FROM PERCEPTION TO PUNCHLINE: EMPOWERING VLM WITH THE ART OF IN-THE-WILD MEME

Anonymous authors
Paper under double-blind review

ABSTRACT

Generating humorous memes is a challenging multimodal task that moves beyond direct image-to-caption supervision. It requires a nuanced reasoning over visual content, contextual cues, and subjective humor. To bridge this gap between visual perception and humorous punchline creation, we propose *HUMOR*, a novel framework that guides VLMs through hierarchical reasoning and aligns them with group-wise human-like preferences. First, HUMOR employs a hierarchical, multi-path Chain-of-Thought (CoT): the model begins by identifying a template-level intent, then explores diverse reasoning paths under different contexts, and finally anchors onto a high-quality, context-specific path. This CoT supervision, which traces back from ground-truth captions, enhances reasoning diversity. We further analyze that this multi-path exploration with anchoring maintains a high expected humor quality, under the practical condition that high-quality paths retain significant probability mass. Second, to capture subjective humor, we train a pairwise reward model that operates within groups of memes sharing the same template. Following established theory, this approach ensures a consistent and robust proxy for human preference, even with noisy labels. The reward model then enables a group-wise reinforcement learning optimization, guaranteeing that the model's humor quality does not degrade beyond a bounded amount. Experiments show that HUMOR empowers various base VLMs with superior reasoning diversity, more reliable preference alignment, and higher overall meme quality compared to strong baselines. Beyond memes, our work presents a general training paradigm for open-ended, human-aligned multimodal generation, where success is guided by comparative judgment within coherent output groups.

1 INTRODUCTION

Creativity in multimodal generation increasingly moves beyond literal description to subjective and context-dependent outputs, such as humor, aesthetics, style, and social alignment, where quality is not defined by a single ground-truth but instead guided by human preference (Yadav et al., 2025; Burn & Kress, 2018). While recent vision–language models (VLMs) achieve strong results on captioning and visual question answering (Kuang et al., 2025; Ghandi et al., 2023), these tasks still admit relatively objective targets (Yan et al., 2023), leaving open how to train systems for goals that are open-ended and preference-driven (Bhatia et al., 2024). Current approaches often model meme generation as a direct image-to-caption task optimized with a fixed loss. This collapses the reasoning process into the decoder, suppresses intermediate interpretation, and tends to produce captions that are fluent yet shallow or not humorous (Yadav et al., 2025).

Meme generation provides a demanding testbed for this challenge. To succeed, a model must identify a template's latent intent, ground it in context-specific details of the image (objects, expressions, layout), and produce a caption that completes a metaphor or subverts expectation in a way humans find funny. This requires both **hierarchical reasoning** and **alignment with subjective humor**. Prior work typically uses text-only humor cues or global regression-style funniness scores (Baluja, 2024; Kalloniatis & Adamidis, 2024; Zhu et al., 2025a), assuming humor is directly comparable across templates. In practice, however, human judgments are more reliable within a group of memes that share the same template or theme, and far less stable across groups with different conventions. Ignoring this structure introduces noise, harms generalization, and encourages shortcuts that reward superficial overlap instead of genuine humor fit.

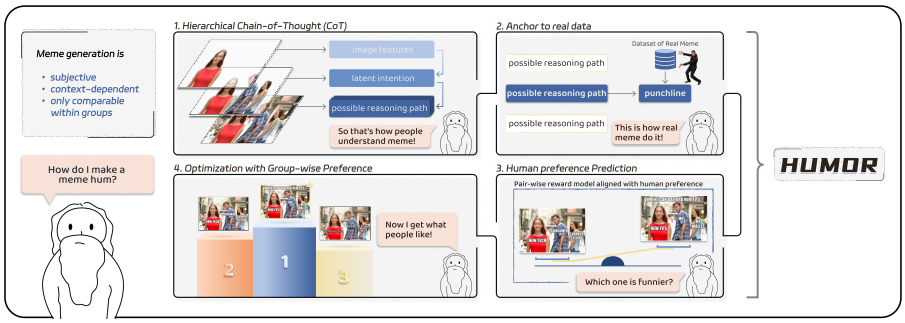

Figure 1: Overview of the HUMOR framework. Given a template image, it first performs hierarchical reasoning with a multi-path CoT: a template-level stage infers latent intent, and a context-level stage explores multiple paths grounded in visual content. One high-quality path is anchored by tracing back from ground-truth captions, supporting diversity while ensuring a conditional humor lower bound. A pairwise reward model then compares memes only within groups sharing the same template, maintaining rank consistency and providing a proxy signal of human-like preference. This reward enables group-wise RL to update the generation model in a stable way, ensuring expected humor does not degrade. Together, these components show how HUMOR combines structured reasoning, group-wise preference modeling, and stable optimization for meme generation.

A second limitation is the lack of an explicit reasoning-then-realization view. Directly sampling captions from images removes control over the interpretive process and makes it difficult to steer generation. Recent evidence shows that chain-of-thought (CoT) intermediates improve reasoning in VLMs. We argue that meme generation requires not just a single trace but a **hierarchical, multi-path reasoning process**: a template-level stage that infers canonical intent, followed by a context-level stage that grounds the intent in specific visual details. Different reasoning paths may lead to distinct metaphor bindings or punchlines. Exploring multiple paths and then anchoring one path with ground-truth data ensures diversity while, as our analysis shows, preserving a conditional lower bound on expected humor whenever high-quality paths keep a meaningful share of probability and the remaining paths are not much worse. Meeting these conditions requires optimizing generation toward human-preferred humor. Since humor cannot be directly measured, we design a **pairwise reward model** that maintains rank consistency within groups and prove that it inherits theoretical guarantees. This model provides a stable proxy signal of human-like preference, and further enables group-wise RL to ensure that expected humor cannot degrade beyond a bounded amount.

Figure 1 provides a high-level overview of *HUMOR*. It illustrates the main challenges in meme generation and how our framework addresses them: hierarchical reasoning with multi-path CoT, group-wise preference modeling, and stable optimization via RL. Taken together, these insights motivate our framework **HUMOR**: Hierarchical Understanding and Meme Optimization via Reinforcement learning. *HUMOR* separates reasoning from realization, respects group-wise comparability, and turns preference signals into stable policy updates. In summary, our contributions are:

1. **A new formulation of meme generation** as an open-ended, group-wise reasoning problem, together with a hierarchical multi-path CoT supervision scheme that separates template-level intent from context-level grounding. This framing exposes interpretable reasoning traces and lays the foundation for preference optimization.
2. **Theoretical analysis** showing that multi-path CoT supervision preserves a conditional humor lower bound and preference learning ensures consistent within-group ordering with provable stability. These results not only explain why our approach remains robust under noisy and subjective labels, also provide transferable insights for other open-ended, human-aligned generation tasks.
3. **Comprehensive experiments** across multiple base models showing that *HUMOR* improves reasoning diversity, preference alignment, and overall meme quality.

2 Related Work

2.1 Evolution of vision-language models for multi-modal process

The pursuit of unified vision-language modeling has progressed through three distinct phases of architectural innovation. Early foundational work established bidirectional frameworks for cross-

modal understanding: ERNIE-ViLG Zhang et al. (2021) and the Unifying Multi-modal Transformer Huang et al. (2021) pioneered transformer-based architectures that jointly optimized text-to-image and image-to-text generation through multi-modal tokenization and autoregressive objectives. Concurrently, Ramesh et al. Ramesh et al. (2021) demonstrated the scalability potential of such approaches through their zero-shot text-to-image generation framework, establishing critical baselines for large-scale multi-modal pretraining. Subsequent advancements focused on enhancing output quality and semantic alignment. Discrete diffusion architectures like Unified Discrete Diffusion Hu et al. (2022) and ERNIE-ViLG 2.0 Feng et al. (2023) introduced specialized denoising experts and semantic regularization techniques, significantly improving image fidelity and text-image correspondence. Contemporary breakthroughs have redefined architectural paradigms through multimodal unification. Models like Show-o Xie et al. (2024) and MonoFormer Zhao et al. (2024) successfully fused autoregressive and diffusion mechanisms within singular architectures via shared attention layers, achieving synergistic improvements in both generation quality and training efficiency. Building upon these advancements, our work leverages multi-modal comprehension capabilities to address the unique challenges of meme generation - particularly its requirement an understanding of metaphor, and subjective humor.

2.2 Meme Analysis and Generation

Internet memes have emerged as a vital component of digital culture, prompting substantial scholarly attention to their multi-modal communications. Extensive research has focused on analyzing topics Du et al. (2020) , semantics Xu et al. (2022) , and emotions Sharma et al. (2020) conveyed in memes. The evolution of meme generation techniques has progressed through distinct technological phases. Initial systems employed rule-based architectures, exemplified by Oliveira et al. Oliveira et al. (2016)'s template-driven approach using standardized structures like *"One does not simply X"*, and Wang et al. Wang & Wen (2015)'s dual-channel model integrating textual and visual features. The advent of deep learning catalyzed more sophisticated generation paradigms. Peirson and Tolunay pioneered this transition with Dank Learning Peirson V & Tolunay (2018), combining Inception V3 image encoders with attention-enhanced LSTM decoders to produce contextually humorous captions. Subsequent innovations introduced transformer architectures: Sadasivam et al.'s MemeBot Sadasivam et al. (2020) and Vyalla et al.'s Memeify Vyalla & Udandarao (2020) demonstrated enhanced text-image alignment through multi-modal fusion techniques. Recent breakthroughs leverage large language models (LLMs) and vision-language models (VLMs) to achieve unprecedented scale and specificity. Wang et al.'s Memecraft Wang & Lee (2024) enables targeted meme creation for social advocacy through cross-modal prompting. Addressing multi-image complexity, Chen et al. proposed XMeCap Chen et al. (2024b), introducing a two-stage framework with supervised fine-tuning and reinforcement learning guided by novel similarity metrics that evaluate both global contexts and localized visual-textual interactions. Concurrently, benchmark datasets have emerged to evaluate multi-modal understanding capabilities. The MemeCap Hwang & Shwartz (2023) provides 6.3K annotated memes with metaphor annotations, while the New Yorker benchmarks Hessel et al. (2022) assess humor comprehension through caption matching and explanation tasks. Expanding contextual understanding, the MCC dataset (MEMEX) Sharma et al. (2023) incorporates external knowledge sources to facilitate abstraction analysis and semantic dependency mining.

3 Problem Formulation

In this section, we specify the objects, signals, and assumptions used throughout the paper. We first define the meme space and its group structure, then describe local pairwise preference data and the latent humor within a group. We introduce a generic observation model for pairwise labels , followed by the generator-level objective and evaluation quantities. The goal is a self-contained problem formulation that highlights group-wise comparability without assuming any particular training method.

Meme Space and Group-wise Comparability: Let $\mathcal{M}$ denote the set of memes under consideration. A meme is a multimodal pair $m = (I, c)$, where $I \in \mathcal{I}$ is an image and c is a textual caption rendered at designated positions. Many memes are created from widely shared *templates* and interpreted through context-dependent associations. Absolute, cross-template comparisons of humor are often ill-posed. Therefore, we assume a collection of disjoint groups

$$\mathcal{G} = \{G_1, \ldots, G_K\}, \qquad G_k \subset \mathcal{M}, \qquad G_k \cap G_\ell = \emptyset \ (k \neq \ell),$$

such that memes within the same group share a comparable structure (e.g., the same template, topic, or punchline schema). We assume human judgments of humor are meaningful and more reliable *within* a fixed group $G \in \mathcal{G}$, while making no claim of comparability *across* groups.

Local Preference Data: For a given group G, human annotators provide pairwise labels indicating which of two memes is funnier. For $m_i, m_j \in G$, define $y_{ij}^G = \mathbb{I}[m_i \succ m_j] \in \{0, 1\}$ where $m_i \succ m_j$ denotes a local preference that m_i is judged funnier than m_j. The dataset comprises triples $(G, (m_i, m_j), y_{ij}^G)$ sampled from a pairing distribution over G. We allow for incompleteness (not all pairs are labeled) and noise (annotators may disagree). We adopt two weak but standard assumptions from preference learning : (i) *local comparability*: preferences are elicited and interpreted only within a fixed group G; (ii) *weak stochastic transitivity*: in expectation, if $m_i \succ m_j$ and $m_j \succ m_\ell$, then $m_i \succ m_\ell$ is more likely than its reversal, without requiring a strict total order.

Latent Humor within a Group: Within each group G, we posit an unobserved latent humor functional $h_G : G \to [0, 1]$, which maps each meme $m \in G$ to a scalar reflecting its relative likelihood of being judged funny by humans in that group. We do not assume that h_G is calibrated across groups, nor that h_G and $h_{G'}$ are directly comparable when $G \neq G'$.

Observation Model for Pairwise Labels: Pairwise labels are treated as noisy observations of differences in latent humor. We assume

$$\Pr\big[\, m_i \succ m_j \mid G \,\big] \;=\; \Lambda\big(\, h_G(m_i) - h_G(m_j) \,\big), \tag{1}$$

where $\Lambda : \mathbb{R} \to (0, 1)$ is a strictly increasing link (e.g., logistic or probit) . Intuitively, Eq. equation 1 states that the probability of preferring m_i to m_j depends *only* on their latent humor gap within the same group: when $h_G(m_i) \approx h_G(m_j)$, the choice is essentially ambiguous (probability $\approx 1/2$); as the gap grows, the probability moves smoothly toward 1 (if $h_G(m_i) > h_G(m_j)$) or 0 (otherwise), capturing that larger humor gaps yield more confident comparisons.

Generative Goal and Evaluation Quantities: A generation model produces captions conditioned on an image: $\pi_\theta(\cdot \mid I) : \quad I \in \mathcal{X} \quad \mapsto$ distribution over captions c. When combined with I, a sample $c \sim \pi_\theta(\cdot \mid I)$ instantiates a meme $m = (I, c)$. For any target group G containing I-based candidates, the expected within-group humor of π_θ is $\mathcal{H}_G(\theta) \;=\; \mathbb{E}_{c \sim \pi_\theta(\cdot|I)}\big[\, h_G\big((I, c)\big) \,\big]$, ,and the population objective aggregates over groups according to a task-specific distribution over (I, G):

$$\mathcal{H}(\theta) \;=\; \mathbb{E}_{(I,G)}\big[\, \mathcal{H}_G(\theta) \,\big]. \tag{2}$$

4 HUMOR FRAMEWORK

We propose **HUMOR**: **H**ierarchical **U**nderstanding and **M**eme **O**ptimization with group-wise reinforcement learning. The framework integrates three components: hierarchical chain-of-thought (CoT) supervision, reward modeling from pairwise preferences, and group-wise policy optimization. Together, these stages ensure that reasoning remains diverse, preferences are consistently captured, and optimization improves expected humor in a stable manner.

4.1 HIERARCHICAL CHAIN-OF-THOUGHT SUPERVISION

Meme generation requires reasoning over both a template's latent intent and its context-specific realization. Training a direct mapping $P_\theta(c \mid I)$ collapses this process into a single decoder, often leading to superficial captions. We instead represent reasoning as a hierarchical chain-of-thought $r = (r_{\text{tmpl}}, r_{\text{scene}})$, which separates template-level interpretation from context-level grounding. Captions are then realized by sampling from $P_\phi(c \mid r, I)$.

To approximate human authorship, we supervise CoT in two stages. In Stage 1, the model explores multiple reasoning paths conditioned only on I, while implicitly hypothesizing a *multiple potential user contexts* $\hat{U}$ (e.g., emotions, intentions, or scenarios a user might want to express). Concretely, the model generates reasoning candidates $\{r^{(i)}\} \sim P_\phi(r \mid I, \hat{U})$, encouraging coverage of diverse interpretations similar to how humans brainstorm several possible jokes before committing. In Stage 2, we anchor one path $\tilde{r}$ consistent with annotated captions, by incorporating the *actual user*

context U that is inferred from ground-truth captions (e.g., their sentiment, intention). Formally, we select $\tilde{r} = \arg\max_r P_\phi(c \mid r, I, U)$, which ensures stability while preserving the diversity gained in Stage 1. The hierarchical CoT framework and its details are provided in Appendix A.

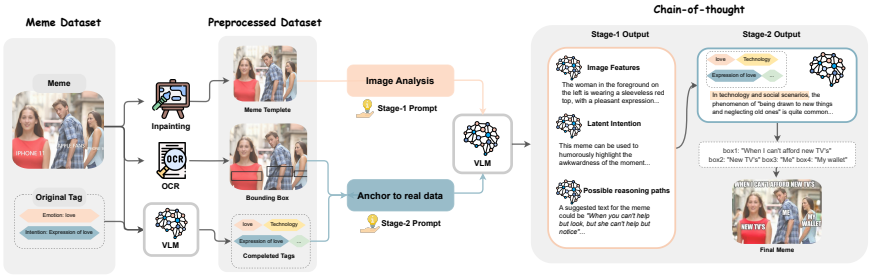

Figure 2: Hierarchical CoT supervision. Stage 1 explores multiple reasoning paths that bind a template intent to different context-specific details. Stage 2 anchors one high-quality path traced from ground-truth captions, preserving diversity while preventing collapse.

The benefit of this design can be formalized. Let $\tilde{h}_G : \mathcal{R} \to [0, 1]$ denote group-relative humor defined over reasoning traces. Suppose there exists a set of "star" paths $R^\star$ with probability mass $\alpha > 0$, and that the average humor gap between non-star paths and the best paths is at most δ. Then:

Proposition 1 (Conditional humor lower bound)**.** *Normalizing* $\max \tilde{h}_G = 1$*, the expected humor after CoT supervision satisfies*

$$\mathbb{E}_{r\sim P_\theta}[\tilde{h}_G(r)] \;\geq\; 1 - (1 - \alpha)\delta.$$

Intuitively, as long as promising reasoning paths retain nontrivial probability (α not too small) and the remaining paths are only mildly worse (small δ), exploration and anchoring preserve a nontrivial lower bound on expected humor. CoT thus broadens the breadth of interpretations without sacrificing quality. However, while α is naturally ensured by anchoring toward ground-truth paths, δ remains uncontrolled: some paths may still be substantially less funny. To minimize δ, we need a mechanism that reflects human humor preferences and can guide optimization beyond imitation.

4.2 Reward Modeling from Pairwise Preferences

The ideal objective would be to recover the latent humor $h_G(m)$ for each meme m. Since humor is subjective and lacks a global scale, this is infeasible. We therefore adopt an *order-consistent* view of reward modeling (following established theory (Sun et al., 2025)) and instantiate it in our *group-wise* meme setting: the reward acts as a *within-group surrogate* of h_G, trained only from relative judgments, avoiding ill-posed cross-group calibration. Intuitively, hierarchical CoT has ensured that high-quality paths keep a meaningful probability mass (the α condition via Stage 2 anchoring), while the reward model supplies the preference signal needed to *shrink the average gap among plausible paths* (the δ condition), turning open-ended exploration into learnable selection.

Each meme $m = (I, c)$ is encoded to a feature vector $\Psi(m) \in \mathbb{R}^d$ using a vision–language encoder. A scoring head $f_\phi : \mathbb{R}^d \to \mathbb{R}$ outputs $s_\phi(m)$, and for a pair (m_i, m_j) from group G we define

$$\widehat{p}_{ij}^{\,G} = \sigma\big(s_\phi(m_i) - s_\phi(m_j)\big), \tag{3}$$

with a logistic link $\sigma(\cdot)$; training minimizes the binary cross-entropy over labeled pairs.

Order consistency and stability in our setting. We make two statements precise for the *within-group* meme space (full proofs in Appendix B).

Proposition 2 (Rank consistency (following established theory))**.** *Under the observation model of Eq. 1 with any strictly increasing link, minimizing* $\mathcal{L}_{pair}$ *recovers the same within-group ordering as the latent humor* h_G*.*

Proposition 3 (Robustness to label noise (margin-aware))**.** *Let* $\Delta_{ij}^G = h_G(m_i) - h_G(m_j)$ *be the true humor gap, and suppose the classifier has pairwise error rate* ε*. Then for pairs with* $|\Delta_{ij}^G| \geq \delta$*, the probability of reversal is bounded above by a function decreasing in* δ *and increasing in* ε*; large humor gaps are therefore preserved even under noisy labels.*

The two propositions above follow the order-consistent analysis of Sun et al. (2025) but are instantiated under our group-wise comparability and used as *drivers* to reduce δ after CoT has secured α. Since pairwise data can be sparse, we aggregate $\widehat{p}_{ij}^{G}$ into a coherent within-group ranking via *Expected Borda Count (EBC)* (Appendix F): for a candidate set $\mathcal{S}_G$, each meme's score equals its expected number of wins against others under Eq. 3. This provides a stable target for training, and inherits expected order consistency when the pairwise model is consistent (Appendix B).

4.3 Group-wise Policy Optimization

We fine-tune the generator to *increase* the probability of higher-ranked captions within each group while *penalizing* deviations from a reference policy, balancing preference alignment with stability. Concretely, we adopt a **Group-wise Relative Policy Optimization (GRPO)** objective. For a candidate set $\mathcal{S}_G$ with ranking q_G from EBC, the reinforcement fine-tuning loss is:

$$\mathcal{L}_{\text{GRPO}}(\theta) = \mathbb{E}_{(I,G)}\Big[- \sum_{m_k \in \mathcal{S}_G} q_G(m_k)\, \log \pi_\theta(c_k \mid I)\Big] + \beta\, \mathbb{E}_I\big[\mathrm{KL}(\pi_\theta(\cdot \mid I) \,\|\, \pi_{\text{ref}}(\cdot \mid I))\big], \quad (4)$$

where π_{ref} is the SFT policy. The listwise term aligns π_θ with the group-local preference distribution q_G (rank-consistent with h_G), and the KL term limits drift, matching our comparability assumptions.

While prior analyses often state optimistic lower bounds for preference-optimized policies, we adopt a corrected, KL-controlled guarantee that holds under our setting and noise model; it is more conservative but faithful to the actual constraints (proof in Appendix C).

Proposition 4 (Bounded change of expected humor under GRPO)**.** *Assume Proposition 2 and $h_G \in [0,1]$. Let $\Delta_{\mathrm{KL}} = \mathbb{E}_I[\mathrm{KL}(\pi_\theta(\cdot \mid I) \,\|\, \pi_{ref}(\cdot \mid I))]$. Then*

$$\mathbb{E}_{(I,G)}\Big[\mathbb{E}_{c\sim\pi_\theta(\cdot|I)}\, h_G((I,c))\Big] \;\geq\; \mathbb{E}_{(I,G)}\Big[\mathbb{E}_{c\sim\pi_{ref}(\cdot|I)}\, h_G((I,c))\Big] \;-\; \sqrt{\tfrac{1}{2}\,\Delta_{\mathrm{KL}}}.$$

Hence, if GRPO enforces $\Delta_{\mathrm{KL}} \leq \tau$, the expected humor cannot drop by more than $\sqrt{\tau/2}$; with the listwise pull toward q_G, this yields non-decreasing behavior within a bounded KL neighborhood.

This bound (via Pinsker's inequality) formalizes the stability we rely on in practice: CoT supplies support (α), the reward model and EBC induce a group-local order that reduces δ, and GRPO turns this order into controlled policy updates. In sum, our use of order-consistent surrogates follows established theory where appropriate, but the *group-wise instantiation*, the *corrected KL-based bound*, and the *integration with multi-path CoT for open-ended generation* are key ingredients that make the approach effective and verifiable for meme generation.

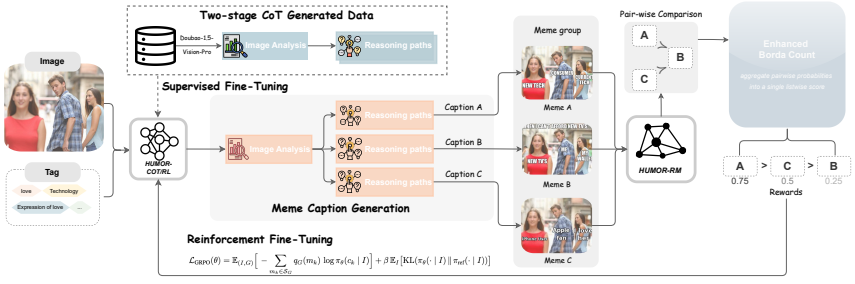

Figure 3: Training Pipeline of HUMOR. Multi-path CoT expands reasoning coverage and anchors a canonical path; the reward model translates pair data into a rank-consistent group-level signal (via EBC); GRPO then updates the generator toward higher-ranked captions.

4.4 Summary

CoT supervision establishes a conditional lower bound on expected humor (Proposition 1) by exploring multiple reasoning paths for coverage and anchoring a canonical path to prevent collapse. The reward model then supplies a rank-consistent and noise-robust surrogate for the (group-local) humor function (Proposition 2, Proposition B.3), and aggregates sparse pairwise labels into coherent within-group rankings. Finally, GRPO turns these rankings into stable policy updates with KL control and improvement guarantees (Proposition 4). Together, these components form HUMOR.

5 EXPERIMENT

Our experiments evaluate whether *HUMOR* performs as intended: (i) fine-tuning with hierarchical CoT can improve generation quality compared to direct mapping and naive CoT approaches; and (ii) learning within-group preferences and translating them into a consistent group ranking can be effectively utilized for further optimization. We present our findings by addressing the following research questions: RQ1: Does *HUMOR* enhance meme quality and diversity compared to strong baseline methods? RQ2: Can VLMs serve as reliable judges for meme, and how should they be appropriately applied to train reward models? RQ3: Does the proposed reward model align with human rankings within a group, and how effective is subsequent RL training using this model? RQ4: What does the reward model learn, and what insights can be gained through further visualizations?

5.1 MEME QUALITY AND DIVERSITY WITH HUMOR

Settings: we compare models trained under the *HUMOR* framework against several strong baselines and variants. Concretely, our evaluation covers multiple open-source and closed-source VLMs, as well as our *HUMOR-CoT* model, which is fine-tuned using only the hierarchical CoT design. Given the highly open-ended and human-aligned nature of meme generation, we prioritize human evaluation. Human raters assign scores to generated memes across four predefined axes. In addition, we adopt the conventional metric of text-level similarity between generated captions and their original reference texts. To further characterize diversity, we introduce a novel metric called **Distance under Context Swap**. This measure replaces the training-set context with a randomly selected one—kept consistent across models—and computes the textual distance from the original caption. A larger distance suggests reduced overfitting to SFT labels and better adaptability to new contexts. Due to observed instability in VLM-based rubric scores for meme evaluation, we incorporate only a single VLM-based metric: a human-likeness score. This is formulated as a binary classification estimate of the probability that a meme was created by a human, with higher values indicating better.

Table 1: **Evaluation results across open-source models, closed-source models, and Qwen2.5-7B-Instruct fine-tuned with different CoT generation methods.** Metrics include context-swap distance (diversity), text-level similarity (sim. to original meme text), human evaluation (Humor, Readability, Relevance, Originality), and Human Rates. Note that Human Rate for **Gemini-2.5-flash** is omitted since this metric is evaluated with itself, making it unavailable for this variant.

Category / Model	Human Evaluation (0-5) ↑				Sim. ↑	Distance ↑	Human Rate (%)↑
	Humor	Readability	Relevance	Originality			
Open-source Models							
Qwen2.5-7B-Instruct (Bai et al., 2025)	2.39	3.35	2.91	2.57	0.549	0.564	75.7
Qwen2.5-32B-Instruct (Bai et al., 2025)	2.54	3.52	3.09	2.76	0.532	0.566	82.2
InternVL3-8B (Zhu et al., 2025b)	2.39	2.79	3.04	2.79	0.507	0.564	62.7
GLM-4.1V-9B-Thinking (Hong et al., 2025)	1.73	2.62	2.75	2.71	0.556	0.572	45.1
Keye-VL-8B-preview (Team et al., 2025)	2.35	3.19	2.99	2.71	0.526	0.580	69.0
Closed-source Models							
GPT-4o (OpenAI, 2024)	2.70	2.99	3.21	**2.97**	0.578	0.552	91.3
Gemini-2.5-flash (Comanici et al., 2025)	2.81	3.29	3.25	2.88	0.565	0.561	-
Fine-tuned Model							
HUMOR-CoT	2.68	**3.70**	3.50	2.90	**0.591**	0.590	91.5
CoT with Single Path (Kim et al., 2023)	1.87	2.79	2.68	2.45	0.583	0.570	86.0
CoT with Self-Improve (Chen et al., 2024a)	2.38	3.68	3.00	2.65	0.579	0.578	89.1
CoT with Subquestion (Wei et al., 2022)	1.85	3.32	2.58	2.47	0.579	**0.597**	87.2
HUMOR-RL (preview)	**2.83**	3.67	**3.55**	2.79	0.582	0.588	**92.3**

Results and Discussion: Table 1 summarizes the overall performance of meme generation across various models The results indicate that our *HUMOR* framework achieve substantial improvements across multiple dimensions, which demonstrates the effectiveness of the *HUMOR* framework for humor-oriented meme generation. Specifically, in terms of *Humor*, *HUMOR-CoT* attains a score of 2.68, substantially surpassing the base model Qwen2.5-7B-Instruct, which scores only 2.39. Qualitative analysis suggests that *HUMOR*-improved models better capture nuanced humor mechanisms such as sarcasm and self-mockery, with *HUMOR-RL* further enhancing this capability. For **Readability**, *HUMOR-CoT* achieves a score of 3.70, outperforming all compared variants—including powerful closed-source models. It can generate captions with appropriate length and engaging structure, avoiding the verbosity common in many VLMs while maintaining humor expressivity, thereby

better aligning with human writing conventions. In **Theme Relevance** and **Originality**, *HUMOR-CoT* also performs strongly, demonstrating an ability to align with deeper user intent and keywords rather than merely referencing superficial visual elements. Although semantic similarity is less indicative for meme captions—which often consist of short phrases, *HUMOR-CoT* still achieves the closest alignment to reference captions among all models. For our proposed **Context-Swap Distance** metric, *HUMOR-CoT* scores 0.590, compared to 0.564 for the baseline, indicating a stronger capacity to produce diverse and context-sensitive outputs when user inputs are altered. This result supports the hypothesis that *HUMOR* reduces overfitting to concrete training labels. Finally, in the Human-Likeness Score, *HUMOR-CoT* exceeds 91%, significantly outperforming the base model (75.7%) and even surpassing the closed-source GPT-4o (91.3%).

5.2 VLM Reliability Evaluation

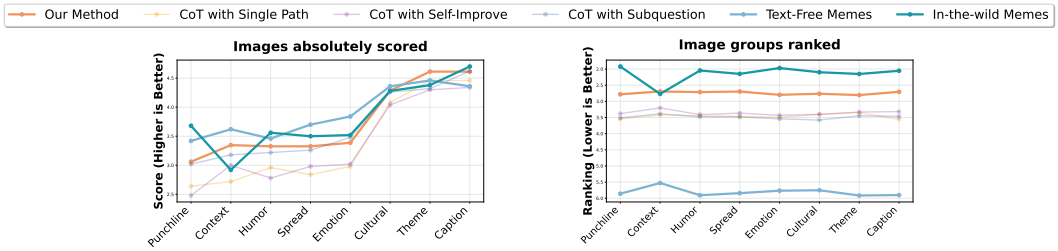

Figure 4: (left) VLM-based **absolute scoring** fails to distinguish meme quality. (right) **Group-wise ranking** produces more reliable distinctions, better aligned with human.

Figure 4 compares evaluation strategies for assessing meme-generation quality. When a VLM is used to assign absolute scores to individual memes (Fig. 4a), it fails to meaningfully distinguish between methods of clearly different quality levels. For instance, *In-the-wild Memes* (human-created and high-quality) and *Text-Free Memes* (text removed) receive similar scores across most dimensions, despite their evident disparity. This result underscores a key limitation of absolute scoring: since humor and cultural resonance are inherently relative and context-sensitive, evaluating memes in isolation proves unreliable. To overcome this issue, we introduce a **group-wise ranking** strategy, in which memes generated from the same base image are compared and ranked collectively across methods. As shown in Fig. 4b, this relative assessment successfully separates high- and low-quality examples, yielding a ranking that aligns more closely with human judgment. Further supported by the radar chart in Fig. 5, the relative evaluation reveals that our method performs second only to *In-the-wild Memes*, consistently surpassing all other generation strategies across every metric.

5.3 Reward Model Rank Consistency and RL Training

In Table 2, we evaluate how reward models fine-tuned on different base models align with human rankings. *Image1–Image5* are five templates (10–15 candidates each; Figure 8). For each template, we obtain a *group-level* human ranking via MaxDiff (Appendix H.1). Model rankings are produced by (i) collecting within-group pairwise probabilities from either the *base* model or the fine-tuned reward model (*HUMOR-RM*), and (ii) aggregating them with Expected Borda Count (EBC). We report Kendall's τ and its p-value to test the *rank consistency* objective (Section 4.2). *HUMOR-RM* on *Keye-VL* achieves consistently high τ with significant p-values (often $p \leq 10^{-3}$) across Image1–Image5, indicating strong within-group agreement with human preferences. On *Qwen2.5-VL-7B*, results are mixed (some moderate, some near-chance, significance not always reached). *Qwen2.5-VL-32B* and other backbones show limited or unstable gains. Overall, under the same fine-tuning and rank-only supervision, stronger, semantically aligned backbones yield reliable rank consistency, whereas weaker or less aligned ones align less steadily. We also validate the effectiveness of the combination between newly-designed content reward (Appendix E) and our pairwise reward model for RL training, where the preview version (*HUMOR-RL*) is shown in Table 1.

5.4 Base Model Comparison and Visualization

Across all evaluated templates (Image 1–5), the *Keye-VL* base model achieves higher within-group ranking consistency with human preferences than *Qwen-VL*. This performance gap is not attributable to the reward model—which is rank-based and group-local—but rather reflects inherent differences

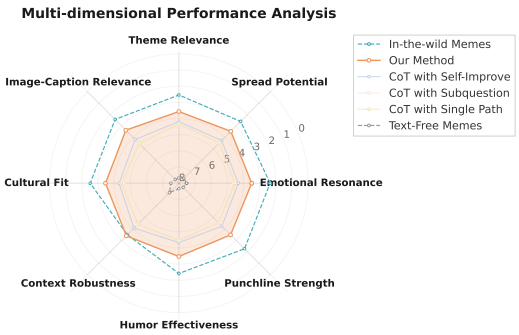

Figure 5: Radar chart comparing methods across multiple dimensions. Our method approaches the quality of human-created memes.

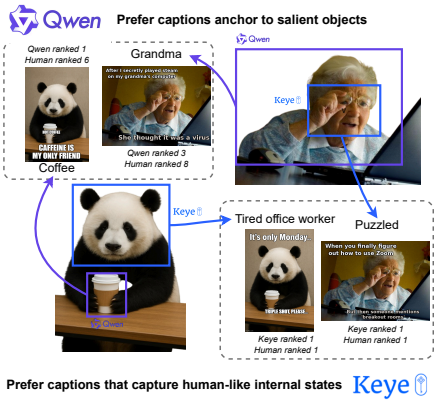

Figure 6: *Qwen* prefers captions that mention direct objects, whereas *Keye* prefers captions reflecting the human-like state.

Table 2: Ranking results of different baselines among distinct template images. It indicates the change after fine-tuning relative to the baseline: an increase in Kendall tau τ and a decrease in p-value p represent improvements (highlighted in green), while the opposite indicates deterioration (shown in red). Significance levels: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Model	Image 1		Image 2		Image 3		Image 4		Image 5	
	$\tau \uparrow$	$p \downarrow$	$\tau \uparrow$	$p \downarrow$	$\tau \uparrow$	$p \downarrow$	$\tau \uparrow$	$p \downarrow$	$\tau \uparrow$	$p \downarrow$
Qwen2.5-VL-7B (Base)	0.16	0.60	0.28	0.17	0.47	0.07	-0.10	0.63	0.29	0.29
Qwen2.5-VL-7B (Finetuned)	0.47	0.07	0.56	0.03*	0.42	0.11	0.14	0.50	0.47	0.07
Δ vs Base	+0.31	−0.53	+0.28	−0.14	−0.04	+0.04	+0.25	−0.13	+0.18	−0.22
Qwen2.5-VL-32B (Base)	0.16	0.61	0.16	0.44	-0.02	1.00	0.14	0.50	0.29	0.29
Qwen2.5-VL-32B (Finetuned)	0.29	0.29	0.47	0.02*	0.07	0.86	0.30	0.14	0.42	0.11
Δ vs Base	+0.13	−0.32	+0.30	−0.42	+0.09	−0.14	+0.15	−0.36	+0.13	−0.18
Keye-VL-8B (Base)	0.05	0.85	0.09	0.70	0.16	0.60	0.29	0.29	0.16	0.60
Keye-VL-8B (Finetuned)	0.78	0.00***	0.77	0.00***	0.78	0.00***	0.78	0.00***	0.78	0.00***
Δ vs Base	+0.73	−0.84	+0.69	−0.70	+0.62	−0.60	+0.49	−0.29	+0.62	−0.60

in representational capacity between the two base models. For example, as illustrated in Figure 6, when Image 5 depicts a panda holding a coffee cup, *Qwen-VL* favors captions containing the word "coffee". Similarly, for Image 2, which shows an older woman looking at a laptop, the model exhibits a preference for captions referencing "grandma" or computer-related terms. In contrast, *Keye-VL* more consistently captures implied internal states or situational cues within the scene and aligns them with the template's communicative intent. In the same examples, *Keye-VL* interprets the panda as resembling a "tired office worker" and the woman as appearing "puzzled", interpretations that correspond more closely with human rankings under our within-group evaluation protocol. These observations aligns with our theoretical expectation: the reward model supplies only a preference ordering; a model's ability to ascend that ordering depends fundamentally on its capacity to represent the nuanced cues that humans use in evaluating humor.

6 Conclusion

In this work, we tackled the complex challenge of teaching VLMs the art of in-the-wild meme generation, a task that requires nuanced reasoning beyond standard image captioning. Our proposed framework, *HUMOR*, successfully bridges the gap from visual perception to humorous punchline by instituting a two-stage process of hierarchical reasoning and preference alignment. Through a novel hierarchical CoT, the model learns to explore diverse creative paths while anchoring on high-quality outcomes. Furthermore, by leveraging group-wise preference modeling and RL, we ensure the generated humor aligns with human judgment in a stable and consistent manner. This work establishes a general and effective paradigm for open-ended multimodal generation tasks.

LLM USAGE STATEMENT

We employ vision–language models (VLMs) for data preprocessing and evaluation. Specifically, we use *Doubao* to perform label assignment and generate hierarchical CoT traces for training data; at evaluation time, we use *Qwen-VL*, *Keye-VL*, and *Gemini-2.5-pro* as VLM judges to assess generated memes. For writing clarity only, we use *GPT-5* to polish the paper's wording without changing technical content or claims.

ETHIC STATEMENT

All training data are drawn from publicly available datasets and contain no voiceprint/biometric audio information. During preprocessing, we filter violent content to the extent possible. However, we cannot guarantee that a model trained for open-ended meme generation will never produce violent or sensitive content at inference time. We therefore recommend deploying standard safety measures (content filters, human-in-the-loop review, and usage policies) to mitigate potential misuse and reduce exposure to harmful outputs.

REPRODUCIBILITY STATEMENT

Upon acceptance, we will release: (i) the full list of dataset sources we use; (ii) our constructed CoT supervision data and the pairwise/reward datasets; and (iii) the complete training and inference codebase. We will also provide prompts, hyperparameters, random seeds, model checkpoints (or scripts to reproduce them), and evaluation scripts to enable end-to-end replication.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Ashwin Baluja. Text is not all you need: Multimodal prompting helps llms understand humor. *arXiv preprint arXiv:2412.05315*, 2024.

Mehar Bhatia, Sahithya Ravi, Aditya Chinchure, Eunjeong Hwang, and Vered Shwartz. From local concepts to universals: Evaluating the multicultural understanding of vision-language models. *arXiv preprint arXiv:2407.00263*, 2024.

Andrew Burn and Gunther Kress. Multimodality, style, and the aesthetic: The case of the digital werewolf. In *Multimodality and aesthetics*, pp. 15–36. Routledge, 2018.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*, 2024a.

Yuyan Chen, Songzhou Yan, Zhihong Zhu, Zhixu Li, and Yanghua Xiao. Xmecap: Meme caption generation with sub-image adaptability. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 3352–3361, 2024b.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Yuhao Du, Muhammad Aamir Masood, and Kenneth Joseph. Understanding visual memes: An empirical analysis of text superimposed on memes shared on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pp. 153–164, 2020.

Zhida Feng, Zhenyu Zhang, Xintong Yu, Yewei Fang, Lanxin Li, Xuyi Chen, Yuxiang Lu, Jiaxiang Liu, Weichong Yin, Shikun Feng, et al. Ernie-vilg 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10135–10145, 2023.

Taraneh Ghandi, Hamidreza Pourreza, and Hamidreza Mahyar. Deep learning approaches on image captioning: A review. *ACM Computing Surveys*, 56(3):1–39, 2023.

Jack Hessel, Ana Marasović, Jena D Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. Do androids laugh at electric sheep? humor" understanding" benchmarks from the new yorker caption contest. *arXiv preprint arXiv:2209.06293*, 2022.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv e-prints*, pp. arXiv–2507, 2025.

Minghui Hu, Chuanxia Zheng, Heliang Zheng, Tat-Jen Cham, Chaoyue Wang, Zuopeng Yang, Dacheng Tao, and Ponnuthurai N Suganthan. Unified discrete diffusion for simultaneous vision-language generation. *arXiv preprint arXiv:2211.14842*, 2022.

Yupan Huang, Hongwei Xue, Bei Liu, and Yutong Lu. Unifying multimodal transformer for bi-directional image and text generation. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, pp. 1138–1147, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450386517. doi: 10.1145/3474085.3481540. URL https://doi.org/10.1145/3474085.3481540.

EunJeong Hwang and Vered Shwartz. Memecap: A dataset for captioning and interpreting memes. *arXiv preprint arXiv:2305.13703*, 2023.

Antonios Kalloniatis and Panagiotis Adamidis. Computational humor recognition: a systematic literature review. *Artificial Intelligence Review*, 58(2):43, 2024.

Seungone Kim, Se June Joo, Doyoung Kim, Joel Jang, Seonghyeon Ye, Jamin Shin, and Minjoon Seo. The cot collection: Improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning. *arXiv preprint arXiv:2305.14045*, 2023.

Jiayi Kuang, Ying Shen, Jingyou Xie, Haohao Luo, Zhe Xu, Ronghao Li, Yinghui Li, Xianfeng Cheng, Xika Lin, and Yu Han. Natural language understanding and inference with mllm in visual question answering: A survey. *ACM Computing Surveys*, 57(8):1–36, 2025.

Jordan J Louviere and George G Woodworth. Best-worst scaling: A model for the largest difference judgments. Technical report, working paper, 1991.

Jordan J Louviere, Terry N Flynn, and Anthony Alfred John Marley. *Best-worst scaling: Theory, methods and applications*. Cambridge University Press, 2015.

Hugo Gonçalo Oliveira, Diogo Costa, and Alexandre Miguel Pinto. One does not simply produce funny memes!–explorations on the automatic generation of internet humor. In *Proceedings of the Seventh International Conference on Computational Creativity (ICCC 2016). Paris, France*, 2016.

OpenAI. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024. Accessed: 2025-09-25.

Abel L Peirson V and E Meltem Tolunay. Dank learning: Generating memes using deep neural networks. *arXiv preprint arXiv:1806.04510*, 2018.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *ArXiv*, abs/2102.12092, 2021. URL https://api.semanticscholar.org/CorpusID:232035663.

Aadhavan Sadasivam, Kausic Gunasekar, Hasan Davulcu, and Yezhou Yang. Memebot: Towards automatic image meme generation. *arXiv preprint arXiv:2004.14571*, 2020.

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas Pykl, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Bjorn Gamback. Semeval-2020 task 8: Memotion analysis–the visuo-lingual metaphor! *arXiv preprint arXiv:2008.03781*, 2020.

Shivam Sharma, Udit Arora, Md Shad Akhtar, Tanmoy Chakraborty, et al. Memex: Detecting explanatory evidence for memes via knowledge-enriched contextualization. *arXiv preprint arXiv:2305.15913*, 2023.

Hao Sun, Yunyi Shen, and Jean-Francois Ton. Rethinking reward modeling in preference-based large language model alignment. In *The Thirteenth International Conference on Learning Representations*, 2025.

Kwai Keye Team, Biao Yang, Bin Wen, Changyi Liu, Chenglong Chu, Chengru Song, Chongling Rao, Chuan Yi, Da Li, Dunju Zang, et al. Kwai keye-vl technical report. *arXiv preprint arXiv:2507.01949*, 2025.

Suryatej Reddy Vyalla and Vishaal Udandarao. Memeify: A large-scale meme generation system. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*, pp. 307–311. 2020.

Han Wang and Roy Ka-Wei Lee. Memecraft: Contextual and stance-driven multimodal meme generation. In *Proceedings of the ACM Web Conference 2024*, pp. 4642–4652, 2024.

William Yang Wang and Miaomiao Wen. I can has cheezburger? a nonparanormal approach to combining textual and visual information for predicting and generating popular meme descriptions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 355–365, 2015.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024.

Bo Xu, Tingting Li, Junzhe Zheng, Mehdi Naseriparsa, Zhehuan Zhao, Hongfei Lin, and Feng Xia. Met-meme: A multimodal meme dataset rich in metaphors. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pp. 2887–2899, 2022.

Srishti Yadav, Zhi Zhang, Daniel Hershcovich, and Ekaterina Shutova. Beyond words: Exploring cultural value sensitivity in multimodal models. *arXiv preprint arXiv:2502.14906*, 2025.

Ming Yan, Haiyang Xu, Chenliang Li, Junfeng Tian, Bin Bi, Wei Wang, Xianzhe Xu, Ji Zhang, Songfang Huang, Fei Huang, et al. Achieving human parity on visual question answering. *ACM Transactions on Information Systems*, 41(3):1–40, 2023.

Han Zhang, Weichong Yin, Yewei Fang, Lanxin Li, Boqiang Duan, Zhihua Wu, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vilg: Unified generative pre-training for bidirectional vision-language generation. *ArXiv*, abs/2112.15283, 2021. URL `https://api.semanticscholar.org/CorpusID:245634812`.

Chuyang Zhao, Yuxing Song, Wenhao Wang, Haocheng Feng, Errui Ding, Yifan Sun, Xinyan Xiao, and Jingdong Wang. Monoformer: One transformer for both diffusion and autoregression. *arXiv preprint arXiv:2409.16280*, 2024.

Haohao Zhu, Junyu Lu, Zeyuan Zeng, Zewen Bai, Xiaokun Zhang, Liang Yang, and Hongfei Lin. Commonality and individuality! integrating humor commonality with speaker individuality for humor recognition. *arXiv preprint arXiv:2502.04960*, 2025a.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025b.

A HIERARCHICAL CHAIN-OF-THOUGHTS OF METAPHOR

To enhance our model's understanding of humor, we replicated the human meme creation process. Through extensive analysis of human meme creation, we extracted a paradigm for hierarchical meme feature analysis.

Take the "Distracted Boyfriend" meme as an example. Humans first capture: the delighted expression of the woman on the left, the action of the man in the center looking back and his subtle flirtatious gaze, the annoyed posture of the woman on the right, and the triangular compositional relationship and explicit emotional direction formed by the three individuals. Humans further abstract this scene and discover that it can be applied to any scenario of infatuation with something new and abandonment of the old, establishing entity mapping relationships. Thus, when the user's request is workplace culture, this template can be adapted to depict a leader being attracted by a new employee during a meeting, with a senior employee showing an expression of helplessness, vividly illustrating the workplace "new vs. old" relationship and generating humor.

How would humans fill in the text? Through statistical analysis of 5,000 classic memes, we found that the text positions in common meme templates are fixed, and the text content is highly correlated with its position. For instance, in the "Distracted Boyfriend" template, the position corresponding to the woman on the right is often used to represent the neglected object, the position corresponding to the man in the center represents the subject of attention shift, and the position corresponding to the woman on the left is the newly focused entity. Therefore, we integrate "text content generation" and "text position allocation" in the meme generation process. By annotating text box positions in the image, the model only needs to use its inherent visual localization ability to find the boxes, understand that text needs to be written in specific areas, and then combine spatial semantic mapping relationships to generate text with greater humorous effects in these positions.

We aim to imitate this thought process to construct Chain-of-Thought (CoT) data:

Data Collection and Preprocessing

Meme Images We collected over 80,000 meme images from platforms such as imgflip, quickmeme, and know your meme, and established a multi-dimensional labeling system:

1. **Emotion Classification**: Covers 7 basic emotions and intensity levels.
2. **Intent Detection**: Differentiates between 10 creation intents such as offense and entertainment.
3. **Metaphor Analysis**: Records metaphorical entities and cross-domain mapping relationships.

Base Images and Text Content/Position Information The FLUX.1-dev-Controlnet-Inpainting-Beta model is used to erase and restore the text areas in original memes, obtaining text-free base images. Meanwhile, OCR technology precisely records the (position, content) pairs of text, providing spatial semantic data for subsequent training.

User Requirements We reconstructed user requirements in reverse using APIs. Taking the meme's labels and final text as inputs, we utilized prompts to reverse-engineer the user's initial request. We analyzed the following dimensions of user requirements: emotion category, emotion intensity, intention, Scene or theme , style preference, and keywords.

CoT Data Generation

Stage One Using the base image as input, we extract high-level semantics of the meme.

First, we perform visual element decomposition. Our framework systematically deconstructs meme templates from four key visual dimensions:

1. **Main Subject Characteristics**: Analyze facial expressions, poses, clothing, and dynamic relationships between characters.
2. **Composition Logic**: Identify visual focal points, color contrasts, and spatial relationships.
3. **Cultural Markers**: Recognize identifiable meme formats and pop culture references.
4. **Narrative Threads**: Interpret body language implications and prop symbolism.

Then, we conduct scenario association and humor construction based on visual analysis:

1. **Social Contexts**: Identify scenarios suitable for group chats, comment sections, and private conversations.
2. **Topic Relevance**: Establish connections with workplace culture, life dilemmas, and internet hotspots.
3. **Emotional Mapping**: Determine appropriate humor techniques, including satire, self-deprecation, exaggeration, and contrast.

Stage Two Using the base image analysis from Stage One, user requirements, and final text as inputs, we infer the customized creation process for specific requests.

We provide few-shot examples of this parsing process. For instance, for the "Distracted Boyfriend" meme, when Stage One yields the semantic pattern of infatuation with something new and abandonment of the old, and identifies three entity positions: A [attention-shifting subject], B [newly focused entity], and C [neglected object], the user's request is a technology theme with the keyword "Apple fanatic." We consider how to align the expression of infatuation with something new and abandonment of the old with the context of technology product updates to reflect being an Apple fanatic. We infer that the semantic mapping of new and old phones is similar. Therefore, combining this image, we deduce that the text should be filled as: "A: APPLE FANS, B: IPHONE 11, C: IPHONE 10," humorously expressing enthusiasm for Apple's new technological products.

Training Rationale and Process We conduct instruction-tuning training using CoT data as supervisory signals. Since our training data contains numerous instances of the same base image, the two-stage CoT process essentially learns metaphorical semantic relationships across different scenarios. It is a divergent associative thinking training where one base image corresponds to multiple scenarios. This CoT approach not only enables the model to understand the high-level semantics of the image itself but also establishes multi-scenario associative capabilities.

Determination and Extraction of Generated Text Format Text boxes in the image are marked using a top-to-bottom, left-to-right coordinate sorting rule, and text content is recorded in the labels in order and in box format. The prompt explicitly requires the model to output in the format "box1:text1, box2:text2."

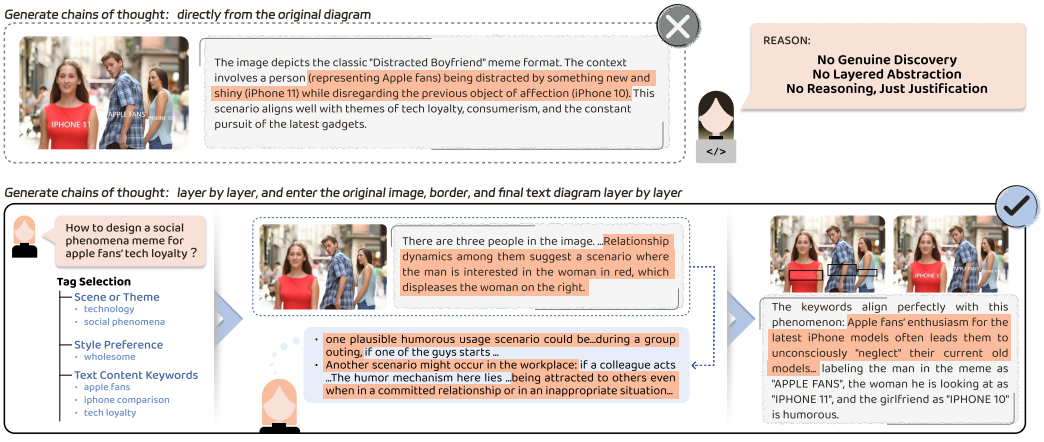

Figure 7: Comparison between direct CoT generation from the original image and our hierarchical CoT generation approach.

Critical Comparison: Direct vs. Hierarchical CoT The direct approach of generating chains of thought from the original image is essentially reverse engineering rather than genuine reasoning. It suffers from four critical flaws: 1) **No Genuine Discovery**: it skips the exploratory stage where humor emerges from active associative search, jumping straight to a fixed answer; 2) **No Layered Abstraction**: it leaps from raw visual details to a specific conclusion without building transferable intermediate metaphors; 3) **No Reasoning, Just Justification**: instead of true inference, it merely defends a predetermined conclusion.

In contrast, our layered CoT framework mirrors human reasoning by progressively abstracting from visual description to general metaphorical patterns and then to domain-specific humor instantiations, thereby enabling genuine creativity and robust generalization.

B Reward Modeling: Assumptions and Proofs

B.1 Setup and Assumptions

For a fixed group G, the latent humor functional is $h_G : G \to [0, 1]$. Pairwise labels follow the observation model of Eq. (1):

$$\Pr[m_i \succ m_j \mid G] \;=\; \Lambda\big(h_G(m_i) - h_G(m_j)\big),$$

where $\Lambda : \mathbb{R} \to (0, 1)$ is strictly increasing. A reward model maps a meme $m = (I, c)$ to a score $s_\phi(m)$; the pairwise probability is

$$\hat{p}_{ij}^{G} \;=\; \sigma\big(s_\phi(m_i) - s_\phi(m_j)\big),$$

and ϕ is learned by minimizing the empirical pairwise cross-entropy $\mathcal{L}_{\text{pair}}$. We assume (A1) the data contains i.i.d. pairs drawn within G with non-degenerate coverage; (A2) the model class for s_ϕ is rich enough to fit the Bayes-optimal decision boundary; (A3) identifiability is up to an additive constant per group (sufficient for ranking).

B.2 Rank Consistency (Proposition 1) — Proof

Proposition (Rank consistency (main text Proposition 1)). Under Eq. (1) with strictly increasing Λ, any risk minimizer of the logistic pairwise loss recovers the same within-group ordering as h_G.

Proof. Let $\eta_{ij} = \Pr[m_i \succ m_j \mid G] = \Lambda(\Delta_{ij})$ with $\Delta_{ij} = h_G(m_i) - h_G(m_j)$. The Bayes-optimal pairwise classifier for logistic loss satisfies $\sigma(s_i^\star - s_j^\star) = \eta_{ij}$, hence

$$s_i^\star - s_j^\star \;=\; \sigma^{-1}(\eta_{ij}) \;=\; \sigma^{-1}\big(\Lambda(\Delta_{ij})\big) \;=:\; \psi(\Delta_{ij}),$$

where ψ is strictly increasing as a composition of strictly increasing functions. Therefore

$$s_i^\star - s_j^\star > 0 \iff \Delta_{ij} > 0 \iff h_G(m_i) > h_G(m_j).$$

Thus any minimizer (up to additive constants) induces the same strict order as h_G inside G. □

B.3 Noise Robustness (Proposition 2) — Proof

Proposition (Noise robustness (main text Proposition 2)). Let $\Delta_{ij}^G = |h_G(m_i) - h_G(m_j)|$. Suppose the learned classifier has average pairwise error ε. If we split pairs into "small-margin" ($\Delta_{ij}^G < \delta$) and "large-margin" ($\Delta_{ij}^G \geq \delta$), then the reversal probability obeys

$$\Pr[\text{reversal}] \;\leq\; \Pr[\Delta_{ij}^G < \delta] \;+\; \Pr[\text{reversal} \mid \Delta_{ij}^G \geq \delta] \;\leq\; \Pr[\Delta_{ij}^G < \delta] \;+\; \varepsilon_\delta,$$

where ε_δ decreases as δ increases and increases with the classifier error ε; in particular, under the observation model Eq. (1), the conditional flipping probability on large-margin pairs is upper-bounded by a monotonically decreasing function of δ.

Proof. Let K be the event "classifier reverses the true order". Decompose by a margin threshold $\delta > 0$:

$$\Pr[K] = \Pr[K \wedge (\Delta_{ij}^G < \delta)] + \Pr[K \wedge (\Delta_{ij}^G \geq \delta)] \leq \Pr[\Delta_{ij}^G < \delta] + \Pr[K \mid \Delta_{ij}^G \geq \delta].$$

The second term is at most the classifier's conditional error on large-margin pairs, denoted ε_δ. Under Eq. (1), the Bayes error on a pair decreases monotonically with $|\Delta_{ij}^G|$, hence ε_δ decreases in δ. If the global average error is ε, then $\varepsilon_\delta \leq \varepsilon$ and often much smaller. Thus large true gaps are stably preserved, while flips concentrate on small-margin pairs. □

B.4 From Pairwise to Group Ranking (EBC)

Given sparsity, we aggregate pairwise probabilities into a within-group ranking via Expected Borda Count (EBC): each item's score equals its expected number of wins against others according to $\hat{p}_{ij}^{G}$. EBC is a monotone transformation of the empirical pairwise preferences and inherits rank consistency in expectation when the pairwise model is consistent, providing a coherent group-wise order for evaluation and optimization. (Operational details as in Sec. 4.2.)

C Group-wise Policy Optimization (GRPO): Guarantees and Proofs

C.1 Objective and Notation

For a candidate set $\mathcal{S}_G$ with group ranking distribution q_G (from EBC), the GRPO loss is

$$\mathcal{L}_{\text{GRPO}}(\theta) = \mathbb{E}_{(I,G)}\Big[- \sum_{m_k \in \mathcal{S}_G} q_G(m_k)\, \log \pi_\theta(c_k \mid I)\Big] + \beta\, \mathbb{E}_I\big[\mathrm{KL}\big(\pi_\theta(\cdot \mid I) \,\|\, \pi_{\text{ref}}(\cdot \mid I)\big)\big].$$

Intuitively, the first term pushes π_θ toward q_G within the group (listwise), and the KL term limits drift from a safe reference policy π_{ref}; both are group-local, matching comparability in our formulation (Sec. 3).

C.2 Bounded Degradation via KL Control

We formalize the "cannot degrade beyond a bounded amount" claim under bounded KL.

Proposition (Bounded improvement under GRPO (main text Proposition 2)). Assume the reward model is rank-consistent (Proposition B.2) and $h_G \in [0,1]$. Let $\Delta_{\mathrm{KL}} = \mathbb{E}_I\big[\mathrm{KL}\big(\pi_\theta(\cdot \mid I) \,\|\, \pi_{\text{ref}}(\cdot \mid I)\big)\big]$. Then the expected within-group humor satisfies

$$\mathbb{E}_{(I,G)}\Big[\mathbb{E}_{c\sim\pi_\theta(\cdot|I)}\, h_G\big((I,c)\big)\Big] \;\geq\; \mathbb{E}_{(I,G)}\Big[\mathbb{E}_{c\sim\pi_{\text{ref}}(\cdot|I)}\, h_G\big((I,c)\big)\Big] \;-\; \sqrt{\tfrac{1}{2}\,\Delta_{\mathrm{KL}}}.$$

Consequently, if GRPO enforces $\Delta_{\mathrm{KL}} \leq \tau$ (by choosing β or an explicit trust region), the expected humor cannot drop by more than $\sqrt{\tau/2}$; with rank-consistent q_G, optimization increases the probability of higher-h_G captions, so the net effect is non-decreasing or improved expected humor once the pull toward q_G outweighs this bound.

Proof. For any fixed (I, G), Pinsker's inequality gives

$$\big\|\pi_\theta(\cdot \mid I) - \pi_{\text{ref}}(\cdot \mid I)\big\|_{\mathrm{TV}} \;\leq\; \sqrt{\tfrac{1}{2}\,\mathrm{KL}\big(\pi_\theta(\cdot \mid I) \,\|\, \pi_{\text{ref}}(\cdot \mid I)\big)}.$$

Since $h_G \in [0,1]$, by the variational characterization of total variation for bounded functions,

$$\Big|\mathbb{E}_{\pi_\theta}[h_G] - \mathbb{E}_{\pi_{\text{ref}}}[h_G]\Big| \;\leq\; \big\|\pi_\theta - \pi_{\text{ref}}\big\|_{\mathrm{TV}} \;\leq\; \sqrt{\tfrac{1}{2}\,\mathrm{KL}(\pi_\theta \| \pi_{\text{ref}})}.$$

Averaging over (I, G) yields the stated bound. During GRPO, the cross-entropy term $-\sum q_G \log \pi_\theta$ (with rank-consistent q_G) increases mass on higher-h_G captions within the group, while the KL term keeps the deviation controlled. Thus expected humor cannot deteriorate beyond the Pinsker bound and, in practice, improves as the listwise alignment progresses. □

C.3 Discussion: Why Listwise q_G Matters

Because q_G aggregates pairwise signals into a coherent group distribution consistent with h_G's ordering, the CE term directly performs a proximal step toward the better subset of captions *without* inventing any cross-group scale. This matches our problem scope and the guarantees in Sec. 4.2–4.3 of the main text.

D PAIR-WISE DATASET CONSTRUCTION

Our reward model is trained on *pairwise* comparisons. Intuitively, pairs whose ordering is both *reliably correct* and *increasingly challenging* drive the model toward more consistent ranks. We therefore construct a curriculum of **five difficulty tiers**, guaranteeing correct orderings while progressively raising difficulty (from trivial mismatches to near-ties within the same template/scene). To span both trivial and subtle distinctions, we sample pairs across all tiers and upweight harder tiers during training, yielding a supervision signal that is confident yet discriminative:

1. Wrong Text Meme ($\star$): This is the most straightforward case, where the original text is replaced with unrelated content, completely removing the humor. This type of meme is easy for the model to classify as "non-humorous" and acts as a baseline.
2. Wrong Location Meme ($\star\star$): A slightly more complex case involves shifting the position of the text in the image. While the metaphor may still exist, the humor diminishes due to the misplacement of text. The model must learn that small positional changes can significantly impact the meme's humor, reflecting a higher degree of difficulty.
3. Boring Meme ($\star\star$): Here, the meme is altered to include a more mundane or less engaging version of the original text. This teaches the model to distinguish between "humorous" and "boring" versions of the same meme. Although the content still aligns with the original, the humor is less impactful, presenting a challenge for classification.
4. Detailed Boring Meme ($\star\star\star$): This is a more nuanced case where only one or two words are changed to make the meme less funny. Despite the minimal changes, the meme's humor is significantly affected. The classifier must be able to identify these subtle shifts in humor, marking this as a more difficult classification task.
5. Generated Meme ($\star \sim \star\star\star$): Finally, memes generated by the fine-tuned VLM represent the highest difficulty level. These memes are intended to be as humorous as the original meme, requiring the classifier to discern fine-grained differences in humor between the generated meme and the original. This provides the model with an opportunity to improve its sensitivity to subtle differences in meme quality.
   By constructing a dataset with pairs of memes across these varying levels of humor, we enable the classifier to learn not only to distinguish obviously bad memes from good ones but also to understand the nuanced differences that make one meme more humorous than another. This rich dataset plays a crucial role in refining the reward model, allowing it to classify memes based on subtle human preferences.

We stratify training so each mini-batch contains an equal number from each tier.

E AUXILIARY REWARDS FOR REASONING-PATH OPTIMIZATION

While optimizing toward the group-wise reward induced by the reward model (Sec. 4.2) is theoretically sufficient to improve the quality of generated memes, the reinforcement learning stage does not directly supervise the internal reasoning path $r = (r_{\text{tmpl}}, r_{\text{scene}})$ because the primary feedback is attached to the realized meme (I, c). To explicitly shape the quality of the reasoning process itself, we introduce two auxiliary rewards that operate on r: a *format reward* and a *content reward*.

E.1 FORMAT REWARD

The format reward enforces structural completeness of the CoT to ensure that essential modules appear and are well-formed. It is computed by deterministic string/structure matching without using LLM-as-judge. Concretely, given a sampled reasoning trace r for (I, U), we check:

1. **Presence of mandatory sections** (e.g., a `Comprehensive Description` section that summarizes visual content and intended template-level intent).
2. **Two-stage structure** (explicit evidence of both template-level intent and context-level grounding consistent with Sec. 4.1).
3. **Text-on-Meme box formatting** (the `Text on the Meme` block must specify box–text mappings consistent with the bounding boxes $B = \{b_i\}$ so that rendered text $T = \{t_i\}$ aligns with B).

The format reward $R_{\text{fmt}}(r) \in [0, 1]$ is the normalized sum of satisfied checks. It shapes r toward complete and renderable reasoning without requiring any subjective judgment.

E.2 Content Reward

The content reward evaluates the informativeness and plausibility of the CoT content via an *LLM-as-judge*. We prompt an evaluation model to score r along four interpretable dimensions (e.g., visual grounding, template intent clarity, metaphorical mapping, and punchline coherence), each with discrete bands (e.g., 1/4/7 points with band descriptors such as "no object description / coarse description / detailed object attributes"). Scores are summed and rescaled to $R_{\text{cnt}}(r) \in [0, 1]$.

To calibrate the judge, we curated CoT traces spanning 0–50 points and assessed several candidates (e.g., **Qwen2.5-VL-7B** and **Keye-VL**). We found **Keye-VL** exhibits the clearest monotonic trend across bands, and thus adopt it as the judge for computing R_{cnt}. Ablations and prompt templates are released for reproducibility.

E.3 Integration with GRPO

Let $s_{\text{RM}}(m)$ denote the reward-model score that induces the group-wise ranking distribution q_G via EBC in Sec. 4.2. For a candidate set $\mathcal{S}_G = \{m_k = (I, c_k)\}$ with associated reasoning traces $\{r_k\}$, we construct an *augmented* group-wise target $\tilde{q}_G$ by combining the primary signal with auxiliary rewards on r_k:

$$\tilde{q}_G(m_k) \;\propto\; \exp\Big(\tfrac{1}{\tau}\Big[s_{\text{RM}}(m_k) \;+\; \lambda_{\text{fmt}} R_{\text{fmt}}(r_k) \;+\; \lambda_{\text{cnt}} R_{\text{cnt}}(r_k)\Big]\Big), \qquad \sum_{m_k \in \mathcal{S}_G} \tilde{q}_G(m_k) = 1, \tag{5}$$

where $\tau > 0$ is a temperature and $\lambda_{\text{fmt}}, \lambda_{\text{cnt}} \geq 0$ are weights. The GRPO objective in Eq. equation 4 is then used with q_G replaced by $\tilde{q}_G$.

Remark (Isotonic shaping and theoretical guarantees). If $(\lambda_{\text{fmt}}, \lambda_{\text{cnt}})$ are chosen such that Eq. 5 is an *isotonic* transformation of the reward-model ranking (i.e., it does not invert the order implied by s_{RM} except to break ties among near-equal items), then the rank consistency guarantees stemming from Proposition 2 are preserved in expectation. Moreover, the KL-bounded improvement in Proposition 4 continues to hold because the proof relies on boundedness of h_G and a KL constraint, both unaffected by auxiliary shaping. In practice we set $\lambda_{\text{fmt}}, \lambda_{\text{cnt}}$ small and use them primarily as tie-breakers and regularizers over r, which empirically reduces variance and accelerates convergence without altering the main ordering.

F EBC Aggregation

Definition (Expected Borda Count). Given a group G and a finite candidate set $\mathcal{S}_G = \{m_1, \ldots, m_n\}$ with pairwise preference probabilities $\widehat{p}_{ij}^{\,G} = \Pr[m_i \succ m_j]$, the Expected Borda Count of item m_i is

$$\text{EBC}_G(m_i) \;=\; \sum_{\substack{j=1 \\ j \neq i}}^{n} \widehat{p}_{ij}^{\,G}.$$

Ties or missing edges are handled by omitting terms (equivalently, treating $\widehat{p}_{ij}^{\,G}$ as undefined); in evaluation we normalize by the number of available opponents for m_i.

Basic properties. (i) If all $\widehat{p}_{ij}^{\,G} \in \{0, 1\}$, EBC reduces to the classical Borda score (number of wins). (ii) If there exists a latent utility $u : \mathcal{S}_G \to \mathbb{R}$ such that $\widehat{p}_{ij}^{\,G} = \sigma(u(m_i) - u(m_j))$ with strictly increasing σ, then sorting by EBC is order-equivalent to sorting by $\sum_{j \neq i} \sigma(u(m_i) - u(m_j))$; in particular, when gaps are consistent across pairs, the EBC order agrees with the order of u. (iii) Under independent edge noise and bounded missingness, the variance of $\text{EBC}_G(m_i)$ decreases with the number of observed pairs, making the aggregate rank more stable than any single comparison.

Listwise normalization (optional). For downstream use, one may define a soft distribution over $\mathcal{S}_G$ via a temperature $T > 0$:

$$q_G(m_i) \;=\; \frac{\exp\big(\mathrm{EBC}_G(m_i)/T\big)}{\sum_{k=1}^{n} \exp\big(\mathrm{EBC}_G(m_k)/T\big)},$$

which converts EBC scores into smooth listwise targets for within-group reweighting. This preserves the group-local nature of the signal and avoids inventing cross-group scales.

Notes on implementation. We compute $\widehat{p}_{ij}^{G}$ only within groups and on the (usually small) candidate sets used for evaluation or optimization. When the pair graph is sparse, we keep EBC unbiased by summing over observed opponents and normalizing by their count; when required, we add small-degree regularization to avoid over-confident ranks for items with very few edges.

G Training settings

G.1 CoT Supervised Fine-tuning settings

Table 3: Training Setup for Finetuning `Qwen2.5-7B-Instruct` with LoRA

Hyperparameter	Value
Finetuning Stage	`sft`
Finetuning Type	`lora`
LoRA Rank	`128`
LoRA Target	`all`
Per Device Train Batch Size	`1`
Gradient Accumulation Steps	`8`
Learning Rate	`3.0e-5`
Num Train Epochs	`5.0`
LR Scheduler Type	`cosine`
Warmup Ratio	`0.1`
bf16	`true`
Dataset	`Eimage`
Total Dataset Size	3,713 crawled memes
Training Instances	3,345
Testing Instances	368
CoT Generation Model	`doubao-1.5-vision-pro`
CoT Variants	*HUMOR-CoT*, *CoT with Single Path*, *CoT with Self-Improve*, *CoT with Subquestion*

G.2 Reward Model Training Settings

Our reward model is implemented as a lightweight extension on top of the base vision–language models. Concretely, we take the final hidden embedding of the last transformer layer and append a two-way classification head. This simple design allows the model to learn preference signals while reusing the representational power of the pretrained backbone.

Based on the dataset constructed in Appendix D, we train reward models using the *LLaMA-Factory* framework with the following backbones: *Keye-VL*, *Qwen2.5-VL-7B*, and *Qwen2.5-VL-32B*. All models are fine-tuned with LoRA (r = 8, lora target is all) to reduce memory and computation overhead. We adopt a learning rate of 1×10^{-4}, with a warmup ratio of 0.1. Each model is trained on a single NVIDIA A800 GPU.

H Evaluation settings

Evaluation Setup. For text generation, we set the decoding temperature to 0 for all models to ensure deterministic outputs. Objective textual evaluation involves two metrics: (1) **Similarity**: we

Figure 8: Template images of each rannking dataset.

extract the final meme caption via regular expressions and compute cosine similarity between generated and reference texts using *bge-base-zh-v1.5*, averaged over the full test set. (2) **Distance**: we randomly select 50 test samples, replace their user contexts with mismatched content, and regenerate three times per sample. The averaged textual dissimilarity across 50 samples is reported.

For multimodal evaluation, we embed captions into corresponding bounding boxes and obtain holistic meme-level judgments from *Gemini-2.5-pro*. We consider three perspectives: (i) human/AI discriminability, (ii) absolute scoring, and (iii) relative ranking.

VLM Scoring. Each meme is evaluated independently on an absolute 1–5 scale under the following eight criteria: 1) Punchline Strength: clarity and impact of the joke/twist; 2) Context Robustness: generalizability across social contexts; 3) Humor Effectiveness: quality of humor, sarcasm, or self-mockery; 4) Spread Potential: universal appeal and memorability; 5) Emotional Resonance: capacity to elicit laughter, surprise, or empathy; 6) Cultural Fit & Relatability: alignment with audience familiarity; 7) Theme Relevance: consistency with keywords and intentions; 8) Image-Caption Relevance: coherence between text and image.

VLM Ranking. Multiple meme candidates are jointly provided, and the model is prompted to rank them under the same eight dimensions, producing a relative quality ordering.

H.1 MaxDiff Ordering

Maximum Difference Scaling (MaxDiff), also known as best–worst scaling, is a widely used method in marketing science and preference elicitation Louviere & Woodworth (1991); Louviere et al. (2015). In a typical MaxDiff task, respondents are repeatedly presented with small subsets of items (e.g., 3–5 candidates) and asked to indicate which option they consider the ”best” and which the ”worst.” Compared to traditional rating scales, MaxDiff provides more discriminative and reliable preference estimates because each choice yields two pieces of information: a positive preference for the selected ”best” item and a negative preference for the ”worst.”

The required number of tasks in MaxDiff depends on the total number of items J to be evaluated and the subset size k. A common guideline is that each item should appear across multiple choice sets to ensure stable estimation. For example, using balanced incomplete block designs (BIBD), each respondent typically completes between $\frac{3J}{k}$ and $\frac{5J}{k}$ choice tasks to achieve acceptable reliability **?**. Thus, the total number of questions can be determined systematically to balance respondent burden and statistical efficiency.

In our study, we adopted a MaxDiff-inspired procedure to construct human preference rankings over memes. Specifically, rather than asking annotators to rate memes on absolute scales, we designed tasks where memes were compared in small groups, and annotators selected the most and least humorous instances. Aggregating these best–worst choices yields a consistent human-validated ranking dataset, which serves as a training and evaluation benchmark for our reward model.