HumorDB: Can AI understand graphical humor?

Supplementary Material

7. Appendix

7.1. Training details

Models were trained using the Adam optimization algorithm with weight decay. We conducted a hyperparameter grid search across learning rates in the set 0.01, 0.001, 0.0001, 0.00001, batch sizes in the set 4, 8, 16, and weight decay parameters in the set 0.1, 0.01, 0.001. Model training proceeded for a fixed number of 10 epochs, with periodic checkpoints. For the final evaluation on the unseen test set, we used the model iteration exhibiting optimal performance on the validation set. We used cross-entropy loss for the Binary classification and Comparison tasks, while mean square loss was used for the Regression task. For all architectures except GPT-4o and Gemini-Flash, we fine-tuned the models. For LLaVA, we performed LoRA fine-tuning instead of full fine-tuning. Due to the uneven distribution of range ratings, we employed a sampling strategy that grouped images into bins according to their funniness ratings. This allowed us to randomly select a balanced number of images from each bin for every training epoch, ensuring a uniform distribution of sample images across all ratings in the training set. We applied the same sampling strategy for the Comparison task, which also contained a slightly uneven distribution.

Models were trained using the Adam optimization algorithm with weight decay. A hyperparameter grid search was conducted across learning rates in the set $\{0.01, 0.001, 0.0001, 0.00001\}$, batch sizes in the set $\{4, 8, 16\}$, and weight decay parameters in the set $\{0.1, 0.01, 0.001\}$. Model training proceeded for a fixed number of 10 epochs, with periodic checkpoints. For the final evaluation on the unseen test set, we used the model iteration exhibiting optimal performance on the validation cohort. We used cross-entropy loss for the Binary classification and Comparison tasks while mean square loss was used for the Regression task. For all architectures except GPT-4o and Gemini-Flash, we fine-tuned the models(Pretraining details are mentioned in Section 4.1). For LlaVA we did lora fine-tuning instead of full fine-tuning. To ensure statistical robustness, each experiment was conducted 5 times for all the experiments except for GPT-4o and Gemini-Flash which were run only one time.

Most of the experiments were run on 4 Nvidia GeForce RTX 2080 Ti GPUs which were part of an internal cluster. However, for training LLaVA and some models for the Comparison task, we used an Nvidia A100 GPU.

7.2. External assets used

We utilized the following assets: The LLaVA repository (Apache-2.0 license) [24, 25], PyTorch [2], huggingface transformers (Apache-2.0 license) [40], and huggingface accelerate (Apache-2.0 license) [12].

Additionally, for the images collected from the internet we provide reference links in the repository.

7.3. Attention maps

We examined the attention maps using the attention rollout technique [1] on the ViT-Huge model [10]. This helped us understand whether the models focused on the actual humorous parts of images or other biases in the dataset. The attention maps may help to better understand how the models classify the images and identify potential shortcomings (**Fig. 11**).

As an example, consider the case of **Fig. 1**. The attention maps for the vit huge model are shown in **Fig.11**. The model fails to pay attention to the most humorous part of the image (the phone, black rectangle), which is critical to assess whether the image is funny or not. Therefore the model is not able to correctly classify both images.

7.4. Crowdsourcing details

There were 850 participants: 200 for binary task, 215 for the range task and 435 for the comparison task. The interfaces used by the participants for the three tasks are shown in **Fig. 12**. The generic instructions given for all tasks were:

- **Binary Task:** Please rate if the image is funny or not.
- **Range Task:** Please rate the degree of funniness of the image on a scale from 1 (not funny) to 10 (very funny).
- **Comparison Task:** Please indicate which of the two images is funnier.
- For funny images write a word that makes the image funny, for not funny images, write a word about the most prominent feature of the image.
- The time required to rate all the images is approximately 9-11 minutes
- Only click on the rating buttons once, and wait till the next image loads (maximum 1 second), a message will show you when the next image is being processed.
- Please do not refresh the page. You will lose progress and will have to start again.
- There are 100-120 images in this survey.
- At the end of the survey, we will provide you a code, please store it and use it appropriately to get the reward.
- Click the button below to begin.

Some participants were discarded due to reliability and

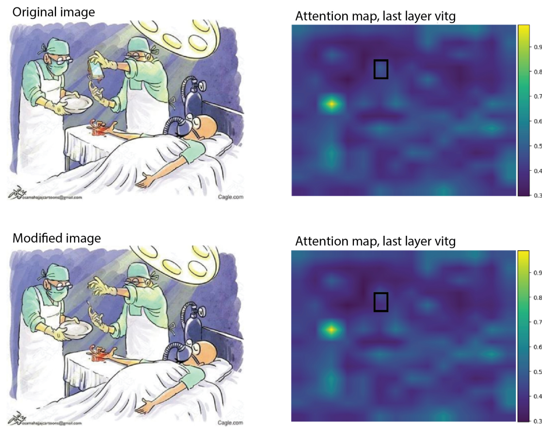
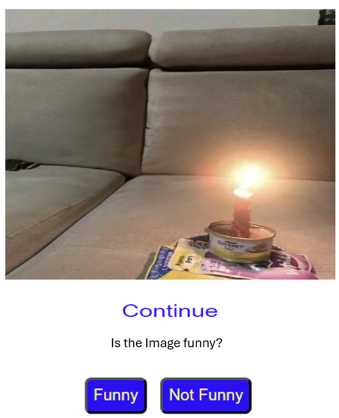
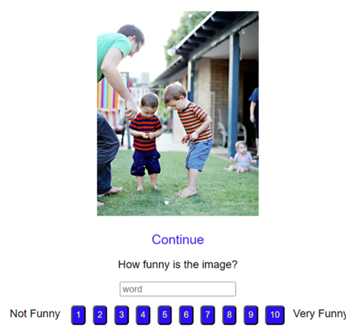


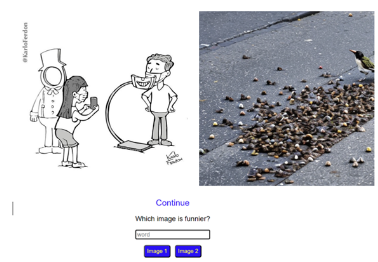
Figure 11. **Attention maps fail to capture elements critical to discern humor**. Attention maps based on the last layer of the vit huge model for the example images from **Fig. 1**. The black rectangle in the attention maps indicates the location of the phone. The maximum attention activation highlights the plate, which does not help distinguish between the original and modified images. Indeed, the model classified both images as funny.



(a) Binary Task Interface



(b) Range Task Interface



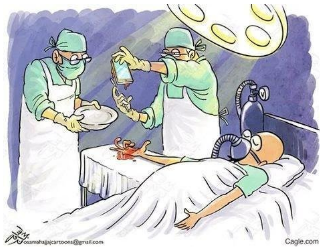
(c) Compare Task Interface

Figure 12. **Crowd sourcing interfaces for the three tasks.**

the amount of outliers in their responses as detailed in Section3.2. All participants were required to submit all questions in the survey for the response to count.

7.5. Multimodal models' answer explanations

For evaluating zero-shot performance of the large multimodal models we test them on testAllSet and testAllPairs sets. We do this to compare the performance of these models with the other fine tuned models on the same test set. The two variants of prompts in binary task as mentioned in section 4.1 were: : (i) "Is the image funny?", and (ii) "Is the image not funny?". The performance on both prompts were similar so we reported the average for the results. In addition for succinct explanations in a particular format for the figures like Fig. 8, we add a suffix 'start answer with yes/no then explain'. In this section we present similar figures to Fig. 8 for Gemini-Flash and LlaVA on zero-shot prompting. We also mention the range ratings these models give for the images and the range rating prompt is mentioned in section 4.1. Gemini-Flash answer explanations are presented in **Fig. 13**. The answer explanation from LlaVA zero-shot are presented in **Fig. 14**.



(a) LLM response: yes. the humor lies in the fact that the patient's hand is being operated on with a smartphone. Range rating: 7

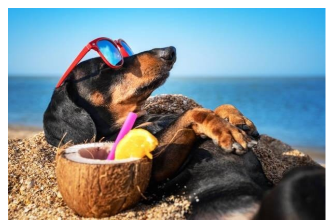
(b) LLM response: yes. the patient's hand is detached from their body. Range rating: 8

(c) LLM response: yes. the shadow of the women looks like two snakes. it is a dark and ironic humor. Range rating: 7

(d) LLM response: no. the image is not inherently funny. it depicts two people sitting on a couch, and there's no clear comedic element. Range rating: 3

(e) LLM response: no. it's a standard photo of a person speaking. there is nothing humorous. Range rating: 2

(f) LLM response: no. it is a picture of a beach with some people on it. Range rating: 1



(g) LLM response: yes. the dog is wearing sunglasses and enjoying a coconut drink. Range rating: 8

(h) LLM response: no. the image is of a child sleeping. there is no humor in the image. Range rating: 2

Figure 13. **Example Gemini-Flash binary classification, range ratings, and explanations.** The human binary and range ratings are as follows for the images: (a) *83%, 6.75*, (b) *14%, 5.70*, (c) *60%, 3.60*, (d) *33%, 3.00*, (e) *60%, 2.80*, (f) *0%, 5.57*, (g) *100%, 5.60*, (h) *33%, 4.86*.

7.6. Participants' Demographics

The demographics of the participants are described in the table **Tab. 4**.

8. Scoring for VLM explanation evaluation

The scoring process for all words and common words scores involved:

- Stemming words from human and VLM explanations using the spacy library.
- Matching synonyms using the nltk library.
- Marking an explanation as satisfactory if any stemmed word synonym from human raters is contained in the model explanation.

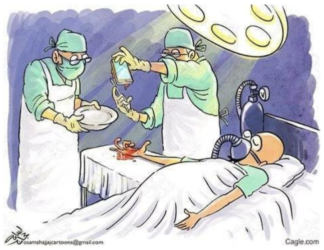
8.1. Results for automated VLM explanation evaluation

VLM explanations for correctly classified funny images showed varying alignment with human annotations (**Table 5**). Gemini provided the most accurate explanations (87.6% All Words Score, 74% Common Words Score), followed by GPT-4o and LLaVA.

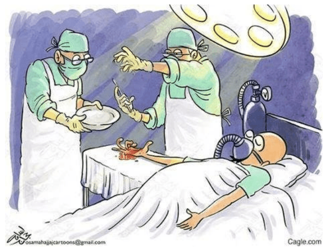
9. Attention Map Evaluation

To gain insight into what regions of an image the model relies on to infer humor, we analyzed attention maps in vision transformer models [1]. Specifically, we examined the attention on images from the test set. We used an *attention rollout* technique that iteratively multiplies the raw attention matrices (augmented by the identity matrix to include residual connections), producing a global attention map over image tokens. We then reshaped this map into a 2D image by mapping each token to its position in the image grid and upsampling it to the original resolution. Attention Rollout does not work for SwinV2 because of its hierarchical attention architecture.

After computing the final attention map, we optionally convert this attention to a *segmentation-like* mask. We smooth the attention map, detect local peaks, and ex-



(a) LLM response: Yes. The humor comes from the juxtaposition of a serious medical situation with the casual and nonchalant behavior of the doctors, who seem more interested in using their phones than focusing on the patient. Range rating: 6



(b) LLM response: Yes, the image is intended to be humorous. It depicts a cartoon of a patient in a hospital bed with two doctors, one holding a tray of food and the other holding a tray of surgical instruments, both seemingly confused about which tray to use. Range rating: 6

(c) LLM response: Yes, the image is funny. It depicts two people sitting on a couch, seemingly engaged in a conversation, but the shadow on the wall behind them forms a shape that looks like a snake eating its own tail. Range rating: 5

(d) LLM response: Yes, the image is funny. It depicts two people sitting on a couch, but their bodies are mirrored, with their heads on the wrong sides of their bodies, creating a humorous and surreal visual effect. Range rating: 6

(e) LLM response: Yes, the image is funny. It depicts a cartoon of a person inside a box, seemingly being ”pulled out” by a hand using a toothpick. The humor comes from the absurdity and unexpectedness of the situation. Range rating: 6

(f) LLM response: Yes, the image is funny because it shows a hand pushing a button on a box that is labeled ”Emergency Button,” but instead of a typical emergency response, the box is filled with a bunch of pencils, which is an unexpected and humorous outcome. Range rating: 5

(g) LLM response:Yes, the image is funny because it shows a car with a large inflatable pool toy, specifically a white inflatable animal, strapped to its roof. This is an unusual and humorous sight because it’s not common to see such a large and whimsical item being transported in this manner. Range rating: 7

(h) LLM response: No. The image shows a car parked in a snowy environment, which is a common sight during winter. There is nothing inherently funny about the image; it appears to be a typical winter scene. Range rating: 2

Figure 14. **Example LlaVA binary classification, range ratings, and explanations.** The human binary and range ratings are as follows for the images: (a) *83%, 6.75*, (b) *14%, 5.70*, (c) *60%, 3.60*, (d) *33%, 3.00*, (e) *60%, 2.80*, (f) *0%, 3.00*, (g) *50%, 5.25*, (h) *0%, 1.00*.

pand those peaks into connected blobs whose intensities exceed a fraction of the peak value. This yields “peak-centered blobs,” each one capturing a region of high attention. Specifically:

Peak-Centered Blob Detection. We developed a systematic approach to identify and analyze significant attention regions:

1. Applied Gaussian smoothing (σ=2.0) to reduce noise in attention maps
2. Detected local maxima using peak_local_max with rela-

Demographic Category	Percentage
Age	
20-29	47.0%
30-39	30.3%
40-49	13.8%
50-59	5.2%
0-19	2.1%
60-69	1.6%
Education	
Undergraduate	56.3%
Postgraduate	29.5%
High School	14.2%
Gender	
Male	52.6%
Female	46.6%
Non-binary	0.8%
Nationality	
United States	54.0%
South Africa	9.0%
Other (41 total)	37.0%

Table 4. Demographic Data

Model	All Words Score	Common Words Score
Gemini	0.876	0.74
GPT4o	0.840	0.724
LLaVA	0.74	0.50

Table 5. VLM explanation accuracy compared to humans annotations

tive threshold 0.5

3. Generated connected components around peaks using intensity-based thresholding
4. Filtered small regions (¡50 pixels) to focus on significant attention areas

Since our dataset contains minimally contrastive pairs, one can also approximate the "true" humorous region by taking the pixel-level difference between the original (funny) and modified (not funny) image in each pair.

Difference Map Generation. For paired images (funny/not-funny versions), we:

1. Computed pixel-wise differences across RGB channels
2. Applied Gaussian smoothing (σ=1.0) to the difference map
3. Identified connected components with significant differences (threshold ¿ 0.1)
4. Generated binary masks highlighting modified regions

Evaluation Metrics. We evaluated attention map quality using three primary metrics:

- Recall: Proportion of ground truth funny regions captured by attention
- Strict Box Containment: Binary measure of whether attention stays within ground truth regions
- Outside Box Ratio: Proportion of attention allocated outside ground truth regions

We could only do this for the ViT and DinoV2 models as attention rollout cannot be directly be applied to SwinV2 type models which is the only other vision only transformer.

10. Logit Attribution Details

Transformer-based NLP models can be probed using *logit attribution* to evaluate how each layer's hidden state contributes to the final output. Here, we adapt the same concept to our ViT-based model, treating each ViT block (and the final classifier) as a "layer."

Layer-wise Analysis. For each layer in the transformer models (ViTs and DINOv2):

1. Extracted hidden states from each transformer layer
2. Applied the classification head to each layer's output regarding that layer as the "last layer" before layernorm and classification head.
3. Computed softmax probabilities for binary humor classification
4. Measured classification accuracy using each layer's predictions

11. Future Directions.

Beyond expanding the cultural breadth of **HumorDB**, we envision several promising research avenues:

- **Interpreting Vision Models.** Recent progress in interpreting transformer-based language models can inform the study of multimodal and vision-only architectures. HumorDB's carefully constructed minimal pairs provide an ideal testbed for *mechanistic interpretability* in vision, offering the kinds of subtle input differences that are otherwise hard to curate for images.
- **Expanded Evaluation Metrics.** Novel benchmarks could explore multi-modal inputs (e.g., text, video, audio) to capture richer humor contexts. This would help evaluate how well models integrate multiple information streams to detect incongruities or comedic timing.
- **Personalized Humor.** Because individual comedic tastes vary, it would be valuable to test models on how well they adapt to personal preferences. Such personalization could move beyond majority voting to reflect diverse humor perceptions.
- **Cultural and Linguistic Diversity.** Truly universal humor comprehension requires sampling across diverse cultural and linguistic backgrounds. Curating a broader spectrum of comedic tropes—slapstick, satire, wordplay,

and so on—will challenge models to generalize beyond Western-centric contexts.

12. In-Lab Validation of Crowdsourced Data

To validate our primary crowdsourced annotations, we conducted a separate in-lab (non-crowdsourced) study. Participants provided ratings for 400 images, yielding $\geq$5 ratings per image. We found high correlation between the in-lab and online data for both the binary task ($\rho = 0.78$) and the range task ($\rho = 0.72$), which reinforces the reliability of our HumorDB.

13. Image categories results

The results of models on various image categories are described in **Tab.** 6.

Model Name	Photos	Photoshopped	Sketches	Cartoons	AI-Gen
dinov2 large	59 ± 3	60 ± 1	47 ± 1	59 ± 2	51 ± 2
vit huge	64 ± 3	61 ± 2	48 ± 2	62 ± 2	52 ± 2
vit large	58 ± 2	58 ± 2	47 ± 1	59 ± 2	51 ± 2
swin2 large	61 ± 2	60 ± 2	47 ± 1	60 ± 2	52 ± 2
convnext large	57 ± 2	57 ± 1	46 ± 1	57 ± 2	50 ± 0
vitg 14	72 ± 3	70 ± 3	51 ± 2	68 ± 3	53 ± 2
resnet152	56 ± 1	55 ± 2	46 ± 1	56 ± 2	50 ± 1
LLaVA (Zero-Shot)	63 ± 5	66 ± 4	46 ± 1	52 ± 2	66 ± 3
LLaVA (fine-tuned)	72 ± 2	76 ± 2	54 ± 2	65 ± 2	69 ± 3
LLaVA (words fine-tuned)	79 ± 2	83 ± 1	54 ± 1	69 ± 2	73 ± 1
BLIP (fine-tuned)	59 ± 1	59 ± 2	48 ± 1	59 ± 2	52 ± 2
BLIP (words fine-tuned)	63 ± 2	66 ± 2	49 ± 1	61 ± 2	55 ± 1
GPT-4o (Zero-Shot)	75	69	50	58	76
Gemini-Flash	73	84	53	74	82

Table 6. Binary Classification Results on various image types present in the dataset.