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

IIT Kharagpur Univ. of Massachusetts Haldia Institute of Tech.

EMNLP 2024

Background

- **Satire in Images:** Satirical images blend humor and irony, making them hard to interpret
- **Research Gap:** Few studies focus on full satire detection in images
- **Model Challenges:** VL models struggle with satire due to complex visual-textual cues

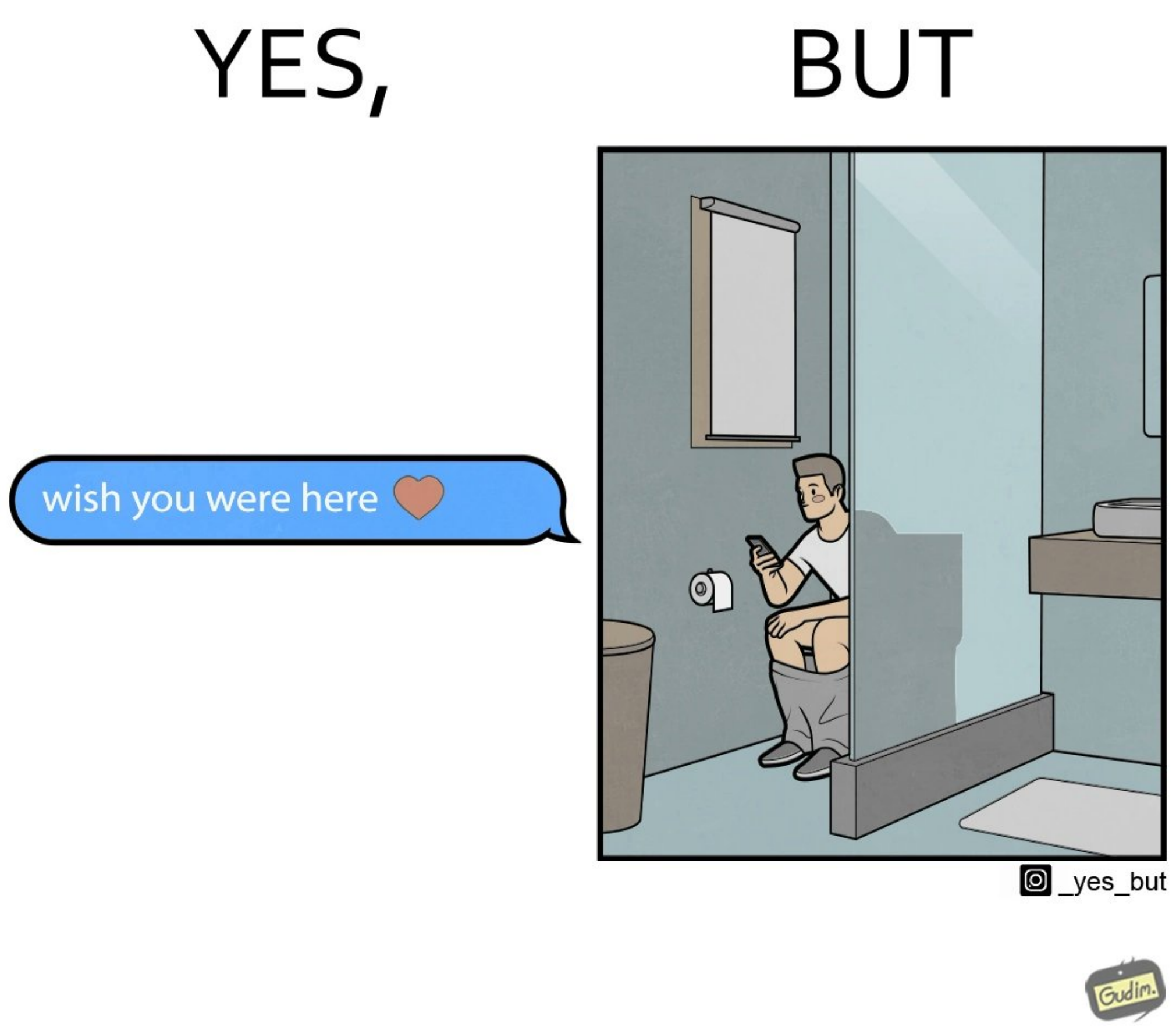
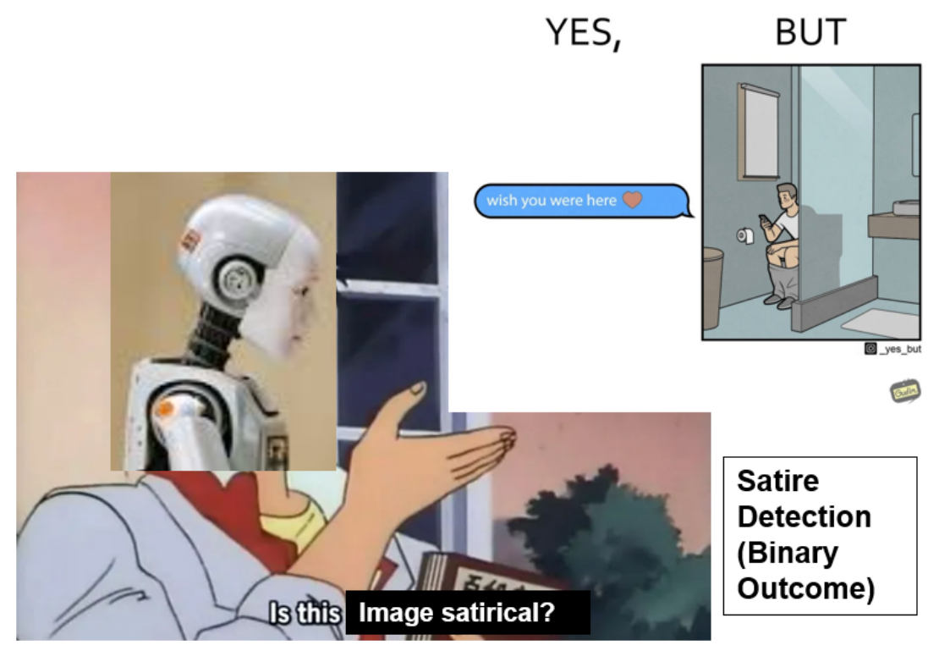


Figure: Satire conveyed through a social media image

Okay, Yes, But what is YesBut?

- **YesBut Dataset**: A one-of-a-kind multimodal dataset with 2,547 images, containing both satirical and non-satirical images, enriched with diverse artistic styles!
- **Challenges**: We introduced unique tasks like Satirical Image Detection, Satirical Image Understanding, and Satirical Image Completion—each pushing the boundaries of current Vision-Language (VL) models.
- **Benchmarking Results**: Even cutting-edge VL models struggle with our tasks, showing the complexity of understanding irony, humor, and societal satire!

Tasks Evaluated



Satire detection

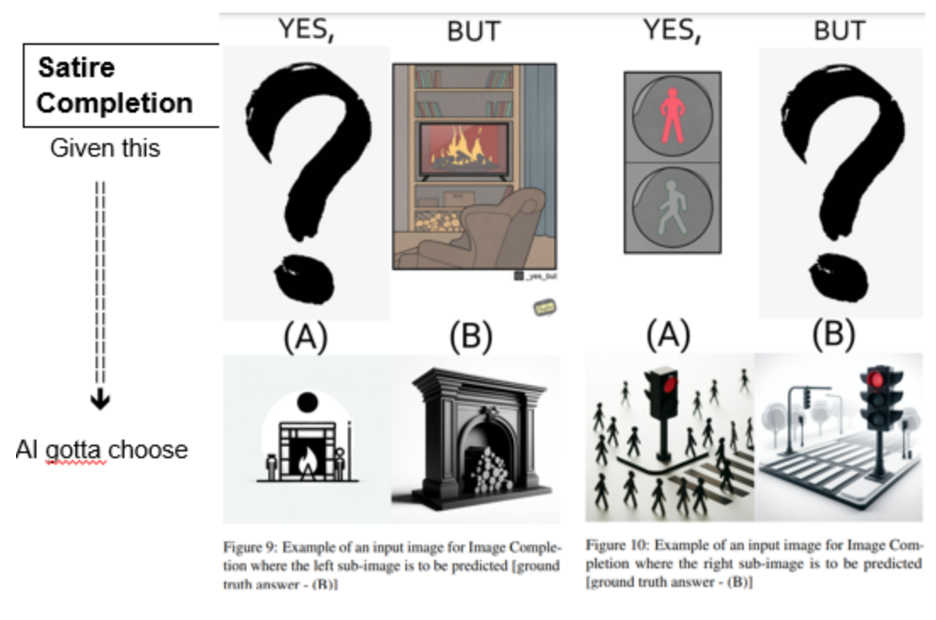
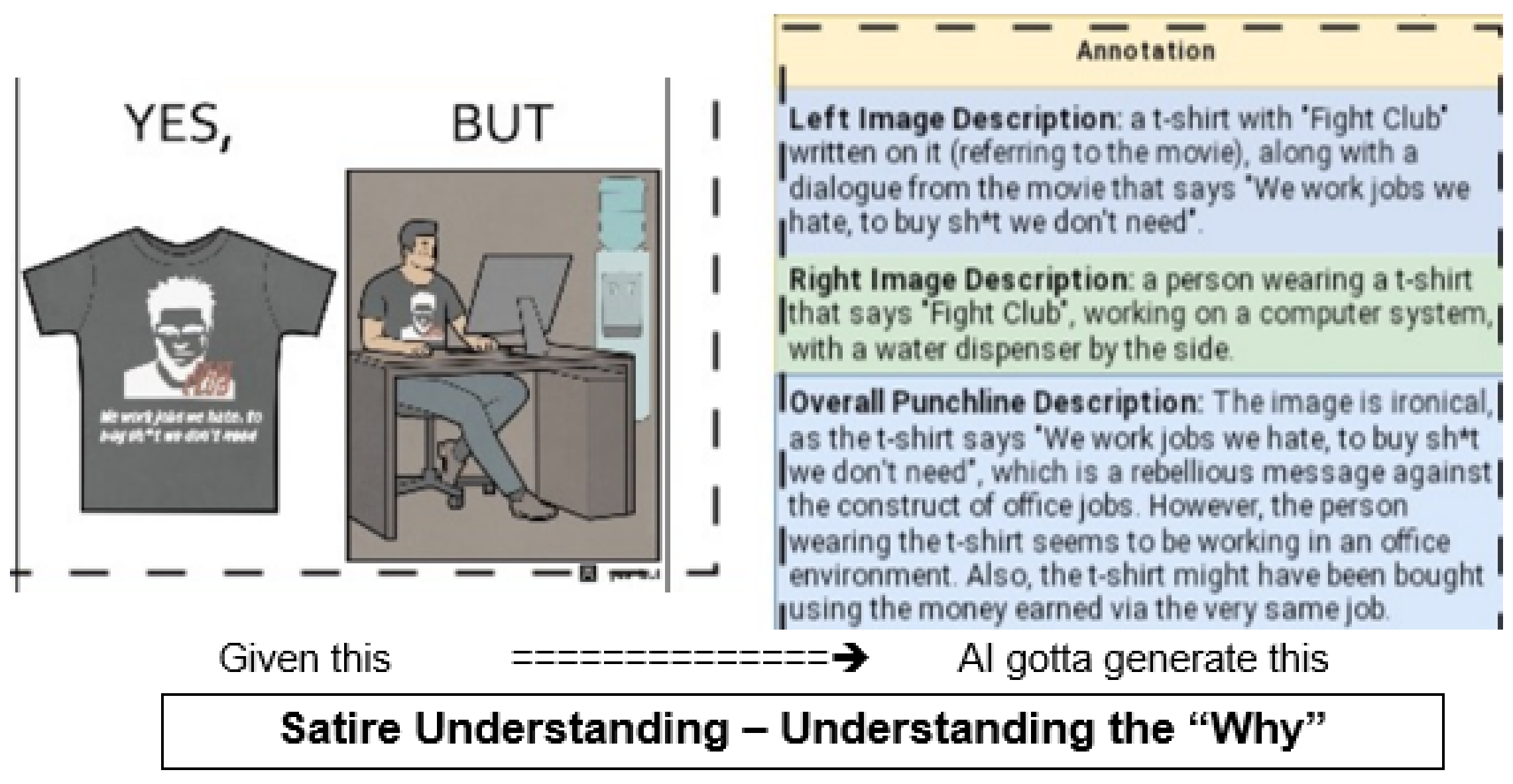


Figure 9: Example of an input image for Image Completion where the left sub-image is to be predicted [ground truth answer - (B)]

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

Satire Completion



Left Image Description: a t-shirt with 'Fight Club' written on it (referring to the movie), along with a dialogue from the movie that says 'We work jobs we hate, to buy sh*t we don't need'.

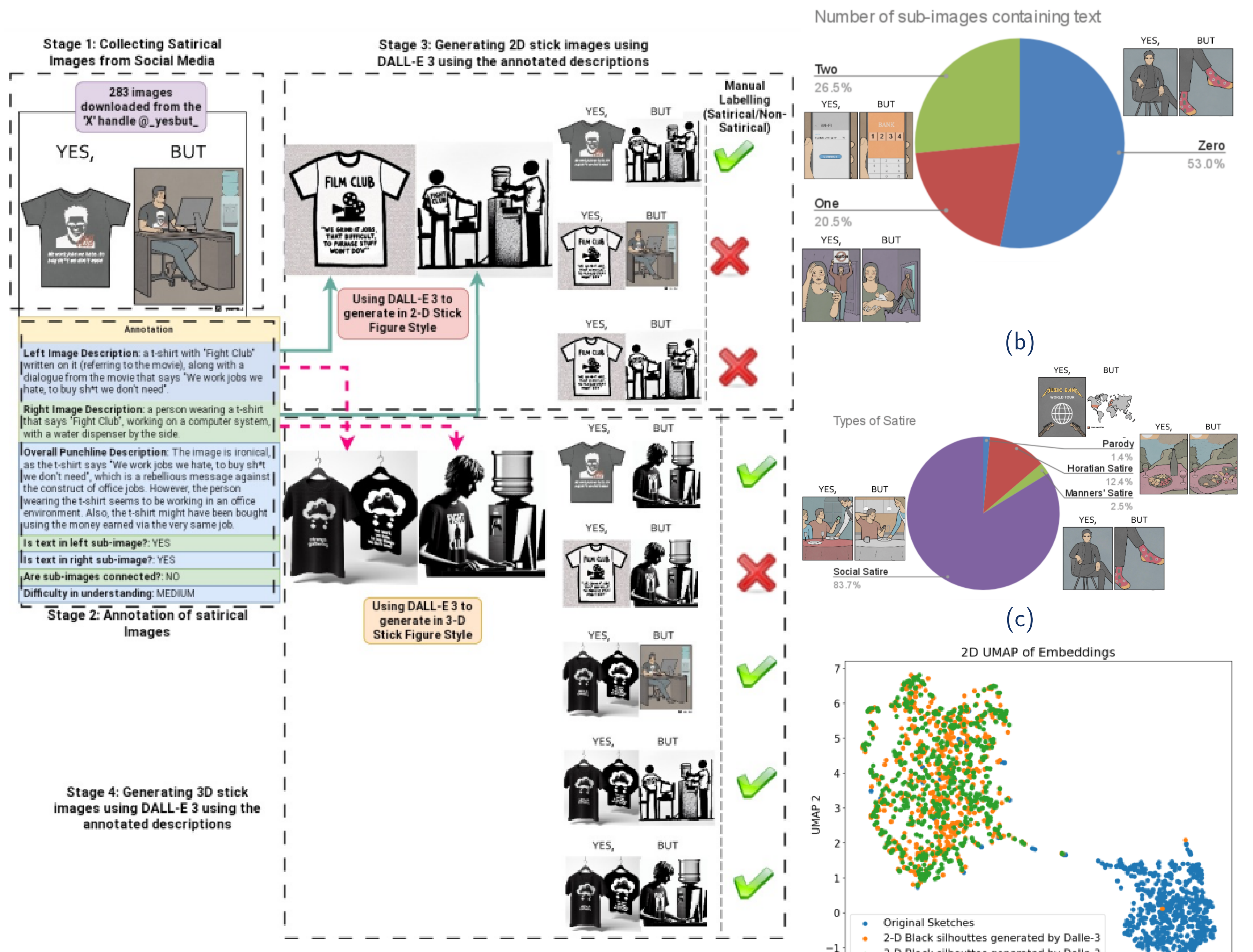
Right Image Description: a person wearing a t-shirt that says 'Fight Club', working on a computer system, with a water dispenser by the side.

Overall Punchline Description: The image is ironical, as the t-shirt says 'We work jobs we hate, to buy sh*t we don't need', which is a rebellious message against the construct of office jobs. However, the person wearing the t-shirt seems to be working in an office environment. Also, the t-shirt might have been bought using the money earned via the very same job.

Satire Understanding – Understanding the "Why"

Satire Understanding

How was YesBut curated?


Stage 1: Collecting Satirical Images from Social Media
283 images downloaded from the 'X' handle @_yesbut_
YES,
BUT
Annotation
Stage 2: Annotation of satirical Images
Stage 3: Generating 2D stick images using DALL-E 3 using the annotated descriptions
Using DALL-E 3 to generate in 2-D Stick Figure Style
Manual Labelling (Satirical/Non-Satirical)
FILM CLUB
Using DALL-E 3 to generate in 3-D Stick Figure Style
Stage 4: Generating 3D stick images using DALL-E 3 using the annotated descriptions
✓
✗
Number of sub-images containing text
Two 26.5%
One 20.5%
Zero 53.0%
Types of Satire
Parody 1.4%
Horatian Satire 12.4%
Manners' Satire 2.5%
Social Satire 83.7%
2D UMAP of Embeddings
UMAP 1
UMAP 2
Original Sketches
2-D Black silhouttes generated by Dalle-3
3-D Black silhouttes generated by Dalle-3
(b)
(c)

Left Image Description: a t-shirt with 'Fight Club' written on it (referring to the movie), along with a dialogue from the movie that says 'We work jobs we hate, to buy sh*t we don't need'.

Right Image Description: a person wearing a t-shirt that says 'Fight Club', working on a computer system, with a water dispenser by the side.

Overall Punchline Description: The image is ironical, as the t-shirt says 'We work jobs we hate, to buy sh*t we don't need', which is a rebellious message against the construct of office jobs. However, the person wearing the t-shirt seems to be working in an office environment. Also, the t-shirt might have been bought using the money earned via the very same job.

Is text in left sub-image?: YES

Is text in right sub-image?: YES

Are sub-images connected?: NO

Difficulty in understanding: MEDIUM

(a) 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

(d) 2D UMAP Representations of CLIP Image representations of YesBut sub-images

Figure: Left: Our annotation pipeline. Right: Distribution of satirical images based on content and annotated descriptions, and UMAP representations.

Experiments and 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

	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

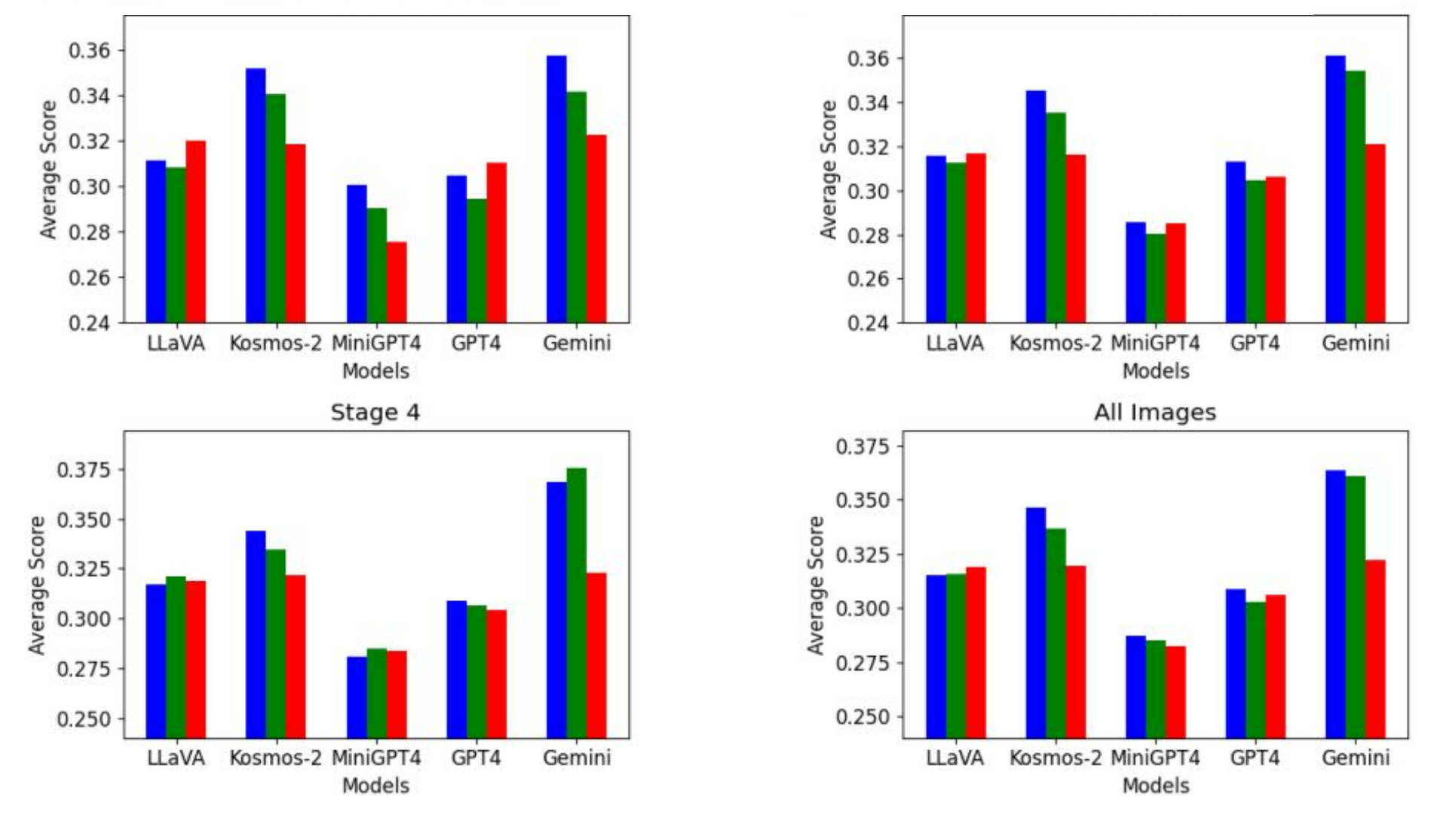


Figure 5: Evaluation of Satirical Image Understanding Capability using multiple VL models at different stages (Stages 2, 3, 4) of annotation of *YesBut*, as well as, for all *YesBut* images

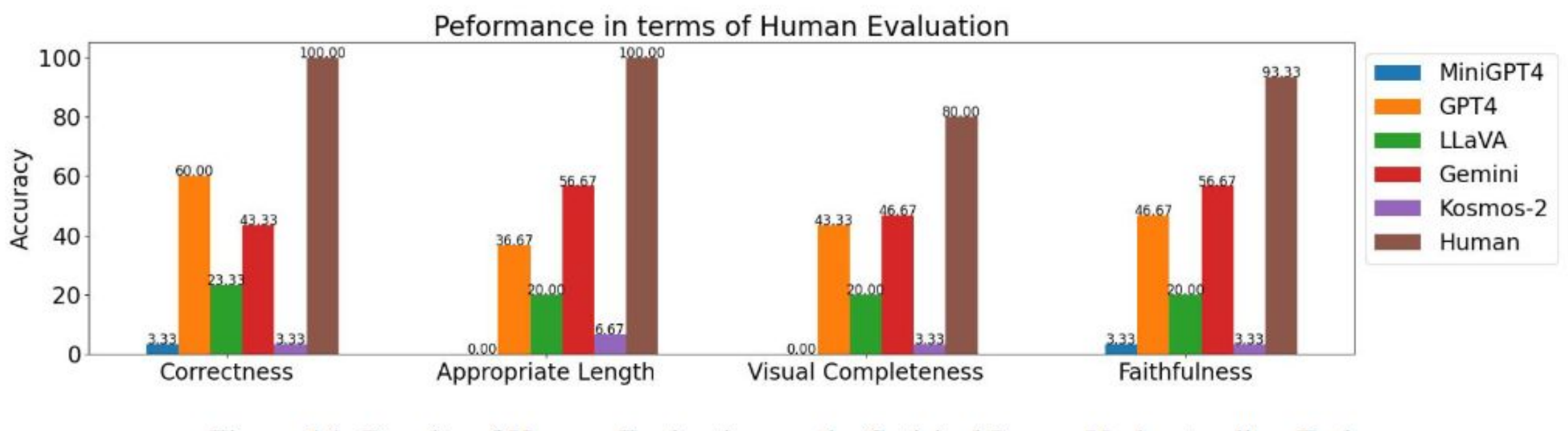


Figure 14: Results of Human Evaluation on the Satirical Image Understanding Task

Results on Satire Detection, Understanding, and Comprehension on the YesBut Dataset

For Further Information

Paper Link YesBut Dataset Code Base Link

Arxiv: https://arxiv.org/abs/2409.13592