette against a white background) - <SUB-IMAGE DESCRIPTION>". Given the original sub-images, 3 new combinations of sub-images are obtained ([original left sub-image, generated right 2D stick sub-image], [generated left 2D stick sub-image, original right sub-image], [generated left 2D stick sub-image, generated right 2D stick sub-image]). We manually label each new combined image as satirical or non-satirical (details of this manual labelling is given *in Section C.3 of Appendix*). At the end of the image generation followed by manual labelling, we end up adding 302 satirical and 547 non-satirical images. Each satirical image generated is assigned the same textual descriptions as the original image.

3.4 Stage 4: Generating 3D stick images using DALL-E 3 on the annotated descriptions

Similar to Stage 3, we further increase the size and diversity using DALL-E-3. We use the following prompt - "Draw using 3D black silhouettes against a white background - <SUB-IMAGE DESCRIPTION>". Given the original sub-images and the sub-images generated in Stage 3, 5 new combinations of sub-images are obtained ([original left sub-image, generated right 3D stick sub-image], [generated left 3D stick sub-image, original right sub-image], [generated left 2D stick sub-image, generated right 3D stick sub-image], [generated left 3D stick sub-image, generated right 2D stick sub-image], [generated left 3D stick sub-image, generated right 3D stick sub-image]). We manually label each new combined image as satirical or non-satirical. At the end of the image generation followed by manual labelling, we end up adding 499 satirical and 916 non-satirical images. Each satirical image generated is assigned the same textual descriptions as the original image.

To ensure that the *YesBut* dataset is of high quality, we validate 25 randomly-sampled satirical and 25 randomly-sampled non-satirical samples (from the images obtained in Stages 3 and 4 of annotation) using another annotator. We find that the 2 annotators agree on 94% of samples, which is a very high agreement, ensuring that the dataset is of a very high quality.

4 The *YesBut* Dataset

The *YesBut* dataset has a total of 2,547 images, 1,084 of which are satirical, the rest 1,463 images being non-satirical. These images spread across 3 diverse artistic styles - colored sketch, 2D stick figure, 3D stick figure.

Dataset	Absence of text in image	Presence of sub-images	Presence of multiple artistic styles	Tasks relevant to multimodal humor comprehension
MemeCap	8.05%	40.61%	15.56%	Meme Captioning
MET-Meme	1.43%	17.92%	9.52%	Meme Understanding, Sentiment Analysis
<i>YesBut</i>	53% ²	100%	88.89%	Satirical Image Detection, Understanding, Completion

Table 1: Statistics of the presence/absence of text, sub-images, and multiple image styles and tasks evaluated in prior datasets vs. *YesBut*.

²ignoring 'YES', 'BUT' Markers at the top of each image, as these markers are not related to the content of the image (see Fig. 1)

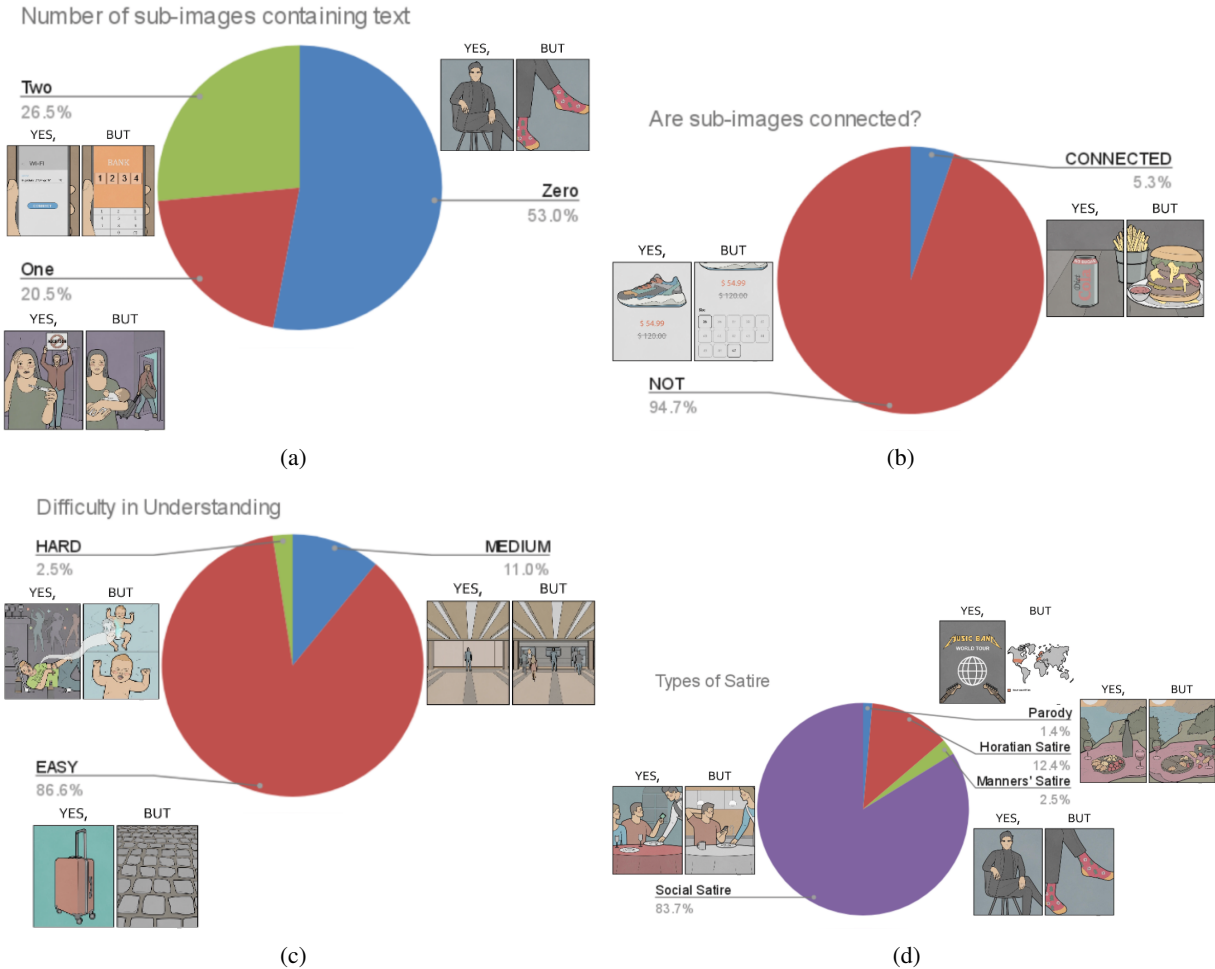


Figure 3: Distribution of the original 283 satirical images downloaded from Social Media based on different aspects of image content and annotated descriptions

Table 1 compares *YesBut* with multimodal satirical and humor datasets from prior art. We can see that *YesBut* has a much larger fraction of images that (1) do not have text, (2) have sub-images, (3) have multiple artistic styles within the image, in comparison to MemeCap (Hwang and Schwartz, 2023) and MET-Meme (Xu et al., 2022) datasets. Lack of text and presence of multiple artistic styles across sub-images makes it challenging for the VL Models to comprehend satire in the images present in *YesBut*. Additionally, the tasks in *YesBut* ensure a more holistic evaluation of satire and humor compared to MemeCap and MET-Meme.

The satirical images cover several aspects of societal satire. To analyze this, we use topic modeling on the left and right sub-image descriptions using BERTopic (Grootendorst, 2022). We get 7 topics (each topic being an unordered set of representative words), which are further elaborated using ChatGPT to get intuitive descriptions for each topic

(refer to *Section D of Appendix*).

We further visualize the diversity of these sub-images by plotting the compressed 2D image representations obtained by applying UMAP (McInnes et al., 2018) on the pre-trained CLIP (Radford et al., 2021) (MIT License) image representations in Fig. 4. The image samples are plotted in different colors based on their artistic style. The original 283 images are not very diverse. However, the generated images of the 2D and 3D stick figure styles are comparatively much more diverse and are semantically distant from the original images, even though they have the same sub-image descriptions. Hence, all the satirical images are highly diverse and cover various scenarios of societal satire.

5 Experimental Setup

We report the performance of various SOTA VL Models (described in Sec. 5.1) for performance evaluation on the tasks (described in Sec. 5.2) de-

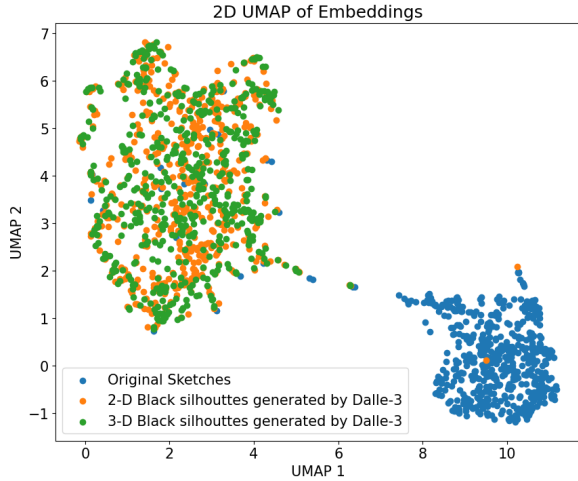


Figure 4: 2D UMAP Representations of CLIP Image representations of *YesBut* sub-images

vised for the *YesBut* Dataset. The evaluation setup and experimental results are described in Sec. 5.3 and 5.4, respectively.

5.1 Models

Gemini. Gemini (Team, 2023) is a closed-source family of Large Multimodal Models (LMMs) from Google. The Gemini project comprises Ultra, Pro, and Nano variants, designed to excel in image and text comprehension. These models cater to diverse applications, from intricate reasoning tasks to memory-constrained on-device scenarios. Notably, the Gemini Ultra model demonstrates SOTA performance across 30/32 benchmarks. Furthermore, it outperforms existing models in all 20 multimodal benchmarks examined. The Gemini models showcase remarkable capabilities in cross-modal reasoning and language understanding. We leverage Gemini Pro Vision API for all tasks in our paper.

GPT4. GPT4 (OpenAI, 2023) is an advanced, closed-source multimodal model capable of processing both image, text inputs to generate coherent textual outputs. GPT4 demonstrates human-level proficiency across professional, academic benchmarks. It achieves commendable performance, ranking within the top 10% of test takers in a simulated bar exam. Operating on an Autoregressive Transformer-based architecture (Vaswani et al., 2017), GPT4 undergoes pre-training to predict subsequent tokens in a document. The subsequent post-training alignment enhances its performance in terms of factuality and adherence to desired behavior. We use gpt-4-vision-preview API for all tasks in our paper.

LLaVA. LLaVA (Large Language and Vision Assistant), proposed by Liu et al. (2023), utilizes visual encoder from pre-trained CLIP (Radford et al., 2021) along with LLaMA (Touvron et al., 2023) language model. The approach involves instruction tuning on visual instruction data assisted by GPT4 (OpenAI, 2023) for enhanced performance.

MiniGPT4. MiniGPT4 (Zhu et al., 2023) has frozen pre-trained language and vision components. It utilizes a singular projection layer to align visual and language features. Notably, it exhibits analogous capabilities to GPT4 in comprehending context. MiniGPT4 uses Vicuna (Chiang et al., 2023) language model, built upon LLaMA-13B, demonstrating performance on par with ChatGPT. In the domain of vision, it integrates BLIP-2 (Li et al., 2023), comprising CLIP ViT-G/14 (Radford et al., 2021) and a Q-Former (Zhang et al., 2024) architecture. Training MiniGPT4 encompasses diverse multimodal datasets, incorporating images from LAION (Schuhmann et al., 2022), Conceptual Captions (Sharma et al., 2018), and SBU (Ordonez et al., 2011).

Kosmos-2. Equipped with a robust capability to comprehend diverse modalities, Kosmos-2 (Peng et al., 2023) excels in undertaking an extensive array of tasks, ranging from zero-shot and few-shot scenarios to intricate multimodal chain-of-thought prompting situations. The model leverages textual instructions for enhanced comprehension of downstream tasks. In the context of chain-of-thought prompting, Kosmos-2 refines its approach by integrating grounding and referring capabilities, utilizing a structured format comprising text spans and bounding boxes as prompts. This innovative approach enhances the model’s effectiveness in generating coherent and contextually grounded responses, exemplifying the evolution from Kosmos-1 (Huang et al., 2023).

Table 2 shows the number of parameters of the open-source VL Models³.

	Number of Parameters (In Billions)
LLaVA (Liu et al., 2023)	7
Kosmos-2 (Peng et al., 2023)	1.66
MiniGPT-4 (Zhu et al., 2023)	15.7

Table 2: Number of Parameters - Open-Source Models

³Compute Details are in Section E.1 of Appendix

5.2 Tasks

We describe the tasks that are evaluated on the *YesBut* Dataset -

Satirical Image Detection: This is a binary classification task, where given an image, the model needs to predict whether the image is satirical or not. This task is carried out on all the 2547 images. Some example input images, along with input the text prompt used for all images is mentioned *in Section E.2 of Appendix.*

Satirical Image Understanding: Given a satirical image, we evaluate the model’s satire understanding capability in images by (1) prompting the model to generate a textual description of each sub-image as input, using the prompt “Describe the image”. (2) prompting the model to generate the punchline in the image using the following prompt (referred to as “WHYFUNNY_PROMPT” hereafter) - “Why is this image funny/satirical?”. This task is carried out on only the 1084 satirical images of the *YesBut* Dataset.

Satirical Image Completion: Given either the left or right sub-image having the style of a colorized sketch, the other sub-image needs to be chosen from two options, one having a 2D, and the other having a 3D stick figure style, such that the entire image so formed is meaningful and satirical. The options are curated based on existing satirical and non-satirical images from the *YesBut* Dataset. We curate 150 such samples for evaluation. Some example input images, along with input the text prompt used for all images is mentioned *in Section E.2 of Appendix.*

5.3 Evaluation Setup

Satirical Image Detection: We use Zero-Shot and Zero-Shot Chain-of-Thought (CoT) (Kojima et al., 2022) setups for inference, and metrics used for binary classification such as Accuracy and F1-Score for evaluation.

Satirical Image Understanding: We use Zero-Shot setup for inference, and standard metrics for automatic evaluation of text generation-based tasks - lexical overlap metrics such as BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), and METEOR (Banerjee and Lavie, 2005), and semantic similarity metrics such as BERTScore (Zhang* et al., 2020) to evaluate the image understanding capabilities of the images and corresponding sub-images (we also experiment with an image-based evaluation metric Polos (Wada et al., 2024), whose results

are shown *in Section E.4 of the Appendix*). Additionally, we randomly sample 30 images (10 images from each obtained in Stage 2, Stage 3, Stage 4) along with their model-generated and human-written overall image descriptions. Each image description is human-evaluated based on the following (binary) criteria (adopted from (Hwang and Schwartz, 2023) and slightly changed⁴ to better suit evaluation on *YesBut*) - (1) **Correctness:** Is the image description correctly able to convey the satire the image wanted to convey? (2) **Appropriate Length:** Is the image description length appropriate for conveying the meaning (i.e. it is not too verbose)? (3) **Visual Completeness:** Does the image description describe all the important elements in the image? (4) **Faithfulness:** Are all the elements of the image description supported by either the visual or text elements (i.e. there are no made-up elements)? - The annotation is carried out by 3 students in the lab⁵, and the majority vote is taken for each image.

Satirical Image Completion: We use Zero-Shot and Zero-Shot CoT setups for inference. and accuracy as the evaluation metric.

Note that we do not use In-Context Learning Setting for inference because this would make the tasks less challenging for the models. Also, we want to analyze how well VL models can comprehend satire on their own *without any support from other exemplars.*

5.4 Results

	TEST ACC.	F1 SCORE
LLaVA (0-shot)	53.67	48.64
LLaVA (0-shot, CoT)	52.22	46.87
Kosmos-2 (0-shot)	42.56	59.71
Kosmos-2 (0-shot, CoT)	56.97	20.35
MiniGPT4 (0-shot)	48.29	49.33
MiniGPT4 (0-shot, CoT)	48.88	50.61
GPT4 (0-shot)	55.44	55.13
GPT4 (0-shot, CoT)	48.29	42.32
Gemini (0-shot)	50.82	48.29
Gemini (0-shot, CoT)	46.36	38.93

Table 3: Evaluation of different VL models on the Satirical Image Detection task

Satirical Image Detection: Table 3 shows the results of satirical image detection capability of VL Models on the *YesBut* Dataset. We can infer that -

⁴We do not use ‘Textual Completeness’ from Hwang and Schwartz (2023), as many images in *YesBut* do not contain text

⁵The annotators who annotated *YesBut* were not a part of the human evaluation

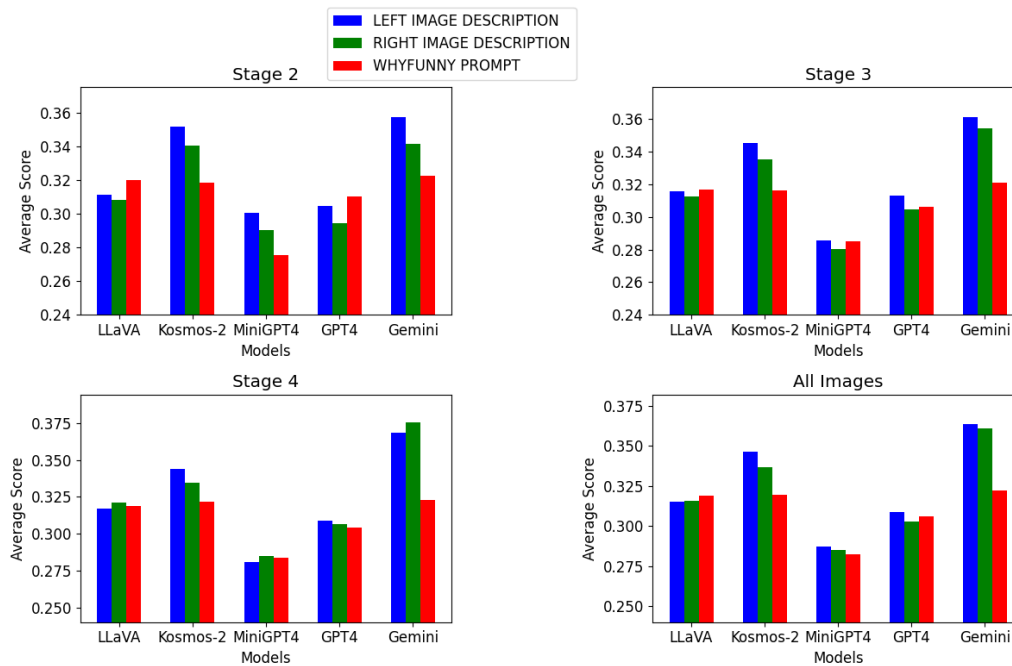


Figure 5: Evaluation of Satirical Image Understanding Capability using multiple VL models at different stages (Stages 2, 3, 4) of annotation of *YesBut*, as well as, for all *YesBut* images

(1) Kosmos-2 in zero-shot CoT and zero-shot settings give the best test accuracy and F1 Score respectively due to its superior visual grounding capabilities (2) Improvement in test accuracy and F1 Score due to CoT is seen only in 2/5 and 1/5 models respectively, suggesting that SOTA VL Models are unable to properly reason/rationalize whether a given image has an element of satire in it (3) Both test accuracy and F1 Score do not cross 60% for any SOTA VL Model, suggesting that there is a significant scope for improvement when it comes to detecting satire/humor in a given image.

Satirical Image Understanding: Fig. 5 shows the average value of the 4 automated metrics (discussed in Sec. 5.3) to evaluate satirical image understanding capability of VL Models at different stages of annotation of *YesBut* (see Table 6 in Section E.4 of Appendix to get individual values of the evaluation metrics, along with performance variation w.r.t annotation difficulty and presence of text in images). We observe that - (1) There is a reduction in the overall understanding capability (average metric corresponding to ‘WHYFUNNY PROMPT’) of the majority of models in Stages 3 and 4 compared to Stage 2, as images in Stages 3 and 4 have different artistic styles in the same image, unlike Stage 2 (2) Kosmos-2 almost always performs better than other open-source models LLaVA and MiniGPT4, as Kosmos-2 has multimodal ground-

ing and referring capabilities, which LLaVA and MiniGPT4 do not have (3) 4 out of 5 models do not understand the entire image better than sub-images within the image across the entire *YesBut* Dataset. Gemini, Kosmos-2 encounter a huge drop in overall reasoning compared to sub-image reasoning, despite showing remarkable cross-modal reasoning and visual grounding capability respectively (4) MiniGPT4 gives the worst performance among all models due to restricted leverage of visual modality compared to textual modality, as stated in Hwang and Shwartz (2023). (5) All average metric values (normalized between 0 and 1) are below 0.4, which shows that there is a lot of scope for improvement in satire understanding capability of SOTA VL Models.

	TEST ACC.
LLaVA (0-shot)	51.33
LLaVA (0-shot, CoT)	56.55
Kosmos-2 (0-shot)	54.67
Kosmos-2 (0-shot, CoT)	53.33
MiniGPT4 (0-shot)	40
MiniGPT4 (0-shot, CoT)	60.67
GPT4 (0-shot)	58.67
GPT4 (0-shot, CoT)	57.33
Gemini (0-shot)	61.11
Gemini (0-shot, CoT)	61.81

Table 4: Evaluation of different VL models on the Satirical Image Completion task

Figure 14 in Section E.4 of Appendix compares the overall image descriptions generated by 5 SOTA Models with ones written by human annotators based on human evaluation (see **Table 10 in Section E.4 of Appendix** for an example). We can see that Gemini and GPT4 perform satisfactorily among the 5 models. However, performance of the (aspect-wise) best model is 40, 43.33, 33.33, 36.66 points less compared to human-level performance on Correctness, Appropriate Length, Visual Completeness, and Faithfulness respectively.

Satirical Image Completion: Table 4 shows results of the satirical image completion task on **YesBut**. We observe that - (1) CoT improves results in 3/5 models, as reasoning is needed to understand the relation between sub-images better (2) Among open-source models, improvement due to CoT is the highest for MiniGPT4, which is the largest open-source model in our study (see Table 2). This is consistent with the observation of Zero-Shot CoT working better for larger models (Kojima et al., 2022) (3) Gemini performs best in both zero-shot and zero-shot CoT settings among all the models.

6 Summary and Conclusion

We present **YesBut**, a high-quality annotated multimodal dataset for Satire Comprehension Evaluation. Our work is one of the first to systematically benchmark multimodal Satire Comprehension ability of SOTA VL Models by proposing 3 non-trivial tasks of Satire Detection, Understanding, and Completion. We observe that SOTA VL Models struggle in those tasks, as **YesBut**, unlike other benchmarks, contains images with sub-images having different artistic styles and no text in most cases, making **YesBut** a challenging multimodal dataset for satire detection and comprehension.

7 Limitations

Subjectivity of annotations: The annotation task involves utilizing background knowledge that may differ among annotators. Consequently, we manually reviewed the annotations to minimize the number of incorrect annotations in the dataset. However, some subjectivity still remains.

Extension to languages other than English: This work is in the English Language. However, we plan to extend our work to languages other than English.

References

- Salvatore Attardo, Jodi Eisterhold, Jennifer Hay, and Isabella Poggi. 2003. **Multimodal markers of irony and sarcasm**. *HUMOR*, 16(2):243–260.
- Satanjeev Banerjee and Alon Lavie. 2005. **METEOR: An automatic metric for MT evaluation with improved correlation with human judgments**. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. **Improving image generation with better captions**. *Computer Science*. <https://cdn.openai.com/papers/dall-e-3.pdf>, 2(3):8.
- Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, and Roy Schwartz. 2023. **Breaking common sense: Whoops! a vision-and-language benchmark of synthetic and compositional images**.
- Branislav Buchel. 2012. **Internet memes as means of communication**. *Brno: Masaryk University*.
- Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. **Towards multimodal sarcasm detection (an _obviously_ perfect paper)**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Florence, Italy. Association for Computational Linguistics.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. **Vicuna: An open-source chatbot impressing gpt-4 with 90% chatgpt quality**.
- Maarten Grootendorst. 2022. **Bertopic: Neural topic modeling with a class-based tf-idf procedure**. *arXiv preprint arXiv:2203.05794*.

- Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. 2019. [UR-FUNNY: A multimodal language dataset for understanding humor](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2046–2056, Hong Kong, China. Association for Computational Linguistics.
- Kazuki Hayashi, Yusuke Sakai, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. 2024. [Artwork explanation in large-scale vision language models](#). *arXiv preprint arXiv:2403.00068*.
- Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023. [Do androids laugh at electric sheep? humor “understanding” benchmarks from the new yorker caption contest](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 688–714, Toronto, Canada. Association for Computational Linguistics.
- Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. 2023. [Language is not all you need: Aligning perception with language models](#). *arXiv preprint arXiv:2302.14045*.
- EunJeong Hwang and Vered Shwartz. 2023. [MemeCap: A dataset for captioning and interpreting memes](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1433–1445, Singapore. Association for Computational Linguistics.
- Radu Tudor Ionescu and Adrian Gabriel Chifu. 2021. [Fresada: A french satire data set for cross-domain satire detection](#). In *2021 International Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.
- Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. [Large language models are zero-shot reasoners](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. [Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models](#).
- Lily Li, Or Levi, Pedram Hosseini, and David Broniatowski. 2020. [A multi-modal method for satire detection using textual and visual cues](#). In *Proceedings of the 3rd NLP4IF Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 33–38, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. [Visual instruction tuning](#).
- Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. [Umap: Uniform manifold approximation and projection](#). *Journal of Open Source Software*, 3(29).
- OpenAI. 2021. [Gpt-3.5 turbo documentation](#).
- OpenAI. 2023. [Gpt-4 technical report](#).
- Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. [Im2text: Describing images using 1 million captioned photographs](#). In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. [Kosmos-2: Grounding multimodal large language models to the world](#). *arXiv preprint arXiv:2306.14824*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. [Learning transferable visual models from natural language supervision](#). In *International conference on machine learning*, pages 8748–8763. PMLR.
- Vilayanur S Ramachandran. 1998. [The neurology and evolution of humor, laughter, and smiling: the false alarm theory](#). *Medical hypotheses*, 51(4):351–354.
- Ana-Cristina Rogoz, Gaman Mihaela, and Radu Tudor Ionescu. 2021. [SaRoCo: Detecting satire in a novel Romanian corpus of news articles](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1073–1079, Online. Association for Computational Linguistics.
- Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. [Laion-5b: An open large-scale dataset for training next generation image-text models](#).

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. [Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Kohtaro Tanaka, Hiroaki Yamane, Yusuke Mori, Yusuke Mukuta, and Tatsuya Harada. 2022. [Learning to evaluate humor in memes based on the incongruity theory](#). In *Proceedings of the Second Workshop on When Creative AI Meets Conversational AI*, pages 81–93, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Gemini Team. 2023. [Gemini: A family of highly capable multimodal models](#).

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. [Llama: Open and efficient foundation language models](#).

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

Yuiga Wada, Kanta Kaneda, Daichi Saito, and Komei Sugiura. 2024. Polos: Multimodal Metric Learning from Human Feedback for Image Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Bo Xu, Tingting Li, Junzhe Zheng, Mehdi Naseriparsa, Zhehuan Zhao, Hongfei Lin, and Feng Xia. 2022. [Met-meme: A multimodal meme dataset rich in metaphors](#). In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’22*, page 2887–2899, New York, NY, USA. Association for Computing Machinery.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. *arXiv preprint arXiv:2303.15343*.

Qiming Zhang, Jing Zhang, Yufei Xu, and Dacheng Tao. 2024. Vision transformer with quadrangle attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

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

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. [Minigt-4: Enhancing vision-language understanding with advanced large language models](#).

Appendix

The Appendix mirrors the sectional structure of the main paper, placing supplementary material for each section in its corresponding appendix section for easy reference. If some sections or subsections lack additional material, only their titles are listed.

A Introduction



Figure 6: Example of a real photograph following the ‘Yes, But’ Theme

Dataset of real, satirical images: We collected a dataset of 119 images containing irony, satire from instagram posts by different users, who resort to using “Yes, But” theme over real photos (e.g. see Figure 6). We perform following 2 tasks on these images - (1) Satirical Image Detection, where we report detection accuracy, as all images have ground truth of “Satirical” (2) Satirical Image Understanding, where we use the WHYFUNNY text prompt and the image as input to the VL Models. The output is evaluated using human evaluation, where the annotator needs to answer whether model-generated text correctly describes satire in the image, and the corresponding accuracy for each

VL Model is reported. The results are shown in Table 5. We can infer that 3 out of 5 models give less than satisfactory performance on Detection, and all models give an accuracy of less than 50% on Image Understanding. Hence, even on real photographs, SOTA VL Models fail to perform well.

Model	Detection	Understanding
LLaVA	18.48	22.69
Kosmos-2	66.39	10.92
MiniGPT4	52.94	10.92
GPT4	93.27	46.22
Gemini	80.67	19.33

Table 5: Performance of different SOTA VL Models on Satirical Detection and Understanding Tasks on real photographs

B Background

B.1 Satirical and Humor Datasets

B.2 Other Image Datasets

C Our Annotation Pipeline

C.1 Stage 1: Collecting Satirical Images from Social Media

C.2 Stage 2: Annotation of Satirical Images

C.3 Stage 3: Generating 2D stick images using DALL-E 3 on the annotated descriptions

Details of the manual labelling: The manual labelling of whether an image with one or more generated sub-images is satirical or not is carried out by a graduate student in our lab. The annotator was given 10 satirical and 10 non-satirical images prior to the manual labelling to provide assistance for the labelling.

C.4 Stage 4: Generating 3D stick images using DALL-E 3 on the annotated descriptions

D The *YesBut* Dataset

Topics obtained after topic-modelling on the left and right sub-image descriptions of satirical images in *YesBut*, along with topic descriptions from ChatGPT -

- *gate_shorts_step_allowed_person*: Likely related to airport security procedures or access control systems, involving individuals wearing shorts being allowed to proceed through a gate or checkpoint.
- *phone_screen_mobile_smartphone_person*: Refers to activities or interactions involving

individuals using their smartphones, possibly related to mobile technology, communication, or digital engagement.

- *woman_image_shows_saying_text*: Implies content featuring women in images, possibly conveying messages or text, suggesting contexts such as advertisements, social media posts, or presentations.
- *plate_table_food_box_cup*: Indicates elements commonly found in dining or food service settings, encompassing plates, tables, various food items, boxes, and cups, suggesting scenarios like restaurants or meal preparation.
- *person_wearing_hair_tattoos_pants*: Describes characteristics of individuals including their clothing choices (pants), hairstyles, and tattoos, likely relevant in contexts such as fashion, identity expression, or cultural representations.
- *car_light_traffic_road_image*: Depicts scenes involving cars, traffic conditions, and roads, possibly associated with transportation, urban environments, or traffic management, often visualized through images.
- *dog_hole_cat_two_throw*: Suggests actions or scenarios involving dogs, cats, and interactions such as throwing, possibly indicating playful or behavioral aspects of these animals, possibly related to pet ownership or animal behavior studies.

E Experimental Setup

E.1 Models

Compute Details: We use an NVIDIA A40 GPU for experiments using the open-source models. The inference time per sample on the GPU for the Satirical Image Detection, Understanding and Completion Tasks for the open-source models go upto around 10 seconds, 1 minute, and 10 seconds respectively.

E.2 Tasks

Text Prompt for *Satirical Image Detection*:

You are an AI expert in detecting humour or satire. User gives you an image, and you have to make a choice "Y" or "N". Instructions: Users image has 2 halves called yes and but, and the combination

of those might make no sense at all, or be extremely funny. Your job is to find out which one it is and output Y if its EXTREMELY funny and N for otherwise. Output format: one character, exactly either "Y" or "N"

Example Image Inputs for Satirical Image Detection:

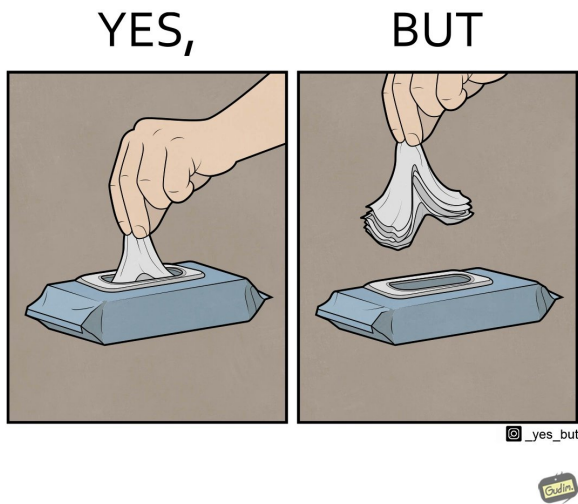


Figure 7: Example of a Satirical Image as input for Satirical Image Detection



Figure 8: Example of a Non-Satirical Image as input for Satirical Image Detection

Figures 7 and 8 are examples of satirical and non-satirical image inputs (for Satirical Image Detection) respectively. For each such image as input to the model, the aforementioned text prompt is used for Satirical Image Detection, and the output is either "Y" (predicting the image is satirical) or "N" (predicting the image is non-satirical).

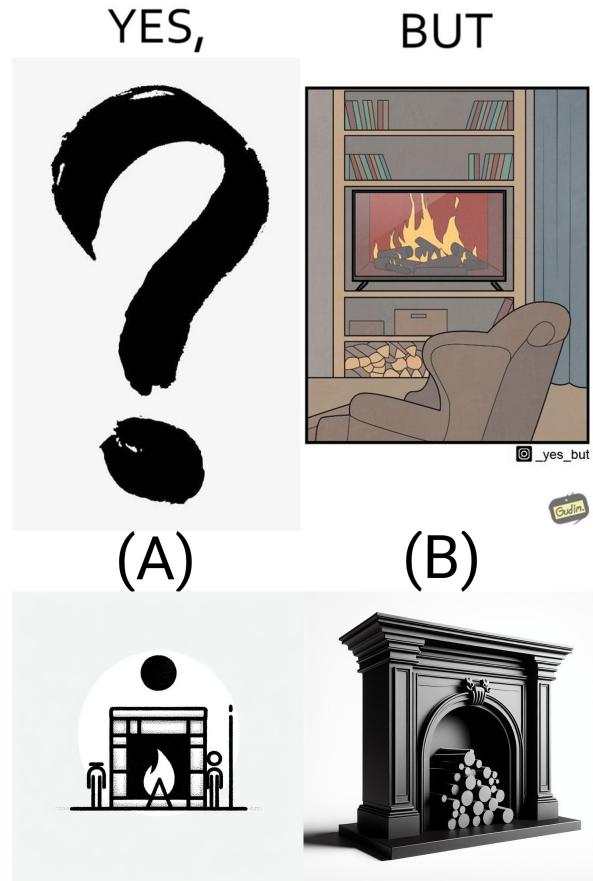


Figure 9: Example of an input image for Image Completion where the left sub-image is to be predicted [ground truth answer - (B)]

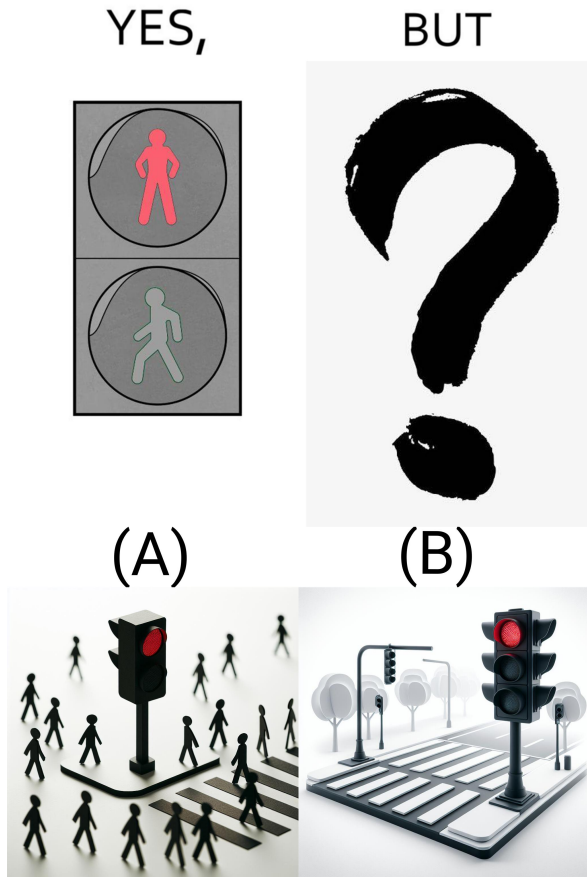


Figure 10: Example of an input image for Image Completion where the right sub-image is to be predicted [ground truth answer - (B)]

Text Prompt for Satirical Image Completion:

You are an AI expert in creating humour or satire. User gives you an image, and you have to make a choice "A" or "B".

Instructions: The image is a 2x2 table with the labels "yes" (top left), "but" (top right), "A" (bottom left), and "B" (bottom right). Either the "yes" cell or the "but" cell will have a question mark in it. Your job is to replace the question mark with either cell "A" or cell "B" so that the resulting [yes,but] pair is funny or satirical. Make a choice "A" or "B":

Output format: one character, exactly either "A" or "B".

Example Image Inputs for Satirical Image Completion:

Figures 9 and 10 are examples of input images for Image Completion where the left and right sub-image is to be predicted respectively. For each such image as input to the model, the aforementioned text prompt is used for Satirical Image Comple-

tion, and the output is either "A" or "B", denoting the sub-image predicted to come in place of the question mark in the input image.

E.3 Evaluation Setup

E.4 Results

Stage of Annotation	Model	BLEU	ROUGE-L	METEOR	BERTScore	Average Score
Stage 1	LEFT IMAGE DESCRIPTION	LLaVA 0.0048	0.1471	0.2352	0.8588	0.3115
		Kosmos-2 0.0113	0.2444	0.271	0.881	0.3519
		MiniGPT4 0.0071	0.1519	0.2075	0.8352	0.3004
		GPT4 0.0071	0.1207	0.2401	0.8506	0.3046
		Gemini 0.0205	0.2513	0.2959	0.862	0.3574
Stage 2	RIGHT IMAGE DESCRIPTION	LLaVA 0.0054	0.1465	0.225	0.8564	0.3083
		Kosmos-2 0.01	0.2358	0.2363	0.88	0.3405
		MiniGPT4 0.0047	0.1427	0.183	0.8317	0.2905
		GPT4 0.0052	0.114	0.2155	0.8435	0.2945
		Gemini 0.0087	0.2259	0.2658	0.8655	0.3415
Stage 3	WHYFUNNY PROMPT	LLaVA 0.0098	0.1778	0.2329	0.8601	0.3201
		Kosmos-2 0.0118	0.2103	0.1802	0.8708	0.3183
		MiniGPT4 0.0013	0.1237	0.1671	0.8108	0.2757
		GPT4 0.0033	0.1684	0.208	0.8607	0.3101
		Gemini 0.0032	0.1973	0.2478	0.8425	0.3227
Stage 4	LEFT IMAGE DESCRIPTION	LLaVA 0.0062	0.1617	0.2327	0.8617	0.3156
		Kosmos-2 0.0119	0.2284	0.2643	0.8769	0.3454
		MiniGPT4 0.0061	0.1182	0.1806	0.8378	0.2857
		GPT4 0.0081	0.1427	0.2457	0.8554	0.313
		Gemini 0.0145	0.2714	0.2714	0.8885	0.3615
Stage 5	RIGHT IMAGE DESCRIPTION	LLaVA 0.0032	0.1567	0.23	0.8603	0.3125
		Kosmos-2 0.0083	0.2092	0.2499	0.8738	0.3353
		MiniGPT4 0.0031	0.1092	0.1858	0.8231	0.2803
		GPT4 0.0044	0.133	0.2312	0.8509	0.3049
		Gemini 0.009	0.2608	0.2593	0.8894	0.3546
Stage 6	WHYFUNNY PROMPT	LLaVA 0.0097	0.1795	0.2189	0.8588	0.3167
		Kosmos-2 0.0093	0.1955	0.1943	0.8657	0.3162
		MiniGPT4 0.0028	0.1499	0.1534	0.835	0.2853
		GPT4 0.0027	0.1466	0.225	0.8498	0.306
		Gemini 0.0081	0.1874	0.2321	0.8565	0.321
Stage 7	LEFT IMAGE DESCRIPTION	LLaVA 0.0064	0.1598	0.2388	0.8621	0.3168
		Kosmos-2 0.0078	0.2269	0.2662	0.8756	0.3441
		MiniGPT4 0.0063	0.1098	0.1836	0.8242	0.281
		GPT4 0.0073	0.1357	0.2378	0.8544	0.3088
		Gemini 0.0233	0.2743	0.29	0.8869	0.3686
Stage 8	RIGHT IMAGE DESCRIPTION	LLaVA 0.0057	0.1709	0.2437	0.8646	0.3212
		Kosmos-2 0.0104	0.2093	0.2452	0.8738	0.3347
		MiniGPT4 0.0043	0.1152	0.1866	0.8342	0.2851
		GPT4 0.007	0.136	0.2309	0.8529	0.3067
		Gemini 0.0311	0.2856	0.2928	0.8923	0.3755
Stage 9	WHYFUNNY PROMPT	LLaVA 0.012	0.1809	0.2233	0.8596	0.319
		Kosmos-2 0.0114	0.2007	0.2092	0.865	0.3216
		MiniGPT4 0.0014	0.1494	0.1519	0.8327	0.2838
		GPT4 0.0024	0.1444	0.2211	0.8488	0.3042
		Gemini 0.0101	0.1875	0.2355	0.8579	0.3227

Table 6: Evaluation of Satire Understanding on images curated at different Stages of annotation of *YesBut*

		BLEU	ROUGE	METEOR	BERTScore
E	LLaVA	0.0094	0.1779	0.2337	0.8607
M		0.0113	0.1741	0.2235	0.8558
D		0.0167	0.1875	0.2434	0.8577
E	Kosmos-2	0.0121	0.2126	0.1841	0.8714
M		0.0089	0.1969	0.1540	0.8683
D		0.0128	0.1894	0.1588	0.8604
E	MiniGPT4	0.0014	0.1316	0.1696	0.8157
M		0.0012	0.1198	0.1412	0.7670
D		0	0.1284	0.1887	0.8240
E	GPT4	0.0028	0.1684	0.2091	0.8615
M		0.0080	0.1773	0.2124	0.8576
D		0	0.1355	0.1566	0.8498
E	Gemini	0.0028	0.1986	0.2507	0.8423
M		0.0031	0.1927	0.2232	0.8659
D		0.0170	0.1775	0.2512	0.7604

Table 7: Effect of annotation difficulty on Satirical Understanding Performance using the WHYFUNNY prompt across several SOTA VL Models (E - EASY, M - MEDIUM, D - DIFFICULT)

Table 7 shows the effect of annotation difficulty on Satirical Understanding Performance. We infer that except for MiniGPT4, no other model performs se-

mantically well (BERTScore) for difficult images. Also, in 12 out of 20 cases (5 VLMs x 4 metrics), VL Models fail to perform well for difficult images. Hence, there is a positive correlation between VLMs and Humans regarding what is difficult, especially from a semantic point of view.

		BLEU	ROUGE	METEOR	BERTScore
Y	LLaVA	0.0103	0.1843	0.2436	0.8612
N		0.0093	0.1720	0.2234	0.8591
Y	Kosmos-2	0.0129	0.2079	0.1801	0.8698
N		0.0109	0.2124	0.1803	0.8716
Y	MiniGPT4	0.0014	0.1354	0.1677	0.8093
N		0.0012	0.1257	0.1665	0.8121
Y	GPT4	0.0052	0.1786	0.2182	0.8653
N		0.0016	0.1594	0.1989	0.8567
Y	Gemini	0.0049	0.2056	0.2533	0.8364
N		0.0017	0.1902	0.2429	0.8479

Table 8: Effect of the presence of text in images on Satirical Understanding Performance using the WHY-FUNNY prompt across several SOTA VL Models (Y - Text is present in the image, N - Text is absent in the image)

Table 8 shows the effect of the presence of text in images on Satirical Understanding Performance. We see that in 15 out of 20 (5 VLMs x 4 metrics) cases, VL Models perform better on images with text vs. no text, suggesting that the absence of text in images makes it difficult to understand satire. This is supported by Hayashi et al. (2024).

Polos Metric

Stage of Annotation	Model	Polos
Stage-2	GPT4	0.4823
	Gemini	0.5278
	LLaVA	0.4623
	MiniGPT4	0.3359
	Kosmos-2	0.4462
Stage-3	GPT4	0.4122
	Gemini	0.4849
	LLaVA	0.3721
	MiniGPT4	0.3161
	Kosmos-2	0.4468
Stage-4	GPT4	0.4146
	Gemini	0.488
	LLaVA	0.3758
	MiniGPT4	0.311
	Kosmos-2	0.4476

Table 9: Evaluation of Satirical Understanding Performance across several SOTA VL Models using the WHY-FUNNY Prompt and the image-based metric Polos.

Table 9 evaluates the Satirical Understanding Performance across several SOTA VL Models us-

ing the image-based Polos Metric. We can infer that all SOTA VL Models fail to perform well on the Polos Metric.

Overall image descriptions (human-written and predicted by 5 SOTA Models)

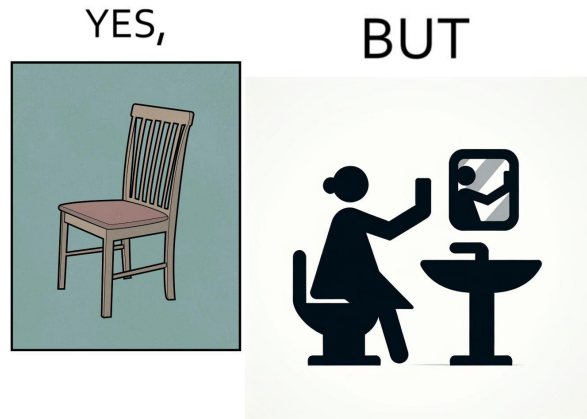


Figure 11: Example of a satirical image from YesBut

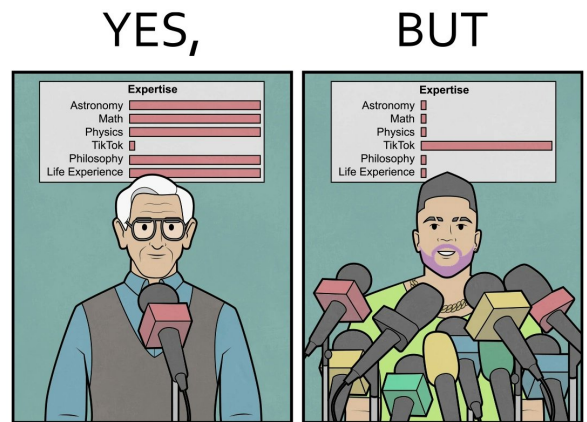
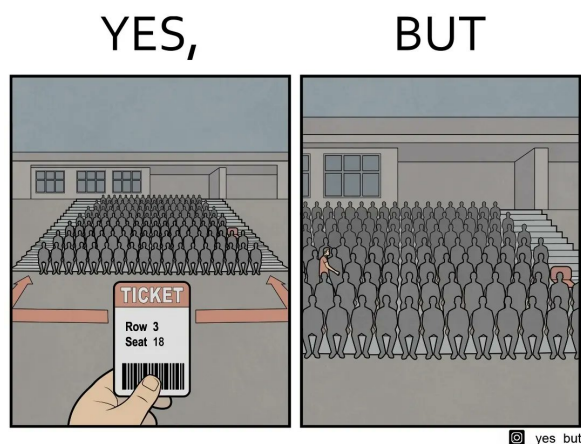


Figure 12: Example of a satirical image from YesBut

	Overall Image Description
Human-written	The image is ironical, as a woman is sitting by the sink taking a selfie using a mirror, while not using a chair that is actually meant for sitting.
LLaVA	The image is funny because it depicts a woman in a dress standing over a toilet in a bathroom, which is an unusual and unexpected setting for a woman dressed in a dress. This is a play on the common expectation that women should be dressed in more formal attire, such as a dress, when they are in a more polite or sophisticated environment, like a restaurant or a formal event. In this case, the woman's attire contrasts with the setting, creating a humorous and unexpected scene.
Kosmos-2	The image is funny/sarcastic because it depicts a woman sitting on a toilet, holding a checkmark, and saying "Yes, but." The image is meant to poke fun at the idea that women are often expected to perform tasks that are considered menial or unnecessary. The woman's actions are considered humorous because they are not typical for a woman to be doing, and the image is intended to make a point about the gender roles and expectations in the workplace.
GPT4	"The image you've provided plays on the typical format of public signs and presents a humorous twist. On the left side, you see a plain chair with the word ""YES"" above it, and on the right side, there's an icon-style figure of a person placing a voting ballot into a box with the word ""BUT"" above it. The satirical element comes from the juxtaposition of the unequivocal ""YES"" with the more ambiguous ""BUT,"" suggesting a contrast between the straightforwardness of sitting in a chair versus the complexity or potential reservations one might have when casting a vote. It could be interpreted as a commentary on"
MiniGPT4	[TXT] "A cartoon depicting a person sitting at a desk with a toilet in the background. The person is holding a magnifying glass up to their face and looking at their reflection in the toilet bowl. The caption reads "yes, but. "
Gemini	The image is funny and satirical because it's a commentary on the way social media makes us feel. The woman in the image is sitting on a toilet but taking a selfie with a weird angle to try and make her look good for social media. In reality, we all know that no one looks good when taking a selfie on the toilet. The image is a reminder that social media is not always a true reflection of reality and that we should be careful not to compare ourselves to others.

Table 10: Overall Image Descriptions (human-written and predicted by 5 SOTA Models) corresponding to Figure 11



16893

Figure 13: Example of a satirical image from *YesBut*

Tables 10, 11, and 12 contain the overall image descriptions (human-written and predicted by 5 SOTA Models) corresponding to Figures 11, 12, and 13 respectively. We perform the following qualitative analysis on these 3 images as follows -

- Table 10 shows that no model gives correct reasoning behind why Figure 11 is ironical, and each model makes different mistakes, For instance, GPT4 makes a blatant mistake of describing the right hand sub-image as a person placing a vote into a box. This shows the *inability of SOTA VL Models to recognize objects properly when there is a mixture of artistic styles.*

	Overall Image Description
Human-written	The image is satirical because it shows that people with more tiktok expertise are treated more importantly than the ones who are expert in all the other areas but tiktok. Here, the number of microphones a man speaks into is indicative of his importance.
LLaVA	The image is funny and satirical because it features a cartoon drawing of a man with a purple mohawk and glasses, surrounded by microphones, and the phrase "Expert, Astronomy" is displayed above him. This juxtaposition creates a humorous and ironic situation, as the man's appearance and occupation seem unrelated or even contradictory. The image likely intends to make a commentary on the absurdity of certain labels or assumptions people might make based on appearance or stereotypes.
Kosmos-2	The image is funny because it depicts an older man sitting in front of a row of microphones, asking a question to a young man who is wearing a mask. The older man is wearing glasses and has a beard, while the young man is not wearing a beard. The scene is set in front a podium, and the older man and the young boy are both wearing masks. This humorous scene is meant to poke fun at the idea that people often assume that older people are wiser
GPT4	The image is humorous because it shows an exaggerated contrast between two individuals' expertise levels across various disciplines, with one having significantly more experience and knowledge in each area. The satirical point lies in suggesting that sometimes people with lesser expertise might present themselves as equally or more knowledgeable, often leading to comical or ironic situations.
MiniGPT4	? [INST
Gemini	The image is funny because it shows how people with a lot of expertise in a particular subject will talk about it in a very technical and complicated way, using a lot of jargon, while people with less expertise in the same subject will talk about it in a more general and relatable way.

Table 11: Overall Image Descriptions (human-written and predicted by 5 SOTA Models) corresponding to Figure 12

- Figure 12 shows that society views people as worthy today based on social media presence, rather than knowledge (worth is represented by number of microphones). Only GPT4 gives a close-to-correct reasoning. This shows the *inability of SOTA VL Models to correlate objects in the image* (in this case, the number of microphones) to *societal constructs* (in this case, worth).
- No VLM is able to decipher Figure 13, where the person looking for the assigned seat in a nearly-filled audience takes path of maximum resistance instead of going from the other side. This shows that *SOTA VL Models are unable to understand miniature sketches of people/objects, as well as numbers* (the row, seat number here)

tions generated by 5 SOTA Models with ones written by human annotators based on human evaluation.

F Summary and Conclusion

Figure 14 compares the overall image descrip-

	Overall Image Description
Human-written	The image is funny, as the person has a ticket of a seat that is on the leftmost end of the gallery. However, the person has entered from the other end, and is going through a sea of seated people to get to the vacant seat. This would cause inconvenience to the people in the row, and would probably take a longer time to reach the seat.
LLaVA	The image is funny and satirical because it shows a person holding a ticket to a stadium with a large number of empty seats. The caption "Yes, but" is displayed above the image, suggesting that despite the empty seats, the event is still happening. This contrasts with the actuality of the event being poorly attended, which is further emphasized by the person being the only one in the image. The humor comes from the irony of the situation and the contrast between the expectation of a crowded event and the actuality of an empty stadium.
Kosmos-2	The image is funny because it depicts a theater full of people, but it's also satirical because it'll make you think about the audience's behavior during a movie. In the image, a person is holding a sign that says "Movies: Sold Out" and is pointing at the empty seats. This suggests that the theater is full, but the audience is not showing up. The person is trying to make a point about the lack of interest in the movie, and the empty chairs are a reflection of that.
GPT4	This image plays on the common frustration of attending an event with assigned seating, only to find that your seat is directly behind a pole or obstruction, blocking the view. It humorously exaggerates the issue by showing an entire audience perfectly aligned except for one obstructed seat, suggesting a comically unfortunate luck that the ticket holder has the only bad seat in the house.
MiniGPT4	Why is this image funny/satirical Title: Why is this image funny/satirical Author: AnonyMousse Date: March 22, 2023 Tags: comedy, satire Comments: The image shows a graph with two lines, one labeled "yes" and the other labeled "but". The graph appears to show a comparison between two different statements or ideas, but it is unclear what they are. The title and caption suggest that the image is meant to be humorous or satirical, but it is not clear
Gemini	The image is funny because the person holding the ticket looks very pleased with their seat assignment, while the image shows that there are plenty of open seats and the entire row is empty except for that one person.

Table 12: Overall Image Descriptions (human-written and predicted by 5 SOTA Models) corresponding to Figure 13

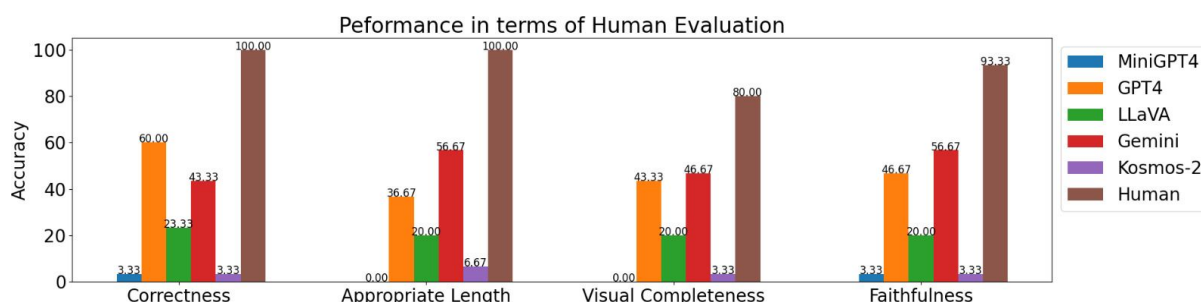


Figure 14: Results of Human Evaluation on the Satirical Image Understanding Task