

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

Hatice Merve Vural^{1,†} Doga Kukul^{1,2} Ege Erdem Ozlu¹ Demir Ekin Arikan¹

Bob Mankoff³ Erkut Erdem^{2,4} Aykut Erdem^{1,2}

¹Koç University, Istanbul, Turkey ²KUIS AI Center, Istanbul, Turkey

³Air Mail and Cartoon Collections ⁴Hacettepe University, Ankara, Turkey

[†]Correspondence: havural24@ku.edu.tr

 [Project page](#)

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 improves performance across caption matching and ranking, with IRS-72B achieving the strongest model performance on ranking (76.10%), surpassing both non-expert human performance and all evaluated open- and closed-source multimodal baselines. Zero-shot transfer further shows that IRS learns generalizable reasoning patterns.

1 Introduction

Humor is a demanding form of human intelligence that requires visual perception, cultural knowledge, and creative reasoning (Kazemi et al., 2025). Unlike standard benchmarks, humor depends not only on choosing an answer but on 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 for 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 treats NYCC as a classification or ranking benchmark, it also offers insight into the reasoning processes underlying visual humor.

In this work, we propose *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. Grounded in incongruity-resolution theory and informed by expert analyses of professional captionist practice (Wood, 2024; Mankoff, 2002), IRS uses *captionist reasoning traces* to supervise intermediate reasoning steps. Across 7B, 32B, and 72B models, IRS 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 show 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 and Raskin, 1991), frame shifting (Coulson and 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 and 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 and 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 and Raskin, 1991), showing improved caption quality over prompting baselines. IRS is complementary: while HOMER 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). Yet, 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., 2026), Visual-RFT (Liu et al., 2025b), and Perception-R1 (Xiao et al., 2026) 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

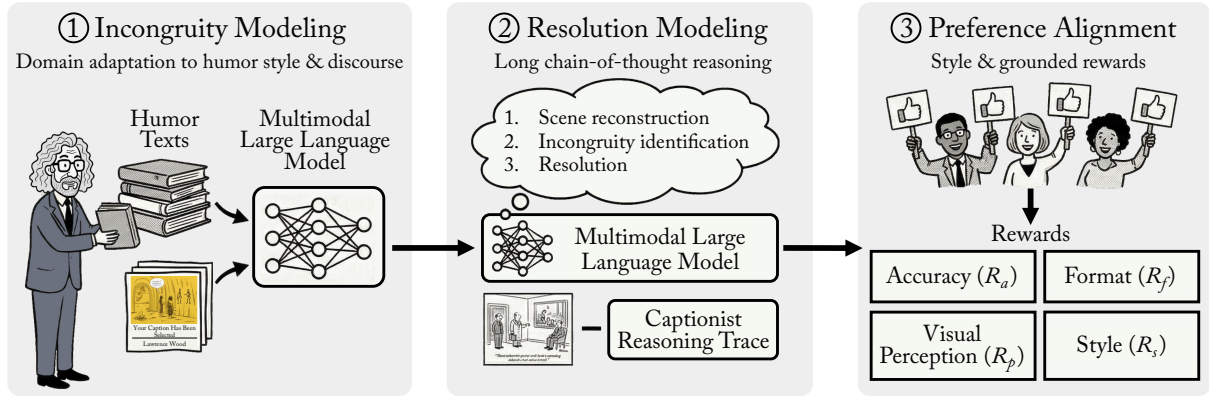


Figure 1: **Overview of Incongruity-Resolution Supervision (IRS).** IRS models humor understanding as a structured reasoning process with three components: (1) *Incongruity Modeling (IM)*, which identifies mismatches in the visual scene; (2) *Resolution Modeling (RM)*, which constructs coherent reinterpretations of these mismatches; and (3) *Preference Alignment (PA)*, 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.

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 often 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 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 show 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 App. B.*

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, 2024a) 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. We assess final trace quality using a structured cross-model audit with DeepSeek-R1 (DeepSeek-AI et al., 2025), NVLM-D-72B (Dai et al., 2024), and Gemini-2.5-Flash (Comanici et al., 2025) as independent external judges. The traces receive strong audit scores, suggesting that RM provides substantive humor-reasoning supervision rather than generic ratio-

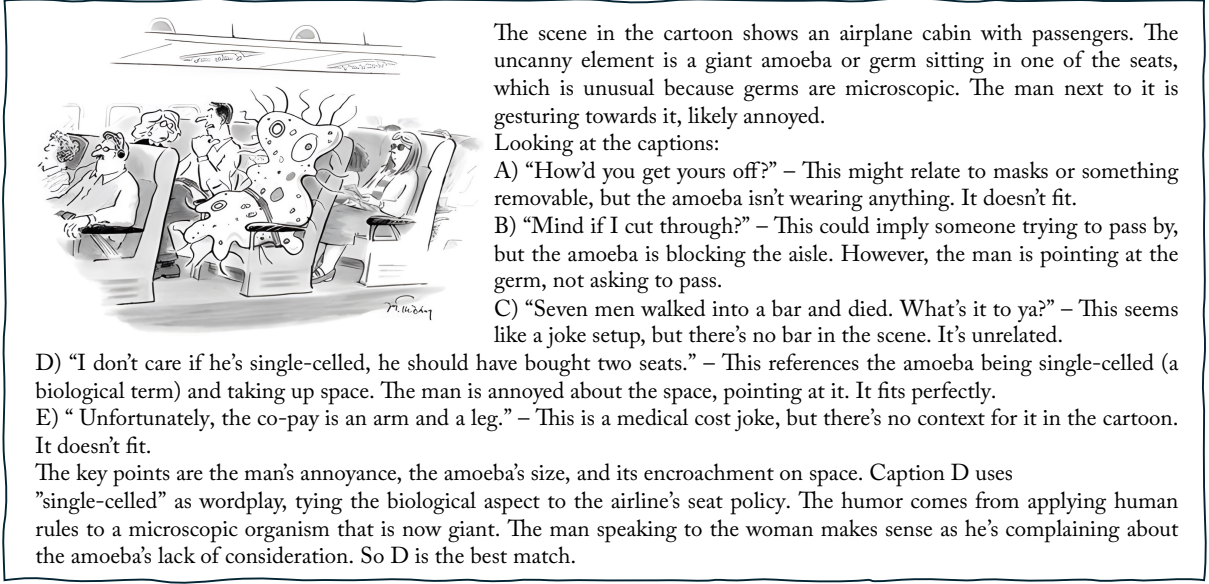


Figure 2: **Example captionist reasoning trace under IRS (matching).** Given a cartoon and five candidate captions (A-E), the trace shows structured reasoning: reconstructing the scene, identifying the key incongruity, evaluating alternatives, and selecting the caption best resolving the mismatch. The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

nales.

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: $\langle \text{think} \rangle \text{reasoning steps} \langle / \text{think} \rangle \langle \text{answer} \rangle \text{final choice} \langle / \text{answer} \rangle$. This structured supervision encourages models to internalize captionist-style reasoning rather than relying on surface-level patterns. Fig. 2 shows 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 and the self-audit analysis are respectively provided in App. F and App. K, while additional examples are given in App. I.*

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:

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

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 combines standard correctness signals with humor-specific components. The first two use common multimodal RLHF practice:

- (i) **Accuracy** (R_a): verifies whether the final caption 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:

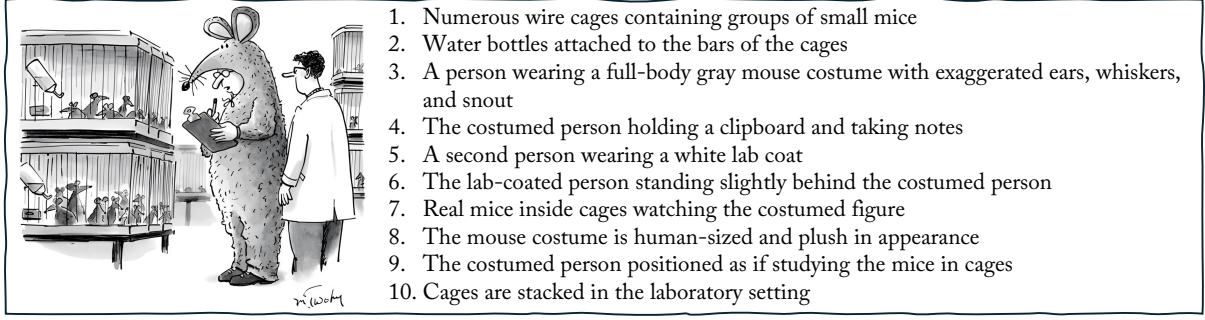


Figure 3: **Curated visual references for perception alignment in IRS.** For each cartoon, we collect concise descriptions of entities, scene context, and key incongruities. These references serve as anchors for evaluating visual grounding and computing the perception reward.

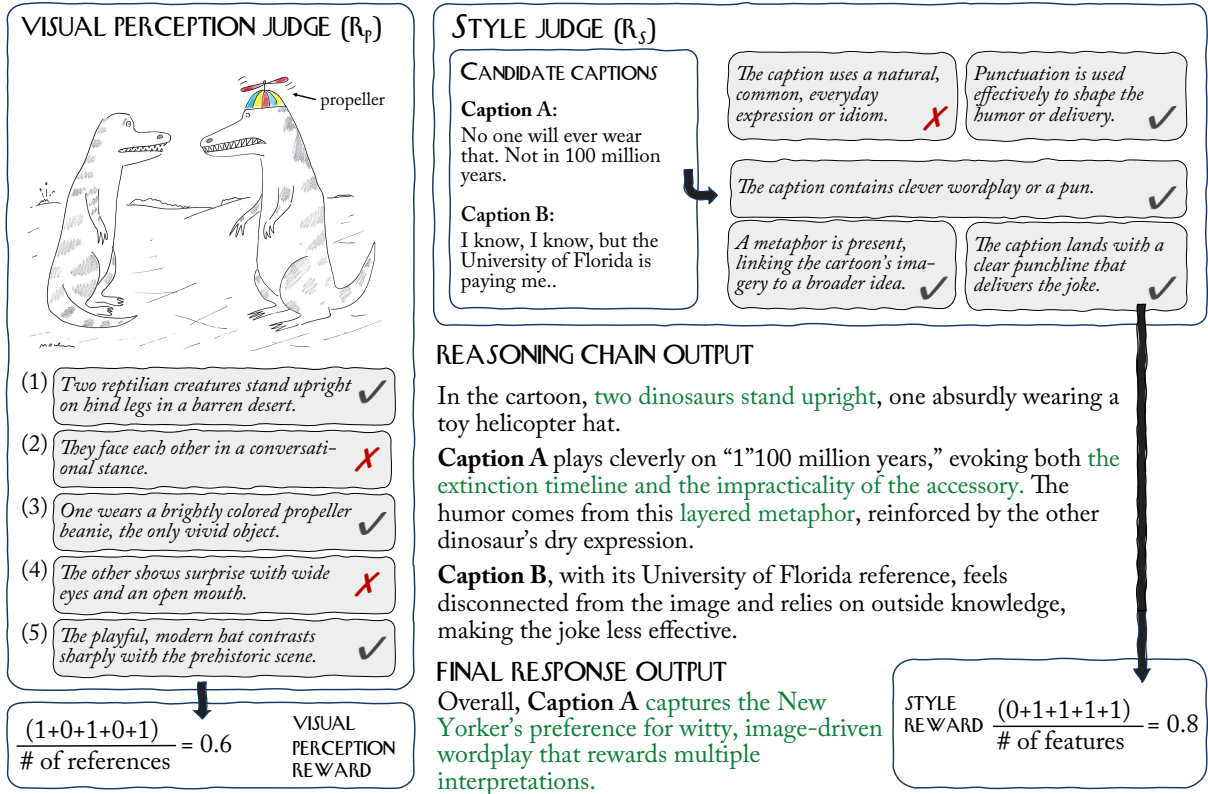


Figure 4: **Preference alignment in IRS via judge-based rewards.** Given a cartoon and candidate captions, two judges evaluate reasoning quality: a *visual perception* judge checks grounding in salient elements and incongruities, while a *style* judge assesses linguistic quality. Their binary outputs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

(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:

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

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., 2026), 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 App. F.*

4 Experimental Setup

We evaluate IRS in two settings: in-domain NYCC caption selection and zero-shot cross-dataset generalization. The in-domain evaluation covers four benchmark settings spanning caption matching and humor ranking, while the cross-dataset evaluation tests whether the learned reasoning patterns transfer beyond NYCC. *Training configurations and optimization details are provided in App. E.*

Datasets and Tasks. Our in-domain evaluation uses four NYCC-based settings. From Hessel et al. (2023), we evaluate: (i) **Matching**, where a model selects the gold caption from five candidates given a cartoon image; and (ii) **Ranking**, where a model selects the crowd-preferred caption from two options. From Zhou et al. (2025), we evaluate: (iii) **10-vs-1000**, which contrasts a top-10 caption with one ranked 1000–1009; and (iv) **30-vs-300**, which contrasts a caption ranked 30–39 with one ranked 300–309. These tasks probe complementary aspects of humor reasoning. Matching tests whether models can identify the visual interpretation that best matches the cartoon, while ranking emphasizes preference alignment between plausible captions. The 30-vs-300 split is especially challenging because both captions are crowd-ranked and stylistically competent, requiring fine-grained preference discrimination.

For cross-dataset generalization, we additionally evaluate IRS zero-shot on two external humor-related benchmarks: **YesBut** (Hu et al., 2024), a two-panel visual humor dataset centered on contrastive reasoning, where we use the classification-based *Philosophy* and *Title* subtasks; and **DeepEval** (Yang et al., 2024), a broader multimodal benchmark with a small humor subset, where we evalu-

ate *DeepSemantics*, *Description*, and *Title*. These datasets differ from NYCC in visual structure, task format, and annotation scheme, allowing us to test whether IRS learns transferable reasoning patterns rather than NYCC-specific heuristics. 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 (OpenAI, 2025) as representative high-performing proprietary systems. Their performance provides a useful reference point, although such models may benefit from data and scale advantages unavailable to open-weight models.
- **Open multimodal reasoning models.** We benchmark against recent open-weight models trained for multimodal reasoning: GLM-4.1V-9B-Thinking (V Team et al., 2026), Kimi-VL-A3B-Thinking-2506 (Kimi Team et al., 2025), LlamaV-o1 (Thawakar et al., 2025), and Qwen2.5-VL at 7B, 32B, and 72B scales (Bai 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 pre-training, 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 sus-

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-Think	59.40	55.60	52.28	49.14
Kimi-VL-A3B-Think	54.00	57.45	52.30	52.01
LlamaV-o1	43.67	50.90	48.57	49.42
Qwen2.5-VL-7B-Inst	42.67	55.06	50.57	47.99
Qwen2.5-VL-32B-Inst	46.67	49.87	52.00	44.57
Qwen2.5-VL-72B-Inst	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

tained 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 App. F.

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 baselines, o3 and o4-mini remain highly competitive, with o3 achieving the highest matching accuracy overall. Their strong performance may reflect both architectural

sophistication and possible data advantages, including broader familiarity with New Yorker-style editorial content 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. *Qualitative results demonstrate that IRS-trained models can generate plausible, captionist-style captions despite not being explicitly trained for this task (App. L). We also audit inference-time reasoning quality, showing that IRS improves humor-specific rationale dimensions beyond answer accuracy (App. K).*

5.2 Comparison with Inference-Time CoT Prompting

A natural question is whether the gains from IRS could be recovered simply by eliciting step-by-step reasoning at inference time, without any of the proposed supervision. To test this, we evaluate controlled zero-shot and few-shot CoT baselines on the same Qwen2.5-VL-7B and 32B backbones, using identical evaluation sets, decoding protocol, and

¹OpenAI had announced a partnership with Condé Nast to display content from brands including *The New Yorker* within their products (OpenAI, 2024b). This does not imply direct training on NYCC captions, but it may provide broader exposure to New Yorker-style editorial language and humor.

Table 2: **Comparison of different prompting strategies and IRS.** Accuracy (%) on caption matching and ranking tasks. Bold values indicate the best performance in each column.

Method	(Hessel et al., 2023)		(Zhou et al., 2025)	
	Matching	Ranking	10-vs-1000	30-vs-300
<i>Qwen2.5-VL-7B</i>				
Direct Prompting	42.67	55.06	50.57	47.99
Zero-shot CoT	46.33	55.32	48.29	50.86
Few-shot CoT	44.67	51.69	47.43	51.14
IRS (Ours)	59.67	64.42	56.29	53.14
<i>Qwen2.5-VL-32B</i>				
Direct Prompting	46.67	49.87	52.00	44.57
Zero-shot CoT	49.33	57.66	49.71	48.86
Few-shot CoT	50.67	54.03	55.43	48.00
IRS (Ours)	62.67	68.05	62.86	53.14

structured output format. As Table 2 shows, IRS achieves the best performance on all four tasks at both backbone scales, improving over the strongest CoT configuration per task by 2.00–13.34 points at 7B and 4.28–12.00 points at 32B. Notably, few-shot CoT does not consistently outperform zero-shot CoT, indicating that additional in-context reasoning demonstrations alone provide no reliable benefit on these tasks.

These controls indicate that inference-time reasoning elicitation alone cannot account for the gains. Rather, the improvement comes from explicitly supervising *what* the model reasons about—the visual incongruity, its plausible resolution, and the comparative quality of humorous interpretations—and aligning that reasoning with humor-specific grounding and stylistic signals.

5.3 Ablation on Sources of Supervision

Table 3 isolates the contribution of each IRS component on the 7B backbone. Three findings stand out. 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. IM, by contrast, is only helpful when combined with RM. On its own, it does not 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 with-

Table 3: **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
+ IM	41.00	51.69	50.29	51.43
+ RM	47.00	56.88	49.71	50.86
+ IM + RM	49.00	56.88	54.57	49.14
+ 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

out resolution supervision can introduce instability. 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.

5.4 Ablation on Reward Functions

Table 4 isolates the contribution of reward signals within Preference Alignment (PA). Accuracy and format rewards ($R_a + R_f$) already provide a strong baseline, improving matching performance over the IM+RM model. Adding humor-aware rewards yields more targeted gains rather than uniform improvements across all tasks. The visual perception reward (R_p) improves ranking and 10-vs-1000 performance, suggesting that explicit grounding is most useful when the model must compare plausible captions. The style reward (R_s) provides smaller but consistent benefit on ranking, aligning with its role in encouraging captionist-consistent expression. These results indicate that R_p and R_s should be interpreted primarily as reasoning-shaping rewards: they are designed to improve visual grounding and stylistic quality, rather than to yield uniform accuracy improvements across all tasks. While R_a directly supervises caption selection, R_p encourages explanations grounded in visual incongruities and R_s encourages stylistically appropriate humor analysis. Thus, the humor-aware rewards complement correctness-based RL by guiding the structure and quality of the rationale.

5.5 Cross-Dataset Generalization

We evaluate zero-shot transfer beyond NYCC using the external benchmarks described in Sec. 4. Despite differences in visual structure, task format, and annotation scheme, IRS consistently improves

Table 4: **Ablation on reward functions.** Humor-aware rewards 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

Table 5: **Zero-shot Generalization to YesBut (Hu et al., 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-Inst (Base)	43.19	29.11
IRS-7B (Ours)	74.90	63.32

Table 6: **Zero-shot Generalization to DeepEval (Yang et al., 2024)** Performance on the *humorous* subset (29 images, 2.9% of the benchmark). We report accuracy on three tasks: DeepSemantics, Description, and Title.

Model	DeepSem	Descr	Title
Qwen2.5-VL-7B-Inst (Base)	10.34	24.13	34.48
IRS-7B (Ours)	54.43	100.00	63.18

over its base model. On YesBut (Hu et al., 2024), it gains +31.7 and +34.2 points on *Philosophy* and *Title*, respectively (Table 5), indicating transfer to contrastive humor reasoning. On DeepEval (Yang et al., 2024), it improves across all evaluated tasks (Table 6), especially semantic alignment and descriptive accuracy, despite minimal stylistic overlap with NYCC.

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 pre-training, 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 ap-

proaching 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.

Limitations

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. 1.7, 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. 1.8, 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 Qwen2.5-VL backbones with 7B, 32B, and 72B parameters. While this range provides a fair and reproducible open baseline, it generally underperforms large proprietary systems such as o3 and GPT-4o on matching and harder ranking tasks. Our

focus in this work is not raw scale, but the development of humor-specific priors, reasoning traces, and reward functions. Scaling IRS to the Qwen2.5-VL-72B backbone further improves performance, particularly on the standard ranking task, 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 books). 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.

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, deploy-

ing such models in real-world applications requires careful consideration of content moderation, cultural sensitivity, and user context.

Acknowledgements

We thank all reviewers for their valuable comments and suggestions, which helped improve the final version of this work. This work was supported in part by the KUIS AI Center Research Awards to Doga Kukul. Erkut Erdem acknowledges funding from the TUBITAK 2247-A Program under Award 123C542. 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. The authors also thank Lambda AI for providing compute credits, and Google for the Google Cloud Platform credits awarded through its Gemini Academic Program.

References

- AllenAI. 2024. Olmo-mix-1124 dataset. <https://huggingface.co/datasets/allenai/olmo-mix-1124>.
- Salvatore Attardo. 1994. *Linguistic Theories of Humor*. Approaches to Semiotics. Mouton de Gruyter.
- Salvatore Attardo and Victor Raskin. 1991. *Script theory revis(it)ed: Joke similarity and joke representation model*. *Humor: International Journal of Humor Research*, 4(3-4):293–347.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. *Qwen2.5-vl technical report*. *Preprint*, arXiv:2502.13923.
- Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. *Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities*. *Preprint*, arXiv:2507.06261.
- Seana Coulson and Marta Kutas. 2001. *Getting it: Human event-related brain response to jokes in good and poor comprehenders*. *Neuroscience Letters*, 316(2):71–74.

- Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuolin Yang, Zihan Liu, Jon Barker, Tuomas Rintamaki, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. [Nvlm: Open frontier-class multimodal llms](#). *Preprint*, arXiv:2409.11402.
- 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, and 181 others. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *Preprint*, arXiv:2501.12948.
- Samuel Doogan, Aniruddha Ghosh, Hanyang Chen, and Tony Veale. 2017. [Idiom savant at Semeval-2017 task 7: Detection and interpretation of English puns](#). In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 103–108, Vancouver, Canada. Association for Computational Linguistics.
- Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023. [Do androids laugh at electric sheep? humor “understanding” benchmarks from the new yorker caption contest](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 688–714, Toronto, Canada. Association for Computational Linguistics.
- Zachary Horvitz, Jingru Chen, Rahul Aditya, Harshvardhan Srivastava, Robert West, Zhou Yu, and Kathleen McKeown. 2024. [Getting serious about humor: Crafting humor datasets with unfunny large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 855–869, Bangkok, Thailand. Association for Computational Linguistics.
- Nabil Hossain, John Krumm, and Michael Gamon. 2019. [“president vows to cut <taxes> hair”: Dataset and analysis of creative text editing for humorous headlines](#). In *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)*, pages 133–142, Minneapolis, Minnesota. Association for Computational Linguistics.
- Nabil Hossain, John Krumm, Tanvir Sajed, and Henry Kautz. 2020. [Stimulating creativity with FunLines: A case study of humor generation in headlines](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 256–262, Online. Association for Computational Linguistics.
- Zhe Hu, Tuo Liang, Jing Li, Yiren Lu, Yunlai Zhou, Yiran Qiao, Jing Ma, and Yu Yin. 2024. [Cracking the code of juxtaposition: Can ai models understand the humorous contradictions](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 47166–47188. Curran Associates, Inc.
- Vedaant V Jain, Gabriel Kreiman, and Felipe dos Santos Alves Feitosa. 2025. [Humordb: Can ai understand graphical humor?](#) In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 604–613.
- 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. 2025. [BIG-bench extra hard](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26473–26501, Vienna, Austria. Association for Computational Linguistics.
- 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, and 76 others. 2025. [Kimi-vl technical report](#). *Preprint*, arXiv:2504.07491.
- Yuan-Hong Liao, Sven Elfle, Liu He, Laura Leal-Taixé, Yejin Choi, Sanja Fidler, and David Acuna. 2025. [Longperceptualthoughts: Distilling system-2 reasoning for system-1 perception](#). In *Second Conference on Language Modeling*.
- Qianchu Liu, Sheng Zhang, Guanghui Qin, Timothy Osowski, Yu Gu, Ying Jin, Sid Kiblawi, Sam Preston, Mu Wei, Paul Vozila, Tristan Naumann, and Hoifung Poon. 2025a. [X-reasoner: Towards generalizable reasoning across modalities and domains](#). *Preprint*, arXiv:2505.03981.
- Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. 2025b. [Visual-rft: Visual reinforcement fine-tuning](#). In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2034–2044.
- Bob Mankoff. 2002. *The Naked Cartoonist: A New Way to Enhance Your Creativity*. Black Dog & Leventhal, New York, NY.
- Bob Mankoff. 2014. *How About Never—Is Never Good for You?: My Life in Cartoons*. Henry Holt and Company, New York, NY.
- Rada Mihalcea and Carlo Strapparava. 2006. [Learning to laugh \(automatically\): Computational models for humor recognition](#). *Computational Intelligence*, 22(2):126–142.
- OpenAI. 2024a. [Gpt-4o system card](#). *arXiv preprint arXiv:2410.21276*. Revision v1.
- OpenAI. 2024b. [Openai partners with condé nast](#). <https://openai.com/index/conde-nast/>. Accessed: 2025-02-21.

- OpenAI. 2025. [Openai o3 and o4-mini system card](#). System card, OpenAI.
- Guilherme Penedo, Hynek Kydlíček, Loubna Ben al-lal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. [The fineweb datasets: Decanting the web for the finest text data at scale](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 30811–30849. Curran Associates, Inc.
- 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, Jiayi Yang, Jingren Zhou, and 25 others. 2025. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Graeme Ritchie. 2001. [Current directions in computational humour](#). *Artificial Intelligence Review*, 16:119–135.
- Andrea C. Samson. 2012. [The influence of empathizing and systemizing on humor processing: Theory of mind and humor](#). *Humor*, 25(1):75–98.
- Wenbo Shang, Yuxi Sun, Jing Ma, and Xin Huang. 2026. [On the wings of imagination: Conflicting script-based multi-role framework for humor caption generation](#). In *The Fourteenth International Conference on Learning Representations*.
- Jerry M. Suls. 1972. [A two-stage model for the appreciation of jokes and cartoons: An information-processing analysis](#). In Jeffrey Goldstein and Paul E. McGhee, editors, *The psychology of humor: Theoretical perspectives and empirical issues*, pages 81–100. Academic Press.
- 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. 2025. [LlamaV-o1: Rethinking step-by-step visual reasoning in LLMs](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24290–24315, Vienna, Austria. Association for Computational Linguistics.
- Sean Trott, Drew E. Walker, Samuel M. Taylor, and Seana Coulson. 2025. [Turing jest: Distributional semantics and one-line jokes](#). *Cognitive Science*, 49(5):e70066.
- V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihang Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, and 74 others. 2026. [Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning](#). *Preprint*, arXiv:2507.01006.
- Robert West and Eric Horvitz. 2019. [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.
- Lawrence Wood. 2024. *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.
- Tong Xiao, Xin Xu, Zhenya Huang, Hongyu Gao, Quan Liu, Qi Liu, and Enhong Chen. 2026. [Perception-r1: Advancing multimodal reasoning capabilities of MLLMs via visual perception reward](#). In *The Fourteenth International Conference on Learning Representations*.
- Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. 2025. [LLaVA-CoT: let vision language models reason step-by-step](#). In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2087–2098.
- Yixin Yang, Zheng Li, Qingxiu Dong, Heming Xia, and Zhifang Sui. 2024. [Can large multimodal models uncover deep semantics behind images?](#) In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1898–1912, Bangkok, Thailand. Association for Computational Linguistics.
- Yufei Zhan, Yousong Zhu, Hongyin Zhao, Fan Yang, Shurong Zheng, Ming Tang, and Jinqiao Wang. 2026. [Vision-r1: Evolving human-free alignment in large vision-language models via vision-guided reinforcement learning](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pages 5807–5817.
- 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. 2024. [Humor in ai: Massive scale crowd-sourced preferences and benchmarks for cartoon captioning](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 125264–125286. Curran Associates, Inc.
- Yaowei Zheng, Juntao Lu, Shenzhi Wang, Zhangchi Feng, Dongdong Kuang, Yuwen Xiong, and Richong Zhang. 2025. [EasyR1: An efficient, scalable, multi-modality rl training framework](#). <https://github.com/hiyouga/EasyR1>.
- Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyang Luo, Zhangchi Feng, and Yongqiang Ma. 2024. [Llamafactory: Unified efficient fine-tuning of 100+ language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410, Bangkok, Thailand. Association for Computational Linguistics.
- Shanshan Zhong, Zhongzhan Huang, Shanghua Gao, Wushao Wen, Liang Lin, Marinka Zitnik, and Pan

Zhou. 2024. 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)*, pages 13246–13257.

Kuan Lok Zhou, Jiayi Chen, Siddharth Suresh, Reuben Narad, Timothy T. Rogers, Lalit K Jain, Robert D Nowak, Bob Mankoff, and Jifan Zhang. 2025. Bridging the creativity understanding gap: Small-scale human alignment enables expert-level humor ranking in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 16273–16287, Suzhou, China. Association for Computational Linguistics.

Appendix

This appendix provides extra 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, statistical, qualitative, and quantitative analyses, as well as additional experiments that further validate our findings.

The content is organized as follows:

- **Section A – Dataset and Benchmark Statistics:** A consolidated summary of the data used at every training stage and evaluation split, reporting the number of examples, task composition, and unique cartoons for the IM, RM, and PA training sets, the in-domain NYCC matching and ranking benchmarks, and the zero-shot transfer benchmarks.
- **Section B – 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 C – Resolution Modeling Corpora:** A detailed description of the Resolution Modeling (RM) dataset, including its composition, sources, human-in-the-loop pipeline, cross-family trace rewriting, quality-filtering, and annotator qualifications. This section provides insight into how captionist reasoning traces are generated to support resolution modeling.
- **Section D – Reward Composition and Weight Allocation:** An analysis of the reward components used for Preference Alignment (PA), including ablations of accuracy, format, perception, and style rewards and comparisons of alternative weight configurations. This section provides insight into the contribution of each reward and the basis for the final reward allocation.
- **Section E – 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 F – 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 G – Inference Configurations:** Inference configurations for all models used in evaluation, including model identifiers, access methods, and decoding configurations. API-accessed models use provider defaults, while configurations for locally hosted models are reported where available.
- **Section H – Statistical Analysis:** Bootstrap confidence intervals for model performance across evaluation settings.
- **Section I – 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. They provide a qualitative view of how structured reasoning evolves. Additional qualitative examples illustrating the limitations discussed in Section 6 are also provided.
- **Section J – Quantitative Error Taxonomy:** A breakdown of IRS-72B model errors across the benchmark splits, with seven error categories and their frequencies reported separately for Matching and Ranking tasks.
- **Section K – Reasoning-Chain Quality Audit:** A structured audit of reasoning-chain quality using a shared set of humor and reasoning-specific evaluation dimensions, assessing evaluator consistency across independent model judges and comparing IRS against zero-shot and few-shot CoT baselines, with full audit prompts and metric definitions.
- **Section L – 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.
- **Section M – User Study Details:** Documentation of the user study, including participant instructions, consent, recruitment, and how the expert and non-expert references were collected.

A Dataset and Benchmark Statistics

Table A.1 consolidates the data volumes used across all training stages and evaluation splits, reporting for each the number of examples, its task composition, and the number of unique cartoons it covers. We report unique-cartoon counts alongside example counts because several benchmark questions or caption comparisons may derive from the same underlying cartoon; separating the two makes the effective visual diversity of each split explicit.

The humor-specific training data consists of NYCC-derived sources, supplemented with general-purpose corpora for broad linguistic and world-knowledge coverage. The Incongruity Modeling (IM) corpus consists of 4,123 documents (6.29M tokens) of humor-specific and general text used for continual pretraining (its composition and temporal coverage are detailed in Appendix B). Resolution Modeling (RM) and Preference Alignment (PA) share the same pool of 5,163 samples (2,208 matching and 2,955 ranking), spanning 736 unique cartoons.

In-domain evaluation uses the full established test splits rather than subsampled subsets: the NYCC test set of 685 examples (300 matching and 385 ranking) over 100 unique cartoons from Hessel et al. (2023), together with the harder 10-vs-1000 and 30-vs-300 ranking splits, similar to those from Zhou et al. (2025) (350 comparisons each, over 35 cartoons). The zero-shot transfer benchmarks contribute no training data: IRS is trained only on the NYCC-derived data above and evaluated directly on YesBut (Hu et al., 2024) (1,044 Philosophy and 1,044 Title queries over 348 image pairs) and the DeepEval humor subset (Yang et al., 2024) (29 images). To prevent leakage, all evaluation cartoons are excluded from the continual pretraining corpus, and we verify negligible n -gram overlap between the CPT corpus and the evaluation caption text.

B 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 B.1 summarizes the main sources, which fall into three categories:

- **Contest deliberations & roundtables (audio/video)** capture conversational reasoning pro-

cesses 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.

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 B.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).

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.

Table A.1: **Dataset and evaluation benchmark statistics.** Summary of data volume, task composition, and unique cartoons used across all training stages and evaluation splits.

Component / Split	Purpose	Size	Task composition	Unique Cartoons
Training Datasets				
IM corpus	Continual pretraining	4,123 docs; 6.29M tokens	Humor-specific and general text	–
RM training	Reasoning-trace SFT	5,163 traces	2,208 Matching; 2,955 Ranking	736
PA training	GRPO alignment	5,163 samples	2,208 Matching; 2,955 Ranking	736
In-Domain Evaluation				
NYCC test	In-domain evaluation	685 examples	300 Matching; 385 Ranking	100
10-vs-1000	Hard ranking evaluation	350 comparisons	Ranking	35
30-vs-300	Fine-grained ranking evaluation	350 comparisons	Ranking	35
Zero-Shot Transfer Evaluation				
YesBut	Zero-shot transfer	1,044 Philosophy 1,044 Title	Underlying philosophy; Title	348
DeepEval humor subset	Zero-shot transfer	29 images	DeepSemantics; Description; Title	29

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.

C Resolution Modeling Corpora

To construct high-quality, reliable reasoning traces without relying on unverified model outputs or same-family self-audits, we implement a multi-stage human-in-the-loop curation and verification pipeline.

Corpus Composition. Rather than directly adopting raw outputs from DeepSeek-R1 (DeepSeek-AI et al., 2025), reasoning chains undergo rigorous annotation, filtering, and cross-family refinement:

1. **Visual and Contextual Annotation Review:** Annotators manually inspect and correct the base NYCC scene descriptions, entity tags, and incongruity annotations to resolve ambi-

guities or missing visual details.

2. **Visual Grounding Verification:** Automatically expanded descriptions are audited to eliminate hallucinations and retain only details explicitly grounded in the image.
3. **Initial Trace Generation:** DeepSeek-R1 generates preliminary reasoning traces conditioned on candidate captions and the human-verified visual metadata.
4. **Automated Heuristic Screening:** Traces are screened via automated checks to filter out answer leakage, shortcut cues, visual inconsistencies, and deviations from prescribed reasoning structures.
5. **Targeted Human Correction:** Flagged traces undergo manual inspection and are corrected, regenerated, or rejected. Revised traces are re-screened prior to inclusion.
6. **Cross-Family Rewriting:** To prevent family-specific bias, GPT-4o (OpenAI, 2024a) rewrites accepted traces into concise, image-grounded captionist-style prose. A subset of rewritten traces is manually reviewed to verify reasoning preservation.

Table B.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%

Table B.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

Table D.1: **Reward composition and weight allocation.** Performance under different reward components and weight configurations. Reward configurations are reported as (R_a, R_p, R_s, R_f) , corresponding to accuracy, perception, style, and format rewards, respectively.

Reward Configuration	Configuration	(Hessel et al., 2023)		(Zhou et al., 2025)	
		Matching	Ranking	10-vs-1000	30-vs-300
(0.8, 0, 0, 0.2)	Accuracy + Format	60.67	57.99	54.57	50.86
(0.4, 0.4, 0, 0.2)	+ Perception Reward	58.00	60.78	56.57	49.43
(0.4, 0, 0.4, 0.2)	+ Style Reward	60.00	58.44	54.57	49.43
(0.3, 0.3, 0.3, 0.1)	Equal Reward Weights	61.00	59.74	54.57	50.29
(0.4, 0.2, 0.2, 0.2)	Final	59.67	64.42	56.29	53.14

7. Final Quality Pass: A final automated audit validates formatting compliance, reasoning coherence, and label consistency across the dataset.

Annotation Team and Quality Control. The annotation and verification process involved three

annotators: one graduate and two undergraduate researchers with backgrounds in computer science and prior experience with multimodal language models and NYCC-style humor data. All annotators were proficient in English. Uncertain cases were discussed and resolved jointly.

D Reward Composition and Weight Allocation

Table D.1 reports the reward-component ablation used to determine the final GRPO reward composition. The baseline combines the accuracy and format rewards. Adding the perception reward produces the largest individual improvement, increasing Ranking accuracy from 57.99 to 60.78 (+2.79) and 10-vs-1000 accuracy from 54.57 to 56.57 (+2.00). Adding the style reward alone yields a smaller improvement on Ranking (+0.45), indicating that the two auxiliary rewards provide complementary supervision.

We further compare two configurations that combine all reward components. Equal weighting of the three rewards (0.3, 0.3, 0.3, 0.1) achieves the highest Matching accuracy but underperforms on Ranking and both larger candidate-set evaluations. The final configuration (0.4, 0.2, 0.2, 0.2) yields the best overall trade-off, improving Ranking by 4.68 points, 10-vs-1000 by 1.72 points, and 30-vs-300 by 2.85 points over the equal-weight configuration, while incurring a 1.33-point decrease in Matching accuracy. We therefore assign the largest weight to the accuracy reward, while equally weighting the perception and style rewards. This allocation prioritizes answer correctness while retaining complementary supervision for visual incongruity and captionist-style reasoning.

E 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 E.1 summarizes hyperparameters across the stages.

Incongruity Modeling (IM). The goal of IM is to adapt the backbone model to humor-relevant discourse. We train Qwen2.5-VL-7B-Instruct (Bai et al., 2025) 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. B. 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. E.1-E.3 show loss, learning rate, and gradient norm dynamics of our IM models.

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 train using LLaMA-Factory (Zheng et al., 2024) with LoRA (rank 64 for 7B, rank 8 for 32B and 72B), 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 ~ 3 hours on 2xH100 GPUs, while the 32B model requires ~ 6 hours on 4xH100 GPUs. Fig. E.4-E.6 show training dynamics. Loss decreases smoothly after warmup, with stable gradients, indicating effective learning of structured reasoning patterns.

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 4xH100 GPUs (8 for the 32B model, and 16 for 72B model). Fig. E.7-E.9 show reward dynamics. Perception and style rewards saturate early, while accuracy and format rewards continue improving.

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

F.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

Table E.1: **Training hyperparameters across IRS (7B) stages.** Each stage supports a different component of IRS, with distinct optimization regimes and resource requirements.

Stage	Base Model	Adapter	Epochs	Batch	LR	Precision	GPUs	Duration
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

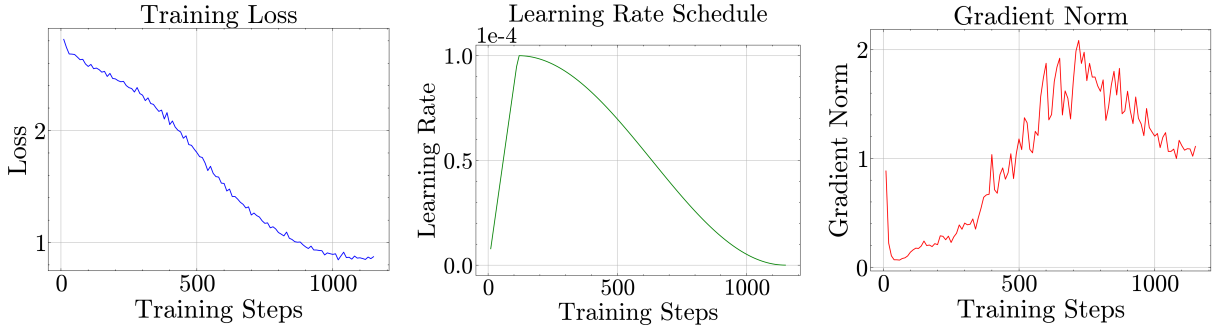


Figure E.1: **IM dynamics of the 7B model.** Loss decreases steadily over training, indicating successful adaptation to humor-specific discourse. The gradient norm remains controlled overall, with larger spikes appearing later in training.

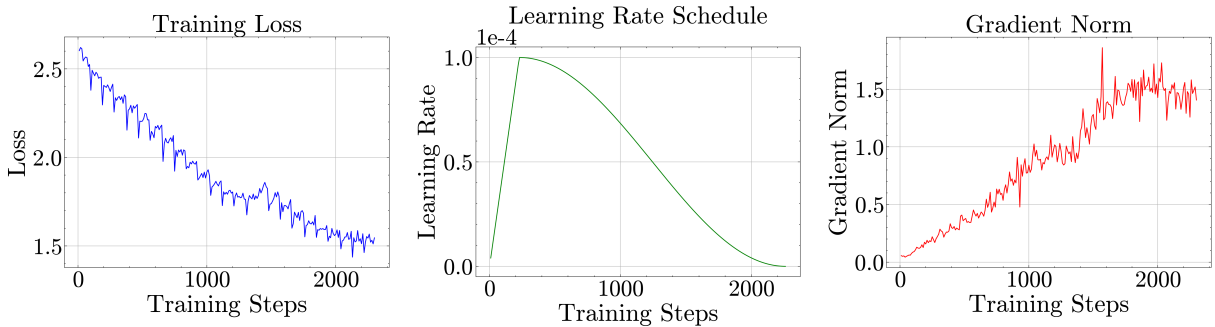


Figure E.2: **IM dynamics of the 32B model.** Loss decreases steadily despite mild early fluctuations, indicating successful adaptation to humor-specific discourse. The gradient norm increases gradually over training while remaining stable overall.

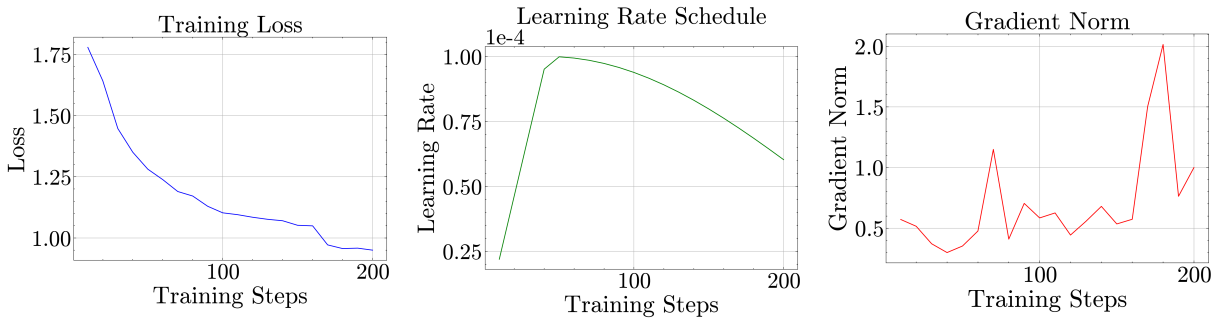


Figure E.3: **IM dynamics of the 72B model.** Loss drops rapidly and then stabilizes, indicating effective early adaptation to humor-specific discourse. The gradient norm remains mostly low but exhibits occasional sharp spikes.

grounding for interpretation.² We use DeepSeek-R1 (DeepSeek-AI et al., 2025) as a teacher model, prompting it to reconstruct the scene, identify

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

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.

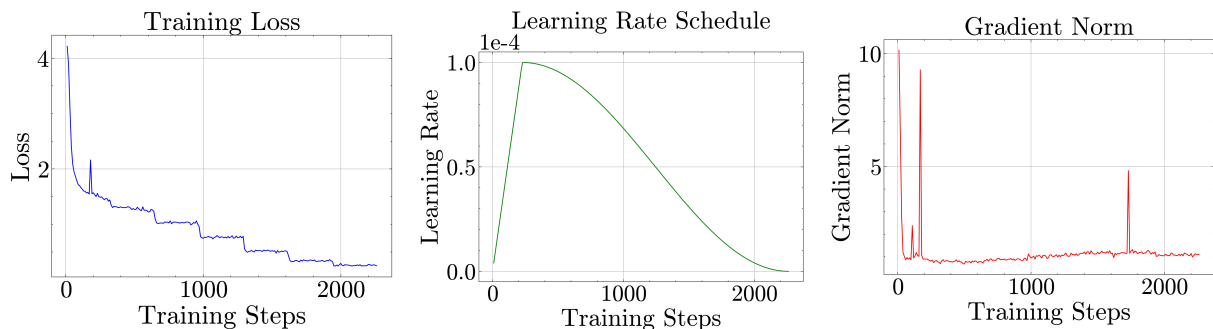


Figure E.4: **RM dynamics of the 7B model.** Rapid loss reduction during warmup followed by stable convergence indicates effective learning of structured reasoning from captionist traces.

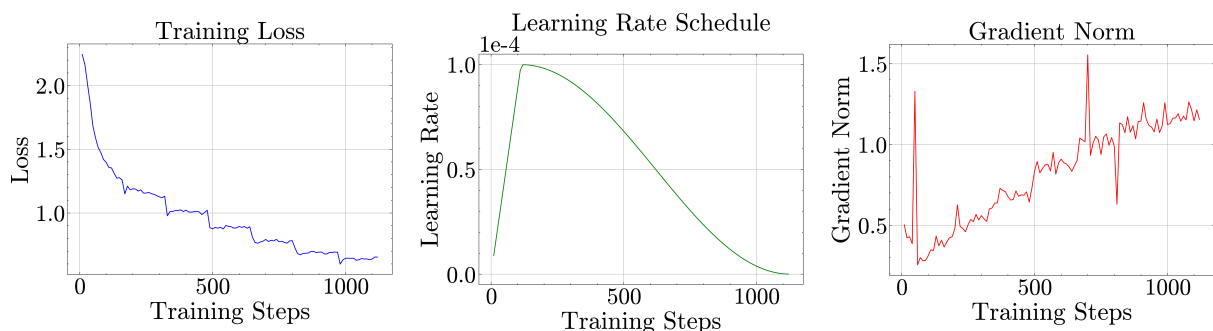


Figure E.5: **RM dynamics of the 32B model.** Loss decreases smoothly after warmup, indicating effective adaptation to structured reasoning supervision. The gradual increase in gradient norm suggests continued refinement of reasoning patterns during training.

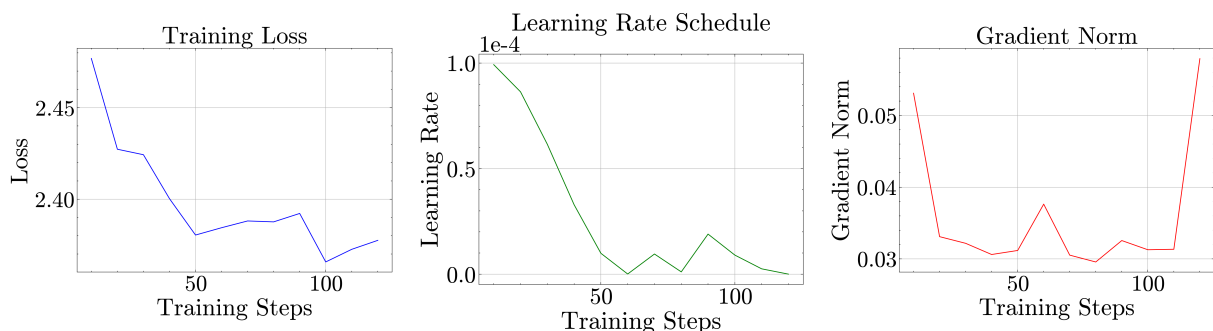


Figure E.6: **RM dynamics of the 72B model.** Loss decreases more gradually with limited training steps, indicating early-stage adaptation to reasoning supervision. The gradient norm remains low overall with occasional spikes, suggesting stable but less extensive optimization.

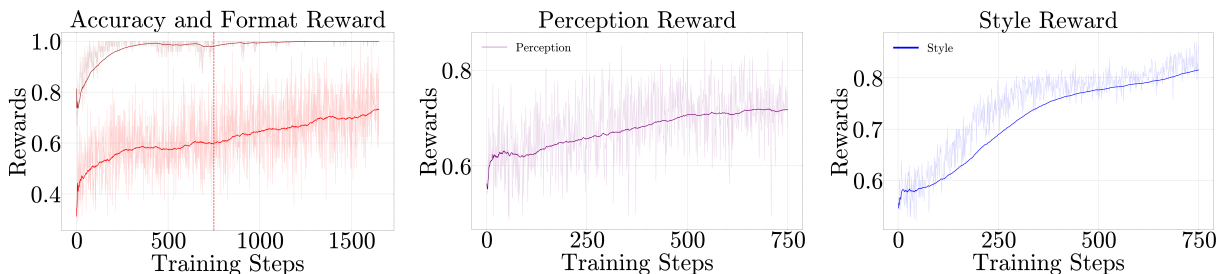


Figure E.7: **PA dynamics of the 7B model.** Perception and style rewards increase rapidly and saturate early, while accuracy and format rewards continue improving, indicating progressive alignment with human judgment and reasoning quality.

The resulting traces are rephrased using GPT-4o (OpenAI, 2024a) to match the style of professional captionist commentary, emphasizing concise, observational, and image-grounded lan-



Figure E.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.

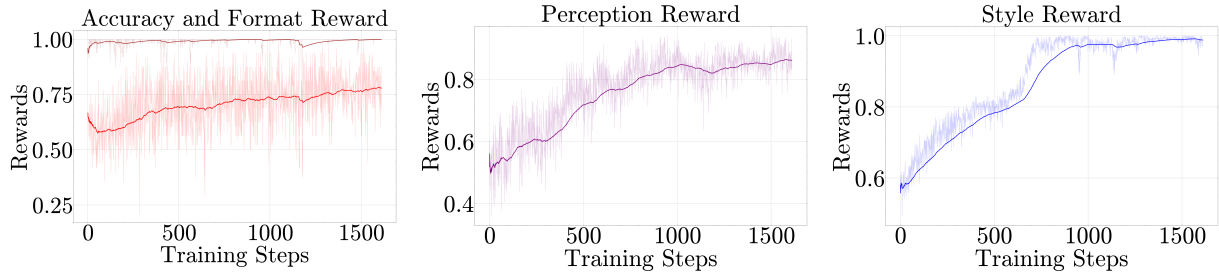


Figure E.9: **PA dynamics of the 72B model.** Perception and style rewards increase rapidly and approach saturation, while accuracy and format rewards continue improving, indicating strong alignment with human judgment and well-calibrated reasoning quality at scale.

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 F.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.

guage. 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. F.1 and F.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. F.3 shows the rephrasing template, used to refine traces into captionist-style commentary.

F.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 (Qwen et al., 2025) 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. F.4 shows the exact prompt.

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. F.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.

F.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 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. F.6 and Fig. F.7 show the full templates.

F.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. F.8 and Fig. F.9 show the full templates.

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 F.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 F.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.

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.

G Inference Configurations

Table G.1 summarizes the models used for evaluation together with their access method and inference settings. We use the default inference settings for models accessed through APIs, while the decoding settings used for locally hosted models are reported in the table. These include temperature and maximum generation length where applicable. All reported results are obtained from a single evaluation run for each model.

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

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 F.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.

H Statistical Analysis

To evaluate model performance, we report accuracy along with 95% confidence intervals (CI) and standard errors. Confidence intervals were calculated using non-parametric percentile bootstrapping with $B = 10,000$ resamples drawn with replacement. For each dataset split of size n , B independent bootstrap resamples were drawn, and sample accuracy was computed per resample. The 95% confidence interval was defined by the 2.5th and 97.5th percentiles of the empirical bootstrap distribution, and standard error was computed as the standard deviation of the resampled means.

Table H.1 presents the comparative evaluation of

IRS-7B, IRS-32B, and IRS-72B across all benchmark splits. IRS-72B improves over IRS-7B by 9.66 and 11.68 percentage points on matching and ranking, respectively; the ranking improvement is accompanied by non-overlapping 95% bootstrap confidence intervals.

A clear performance drop is observed as split difficulty increases. On 30-vs-300 Ranking split, all three models operate near random-chance accuracy (50%), indicating that subtle visual-textual incongruities in this subset remain challenging even for the 72B parameter model.

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

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 F.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.

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 F.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.

Table G.1: **Inference configurations.** Model identifiers, access methods, and inference configurations used for evaluation. API-accessed models use the providers’ default inference settings.

Model	Model Identifier	Access	Inference Configuration
GPT-4o	gpt-4o	OpenAI API	Default
o4-mini	o4-mini	OpenAI API	Default
o3	o3	OpenAI API	Default
DeepSeek-R1	deepseek-reasoner	DeepSeek API	Default
Gemini 2.5 Flash	gemini-2.5-flash	Gemini API	Default
Qwen2.5-VL	...	Local	$T = 0.5$,
{7B,32B,72B}-Instruct	...	Local	Max new tokens = 1024
GLM-4.1V-9B-Think	...	Local	$T = \text{Model Default}$,
			Max new tokens = 2048
Kimi-VL-A3B-Think	...	Local	$T = 0.5$,
			Max new tokens = 1024
LlamaV-o1	...	Local	$T = \text{Model Default}$,
			Max new tokens = 2048
NVLM-D-72B	...	Local	$T = 0.5$,
			Max new tokens = Model Default

I 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.

I.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. I.1, the trace demonstrates how captionists compare alternatives by identifying which caption better exploits the visual incongruity and delivers a more ef-

fective punchline. These traces form the structured supervision used during Resolution Modeling.

I.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. I.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 con-

Table H.1: **Performance with 95% bootstrap confidence intervals.** Accuracy and standard error across the in-domain evaluation splits.

Model	Task	n	Accuracy (%) [95% CI]	Standard Error
IRS-7B	Matching	300	59.67 [54.00, 65.33]	0.0288
	Ranking	385	64.42 [59.48, 69.09]	0.0244
	Ranking (10-vs-1000)	350	56.29 [51.14, 61.43]	0.0264
	Ranking (30-vs-300)	350	53.14 [47.99, 58.29]	0.0268
IRS-32B	Matching	300	62.67 [57.33, 68.33]	0.028
	Ranking	385	68.05 [63.38, 72.47]	0.0238
	Ranking (10-vs-1000)	350	62.86 [57.71, 68.00]	0.0258
	Ranking (30-vs-300)	350	53.14 [47.71, 58.29]	0.0268
IRS-72B	Matching	300	69.33 [64.00, 74.33]	0.0267
	Ranking	385	76.10 [71.69, 80.26]	0.0217
	Ranking (10-vs-1000)	350	62.57 [57.43, 67.71]	0.0260
	Ranking (30-vs-300)	350	50.86 [45.43, 56.00]	0.0269

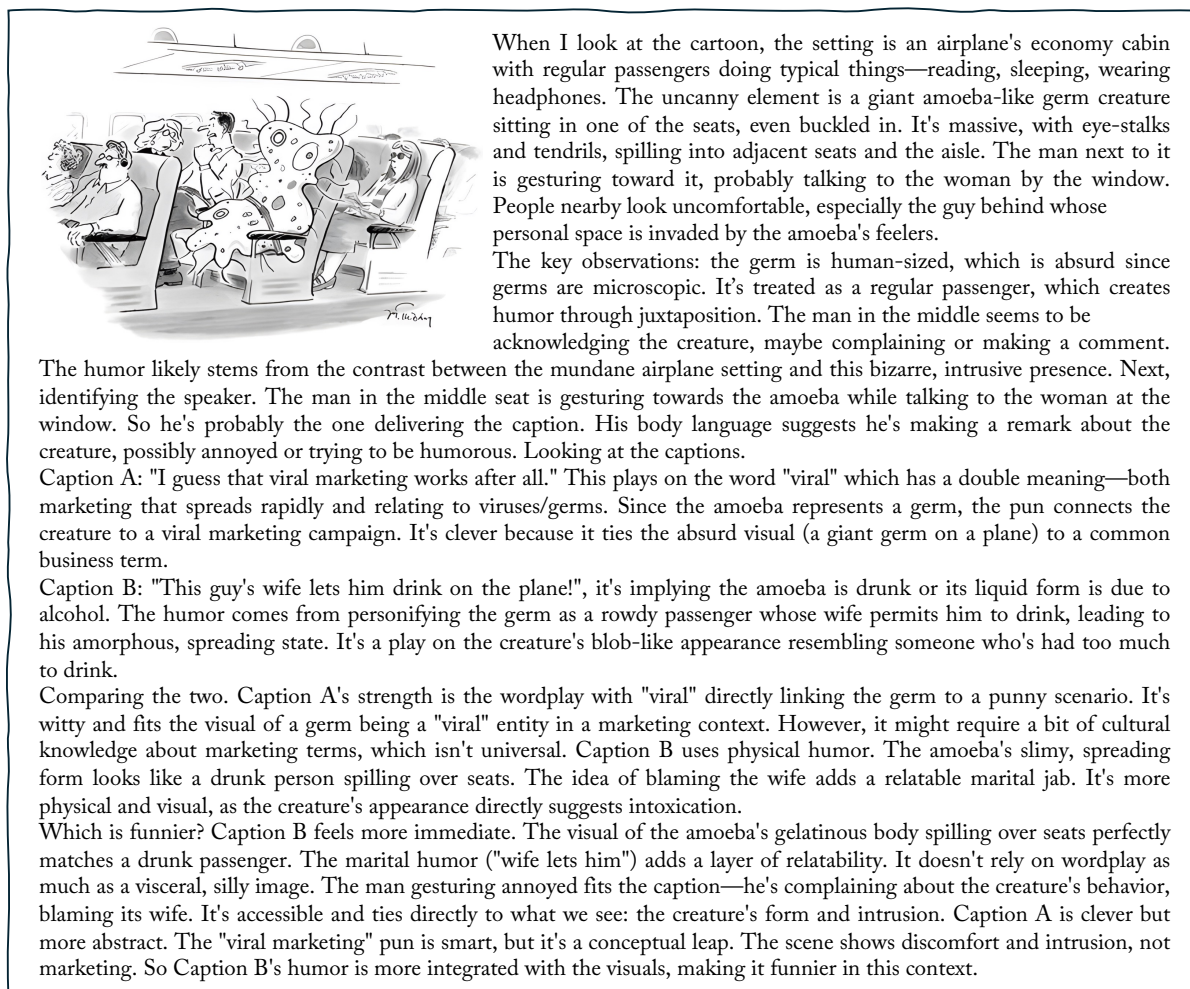
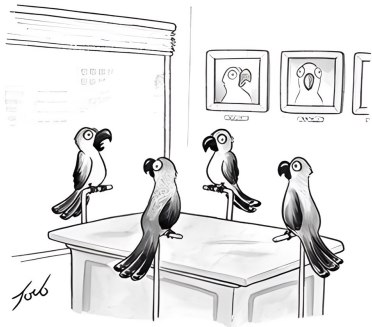


Figure I.1: **Example reasoning trace for a caption ranking task.** The trace compares two captions against the visual setup (a giant amoeba crowding an airplane seat) and explains why Caption B ("This guy's wife lets him drink on the plane!") better resolves the incongruity and captures the visual humor.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



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 I.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.

trasts. The output reads closer to a captionist’s commentary: concise, well-structured, and attuned to comedic effect.

Fig. I.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.

I.3 Judge Responses

Fig. I.4 and I.5 illustrate the outputs of the LLM-as-judge components used to compute perception and style rewards. In Fig. I.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. I.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.

I.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. I.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.

I.5 Discussion

The examples in Appendix I 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.

I.6 Limitations

We provide two qualitative examples (Figure I.7 and Figure I.8) highlighting limitations of IRS and discuss the corresponding failure cases in the Section 6.

J Quantitative Error Taxonomy

To better understand failure cases, we analyze the incorrect predictions of IRS-72B across all benchmark splits. Each error is assigned to one of seven predefined categories by Gemini-2.5-Flash (Comanici et al., 2025). Table J.1 summarizes the resulting error distribution. Errors are classified into seven failure categories:

- **Visual Misperception & Hallucination:** Misidentifying physical objects, characters, or background elements, or hallucinating non-existent scene components.
- **Visual Incongruity Blindness:** Correctly identifying individual objects while failing to recognize the central uncanny visual anomaly or contradiction.
- **Humor Mechanism & Pun Misinterpretation:** Failing to resolve idioms, metaphors, or wordplay in candidate captions.
- **Speaker Misdetection:** Failing to determine which character in the cartoon is speaking or receiving the dialogue line, leading to an inverted perspective on the joke.
- **Overthinking & Post-Hoc Rationale Overfit:** Constructing an overly complex, post-hoc philosophical narrative to justify an incorrect choice.
- **Option Discrimination Failure:** Selecting a plausible caption without properly contrasting it against the funnier, visually-anchored punch-

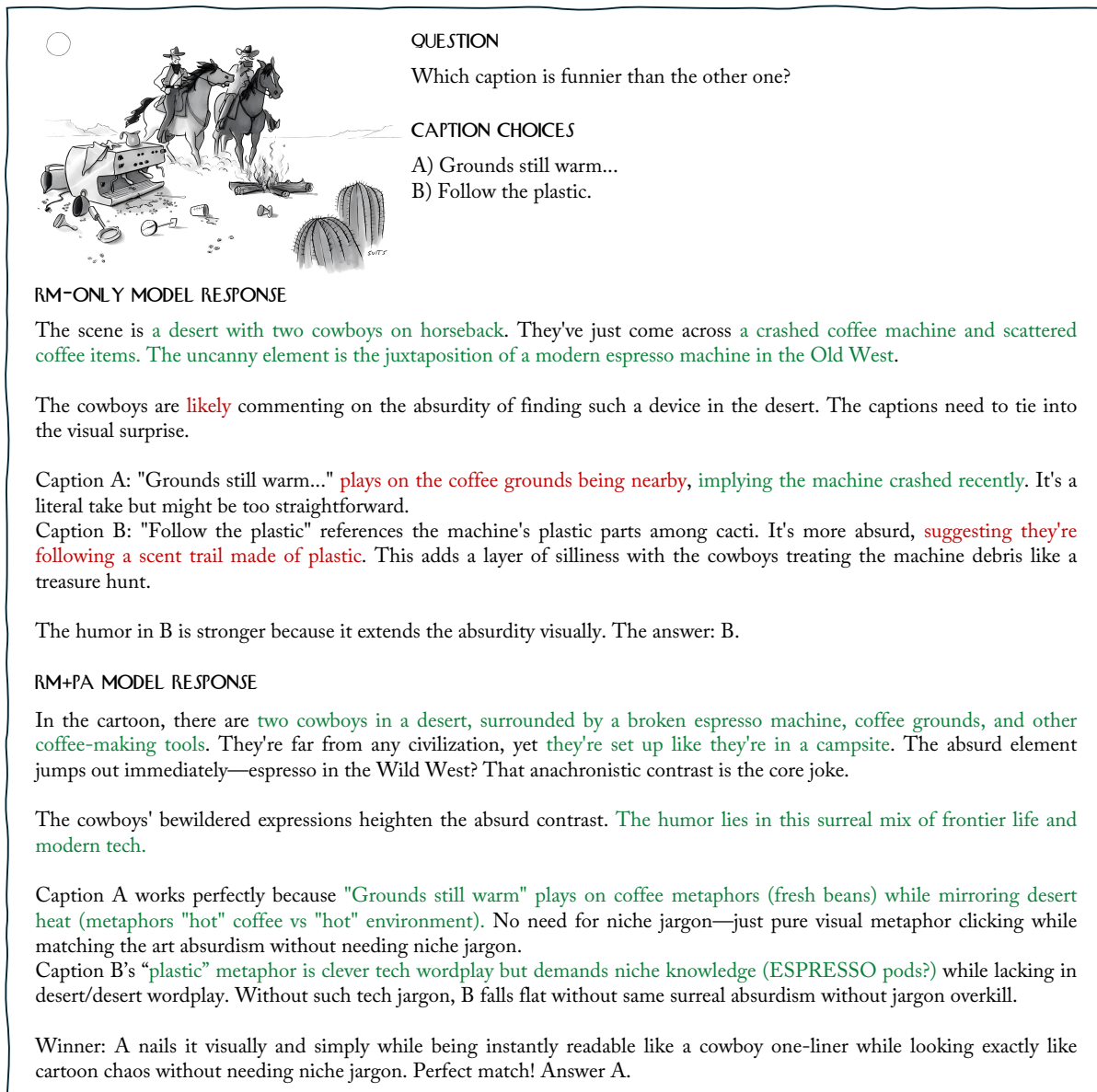


Figure I.3: **Example reasoning on a caption ranking task.** The RM-only model fixates on literal associations (e.g., coffee grounds as evidence of recency), while the RM+PA model integrates the broader absurdity of encountering a modern espresso machine in a cowboy setting. By grounding its reasoning 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.

line.

- **Cultural, Domain, or Anachronism Blindness:** Failing to recognize pop-culture references, historical contexts, brand names, or specific tropes.

In the Matching task, visual misperception and hallucination account for 81 of 92 errors. These errors often involve incorrectly perceived objects, attributes, or other visual details, which can cause the model to select a distractor caption. In contrast, humor interpretation errors are more common in the ranking task. Since both candidate captions are related to the visual content, these errors are

more closely related to distinguishing the intended humorous interpretation.

Cultural and domain knowledge errors account for 71 errors overall, relatively more errors in the Ranking splits. These cases involve captions that depend on information beyond the visual content, such as references to historical figures, cultural concepts, or brands. As the difficulty level increases in the Ranking task, overthinking errors become more frequent. These cases are characterized by increasingly elaborate reasoning that leads to an incorrect caption choice. The remaining categories occur

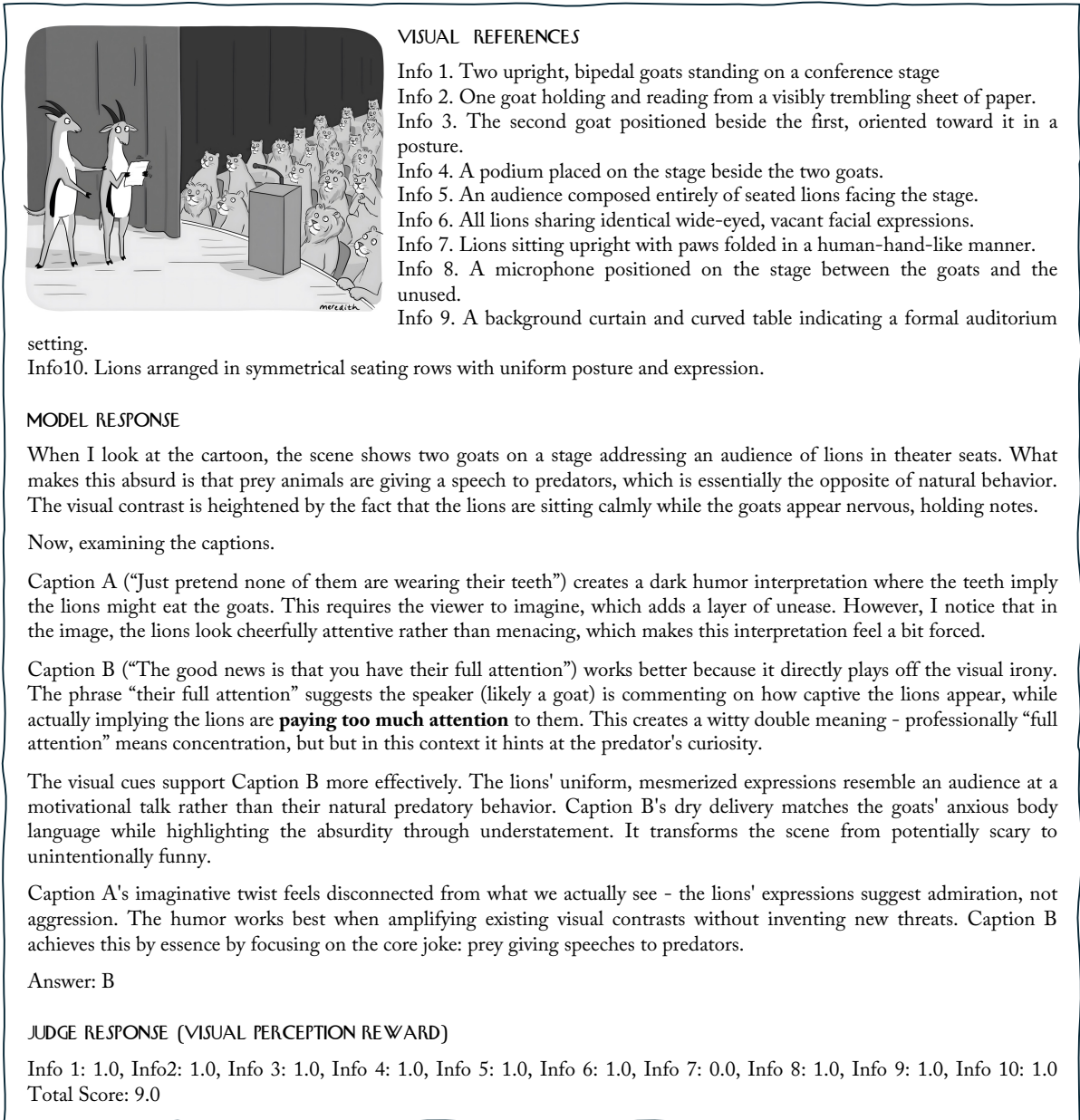


Figure I.4: **Visual perception judge response.** Binary scores indicating whether the reasoning reflects curated visual references, providing a direct measure of grounding in salient scene elements.

Table J.1: **Error Taxonomy Across Matching and Ranking tasks for IRS-72B** ($N = 487$ errors). Columns report raw error counts separated by task types.

Error Category	(Hessel et al., 2023)		(Zhou et al., 2025)	
	Matching	Ranking	10-vs-1000	30-vs-300
Humor Mechanism & Pun Misinterpretation	0	45	49	73
Visual Misperception & Hallucination	81	18	34	19
Cultural / Domain / Anachronism Blindness	3	20	24	24
Overthinking & Post-Hoc Rationale Overfit	5	7	20	28
Option Discrimination Failure	1	1	2	14
Visual Incongruity Blindness	1	1	2	13
Speaker Misdetection	1	0	0	1

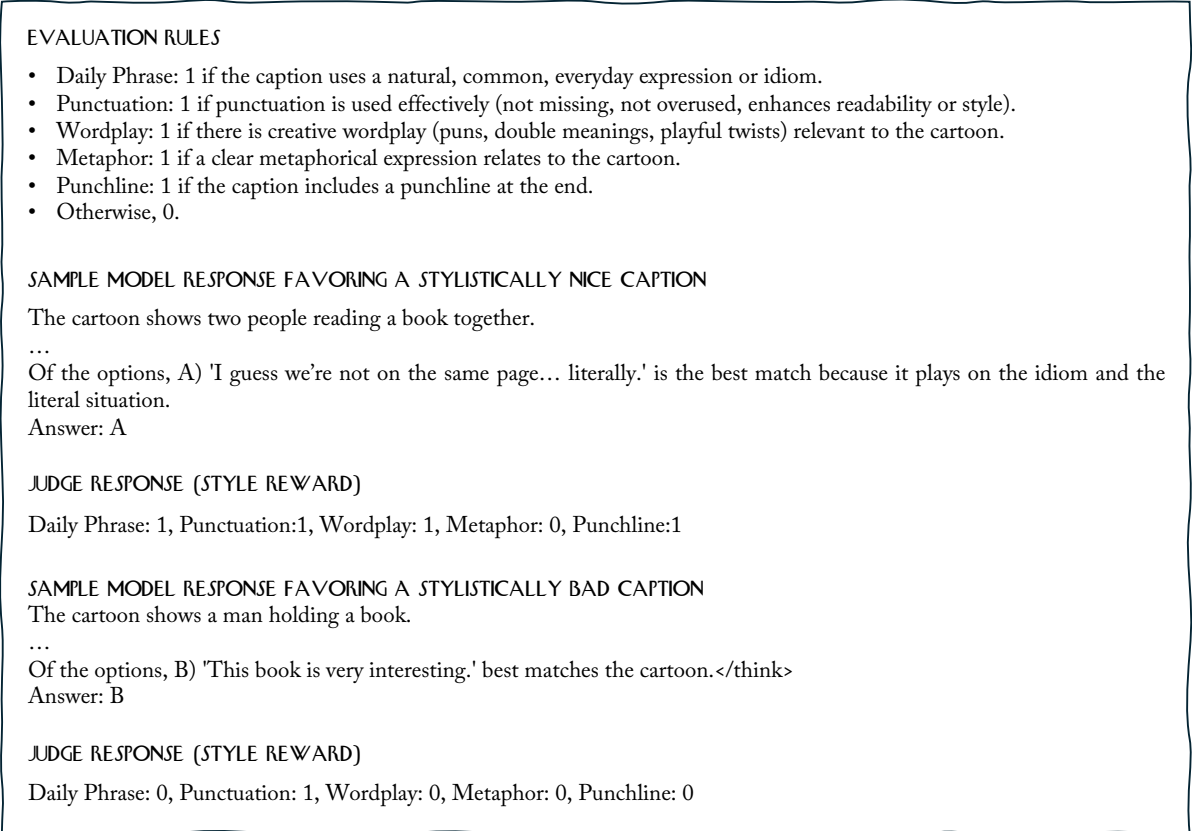


Figure I.5: **Style judge response.** Binary scores across five stylistic dimensions (natural phrasing, punctuation, wordplay, metaphor, and punchline placement), capturing linguistic qualities associated with effective caption writing.

much less frequently and increase as the ranking questions become harder.

Overall, the error distribution shows a clear difference between the two task types: Matching errors are dominated by visual grounding failures, whereas Ranking errors more often involve humor interpretation and comparative reasoning.

K Reasoning-Chain Quality Audit

To assess the quality of the resulting reasoning traces, we perform a structured reasoning-chain audit using DeepSeek (DeepSeek-AI et al., 2025), NVLM-D-72B (Dai et al., 2024), and Gemini-2.5-Flash (Comanici et al., 2025) as external evaluators. Rather than evaluating only answer correctness, the audit scores traces along several dimensions that are listed below.

Matching-task audit dimensions. For matching tasks, where the model selects one caption from five candidates, we evaluate the following dimensions:

- **Incongruity:** whether the reasoning identifies the central mismatch, tension, or absurdity that

drives the humor.

- **Humor Insight:** whether the explanation captures specific humor mechanisms such as wordplay, metaphor, reversal, or social convention.
- **Caption Alignment:** whether the reasoning clearly explains why the chosen caption matches the cartoon.
- **Option Discrimination:** whether the reasoning explains why the selected caption is better than the alternatives.

Ranking-task audit dimensions. For ranking tasks, where the model compares two captions, we evaluate:

- **Incongruity Alignment:** whether it identifies the primary source of humor.
- **Caption-Scene Fit:** whether the reasoning explains how each caption interacts with the visual elements.
- **Humor Sensitivity:** whether the reasoning captures subtle differences in timing, understatement, absurdity, or cultural framing.
- **Funniness Justification:** whether the rationale explains why one caption is funnier, not merely

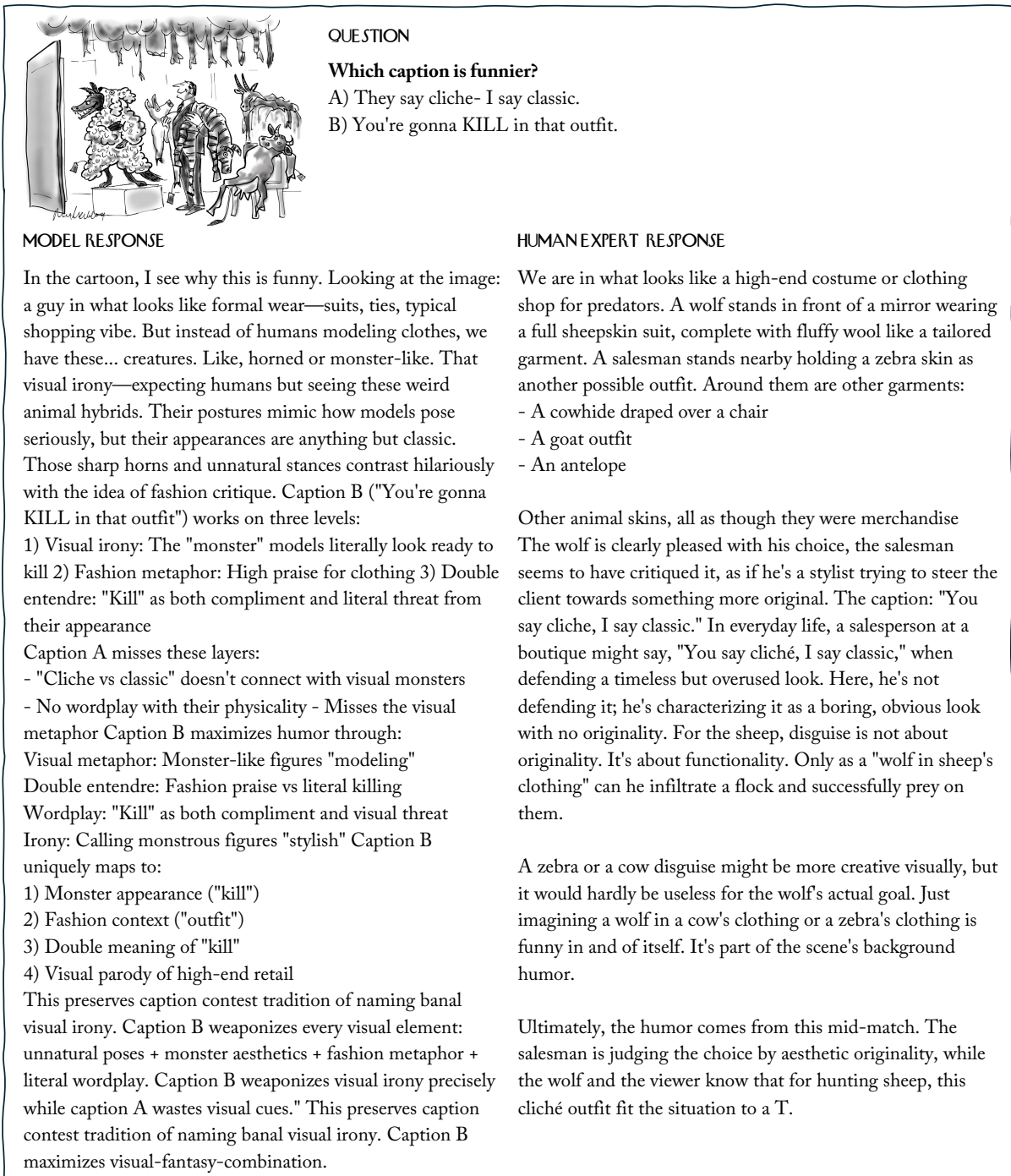


Figure I.6: **Model vs. expert reasoning.** The model and expert reach different conclusions but follow similar reasoning steps, indicating alignment in how the cartoon is analyzed even when preferences differ.

more coherent.

- **Comparative Evaluation:** whether the two captions are explicitly compared rather than discussed in isolation.

As shown in Table K.1, we conduct a cross-model quality audit using three independent judges from distinct model families, parameter scales, and training origins: DeepSeek-R1, NVLM-D-72B, and Gemini-2.5-Flash to eliminate potential self-

assessment bias inherent in evaluating reasoning traces using the model family that generated them. Absolute scores vary across judges due to differences in internal calibration applied by each model family to the evaluation rubric. Across both matching and ranking tasks, all three independent judges consistently assess the reasoning traces as high quality.

Crucially, the relative ordering across rubric di-

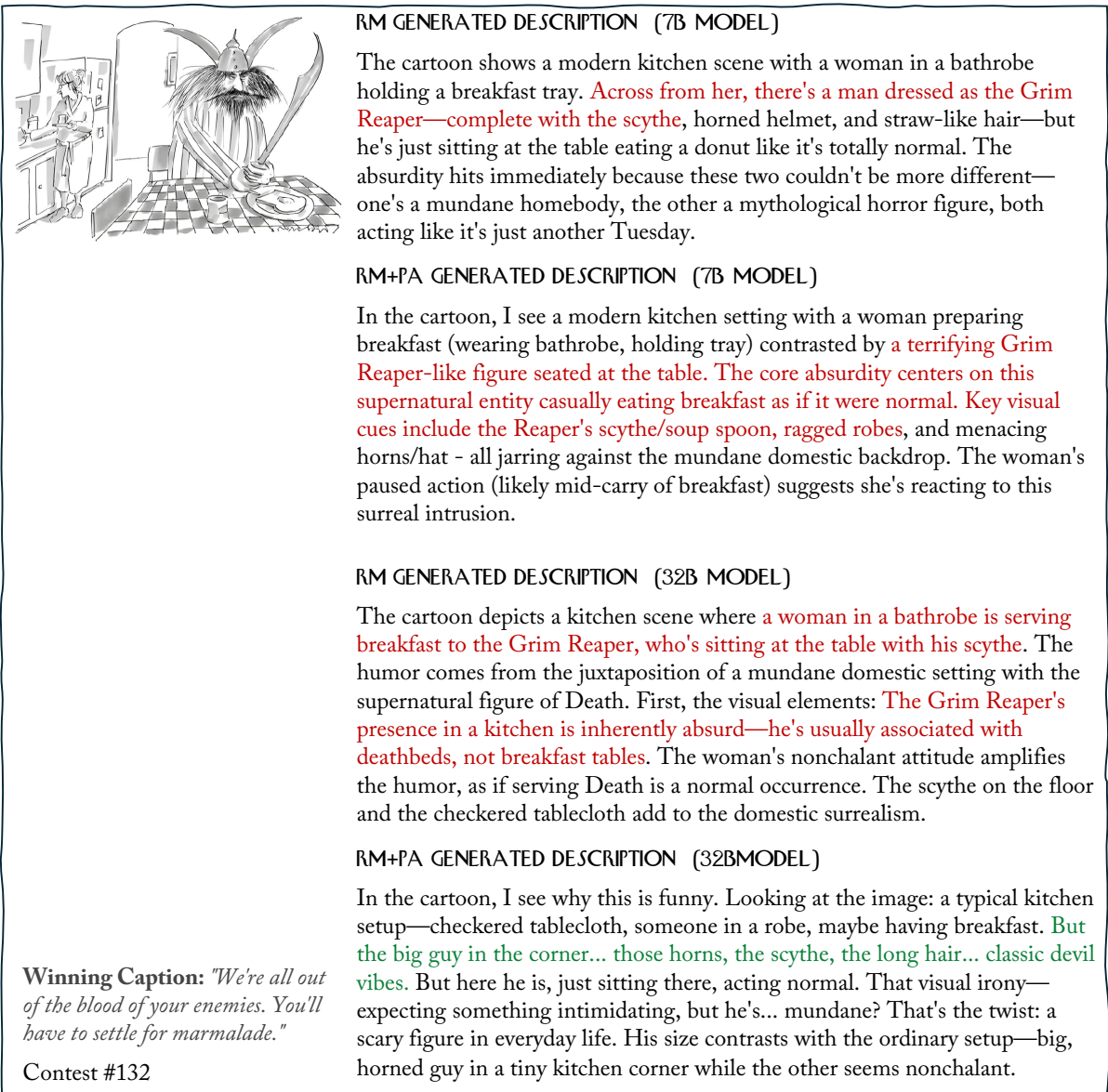


Figure I.7: **Limitation: Visual perception errors.** All 7B-scale variants misidentify the foreground figure as the Grim Reaper (with scythe), though the humor relies on the incongruity of a Viking warrior in a domestic kitchen; this error cascades into mismatched reasoning and an incorrect caption. At 32B, the model misidentifies objects early in training but, with the two additional PA rewards, corrects this and identifies the objects correctly.

mensions is strictly preserved under every evaluator: higher-order visual-semantic metrics, such as *Incongruity* and *Incongruity Alignment*, consistently score highest, whereas discriminative metrics like *Option Discrimination* and *Funniness Justification* remain consistently lowest. This convergent ranking across three model families, despite calibration differences, is stronger evidence of trace quality than any single judge’s absolute scores, and indicates the audit reflects genuine properties of the traces rather than the idiosyncrasies of one evaluator.

To evaluate whether IRS enhances underlying reasoning mechanisms rather than merely optimiz-

ing final-answer accuracy, we complement our performance metrics with a component-level rationale quality assessment on inference-time rationales generated on the held-out test set. We compare IRS (32B) against its underlying base architecture (Qwen2.5-VL-32B) configured with zero-shot and few-shot CoT prompting. All rationales are independently assessed using Gemini-2.5-Flash.

Table K.2 summarizes the results. IRS achieves superior scores across all evaluated dimensions and overall averages in both matching and ranking tasks.

Overall, these results suggest that IRS improves not only answer accuracy but also the quality and

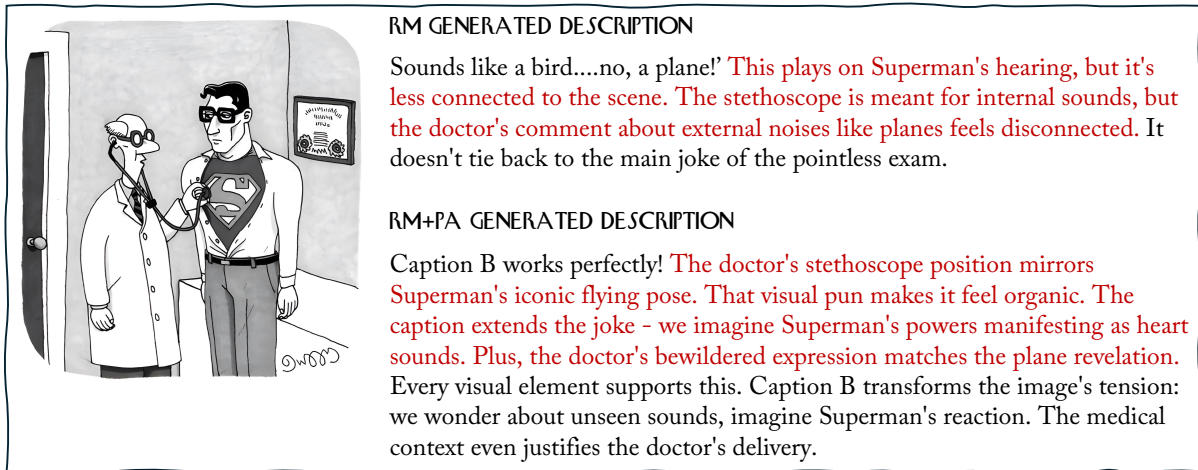


Figure I.8: **Limitation: Shallow cultural understanding.** Sample question from ranking task with the choices A) "SORRY, YOU DON'T QUALIFY FOR OBAMACARE.", B) "Sounds like a bird....no, a plane!". The humor depends on a Superman reference (the "it's a bird... it's a plane..." trope). The RM-only model fixates on superficial wordplay; RM+PA recognizes the pop-culture context but hallucinates a visual link and does not fully recover the cultural script.

structure of the generated reasoning itself. The strong incongruity-related scores are particularly consistent with the central objective of IRS: explicitly supervising incongruity-resolution reasoning rather than relying solely on implicit reasoning emerging from scale.

We provide full audit prompts below.

Audit prompts. The audit prompts instruct models to act as a skeptical expert reviewer rather than to regenerate a new interpretation. The evaluator is asked to identify weaknesses such as superficial humor analysis, and to return only a structured JSON object containing scores, flaws, strengths, reflection, revised answer, and revised rationale.

Matching Task Audit Prompt

You are a senior cartoon-caption analyst reviewing a model's reasoning trace for a New Yorker Caption Contest matching task. You are given:

- a cartoon scene description,
- the unusual or uncanny element,
- key observations,
- five candidate captions,

and a step-by-step reasoning chain produced by a model, and the model's selected answer.

Your job now is not to generate a new interpretation from scratch, but to critically audit the provided reasoning trace as an expert reviewer.

You must assess whether the model's reasoning:

1. identified the central incongruity that actually drives the humor,
2. explained why the selected caption matches better than the alternatives,
3. avoided superficial, generic, or unsupported interpretations.

Important:

- Do not simply restate the model's reasoning.
- Look for overreach, hallucinated visual assumptions, weak humor analysis, and poor discrimination among options.
- Be strict and skeptical.
- If the reasoning is weak, say so clearly.
- If needed, revise the final choice.

Input:

Scene: [SCENE]

Description: [DESCRIPTION]

Uncanny Element: [UNCANNY ELEMENT]

Observations: [OBSERVATIONS]

Candidate Captions:

A) [CAPTION A]

B) [CAPTION B]

C) [CAPTION C]

D) [CAPTION D]

E) [CAPTION E]

Model's Reasoning:

[REASONING CHAIN]

Model's Selected Answer:

[ANSWER]

Perform a professional audit along the following dimensions:

1. Incongruity Diagnosis: Did it identify the real mismatch, tension, or absurdity that makes the cartoon funny?
2. Caption-Image Alignment: Did it clearly explain why the chosen caption matches the cartoon?
3. Humor Interpretation: Did it explain the humor in a specific and non-trivial way, including metaphor, reversal, social convention, or wordplay where relevant?
4. Option Discrimination: Did it explain why the chosen caption is better than the other four, rather than merely sounding plausible on its own?

Respond ONLY in valid JSON with the following format:

```
{
  "scores": {
    "incongruity": 0.0,
    "humor_insight": 0.0,
  }
}
```

```

    "caption_alignment": 0.0,
    "option_discrimination": 0.0,
  },
  "major_flaws": [
    "...",
    "..."
  ],
  "minor_flaws": [
    "...",
    "..."
  ],
  "best_supported_elements": [
    "...",
    "..."
  ],
  "reflection": "A concise expert assessment of whether the original reasoning is trustworthy.",
  "revised_answer": "A/B/C/D/E",
  "revised_rationale": "A short explanation of whether the original answer should stand or be changed."
}

```

Ranking Task Audit Prompt

You are a senior cartoon-caption editor reviewing a model's reasoning trace for a New Yorker Caption Contest ranking task.

You are given:

- a cartoon scene description,
- an uncanny or unusual element,
- key observations,
- two candidate captions,

and a step-by-step reasoning chain produced by a model, and the model's selected answer.

Your task now is to critically audit the provided reasoning trace as an expert evaluator of humor.

You must assess whether the model's reasoning:

1. correctly identified the central incongruity driving the humor,
2. evaluated both captions comparatively (not in isolation),
3. captured subtle differences in tone, timing, and relevance,
4. justified why one caption is *funnier*, not just plausible.

Important:

- Do not simply restate the model's reasoning.
- Be skeptical and look for weaknesses.
- Focus on comparative humor judgment, not absolute interpretation.
- If the reasoning is weak or biased, say so clearly.
- Revise your decision if necessary.

Input:

Scene: [SCENE]

Description: [DESCRIPTION]

Uncanny Element: [UNCANNY ELEMENT]

Observations: [OBSERVATIONS]

Captions:

A) [CAPTION A]

B) [CAPTION B]

Model's Reasoning:

[REASONING CHAIN]

Model's Selected Answer:

[ANSWER]

Perform a professional audit along the following di-

mensions:

1. Incongruity Alignment: Did it correctly identify the *main* source of humor in the cartoon?
 2. Caption-Scene Fit: Did it clearly explain how each caption interacts with the visual?
 3. Humor Sensitivity: Did it capture nuances such as timing, understatement, absurdity, or cultural framing?
 4. Funniness Justification: Did it justify *why one is funnier*, not just more coherent?
 5. Comparative Evaluation: Did it explicitly compare A vs B, or just explain one in isolation?
- Respond ONLY in valid JSON:

```

{
  "scores": {
    "incongruity_alignment": 0.0,
    "comparative_evaluation": 0.0,
    "humor_sensitivity": 0.0,
    "caption_scene_fit": 0.0,
    "funniness_justification": 0.0,
  },
  "major_flaws": [
    "...",
    "..."
  ],
  "minor_flaws": [
    "...",
    "..."
  ],
  "strengths": [
    "...",
    "..."
  ],
  "reflection": "Concise expert judgment on whether the original reasoning is reliable.",
  "revised_answer": "A/B",
  "revised_rationale": "Short explanation of whether and why the choice should change."
}

```

L 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. L.1-L.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

Table K.1: **Cross-Model Reasoning chain audit of RM supervision traces.** DeepSeek-R1, NVLM-D-72B, and Gemini-2.5-Flash evaluate the traces across multiple rationale-quality dimensions. The generated traces consistently exhibit stronger higher-level humor reasoning, particularly in incongruity alignment, humor sensitivity, caption-scene fit, and funniness justification.

Matching Tasks			
Metric	DeepSeek-R1	NVLM-D-72B	Gemini-2.5-Flash
Incongruity	0.7525	0.9892	0.9629
Caption Alignment	0.6310	0.8469	0.9120
Humor Insight	0.6072	0.8322	0.8786
Option Discrimination	0.4903	0.7349	0.8516
Overall Score	0.6202	0.8508	0.9013

Ranking Tasks			
Metric	DeepSeek-R1	NVLM-D-72B	Gemini-2.5-Flash
Incongruity Alignment	0.7175	0.9305	0.9303
Caption Scene Fit	0.6554	0.7202	0.8334
Humor Sensitivity	0.5464	0.7817	0.7224
Funniness Justification	0.4532	0.6064	0.6914
Comparative Evaluation	0.5088	0.5428	0.8761
Overall Score	0.5762	0.7163	0.8107

Table K.2: **Reasoning chain audit on the test set.** We report the Gemini-2.5-Flash scores of inference-time rationales satisfying component-level rubric criteria across matching and ranking tasks.

Matching Tasks			
Metric	Zero-shot	Few-shot	IRS
Incongruity	0.3470	0.3637	0.4072
Caption Alignment	0.3030	0.3377	0.3551
Humor Insight	0.3671	0.3775	0.4042
Option Discrimination	0.3441	0.4065	0.4342
Overall Score	0.3403	0.3714	0.4002

Ranking Tasks			
Metric	Zero-shot	Few-shot	IRS
Incongruity Alignment	0.2914	0.3089	0.3452
Caption Scene Fit	0.3015	0.2711	0.3587
Humor Sensitivity	0.2693	0.2649	0.3160
Funniness Justification	0.2307	0.2227	0.2810
Comparative Evaluation	0.4985	0.5339	0.5976
Overall Score	0.3183	0.3203	0.3797

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.

M User Study Details

To establish a human reference point for the matching and ranking tasks, we conducted a small user study. Participants were shown New Yorker Cartoon Caption Contest cartoons and asked to select the most appropriate caption from the provided options, following the same task formats used in our automated evaluation (five-way selection for matching, two-way selection for ranking).

Instructions and interface. Participants completed the study through an online form. The full text of the instructions, disclaimers, and consent statement shown to participants, along with a representative matching question, is reproduced in Fig. M.1 and M.2. The instructions asked participants to view each cartoon carefully, read all candidate captions, and choose the one that best matches (matching task) or is funnier (ranking task). The matching and ranking interfaces were analogous, differing only in the number of candidate captions and the framing of the question. The study comprised 15 questions per task, and participants were asked to answer all of them.

Recruitment and payment. The non-expert participants (21 in total) were recruited on a voluntary basis from among graduate students in our research group. Participation was entirely optional and unpaid, was not tied to course credit, compensation, or any form of academic evaluation, and participants were free to decline or withdraw at any time without consequence. As stated in the consent text shown to participants (Fig. M.1), the study was an entirely voluntary, unpaid academic research effort with no financial compensation. Because participation was voluntary and unpaid rather than conducted through a paid crowdsourcing platform, no payment-adequacy considerations relative to participant demographics apply. The task involved only minimal risk (standard computer use), as disclosed to participants. We note that, while these participants are not professional captionists, they have some familiarity with this line of research; we therefore interpret their results as a non-expert human reference rather than as a sample drawn from the general population.

Expert reference. The expert human reference was provided by a domain expert with extensive professional experience evaluating New Yorker-style cartoon captions, who is a member of the research team for this work. The expert



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 L.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"

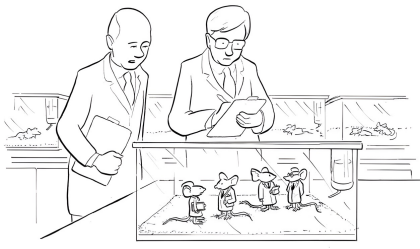
Irony: High expectations vs Low-function

Precisely subverting fairy tale tropes."

This maximizes visual irony and metaphorical clash.

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

Figure L.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 L.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.

completed the same held-out matching and ranking items used in the automated evaluation, under the same conditions, with no access to gold captions or crowd preference statistics at the time of eval-

uation. Because the expert is affiliated with the authors, we report the expert results as a qualitative reference for what human-level humor reasoning looks like rather than as a controlled, independent

NYCC - Matching (Hessel et al., 2023)

Instructions: Help us identify which captions best match the cartoons. For each question, view the cartoon image and select the most appropriate caption from the 5 options provided.

Disclaimers & Consent:

- **Voluntary & Unpaid:** This is an entirely voluntary, unpaid academic research effort. There is no financial compensation, and you may withdraw at any time.
- **Data Use:** Your responses will be completely anonymized and used strictly for research purposes, including potential publication and open-science data sharing.
- **Risks:** This task involves minimal risk (standard computer use).

By proceeding to the task, you confirm you are at least 18 years old and consent to your anonymized data being used for this research.

* Indicates required question

Figure M.1: Consent form and disclaimers shown to participants before the user study.

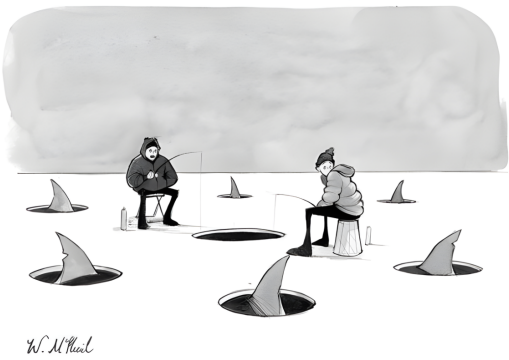
Instructions

Instructions:

- Look at each cartoon image carefully
- Read all five caption options (A through E)
- Choose which caption best matches or fits the cartoon
- There are 15 questions total
- Please answer all questions

Thank you for participating in our caption contest research!

Contest 471 Cartoon



Question 1: Contest 471: Which caption matches the cartoon? *

Contest 471

☐ A) Watch out for one with a saw.

☐ B) I'm afraid they'll give you life.

☐ C) They're very intelligent, dear. Try reasoning with him.

☐ D) Remember, it's all in the timing.

☐ E) Have you always felt like a victim?

Figure M.2: Task instructions and an example matching question.

population-level baseline.

Consent and data use. Before beginning the task, participants were shown a consent statement explaining that their responses would be completely anonymized and used strictly for research purposes, including potential publication and open-science

data sharing. By proceeding to the task, participants confirmed that they were at least 18 years old and consented to their anonymized data being used for this research. No personally identifying information was collected.

N 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.