Learning to Think Like a Cartoon Captionist: Incongruity–Resolution Supervision for Multimodal Humor Understanding

Hatice Merve Vural[1] **Doga Kukul**[1,2] **Ege Erdem Ozlu**[1] **Demir Ekin Arikan**[1]
Bob Mankoff[3] **Erkut Erdem**[2,4] **Aykut Erdem**[1,2]
[1]Koç University, Istanbul, Turkey [2] KUIS AI Center, Istanbul, Turkey
[3] Air Mail and Cartoon Collections [4] Hacettepe University, Ankara, Turkey

https://cyberiada.github.io/NYCC-Thinking/

Abstract

Humor is one of the few cognitive tasks where getting the reasoning right matters as much as getting the answer right. While recent work evaluates humor understanding on benchmarks such as the New Yorker Cartoon Caption Contest (NYCC), it largely treats it as black-box prediction, overlooking the structured reasoning processes underlying humor comprehension. We introduce IRS (Incongruity-Resolution Supervision), a framework that decomposes humor understanding into three components: *incongruity modeling*, which identifies mismatches in the visual scene; *resolution modeling*, which constructs coherent reinterpretations of these mismatches; and *preference alignment*, which evaluates candidate interpretations under human judgments. Grounded in incongruity-resolution theory and expert captionist practice, IRS supervises intermediate reasoning process through structured traces that make the path from visual perception to humorous interpretation explicit and learnable. Across 7B, 32B, and 72B models on NYCC, IRS outperforms strong open and closed multimodal baselines across caption matching and ranking tasks, with our largest model approaching expert-level performance on ranking. Zero-shot transfer to external benchmarks shows that IRS learns generalizable reasoning patterns. Our results suggest that supervising reasoning structure, rather than scale alone, is key for reasoning-centric tasks.

1 Introduction

Humor is a demanding facet of human intelligence, requiring the integration of visual perception, cultural knowledge, and creative reasoning (Kazemi et al., 2025). Unlike standard benchmark tasks, humor is not defined by a single correct answer but by a reasoning process: identifying a mismatch between expectation and observation, and resolving it in a coherent yet surprising way. This incongruity-resolution dynamic, long studied in cognitive science (Suls, 1972; Attardo, 1994), suggests that humor understanding is fundamentally a structured reasoning problem — yet most computational approaches treat it as a prediction task, training models to select or rank captions without modeling the interpretive steps that make a caption funny. Cartoon captioning makes this gap concrete. A strong caption does not merely describe a scene; it identifies a salient visual tension and reframes it into a meaningful punchline. This requires knowing what the incongruity is, how it can be resolved, and why one resolution is funnier than another; capabilities that current multimodal language models only partially capture, as reflected in their gap to human performance on humor benchmarks (Hessel et al., 2023; Zhou et al., 2025).

The New Yorker Cartoon Caption Contest (NYCC) provides a structured setting to study this problem, pairing each image with thousands of crowd-submitted captions, expert editorial curation, and large-scale audience judgments and creating a rare alignment between visual input, linguistic creativity, expert reasoning, and human preferences. While prior work primarily treats NYCC as a classification or ranking benchmark, it also offers insight into the reasoning processes underlying visual humor.

We argue that improving humor understanding requires supervising these reasoning processes directly. To this end, we introduce *Incongruity-Resolution Supervision (IRS)*, a framework that decomposes humor understanding into three components: *incongruity modeling*, which identifies mismatches in the visual scene; *resolution modeling*, which constructs coherent reinterpretations of those mismatches; and *preference alignment*, which evaluates candidate interpretations under human judgments. This decomposition is grounded in incongruity-resolution theory and informed by expert analyses of professional captionist practice (Wood, 2024; Mankoff, 2002), and implemented through *captionist reasoning traces*, which provide structured supervision over intermediate reasoning steps. Across 7B, 32B, and 72B multimodal models, IRS consistently improves performance on NYCC matching and ranking tasks. Our largest model approaches expert-level ranking performance, and zero-shot transfer to external humor benchmarks suggests that IRS learns generalizable reasoning patterns rather than dataset-specific heuristics.

Our contributions are as follows:

- We introduce Incongruity-Resolution Supervision (IRS), a theoretically grounded framework that decomposes humor understanding into three explicit, learnable stages bridging cognitive theory and multimodal learning. IRS is implemented through *captionist reasoning traces*, structured supervision signals derived from expert practice that make intermediate reasoning over visual humor explicit and learnable.
- We demonstrate that IRS substantially improves both performance and reasoning quality across model scales on NYCC, and that the learned reasoning patterns transfer zero-shot to out-of-domain humor benchmarks, establishing explicit reasoning supervision as a scalable and generalizable approach to creative multimodal understanding.

2 Related Work

Computational Humor. Humor is commonly modeled through incongruity-resolution theory (Suls, 1972; Attardo, 1994), which frames it as the violation of an expectation followed by a coherent reinterpretation. Related perspectives, including script opposition (Attardo & Raskin, 1991), frame shifting (Coulson & Kutas, 2001), and Theory of Mind accounts (Samson, 2012), further highlight its structured cognitive nature. While large language models (LLMs) capture surface-level patterns of humor, they remain limited on tasks requiring deeper reasoning (Trott et al., 2025). Early computational approaches relied on rule-based templates for jokes and puns (Ritchie, 2001; Mihalcea & Strapparava, 2006; Doogan et al., 2017), later replaced by data-driven methods with curated humor corpora (Hossain et al., 2019; 2020) and large-scale datasets (West & Horvitz, 2019; Horvitz et al., 2024). Recent benchmarks extend to visually grounded humor, including HumorDB (Jain et al., 2025), YesBut (Hu et al., 2024), DeepEval (Yang et al., 2024), and Oogiri (Zhong et al., 2024). However, these datasets primarily evaluate outcomes (e.g., selecting the funnier caption) rather than modeling the underlying reasoning processes emphasized in cognitive theory.

The New Yorker Cartoon Caption Contest. The New Yorker Cartoon Caption Contest (NYCC) is a key resource for studying multimodal humor, aligning visual input, linguistic creativity, expert judgment, and crowd preferences. Hessel et al. (2023) introduced a benchmark derived from NYCC spanning caption matching, ranking, and explanation tasks, with later work extending evaluation through harder ranking splits (Zhou et al., 2025) and large-scale preference analysis (Zhang et al., 2024). Recent approaches incorporate humor theory into modeling. Shang et al. (2026) propose HOMER, a generation framework grounded in the General Theory of Verbal Humor (Attardo & Raskin, 1991), showing improved caption quality over prompting baselines. IRS is complementary: while HOMER

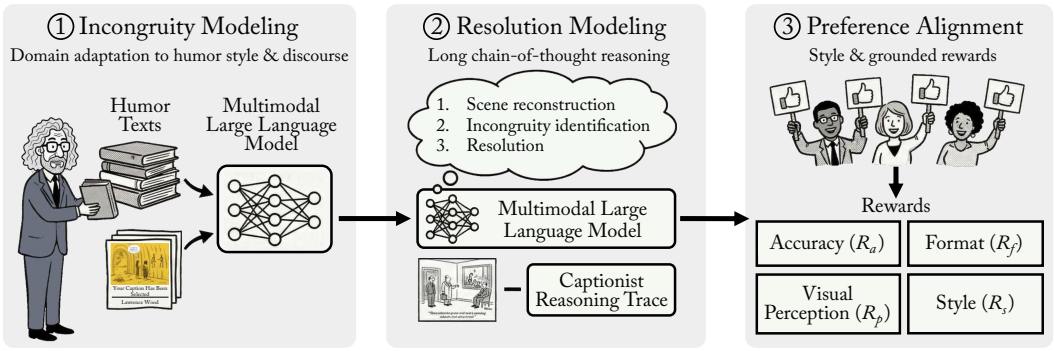


Figure 1: **Overview of Incongruity-Resolution Supervision (IRS).** IRS models humor understanding as a structured reasoning process with three components: (1) *incongruity modeling*, which identifies mismatches in the visual scene; (2) *resolution modeling*, which constructs coherent reinterpretations of these mismatches; and (3) *preference alignment*, which evaluates candidate interpretations under human judgments. The training stages, continual pretraining, supervised learning on captionist reasoning traces, and reinforcement learning with grounded rewards, support these components.

focuses on caption generation with closed models, IRS targets humor understanding via explicit supervision of intermediate reasoning in open-weight models.

Multimodal Reasoning and Alignment. Recent work explores equipping multimodal LLMs with explicit reasoning via structured traces distilled from stronger models, improving performance and reducing hallucination (Xu et al., 2025; Thawakar et al., 2025; Liu et al., 2025a; Liao et al., 2025). However, these approaches generally rely on generic chain-of-thought formats rather than task-specific reasoning structures. Reinforcement learning with grounded rewards offers a complementary alignment strategy. Methods such as Vision-R1 (Zhan et al., 2025), Visual-RFT (Liu et al., 2025b), and Perception-R1 (Xiao et al., 2025) improve visual grounding and reduce spurious correlations through task-specific reward signals, but do not explicitly model the structure of reasoning itself.

3 Approach

Analyses of professional captionist practice (Wood, 2024; Mankoff, 2002) reveal a consistent interpretive process: identifying a salient visual incongruity, constructing a coherent reinterpretation, and evaluating punchlines by how effectively they resolve the tension in a surprising yet meaningful way. IRS models this process through three learnable components: *Incongruity Modeling (IM)*, which identifies mismatches in the visual scene; *Resolution Modeling (RM)*, which constructs coherent reinterpretations of those mismatches; and *Preference Alignment (PA)*, which evaluates candidate interpretations against human judgments of humor (Fig. 1). We implement this via *captionist reasoning traces*, structured sequences that make intermediate reasoning from visual perception to humorous interpretation explicit. These traces capture how incongruities are identified, interpretations are constructed, and captions are evaluated, enabling direct supervision of reasoning.

3.1 Incongruity Modeling via Domain-Adaptive Pretraining

Recognizing humor-relevant incongruities requires background knowledge that is largely absent from standard pretraining corpora: stylistic conventions of cartoon humor, editorial heuristics for what makes a caption work, and culturally grounded expectations about narrative and visual surprise. In the IM stage, we perform *continual pretraining* (CPT) on a curated corpus of captionist discussions, editorial analyses, caption-writing guides, and complementary general-knowledge text. CPT uses a standard causal language modeling objective and does not optimize downstream tasks directly. Its role is preparatory: biasing the

The scene in the cartoon shows an airplane cabin with passengers. The uncanny element is a giant amoeba or germ sitting in one of the seats, which is unusual because germs are microscopic. The man next to it is gesturing towards it, likely annoyed.
Looking at the captions:
A) "How'd you get yours off?" – This might relate to masks or something removable, but the amoeba isn't wearing anything. It doesn't fit.
B) "Mind if I cut through?" – This could imply someone trying to pass by, but the amoeba is blocking the aisle. However, the man is pointing at the germ, not asking to pass.
C) "Seven men walked into a bar and died. What's it to ya?" – This seems like a joke setup, but there's no bar in the scene. It's unrelated.
D) "I don't care if he's single-celled, he should have bought two seats." – This references the amoeba being single-celled (a biological term) and taking up space. The man is annoyed about the space, pointing at it. It fits perfectly.
E) " Unfortunately, the co-pay is an arm and a leg." – This is a medical cost joke, but there's no context for it in the cartoon. It doesn't fit.
The key points are the man's annoyance, the amoeba's size, and its encroachment on space. Caption D uses "single-celled" as wordplay, tying the biological aspect to the airline's seat policy. The humor comes from applying human rules to a microscopic organism that is now giant. The man speaking to the woman makes sense as he's complaining about the amoeba's lack of consideration. So D is the best match.

Figure 2: **Example captionist reasoning trace under IRS (matching).** Given a cartoon and five candidate captions (A-E), the trace shows structured reasoning: reconstructing the scene, identifying the key incongruity, evaluating alternatives, and selecting the caption best resolving the mismatch. The final choice links the character's annoyance to the wordplay on "single-celled", illustrating both visual grounding and humor-specific reasoning.

model's representations toward humor-relevant concepts, incongruity-resolution structure, lexical economy, narrative framing, and culturally grounded references, so that subsequent supervised training can build on a more appropriate prior. Models incorporating this stage exhibit more stable and coherent reasoning during fine-tuning, suggesting that domain adaptation meaningfully shapes what the model attends to rather than merely what it predicts. *Full details of the IM corpus are provided in Appendix A.*

3.2 Resolution Modeling via Captionist Reasoning Traces

The core of IRS is the supervision of resolution modeling; teaching the model how incongruities are reinterpreted into coherent, humorous readings. We construct a dataset of *captionist reasoning traces* for both caption matching and ranking tasks. Starting from human-annotated cartoon descriptions provided by Hessel et al. (2023), we generate structured step-by-step analyses using DeepSeek-R1 (DeepSeek-AI et al., 2025), following a fixed expert-derived template that covers scene reconstruction, incongruity identification, and narrative resolution. The generated traces are further verified under human supervision to ensure consistency with expert reasoning patterns. To reflect professional captionist discourse, traces are rephrased using GPT-4o (OpenAI, 2024) to emphasize concise, observational, image-grounded commentary, replacing description-based phrasing with direct visual reference while preserving the underlying reasoning structure. In rare cases where annotations do not fully support the correct caption, we minimally adjust the teacher prompt to ensure alignment with the intended reasoning.

Formally, given an input pair (I, Q) consisting of a cartoon image and its associated question, the model is supervised to produce outputs of the form:

$\langle$`think`$\rangle$ *reasoning steps* $\langle$`/think`$\rangle\langle$`answer`$\rangle$ *final choice* $\langle$`/answer`$\rangle$.

This structured supervision encourages models to internalize captionist-style reasoning rather than relying on surface-level pattern matching. Fig. 2 illustrates a representative trace: the model reconstructs the scene (a giant amoeba on an airplane), rules out off-topic captions, and identifies the correct option by recognizing the wordplay on "single-celled" as the resolution of the visual incongruity. *Prompt templates for reasoning trace generation are provided in Appendix C. Additional examples are included in Appendix D.*

1. Numerous wire cages containing groups of small mice
2. Water bottles attached to the bars of the cages
3. A person wearing a full-body gray mouse costume with exaggerated ears, whiskers, and snout
4. The costumed person holding a clipboard and taking notes
5. A second person wearing a white lab coat
6. The lab-coated person standing slightly behind the costumed person
7. Real mice inside cages watching the costumed figure
8. The mouse costume is human-sized and plush in appearance
9. The costumed person positioned as if studying the mice in cages
10. Cages are stacked in the laboratory setting

Figure 3: **Curated visual references for perception alignment in IRS.** For each cartoon, we collect concise descriptions of entities, scene context, and key incongruities. These references serve as anchors for evaluating whether model reasoning is grounded in salient visual elements and are used to compute the visual perception reward.

3.3 Preference Alignment via Humor-Aware Rewards

Resolution Modeling provides reasoning structure, but does not ensure visual grounding or stylistic consistency. We address this through reinforcement learning using humor-specific rewards, optimized via GRPO (DeepSeek-AI et al., 2025), which directly optimizes the reasoning process without a value network, using the following objective:

$$\mathcal{J}(\theta) = \mathbb{E}_{(I,q)\sim\mathcal{D},\, \{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(O|I,q)}$$
$$\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\left(\frac{\pi_\theta(o_{i,t}\mid q,\, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}\mid q,\, o_{i,<t})}\,\hat{A}_{i,t},\, \text{clip}\left(\frac{\pi_\theta(o_{i,t}\mid q,\, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}\mid q,\, o_{i,<t})},\, 1-\epsilon,\, 1+\epsilon\right)\hat{A}_{i,t}\right)$$
$$-\,\beta\,\text{KL}\left[\pi_\theta \,\|\, \pi_{\text{ref}}\right],$$

where ϵ is the clipping hyperparameter, β is the KL penalty, and π_{ref} is the reference model. Advantages are computed by normalizing rewards across sampled responses:

$$\hat{A}_i = \frac{r_i - \text{mean}(\{r_i\}_{i=1}^{G})}{\text{std}(\{r_i\}_{i=1}^{G})}.$$

Our composite reward integrates standard correctness signals with two humor-specific components. The first two follow common multimodal RLHF practice:

(i) **Accuracy** (R_a): verifies whether the final caption choice is correct (gold caption for matching; crowd-preferred caption for ranking).

(ii) **Format** (R_f): enforces adherence to the structured reasoning format, ensuring outputs remain parseable and interpretable.

These signals provide necessary scaffolding but are insufficient for humor understanding. Our key additions are two humor-specific rewards:

(iii) **Visual Perception** (R_p): rewards reasoning grounded in salient visual elements and incongruities. As illustrated in Fig. 3, we curate up to ten reference descriptions per cartoon, extending the annotations of Hessel et al. (2023), and use Qwen2.5-7B-Instruct (Qwen et al., 2025) as an LLM judge to assess visual grounding.

(iv) **Style** (R_s): evaluates linguistic quality using an LLM-as-judge based on captionist guidelines (Wood, 2024), assessing everyday phrasing, punctuation, wordplay, metaphor, and punchline placement via an LLM-as-judge. A binary score is aggregated into the style reward using the same Qwen2.5-7B-Instruct judge.

Fig. 4 illustrates how these judges operate in practice. The four signals are combined as:

$$R = \lambda_a R_a + \lambda_f R_f + \mathbb{I}[R_a = 1]\left(\lambda_p R_p + \lambda_s R_s\right), \tag{1}$$

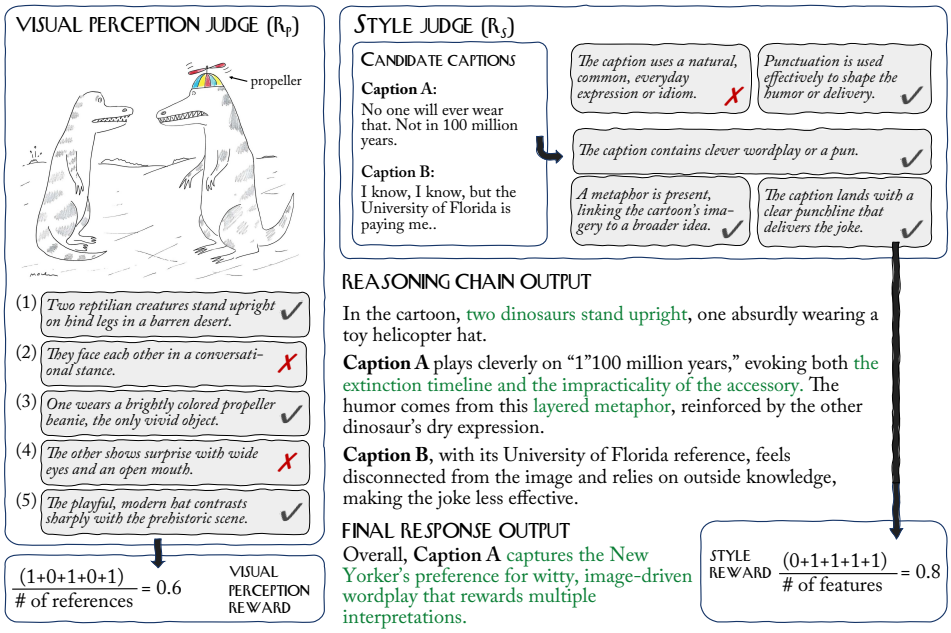


Figure 4: **Preference alignment in IRS via judge-based rewards.** Given a cartoon and candidate captions, two judges evaluate reasoning quality: a *visual perception* judge checks grounding in salient elements and incongruities, while a *style* judge assesses linguistic quality. Their binary outputs are aggregated into perception and style rewards (R_p, R_s), guiding learning toward visually grounded, captionist-consistent reasoning.

conditioning perception and style rewards on correctness to ensure they reinforce rather than distract from accurate caption selection. Following prior work (Liu et al., 2025a; Xiao et al., 2025), we remove the KL penalty, which improves reasoning coherence. When either R_p or R_s exceeds 80% of its maximum, over 70% of cases are correct, indicating that the LLM-based judge provides meaningful guidance rather than spurious optimization signals. Importantly, the judge operates solely on reasoning traces, without access to ground-truth captions, ensuring reward signals reflect reasoning quality rather than answer matching. *Reward-judge prompts are provided in Appendix C.*

4 Experimental Setup

We evaluate Incongruity-Resolution Supervision (IRS) across four benchmark settings spanning caption matching and humor ranking, designed to test complementary aspects of humor reasoning at varying levels of difficulty. Two setups are drawn from Hessel et al. (2023) (matching and ranking), and two from Zhou et al. (2025) (10-vs-1000 and 30-vs-300), which vary the difficulty of preference discrimination. *Training configurations and optimization details for all stages are provided in Appendix B.*

Datasets and Tasks. Our evaluation covers four settings drawn from two NYCC-based benchmarks: From Hessel et al. (2023): (i) **Matching**, in which a model selects the gold caption from five candidates given a cartoon image; and (ii) **Ranking**, in which a model selects the crowd-preferred caption from two options. From Zhou et al. (2025): (iii) **10-vs-1000**, which requires discriminating between a top-10 caption and one ranked 1000–1009; and (iv) **30-vs-300**, which requires discriminating between a caption ranked 30–39 and one ranked 300–309. These tasks target distinct reasoning competencies. Matching tests the ability to identify the correct visual interpretation from a set of distractors, rewarding models that can localize salient incongruities and rule out off-topic options. Ranking emphasizes preference alignment, determining which of two plausible captions better resolves the incongruity, and becomes increasingly demanding as the quality gap between

options narrows. The 30-vs-300 setting is the most challenging, as both captions are crowd-ranked and stylistically competent, requiring fine-grained discrimination that goes beyond incongruity detection alone. To prevent data leakage, all evaluation cartoons are excluded from the continual pretraining corpus, and we verify negligible n-gram overlap between the CPT corpus and evaluation caption text.

Models. We compare IRS against a diverse set of baselines spanning text-only, closed-source, and open-weight systems, allowing us to assess gains attributable to explicit reasoning supervision independently of scale and data access advantages.

- **Text-only reasoning.** DeepSeek-R1 (DeepSeek-AI et al., 2025) is evaluated on ground-truth textual annotations of cartoons rather than raw images. Since it receives curated scene descriptions that multimodal models must infer from pixels, its performance represents an approximate upper bound on reasoning given perfect visual perception rather than a true multimodal comparison.
- **Closed multimodal models.** We include o3 and o4-mini as representative high-performing proprietary systems. Their strong performance may partly reflect access to large-scale paywalled or proprietary corpora during pretraining, including, potentially, New Yorker editorial content via OpenAI's content licensing partnership with Condé Nast (OpenAI, 2024), making them a useful but not strictly comparable reference point for open-weight approaches.
- **Open multimodal reasoning models.** We benchmark against recent open-weight models trained for multimodal reasoning: GLM-4.1V-9B-Thinking (Hong et al., 2025), Kimi-VL-A3B-Thinking-2506 (Team et al., 2025), LlamaV-o1 (Thawakar et al., 2025), and Qwen2.5-VL at 7B, 32B, and 72B scales (Qwen et al., 2025). These models apply general-purpose reasoning strategies without humor-specific supervision.
- **IRS (ours).** We apply IRS to Qwen2.5-VL backbones at 7B, 32B, and 72B scales, implementing all three training stages: domain-adaptive pretraining, supervised fine-tuning on captionist reasoning traces, and preference-based alignment with perceptual and stylistic rewards. Comparing IRS against the corresponding Qwen2.5-VL base models isolates the contribution of IRS-specific supervision from backbone capacity.
- **Human baselines.** We report expert and non-expert human performance on a subset of tasks. The expert captionist, a professional with sustained experience in NYCC, serves as a qualitative reference for what human-level humor reasoning looks like, while a user study with 21 participants provides a measure of alignment with crowd preferences.

Evaluation prompts are provided in Appendix C.

5 Experimental Results

5.1 Comparison with Baselines

Table 1 reports accuracy across all models.

Human performance reveals task-specific difficulty. The expert captionist achieves perfect scores on *matching* and *ranking*, tasks built from NYCC finalist captions — but shows lower agreement with crowd preferences on the 10-vs-1000 and 30-vs-300 settings (60% and 40%, respectively). This is consistent with the known divergence between editorial judgment and popular vote: expert captionists optimize for originality and craft, while crowd preferences reflect broader accessibility. Non-expert human performance is substantially lower across all tasks, confirming that NYCC-style humor reasoning requires sustained domain familiarity rather than general common sense.

Closed models benefit from data advantages that open models lack. Among multimodal baselines, o3 and o4-mini remain highly competitive, with o3 achieving the highest matching accuracy overall. Their strong performance likely reflects both architectural sophistication

Table 1: **Performance across models.** Accuracy (%) on caption matching and ranking tasks. Bold values indicate the best performance in each column, and underlined values indicate the second-best performance (excluding human baselines).

Model	(Hessel et al., 2023)		(Zhou et al., 2025)	
	Matching	Ranking	10-vs-1000	30-vs-300
Human (Expert)	100.00	100.00	60.00	40.00
Human (Non-Expert)	53.03	65.61	54.70	52.27
DeepSeek-R1	74.00	64.67	56.86	47.14
o4-mini	<u>75.08</u>	62.59	60.17	51.42
o3	**83.33**	62.85	**69.05**	**54.57**
GLM-4.1V-9B-Thinking	59.40	55.60	52.28	49.14
Kimi-VL-A3B-Thinking	54.00	57.45	52.30	52.01
LlamaV-o1	43.67	50.90	48.57	49.42
Qwen2.5-VL-7B-Instruct	42.67	55.06	50.57	47.99
Qwen2.5-VL-32B-Instruct	46.67	49.87	52.00	44.57
Qwen2.5-VL-72B-Instruct	56.00	55.58	53.71	50.29
IRS-7B (Our model)	59.67	64.42	56.29	<u>53.14</u>
IRS-32B (Our model)	62.67	<u>68.05</u>	<u>62.86</u>	<u>53.14</u>
IRS-72B (Our model)	69.33	**76.10**	62.57	50.86

and pretraining data advantages, including potential exposure to New Yorker editorial content via licensing arrangements unavailable to open-weight models.[1] DeepSeek-R1, despite operating on curated textual annotations rather than images, performs competitively on matching and ranking, confirming that high-quality perceptual grounding remains a bottleneck for open multimodal models. Among open-weight baselines, scaling alone provides limited benefit: Qwen2.5-VL improves only modestly from 7B to 72B, and even at 72B falls well short of closed-model performance without task-specific reasoning supervision.

IRS consistently improves performance across scales. Models trained with IRS outperform their base counterparts at every scale, with improvements growing as backbone capacity increases. At 7B, IRS narrows the gap to closed models substantially, demonstrating that explicit reasoning supervision transfers even at smaller scales. At 32B, IRS produces the strongest open-weight model across most tasks, with particularly large gains on ranking — the task most sensitive to preference alignment between competing interpretations. At 72B, IRS achieves the highest ranking accuracy overall (76.10%), surpassing all baselines including o3, and approaches expert-level performance on that task. Performance on the 30-vs-300 setting is comparatively less stable, as semantically similar candidates make fine-grained preference distinctions inherently ambiguous.

Models trained with the full IRS pipeline also exhibit improved zero-shot generalization to out-of-domain humor benchmarks (YesBut (Hu et al., 2024), DeepEval (Yang et al., 2024)), suggesting that IRS captures transferable reasoning patterns rather than dataset-specific heuristics (Appendix F). We also report qualitative caption generation results, showing that IRS models produce plausible, captionist-style captions despite not explicitly training for this task (Appendix G).

5.2 Ablation on Sources of Supervision

Table 2 isolates the contribution of each IRS component on the 7B backbone. Three findings stand out. Resolution Modeling (RM) is the main source of improvement. Adding RM to the base model leads to consistent gains across all tasks, highlighting the importance of explicitly modeling how incongruities are interpreted into coherent humorous readings. Incongruity Modeling (IM), by contrast, is only helpful when combined with RM. On its own, it does not

[1] OpenAI has content licensing agreements with Condé Nast (OpenAI, 2024). While this does not imply direct training on NYCC, it may confer stylistic familiarity with New Yorker-style humor.

Table 2: **Ablation on IRS components.** Each component contributes complementary gains; the full IRS pipeline yields the strongest results.

Approach	(Hessel et al., 2023)		(Zhou et al., 2025)	
	Matching	Ranking	10-vs-1000	30-vs-300
Base Model	42.67	55.06	50.57	47.99
+ Incongruity Modeling (IM)	41.00	51.69	50.29	51.43
+ Resolution Modeling (RM)	47.00	56.88	49.71	50.86
+ IM + RM	49.00	56.88	54.57	49.14
+ Preference Alignment (PA)	56.67	58.96	54.86	49.14
+ RM + PA	57.33	58.18	56.29	44.86
+ IM + PA	46.00	51.95	49.43	50.29
+ IM + RM + PA (IRS)	59.67	64.42	56.29	53.14

Table 3: **Ablation on reward functions.** Humor-aware rewards substantially improve ranking and generalization.

Approach	(Hessel et al., 2023)		(Zhou et al., 2025)	
	Matching	Ranking	10-vs-1000	30-vs-300
Base Model + IM + RM	49.00	56.88	54.57	49.14
+ PA (w/ $R_a + R_f$)	60.67	57.99	54.57	50.86
+ PA (w/ $R_a + R_f + R_p$)	58.00	60.78	56.57	49.43
+ PA (w/ $R_a + R_f + R_s$)	60.00	58.44	54.57	49.43

improve accuracy and can even degrade performance, suggesting that humor-specific priors are most effective when paired with a structured reasoning process. When combined with RM, however, it provides additional gains, particularly on the more challenging ranking tasks (Zhou et al., 2025). In contrast, IM+PA underperforms Base+PA, indicating that domain adaptation without resolution supervision can introduce instability. Preference Alignment (PA) provides complementary benefits, especially on ranking tasks that depend on perceptual grounding and stylistic quality. The full IRS pipeline (IM+RM+PA) achieves the strongest overall performance, indicating that each component contributes a distinct and necessary aspect of humor reasoning.

5.3 Ablation on Reward Functions

Table 3 isolates the contribution of individual reward signals within the Preference Alignment (PA) component. Accuracy and format rewards alone ($R_a + R_f$) provide a strong baseline, while adding the visual perception (R_p) and style (R_s) rewards improves performance, particularly on harder ranking tasks (Zhou et al., 2025), where visual grounding and stylistic quality are more critical. These improvements align with the IRS decomposition: R_a supervises caption selection, R_p reinforces visual grounding, and R_s encourages captionist-consistent expression. The pattern of improvements–with humor-specific rewards contributing most on the harder, more ambiguous ranking tasks–confirms that what distinguishes PA from generic RL alignment is precisely its humor-specific reward structure, designed to enforce the full IRS reasoning chain rather than optimizing for correctness alone.

6 Conclusion

We introduced Incongruity-Resolution Supervision (IRS), a framework that models humor understanding as a structured reasoning process grounded in incongruity-resolution theory and expert captionist practice. Rather than supervising caption selection alone, IRS supervises the intermediate steps that make a caption funny–identifying visual mismatches, constructing coherent reinterpretations, and evaluating candidate interpretations against human judgments–through domain-adaptive pretraining, captionist reasoning traces, and

preference alignment with perceptual and stylistic rewards. Across 7B, 32B, and 72B models, IRS consistently outperforms strong baselines on matching and ranking tasks, with the largest model approaching expert-level performance and generalizing to out-of-domain benchmarks. These results suggest that explicitly supervising reasoning structure, rather than relying on scale alone, is key for complex, subjective tasks such as humor.

Ethics Statement

This work aims to advance the field of machine learning by improving multimodal reasoning in creative and subjective domains, using visual humor as a testbed. By modeling expert-inspired reasoning processes rather than relying solely on scale or pattern matching, our approach contributes to more interpretable and transparent multimodal language models.

The methods and findings presented here are not intended for high-stakes decision-making or automated content moderation, but rather for understanding how models can reason about nuanced, culturally grounded phenomena such as humor. Potential applications include creative assistance tools, educational systems, and research on human–AI interaction, where explainability and stylistic awareness are desirable.

At the same time, we recognize that humor is subjective and culturally dependent. Models trained on specific humor traditions, such as the New Yorker Cartoon Caption Contest, may reflect the stylistic norms and biases of that context and may not generalize uniformly across cultures or communities. We discuss these limitations, along with broader ethical considerations related to data sources, subjectivity, and evaluation practices, in Appendix E.

Overall, we believe this work presents a low-risk contribution whose primary impact is to advance methodological understanding and reproducibility in multimodal reasoning research.

Acknowledgement

This work was partly supported by the KUIS AI Center Research Awards to Doga Kukul. The authors gratefully acknowledge that the numerical computations reported in this work were performed in part using the TRUBA resources of the TUBITAK ULAKBIM High Performance and Grid Computing Center.

References

AllenAI. Olmo-mix-1124 dataset. https://huggingface.co/datasets/allenai/olmo-mix-1124, 2024.

Salvatore Attardo. *Linguistic Theories of Humor*. Approaches to Semiotics. Mouton de Gruyter, 1994. ISBN 9783110142556.

Salvatore Attardo and Victor Raskin. Script theory revis(it)ed: Joke similarity and joke representation model. *Humor: International Journal of Humor Research*, 4(3-4):293–347, 1991. ISSN 0933-1719.

Seana Coulson and Marta Kutas. Getting it: Human event-related brain response to jokes in good and poor comprehenders. *Neuroscience Letters*, 316(2):71–74, 2001. ISSN 0304-3940.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin

Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Samuel Doogan, Aniruddha Ghosh, Hanyang Chen, and Tony Veale. Idiom savant at Semeval-2017 task 7: Detection and interpretation of English puns. In Steven Bethard, Marine Carpuat, Marianna Apidianaki, Saif M. Mohammad, Daniel Cer, and David Jurgens (eds.), *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 103–108, Vancouver, Canada, August 2017. Association for Computational Linguistics.

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.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, Baoxu Wang, Bin Chen, Boyan Shi, Changyu Pang, Chenhui Zhang, Da Yin, Fan Yang, Guoqing Chen, Jiazheng Xu, Jiale Zhu, Jiali Chen, Jing Chen, Jinhao Chen, Jinghao Lin, Jinjiang Wang, Junjie Chen, Leqi Lei, Letian Gong, Leyi Pan, Mingdao Liu, Mingde Xu, Mingzhi Zhang, Qinkai Zheng, Sheng Yang, Shi Zhong, Shiyu Huang, Shuyuan Zhao, Siyan Xue, Shangqin Tu, Shengbiao Meng, Tianshu Zhang, Tianwei Luo, Tianxiang Hao, Tianyu Tong, Wenkai Li, Wei Jia, Xiao Liu, Xiaohan Zhang, Xin Lyu, Xinyue Fan, Xuancheng Huang, Yanling Wang, Yadong Xue, Yanfeng Wang, Yanzi Wang, Yifan An, Yifan Du, Yiming Shi, Yiheng Huang, Yilin Niu, Yuan Wang, Yuanchang Yue, Yuchen Li, Yutao Zhang, Yuting Wang, Yu Wang, Yuxuan Zhang, Zhao Xue, Zhenyu Hou, Zhengxiao Du, Zihan Wang, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Minlie Huang, Yuxiao Dong, and Jie Tang. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025. URL https://arxiv.org/abs/2507.01006.

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.

Nabil Hossain, John Krumm, and Michael Gamon. "president vows to cut <taxes> hair": Dataset and analysis of creative text editing for humorous headlines. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 133–142, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

Nabil Hossain, John Krumm, Tanvir Sajed, and Henry Kautz. Stimulating creativity with FunLines: A case study of humor generation in headlines. In Asli Celikyilmaz and Tsung-Hsien Wen (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 256–262, Online, July 2020. Association for Computational Linguistics.

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 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. 47166–47188. Curran Associates, Inc., 2024.

Veedant Jain, Gabriel Kreiman, and Felipe dos Santos Alves Feitosa. Humordb: Can ai understand graphical humor?, 2025.

Mehran Kazemi, Bahare Fatemi, Hritik Bansal, John Palowitch, Chrysovalantis Anastasiou, Sanket Vaibhav Mehta, Lalit K Jain, Virginia Aglietti, Disha Jindal, Peter Chen, Nishanth Dikkala, Gladys Tyen, Xin Liu, Uri Shalit, Silvia Chiappa, Kate Olszewska, Yi Tay, Vinh Q. Tran, Quoc V Le, and Orhan Firat. BIG-bench extra hard. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26473–26501, Vienna, Austria, July 2025. Association for Computational Linguistics.

Yuan-Hong Liao, Sven Elflein, Liu He, Laura Leal-Taixé, Yejin Choi, Sanja Fidler, and David Acuna. Longperceptualthoughts: Distilling system-2 reasoning for system-1 perception. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=SrKdi4MsUW.

Qianchu Liu, Sheng Zhang, Guanghui Qin, Timothy Ossowski, Yu Gu, Ying Jin, Sid Kiblawi, Sam Preston, Mu Wei, Paul Vozila, Tristan Naumann, and Hoifung Poon. X-Reasoner: towards generalizable reasoning across modalities and domains, 2025a. URL https://arxiv.org/abs/2505.03981.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2025b.

Bob Mankoff. *The Naked Cartoonist: A New Way to Enhance Your Creativity*. Black Dog & Leventhal, New York, NY, 2002.

Rada Mihalcea and Carlo Strapparava. Learning to laugh (automatically): Computational models for humor recognition. *Computational Intelligence*, 22(2):126–142, 2006.

OpenAI. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, October 2024. URL https://arxiv.org/abs/2410.21276. Revision v1.

OpenAI. Openai partners with condé nast. https://openai.com/index/conde-nast/, 2024. Accessed: 2025-02-21.

OpenAI. Openai o3 and o4-mini system card. System card, OpenAI, April 2025. URL https://openai.com/index/o3-o4-mini-system-card/.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The fineweb datasets: Decanting the web for the finest text data at scale. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=n6SCkn2QaG.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025.

Graeme Ritchie. Current directions in computational humour. *Artificial Intelligence Review*, 16:119–135, 2001.

Andrea C. Samson. The influence of empathizing and systemizing on humor processing: Theory of mind and humor. *Humor*, 25(1):75–98, 2012. ISSN 0933-1719.

Wenbo Shang, Yuxi Sun, Jing Ma, and Xin Huang. On the wings of imagination: Conflicting script-based multi-role framework for humor caption generation. In *The Fourteenth International Conference on Learning Representations*, 2026.

Jerry M. Suls. A two-stage model for the appreciation of jokes and cartoons: An information-processing analysis. In Jeffrey Goldstein and Paul E. McGhee (eds.), *The psychology of humor: Theoretical perspectives and empirical issues*, pp. 81–100. Academic Press, 1972.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, Congcong Wang, Dehao Zhang, Dikang Du, Dongliang Wang, Enming Yuan, Enzhe Lu, Fang Li, Flood Sung, Guangda Wei, Guokun Lai, Han Zhu, Hao Ding, Hao Hu, Hao Yang, Hao Zhang, Haoning Wu, Haotian Yao, Haoyu Lu, Heng Wang, Hongcheng Gao, Huabin Zheng, Jiaming Li, Jianlin Su, Jianzhou Wang, Jiaqi Deng, Jiezhong Qiu, Jin Xie, Jinhong Wang, Jingyuan Liu, Junjie Yan, Kun Ouyang, Liang Chen, Lin Sui, Longhui Yu, Mengfan Dong, Mengnan Dong, Nuo Xu, Pengyu Cheng, Qizheng Gu, Runjie Zhou, Shaowei Liu, Sihan Cao, Tao Yu, Tianhui Song, Tongtong Bai, Wei Song, Weiran He, Weixiao Huang, Weixin Xu, Xiaokun Yuan, Xingcheng Yao, Xingzhe Wu, Xinhao Li, Xinxing Zu, Xinyu Zhou, Xinyuan Wang, Y. Charles, Yan Zhong, Yang Li, Yangyang Hu, Yanru Chen, Yejie Wang, Yibo Liu, Yibo Miao, Yidao Qin, Yimin Chen, Yiping Bao, Yiqin Wang, Yongsheng Kang, Yuanxin Liu, Yuhao Dong, Yulun Du, Yuxin Wu, Yuzhi Wang, Yuzi Yan, Zaida Zhou, Zhaowei Li, Zhejun Jiang, Zheng Zhang, Zhilin Yang, Zhiqi Huang, Zihao Huang, Zijia Zhao, Ziwei Chen, and Zongyu Lin. Kimi-vl technical report, 2025. URL https://arxiv.org/abs/2504.07491.

Omkar Thawakar, Dinura Dissanayake, Ketan Pravin More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Ilmuz Zaman Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, Hisham Cholakkal, Ivan Laptev, Mubarak Shah, Fahad Shahbaz Khan, and Salman Khan. LlamaV-o1: Rethinking step-by-step visual reasoning in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 24290–24315, Vienna, Austria, July 2025. Association for Computational Linguistics.

Sean Trott, Drew E. Walker, Samuel M. Taylor, and Seana Coulson. Turing jest: Distributional semantics and one-line jokes. *Cognitive Science*, 49(5):e70066, 2025.

Robert West and Eric Horvitz. Reverse-engineering satire, or "paper on computational humor accepted despite making serious advances". *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7265–7272, Jul. 2019.

L. Wood. *Your Caption Has Been Selected: More Than Anyone Could Possibly Want to Know About The New Yorker Cartoon Caption Contest*. St. Martin's Publishing Group, 2024. ISBN 9781250333414.

Tong Xiao, Xin Xu, Zhenya Huang, Hongyu Gao, Quan Liu, Qi Liu, and Enhong Chen. Advancing multimodal reasoning capabilities of multimodal large language models via visual perception reward, 2025. URL https://arxiv.org/abs/2506.07218.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. LLaVA-CoT: let vision language models reason step-by-step. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2025.

Yixin Yang, Zheng Li, Qingxiu Dong, Heming Xia, and Zhifang Sui. Can large multimodal models uncover deep semantics behind images? In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 1898–1912, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.113. URL https://aclanthology.org/2024.findings-acl.113/.

Yufei Zhan, Yousong Zhu, Shurong Zheng, Hongyin Zhao, Fan Yang, Ming Tang, and Jinqiao Wang. Vision-r1: Evolving human-free alignment in large vision-language models via vision-guided reinforcement learning, 2025. URL https://arxiv.org/abs/2503.18013.

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., 2024.

Yaowei Zheng, Junting Lu, Shenzhi Wang, Zhangchi Feng, Dongdong Kuang, and Yuwen Xiong. Easyr1: An efficient, scalable, multi-modality rl training framework. https://github.com/hiyouga/EasyR1, 2025.

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.

Kuan Lok Zhou, Jiayi Chen, Siddharth Suresh, Reuben Narad, Timothy T. Rogers, Lalit K Jain, Robert D Nowak, Bob Mankoff, and Jifan Zhang. Bridging the creativity understanding gap: Small-scale human alignment enables expert-level humor ranking in llms, 2025.

Appendix

This appendix provides comprehensive supplementary material to support transparency, reproducibility, and a deeper understanding of our approach. It expands on key components of the main paper, including dataset construction, training procedures, prompt design, and qualitative analysis, as well as additional experiments that further validate our findings. Where appropriate, these materials offer a more detailed view of the individual components underlying *Incongruity–Resolution Supervision (IRS)*.

The content is organized as follows:

- **Section A - Incongruity Modeling Corpora:** A detailed description of the Incongruity Modeling (IM) dataset, including its composition, sources, temporal coverage, and licensing considerations. This section provides insight into how domain-relevant knowledge is incorporated to support incongruity modeling.
- **Section B - Training Details:** Complete documentation of model configurations, hyperparameters, optimization strategies, and implementation details across all training stages. This includes specifics for continual pretraining, supervised learning with reasoning traces, and reinforcement learning with humor-aware rewards.
- **Section C - Prompt Templates:** Full prompt specifications used throughout the pipeline, including those for generating captionist reasoning traces, evaluating outputs via reward models, and performing inference during testing. These templates are critical for reproducing the structured reasoning behavior induced by IRS.
- **Section D - Qualitative Examples:** Additional examples illustrating model behavior at different stages of training. These include generated reasoning traces, outputs refined through reinforcement learning, curated visual references, judge responses, and comparisons with expert captionist reasoning. Together, these examples provide a qualitative view of how structured reasoning evolves.
- **Section E - Limitations and Ethical Considerations:** An extended discussion of the limitations of our approach, including failure modes related to perception and reasoning, the inherent subjectivity of humor, cultural variability, and broader ethical considerations associated with modeling creative human judgments.
- **Section F - Cross-Dataset Generalization:** Additional experiments evaluating our model on external humor-related datasets with different formats and annotation schemes. These results demonstrate that the reasoning patterns learned from NYCC generalize beyond the original domain.
- **Section G - Zero-shot Caption Generation:** Qualitative results demonstrating that IRS-trained models can generate plausible, captionist-style captions despite not being explicitly trained for this task. These examples highlight how structured reasoning learned through IRS transfers to open-ended caption generation, producing outputs grounded in visual incongruity and consistent with editorial humor conventions.

A Incongruity Modeling Corpora

Corpus Composition. Our corpus is designed to support the *incongruity modeling* component of **Incongruity-Resolution Supervision (IRS)** by exposing the model to the discourse, reasoning patterns, and stylistic conventions underlying cartoon humor. Table A.1 summarizes the main sources, which fall into three categories:

- **Contest deliberations & roundtables (audio/video)** capture conversational reasoning processes involved in caption generation and evaluation. Sources such as the *New Yorker Cartoon Caption Contest Podcast*, *Official New Yorker Cartoon Podcast*, *CartoonStock YouTube Panel Discussions*, and *The Cartoon Pad* provide rich examples of how incongruities are identified, debated, and refined in practice.
- **Editorial & captionist commentary (written)** provides reflective analyses explaining why specific captions succeed or fail. We include Lawrence Wood's commentaries on

CartoonStock.com, which offer explicit reasoning about humor effectiveness and stylistic choices.

- **Books: craft, process, and editorial perspective** distill expert knowledge into concise heuristics and principles. Works such as *The Naked Cartoonist* (Mankoff, 2002), *Your Caption Has Been Selected* (Wood, 2024), and *How About Never—Is Never Good for You?* (Mankoff, 2014) provide structured guidance on creativity, interpretation, and editorial decision-making.

Table A.1: **Composition of the IM corpus.** The corpus is dominated by conversational and editorial sources, reflecting the importance of reasoning processes and stylistic conventions in humor understanding.

Source	Instances	Words	Tokens	% of Tokens
Contest Deliberations and Commentaries				
New Yorker Caption Contest Podcast	173	2,292,458	2,921,356	46.43%
The Cartoon Pad	43	443,774	568,849	9.04%
Official New Yorker Cartoon Podcast	34	349,588	465,701	7.40%
CartoonStock YouTube Panel Discussions	36	224,864	281,295	4.47%
CartoonStock Lawrence Wood Commentaries	175	128,635	170,978	2.72%
Books on Caption Writing				
How About Never	1	40,621	56,010	0.89%
Your Caption Has Been Selected	94	8,087	11,057	0.18%
The Naked Cartoonist	1	3,062	4,076	0.06%
General-Purpose Corpora				
FineWeb	1759	979,418	1,308,241	20.79%
Olmo-Mix-1124	1807	353,705	504,177	8.01%
Total	4123	4,824,212	6,291,740	100.00%

To ensure stable optimization and prevent overfitting to narrow humor-specific distributions, we additionally incorporate subsets of FineWeb (Penedo et al., 2024) and OLMo-Mix-1124 (AllenAI, 2024). These data provide broad linguistic coverage and general world knowledge, which are essential for interpreting cultural references and contextual nuances frequently used in cartoon captions. All items overlapping evaluation cartoons or captions are removed to avoid leakage.

Temporal Coverage. As summarized in Table A.2, our podcast and commentary sources span nearly a decade of caption contest discourse. The *Official New Yorker Cartoon Podcast* provides the earliest coverage (2015–2021), while more recent series such as the NYCC Podcast (2021–2024) and *The Cartoon Pad* (2021–2024) extend into the present. More recent commentary comes from CartoonStock YouTube Panels through July 2025, complemented by Lawrence Wood's weekly contest analyses (2019–2024).

Table A.2: **Temporal coverage of humor-specific sources.** The corpus spans nearly a decade of caption contest discourse, capturing both stable stylistic conventions and evolving humor preferences.

Source	First episode	Last episode	Coverage
NYCC Podcast	Mar 18, 2021 (Ep. 1)	Sep 25, 2024 (Ep. 173)	3.5 years
The Cartoon Pad	Apr 2, 2021 (Ep. 1)	Jul 24, 2024 (Ep. 44)	3 years
Official NYer Cartoon Podcast	Nov 14, 2015 (Ep. 29)	Jan 22, 2021 (Ep. 289)	5+ years
Cartoon Stock YouTube Podcasts	Dec 20, 2022	Jul 25, 2025	2.5 years
Lawrence Wood Commentaries	Jul 12, 2019	Oct 22, 2024	5 years

CartoonStock is an independent cartoon archive and marketplace that hosts a caption contest similar to NYCC. Its panel discussions and Wood's critiques are particularly valuable, as they explicitly analyze finalist and winning captions, highlighting both successful and

unsuccessful interpretations. This complements NYCC data by exposing the model to diverse reasoning styles and evaluative criteria across platforms.

Licensing note. Parts of the IM corpus are derived from copyrighted sources (e.g., books and podcast transcripts). Due to licensing restrictions, these materials cannot be directly released. To ensure reproducibility, we will release all train/test splits used for evaluation, along with preprocessing code and prompt templates required to reconstruct the pipeline. This balances transparency with respect for intellectual property constraints.

B Training Details

We describe the training procedures used to implement *Incongruity-Resolution Supervision (IRS)*. These consist of three stages: Incongruity Modeling (IM), Resolution Modeling (RM), and Preference Alignment (PA). Table B.1 summarizes hyperparameters.

Table B.1: **Training hyperparameters across IRS stages.** Each stage supports a different component of IRS, with distinct optimization regimes and resource requirements.

Stage	Base Model	Adapter	Epochs	Batch	LR	Prec.	GPUs	Dur.
IM	Qwen2.5-VL-7B	LoRA (rank 32)	50	1	1e-4	bf16	2× H100	∼4 h
RM	IM model	LoRA (rank 64)	7	16	1e-4	bf16	2× H100	∼3 h
PA	RM model	Full model	5	16 (roll.)	1e-6	bf16	4× H100	∼1.5 d

Incongruity Modeling (IM). The goal of IM is to adapt the backbone model to humor-relevant discourse. We train Qwen2.5-VL-7B-Instruct using LoRA adapters (rank 32, scaling factor 64, dropout 0.05), and Qwen2.5-VL-32B-Instruct using LoRA with 4-bit quantization (rank 8, scaling factor 16), on the curated corpus described in Sec. A. Training the 7B model runs for 50 epochs with batch size 1 and learning rate 1×10^{-4} on 2 NVIDIA H100 GPU (∼4 hours). The 32B model is trained for 25 epochs with batch size 1 and learning rate 1×10^{-4} on 4 NVIDIA H100 GPUs (∼6 hours). Similarly, the 72B model is trained for 10 epochs with batch size 1 and learning rate 1×10^{-4} on 4 NVIDIA H200 GPUs (∼6 hours). Since the IM corpus is predominantly textual, we freeze the visual encoder and multimodal projection layers, updating only the language component. This prevents degradation of visual understanding while adapting the model to humor-specific reasoning patterns. Fig. B.1-B.3 show loss, learning rate, and gradient norm dynamics of our IM models.

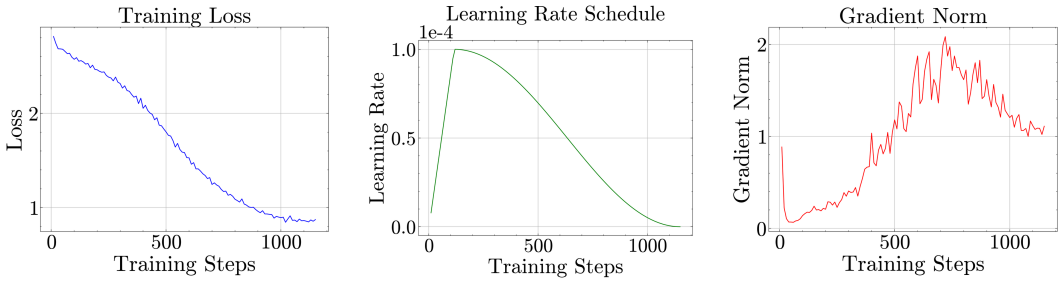


Figure B.1: **IM dynamics of the 7B model.** Loss decreases steadily over training, indicating successful adaptation to humor-specific discourse. The gradient norm remains controlled overall, with larger spikes appearing later in training.

Resolution Modeling (RM). RM teaches the model to construct structured interpretations using captionist reasoning traces (Sec. 3.2). We initialize from the IM-adapted model and

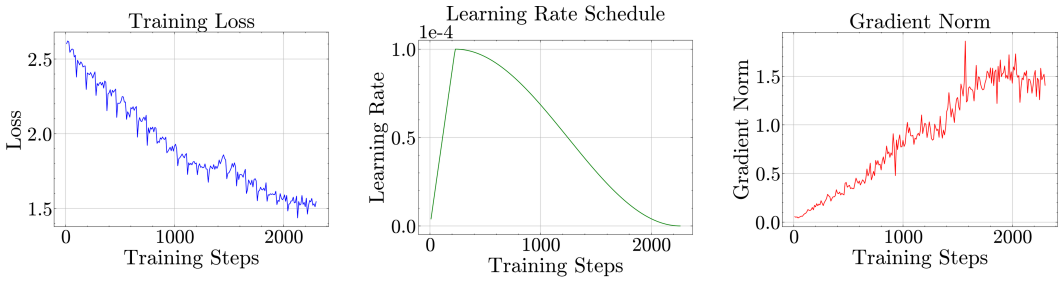


Figure B.2: **IM dynamics of the 32B model.** Loss decreases steadily despite mild early fluctuations, indicating successful adaptation to humor-specific discourse. The gradient norm increases gradually over training while remaining stable overall.

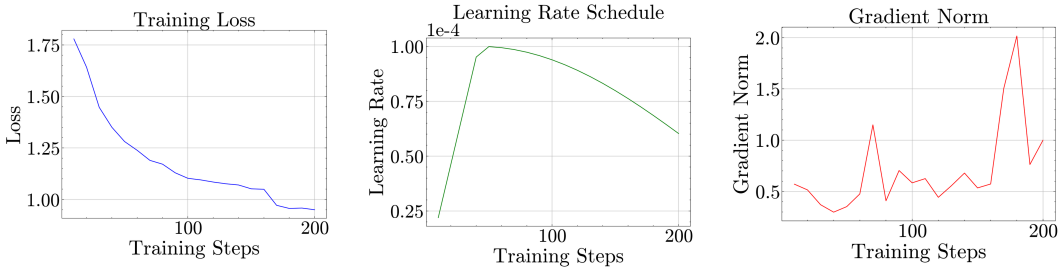


Figure B.3: **IM dynamics of the 72B model.** Loss drops rapidly and then stabilizes, indicating effective early adaptation to humor-specific discourse. The gradient norm remains mostly low but exhibits occasional sharp spikes.

train using `LLaMA-Factory` with LoRA (rank 64 for 7B, rank 8 for 32B), keeping the vision encoder frozen. Training uses distilled reasoning traces and runs for 7 epochs with effective batch size 16, cosine learning-rate schedule with 10% warmup, and peak rate 1×10^{-4}. The 7B model trains in $\sim$3 hours on 2×H100 GPUs, while the 32B model requires $\sim$6 hours on 4×H100 GPUs. Fig. B.4-B.6 show training dynamics. Loss decreases smoothly after warmup, with stable gradients, indicating effective learning of structured reasoning patterns.

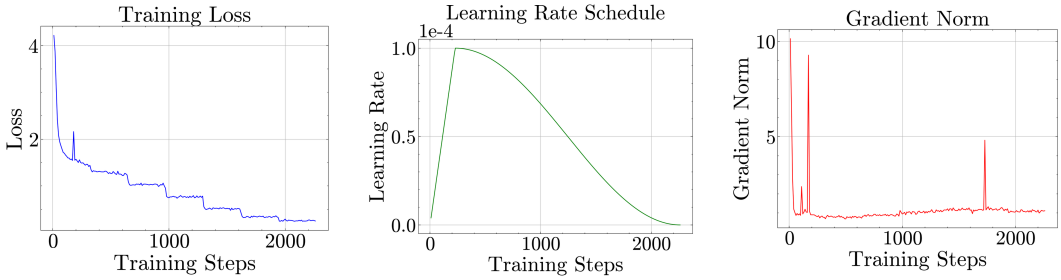


Figure B.4: **RM dynamics of the 7B model.** Rapid loss reduction during warmup followed by stable convergence indicates effective learning of structured reasoning from captionist traces.

Preference Alignment (PA). PA optimizes model outputs with humor-aware rewards (Sec. 3.3). We use EasyR1 (Zheng et al., 2025) with GRPO optimization (DeepSeek-AI et al., 2025), sampling 3 rollouts per prompt with maximum length 512 and temperature 1.0. Training uses rollout batch size 16 and learning rate 1×10^{-6} on 4×H100 GPUs (8 for the 32B model, and 16 for 72B model). Fig. B.7-B.9 show reward dynamics. Perception and style rewards saturate early, while accuracy and format rewards continue improving.

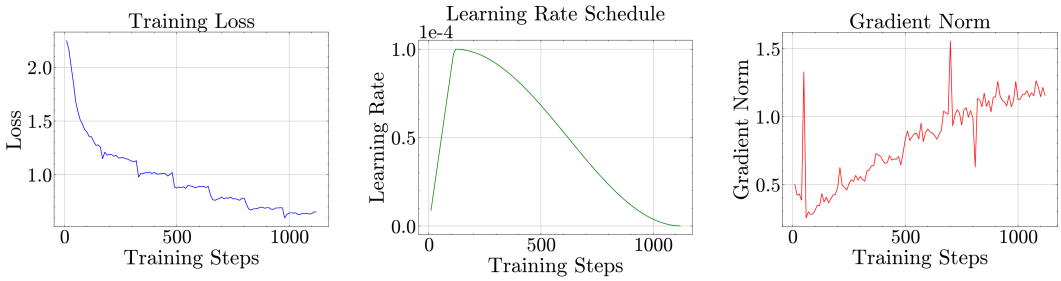


Figure B.5: **RM dynamics of the 32B model.** Loss decreases smoothly after warmup, indicating effective adaptation to structured reasoning supervision. The gradual increase in gradient norm suggests continued refinement of reasoning patterns during training.

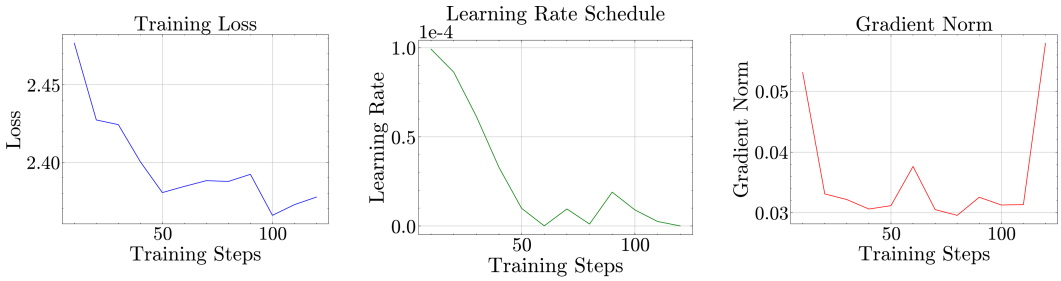


Figure B.6: **RM dynamics of the 72B model.** Loss decreases more gradually with limited training steps, indicating early-stage adaptation to reasoning supervision. The gradient norm remains low overall with occasional spikes, suggesting stable but less extensive optimization.

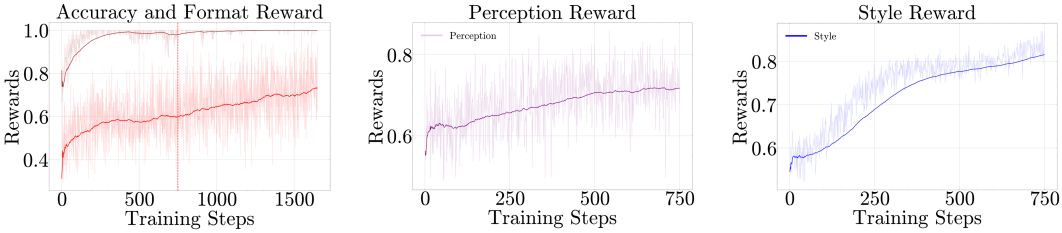


Figure B.7: **PA dynamics of the 7B model.** Perception and style rewards increase rapidly and saturate early, while accuracy and format rewards continue improving, indicating progressive alignment with human judgment and reasoning quality.

Figure B.8: **PA dynamics of the 32B model.** Perception and style rewards increase steadily over training, while accuracy and format rewards improve more gradually, indicating stable and sustained alignment with human judgment and reasoning quality.

C Prompts

This section provides the exact prompts used to implement different components of *Incongruity-Resolution Supervision (IRS)*. For reproducibility, we group them into three categories: (i) reasoning-trace generation, which supports structured interpretation; (ii)

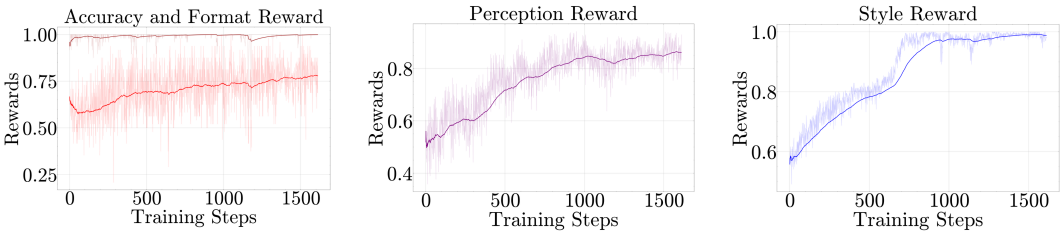


Figure B.9: **PA dynamics of the 72B model.** Perception and style rewards increase rapidly and approach saturation, while accuracy and format rewards continue improving, indicating strong alignment with human judgment and well-calibrated reasoning quality at scale.

reward-judge prompts, which guide preference alignment; and (iii) evaluation prompts for both text-only and multimodal models.

C.1 Reasoning Trace Generation.

We generate captionist reasoning traces from the structured cartoon annotations of Hessel et al. (2023), which include scene descriptions, uncanny elements, and entity references that provide grounding for interpretation.[2] We use DeepSeek-R1 (DeepSeek-AI et al., 2025) as a teacher model, prompting it to reconstruct the scene, identify salient incongruities, infer speaker intent, and analyze humor through mechanisms such as wordplay and cultural references before selecting the correct caption. This process produces structured reasoning traces that explicitly model how captions are interpreted and evaluated.

The resulting traces are rephrased using GPT-4o (OpenAI, 2024) to match the style of professional captionist commentary, emphasizing concise, observational, and image-grounded language. This transformation replaces annotation-based phrasing (e.g., *"the description says"*) with direct visual interpretation (e.g., *"when I look at the cartoon"*), while preserving the underlying reasoning structure. This step ensures stylistic consistency between generated traces and real captionist discourse.

Fig. C.1 and C.2 shows the generation templates for the matching and ranking tasks, respectively, which structures the reasoning process around scene description, incongruity detection, and caption justification. Fig. C.3 shows the rephrasing template, used to refine traces into captionist-style commentary.

C.2 Reward-Judge Prompts

We design two LLM-as-judge templates that produce automatic, fine-grained reward signals during reinforcement learning. These prompts implement *preference alignment* within IRS by evaluating whether model outputs are both visually grounded and stylistically consistent with captionist practice. Each judge returns binary scores for multiple criteria, which are aggregated into the reward function (Sec. 3.3). Together, these rewards capture two core dimensions of humor understanding: (i) perceptual grounding, encouraging reasoning to reference concrete visual incongruities in the cartoon, and (ii) stylistic fidelity, encouraging captions and explanations to reflect the linguistic qualities of professional humor writing.

Visual Perception Judge. This prompt evaluates whether model reasoning is grounded in salient visual details. Each cartoon is associated with up to ten curated reference descriptions (entities, background elements, and incongruities). Qwen2.5-7B-Instruct receives both the model's reasoning and these references, and outputs a binary vector indicating whether each reference is explicitly reflected in the reasoning. This design directly links the perception reward (R_p) to the visual anchors curated in our dataset (Fig. 3), encouraging explanations that are faithful to the observed scene rather than relying on abstract or unsupported interpretations. Fig. C.4 shows the exact prompt.

[2]For cartoons missing these descriptions, we use the OpenAI o3 model OpenAI (2025) to generate them automatically from the image.

REASONING TRACE GENERATION PROMPT (MATCHING TASK)

You are a cartoon analyst evaluating humor in New Yorker Caption Contest. Below is a detailed description of the cartoon, including its visual scene, unusual/uncanny elements, and key observations. Your task is to act as if you're looking directly at the cartoon—not just reading about it. Reason step by step: Think step by step:
1- Understand the visual setting and what makes it strange or surprising.
2- Identify who is most likely speaking in the cartoon.
3- Reconstruct the story or situation behind the scene—what might be going on between the characters?
4- Analyze the humor in each caption: look for metaphors, cultural references, and wordplay.

Finally, from given five captions, one of which matches the cartoon image,
and the others are unrelated, decide which caption that best matches the cartoon image, and justify your choice as if you were analyzing the cartoon visually.

Scene: [SCENE]

Description: [DESCRIPTION]

Uncanny Element: [UNCANNY ELEMENT]

Observations: [OBSERVATIONS]

A) [CAPTION A]
B) [CAPTION B]
C) [CAPTION C]
D) [CAPTION D]
E) [CAPTION E]

Question: Based on what you "see" in the cartoon, which caption best matches the cartoon image? Justify your choice with detailed reasoning based on visual analysis, speaker context, and linguistic play.

{Your answer should support Caption [ANSWER] as the matching caption with strong reasoning.}

Figure C.1: **Prompt for reasoning trace generation (matching).** Enforces structured, step-by-step reasoning that reconstructs the scene, identifies incongruities, infers speaker context, and justifies the correct caption through visual grounding and narrative interpretation. The correct answer is optionally provided.

REASONING TRACE GENERATION PROMPT (RANKING TASK)

You are a cartoon analyst evaluating humor in New Yorker Caption Contest. Below is a detailed description of the cartoon, including its visual scene, unusual/uncanny elements, and key observations. Your task is to act as if you're looking directly at the cartoon—not just reading about it. Reason step by step: Think step by step:
1- Understand the visual setting and what makes it strange or surprising.
2- Identify who is most likely speaking in the cartoon.
3- Reconstruct the story or situation behind the scene—what might be going on between the characters?
4- Analyze the humor in each caption: look for metaphors, cultural references, and wordplay.

Finally, decide which caption is funnier, and justify your choice as if you were analyzing the cartoon visually.

Scene: [SCENE]

Description: [DESCRIPTION]

Uncanny Element: [UNCANNY ELEMENT]

Observations: [OBSERVATIONS]

A) [CAPTION A]
B) [CAPTION B]

Question: Based on what you "see" in the cartoon, which caption is funnier? Justify your choice with detailed reasoning based on visual analysis, speaker context, and linguistic play.

{Your answer should support Caption [ANSWER] as the funnier caption with strong reasoning.}

Figure C.2: **Prompt for reasoning trace generation (ranking).** Enforces structured, step-by-step reasoning that reconstructs the scene, identifies incongruities, infers speaker context, and compares competing captions through visual grounding and narrative interpretation. The correct answer is optionally provided.

REASONING TRACE REPHRASING PROMPT (MATCHING & RANKING TASKS)

You are an assistant tasked with lightly rewriting a cartoon analysis. Below is a response that was written based on a textual description of a cartoon. Your ONLY task is to rephrase any language that refers to "the description", "I imagine", or "visualizing" — and instead replace such phrases with observational ones like "When I look at the cartoon" or "In the image I see".

- Do not change anything else.
- Keep all reasoning, sentences, structure, and wording identical — only modify phrases that imply the analysis was based on a description.

--- ORIGINAL RESPONSE ---
[ORIGINAL RESPONSE]

--- REWRITTEN RESPONSE ---

Figure C.3: **Prompt for reasoning-trace rephrasing.** Rewrites reasoning traces into captionist-style commentary by replacing description-based phrasing with image-grounded language, while preserving the original reasoning structure and content.

VISUAL PERCEPTION REWARD PROMPT

You are an evaluator. Read VISUAL INFO (XML) and the CANDIDATE answer. For each listed tag, output ONLY <tag>0</tag> concatenated in the same order, with no spaces or extra text.
VISUAL INFO: [VISUAL INFO]

CANDIDATE: [ANSWER]

Evaluation rule:
- Output <tag>1</tag> if the candidate explicitly and correctly reflects the fact in VISUAL INFO for that tag.
- Output <tag>0</tag> otherwise (missing, wrong, or contradicted).
Return ONLY these fields concatenated in EXACTLY this order (no spaces/newlines)

Figure C.4: **Prompt for visual perception judge.** Evaluates whether reasoning trace explicitly references curated visual details; outputs binary tags for each attribute.

Style Judge. This prompt evaluates the linguistic quality of captions and explanations. Drawing on captionist guidelines (Wood, 2024), the template checks five stylistic dimensions:

1. **Natural phrasing** (use of idiomatic, everyday expressions),
2. **Punctuation** (effective, neither missing nor overused),
3. **Wordplay** (puns, double meanings, playful twists),
4. **Metaphor** (figurative expressions grounded in the cartoon),
5. **Punchline placement** (delivering the payoff at the end).

The judge outputs a binary vector corresponding to these criteria without additional commentary. The aggregated score forms the style reward (R_s), encouraging models not only to select the correct caption but to justify it in a way that reflects the tone and structure of professional humor writing. Fig. C.5 shows the full template.

Collectively, these LLM-as-judge rewards complement correctness- and format-based signals, enabling reinforcement learning to capture both perceptual grounding and stylistic quality. This enriches training with humor-aware evaluation criteria that go beyond task accuracy alone.

C.3 Text-only Evaluation

We evaluate DeepSeek-R1 (DeepSeek-AI et al., 2025) as a text-only baseline using structured cartoon annotations from Hessel et al. (2023). These annotations provide scene descriptions, uncanny elements, and key observations, enabling reasoning without access to raw visual inputs. This setting isolates reasoning ability under near-perfect perception, since the model is given curated descriptions that abstract away visual ambiguity. We use the original

STYLE REWARD PROMPT

You are an evaluator. Read the CANDIDATE answer and evaluate the caption that is selected as the answer. For each listed tag and each style criterion, output ONLY <tag>0 or 1</tag> concatenated in the same order, with no spaces or extra text.

CANDIDATE: [CANDIDATE]

Evaluation rules:
Determine the caption that is selected for the cartoon. Then, for each of the following style criteria, output <tag>1</tag> if the caption meets the criterion, and <tag>0</tag> otherwise:

<daily_phrase>: 1 if the candidate uses a natural, common, everyday expression or idiom that relates to the cartoon.

<punctuation>: 1 if punctuation is used effectively (not missing, not overused, contributes to readability or style) and enhances the connection to the cartoon.

<wordplay>: 1 if the candidate shows creative wordplay (puns, double meanings, playful twists) that are relevant to the cartoon.

<metaphor>: 1 if the candidate includes a clear metaphorical expression that relates directly to the cartoon.

<punchline>: 1 if the candidate includes a punchline at the end of the caption.

Otherwise output 0.

Return ONLY the tags concatenated in EXACTLY this order with no spaces or newlines:

Figure C.5: **Prompt for style judge.** Evaluates captions against stylistic criteria such as phrasing, punctuation, wordplay, metaphor, and punchline placement; outputs binary tags for each dimension.

prompts from Hessel et al. (2023) for both matching and ranking tasks. The matching prompt presents five candidate captions with one correct option, while the ranking prompt presents two captions with one preferred by the crowd. In both cases, the model is instructed to reason solely from the textual annotations. Fig. C.6 and Fig. C.7 show the full templates.

TEXT-ONLY EVALUATION PROMPT (MATCHING TASK)

In this task, you will see a description of an uncanny situation of a cartoon from the New Yorker Cartoon Caption Contest. Then, you will see five jokes — only one of which was written about the described situation. Pick which of the five choices truly corresponds to the described scene.

The image takes place in the following location:

Image Description: [IMAGE DESCRIPTION]

Image Uncanny Description: [UNCANNY DESCRIPTION]

The scene includes: [SCENE DESCRIPTION]

One of the following funny captions is most relevant to the scene:
A) [CAPTION A]
B) [CAPTION B]
C) [CAPTION C]
D) [CAPTION D]
E) [CAPTION E]

The funny caption that matches the scene is:

Figure C.6: **Prompt for text-only evaluation (matching).** Evaluates DeepSeek-R1 using textual annotations of the cartoon; model selects the caption that best matches the described scene.

TEXT-ONLY EVALUATION PROMPT (RANKING TASK)

In this task, you will see a description of an uncanny situation of a cartoon from the New Yorker Cartoon Caption Contest. Then, you will see two jokes that were written about the situation. One of the jokes is better than the other one. Pick which of the two jokes is the one rated as funnier by people.

The image takes place in the following location:

Image Description: [IMAGE DESCRIPTION]

Image Uncanny Description: [UNCANNY DESCRIPTION]

The scene includes: [SCENE DESCRIPTION]

A) [CAPTION A]
B) [CAPTION B}

The funnier is:

Figure C.7: **Prompt for text-only evaluation (ranking).** Evaluates DeepSeek-R1 using textual annotations; model selects the funnier caption between two options.

C.4 Multimodal Evaluation

We standardize evaluation of vision-language models using task-specific prompts that present the cartoon image (`<image>`) together with candidate captions. All models are required to produce outputs in the controlled format `<think>...</think><answer>...</answer>`, which enforces explicit reasoning traces and enables automatic parsing and evaluation. Fig. C.8 and Fig. C.9 show the full templates.

MULTIMODAL EVALUATION PROMPT (MATCHING TASK)

[CARTOON IMAGE]

The image is a cartoon from the New Yorker Cartoon Caption Contest. I will provide you with five captions, one of which matches the cartoon image, and the others are unrelated. Choose the caption that best matches the cartoon image:
A) [CAPTION A]
B) [CAPTION B]
C) [CAPTION C]
D) [CAPTION D]
E) [CAPTION E]

First think about the reasoning process in the mind and then provide the answer. The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively, i.e., <think> reasoning process here </think><answer> answer here </answer>

Figure C.8: **Prompt for multimodal evaluation (matching).** Cartoon image plus five candidate captions (A–E). Model outputs reasoning and final choice in `<think>` and `<answer>` tags.

To ensure fair comparison, the same prompting protocol is applied to all competing multimodal reasoning models (e.g., GLM-4V, Qwen2.5-VL, Kimi-VL). Models are explicitly instructed to *"think before answering"* and to follow the standardized output format. This helps ensure that performance differences reflect reasoning ability rather than prompt design.

D Additional Examples: Ground Truth Traces, Evolution of Model Reasoning, and Judge Responses

To complement the quantitative results in the main paper, we provide qualitative examples that illustrate how reasoning traces, model outputs, and reward signals interact to shape humor-aware reasoning under IRS.

MULTIMODAL EVALUATION PROMPT (RANKING TASK)

[CARTOON IMAGE]

The image is a cartoon from the New Yorker Cartoon Caption Contest. I will provide you with two captions; one of them is deemed funnier by people. Which of the following captions is funnier:
A) [CAPTION A]
B) [CAPTION B}

First think about the reasoning process in the mind and then provide the answer. The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively, i.e., <think> reasoning process here </think><answer> answer here </answer>

Figure C.9: **Prompt for multimodal evaluation (ranking).** Cartoon image plus two candidate captions (A–B). Model outputs reasoning and final choice in `<think>` and `<answer>` tags.

D.1 Ground-Truth Traces

In addition to the matching example in the main text (Fig. 2), we include a ground-truth reasoning trace for a ranking task. As illustrated in Fig. D.1, the trace demonstrates how captionists compare alternatives by identifying which caption better exploits the visual incongruity and delivers a more effective punchline. These traces form the structured supervision used during Resolution Modeling.

When I look at the cartoon, the setting is an airplane's economy cabin with regular passengers doing typical things—reading, sleeping, wearing headphones. The uncanny element is a giant amoeba-like germ creature sitting in one of the seats, even buckled in. It's massive, with eye-stalks and tendrils, spilling into adjacent seats and the aisle. The man next to it is gesturing toward it, probably talking to the woman by the window. People nearby look uncomfortable, especially the guy behind whose personal space is invaded by the amoeba's feelers.
The key observations: the germ is human-sized, which is absurd since germs are microscopic. It's treated as a regular passenger, which creates humor through juxtaposition. The man in the middle seems to be acknowledging the creature, maybe complaining or making a comment.
The humor likely stems from the contrast between the mundane airplane setting and this bizarre, intrusive presence. Next, identifying the speaker. The man in the middle seat is gesturing towards the amoeba while talking to the woman at the window. So he's probably the one delivering the caption. His body language suggests he's making a remark about the creature, possibly annoyed or trying to be humorous. Looking at the captions.
Caption A: "I guess that viral marketing works after all." This plays on the word "viral" which has a double meaning—both marketing that spreads rapidly and relating to viruses/germs. Since the amoeba represents a germ, the pun connects the creature to a viral marketing campaign. It's clever because it ties the absurd visual (a giant germ on a plane) to a common business term.
Caption B: "This guy's wife lets him drink on the plane!", it's implying the amoeba is drunk or its liquid form is due to alcohol. The humor comes from personifying the germ as a rowdy passenger whose wife permits him to drink, leading to his amorphous, spreading state. It's a play on the creature's blob-like appearance resembling someone who's had too much to drink.
Comparing the two. Caption A's strength is the wordplay with "viral" directly linking the germ to a punny scenario. It's witty and fits the visual of a germ being a "viral" entity in a marketing context. However, it might require a bit of cultural knowledge about marketing terms, which isn't universal. Caption B uses physical humor. The amoeba's slimy, spreading form looks like a drunk person spilling over seats. The idea of blaming the wife adds a relatable marital jab. It's more physical and visual, as the creature's appearance directly suggests intoxication.
Which is funnier? Caption B feels more immediate. The visual of the amoeba's gelatinous body spilling over seats perfectly matches a drunk passenger. The marital humor ("wife lets him") adds a layer of relatability. It doesn't rely on wordplay as much as a visceral, silly image. The man gesturing annoyed fits the caption—he's complaining about the creature's behavior, blaming its wife. It's accessible and ties directly to what we see: the creature's form and intrusion. Caption A is clever but more abstract. The "viral marketing" pun is smart, but it's a conceptual leap. The scene shows discomfort and intrusion, not marketing. So Caption B's humor is more integrated with the visuals, making it funnier in this context.

Figure D.1: **Example reasoning trace for a caption ranking task.** The trace compares two captions against the visual setup (a giant amoeba crowding an airplane seat) and explains why Caption B ("This guy's wife lets him drink on the plane!") better resolves the incongruity and captures the visual humor. This example illustrates how ranking traces encode comparative reasoning over humor mechanisms such as visual grounding and punchline strength.

D.2 Evolution of Model Reasoning.

To examine how IRS shapes reasoning, we compare outputs from RM-only and RM+PA models on matching and ranking tasks.

Fig. D.2 presents a caption matching example. The RM-only model produces fluent but shallow explanations, often identifying surface-level themes (e.g., a "therapy session") without fully connecting them to the central incongruity. Its reasoning is verbose and only partially grounded in the image. In contrast, the RM+PA model anchors its explanation in salient visual cues, filters distractors, and highlights multiple interacting humor mechanisms, including wordplay, speaker roles, and ironic contrasts. The output reads closer to a captionist's commentary: concise, well-structured, and attuned to comedic effect.

Fig. D.3 shows a ranking example, where the difference is more pronounced. The RM-only model focuses on literal associations (e.g., coffee grounds as evidence of recency), while missing the broader incongruity of cowboys interacting with modern espresso machines. The RM +PA model instead integrates visual context and stylistic considerations, selecting the caption that better captures the underlying joke without relying on overly specific cues.

These examples are consistent with the quantitative results: PA improves not only accuracy but also the structure and style of reasoning, shifting outputs toward more grounded and economical explanations.

D.3 Judge Responses

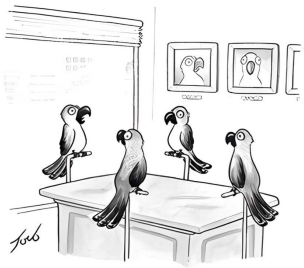
Fig. D.4 and D.5 illustrate the outputs of the LLM-as-judge components used to compute perception and style rewards. In Fig. D.4, the perception judge compares the model's reasoning against curated visual references and returns per-attribute binary scores. This ensures that explanations are explicitly grounded in salient scene elements. In Fig. D.5, the style judge resolves the model's `<answer>` to the selected caption and evaluates it along five dimensions: natural phrasing, punctuation, wordplay, metaphor, and punchline placement. These binary signals provide structured feedback on linguistic quality and humor delivery. Together, these reward components translate abstract notions of grounding and stylistic quality into measurable signals used during PA.

D.4 Comparison with Expert Traces

We include an example where the model's final choice differs from that of a professional captionist. As shown in Fig. D.6, the model nonetheless follows a similar reasoning process, identifying the central incongruity, evaluating alternative interpretations, and providing a structured justification. This alignment in reasoning structure, despite differing conclusions, suggests that the model has learned aspects of the captionist analysis process beyond surface-level pattern matching.

D.5 Discussion

The examples in Appendix D provide qualitative evidence of how IRS enables humor-aware reasoning. Ground-truth traces illustrate the structured supervision signals used during training, while model outputs show how reasoning evolves from surface-level associations to more grounded and stylistically consistent explanations during RM. The judge responses further reveal how PA translates qualitative criteria such as grounding and humor style into explicit reward signals. Together, these examples demonstrate that the improvements observed in quantitative evaluations are accompanied by systematic changes in reasoning behavior. Rather than treating humor as a black-box prediction task, IRS supports a decomposable and interpretable reasoning process, making it possible to analyze how models arrive at their decisions.



QUESTION

Which of the five choices is the matching caption?

CAPTION CHOICES

A) Unfortunately, your time is up, but I think we have a general outline of your problem.
B) There are cheaper ways to find an audience, Jim.
C) Can we talk?
D) It's our new interactive TV. When he's done, I'm having an appendectomy on 'General Hospital.'
E) I hope Skip doesn't ask me to sacrifice.

RM-ONLY MODEL RESPONSE

The cartoon shows four parrots in a doctor's office. Each has a microphone or perch, and they're looking at framed portraits on the wall. The uncanny element is that the parrots appear to be mimicking human behavior—taking turns looking at their own portraits, as if they're patients in a therapy session. When I look at the cartoon, the setting is a doctor's office with typical elements like an examination table and blinds. The portraits are of the parrots themselves, which is unusual. The humor comes from the absurdity of parrots acting like humans, especially in a therapy context. In the image, one parrot is likely speaking, possibly commenting on the situation. The others are engaged in looking at the portraits. The story seems to be a therapy session where the parrots are analyzing their own images, which is funny because they're birds, not humans.

Now, analyzing each caption:
A) Unfortunately, your time is up, but I think we have a general outline of your problem. This sounds like a therapist ending a session. The parrots are in a therapy setup, so this fits. The humor is in applying human therapy to parrots.
B) There are cheaper ways to find an audience, Jim. This might refer to publicizing something, but it doesn't connect to the therapy or portraits. The parrots aren't showing audience interest, so it's unrelated.
C) Can we talk? A common phrase, but here it could be a parrot speaking. However, it's too vague. The therapy context is implied, but the caption is too generic without tying into the portraits.
D) It's our new interactive TV. When he's done, I'm having an appendectomy on 'General Hospital.' This mentions TV and surgery, which doesn't relate to the therapy portraits. The parrots aren't watching TV; they're looking at their own portraits. So, it's off-topic.
E) I hope Skip doesn't ask me to sacrifice. This might imply a religious or ritual context, like sacrificing something. The parrots are birds, and "sacrifice" could be a pun, but it doesn't fit the therapy scene. The portraits aren't related to sacrifice.

Comparing all, caption A fits best. The therapist ending a session with parrots makes sense, and the humor comes from applying human therapy to animals. The portraits are part of the therapy, so the caption ties everything together. The other captions don't connect as well, either because they're off-topic or too vague.

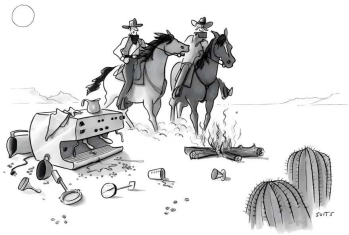
RM+PA MODEL RESPONSE

In the cartoon, I see a corporate office setting with four parrots sitting on stools around a conference table - three facing forward and one turned sideways like a facilitator. Above them are framed self-portraits of another parrot pair. The absurdity hits immediately: birds replacing humans in what should be a human professional environment, with those portraits suggesting this isn't just any office but one where birds literally document their colleagues' "faces" (pun intended).

Looking at the captions:
- Option A feels too human-centric (therapist/human patient scenario)
- Option B implies an audience/celebrity context (no visual cues like microphones/stages)
- Option C suggests personal conversation (only group observation shown)
- Option D introduces medical/TV chaos (absent visual elements)
- Option E hints at sacrificial themes (no predators or threats visible)

Option C stands out perfectly - that sideways parrot could easily be "facilitating" a meeting while others listen! The phrase "Can we talk?" visually mirrors their clustered posture and the meeting context. The humor lands through triple-layered irony: (1) Birds as corporate workers (absurd premise), (2) Parrot language mimicking human small talk (species-appropriate gag), (3) The speaker's position (facilitator/observer) determines who says it. Every visual detail serves the joke: portraits establish "office culture," perches show hierarchy, and the unspoken tension makes "Can we talk?" feel live. The winning caption weaponizes every visual gag point while letting the image breathe. 🦜👍

Figure D.2: **Example reasoning trace for a caption matching task.** The RM-only model produces fluent but superficial descriptions, often framing the parrots' scene as a "therapy session" without fully connecting this to the core incongruity of the humor. The RM+PA model, by contrast, grounds its reasoning in salient visual cues, filters irrelevant distractors, and highlights layered humor mechanisms such as species-appropriate wordplay, speaker positioning, and irony, producing commentary closer to professional captionist style.



QUESTION

Which caption is funnier than the other one?

CAPTION CHOICES

A) Grounds still warm...
B) Follow the plastic.

RM-ONLY MODEL RESPONSE

The scene is a desert with two cowboys on horseback. They've just come across a crashed coffee machine and scattered coffee items. The uncanny element is the juxtaposition of a modern espresso machine in the Old West.

The cowboys are likely commenting on the absurdity of finding such a device in the desert. The captions need to tie into the visual surprise.

Caption A: "Grounds still warm..." plays on the coffee grounds being nearby, implying the machine crashed recently. It's a literal take but might be too straightforward.
Caption B: "Follow the plastic" references the machine's plastic parts among cacti. It's more absurd, suggesting they're following a scent trail made of plastic. This adds a layer of silliness with the cowboys treating the machine debris like a treasure hunt.

The humor in B is stronger because it extends the absurdity visually. The answer: B.

RM+PA MODEL RESPONSE

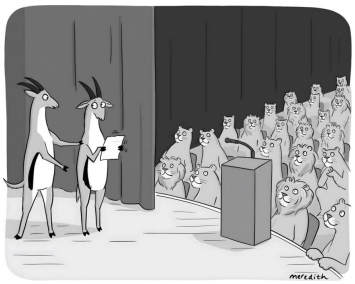
In the cartoon, there are two cowboys in a desert, surrounded by a broken espresso machine, coffee grounds, and other coffee-making tools. They're far from any civilization, yet they're set up like they're in a campsite. The absurd element jumps out immediately—espresso in the Wild West? That anachronistic contrast is the core joke.

The cowboys' bewildered expressions heighten the absurd contrast. The humor lies in this surreal mix of frontier life and modern tech.

Caption A works perfectly because "Grounds still warm" plays on coffee metaphors (fresh beans) while mirroring desert heat (metaphors "hot" coffee vs "hot" environment). No need for niche jargon—just pure visual metaphor clicking while matching the art absurdism without needing niche jargon.
Caption B's "plastic" metaphor is clever tech wordplay but demands niche knowledge (ESPRESSO pods?) while lacking in desert/desert wordplay. Without such tech jargon, B falls flat without same surreal absurdism without jargon overkill.

Winner: A nails it visually and simply while being instantly readable like a cowboy one-liner while looking exactly like cartoon chaos without needing niche jargon. Perfect match! Answer A.

Figure D.3: **Example reasoning on a caption ranking task.** The RM-only model fixates on literal associations (e.g., coffee grounds as evidence of recency), while the RM+PA model integrates the broader absurdity of encountering a modern espresso machine in a cowboy setting. By grounding its reasoning in visual context and avoiding reliance on niche jargon, the RM+PA model produces a more coherent and humor-aware judgment that better reflects the cartoon's visual gag.



VISUAL REFERENCES

Info 1. Two upright, bipedal goats standing on a conference stage
Info 2. One goat holding and reading from a visibly trembling sheet of paper.
Info 3. The second goat positioned beside the first, oriented toward it in a posture.
Info 4. A podium placed on the stage beside the two goats.
Info 5. An audience composed entirely of seated lions facing the stage.
Info 6. All lions sharing identical wide-eyed, vacant facial expressions.
Info 7. Lions sitting upright with paws folded in a human-hand-like manner.
Info 8. A microphone positioned on the stage between the goats and the unused.
Info 9. A background curtain and curved table indicating a formal auditorium setting.
Info10. Lions arranged in symmetrical seating rows with uniform posture and expression.

MODEL RESPONSE

When I look at the cartoon, the scene shows two goats on a stage addressing an audience of lions in theater seats. What makes this absurd is that prey animals are giving a speech to predators, which is essentially the opposite of natural behavior. The visual contrast is heightened by the fact that the lions are sitting calmly while the goats appear nervous, holding notes.

Now, examining the captions.

Caption A ("Just pretend none of them are wearing their teeth") creates a dark humor interpretation where the teeth imply the lions might eat the goats. This requires the viewer to imagine, which adds a layer of unease. However, I notice that in the image, the lions look cheerfully attentive rather than menacing, which makes this interpretation feel a bit forced.

Caption B ("The good news is that you have their full attention") works better because it directly plays off the visual irony. The phrase "their full attention" suggests the speaker (likely a goat) is commenting on how captive the lions appear, while actually implying the lions are **paying too much attention** to them. This creates a witty double meaning - professionally "full attention" means concentration, but but in this context it hints at the predator's curiosity.

The visual cues support Caption B more effectively. The lions' uniform, mesmerized expressions resemble an audience at a motivational talk rather than their natural predatory behavior. Caption B's dry delivery matches the goats' anxious body language while highlighting the absurdity through understatement. It transforms the scene from potentially scary to unintentionally funny.

Caption A's imaginative twist feels disconnected from what we actually see - the lions' expressions suggest admiration, not aggression. The humor works best when amplifying existing visual contrasts without inventing new threats. Caption B achieves this by essence by focusing on the core joke: prey giving speeches to predators.

Answer: B

JUDGE RESPONSE (VISUAL PERCEPTION REWARD)

Info 1: 1.0, Info2: 1.0, Info 3: 1.0, Info 4: 1.0, Info 5: 1.0, Info 6: 1.0, Info 7: 0.0, Info 8: 1.0, Info 9: 1.0, Info 10: 1.0
Total Score: 9.0

Figure D.4: **Visual perception judge response.** Binary scores indicating whether the reasoning reflects curated visual references, providing a direct measure of grounding in salient scene elements.

EVALUATION RULES

- Daily Phrase: 1 if the caption uses a natural, common, everyday expression or idiom.
- Punctuation: 1 if punctuation is used effectively (not missing, not overused, enhances readability or style).
- Wordplay: 1 if there is creative wordplay (puns, double meanings, playful twists) relevant to the cartoon.
- Metaphor: 1 if a clear metaphorical expression relates to the cartoon.
- Punchline: 1 if the caption includes a punchline at the end.
- Otherwise, 0.

SAMPLE MODEL RESPONSE FAVORING A STYLISTICALLY NICE CAPTION

The cartoon shows two people reading a book together.
…
Of the options, A) 'I guess we're not on the same page… literally.' is the best match because it plays on the idiom and the literal situation.
Answer: A

JUDGE RESPONSE (STYLE REWARD)

Daily Phrase: 1, Punctuation:1, Wordplay: 1, Metaphor: 0, Punchline:1

SAMPLE MODEL RESPONSE FAVORING A STYLISTICALLY BAD CAPTION

The cartoon shows a man holding a book.
…
Of the options, B) 'This book is very interesting.' best matches the cartoon.</think>
Answer: B

JUDGE RESPONSE (STYLE REWARD)

Daily Phrase: 0, Punctuation: 1, Wordplay: 0, Metaphor: 0, Punchline: 0

Figure D.5: **Style judge response.** Binary scores across five stylistic dimensions (natural phrasing, punctuation, wordplay, metaphor, and punchline placement), capturing linguistic qualities associated with effective caption writing.

QUESTION

Which caption is funnier?
A) They say cliche- I say classic.
B) You're gonna KILL in that outfit.

MODEL RESPONSE

In the cartoon, I see why this is funny. Looking at the image: a guy in what looks like formal wear—suits, ties, typical shopping vibe. But instead of humans modeling clothes, we have these... creatures. Like, horned or monster-like. That visual irony—expecting humans but seeing these weird animal hybrids. Their postures mimic how models pose seriously, but their appearances are anything but classic. Those sharp horns and unnatural stances contrast hilariously with the idea of fashion critique. Caption B ("You're gonna KILL in that outfit") works on three levels:
1) Visual irony: The "monster" models literally look ready to kill 2) Fashion metaphor: High praise for clothing 3) Double entendre: "Kill" as both compliment and literal threat from their appearance
Caption A misses these layers:
- "Cliche vs classic" doesn't connect with visual monsters
- No wordplay with their physicality - Misses the visual metaphor Caption B maximizes humor through:
Visual metaphor: Monster-like figures "modeling"
Double entendre: Fashion praise vs literal killing
Wordplay: "Kill" as both compliment and visual threat
Irony: Calling monstrous figures "stylish" Caption B uniquely maps to:
1) Monster appearance ("kill")
2) Fashion context ("outfit")
3) Double meaning of "kill"
4) Visual parody of high-end retail
This preserves caption contest tradition of naming banal visual irony. Caption B weaponizes every visual element: unnatural poses + monster aesthetics + fashion metaphor + literal wordplay. Caption B weaponizes visual irony precisely while caption A wastes visual cues." This preserves caption contest tradition of naming banal visual irony. Caption B maximizes visual-fantasy-combination.

HUMAN EXPERT RESPONSE

We are in what looks like a high-end costume or clothing shop for predators. A wolf stands in front of a mirror wearing a full sheepskin suit, complete with fluffy wool like a tailored garment. A salesman stands nearby holding a zebra skin as another possible outfit. Around them are other garments:
- A cowhide draped over a chair
- A goat outfit
- An antelope

Other animal skins, all as though they were merchandise
The wolf is clearly pleased with his choice, the salesman seems to have critiqued it, as if he's a stylist trying to steer the client towards something more original. The caption: "You say cliche, I say classic." In everyday life, a salesperson at a boutique might say, "You say cliché, I say classic," when defending a timeless but overused look. Here, he's not defending it; he's characterizing it as a boring, obvious look with no originality. For the sheep, disguise is not about originality. It's about functionality. Only as a "wolf in sheep's clothing" can he infiltrate a flock and successfully prey on them.

A zebra or a cow disguise might be more creative visually, but it would hardly be useless for the wolf's actual goal. Just imagining a wolf in a cow's clothing or a zebra's clothing is funny in and of itself. It's part of the scene's background humor.

Ultimately, the humor comes from this mid-match. The salesman is judging the choice by aesthetic originality, while the wolf and the viewer know that for hunting sheep, this cliché outfit fit the situation to a T.

Figure D.6: **Model vs. expert reasoning.** The model and expert reach different conclusions but follow similar reasoning steps, indicating alignment in how the cartoon is analyzed even when preferences differ.

E Limitations and Ethical Considerations

While our framework improves multimodal humor understanding, several limitations remain that highlight open challenges in reasoning-based alignment.

Visual perception errors. Despite reinforcement learning with perceptual rewards, the model can misidentify salient visual entities, which propagates through the reasoning process and leads to incorrect caption selection. As shown in Fig. E.1, the model confuses a Viking warrior with the Grim Reaper (scythe) due to superficial visual similarity (e.g., weapon shape), causing a complete mismatch between perceived scene and intended incongruity. Scaling the backbone to 32B substantially reduces such errors, indicating that perceptual grounding benefits from increased representational capacity. However, this improvement is not uniform: stylized cartoons require abstraction over sparse and exaggerated visual cues, which remain underrepresented in standard vision pretraining corpora. This suggests that reasoning supervision alone cannot compensate for systematic perception errors, and that advances in vision encoders or domain-specific visual pretraining are necessary to fully address this limitation.

Shallow cultural grounding. The model sometimes produces fluent but brittle analyses that miss culturally embedded references. As shown in Fig. E.2, an RM-only model latches onto surface wordplay (*"plane"*) and ignores the cartoon's Superman motif—a popular-culture cue. After PA, the model gestures toward the pop-culture context but over-attributes visual evidence (e.g., inventing a *"flying-pose"* link) and still fails to articulate the catchphrase-based joke. This illustrates the challenge of grounding cultural knowledge without hallucination.

Model size and scaling. Our experiments use a 7B-parameter backbone (Qwen2.5-VL-7B). While this provides a fair and reproducible open baseline, it naturally underperforms large proprietary systems such as o3 and GPT-4o. Our focus in this work is not raw scale, but the development of humor-specific priors, reasoning traces, and reward functions. We also scaled our pipeline to a larger open backbone, Qwen2.5-VL-32B model, and observe the improvement in the performances on matching and ranking tasks as presented in the Table 1.

Cultural and linguistic specificity. Our dataset is centered on the New Yorker Cartoon Caption Contest, which reflects predominantly Anglophone and U.S.-centric humor. As a result, the learned reasoning patterns may not transfer to other cultural contexts, where humor relies on different conventions, references, and linguistic structures. Extending this framework to multilingual and culturally diverse datasets is necessary to evaluate the generality of humor-aware reasoning.

Subjectivity in evaluation. Humor is inherently subjective, and even within NYCC, crowd judgments and editorial selections represent only a subset of possible interpretations. Our reward functions approximate these preferences through perceptual and stylistic criteria, but cannot fully capture the diversity of human responses. This introduces an unavoidable gap between automated evaluation and human judgment, motivating future work on more robust and diverse human evaluation protocols.

Data access and licensing. Parts of our IM corpus are derived from copyrighted sources (e.g., podcast transcripts, published boooks). Due to licensing restrictions, we cannot release this material directly, though we will provide evaluation splits and full preprocessing pipelines. This ensures reproducibility while respecting intellectual property.

Despite these limitations, our work represents a step toward aligning multimodal reasoning models with one of the most elusive facets of human intelligence: humor. Addressing cultural, subjective, and computational challenges offers fertile ground for future research.

RM GENERATED DESCRIPTION (7B MODEL)

The cartoon shows a modern kitchen scene with a woman in a bathrobe holding a breakfast tray. Across from her, there's a man dressed as the Grim Reaper—complete with the scythe, horned helmet, and straw-like hair—but he's just sitting at the table eating a donut like it's totally normal. The absurdity hits immediately because these two couldn't be more different—one's a mundane homebody, the other a mythological horror figure, both acting like it's just another Tuesday.

RM+PA GENERATED DESCRIPTION (7B MODEL)

In the cartoon, I see a modern kitchen setting with a woman preparing breakfast (wearing bathrobe, holding tray) contrasted by a terrifying Grim Reaper-like figure seated at the table. The core absurdity centers on this supernatural entity casually eating breakfast as if it were normal. Key visual cues include the Reaper's scythe/soup spoon, ragged robes, and menacing horns/hat - all jarring against the mundane domestic backdrop. The woman's paused action (likely mid-carry of breakfast) suggests she's reacting to this surreal intrusion.

RM GENERATED DESCRIPTION (32B MODEL)

The cartoon depicts a kitchen scene where a woman in a bathrobe is serving breakfast to the Grim Reaper, who's sitting at the table with his scythe. The humor comes from the juxtaposition of a mundane domestic setting with the supernatural figure of Death. First, the visual elements: The Grim Reaper's presence in a kitchen is inherently absurd—he's usually associated with deathbeds, not breakfast tables. The woman's nonchalant attitude amplifies the humor, as if serving Death is a normal occurrence. The scythe on the floor and the checkered tablecloth add to the domestic surrealism.

RM+PA GENERATED DESCRIPTION (32BMODEL)

In the cartoon, I see why this is funny. Looking at the image: a typical kitchen setup—checkered tablecloth, someone in a robe, maybe having breakfast. But the big guy in the corner... those horns, the scythe, the long hair... classic devil vibes. But here he is, just sitting there, acting normal. That visual irony—expecting something intimidating, but he's... mundane? That's the twist: a scary figure in everyday life. His size contrasts with the ordinary setup—big, horned guy in a tiny kitchen corner while the other seems nonchalant.

Winning Caption: *"We're all out of the blood of your enemies. You'll have to settle for marmalade."*

Contest #132

Figure E.1: **Limitation: Visual perception errors.** All 7B-scale variants of our model misidentify the foreground figure as the Grim Reaper (with scythe), while the humor relies on the incongruity of a Viking warrior in a domestic kitchen. This perceptual error cascades into mismatched reasoning and incorrect caption selection. When scaling our approach to a larger model, we observe that it initially misidentifies objects in the early stages of training, however, with our introduced approach, which incorporates two additional rewards in PA setup, the model corrects these shortcomings and begins generating responses that correctly identify the objects.

Ethical considerations. Modeling humor raises additional ethical concerns. Humor often relies on cultural stereotypes, sensitive topics, or implicit social norms, and models trained on such data may reproduce or amplify biases. Furthermore, systems capable of generating humor-like captions could be misused to produce offensive or inappropriate content. While our work focuses on controlled evaluation settings, deploying such models in real-world applications requires careful consideration of content moderation, cultural sensitivity, and user context.

F Cross-Dataset Generalization

Although our model is trained exclusively on the New Yorker Cartoon Caption Contest (NYCC), we evaluate its ability to generalize to humor settings that differ in visual structure,

RM GENERATED DESCRIPTION

Sounds like a bird....no, a plane!' This plays on Superman's hearing, but it's less connected to the scene. The stethoscope is meant for internal sounds, but the doctor's comment about external noises like planes feels disconnected. It doesn't tie back to the main joke of the pointless exam.

RM+PA GENERATED DESCRIPTION

Caption B works perfectly! The doctor's stethoscope position mirrors Superman's iconic flying pose. That visual pun makes it feel organic. The caption extends the joke - we imagine Superman's powers manifesting as heart sounds. Plus, the doctor's bewildered expression matches the plane revelation. Every visual element supports this. Caption B transforms the image's tension: we wonder about unseen sounds, imagine Superman's reaction. The medical context even justifies the doctor's delivery.

Figure E.2: **Limitation: Shallow cultural understanding.** The humor depends on a Superman reference (e.g., the well-known *"it's a bird… it's a plane…"* trope). The RM-only model fixates on superficial wordplay; the RM+PA model recognizes the pop-culture context but hallucinates a visual link and does not fully recover the cultural script.

linguistic form, and task formulation. We conduct zero-shot experiments on two external benchmarks:

- **YesBut** (Hu et al., 2024): A two-panel visual humor dataset centered on contrastive reasoning ("yes…but…"), which differs from NYCC's single-image incongruity-resolution format. Among its four subtasks, we focus on the classification-based *Philosophy* and *Title* tasks, which are most comparable to caption selection.
- **DeepEval** (Yang et al., 2024): A broad multimodal benchmark with a small humor subset (2.9%). The evaluated tasks, *DeepSemantics*, *Description*, and *Title*, respectively measure semantic alignment, descriptive accuracy, and title appropriateness—none are cartoon-captioning–specific.

Despite the substantial shift in task structure and domain, our model shows consistent improvements over its base counterpart. On YesBut, we observe large absolute gains of +31.7 and +34.2 points on the Philosophy and Title tasks, respectively (Table F.1), indicating strong transfer to contrastive humor reasoning. On DeepEval, the model also improves performance across all evaluated tasks (Table F.2), particularly in semantic alignment and descriptive accuracy, despite minimal stylistic overlap with NYCC.

Table F.1: **Zero-shot Generalization to YesBut (Hu et al., NeurIPS 2024).** Accuracy (%) on the Philosophy and Title subtasks. Our models are trained only on NYCC data yet show strong transfer, particularly for the 7B backbone.

Model	Philosophy	Title
Qwen2.5-VL-7B-Instruct (Base)	43.19	29.11
IRS-7B (Ours)	**74.90**	**63.32**

Table F.2: **Zero-shot Generalization to DeepEval (Yang et al., ACL 2024).** Performance on the *humorous* subset (29 images, 2.9% of the benchmark). We report accuracy on three tasks: DeepSemantics, Description, and Title.

Model	DeepSemantics	Description	Title
Qwen2.5-VL-7B-Instruct (Base)	10.34	24.13	34.48
IRS-7B (Ours)	**54.43**	**100.00**	**63.18**

These results provide evidence that the proposed reasoning-trace supervision and humor-aware alignment enable the model to acquire transferable multimodal reasoning patterns rather than overfitting to NYCC-specific stylistic templates.

G Caption Generation without Task-Specific Training

Although IRS is designed for caption selection and ranking, we observe that models trained with IRS can also generate plausible captions in a zero-shot manner. This is notable because caption generation is not explicitly included in our training objectives. We evaluate this capability qualitatively by prompting the model to produce a single caption for a given cartoon, following the New Yorker Caption Contest format.

As shown in Fig. G.1-G.3, the IRS-trained model generates captions that are concise, conversational, and grounded in the visual incongruity, often reflecting captionist-style reasoning patterns such as wordplay, irony, and narrative reinterpretation. Compared to the base model, which often produces plausible but less structured captions, the IRS model more consistently captures the underlying humorous mechanism and its connection to the visual incongruity. This suggests that supervising structured reasoning, rather than optimizing directly for generation, enables the model to internalize transferable patterns that extend naturally to captioning. These results further support our central claim: explicitly modeling and supervising reasoning structure leads to more generalizable capabilities, even in tasks not seen during training.

H LLM Usage

LLMs were used to polish the writing and improve the clarity and flow of the text. All final revisions were reviewed and edited by the authors.

The image is a cartoon from the New Yorker Cartoon Caption Contest. Your task is to generate a funny caption for the cartoon.
Rules:
- The caption MUST be a single line of dialogue spoken by a character.
- It must sound natural and conversational.
- Do NOT describe the image.
- Do NOT explain the joke.
- Keep it concise (ideally under 15 words).

First, think through the reasoning process. Then provide the final answer. Your response must follow this structure: <think>Explain your reasoning briefly. Identify the incongruity and comedic twist briefly.</think> <answer>Return ONLY the caption, exactly as written.</answer>

BASE MODEL RESPONSE

The image depicts a humorous take on the biblical story of Moses receiving the Ten Commandments. Instead of a traditional, divine presence, the commandments are being handed down by a cloud that resembles a sheep, complete with glasses and a relaxed demeanor. The incongruity lies in the unexpected and whimsical portrayal of a divine moment, replacing the usual solemnity with a lighthearted, almost absurd scenario. The comedic twist comes from the juxtaposition of a serious religious narrative with a casual, almost mundane delivery by a cloud that looks like a sheep. The caption should capture this incongruity in a conversational and humorous way, reflecting the absurdity of the situation while maintaining a single line of dialogue.

"Okay, Moses, here's your list—don't lose it this time, okay?"

IRS MODEL RESPONSE

In the cartoon, I see why this is funny. Looking at the visual: a figure in traditional robes, desert-like setting—classic"divine revelation" vibe. That big cloud dominates visually, but instead of the expected heavenly imagery, it's… glasses and paperwork? Big visual irony: effortlessness meets bureaucracy. That cloud "guy" looks anything but majestic—framed, literalized. Meanwhile, the human below awaits… only to get… a form? Key visual irony: sacred expectation meets… admin? The irony is gold:
1) Visual irony: Effortless cloud "authority" vs physical robes
2) Symbolism: Timeless wisdom vs modern triviality
3) Parody: Divine moment reduced to…
4) Metaphor: Clouds=omniscience vs literal cloudiness
5) Double entendre: Clouds/technology

Caption must highlight:
- Power dynamic
- Formality parody – Divine/mechanical irony
- Bureaucracy metaphor

"Can you fill out this form while I write the Ten Commandments?" This works because:
Visual irony: Cloud "authority" holding… papers
Symbolic: Commandments vs bureaucracy
Metaphor: Cloud tech vs tradition
Wordplay: Clouds/technology
Parody: Divine decree meets red tape

The visual irony of sacred expectation vs mundane paperwork. This preserves caption contest tradition of visual irony.

"Can you fill out this form while I write the Ten Commandments?"

Figure G.1: **Zero-shot caption generation with IRS (divine-bureaucratic incongruity).** The IRS model captures the clash between a sacred revelation and mundane paperwork, grounding the humor in a reinterpretation of divine authority as bureaucratic process.

The image is a cartoon from the New Yorker Cartoon Caption Contest. Your task is to generate a funny caption for the cartoon.
Rules:
- The caption MUST be a single line of dialogue spoken by a character.
- It must sound natural and conversational.
- Do NOT describe the image.
- Do NOT explain the joke.
- Keep it concise (ideally under 15 words).

First, think through the reasoning process. Then provide the final answer. Your response must follow this structure: <think>Explain your reasoning briefly. Identify the incongruity and comedic twist briefly.</think> <answer>Return ONLY the caption, exactly as written.</answer>

BASE MODEL RESPONSE

The image depicts a fairy hovering over a bed, holding a plunger and a jug, with a cloud of bubbles and hearts around her. The fairy appears to be in the process of cleaning or "plumbing" the bed. The incongruity lies in the fairy's unusual task—typically, fairies are associated with magic, wishes, and romance, not plumbing. The comedic twist arises from the unexpected and mundane nature of her actions, juxtaposed with her magical appearance and the romantic bubbles and hearts. Given this setup, the caption should reflect the fairy's actions in a humorous and conversational tone, emphasizing the absurdity of her task.

"Just a little bedtime maintenance to keep your dreams flowing smoothly!"

IRS MODEL RESPONSE

In the cartoon, I see why this is funny. Looking at the visual: a typical setup—person in bed, looking normal. But above them, that "fairy" figure with wings, floating on clouds with hearts... classic fairy tale imagery: mythical, hopeful, magical. Except... wait, the visual irony: big-winged, cloud-hovering thing holding something... but instead of a wand, it's a... plunger? That visual irony clicks: expected fairy imagery meets... a plumbing tool? Those big wings, effortful pose... but a plunger? That visual irony—effort meets absurdity: elaborate setup for... fixing clogs?

The core irony:
1) Visual metaphor: Effortless fairy trope (wings/clouds/hearts)
2) Object/subtext: High-minded pose with a mundane plunger
3) Symbolism: Magic (clouds/wings) vs Functionality (plunger)
4) Wordplay potential: "Fairy" vs "Fixer"

Caption should highlight:
Visual irony: Fairy imagery vs utilitarian object
Symbolism: Magic vs Function
Double entendre: "Fairy" (mythical) vs "Plumbing fairy"
Wordplay: "Granting wishes" vs "Unclogging"
Caption delivering irony: "Sorry, I can't grant wishes—just unclog drains."

This works because:
Visual: Fairy above bed with wings/plunger
Metaphor: Effortful magic meets plumbing
Symbolism: Clouds/hearts vs mundane tool
Double play: "Fairy" (mythical) vs "Fixer"
Irony: High expectations vs Low-function
Precisely subverting fairy tale tropes."
This maximizes visual irony and metaphorical clash.

"Sorry, I can't grant wishes—just unclog drains."

Figure G.2: **Zero-shot caption generation with IRS (magic-mundane incongruity).** The IRS model resolves the mismatch between fairy-tale imagery and a utilitarian task, producing a caption that reframes magical expectation as everyday functionality.

The image is a cartoon from the New Yorker Cartoon Caption Contest. Your task is to generate a funny caption for the cartoon.
Rules:
- The caption MUST be a single line of dialogue spoken by a character.
- It must sound natural and conversational.
- Do NOT describe the image.
- Do NOT explain the joke.
- Keep it concise (ideally under 15 words).

First, think through the reasoning process. Then provide the final answer. Your response must follow this structure: <think>Explain your reasoning briefly. Identify the incongruity and comedic twist briefly.</think> <answer>Return ONLY the caption, exactly as written.</answer>

BASE MODEL RESPONSE

The image depicts two scientists observing a group of mice in a laboratory setting. The mice are dressed in lab coats, holding notebooks, and appear to be conducting their own experiment or observing something. The humor arises from the unexpected and anthropomorphic behavior of the mice, mimicking the scientists' actions in a recursive and absurd manner. The incongruity lies in the reversal of roles—typically, scientists observe animals, but here, the animals are observing the scientists, creating a comedic twist.

Given this setup, the caption should reflect the mice's perspective or reaction to being observed while they are also observing. A natural and conversational line spoken by one of the mice would capture the humor effectively.

"I told you they'd notice us watching them. Now what do we do?"

IRS MODEL RESPONSE

In the cartoon, I see why this is funny. The visual irony jumps out: lab setup, humans in lab coats looking serious—typical sci-fi mice testing vibe. But wait... those "subjects" under the microscope? Tiny figures acting just like the "real" scientists: coats, notebooks, posture. Total role reversal visual irony—effortless irony of expected lab dynamics mirrored. Those mice look humanized while humans observe... but the mice are doing the exact same thing. That visual irony: effortful observation meets... observed observers. Caption should highlight this irony:
- Human scientists meticulously studying...
- ...mice who look equally scientific
- Parody of lab hierarchy: observer vs "subject"
- Double layer of irony: mice mimicking human postures
- Metaphor: Experimenting on... experimenters?

Caption needs to verbalize the visual irony:
Scientists studying mice... who study them back. That meta-layer:
1) Expected lab power dynamic
2) Mice performing human-like scrutiny
3) Visual parody: coats/papers vs cages
4) Linguistic irony: "testing the testers"

This preserves caption contest tradition of visual irony.
Maintaining visual metaphor: lab formality meets absurd role reversal.

"Well, our data shows these human subjects exhibit excessive note-taking behavior."

Figure G.3: **Zero-shot caption generation with IRS (role-reversal incongruity).** The IRS model captures the recursive reversal between observers and subjects, grounding the humor in a mirrored interpretation of scientific authority.