MUNCH: BENCHMARKING MULTIMODAL HUMOR UNDERSTANDING IN VISUAL LANGUAGE MODELS THROUGH MEME COMPREHENSION

Anonymous authors
Paper under double-blind review

ABSTRACT

A meme template is a reusable image that stands for a situation or stance, and a caption turns it into a joke by filling that slot with a new case. Recognizing which caption fills a template is an abstract form of multimodal humor understanding that evaluations of vision–language models rarely isolate. We introduce MUNCH, a benchmark of 123K Reddit memes whose captions have been inpainted out and recast as four-way multiple-choice questions, with test partitions that hold out templates and entire semantic clusters. Across 23 zero-shot multimodal large language models (MLLMs), the strongest scores 14 to 21 points below our human performance estimate, although these models describe images fluently. A contrastive model fine-tuned on MUNCH reaches the human performance estimate on templates seen in training (86%), but its gains shrink as templates and their meanings move away from training, and on held-out semantic clusters it scores 13 points below the estimate (62%). Our results show that multimodal image–text humor understanding is challenging for otherwise highly-performant models but can be learned with training.

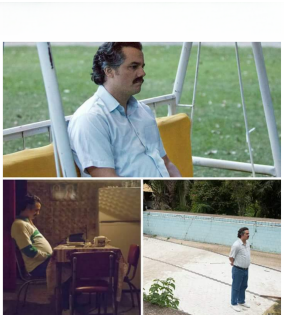
1 INTRODUCTION

Computational humor asks whether a model can recognize what makes something funny, which often requires recovering an implicit intent (Oliveira et al., 2016). Humor carried jointly by an image and its text is harder still, because neither channel holds the joke alone. Memes are the everyday form of image–text humor. A meme template is a reusable image that stands for a situation or stance, and each caption turns the template into a joke by filling that slot with a new case (Nissenbaum & Shifman, 2018). Recognizing the joke means representing the semantics a community has attached to an otherwise arbitrary image, and then reading the intent a writer signals by pairing one particular caption with the template.

Computational humor has been studied most thoroughly in text, where puns, one-liners, and humorous headlines offer clean supervision (Weller & Seppi, 2019; Horvitz et al., 2024). Multimodal humor has been approached mainly through professionally produced cartoons and comics (Hu et al., 2024; Ryan et al., 2025), most closely by Hessel et al. (2023), who ask models to match New Yorker Caption Contest cartoons to their winning captions. While memes have been studied at scale across natural language processing and computer vision, most work has focused on offensiveness, hate speech, and open-ended interpretation rather than for what makes an image and a caption funny together (Jha et al., 2024; Agarwal et al., 2024; Hwang & Shwartz, 2023). Whether models understand the humor of text–image pairings that ordinary users produce remains open.

We introduce MUNCH (**M**eme **U**nderstanding for **N**atural-language **C**omputational **H**umor), a benchmark of 123,118 Reddit memes drawn from 886 templates in 275 semantic clusters. Each instance removes a meme's original caption from its image and asks a model to pick, out of four options, the caption the meme's author wrote (Figure 1). Because the caption is gone from the image and every distractor was written for a template that means something else, answering requires reading a template's meaning off the image rather than off the option text. Three test partitions withhold progressively more of the training data—first the individual templates and then the semantic clusters

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



A: 6 year old me eating spinach after watching Popeye: where muscles

B: The guy who gave his phone for repair before quarantine be like:

C: When you play Minecraft in a park

D: When an Owen Wilson character says "wow"

Figure 1: A MUNCH instance. The model sees the template with its caption inpainted out and picks the caption a Reddit user wrote for it (answer B). The three distractors are captions from templates in other semantic clusters (meme categories with different meanings).

those templates belong to—so that accuracy reports how far a model's grasp of image–text humor generalizes beyond what the model was trained on (§3.3).

We ask whether current models represent text–image pairings well enough to recognize which caption makes a template funny, testing models both zero-shot and after fine-tuning on MUNCH. Models that never see the image score near 25% chance, and a prior built from a template's usual captions succeeds only where that template appeared in training. Even the strongest zero-shot model we evaluate, a closed model, scores 14 to 21 points below our human performance estimate, although these models describe images fluently (Li et al., 2025; Bai et al., 2025a). Fine-tuning closes that gap on familiar templates and on new templates whose meaning is familiar, far enough for a LoRA-tuned 7B open model to surpass every closed model that we test. The gains shrink sharply once the meaning of a novel image is itself unfamiliar, where the strongest system scores 13 points below the human performance estimate.

Our contributions are threefold. First, we release MUNCH, a benchmark whose partitions hold out templates and the meanings templates carry, with audits of both construction steps (§3). Second, we show that current models' multimodal humor understanding transfers to new templates carrying familiar meanings far better than to meanings a model has never seen (§4, §6). Third, we disentangle recognizing a template from understanding a pairing, using systems that never receive the image and priors built from a template's usual captions (§5).

2 RELATED WORK

Multimodal humor Benchmarks of multimodal humor ask whether a model can tell a funny image–text pairing from an unfunny one, and most draw their material from professionally produced or commissioned sources. The closest precursor to our work is Hessel et al. (2023), who evaluate models on New Yorker Caption Contest cartoons via multiple-choice caption matching, and a follow-up releases more than 250M human ratings of 2.2M contest captions as a benchmark for ranking and generating captions (Zhang et al., 2024a). Recent benchmarks extend image–text humor in several directions. YesBut (Hu et al., 2024) probes visual-contradiction humor in two-panel comics, PunchBench (Ouyang et al., 2025) targets multimodal punchlines while explicitly attacking language-shortcut bias, PixelHumor (Ryan et al., 2025) tests multi-panel sequencing, HumorDB (Jain et al., 2025) contrasts minimally edited funny and unfunny images, and v-HUB (Shi et al., 2026) tests video humor carried mostly by visual cues. Generation-oriented work pairs 2.9M images with humorous captions (Li et al., 2023b) and trains models for witty Oogiri responses, arguing that step-by-step reasoning is ill-suited to creative humor (Zhong et al., 2024). Earlier multimodal humor work pairs text with images (Chandrasekaran et al., 2016) or video (Patro et al., 2021), and text-humor recognition has been studied by Weller & Seppi (2019) and Horvitz et al. (2024) among others. These benchmarks all draw on material produced professionally or commissioned for the task, where a single author invented the humorous pairing. MUNCH instead measures humor in pairings that circulated on their own, so that what a model must recover is the convention a community attached to a template rather than one author's invention.

Figurative and cultural meaning in images A meme template carries its meaning by convention, which links MUNCH to benchmarks of non-literal and culture-dependent image understanding. Visual metaphor benchmarks ask models to interpret or retrieve images whose meaning differs from their literal content (Akula et al., 2023; Chakrabarty et al., 2023). II-Bench (Liu et al., 2024b) and DeepEval (Yang et al., 2024) cast image implication and figurative semantics as VQA, and Saakyan et al. (2025) cover humor as one of five figurative phenomena. WHOOPS! (Bitton-Guetta et al., 2023) and Visual Riddles (Bitton-Guetta et al., 2024) require commonsense and world knowledge, and on Visual Riddles the best model reaches 40% against 82% for humans. Abstract visual puzzles leave current MLLMs near chance (Jiang et al., 2024), and CulturalVQA and CVQA show that models answer questions about some cultures far better than others (Nayak et al., 2024; Romero et al., 2024). Meme templates are a convention of this kind, learned through exposure to a community's use of an image rather than from the image alone.

Meme corpora in vision and language Memes have been studied as multimodal media for tasks other than humor, including disinformation (Qu et al., 2022), hateful content (Kiela et al., 2020; Jha et al., 2024), and emotion or sentiment (Sharma et al., 2024; Mishra et al., 2023). The closest meme work is MemeCap (Hwang & Shwartz, 2023), which tasks models with generating meme interpretations from the image, title, and surrounding text, and its follow-up BottleHumor (Hwang et al., 2025), a humor-explanation method evaluated on MemeCap. Because the caption stays legible in those tasks, a model can succeed by describing a meaning the model is able to read; MUNCH asks the complementary question of whether a model can identify the caption that makes a template funny when neither a caption nor the surrounding post is available to read, which requires inferring the template's meaning itself. The MemeReaCon dataset (Zhao et al., 2025) pairs 1,565 Reddit memes with reader reactions, and Tanaka et al. (2022) train CLIP and ViLT for incongruity-based humor detection on 7.5K memes. MemeCLIP adapts CLIP to classify hate, stance, and humor in memes (Shah et al., 2024), and a concurrent template-centric resource annotates 33K memes from 50 templates and fine-tunes CLIP for meme–text retrieval (Deng et al., 2025).

3 THE MUNCH BENCHMARK

MUNCH turns in-the-wild memes into multiple-choice instances that test whether a model understands the humor of a text–image pairing. A meme is a template t paired with the caption c its author wrote for the template, and every template belongs to a semantic cluster $s(t)$ that groups the templates conveying the same meaning. An instance shows the template image with the caption inpainted out, written $\tilde{t}$, and offers four options $\{c, d_1, d_2, d_3\}$ in which each distractor d_j is a caption written for a template from a cluster other than $s(t)$. A model answers by choosing one option, and accuracy is how often the model chooses c. Further details are in Appendix B.

3.1 SOURCE DATA AND FILTERING

We curate MUNCH from the 3.8M memes in the SEMANTICMEMES dataset (Zhou et al., 2024), which provides the text of each meme and its bounding box from Optical Character Recognition (OCR). SEMANTICMEMES also supplies the key meme-interpretation structure that MUNCH is built on. Every meme carries a template ID, which is shared by memes with the same image format, and a cluster ID that groups templates whose semantics are interchangeable (e.g,. templates for expressing unpopular opinions). Because the cluster labels record which templates mean the same thing, drawing each distractor from a cluster other than the gold's sharply lowers the risk that a distractor is also a valid filler for the template (§3.2), and lets us hold out meanings rather than only images at test time (§3.3). We start from the 3.3M memes still downloadable in July 2024.

We filter the memes in six steps (Appendix B), keeping single-caption English memes, one meme per unique caption, and templates with ≥ 50 instances. Captions failing a dictionary check are re-OCRed with `idefics2-8b` (Laurençon et al., 2023; 2024) and kept if the new text passes; a manual check removes the 82 templates that are not a single text fill, leaving 127,204 memes from 886 templates.

Partition	Instances	Clusters	Templates
Training	82,908	254	636
Validation	13,689	270	881
$\mathcal{KT}$	13,836	249	630
$\mathcal{HT}$	6,247	38	179
$\mathcal{HC}$	6,438	21	71
Total	123,118	275	886

Table 1: Statistics for MUNCH. $\mathcal{HC}$ holds the fewest clusters and is the most concentrated of the three test partitions, with its largest cluster alone holding a third of its instances, so we report cluster-bootstrap intervals and macro-over-cluster accuracy alongside pooled accuracy.

3.2 INSTANCES AND DISTRACTORS

Each instance in MUNCH is curated to avoid potential shortcuts in evaluation. Question answering benchmarks risk having the correct answer inferred from the options alone (Gururangan et al., 2018; Poliak et al., 2018). Image–caption matching benchmarks now guard against this risk (Thrush et al., 2022; Hsieh et al., 2023; Li et al., 2024a; Chen et al., 2024; Tong et al., 2024)and we take the following steps to prevent shortcut learning. Later, §5 demonstrate the effectiveness of these efforts by the robustness tests that measure how much of a model's score survives when the image or the training data is withheld.

All original meme instances contain their caption, which must be removed so a model cannot read the answer from the image. We mask the OCR bounding box and fill the box with the LaMa inpainting model (Suvorov et al., 2022), producing one caption-free image $\tilde{t}$ per template t (Appendix C). Each instance pairs $\tilde{t}$ with four options $\{c, d_1, d_2, d_3\}$, of which only the gold caption c was written for t, and the options are shuffled so answer positions are balanced (Zheng et al., 2024; Liu et al., 2024a). Each distractor d_j is a caption written for a template drawn from a cluster other than $s(t)$.

Drawing distractors from other clusters gives each instance exactly one correct answer. A caption written for one template also fits the other templates of its cluster, so a distractor drawn from $s(t)$, and still more from t itself, would be a second option fitting $\tilde{t}$ equally well. Every caption is the gold answer exactly once, and always picking the longest option does no better than chance (Appendix H).

3.3 PARTITIONS

The partitions of MUNCH test how far a model's grasp of meme humor reaches beyond the training templates by withholding, at test time, first the template t itself and then the meaning that template carries, $s(t)$. The Known Template ($\mathcal{KT}$) partitions contains all instances whose templates appear in training. Heldout Template ($\mathcal{HT}$) instances use templates absent from training whose semantic cluster does appear, so a model must infer a template's intent from its image and recognize that intent in a *novel* image. Heldout Cluster ($\mathcal{HC}$) instances use templates t whose cluster, $s(t)$ never appears in training, which most resembles seeing an unfamiliar meme in the wild. The released partitions hold 123,118 instances.

3.4 HARDER DISTRACTORS

MUNCH-HARD keeps the gold captions of the 3,000-instance subset and replaces each distractor with the caption nearest to the gold in Qwen3-Embedding-4B space, one from each of three other clusters of the same partition (i.e., the same pool the standard distractors are drawn from). Candidates with cosine similarity ≥ 0.9 to the gold are excluded as near-paraphrases. The hard distractors have a mean cosine similarity of 0.67–0.69 with the gold caption, against 0.41–0.42 for the standard distractors.

3.5 DATA QUALITY

The two steps most likely to corrupt an instance, caption transcription and caption removal, were each audited. An annotator compared 100 OCR-extracted captions against the text visible in each post's original image, 50 that passed through the original OCR unchanged and 50 recovered by the

idefics2-8b rerun, and meaning was preserved in 99.5% of instances (Appendix B.4). The rerun, applied to 13.3% of instances, recovered the caption correctly in 48 of 50 cases and is the only source of the sample's major errors. A second audit of 50 inpainted templates found no readable trace of a removed caption.

4 Benchmark Results

We evaluate zero-shot MLLMs, contrastive models, and fine-tuned models on the three test partitions, then retest the models with harder distractors (§4.3).

4.1 Experimental Setup

The partitions are assigned by cluster: 7.5% of the data goes to $\mathcal{HC}$ and dev, $\mathcal{HT}$ is drawn from multi-template clusters until the partition reaches roughly 5% of the data, and the remaining instances are split into training, $\mathcal{KT}$, and a validation sample (Appendix D).

Models We evaluate 23 MLLMs zero-shot. Open models come from the Qwen3.5, Qwen3.6, and Qwen3.8 families spanning 0.8B to 35B parameters (Qwen Team, 2026a;b;c;d), the Gemma 4 family from E2B to 31B (Google DeepMind, 2026), LLaVA-OneVision-7B (Li et al., 2025), and five earlier models (Li et al., 2024b; Grattafiori et al., 2024; Bai et al., 2025b; Abouelenin et al., 2025; Wu et al., 2024); the full roster is in Table 2. Closed models are three current models from frontier labs, `claude-haiku-4-5-20251001`, `gpt-5.6-luna`, and `gemini-3.5-flash-lite`. Contrastive baselines are CLIP ViT-B/32 and ViT-L/14-336 (Radford et al., 2021), SigLIP-so400m (Zhai et al., 2023), SigLIP2-so400m (Tschannen et al., 2025), and OpenCLIP ViT-bigG/14 trained on LAION-2B (Cherti et al., 2023), which choose the caption most similar to the image. Image–text matching baselines choose the option scored highest by the BLIP and BLIP-2 matching heads (Li et al., 2022; 2023a), by BLIP-2's captioner, or by VQAScore with CLIP-FlanT5-XXL, the probability that a vision–language model answers yes when asked whether the image shows the caption (Lin et al., 2024). We LoRA fine-tune four MLLMs (Hu et al., 2022) and fine-tune CLIP ViT-L/14-336 with a four-way contrastive objective (Goyal et al., 2023) over each instance's own options, so other captions of the same template are never negatives. Checkpoints are selected by accuracy on validation instances whose template appears in training (Appendix A).

Prompting and scoring Every MLLM receives the same prompt, which was chosen by testing ten variants on a sample of instances (Appendices E, F, and A). Models are scored using an extractor to recover the chosen option from free-form output (Appendix I); responses that the extractor cannot parse are treated as wrong. Gemini 3.5 Flash-Lite declined 0.2% of instances on content-policy grounds; the other two closed models labeled all instance. A declined instance counts as wrong. Chance is 25% in every partition.

Human baseline Because every instance has a known answer, we estimate human performance the way a model is scored, as accuracy against the gold caption (cf., Nangia & Bowman, 2019), rather than as inter-annotator agreement, which measures the reliability of annotators who create a resource's labels (Artstein & Poesio, 2008). Two annotators whose familiarity with meme culture differs, from moderate to high, labeled 200 instances, so that the difference between the two scores shows how much the task depends on that familiarity. The annotators scored 54.7% and 75.2% in macro-over-cluster accuracy, in line with their familiarity, and we use the more familiar annotator's 75.2% as the human performance estimate (Appendix G). We use the same estimate for all partitions, which differ only in what a model saw in training, rather than which memes an annotator knows.

Statistics and Evaluation Instances built on templates from the same semantic cluster share that cluster's meaning and caption conventions, so those instances are not independent draws. All confidence intervals therefore come from a two-stage cluster bootstrap (2,000 replicates) that resamples clusters and then instances within each; comparisons between two systems resample the same clusters for both. To check that no comparison rests on one cluster, we repeat each comparison with each cluster removed in turn. Because clusters differ greatly in size, we report *macro-over-cluster accuracy*, which weights every cluster equally, with pooled accuracy alongside the macro figures in Table 2.

Model	Pooled $\mathcal{KT}$	Pooled $\mathcal{HT}$	Pooled $\mathcal{HC}$ [95% CI]	Macro-over-cluster $\mathcal{KT}$	Macro-over-cluster $\mathcal{HT}$	Macro-over-cluster $\mathcal{HC}$	All
Zero-shot, open MLLMs							
Qwen3.5-0.8B	30.2	31.6	30.9 [28, 33]	32.1	34.1	33.5	30.7
Qwen3.5-2B	32.0	33.2	38.0 [33, 41]	34.7	36.0	37.8	33.8
Qwen3.5-4B	34.2	36.7	41.1 [35, 45]	41.5	40.7	38.7	36.5
Qwen3.5-9B	35.2	38.2	42.6 [38, 45]	42.3	42.8	43.8	37.7
Qwen3.5-27B	42.5	**46.5**	**51.3** [43, 57]	48.4	**51.4**	**49.6**	**45.6**
Qwen3.5-35B-A3B	38.9	41.5	45.2 [40, 48]	43.7	46.8	47.7	41.0
Qwen3.6-27B	43.0	45.2	49.9 [43, 55]	**49.2**	50.0	48.3	45.2
Qwen3.6-35B-A3B	39.5	41.7	45.6 [40, 49]	45.2	46.4	45.1	41.5
Qwen3.8-27B	40.7	44.4	50.0 [43, 56]	46.3	49.8	47.6	43.8
Gemma-4-E2B	29.4	30.0	31.7 [29, 34]	32.5	33.6	32.9	30.1
Gemma-4-E4B	30.8	33.2	39.4 [29, 44]	32.5	35.3	34.4	33.4
Gemma-4-12B	34.9	37.1	42.2 [36, 46]	38.6	38.5	38.9	37.2
Gemma-4-26B-A4B	37.7	38.6	45.1 [36, 49]	39.3	40.9	39.7	39.7
Gemma-4-31B	**43.4**	45.5	49.9 [43, 54]	48.8	49.3	44.7	45.4
LLaVA-OneVision-7B	35.2	38.4	43.6 [35, 48]	38.8	42.6	42.6	38.0
LLaVA-v1.6-Mistral-7B	30.2	31.9	33.3 [28, 37]	31.9	34.9	32.2	31.3
Llama-3.2-11B-Vision	33.4	35.5	40.2 [32, 44]	35.8	35.8	35.3	35.6
Qwen2.5-VL-7B	33.1	35.5	38.1 [32, 42]	36.2	40.4	38.6	34.9
Phi-4-multimodal	28.9	29.9	31.4 [28, 33]	30.6	29.4	31.1	29.7
DeepSeek-VL2-tiny	27.7	27.3	28.2 [26, 30]	27.6	27.4	28.4	27.7
Zero-shot, closed MLLMs							
Claude Haiku 4.5	35.2	37.1	42.1 [34, 46]	39.2	39.3	42.2	37.3
GPT-5.6 Luna	35.7	38.7	41.1 [35, 44]	41.6	44.7	41.5	37.7
Gemini 3.5 Flash-Lite	**60.8**	**57.1**	**58.7** [51, 65]	**61.6**	**58.8**	**53.8**	**59.4**
Zero-shot, contrastive							
CLIP ViT-B/32	44.2	47.1	51.3 [39, 61]	48.9	51.1	47.0	46.6
CLIP ViT-L/14-336	**50.5**	**53.8**	**56.2** [44, 65]	**55.2**	**57.2**	**51.0**	**52.7**
SigLIP-so400m	37.7	39.7	44.3 [37, 49]	43.2	43.4	44.3	39.8
SigLIP2-so400m	34.5	36.3	39.2 [35, 43]	36.2	35.7	39.2	36.0
OpenCLIP ViT-bigG/14	35.4	35.8	46.5 [32, 57]	39.3	38.1	41.1	38.2
***Fine-tuned on* MUNCH**							
LLaVA-v1.6-Mistral-7B, LoRA	51.1	43.9	44.9 [34, 51]	39.4	42.8	40.6	47.9
Llama-3.2-11B-Vision, LoRA	71.2	60.4	53.7 [47, 60]	61.6	58.7	57.0	64.4
Qwen2.5-VL-7B, LoRA	77.1	66.7	53.9 [50, 61]	67.2	68.3	58.2	69.0
Gemma-4-E4B, LoRA	76.8	61.9	36.4 [28, 48]	66.2	59.9	42.6	63.5
CLIP ViT-L/14-336, contrastive	**86.3**	**78.0**	**63.0** [57, 68]	**85.9**	**77.2**	**62.2**	**78.7**
Human performance estimate	–	–	–	75.2 (all partitions)			–

Table 2: Answer accuracy (%) on the three test partitions, pooled over instances and averaged over clusters. Bold marks the highest score in each column within each block of systems. $\mathcal{HC}$ intervals are two-stage cluster-bootstrap 95% intervals on pooled accuracy; macro accuracy averages over the 249, 38, and 21 clusters of $\mathcal{KT}$, $\mathcal{HT}$, and $\mathcal{HC}$, five of which hold 81.5% of $\mathcal{HC}$'s instances. All pools the three partitions. The human performance estimate is one annotator's macro-over-cluster accuracy on 200 instances, one figure for all partitions.

4.2 RESULTS

Zero-shot MLLMs trail humans on every partition The best score any zero-shot MLLM reaches on any partition, 61.6 on $\mathcal{KT}$ by Gemini 3.5 Flash-Lite, sits 13.6 points below the human performance estimate (Table 2, closed block), while its $\mathcal{HC}$ accuracy has an even larger gap (21.4 points). The three closed models do not behave alike. Gemini 3.5 Flash-Lite leads every open MLLM by more than 12 points on $\mathcal{KT}$, while Claude Haiku 4.5 and GPT-5.6 Luna score where the mid-sized open models do. Scale helps within a model family but levels off well short of the gap, e.g., Qwen3.5 gains sixteen points on $\mathcal{KT}$ between 0.8B and 27B. Zero-shot CLIP ViT-L/14-336 outscores every open MLLM we test, so the image–caption signal lies within reach of a model trained only on image–text similarity, consistent with prior work finding that MLLMs classify images less accurately than the contrastive encoders MLLMs are built on (Zhang et al., 2024b). Contrastive training alone doesn't

close the gap, however, as the larger and newer SigLIP, SigLIP2, and OpenCLIP encoders score 12 to 19 points lower than CLIP ViT-L/14-336 on $\mathcal{KT}$. What a contrastive encoder has seen in pretraining matters more here than how large the encoder is, in line with zero-shot accuracy tracking a concept's pretraining frequency (Udandarao et al., 2024).

Fine-tuning helps most on familiar templates Fine-tuning raises every system above the zero-shot accuracy of the same base model, and for all but the weakest system the gain is largest on $\mathcal{KT}$ and smallest on $\mathcal{HC}$. Supervision thus transfers to new templates whose meaning is familiar but largely fails to help transfer the ability to new meanings. Fine-tuned CLIP ViT-L/14-336 leads every LoRA-tuned MLLM, by 18.8 points on $\mathcal{KT}$, 8.8 on $\mathcal{HT}$, and 4.1 on $\mathcal{HC}$. Fine-tuning helps many of the small models outperform the closed model we evaluate; for example, LoRA-tuned Qwen2.5-VL-7B surpasses Gemini on $\mathcal{KT}$ and $\mathcal{HT}$, and fine-tuned CLIP ViT-L/14-336 surpasses Gemini on all three partitions and reaches the human performance estimate on $\mathcal{KT}$ and $\mathcal{HT}$. These results show that understanding the image–text humor in memes can be learned, but most current models lack this humor understanding ability when run in zero-shot.

Reasoning mode does not help at the scales tested Because humor is often described as requiring inference over an incongruity (Suls, 1972), extended reasoning is a natural candidate for closing the zero-shot gap. With a 16K-token budget on a stratified subset of 1,000 instances per partition, the smaller Qwen3.5 models exhaust the budget without producing an answer on a quarter to half of $\mathcal{KT}$ responses, and counting those responses as wrong lowers the accuracy of the smaller models (Appendix Table 8). At 27B parameters, Qwen3.8 finishes nearly every response but gains only a few points of accuracy that is statistically indistinguishable from its no-thinking counterpart's performance. Our result agrees with prior work finding that chain-of-thought helps mainly on math and symbolic reasoning (Sprague et al., 2025) and that longer multimodal traces draw attention away from the image (Xu et al., 2025).

4.3 Results on Munch-hard

Munch-hard tests whether models perform as well when facing more semantically similar options.

Closer distractors lower accuracy for nearly every system Accuracy falls on Munch-hard for almost every system on $\mathcal{KT}$ and $\mathcal{HT}$ (Appendix Table 9). The accuracy of zero-shot contrastive models falls 7.6 to 15.0 points on $\mathcal{KT}$, the largest drop for the highest-performing model in the standard test set, Gemini 3.5 Flash-Lite. On $\mathcal{HC}$, most models had statistically indistinguishable performance. These results suggest models current capabilities are not robust to distractors that have some meaning in common with the actual caption (though are still incorrect).

Fine-tuned systems lose more to closer distractors than zero-shot MLLMs do Fine-tuning does not models more robust to semantically similar distractors (Appendix Table 9, bottom block). This result suggests that fine-tuning helps models learn the topical cues associated with specific templates (which semantically closer distractors remove), rather than learning a deeper representation of the image–text humor. Nevertheless, fined-tuned models are still above chance on $\mathcal{HC}$ (and mostly above their zero-shot counterparts), indicating that models did learn some generalization ability.

5 Ablations and Robustness Checks

Does the answer depend on the image? On the full partitions, systems that never sees the image all score within 2.6 points of chance (25%, Appendix H, Table 6). The baseline of always picking the longest option similarly is at chance, scoring 23.7 to 25.7%. Thus, answering Munch cannot be solved by a text-only shortcut and requires the image to understand which caption is humorous.

Do familiar templates carry conventions a model can exploit? Familiar templates could carry conventions a model can exploit. A template often reflects a theme so the captions people write for a template likely share vocabulary and stance. To test how predictive a template's conventions are, we quantify that convention with a *caption prior*, which ignores the image and picks the option whose Qwen3-Embedding-4B embedding (Zhang et al., 2025) is most semantically similar to the

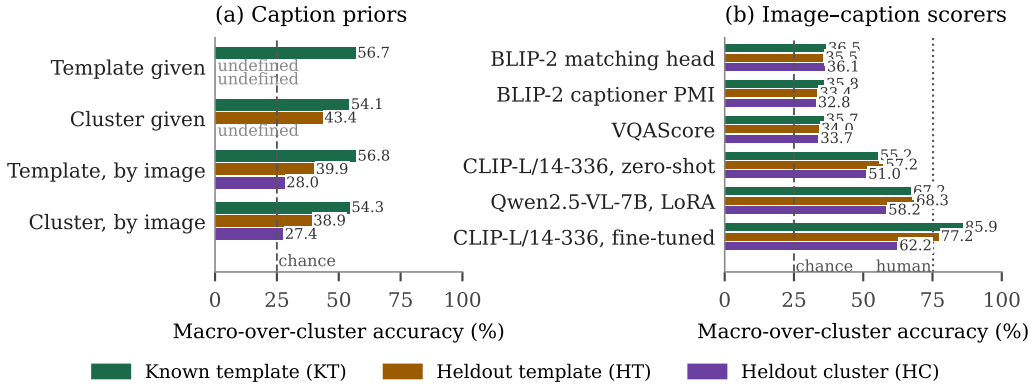


Figure 2: Macro-over-cluster accuracy of the systems that locate what MUNCH measures. (a) Caption priors pick the option most semantically similar to the training captions of a template or cluster. Told the template or cluster, a prior is undefined where that template or cluster is absent from training, and found by nearest-neighbor search over images, a prior is defined on every partition. (b) Scorers built to judge whether a caption describes an image stay near 35%, while fine-tuned CLIP ViT-L/14-336 leads every other system yet scores 13 points below the human performance estimate (dotted) on $\mathcal{HC}$.

embeddings of that template's training captions; analogously, a *cluster prior* does the same over a semantic cluster (Appendix J). Told an instance's template, the caption prior reaches 56.7% on $\mathcal{KT}$, and told the instance's cluster, 54.1%, so the typical content of a template's captions alone is highly informative of what content it selects for (Figure 2a). This result suggests that much of the performance improvements on familiar templates is due to learning caption conventions. On $\mathcal{HT}$ the cluster prior falls to 43.4%, since a new template shares only part of the conventions of the cluster, but we find models surpass this prior, indicating they have learned how to infer a template's selection preferences and can generalize these preferences to new images when the preferences have been seen before.

Can a model recover a template's caption conventions from the image alone? When the template is not given but matched by image to the most similar training template, the caption priors score nearly as well on $\mathcal{HT}$ as the prior told the instance's cluster, and barely above chance on $\mathcal{HC}$ (Figure 2a, priors found by image). On $\mathcal{HT}$ the matched template shares the instance's cluster for 48% of instances, and the matched template and cluster priors score 39.9 and 38.9%, against 43.4% for the prior told the cluster. On $\mathcal{HC}$, where no training template shares the cluster (i.e., the type of content or intent it selects for), the matched template prior scores 28.0%, which LoRA-tuned Gemma-4-E4B outperforms 14.6 points, so template recognition alone does not account for what fine-tuning achieves on unfamiliar meanings.

Is multimodal humor understanding learnable and transferable? Fine-tuning a contrastive image–text model on the training partition yields the strongest system we evaluate. CLIP ViT-L/14-336 trained with a four-way objective over each instance's own options leads every LoRA-tuned MLLM on all three partitions (Figure 2b). The performance of the pretrained model stays in the fifties, even though literal image–caption retrieval is the task contrastive pretraining optimizes (Radford et al., 2021). Further, models built expressly to judge whether a caption describes an image, including BLIP-2's matching head and VQAScore, score lower still. A MUNCH caption *fills* the slot a template opens, rather than *describing* the image, so a model must identify what a template is selecting for before the model can recognize a fitting filler. Supervision teaches that mapping for templates and meanings. Our results show the gain over the pretrained model shrinks steadily with distance from training, from 30.7 points on $\mathcal{KT}$ to 20.0 on $\mathcal{HT}$ and 11.2 on $\mathcal{HC}$, indicating that models struggle with both identifying what a template selects for and identify the mapping between filler and selectional reference. While trained models are still able to transfer this ability to novel templates and meanings ($\mathcal{HC}$), even the strongest system scores below our human performance estimate.

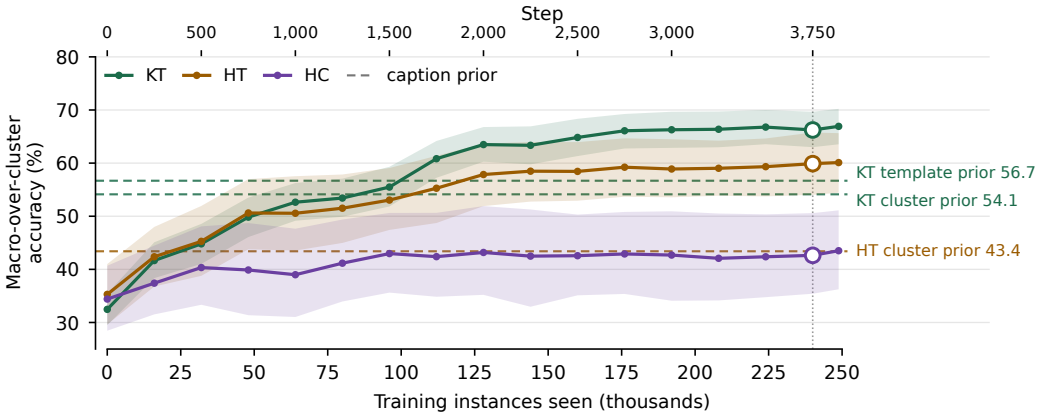


Figure 3: Macro-over-cluster accuracy of LoRA-tuned Gemma-4-E4B over three training epochs, with 95% cluster-bootstrap bands, the embedding caption priors as dashed lines, and the selected checkpoint (step 3,750) marked.

6 Fine-Tuning Dynamics

Does continued training help extend matching to new meanings? For Gemma-4-E4B, accuracy on $\mathcal{KT}$ rises from 32.5 to 66.9 over training, against 56.7 for the template caption prior, and accuracy on $\mathcal{HT}$ follows the same course (Figure 3). Accuracy on $\mathcal{HC}$ instead rises from 34.4 to 43.0 by step 1,500 and stays between 42.1 and 43.5 thereafter, while $\mathcal{KT}$ gains another 11.4 points. Pooled $\mathcal{HC}$ accuracy stays at 33.8 to 39.7 throughout, against 39.4 zero-shot. Continued training therefore strengthens matching for familiar templates and meanings; however, learning for the general ability quickly saturates and additional training does not help the model generalize for new meanings. This result suggest models are likely learning template- or cluster-specific cues, rather than general image–text humor interpretation abilities.

Which memes stay difficult, and what makes a meme difficult? Memes from general communities are easier than memes from topically niche ones. A logistic regression of zero-shot correctness on subreddit, template, caption length, and model (Appendix L) assigns the most negative coefficients to communities about history, a single person, or the Marvel universe, whose captions presuppose community knowledge. The easiest community, r/antimeme, rewards captions that describe the image literally, which is the setting in which zero-shot models are strongest. Templates whose captions quote a statement of opinion on a depicted screen (e.g., the Lisa Simpson in front of a screen meme) are the most difficult, because the same visual frame hosts semantically diverse fillers.

7 Conclusion

Models that work alongside people in social settings encounter humor as a matter of course, so benchmarks need to measure multimodal humor understanding directly. We introduce MUNCH, a large-scale benchmark for multimodal humor understanding built from 123,118 Reddit memes whose original captions have been inpainted out and recast as four-way multiple-choice questions. Even the strongest closed model we evaluate scores 14 to 21 points below our human performance estimate zero-shot, while fine-tuning lets a 7B open model surpass closed-weight model on familiar templates and familiar meanings. The improvement from fine-tuning help performance on new templates far more readily than new meanings, where the strongest system scores 13 points below the human performance estimate. MUNCH offers a controlled testbed for multimodal humor. We release MUNCH, its partitions, the inpainted templates, and the evaluation code at `http://url` under CC-BY-SA 4.0.

ETHICS STATEMENT

Memes are widely used and some memes are created with malicious or hurtful intent, which prior work has sought to identify. Here, MUNCH is curated from the SEMANTICMEMES corpus (Zhou et al., 2024) which was explicitly constructed to avoid containing content that may be objectionable. SEMANTICMEMES draws from actively moderated subreddits with rules against explicitly hateful content (e.g., not 4chan-like sources). Further, we recrawl the URLs from their data, which, as a result, omits any memes the original authors deleted or made private, thereby preserving meme authors' right to privacy. To further assess any potential risk of offensiveness, our manual annotation and error analysis of several thousand memes (Appendix B.5.1), we did not encounter overtly offensive content. We additionally restrict to widely circulated, neutral templates (§B). While these choices do not eliminate the risk of harmful content, this construction process should make it rare in practice. Our use of the data complies with Reddit's Terms of Service.

The task in MUNCH is discriminative where models *select* a caption rather than generate one. However, some models have been trained to generate humorous captions, which MUNCH could be used for as well due to its image–text pairs. If MUNCH is used in future generative training, we recommend pairing it with hate- and disinformation-detection tools from meme-safety research (Kiela et al., 2020; Jha et al., 2024; Qu et al., 2022). Although our discriminative task and data selection process would not optimize models to produce offensive captions, the dataset's captions come from in-the-wild posts and may nevertheless reflect their biases, so care should be taken in any pipeline that uses them for training models to generate.

AI USE STATEMENT

We used AI assistants to search for related work, check grammar and copyedit, and simulate reviewer feedback. An AI coding assistant helped implement model training and inference, and parts of the scoring and analysis code (including making the answer extractor more robust, adding cluster bootstrap, and generating tables from results files to avoid copy-paste errors). AI assistants help revise parts of the text for space and readability. The authors reviewed, revised where necessary, and verified all AI-assisted code, analyses, and text, and take full responsibility for the content of the submission.

REPRODUCIBILITY STATEMENT

We will release the MUNCH instances, partitions, inpainted templates, and all model predictions. All code used for the project will be released on github. Final model checkpoints will be released on Huggingface and all intermediate models checkpoints will be available upon request as long as storage allows us to keep them. Prompts are in Appendix F, inference settings in Appendix A, and the answer extractor in Appendix I. Subsets used for reasoning-mode and closed-model evaluation are fixed by seed and released.

REFERENCES

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yunsheng Li, Chen Liang, Xihui Lin, Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Arindam Mitra, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyna Zhang, Yunan Zhang, and Xiren Zhou. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras, 2025. URL https://arxiv.org/abs/2503.01743.

Siddhant Agarwal, Shivam Sharma, Preslav Nakov, and Tanmoy Chakraborty. MemeMQA: Multimodal question answering for memes via rationale-based inferencing. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 5042–5078, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.300. URL https://aclanthology.org/2024.findings-acl.300/.

Arjun R. Akula, Brendan Driscoll, Pradyumna Narayana, Soravit Changpinyo, Zhiwei Jia, Suyash Damle, Garima Pruthi, Sugato Basu, Leonidas Guibas, William T. Freeman, Yuanzhen Li, and Varun Jampani. Metaclue: Towards comprehensive visual metaphors research. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23201–23211, June 2023.

Ron Artstein and Massimo Poesio. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008. doi: 10.1162/coli.07-034-R2. URL https://aclanthology.org/J08-4004/.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *CoRR*, abs/2511.21631, November 2025a. URL https://doi.org/10.48550/arXiv.2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

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 (ICCV)*, pp. 2616–2627, October 2023.

Nitzan Bitton-Guetta, Aviv Slobodkin, Aviya Maimon, Eliya Habba, Royi Rassin, Yonatan Bitton, Idan Szpektor, Amir Globerson, and Yuval Elovici. Visual riddles: a commonsense and world knowledge challenge for large vision and language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 139561–139588. Curran Associates, Inc., 2024. doi: 10.52202/079017-4428. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/fbf5efe979e6754dc06a0869233f2510-Paper-Datasets_and_Benchmarks_Track.pdf.

Tuhin Chakrabarty, Arkadiy Saakyan, Olivia Winn, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki, and Smaranda Muresan. I spy a metaphor: Large language models and diffusion models co-create visual metaphors. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 7370–7388, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.465. URL https://aclanthology.org/2023.findings-acl.465/.

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 (CVPR)*, 2016. URL https://arxiv.org/abs/1512.04407.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems*, 2024.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible Scaling Laws for Contrastive Language-Image Learning. In *Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, 2023. doi: 10.1109/CVPR52729.2023.00276. URL https://mlanthology.org/cvpr/2023/cherti2023cvpr-reproducible/.

Shiling Deng, Serge Belongie, and Peter Ebert Christensen. Large vision-language models for knowledge-grounded data annotation of memes, 2025. URL https://arxiv.org/abs/2501.13851.

Google DeepMind. Gemma 4 model card and core documentation. https://ai.google.dev/gemma/docs/core/model_card_4, 2026. Accessed: May 2026.

Sachin Goyal, Ananya Kumar, Sankalp Garg, Zico Kolter, and Aditi Raghunathan. Finetune like You Pretrain: Improved Finetuning of Zero-Shot Vision Models. In *Conference on Computer Vision and Pattern Recognition*, pp. 19338–19347, 2023. doi: 10.1109/CVPR52729.2023.01853. URL https://mlanthology.org/cvpr/2023/goyal2023cvpr-finetune/.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu,

Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 107–112, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2017. URL https://aclanthology.org/N18-2017/.

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 (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 688–714, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.41. URL https://aclanthology.org/2023.acl-long.41/.

Zachary Horvitz, Jingru Chen, Rahul Aditya, Harshvardhan Srivastava, Robert West, Zhou Yu, and Kathleen McKeown. Getting serious about humor: Crafting humor datasets with unfunny large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 855–869, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-short.76. URL https://aclanthology.org/2024.acl-short.76/.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 31096–31116. Curran Associates, Inc., 2023. doi: 10.52202/075280-1355. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/63461de0b4cb760fc498e85b18a7fe81-Paper-Datasets_and_Benchmarks.pdf.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://arxiv.org/abs/2106.09685.

Zhe Hu, Tuo Liang, Jing Li, Yiren Lu, Yunlai Zhou, Yiran Qiao, Jing Ma, and Yu Yin. Cracking the code of juxtaposition: Can AI models understand the humorous contradictions. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024. Oral.

EunJeong Hwang and Vered Shwartz. MemeCap: A dataset for captioning and interpreting memes. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 1433–1445, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.89. URL https://aclanthology.org/2023.emnlp-main.89/.

EunJeong Hwang, Peter West, and Vered Shwartz. BottleHumor: Self-informed humor explanation using the information bottleneck principle. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025.

Vedaant V Jain, Gabriel Kreiman, and Felipe dos Santos Alves Feitosa. Humordb: Can ai understand graphical humor? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 604–613, October 2025.

Prince Jha, Krishanu Maity, Raghav Jain, Apoorv Verma, Sriparna Saha, and Pushpak Bhattacharyya. Meme-ingful analysis: Enhanced understanding of cyberbullying in memes through multimodal explanations. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 930–943, St. Julian's, Malta, 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.eacl-long.56.

Yifan Jiang, Jiarui Zhang, Kexuan Sun, Zhivar Sourati, Kian Ahrabian, Kaixin Ma, Filip Ilievski, and Jay Pujara. MARVEL: Multidimensional abstraction and reasoning through visual evaluation and learning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 46567–46592. Curran Associates, Inc., 2024. doi: 10.52202/079017-1477. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/529d8b3a23991e83db07f21727256374-Paper-Datasets_and_Benchmarks_Track.pdf.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020. URL https://arxiv.org/abs/2005.04790.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. Obelics: An open web-scale filtered dataset of interleaved image-text documents. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track*, 2023.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024.

Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. Naturalbench: Evaluating vision-language models on natural adversarial samples. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 17044–17068. Curran Associates, Inc., 2024a. doi: 10.52202/079017-0542. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/1e69ff56d0ebff0752ff29caaddc25dd-Paper-Datasets_and_Benchmarks_Track.pdf.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *Trans. Mach. Learn. Res.*, 2025, 2025. URL https://openreview.net/forum?id=zKv8qULV6n.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024b.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12888–12900. PMLR, 2022. URL https://proceedings.mlr.press/v162/li22n.html.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023a. URL https://proceedings.mlr.press/v202/li23q.html.

Runjia Li, Shuyang Sun, Mohamed Elhoseiny, and Philip Torr. Oxfordtvg-hic: Can machine make humorous captions from images? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20293–20303, October 2023b.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *Computer Vision – ECCV 2024*, Lecture Notes in Computer Science, pp. 366–384. Springer, 2024. doi: 10.1007/978-3-031-72673-6_20.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? In *ECCV (6)*, pp. 216–233, 2024a. URL https://doi.org/10.1007/978-3-031-72658-3_13.

Ziqiang Liu, Feiteng Fang, Xi Feng, Xinrun Du, Chenhao Zhang, Zekun Wang, Yuelin Bai, Qixuan Zhao, Liyang Fan, Chengguang Gan, Hongquan Lin, Jiaming Li, Yuansheng Ni, Haihong Wu, Yaswanth Narsupalli, Zhigang Zheng, Chengming Li, Xiping Hu, Ruifeng Xu, Xiaojun Chen, Min Yang, Jiaheng Liu, Ruibo Liu, Wenhao Huang, Ge Zhang, and Shiwen Ni. II-Bench: An image implication understanding benchmark for multimodal large language models. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2024b.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, Benjamin Bossan, and Marian Tietz. Peft: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft, 2022.

Shreyash Mishra, S Suryavardan, Parth Patwa, Megha Chakraborty, Anku Rani, Aishwarya Reganti, Aman Chadha, Amitava Das, Amit Sheth, Manoj Chinnakotla, Asif Ekbal, and Srijan Kumar. Memotion 3: Dataset on sentiment and emotion analysis of codemixed hindi-english memes. In *Proceedings of De-Factify 2: 2nd Workshop on Multimodal Fact Checking and Hate Speech Detection, co-located with AAAI 2023*, volume 3555 of *CEUR Workshop Proceedings*, 2023. URL https://ceur-ws.org/Vol-3555/.

Nikita Nangia and Samuel R. Bowman. Human vs. muppet: A conservative estimate of human performance on the GLUE benchmark. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4566–4575, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1449. URL https://aclanthology.org/P19-1449/.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Van Steenkiste, Lisa Anne Hendricks, Karolina Stanczak, and Aishwarya Agrawal. Benchmarking vision language models for cultural understanding. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5769–5790, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.329. URL https://aclanthology.org/2024.emnlp-main.329/.

Asaf Nissenbaum and Limor Shifman. Meme templates as expressive repertoires in a globalizing world: A cross-linguistic study. *Journal of Computer-Mediated Communication*, 23(5):294–310, 2018.

Hugo Gonçalo Oliveira, Diogo Costa, and Alexandre Miguel Pinto. One does not simply produce funny memes! - explorations on the automatic generation of internet humor. In *Proceedings of the Seventh International Conference on Computational Creativity (ICCC)*, 2016. URL https://www.computationalcreativity.net/iccc2016/wp-content/uploads/2016/06/paper_32-1.pdf.

Kun Ouyang, Yuanxin Liu, Shicheng Li, Yi Liu, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. PunchBench: Benchmarking MLLMs in multimodal punchline comprehension. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 986–1008, Vienna, Austria, 2025. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, 2019. URL https://arxiv.org/abs/1912.01703.

Badri N. Patro, Mayank Lunayach, Deepankar Srivastava, Sarvesh Sarvesh, Hunar Singh, and Vinay P. Namboodiri. Multimodal humor dataset: Predicting laughter tracks for sitcoms. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 576–585, 2021. doi: 10.1109/WACV48630.2021.00062.

Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. POTATO: The portable text annotation tool. In Wanxiang Che and Ekaterina Shutova (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 327–337, Abu Dhabi, UAE, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-demos.33. URL https://aclanthology.org/2022.emnlp-demos.33/.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. In Malvina Nissim, Jonathan Berant, and Alessandro Lenci (eds.), *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pp. 180–191, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-2023. URL https://aclanthology.org/S18-2023/.

Jingnong Qu, Liunian Harold Li, Jieyu Zhao, Sunipa Dev, and Kai-Wei Chang. Disinfomeme: A multimodal dataset for detecting meme intentionally spreading out disinformation, 2022. URL https://arxiv.org/abs/2205.12617.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a. URL https://qwen.ai/blog?id=qwen3.5.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, April 2026b. URL https://qwen.ai/blog?id=qwen3.6-27b.

Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all, April 2026c. URL https://qwen.ai/blog?id=qwen3.6-35b-a3b.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026d. URL https://qwen.ai/blog?id=qwen3.8.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021. URL https://arxiv.org/abs/2103.00020.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, Frederico Belcavello, Ganzorig Batnasan, Gisela Vallejo, Grainne Caulfield, Guido Ivetta, Haiyue Song, Henok Biadglign Ademtew, Hernán Maina, Holy Lovenia, Israel Abebe Azime, Jan Christian Blaise Cruz, Jay Gala, Jiahui Geng, Jesus-German Ortiz-Barajas, Jinheon Baek, Jocelyn Dunstan, Laura Alonso Alemany, Kumaranage Ravindu Yasas Nagasinghe, Luciana Benotti, Luis Fernando D' Haro, Marcelo Viridiano, Marcos Estecha-Garitagoitia, Maria Camila Buitrago Cabrera, Mario Rodríguez-Cantelar, Mélanie Jouitteau, Mihail Mihaylov, Naome Etori, Mohamed Fazli Mohamed Imam, Muhammad Farid Adilazuarda, Munkhjargal Gochoo, Munkh-Erdene Otgonbold, Olivier Niyomugisha, Paula Mónica Silva, Pranjal Chitale, Raj Dabre, Rendi Chevi, Ruochen Zhang, Ryandito Diandaru, Samuel Cahyawijaya, Santiago Góngora, Soyeong Jeong, Sukannya Purkayastha, Tatsuki Kuribayashi, Teresa Clifford, Thanmay Jayakumar, Tiago Timponi Torrent, Toqeer Ehsan, Vladimir Araujo, Yova Kementchedjhieva, Zara Burzo, Zheng Wei Lim, Zheng Xin Yong, Oana Ignat, Joan Nwatu, Rada Mihalcea, Thamar Solorio, and Alham Fikri Aji. Cvqa: Culturally-diverse multilingual visual question answering benchmark. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 11479–11505. Curran Associates, Inc., 2024. doi: 10.52202/079017-0366. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/1568882ba1a50316e87852542523739c-Paper-Datasets_and_Benchmarks_Track.pdf.

Yuriel Ryan, Rui Yang Tan, Kenny Tsu Wei Choo, and Roy Ka-Wei Lee. Humor in pixels: Benchmarking large multimodal models understanding of online comics. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 14024–14050, Suzhou, China, 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.findings-emnlp.755/.

Arkadiy Saakyan, Shreyas Kulkarni, Tuhin Chakrabarty, and Smaranda Muresan. Understanding figurative meaning through explainable visual entailment. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1–23, Albuquerque, New Mexico, 2025. URL https://aclanthology.org/2025.naacl-long.1/.

Siddhant Bikram Shah, Shuvam Shiwakoti, Maheep Chaudhary, and Haohan Wang. MemeCLIP: Leveraging CLIP representations for multimodal meme classification. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17320–17332, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.959. URL https://aclanthology.org/2024.emnlp-main.959/.

Shivam Sharma, Ramaneswaran S, Md. Shad Akhtar, and Tanmoy Chakraborty. Emotion-aware multimodal fusion for meme emotion detection. *IEEE Transactions on Affective Computing*, 15(3): 1800–1811, 2024. doi: 10.1109/TAFFC.2024.3378698. URL https://arxiv.org/abs/2403.10279.

Zhengpeng Shi, Yanpeng Zhao, Jianqun Zhou, Yuxuan Wang, Qinrong Cui, Wei Bi, Song-Chun Zhu, Bo Zhao, and Zilong Zheng. v-HUB: A benchmark for video humor understanding from vision and sound. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 38544–38567, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1785. URL https://aclanthology.org/2026.acl-long.1785/.

Zayne Sprague, Fangcong Yin, Juan Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 94118–94162, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/ead542f13a38179d1b55b88610f959a1-Paper-Conference.pdf.

Jerry M. Suls. A two-stage model for the appreciation of jokes and cartoons: An information-processing analysis. In Jeffrey H. Goldstein and Paul E. McGhee (eds.), *The Psychology of Humor: Theoretical Perspectives and Empirical Issues*, pp. 81–100. Academic Press, New York, 1972. doi: 10.1016/B978-0-12-288950-9.50010-9.

Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2149–2159, 2022.

Kohtaro Tanaka, Hiroaki Yamane, Yusuke Mori, Yusuke Mukuta, and Tatsuya Harada. Learning to evaluate humor in memes based on the incongruity theory. In Xianchao Wu, Peiying Ruan, Sheng Li, and Yi Dong (eds.), *Proceedings of the Second Workshop on When Creative AI Meets Conversational AI*, pp. 81–93, Gyeongju, Republic of Korea, October 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.cai-1.9/.

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 (CVPR)*, pp. 5238–5248, June 2022.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 87310–87356. Curran Associates, Inc., 2024. doi: 10.52202/079017-2771. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/9ee3a664ccfeabc0da16ac6f1f1cfe59-Paper-Conference.pdf.

Michael Tschannen, Alexey A. Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier J. Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *CoRR*, abs/2502.14786, February 2025. URL https://doi.org/10.48550/arXiv.2502.14786.

Vishaal Udandarao, Ameya Prabhu, Adhiraj Ghosh, Yash Sharma, Philip Torr, Adel Bibi, Samuel Albanie, and Matthias Bethge. No "zero-shot" without exponential data: Pretraining concept frequency determines multimodal model performance. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 61735–61792. Curran Associates, Inc., 2024. doi: 10.52202/079017-1973. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/715b78ccfb6f4cada5528ac9b5278def-Paper-Conference.pdf.

Orion Weller and Kevin Seppi. Humor detection: A transformer gets the last laugh. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3621–3625, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1372. URL https://aclanthology.org/D19-1372/.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6/.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding, 2024. URL https://arxiv.org/abs/2412.10302.

Zhongxing Xu, Chengzhi Liu, Qingyue Wei, Juncheng Wu, James Zou, Xin Wang, Yuyin Zhou, and Sheng Liu. More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 82878–82905. Curran Associates, Inc., 2025. doi: 10.52202/085713-2777. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/777db387a5ccb131ba8c7cd155166b85-Paper-Conference.pdf.

Yixin Yang, Zheng Li, Qingxiu Dong, Heming Xia, and Zhifang Sui. Can large multimodal models uncover deep semantics behind images? In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 1898–1912. Association for Computational Linguistics, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11975–11986, October 2023.

Jifan Zhang, Lalit Jain, Yang Guo, Jiayi Chen, Kuan Lok Zhou, Siddharth Suresh, Andrew Wagenmaker, Scott Sievert, Timothy Rogers, Kevin Jamieson, Robert Mankoff, and Robert Nowak. Humor in ai: Massive scale crowd-sourced preferences and benchmarks for cartoon captioning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 125264–125286. Curran Associates, Inc., 2024a. doi: 10.52202/079017-3978. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/e297fb6cd1690ee5b39c5bb4c58ad801-Paper-Datasets_and_Benchmarks_Track.pdf.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *CoRR*, abs/2506.05176, June 2025. URL https://doi.org/10.48550/arXiv.2506.05176.

Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. Why are visually-grounded language models bad at image classification? In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 51727–51753. Curran Associates, Inc., 2024b. doi: 10.52202/079017-1639. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/5c7024041be305c94d7311cfcc53d93e-Paper-Conference.pdf.

Zhengyi Zhao, Shubo Zhang, Yuxi Zhang, Yanxi Zhao, Yifan Zhang, Zezhong Wang, Huimin Wang, Yutian Zhao, Bin Liang, Yefeng Zheng, Binyang Li, Kam-Fai Wong, and Xian Wu. MemeReaCon:

Probing contextual meme understanding in large vision-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *ICLR*, 2024. URL https://openreview.net/forum?id=shr9PXz7T0.

Shanshan Zhong, Zhongzhan Huang, Shanghua Gao, Wushao Wen, Liang Lin, Marinka Zitnik, and Pan Zhou. Let's think outside the box: Exploring leap-of-thought in large language models with creative humor generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13246–13257, June 2024.

Naitian Zhou, David Jurgens, and David Bamman. Social meme-ing: Measuring linguistic variation in memes. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3005–3024, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.166. URL https://aclanthology.org/2024.naacl-long.166/.

A Model Inference Details

Experiments are conducted on NVIDIA RTX A6000 (48 GB) and NVIDIA A100 (80 GB) GPUs using vLLM 0.17.1 (Kwon et al., 2023), Hugging Face Transformers 5.8.0 (4.37.2 for fine-tuning LLaVA-v1.6-Mistral-7B) (Wolf et al., 2020), PyTorch 2.11.0 (2.1.2 for fine-tuning LLaVA-v1.6-Mistral-7B) (Paszke et al., 2019), and PEFT 0.15.2 (Mangrulkar et al., 2022) with CUDA 13.0. Gemma-4-12B was not supported on the installed vLLM version and was instead run using Transformers with greedy decoding using the same prompt, chat template, and token limits.

The other open models run at temperature 0.01, and zero-shot runs generate at least 5 tokens. The generation cap is 5,000 tokens for LLaVA-v1.6-Mistral-7B, Qwen2.5-VL-7B, DeepSeek-VL2-tiny, and Phi-4-multimodal (1,000 for Llama-3.2-11B-Vision), 256 for the LoRA-tuned LLaVA-v1.6-Mistral-7B, Qwen2.5-VL-7B, and Llama-3.2-11B-Vision, and 512 for the Qwen3.5, Qwen3.6, Qwen3.8, Gemma 4, and LLaVA-OneVision models. Direct answers average under 90 characters, far below every cap. Reasoning-mode runs use a 16,384-token budget.

Checkpoints are selected by accuracy on the validation instances whose template appears in training. Qwen2.5-VL-7B, LLaVA-v1.6-Mistral-7B, and Llama-3.2-11B-Vision are trained for one epoch and evaluated at the final checkpoint, with no selection among checkpoints. Gemma-4-E4B is trained for three epochs (3,888 steps) with a checkpoint every 250 steps, and the selected checkpoint is step 3,750. Its adapter also trains the output head and the input embeddings, which the base model ties, so the merged model keeps the two separate; the merged checkpoint gives the adapter's answer on 98.3% of 300 validation instances. CLIP ViT-L/14-336 is fine-tuned for three epochs (7,770 steps), and the selected checkpoint is the final one.

B Additional Details on Data Preprocessing

This appendix traces the six filtering steps summarized in §3.1 (Figure 4), reports the memes remaining after each step (Table 3), and gives examples for each step. The OCR rerun in step 5 restores 17,936 of the 117,967 captions that fail the dictionary test, and step 6 also drops 188 memes whose recovered caption duplicates another meme's caption.

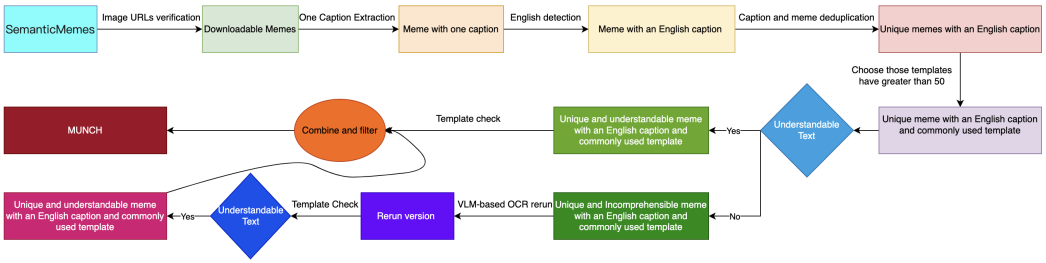


Figure 4: Construction of MUNCH. Six filtering steps reduce 3.3M downloadable memes from SEMANTICMEMES to 127,204 single-caption English memes from 886 templates, whose captions are then inpainted out.

B.1 One Caption Extraction

Memes in SEMANTICMEMES carry zero, one, or several captions inside the image (Figure 5a, Figure 5b, and Figure 5c). We estimate a meme's number of captions as the number of bounding boxes in its OCR output, and keep memes with exactly one bounding box together with the caption that box contains.

B.2 Removing Duplicate Captions and Memes

The same meme is often posted to several subreddits. The meme in Figure 6, for example, appears in both *r/meme* and *r/AmongUsMemes*, identical except for image size. Keeping both copies could place the same meme in training and in a test partition, so step 3 removes duplicate captions and memes.

Step	Memes remaining
Downloadable memes	3,348,252
(1) One caption	433,652
(2) English	385,920
(3) Caption and meme deduplication	280,045
(4) Templates with $\geq$50 instances	232,503
(5) Understandability test, passing captions	114,536
plus captions recovered by the OCR rerun	132,472
(6) Template check and caption re-deduplication	127,204
Released as instances (Table 1)	123,118

Table 3: Memes remaining after each filtering step. Of the 117,967 captions that fail step 5, the OCR rerun recovers 17,936. Step 6 removes 82 templates (5,080 memes) and 188 memes whose recovered caption duplicates another.



(a) No caption (b) One caption (c) Two captions

Figure 5: Memes with zero, one, and two captions. MUNCH keeps memes with exactly one caption.

B.3 NON-STANDARD ENGLISH IN MEMES

Meme captions often use non-standard English, such as jargon, slang, and abbreviations. The caption in Figure 7a names *r/historymemes*, a Reddit community. The caption in Figure 7b uses *sheesh*, a slang term that expresses disbelief or excitement, and the caption in Figure 7c uses *ASAP*, an abbreviation of "as soon as possible."

B.4 RERUNNING OCR WITH AN MLLM

SEMANTICMEMES extracts caption text with the `EasyOCR`[1] package, whose output our manual check found to be garbled for some memes. For Figure 8, EasyOCR returns "When you work at a bank and 3 with mask come in but {leyreihstrobbog ee pace", which is correct only at the start. We therefore rerun OCR with idefics2-8b (Laurençon et al., 2024), prompted with `What is the text in this image? Please directly print out the text.` For Figure 8, idefics2-8b returns "When you work at a bank and 3 guys with mask come in but they're just robbing the place". Weighting the two audited strata by their corpus shares, 86.5% of the sampled captions were transcribed exactly and 12.9% carried minor errors such as casing or punctuation, giving the 99.5% of instances whose meaning was preserved that §3.5 reports.

B.5 TEMPLATE CHECK AND RESULTS

Step 6 checks one meme per template by hand and removes the templates whose format is not a single text fill.

[1] `https://github.com/JaidedAI/EasyOCR`



(a) Posted in *r/meme* (b) Posted in *r/AmongUsMemes*

Figure 6: The same meme posted to two subreddits.



(a) Jargon (b) Slang (c) Abbreviation

Figure 7: Memes that use non-standard English.

B.5.1 Manual Template Annotation

Checking every meme by hand is infeasible, so the check covers one meme per template. Of the 990 templates, 22 had no memes left after the understandability test, and from each of the other 968 we select one meme at random. Each selected meme is annotated as a multiple-choice question in the Potato tool (Pei et al., 2022) (Figure 9). The check kept 886 templates and removed 82 whose format is not a single text fill, for example because OCR merged two captions into one.



Figure 8: A meme from SEMANTICMEMES whose EasyOCR transcription is garbled.

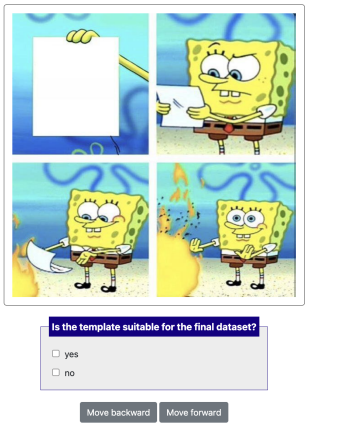
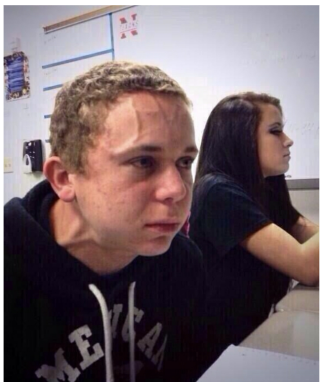


Figure 9: The POTATO interface used for the template check.

B.5.2 TEMPLATE ISSUES

The idefics2-8b rerun improves transcription but leaves two failures that the template check removes. When a meme has no caption, both EasyOCR and idefics2-8b still return one (Figure 10a). When EasyOCR merges several captions into one bounding box, the meme passes as a single-caption meme, and idefics2-8b also returns the merged text as one caption (Figure 10b).



(a) Both methods detect a caption, but the meme has none.



(b) Three separate captions merged into one.

Figure 10: Template issues removed by the manual check.

C Inpainting Process

We remove captions with LaMa, an inpainting network built on fast Fourier convolutions (Suvorov et al., 2022). LaMa takes the color image and a mask, which we draw from the OCR bounding box as a white (255) rectangle on a black (0) background (Figure 11). We run LaMa through the `simple-lama-inpainting`[2] package. Every meme belongs to one of the 886 templates, so inpainting one image per template covers the whole dataset.

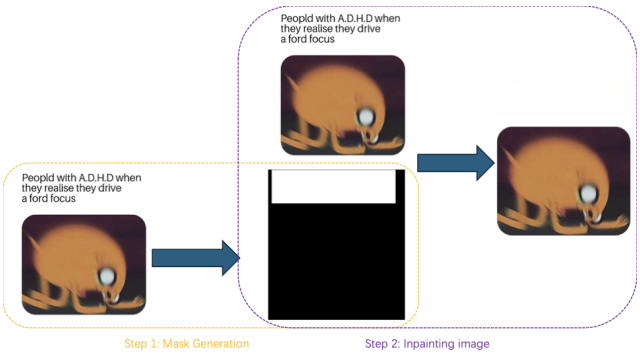


Figure 11: Inpainting a template. The OCR bounding box gives the mask (step 1), and LaMa fills the masked region of the meme (step 2).

D Data Partitioning

The number of memes per template and per cluster constrains how memes can be divided among the training, dev, and test partitions ($\mathcal{KT}$, $\mathcal{HT}$, $\mathcal{HC}$), so the split runs in stages. We first divide the clusters into those with a single template and those with several templates, and then randomly select 92.5% of the dataset and split it as follows.

- For clusters with a single template, we split the data as follows:
  - 80%[3] to Training.
  - 13% to $\mathcal{KT}$.

[2] `https://github.com/enesmsahin/simple-lama-inpainting`

[3] Some data points have to be excluded from $\mathcal{KT}$ and $\mathcal{HT}$, and an 80% share aims for about 65% of the data in training.

  - 7% to Dev.
- For clusters with multiple templates, we process each cluster individually:
  - If a template has fewer than 100 instances[4], its instances are split as follows.
    * 67% to $\mathcal{HT}$.
    * 33% to Dev.
  - This allocation continues until $\mathcal{HT}$ reaches 6,400 instances, about 5% of the dataset. The remaining data in multi-template clusters is split as follows.
    * 80% to Training.
    * 13% to $\mathcal{KT}$.
    * 7% to Dev.

The remaining 7.5% of the dataset is split as follows.

- 67% to $\mathcal{HC}$.
- 33% to Dev.

E PROMPT SELECTION

MLLM accuracy can vary with the wording of the instruction, so we compared 10 prompts in a pilot on a small sample of instances (Table 4). The prompt with the highest average accuracy, 28.0%, asks which caption is highly related to the meme template, and the instruction every model receives is built on that prompt (Table 5).

Table 4: Pilot accuracy of the ten candidate prompts.

Prompt	**Accuracy**
[INST] <image> Which one is the most suitable text for the meme? A: content of option A B: content of option B C: content of option C D: content of option D [/INST]	23.40%
[INST] <image> Which caption is the most suitable text for the meme image? A: content of option A B: content of option B C: content of option C D: content of option D [/INST]	27.00%
[INST] <image> Which caption best describes the meme image? A: content of option A B: content of option B C: content of option C D: content of option D [/INST]	25.00%
[INST] <image> Which of the following captions do you think suits this meme the best? A: content of option A B: content of option B C: content of option C D: content of option D [/INST]	27.60%

[4]This rule places more templates in $\mathcal{HT}$.

Prompt	Accuracy
[INST] <image> Given the meme image, which caption is the most fitting? A: content of option A B: content of option B C: content of option C D: content of option D [/INST]	24.30%
[INST] <image> One of the captions is highly related to the meme template, which one is it? A: content of option A B: content of option B C: content of option C D: content of option D [/INST]	28.00%
[INST] <image> Meme includes one caption and one text. Which one is the correct caption after removing the meme's caption among these four options? A: content of option A B: content of option B C: content of option C D: content of option D [/INST]	24.40%
[INST] <image> The meme's caption is missing, please reason the fit one. A: content of option A B: content of option B C: content of option C D: content of option D [/INST]	26.60%
[INST] <image> Firstly, please try to understand the meme template and find one caption below that is mostly related to the template. A: content of option A B: content of option B C: content of option C D: content of option D [/INST]	26.60%
[INST] <image> You are shown a meme image. Your task is to determine which caption is the most appropriate for this meme. Choose the option that best matches the context and humor of the meme. A: content of option A B: content of option B C: content of option C D: content of option D [/INST]	27.40%

F THE PROMPT FORMAT FOR MLLMS

Table 5 gives the full prompt for six MLLMs, each in its own chat template. The Qwen3.5, Qwen3.6, Qwen3.8, and LLaVA-OneVision models and the other Gemma 4 sizes receive the same instruction through their own chat templates.

Table 5: The full prompt for six MLLMs, each in its own chat template.

Model name	Prompt
LLaVA-v1.6-Mistral-7B	[INST] <image> One of the following four captions is highly related to the meme template. Identify the correct one (output ONLY the answer): A: content of option A B: content of option B C: content of option C D: content of option D Requirements: 1. Strictly start with "Answer: " followed by the option letter (A, B, C, or D) and its full content. 2. No explanations, punctuation, or line breaks allowed. 3. If uncertain, guess the most likely option. [/INST]
Llama-3.2-11B-Vision-Instruct	<|begin_of_text|><|start_header_id|>user<|end_header_id|> <|image|> One of the following four captions is highly related to the meme template. Identify the correct one (output ONLY the answer): A: content of option A B: content of option B C: content of option C D: content of option D Requirements: 1. Strictly start with "Answer: " followed by the option letter (A, B, C, or D) and its full content. 2. No explanations, punctuation, or line breaks allowed. 3. If uncertain, guess the most likely option. <|eot_id|><|start_header_id|>assistant<|end_header_id|>
Qwen2.5-VL-7B-Instruct	<|im_start|>user <|vision_start|><|image_pad|><|vision_end|>One of the following four captions is highly related to the meme template. Identify the correct one (output ONLY the answer): A: content of option A B: content of option B C: content of option C D: content of option D Requirements: 1. Strictly start with "Answer: " followed by the option letter (A, B, C, or D) and its full content. 2. No explanations, punctuation, or line breaks allowed. 3. If uncertain, guess the most likely option.<|im_end|> <|im_start|>assistant
DeepSeek-VL2-tiny	<|User|>: <image> <|ref|>One of the following four captions is highly related to the meme template. Identify the correct one (output ONLY the answer): A: content of option A B: content of option B C: content of option C D: content of option D Requirements: 1. Strictly start with "Answer: " followed by the option letter (A, B, C, or D) and its full content. 2. No explanations, punctuation, or line breaks allowed. 3. If uncertain, guess the most likely option<|/ref|>. <|Assistant|>:

Table 5 (cont.)

Model name	Prompt
Phi-4-multimodal-instruct	<|user|><|image_1|>One of the following four captions is highly related to the meme template. Identify the correct one (output ONLY the answer): A: content of option A B: content of option B C: content of option C D: content of option D Requirements: 1. Strictly start with "Answer: " followed by the option letter (A, B, C, or D) and its full content. 2. No explanations, punctuation, or line breaks allowed. 3. If uncertain, guess the most likely option<|end|><|assistant|>
Gemma-4-E4B	<bos><|turn>user <|image|>One of the following four captions is highly related to the meme template. Identify the correct one (output ONLY the answer): A: content of option A B: content of option B C: content of option C D: content of option D Requirements: 1. Strictly start with "Answer: " followed by the option letter (A, B, C, or D) and its full content. 2. No explanations, punctuation, or line breaks allowed. 3. If uncertain, guess the most likely option<turn|> <|turn>model

G Human Evaluation Details

Two annotators, both fluent in English and familiar with memes, labeled the same 200 instances in the Potato tool (Pei et al., 2022) (Figure 12), with the instruction "Select the most appropriate caption for a meme image from a list of four options." Every instance has a known answer, so we score each annotator against the gold caption, as human-performance estimates for other benchmarks do (Nangia & Bowman, 2019). Inter-annotator agreement suits annotators who construct the ground truth (Artstein & Poesio, 2008), whereas these annotators perform the same task as the models. The two annotators differ in meme familiarity, and each score is an estimate of human performance on that task. The 200 instances include 33 templates that appear more than once (up to nine times). Because the partitions differ only in what a model saw in training, we report each annotator's macro-over-cluster accuracy over all 200 instances, 54.7% and 75.2%, and the human performance estimate in the main text is the more familiar annotator's 75.2%. Pooled over instances, the two annotators score 53.0% and 81.0%.

H Systems Without the Image

Table 6 reports the systems of §5 that never see the image. Every system not trained on Munch scores within 2.6 points of chance on every partition. With the image removed, LoRA-tuned Gemma-4-E4B, whose training showed the model the captions of every training template, scores 26.0, 25.7, and 25.3% on $\mathcal{KT}$, $\mathcal{HT}$, and $\mathcal{HC}$. On the 1,000-instance subsets, seven zero-shot MLLMs score 22.0 to 30.5% with the image removed and 31.3 to 61.2% with it. Scores without the image spread wider on the subsets than on the full partitions, as expected when 1,000 instances are divided among 21 to 249 clusters. Replacing the image with a gray one gives 19.6 to 31.2%.

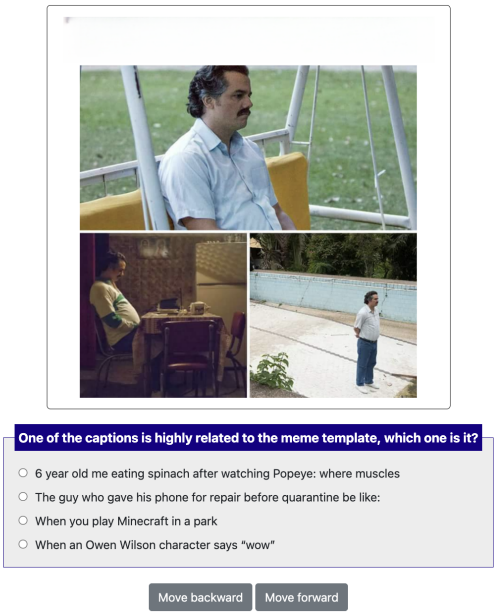


Figure 12: The annotation interface used to label which caption is most appropriate for a meme image.

System	Trained on MUNCH	$\mathcal{KT}$	$\mathcal{HT}$	$\mathcal{HC}$
Full partitions				
Chance	–	25.0	25.0	25.0
Always pick the longest option	no	25.7	25.1	23.7
Text-only LLM (Qwen2.5-7B-Instruct)	no	26.8	24.7	27.6
MLLM, image removed (Qwen2.5-VL-7B)	no	24.1	25.3	25.1
Fine-tuned MLLM, image removed (Gemma-4-E4B, LoRA)	yes	26.0	25.7	25.3
1,000-instance subsets, image removed (with the image in parentheses)				
Qwen3.5-4B	no	26.1 (36.5)	26.9 (41.4)	27.0 (42.0)
Qwen3.6-27B	no	23.8 (46.5)	26.0 (50.4)	24.1 (51.9)
Qwen3.8-27B	no	24.0 (43.9)	26.0 (47.4)	26.4 (49.1)
Gemma-4-E4B	no	23.6 (31.3)	23.1 (33.8)	23.2 (37.8)
LLaVA-OneVision-7B	no	26.6 (40.3)	24.3 (43.5)	26.6 (43.5)
Claude Haiku 4.5	no	23.4 (39.6)	25.2 (42.5)	26.7 (37.8)
Gemini 3.5 Flash-Lite	no	25.2 (61.2)	22.0 (58.3)	30.5 (53.9)

Table 6: Macro-over-cluster accuracy (%) of systems that never see the image, averaged over the 249, 38, and 21 clusters of $\mathcal{KT}$, $\mathcal{HT}$, and $\mathcal{HC}$. On the full partitions every system scores within 2.6 points of chance, including the LoRA-tuned model, which was trained on MUNCH with its images. The 1,000-instance subsets cover the same clusters, and on them every model scores 22.0 to 30.5 with the image removed and 31.3 to 61.2 with it, so the options without the image do not reveal the answer.

I ANSWER EXTRACTION

MLLM responses often depart from the requested `Answer: X <caption>` format, so we recover the chosen option with a five-step extractor applied in order. (1) An explicit answer statement such as "Answer: B" or "The correct option is C" gives the letter. (2) A response consisting only of a letter, such as "A." or "(C)", gives that letter. (3) A response that begins with a letter followed by punctuation and text gives that letter. (4) A response that contains an option's text verbatim gives that option. (5) Otherwise, the option with the highest fuzzy-match ratio to the response is chosen if the ratio is at least 70. Responses that no step matches count as wrong, so accuracy is computed over all instances. Table 7 also reports exact-format accuracy, which requires the response to equal `Answer:` followed by the gold letter and the full gold caption. Every exact-format response is parsed

by step 1, so answer accuracy is never lower than exact-format accuracy. For reasoning-mode runs, the extractor reads only the text after the reasoning, and a response whose reasoning never reaches a final answer counts as wrong.

J Caption Prior Details

The caption priors of §5 never read the image. Each scores an option by its semantic similarity to the training captions of a template or cluster. Writing $C(t)$ for the captions of template t in the training partition, the template prior embeds every caption with Qwen3-Embedding-4B (Zhang et al., 2025), averages the vectors of $C(t)$ into a centroid, and selects the option whose vector has the highest cosine similarity to that centroid. The cluster prior is identical except that it averages over $\bigcup_{t' \in s(t)} C(t')$, every training caption of the instance's semantic cluster. Each prior is defined only where the relevant template or cluster occurs in training, so Figure 2a leaves the template prior undefined on $\mathcal{HT}$ and $\mathcal{HC}$ and the cluster prior undefined on $\mathcal{HC}$.

The priors found by image replace the given template or cluster with one matched from the image. Each training template is represented by the CLIP ViT-L/14-336 embedding of its inpainted image $\tilde{t}$, the test image is matched to the nearest such template by cosine similarity, and the prior of that template or of its cluster is applied. The match is defined on every partition, so these are the only priors reported on $\mathcal{HC}$. A $\mathcal{KT}$ instance shows the same inpainted image as its training template, so on $\mathcal{KT}$ the match is exact, and the priors found by image equal the priors told the template or cluster to within 0.2 points, the difference coming from recomputed caption embeddings.

K Full Results

Table 7 reports exact-format and answer accuracy for every system in Table 2, Table 8 compares reasoning mode with direct answers, and Table 9 compares the standard distractors with MUNCH-HARD on the same instances.

L Error Analyses

Four analyses characterize the memes that MLLMs answer wrongly. They use the 11 MLLMs of our original evaluation, three of them LoRA fine-tuned, rather than the full roster of Table 2.

Are some memes more difficult than others? Figure 13 counts, for each instance, how many of the 11 models answer it correctly. Few instances sit at the extremes. All 11 models answer about 3% of instances wrongly and about 1% correctly. The distribution peaks at three and four correct models, and most instances are answered correctly by fewer than half of the models, so many memes are difficult for most of these MLLMs, though few are missed by all of them.

Does fine-tuning help MLLMs on difficult memes? Figure 14 groups instances by how many models answer them correctly and shows the share of those correct answers that come from the three LoRA fine-tuned models. If correct answers were spread evenly across the 11 models, the LoRA share would be 3 of 11, about 27%. The share is higher in every group but the last, where all 11 models are correct and the share is 27% by construction, because the LoRA models answer more instances correctly overall. The share is highest, about half, among instances that only one to four models answer correctly, suggesting that fine-tuning helps on memes that most models miss. Zero-shot models supply the other half of the correct answers on those instances, so these difficult memes are not out of reach without fine-tuning.

Are memes from some subreddits more difficult? Reddit, the source of our memes, hosts many meme communities, some focused on a topic such as trebuchets or anime, and memes from a topical community may be more difficult for MLLMs. We fit a logistic regression of whether a zero-shot model's answer is correct on the subreddit, the template, the length of the correct caption, and the model. Subreddits with fewer than 30 instances are pooled into a single "Other" category, as are templates with fewer than 30 instances.

	Exact-format accuracy				Answer accuracy			
Model	$\mathcal{KT}$	$\mathcal{HT}$	$\mathcal{HC}$	*All*	$\mathcal{KT}$	$\mathcal{HT}$	$\mathcal{HC}$	*All*
Zero-shot, open MLLMs								
Qwen3.5-0.8B	29.9	31.1	30.5	30.3	30.2	31.6	30.9	30.7
Qwen3.5-2B	30.4	31.0	35.8	31.9	32.0	33.2	38.0	33.8
Qwen3.5-4B	34.1	36.5	41.0	36.3	34.2	36.7	41.1	36.5
Qwen3.5-9B	35.0	38.0	42.2	37.5	35.2	38.2	42.6	37.7
Qwen3.5-27B	42.4	46.5	51.2	45.5	42.5	46.5	51.3	45.6
Qwen3.5-35B-A3B	38.7	41.4	45.0	40.9	38.9	41.5	45.2	41.0
Qwen3.6-27B	43.0	45.2	49.9	45.2	43.0	45.2	49.9	45.2
Qwen3.6-35B-A3B	39.4	41.5	45.5	41.4	39.5	41.7	45.6	41.5
Qwen3.8-27B	40.7	44.4	50.0	43.8	40.7	44.4	50.0	43.8
Gemma-4-E2B	27.2	28.1	29.9	28.1	29.4	30.0	31.7	30.1
Gemma-4-E4B	30.7	33.1	39.3	33.3	30.8	33.2	39.4	33.4
Gemma-4-12B	33.2	35.4	40.6	35.5	34.9	37.1	42.2	37.2
Gemma-4-26B-A4B	37.5	38.3	44.9	39.5	37.7	38.6	45.1	39.7
Gemma-4-31B	43.1	45.3	49.7	45.2	43.4	45.5	49.9	45.4
LLaVA-OneVision-7B	30.5	34.3	38.2	33.3	35.2	38.4	43.6	38.0
LLaVA-v1.6-Mistral-7B	20.1	20.8	22.4	20.8	30.2	31.9	33.3	31.3
Llama-3.2-11B-Vision	28.1	28.9	33.2	29.5	33.4	35.5	40.2	35.6
Qwen2.5-VL-7B	8.8	9.7	11.3	9.6	33.1	35.5	38.1	34.9
Phi-4-multimodal	7.0	7.7	8.2	7.5	28.9	29.9	31.4	29.7
DeepSeek-VL2-tiny	0.4	0.4	0.6	0.4	27.7	27.3	28.2	27.7
Zero-shot, closed MLLMs								
Claude Haiku 4.5	33.6	35.3	39.5	35.4	35.2	37.1	42.1	37.3
GPT-5.6 Luna	35.0	37.9	40.2	36.9	35.7	38.7	41.1	37.7
Gemini 3.5 Flash-Lite	60.3	56.8	58.4	59.0	60.8	57.1	58.7	59.4
Zero-shot, contrastive								
CLIP ViT-B/32	–	–	–	–	44.2	47.1	51.3	46.6
CLIP ViT-L/14-336	–	–	–	–	50.5	53.8	56.2	52.7
SigLIP-so400m	–	–	–	–	37.7	39.7	44.3	39.8
SigLIP2-so400m	–	–	–	–	34.5	36.3	39.2	36.0
OpenCLIP ViT-bigG/14	–	–	–	–	35.4	35.8	46.5	38.2
Fine-tuned on MUNCH								
LLaVA-v1.6-Mistral-7B, LoRA	50.8	43.4	44.5	47.6	51.1	43.9	44.9	47.9
Llama-3.2-11B-Vision, LoRA	70.0	59.2	52.9	63.3	71.2	60.4	53.7	64.4
Qwen2.5-VL-7B, LoRA	77.0	66.7	53.9	69.0	77.1	66.7	53.9	69.0
Gemma-4-E4B, LoRA	76.5	61.7	36.3	63.3	76.8	61.9	36.4	63.5
CLIP ViT-L/14-336, contrastive	–	–	–	–	86.3	78.0	63.0	78.7

Table 7: Exact-format and answer accuracy (%) for every system in Table 2, computed on the same predictions. Exact-format accuracy requires the response to equal `Answer:` followed by the gold letter and the full gold caption. Answer accuracy scores the option recovered by the extractor of Appendix I, which parses every exact-format response, so answer accuracy is never lower. Contrastive models select an option without writing a response, so exact-format accuracy does not apply to them (–).

Memes from general communities are easier to answer (Figure 15). The subreddits with the most negative coefficients are topically niche, focusing on history, a single person, or the Marvel comics universe, and their memes can require background knowledge to read the caption in light of the image. The easiest community, r/antimeme, is unusual. Its captions deliberately lack humor[5] and

[5]This anti-humor format draws its humor from subverting expectations.

	$\mathcal{KT}$			$\mathcal{HT}$			$\mathcal{HC}$		
Model	direct	reason	trunc.	direct	reason	trunc.	direct	reason	trunc.
Qwen3.5-0.8B	31.9	12.8	51	32.9	15.2	51	31.4	17.0	48
Qwen3.5-2B	30.7	17.6	45	33.3	20.7	46	38.9	28.4	38
Qwen3.5-4B	33.2	29.4	23	38.0	29.8	26	41.8	40.1	18
Gemma-4-E4B	30.4	30.8	0	33.2	33.2	0	41.3	37.4	0
Qwen3.5-27B	41.9	40.6	15	44.2	41.3	17	52.6	52.1	13
Qwen3.8-27B	40.6	41.5	1	43.2	43.3	1	49.1	53.7	1

Table 8: Accuracy (%) with direct answers (direct) and in reasoning mode with a 16K-token budget (reason), on a stratified subset of 1,000 instances per partition. The trunc. column gives the percentage of reasoning-mode responses that never reached a final answer, which count as wrong.

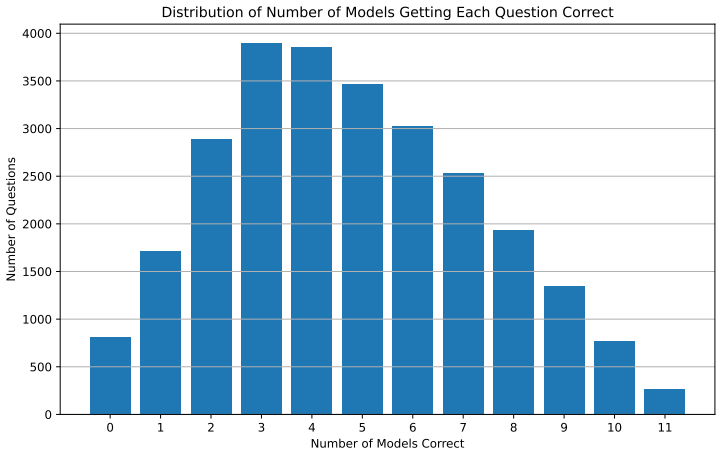


Figure 13: Number of instances answered correctly by a given number of the 11 models (x-axis).

often describe elements of the image, which is closer to the image–text pairs MLLMs are trained on than a joke caption is.

Are some templates more difficult? The template coefficients of the same regression identify the most and least difficult templates. In the templates with the most negative coefficients (Figure 17), the caption appears on a depicted screen or sign as a statement, often arguing for an outcome or highlighting a fact the poster wants more widely known. Such statements can be about anything, so these templates carry semantically diverse captions and are difficult to match to them. The templates with the most positive coefficients (Figure 16) instead have distinctive visual features that their captions refer to directly. Captions for Top 5 frequently mention "the future," and captions for Top 6 describe a situation with "me and the boys," though not every caption does.

	$\mathcal{KT}$			$\mathcal{HT}$			$\mathcal{HC}$		
System	std	hard	Δ	std	hard	Δ	std	hard	Δ
Zero-shot, contrastive									
CLIP ViT-B/32	52.2	39.1	−13.1	51.6	32.0	−19.5	44.8	34.2	−10.7†
CLIP ViT-L/14-336	59.6	44.6	−15.0	56.1	44.3	−11.8	52.1	37.8	−14.3
SigLIP-so400m	39.2	29.0	−10.2	40.6	27.0	−13.7	41.6	24.8	−16.8
SigLIP2-so400m	35.0	27.4	−7.6	37.0	30.1	−6.9	39.5	32.9	−6.5†
OpenCLIP ViT-bigG/14	38.5	28.9	−9.7	37.8	23.4	−14.5	43.0	21.9	−21.2
Zero-shot, image–text matching									
BLIP matching head	36.6	24.1	−12.5	35.1	23.4	−11.7	41.5	25.9	−15.6
BLIP-2 matching head	35.1	21.9	−13.3	35.0	28.5	−6.5†	30.4	26.5	−3.9†
BLIP-2 captioner (PMI)	35.6	26.6	−9.0	31.2	29.2	−2.0†	35.3	30.5	−4.8†
VQAScore	30.5	22.8	−7.8	33.4	23.2	−10.2	30.5	27.0	−3.5†
Zero-shot, open MLLMs									
Gemma-4-E4B	31.3	37.9	+6.7	33.8	33.6	−0.2†	37.8	26.4	−11.4
Qwen3.5-4B	36.5	36.4	−0.1†	41.4	37.1	−4.3†	42.0	32.2	−9.8†
Qwen3.5-27B	46.1	41.7	−4.4†	48.6	43.1	−5.6†	51.7	41.2	−10.5†
Qwen3.6-27B	46.5	44.0	−2.4†	50.4	43.6	−6.8†	51.9	38.3	−13.6†
Gemma-4-31B	51.1	47.1	−4.0†	52.5	39.8	−12.8	37.8	43.6	+5.9†
Zero-shot, closed MLLMs									
Claude Haiku 4.5	39.6	36.0	−3.6†	42.5	33.8	−8.7	37.8	28.6	−9.2†
GPT-5.6 Luna	40.6	38.4	−2.2†	44.7	37.3	−7.3	38.8	35.6	−3.3†
Gemini 3.5 Flash-Lite	61.2	44.1	−17.1	58.2	47.5	−10.7	53.9	43.6	−10.3†
Fine-tuned on MUNCH									
CLIP ViT-L/14-336, contrastive	84.1	69.7	−14.4	78.5	62.4	−16.2	65.6	55.9	−9.7†
Gemma-4-E4B, LoRA	69.7	59.4	−10.3	61.0	52.6	−8.5	47.6	41.5	−6.2†
Qwen2.5-VL-7B, LoRA	70.1	59.0	−11.1	69.9	56.9	−12.9	57.9	45.1	−12.8
Llama-3.2-11B-Vision, LoRA	58.6	49.9	−8.6	60.0	48.7	−11.3	54.7	40.2	−14.5

Table 9: Macro-over-cluster accuracy (%) on the same 1,000 instances per partition with the standard distractors (std) and with MUNCH-HARD distractors (hard), and the difference Δ (hard minus std). A dagger marks a difference whose 95% two-stage cluster-bootstrap interval includes zero; $\mathcal{HC}$ averages over 21 clusters, so most of its differences are not resolved at this sample size. Caption priors are omitted because MUNCH-HARD selects its distractors with the embedding model the priors use (§4.3). The LoRA-tuned Qwen2.5-VL-7B and Llama-3.2-11B-Vision rows rerun the released LoRA adapters under current libraries for both distractor sets. The rerun reproduces 99.2–99.7% of the original Qwen2.5-VL-7B answers and 84–90% of the original Llama-3.2-11B-Vision answers, and the rerun accuracy of Llama-3.2-11B-Vision is 0.9–2.4 points below the original.

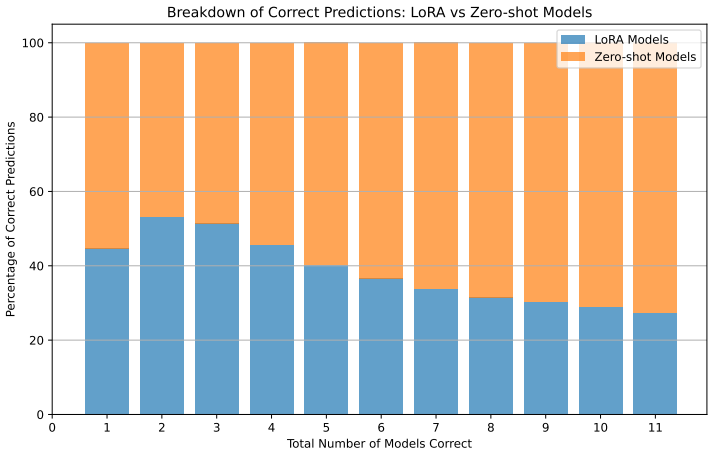


Figure 14: Share of correct answers from LoRA fine-tuned and zero-shot models, for instances grouped by the number of models that answer them correctly (x-axis).

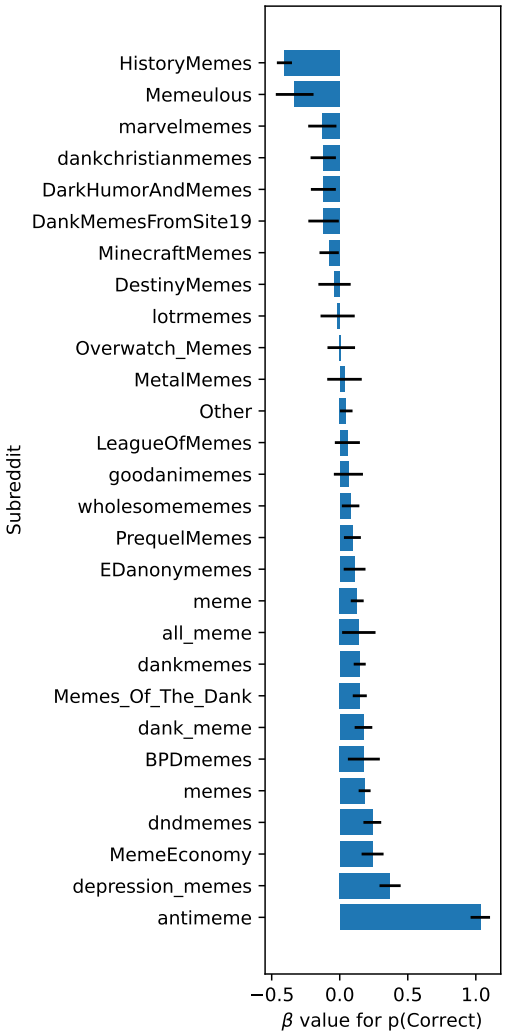
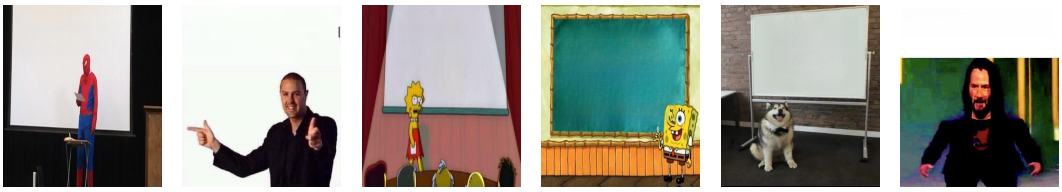


Figure 15: Subreddit coefficients (β) in a logistic regression of whether a zero-shot MLLM answers correctly. A positive coefficient means a correct answer is more likely. Memes from general meme communities are answered correctly more often.



(a) Top 1 (b) Top 2 (c) Top 3 (d) Top 4 (e) Top 5 (f) Top 6

Figure 16: The six templates with the most positive coefficients in the regression of Figure 15, whose memes are most likely to be answered correctly. Their captions often refer to visual cues in the template.



(a) Bottom 1 (b) Bottom 2 (c) Bottom 3 (d) Bottom 4 (e) Bottom 5 (f) Bottom 6

Figure 17: The six templates with the most negative coefficients in the regression of Figure 15, whose memes are least likely to be answered correctly. Their captions post a message to argue for it or highlight its importance, a format that supports diverse uses.