Cross-Case Rubric Learning for Consistent Visual Preference Evaluation

Anonymous authors
Manuscript in preparation

Abstract

Multimodal large language models (MLLMs) can produce plausible rubric-based assessments of visual creatives, yet these often lack cross-case consistency and fail to translate into accurate preference evaluations. To bridge this gap, we introduce a curated dataset of 7,044 same-product image pairs, with preference labels derived directly from aggregate click-through rates (CTR) in real user exposure logs. We then propose Cross-Case Rubric Learning (CCRL), a framework that connects rubric-based visual reasoning with these behavioral outcomes. CCRL decouples preference evaluation into identifying visual differences, selecting relevant criteria, and assigning dimension-level scores that are deterministically aggregated. We optimize this scoring with a leave-one-out policy-gradient (RLOO) estimator, combining local rewards with conditional cross-case feedback. Specifically, independently assessed visual relationships act as relational constraints, penalizing inconsistent scores for equivalent visual elements across different cases. CCRL achieves the highest reported pairwise accuracy among the evaluated baselines. We assess intermediate evaluation quality using expert-reference rubric micro-F1 and full-program compliance.

1 Introduction

As generative AI lowers the barrier to producing diverse visual content, the *a priori* evaluation of user preferences remains a central challenge. While multimodal large language models (MLLMs) can produce plausible rubric-based assessments, their reasoning often lacks grounding in real-world behavioral outcomes. Evaluating a real-world creative requires assessing how varied presentation aspects jointly affect audience response. Despite complementary advances in content assessment and multimodal reasoning (Zhao et al., 2019; Wang et al., 2021; Lin et al., 2026), applying evaluation criteria coherently across diverse visual contexts—while retaining sensitivity to meaningful differences—remains difficult, especially for new creatives facing a cold-start problem.

To ground evaluation in real-world behavior, we curate a dataset of 7,044 pairs of alternative images for the same product, collected under identical time windows for random user exposures. Preference labels are derived from exact-image CTRs, filtered by minimum-exposure and CTR-gap thresholds to ensure behavioral significance. To connect these outcomes with interpretable visual reasoning, domain experts constructed a library of 54 evaluation rubrics. An expert-annotated subset of 723 pairs provides detailed rationales—including visual commonalities, differences, criterion selection, and scoring—produced with AI assistance.

However, optimizing solely for an accurate pairwise decision constrains generalization. Visual preference is driven by multiple concurrent factors; merely learning to predict the CTR winner risks rote memorization rather than acquiring transferable evaluation skills. Decomposing the evaluation into explicit rubric dimensions forces the model to articulate the reasons behind a preference shift. Yet, for this reasoning to be robust, criteria must be applied consistently: if specific visual elements remain unchanged across different creatives, their rubric contributions should remain stable regardless of unrelated contextual changes. A generalizable evaluator must disentangle and consistently attribute these factors across independent cases.

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

This raises a diagnostic question regarding standard training paradigms. As Section 4 explains, our audits reveal that while supervised fine-tuning (SFT) improves format compliance, it struggles to form coherent evaluative judgments. Furthermore, conventional per-case reinforcement learning (RL) improves single-order accuracy but frequently fails logical checks such as A/B swap invariance. These limitations stem from a structural issue: per-case binary feedback is fundamentally underdetermined, failing to fully specify how individual rubric contributions should behave across contextual changes.

While existing works provide foundational components for relational and context-sensitive reasoning (Lin et al., 2026; Zhang et al., 2026; Qiu et al., 2026; Pres et al., 2026b;a), they do not fully resolve this underdetermined evaluation. We therefore propose **Cross-Case Rubric Learning (CCRL)**, which enforces the *conditional consistency of aligned rubric contributions across independently evaluated cases*. We implement this via a relational graph where nodes represent image-pair cases and edges identify comparable criterion contributions. Crucially, the model processes only its isolated case during generation. To optimize the policy, we sample multiple independent outputs per case to form a case-by-rollout matrix. Graph edges induce candidate-to-candidate relational penalties, which are combined with local task rewards using a leave-one-out policy-gradient (RLOO) estimator, providing distinct training signals while preserving proper credit assignment.

In summary, this paper makes three complementary contributions:

- **A behavioral image dataset with expert annotations.** We curate 7,044 same-product image pairs labeled by significant CTR gaps, alongside a structured 54-rubric library and 723 pairs manually annotated with detailed expert rationales.
- **A diagnosis of the rubric evaluation gap.** We demonstrate that standard SFT and per-case RL leave criterion contributions underdetermined, frequently failing logical checks such as A/B swap invariance.
- **Cross-Case Rubric Learning (CCRL).** We introduce a graph-coupled RLOO framework enforcing consistent rubric contributions across independent cases. CCRL achieves the highest reported pairwise accuracy among evaluated baselines, with expert-reference rubric agreement, full-program compliance, and image-identity swap consistency providing complementary evaluation metrics.

2 A Dataset of Real User Feedback on Images

2.1 Data Collection and Filtering

We construct a behavioral dataset of alternative product images sourced from an e-commerce platform. The initial pool comprises 20,000 image pairs; each pair represents the same product and was exposed to users during identical time windows.

To reduce unreliable behavioral comparisons, we exclude: (1) invalid or identical decoded images; (2) pairs in which either image has fewer than 100 sampled exposures; and (3) pairs whose absolute exact-image CTR difference is less than 0.1 percentage points. We assume the retained pairs exhibit meaningful behavioral differences and use the higher observed CTR as the preference label. This filtering yields 7,044 pairs, partitioned into item-disjoint training, validation, and test splits (Table 1).

Table 1: Dataset statistics after minimum-exposure and CTR-gap filtering.

Partition	Image pairs
Train	5,626
Validation	695
Test	723
Total	7,044

2.2 Evaluation Taxonomy and Dual-Tier Annotation

To ground objective click behavior in interpretable visual reasoning, seven domain experts established a visual evaluation taxonomy. This taxonomy categorizes user preferences into 54 specific rubrics across 10 overarching aspects (e.g., *Foundational Presentation*, *Spatial Composition*).

We construct a dual-tier annotation layer based on this taxonomy to support model training and evaluation:

- **Expert Annotations (723 pairs):** Domain experts, with AI assistance, manually annotated a subset of the dataset. They identified visual commonalities and differences, selected relevant rubrics, and assigned dimension-level scores with rationales to serve as reference judgments for supervision and evaluation.
- **Machine-Generated CoT (7,044 pairs):** For scalable training, we utilized Gemini to automatically generate structured Chain-of-Thought (CoT) rationales for the full dataset. These CoTs adhere to the expert taxonomy, decomposing predictions into visual difference identification and criterion scoring.

2.3 Data Structure and Release

The dataset decouples the behavioral layer (image identities, exposure/click counts, and CTR winners) from the auxiliary reasoning layer (taxonomy, rationales, and CoTs). Target labels and behavioral metrics are isolated from the evaluator's input prompt. We will publicly release the dataset—including image features, metadata, taxonomy, and deterministic filtering scripts—to support reproducible research on visual preference alignment. Additional industry statistics, rubric coverage, and illustrative cases appear in Appendix C.

3 Problem Formulation

Input and Target. Let $x_i = (I_i^A, I_i^B, c_i)$ be an evaluation instance comprising two alternative creatives and optional metadata c_i. The objective is to predict their CTR ordering. With c_i^v and n_i^v denoting the clicks and exposures for image $v \in \{A, B\}$, the binary behavioral label $t_i \in \{0, 1\}$ is defined as:

$$t_i = \mathbb{1}[c_i^A n_i^B > c_i^B n_i^A]. \tag{1}$$

These behavioral counts are strictly unobservable during inference.

Structured Evaluator. We formulate the evaluator as an autoregressive MLLM π_θ that generates a structured reasoning path. For input x_i, the output is $y_i = (D_i, \mathcal{R}_i, \mathbf{a}_i, \mathbf{s}_i, H_i)$, where D_i describes visual differences, $\mathcal{R}_i$ is the set of selected rubrics, H_i is the textual rationale, while $\mathbf{a}_i$ and $\mathbf{s}_i$ represent the criterion applicability states and scores.

Deterministic Aggregation. To ensure traceability, the final decision is deterministically aggregated from intermediate outputs. For a criterion $r \in \mathcal{R}_i$ on image v, given its applicability $a_{ir}^v \in \{0, 1\}$ and scalar score s_{ir}^v, the net contribution is $u_{ir}^v = a_{ir}^v s_{ir}^v$. The overall margin is $d_i = \sum_{r \in \mathcal{R}_i} \Delta_{ir}$, where $\Delta_{ir} = u_{ir}^A - u_{ir}^B$. The prediction follows the sign of d_i. Crucially, to maintain logical consistency under an A/B position swap, a robust evaluator must preserve the selected rubrics $\mathcal{R}_i$ while inverting the sign of the comparative scores ($\Delta_{ir}(x_i^{\text{swap}}) = -\Delta_{ir}(x_i)$).

4 Diagnosing the Evaluation Gap: The Limits of Local Supervision

A single A/B preference label supervises only the final holistic outcome, leaving the intermediate reasoning chain—inferring visual facts, selecting criteria, and scoring—loosely constrained. Before introducing our methodology, we analyze the fundamental limitations of standard local training objectives, such as supervised fine-tuning (SFT) and per-case reinforcement learning (RL).

4.1 Empirical Observations from Local Training

Our audits of standard local optimization methods reveal a persistent misalignment between generating executable formatting and forming robust evaluative judgments.

SFT improves output validity, but full evaluation compliance remains limited. On the 723-pair evaluation cohort, SFT achieves 96.96% valid output, compared with 95.16% for the zero-shot model. Its final-label accuracy is 52.14%, full-program compliance is 20.33%, and rubric micro-F1 is 34.30% (Table 4). These metrics separate producing a valid output from satisfying the complete evaluation rules and selecting criteria that agree with expert references. A correct final label does not require the entire intermediate evaluation to pass all compliance checks.

Local RL improves single-order accuracy without preserving swap consistency. Local-RLOO increases original-order accuracy from SFT's 52.14% to 56.71%, while image-identity consistency under an A/B swap decreases from 29.32% to 27.24% (Table 3). Swap-RLOO partially recovers consistency to 28.35%. Thus, better final-label accuracy in the original order does not by itself ensure that the evaluator selects the same underlying image after the order is reversed.

4.2 The Underdetermined Nature of Local Supervision

These empirical failures reflect fundamental structural vulnerabilities in isolated per-case supervision, which we categorize into two primary modes:

V1: Reward Equivalence and Compensating Errors. Recall from Section 3 that the final preference margin is $d_i = \sum_{r \in \mathcal{R}_i} \Delta_{ir}$, where $\Delta_{ir} = u_{ir}^A - u_{ir}^B$. For any non-zero perturbation ϵ, compensating errors leave the final margin unchanged:

$$(\Delta_{ir}, \Delta_{is}) \mapsto (\Delta_{ir} + \epsilon, \Delta_{is} - \epsilon) \implies d_i' = d_i. \tag{2}$$

A local outcome reward based on d_i cannot distinguish a correct evaluation from one where a hallucinated advantage in criterion r masks a missed deduction in criterion s. Consequently, criterion-level credit assignment remains mathematically underdetermined.

V2: Cross-Case Instability and Unanchored Reuse. The A/B swap failure is a localized symptom of unanchored criterion reuse. A model may assign a high score to a specific visual element in one instance but contradict itself on the identical element in another case due to irrelevant background changes. Local RL, operating strictly within isolated cases x_i, provides no gradient signal to penalize these cross-case contradictions.

4.3 Motivating Relational Supervision

Because distinct criterion-level decompositions can produce the same preference margin, outcome-only supervision is ambiguous. While local outcome rewards are necessary, they are insufficient to constrain intermediate judgments; such constraints must be supplied explicitly.

Cross-case relations provide this missing structural supervision. By defining when a criterion should remain applicable and when an aligned judgment should remain stable despite irrelevant contextual changes, these relations regularize the evaluation process. This motivates our proposed framework, which leverages cross-case consistency to resolve the ambiguity of local supervision.

5 Cross-Case Rubric Learning (CCRL)

To address the underdetermined nature of local feedback, CCRL explicitly injects independently assessed visual relations into the training objective. We formulate this as a graph-coupled policy gradient framework, where cross-case relational penalties structurally regularize the intermediate rubric scores.

5.1 Relational Graph over Aligned Contributions

We construct a training block $B = (X, E)$ comprising a batch of evaluation cases $X = \{x_1, \ldots, x_M\}$ and a set of relational edges E. Each edge $e \in E$ specifies a constraint between two rubric contributions originating from different cases, denoted by endpoints (i, v, r) and (j, v', r'). These edges are firmly grounded in visual evidence (e.g., a shared background, an identical subject, or an unchanged promotional badge), dictating how specific dimension-level scores should behave relative to one another.

Table 2 summarizes the primary relation types. Crucially, this graph is constructed *a priori* based on independent visual assessments (e.g., via a frozen vision-language prior or human annotation). This ensures that the relational constraints remain strictly decoupled from both the actor's generated scores and the final CTR labels.

Relation Type	**Constraint Definition**	**Visual Rationale & Example**
Same Conditions	Aligned contributions must stay within an assessed tolerance τ_e.	Visually identical product elements should yield stable utility scores.
Irrelevant Change	A contribution must remain stable when its relevant conditions are unchanged.	Changing a distant region while preserving a badge's size and contrast should preserve its recognizability.
Local Change	Assessed unaffected contributions remain stable; affected ones may vary.	Enlarging text alters readability but not necessarily the overall product aesthetic.
Gate Transition	Applicability strictly matches visual evidence; inactive contributions are zero.	Unreadable fine print provides zero informational utility.

Table 2: Taxonomy of conditional cross-case relations. These constraints dictate the stability and permitted variations of intermediate rubric scores based on observable visual evidence.

5.2 Local Rewards and Cross-Case Penalties

For each generated reasoning path y_i, we define a composite reward function. First, the local objective evaluates the fidelity of the final decision to the behavioral target t_i. We map the aggregated preference margin d_i (Section 3) to a probability $p_i = \sigma(d_i/T)$ with temperature $T > 0$. The local reward $U_i(y_i)$ combines a soft predictive penalty with a structural grounding penalty $G_i \in [-1, 0]$ (penalizing hallucinations or formatting errors):

$$U_i(y_i) = -(p_i - t_i)^2 + \eta G_i(y_i). \tag{3}$$

Second, we enforce the cross-case graph constraints. For an equality-type edge $e = (i, j)$ linking contributions u^v_{ir} and $u^{v'}_{jr'}$, the relational penalty $C_e(y_i, y_j) \in [-1, 0]$ is defined as a bounded hinge loss:

$$C_e(y_i, y_j) = -\min\left\{1, \frac{\left[|u^v_{ir} - u^{v'}_{jr'}| - \tau_e\right]_+}{2S_{\max}}\right\}, \tag{4}$$

where $\tau_e \geq 0$ denotes the assessed tolerance. We bound the per-image criterion scores s^v_{ir} to $[-3, 3]$ and set $S_{\max} = 3$. Since $u^v_{ir} = a^v_{ir} s^v_{ir}$ with binary applicability, contributions also lie in $[-3, 3]$; hence $2S_{\max} = 6$ normalizes their maximum possible absolute difference. If a required visual condition is missing or the output syntax is invalid, the penalty defaults to -1. A zero penalty indicates that the logical constraint is perfectly satisfied.

The block return combines local rewards and relational penalties using fixed node weights α_i and edge weights b_e. We optimize its per-case normalized expectation:

$$F_B(Y) = \sum_{i=1}^{M} \alpha_i U_i(y_i) + \lambda \sum_{e=(i,j)\in E} b_e C_e(y_i, y_j), \tag{5}$$

$$J(\theta) = \mathbb{E}_{B\sim P_B} \mathbb{E}_{Y\sim \prod_i \pi_\theta(\cdot|x_i)}[F_B(Y)/M]. \tag{6}$$

5.3 Graph-Coupled Leave-One-Out Policy Gradient (RLOO)

To optimize Equation 6 while maintaining proper case-level credit assignment, we employ a case-by-rollout matrix formulation (Pres et al., 2026b). For each case i, we independently sample $K > 1$ reasoning paths $\{y_{ik}\}_{k=1}^K$ from the current policy π_{old}. During generation, cases are processed in strict isolation.

For every edge $e = (i, j)$, the interactions between the K samples of case i and the K samples of case j form a $K \times K$ relational penalty matrix, where the entry at (k, ℓ) is $C_e(y_{ik}, y_{j\ell})$. The expected

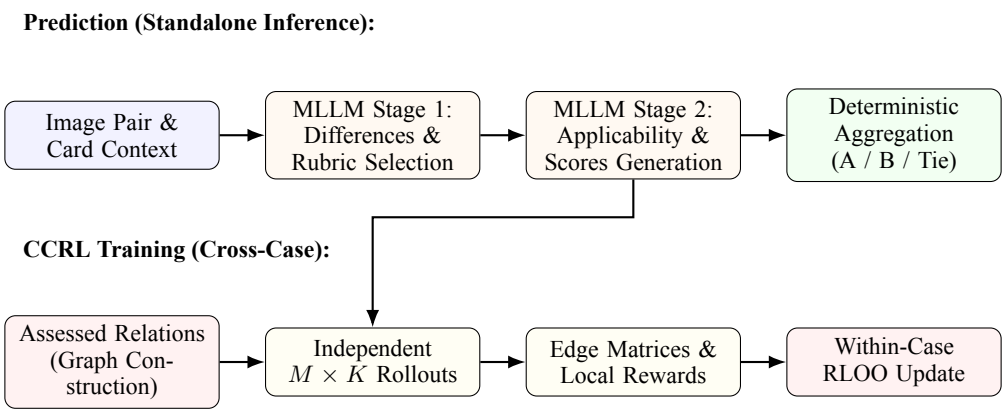

Figure 1: System Architecture of CCRL. **Top (Inference):** The standalone evaluation interface takes paired images and context, generates visual differences and criterion-level judgments via the MLLM, and deterministically computes the pairwise preference. **Bottom (Training):** Independently assessed relations form a cross-case graph. $M \times K$ independent rollouts populate the $K \times K$ edge matrices to compute relational penalties, yielding row/column returns for the leave-one-out (RLOO) policy gradient update.

relational return for candidate k in case i is estimated by averaging over the K independently sampled neighbors in case j. Combining this with the local reward, the total estimated return for candidate k is:

$$\widehat{q}_{ik} = \alpha_i U_i(y_{ik}) + \lambda \sum_{e \ni i} \frac{b_e}{K} \sum_{\ell=1}^{K} C_e(y_{ik}, y_{j\ell}). \tag{7}$$

To center these returns and reduce variance, we apply a Leave-One-Out (RLOO) baseline (Ahmadian et al., 2024). The advantage A_{ik} for candidate y_{ik} is computed by subtracting the mean return of its *siblings* (the other $K-1$ rollouts from the same case):

$$A_{ik} = \widehat{q}_{ik} - \frac{1}{K-1} \sum_{h \neq k} \widehat{q}_{ih}. \tag{8}$$

Conditional on the neighbor samples, this leave-one-out baseline is independent of candidate y_{ik}. The resulting policy-gradient estimator is unbiased for the normalized objective in Equation 6 at the sampling policy (see Appendix A for the proof).

We use an on-policy update with fresh rollouts for each optimizer step. Treating the sampled advantages A_{ik} as fixed constants, we optimize the surrogate objective by aggregating token-level log probabilities over each generated sequence $\mathcal{T}_{ik}$:

$$\mathcal{L}_{\text{PG}} = -\frac{1}{M \cdot K} \sum_{i=1}^{M} \sum_{k=1}^{K} A_{ik} \sum_{t \in \mathcal{T}_{ik}} \log \pi_\theta(y_{ikt} \mid x_i, y_{ik,<t}). \tag{9}$$

6 Experiments

We organize our evaluation around three primary research questions:

- **RQ0 (Failure Diagnosis):** What vulnerabilities remain after SFT and local RL when measuring accuracy and swap consistency?
- **RQ1 (Preference Prediction):** Does relational training across cases improve overall CTR ordering prediction over local-only RL?
- **RQ2 (Evaluation Process):** How well do model criteria align with expert annotations, and how often does the reasoning pass programmatic checks?

6.1 Experimental Setup

Dataset and Metrics. We evaluate all methods on the 723-pair annotated cohort (Section 2). **Accuracy** measures if the final prediction matches the observed CTR winner. **Valid Output** checks syntactic formatting compliance. **Full-Program Compliance** ensures the complete evaluation satisfies all intermediate programmatic checks (e.g., rubric applicability rules); a correct final label does not guarantee full compliance. **Rubric micro-F1** measures agreement between predicted and expert-selected criteria, pooled across all evaluation pairs. For diagnostic purposes, **Identity Consistency** measures the fraction of pairs where predictions across an A/B swap select the identical underlying image.

Implementation Details. The actor π_θ, initialized from Qwen3.5-27B, samples $K = 8$ independent outputs per case. We set the reward temperature to $T = 1.0$ and optimize the RLOO objective (Equation 9) with a learning rate of 1×10^{-6} for 3 epochs.

Baselines. We compare our approach against:

1. **Zero-shot MLLM:** The base model evaluated with a zero-shot prompt.
2. **Creative4U (Lin et al., 2026):** A strong multimodal baseline explicitly rewarding comparative reasoning.
3. **SFT:** The model fine-tuned on versioned reference responses.
4. **Local-GRPO:** The actor trained with local rewards using within-case group-relative normalization, without cross-case feedback.
5. **Local-RLOO:** The actor trained using the RLOO algorithm with only local task and grounding rewards ($\lambda = 0$ in Equation 5).
6. **Ours (Graph-RLOO):** The full framework incorporating both local rewards and cross-case relational penalties (Equation 7).

6.2 RQ0: Failure Diagnosis and Swap Consistency

To empirically validate the structural vulnerabilities identified in Section 4, we isolate the effect of A/B positional swapping. A robust evaluator must maintain identity consistency—selecting the same underlying image regardless of display order.

As Table 3 shows, while Local-RLOO improves original-order accuracy over SFT (52.14% to 56.71%), its swapped-order accuracy drops to 32.23%, and identity consistency falls to 27.24%. Adding a direct swap penalty (Swap-RLOO) slightly recovers consistency to 28.35%, but it remains below the SFT baseline. This confirms that local training and simple outcome penalties fail to eliminate order sensitivity, motivating the need for structural cross-case constraints.

Table 3: A/B swap diagnosis on 723 image pairs. Identity consistency measures if both display orders yield the same underlying image choice. Swap-RLOO adds a swap-consistency penalty to local training.

Method	Original-order accuracy	Swapped-order accuracy	Identity consistency
SFT	0.5214	0.3333	0.2932
Local-RLOO	0.5671	0.3223	0.2724
Swap-RLOO	0.5726	0.3306	0.2835

6.3 Main Results (RQ1 & RQ2)

Table 4 summarizes the performance across all metrics.

RQ1 (Preference Prediction): Graph-RLOO (Ours) achieves the highest final-label accuracy at 58.37%, outperforming the Zero-shot MLLM (41.08%) and Creative4U (49.38%). CCRL improves upon the Local-RLOO baseline by 1.66 percentage points and standard SFT by 6.23 points, demonstrating the benefit of cross-case relational training for objective CTR prediction.

RQ2 (Evaluation Process): Beyond final accuracy, CCRL improves intermediate reasoning quality. It achieves a Rubric micro-F1 of 37.21% (compared to 35.27% for Local-RLOO and 34.30% for SFT), indicating stronger alignment with expert visual assessments. Furthermore, CCRL reaches 100.00% format validity and 25.17% full-program compliance. While the low absolute compliance rate highlights the difficulty of flawless multimodal reasoning, Graph-RLOO still yields the highest complete-check pass rate among all configurations, verifying that cross-case supervision concurrently benefits both behavioral prediction and structural consistency.

Table 4: Main results on 723 image pairs. Accuracy targets the final predicted label; full-program compliance requires passing all intermediate programmatic checks. Rubric F1 is micro-averaged over criterion selections. Dashes denote unreported metrics.

Method	Training objective	Accuracy	Valid output	Full-program compliance	Rubric micro-F1
Zero-shot MLLM	None	0.4108	0.9516	0.1521	0.2174
Creative4U	–	0.4938	–	–	–
SFT	Cross-Entropy	0.5214	0.9696	0.2033	0.3430
Local-GRPO	Local Eq. 3	0.5532	0.9820	0.2061	0.3541
Local-RLOO	Local Eq. 3	0.5671	0.9917	0.2448	0.3527
Ours	Graph Eq. 5	**0.5837**	**1.0000**	**0.2517**	**0.3721**

7 Related Work

7.1 Personalized User Simulation with Large Language Models

LLM-based simulation explores whether general-purpose models can replicate distinct user behaviors. Frameworks like RecAgent and Agent4Rec construct interactive agents for recommendation environments (Wang et al., 2024; Zhang et al., 2023a). SimUSER models behaviors based on interaction profiles to study visual presentation effects (Bougie & Watanabe, 2025), while MEDEA employs Social-CoT to simulate diverse audience reactions (Li et al., 2026). While these studies successfully connect text reasoning to behavioral assessment, they typically entangle visual evaluation with personalized demographic simulation. In contrast, our formulation grounds evaluation in population-level CTR, aggregated over random exposures to same-product alternatives within matched time windows. This provides a practical, robust starting point for visual preference evaluation by reducing sensitivity to individual demographic variations. Because aggregation alone does not guarantee demographic balance, exploring personalized conditioning for cohort-specific predictions remains a complementary direction.

7.2 Rubric-Based Preference Evaluation and Consistency Learning

Rubrics enhance interpretability by exposing the specific criteria underlying a holistic evaluation. Recent approaches generate dynamic or personalized scoring schemes for text evaluation, as demonstrated by PREFINE (Ueda & Takayanagi, 2025), P-GenRM (Zhang et al., 2026), and PARL (Qiu et al., 2026). Rubric-ARM further investigates the alternating optimization of rubric generation and response judging (Xu et al., 2026). While these methods establish context-sensitive criteria, ensuring their numerical stability across distinct cases remains an open challenge, motivating explicit process supervision (Lightman et al., 2023). Self-CTRL represents a key consistency-learning predecessor, aligning independently generated self-explanations and behaviors across related contexts (Pres et al., 2026b). Broader self-consistency frameworks (Pres et al., 2026a) and testing paradigms like CheckList (Ribeiro et al., 2020) similarly formalize invariant behavioral expectations. Our work applies these consistency concepts to conditional alignment of rubric contributions across visual evaluation cases. We optimize grounded cross-case relations using RLOO for within-case baseline centering (Ahmadian et al., 2024), distinguishing our approach from purely local baselines like GRPO (Shao et al., 2024) or off-policy surrogates like PPO (Schulman et al., 2017) (discussed further in Appendix B).

7.3 Visual Preference Prediction and Creative Selection

Selecting optimal creatives under sparse feedback remains a key challenge in online deployment. Early methods predict perceptual quality distributions (Talebi & Milanfar, 2018) or leverage deep content features (Zhao et al., 2019). For online exploration, AES integrates compositional trees with efficient bandits (Chen et al., 2021), while HBM-VAM dynamically updates visual priors using observed feedback (Wang et al., 2021). Parallel Ranking advances creative serving by introducing joint offline optimization under strict latency constraints (Yang et al., 2023). Our closest task-level precedent is Creative4U, which applies multimodal comparative reasoning to creative selection using GRPO-RPR to reward dimensional recall and precision (Lin et al., 2026). While Creative4U optimizes comparative rationales via intra-case feedback, our framework directly resolves the inherent underdetermination of local supervision. By deterministically computing predictions from intermediate criterion scores, we expose the decision process to explicit cross-case relational regularization.

8 Limitations and Scope

While CCRL provides a structured mechanism to enforce cross-case consistency, it relies on specific methodological assumptions. First, conditional consistency alone does not guarantee correctness; a degenerate model could theoretically satisfy the graph by assigning constant scores across all cases. We mitigate this by coupling the relational graph with local grounding rewards, though the risk of uninformative convergence remains if edge weights are improperly balanced. Second, relying on a frozen vision-language prior to construct the relational graph may propagate systematic biases into the actor, as visual similarity does not strictly guarantee equal subjective utility.

Furthermore, while aggregating CTR over matched time windows reduces demographic noise, population averages can obscure heterogeneous preferences. A visual element deemed overly complex by one user cohort might be highly informative to another. Thus, personalized conditioning could improve predictions by adapting rubric selection and contribution weights to specific audiences. Future work will extend this framework toward audience-aware evaluation and user simulation.

9 Conclusion

In this work, we identified and addressed a fundamental bottleneck in multimodal preference alignment: the underdetermined nature of isolated, per-case supervision. We introduced a dataset of 7,044 same-product image pairs labeled by observed CTR ordering, accompanied by a 54-rubric taxonomy and AI-assisted expert annotations. Our empirical diagnosis separates output validity, final-label accuracy, complete evaluation compliance, and identity consistency under an A/B swap. Improved original-order accuracy does not necessarily preserve identity consistency.

To resolve this, we proposed CCRL. By coupling an explicit cross-case penalty graph with an RLOO estimator, CCRL encourages the model to assign stable, grounded scores to visual elements across independent cases. The framework achieves the highest reported final-label accuracy and rubric micro-F1 among the evaluated configurations. Full-program compliance provides a complementary process measure and remains limited, indicating room for more reliable intermediate evaluations. Ultimately, CCRL offers a principled, testable framework for aligning the complex visual reasoning of generative models with objective behavioral outcomes.

AI Assistance

Large language models and other AI tools assisted with literature retrieval, research ideation, drafting sections of the manuscript, and language editing. AI assistance also supported code development, data formatting, annotation record analysis, and the preparation of figures and LaTeX material.

For the 723-pair expert-annotated subset, domain experts performed manual annotation with AI assistance. AI tools supported drafting candidate criteria, scores, and explanations; the experts made the final annotation decisions and manually revised the resulting records. These expert annotations are distinct from the automatically generated Gemini CoTs covering the full dataset. Product images and observed click/exposure logs represent raw behavioral data, not AI-generated content. The au-

thors take full responsibility for the final manuscript and for verifying all AI-assisted text, code, data processing, and analyses.

References

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to Basics: Revisiting REINFORCE Style Optimization for Learning from Human Feedback in LLMs. *arXiv preprint arXiv:2402.14740*, 2024. URL https://arxiv.org/abs/2402.14740.

Nicolas Bougie and Narimasa Watanabe. SimUSER: Simulating User Behavior with Large Language Models for Recommender System Evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pp. 43–60, 2025. doi: 10.18653/v1/2025.acl-industry.5. URL https://aclanthology.org/2025.acl-industry.5/.

Jin Chen, Tiezheng Ge, Gangwei Jiang, Zhiqiang Zhang, Defu Lian, and Kai Zheng. Efficient Optimal Selection for Composited Advertising Creatives with Tree Structure. *arXiv preprint arXiv:2103.01453*, 2021. URL https://arxiv.org/abs/2103.01453.

Tianjiao Li, Kai Zhao, Xiang Li, Yang Liu, and Huyang Sun. Community-Aware Assessment of Social Textual Engagement and Resonance: A Human-Centric Perspective on User-Generated Content Evaluation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21578–21600, 2026. doi: 10.18653/v1/2026.acl-long.985. URL https://aclanthology.org/2026.acl-long.985/.

Jiayi Liao, Sihang Li, Zhengyi Yang, Jiancan Wu, Yancheng Yuan, Xiang Wang, and Xiangnan He. LLaRA: Large Language-Recommendation Assistant. *arXiv preprint arXiv:2312.02445*, 2024. URL https://arxiv.org/abs/2312.02445.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's Verify Step by Step. *arXiv preprint arXiv:2305.20050*, 2023. URL https://arxiv.org/abs/2305.20050.

Yukang Lin, Xiao Zhang, Xiang Zhang, Shichang Jia, Bowen Wan, Chenghan Fu, Xudong Ren, Yueran Liu, Wanxian Guan, Pengji Wang, Chuan Yu, and Baolin Liu. Creative4U: MLLMs-based Advertising Creative Image Selector with Comparative Reasoning. *arXiv preprint arXiv:2508.12628v2*, 2026. URL https://arxiv.org/abs/2508.12628v2.

Itamar Pres, Belinda Z. Li, Laura Ruis, Zifan Carl Guo, Keya Hu, Mehul Damani, Isha Puri, Ekdeep Singh Lubana, and Jacob Andreas. Position: It's Time to Optimize LLMs for Self-Consistency. In *International Conference on Machine Learning, Position Paper Track*, 2026a. URL https://arxiv.org/abs/2608.05188.

Itamar Pres, Laura Ruis, Melat Ghebreselassie, Belinda Z. Li, and Jacob Andreas. Self-CTRL: Self-Consistency Training with Reinforcement Learning. *arXiv preprint arXiv:2606.18327*, 2026b. URL https://arxiv.org/abs/2606.18327.

Yilun Qiu, Xiaoyan Zhao, Yang Zhang, Yuxin Chen, Cilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, Yoko Yamakata, and Tat-Seng Chua. Preference-Aware Rubric Learning for Personalized Evaluation. *arXiv preprint arXiv:2605.31545*, 2026. URL https://arxiv.org/abs/2605.31545.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, 2020. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442/.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Hossein Talebi and Peyman Milanfar. NIMA: Neural Image Assessment. *IEEE Transactions on Image Processing*, 27(8):3998–4011, 2018. doi: 10.1109/TIP.2018.2831899.

Kentaro Ueda and Takehiro Takayanagi. PREFINE: Personalized Story Generation via Simulated User Critics and User-Specific Rubric Generation. *arXiv preprint arXiv:2510.21721*, 2025. URL https://arxiv.org/abs/2510.21721.

Lei Wang, Jingsen Zhang, Hao Yang, Zhiyuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Ruihua Song, Wayne Xin Zhao, Jun Xu, Zhicheng Dou, Jun Wang, and Ji-Rong Wen. User Behavior Simulation with Large Language Model based Agents. *arXiv preprint arXiv:2306.02552v3*, 2024. URL https://arxiv.org/abs/2306.02552v3.

Shiyao Wang, Qi Liu, Tiezheng Ge, Defu Lian, and Zhiqiang Zhang. A Hybrid Bandit Model with Visual Priors for Creative Ranking in Display Advertising. *arXiv preprint arXiv:2102.04033*, 2021. URL https://arxiv.org/abs/2102.04033.

Ran Xu, Tianci Liu, Zihan Dong, Tony Yu, Ilgee Hong, Carl Yang, Linjun Zhang, Tuo Zhao, and Haoyu Wang. Alternating Reinforcement Learning for Rubric-Based Reward Modeling in Non-Verifiable LLM Post-Training. *arXiv preprint arXiv:2602.01511*, 2026. URL https://arxiv.org/abs/2602.01511.

Zhiguang Yang, Lu Wang, Chun Gan, Liufang Sang, Haoran Wang, Wenlong Chen, Jie He, Changping Peng, Zhangang Lin, and Jingping Shao. Parallel Ranking of Ads and Creatives in Real-Time Advertising Systems. *arXiv preprint arXiv:2312.12750*, 2023. URL https://arxiv.org/abs/2312.12750.

An Zhang, Yuxin Chen, Leheng Sheng, Xiang Wang, and Tat-Seng Chua. On Generative Agents in Recommendation. *arXiv preprint arXiv:2310.10108*, 2023a. URL https://arxiv.org/abs/2310.10108.

Pinyi Zhang, Ting-En Lin, Yuchuan Wu, Jingyang Chen, Zongqi Wang, Hua Yang, Ze Xu, Fei Huang, Kai Zhang, and Yongbin Li. P-GenRM: Personalized Generative Reward Model with Test-time User-based Scaling. *arXiv preprint arXiv:2602.12116*, 2026. URL https://arxiv.org/abs/2602.12116.

Yang Zhang, Fuli Feng, Jizhi Zhang, Keqin Bao, Qifan Wang, and Xiangnan He. CoLLM: Integrating Collaborative Embeddings into Large Language Models for Recommendation. *arXiv preprint arXiv:2310.19488*, 2023b. URL https://arxiv.org/abs/2310.19488.

Zhichen Zhao, Lei Li, Bowen Zhang, Meng Wang, Yuning Jiang, Li Xu, Fengkun Wang, and Wei-Ying Ma. What You Look Matters? Offline Evaluation of Advertising Creatives for Cold-start Problem. In *Proceedings of CIKM*, 2019. doi: 10.1145/3357384.3357813.

A Derivation of the Cross-Case RLOO Estimator

In this section, we prove that the graph-coupled Leave-One-Out (RLOO) estimator provides an unbiased gradient for the cross-case relational objective at the sampling policy.

Consider a training block $B = (X, E)$ comprising M cases, its associated graph, fixed edge weights, and reward functions. Let the block objective be $J_B = \mathbb{E}_Y[F_B(Y)/M]$. We assume the following standard regularity conditions: (1) the block sampling distribution P_B is independent of the current policy parameters θ; (2) candidate outputs are sampled independently across cases and are conditionally independent given the case input x_i; and (3) the reward functions lack direct differentiable dependence on θ. Under these conditions, we can exchange differentiation and expectation:

$$\nabla_\theta J_B = \frac{1}{M}\mathbb{E}_Y\left[F_B(Y)\sum_i \nabla_\theta \log \pi_\theta(y_i \mid x_i)\right]. \tag{10}$$

Let $q_i(Y) = \alpha_i U_i(y_i) + \lambda \sum_{e \ni i} b_e C_e(y_i, y_j)$ collect the local reward and incident edge penalties for case i, where j denotes the other endpoint (maintaining the edge's fixed endpoint order). We can decompose the objective as $F_B(Y) = q_i(Y) + F_{-i}(Y)$, where F_{-i} contains only terms that do not involve case i. For a fixed set of responses from all other cases Y_{-i}, the expected product of $F_{-i}(Y)$ with the score function of case i is identically zero:

$$\mathbb{E}_{y_i}\left[F_{-i}\nabla_\theta \log \pi_\theta(y_i \mid x_i)\right] = F_{-i}\nabla_\theta \sum_{y_i} \pi_\theta(y_i \mid x_i) = 0. \tag{11}$$

Therefore, only the incident return contributes to each case's score-function term:

$$\nabla_\theta J_B = \frac{1}{M}\mathbb{E}_Y\left[\sum_i q_i(Y)\nabla_\theta \log \pi_\theta(y_i \mid x_i)\right]. \tag{12}$$

For each generated candidate y_{ik}, averaging the pairwise penalties over the K independent samples of its neighbors in $\widehat{q}_{ik}$ provides a Monte Carlo estimate of its incident relational expectation. Crucially, the leave-one-out baseline for candidate k is computed strictly from the other $K-1$ independent rollouts within the same case row. Conditional on the neighbor samples from other cases, this baseline is completely independent of y_{ik}. Consequently, its expected product with the score function of y_{ik} is strictly zero. Averaging over the K candidates and M cases, and then over sampled blocks, yields an unbiased estimator of the gradient of the normalized objective in Equation 6 at the sampling policy. Note that shared model parameters do not violate the conditional independence of the sampled outputs within a case.

B Off-Policy Limits and Sequence Likelihoods

A fundamental property of the cross-case graph is that evaluating exact importance sampling for off-policy updates requires the joint likelihood of multiple endpoints. Let $\rho_i = \pi_\theta(y_i \mid x_i)/\pi_{\text{old}}(y_i \mid x_i)$ denote the sequence-level likelihood ratio. Factoring the objective yields:

$$J_B(\theta) = \frac{1}{M}\mathbb{E}_{Y\sim\pi_{\text{old}}}\left[\sum_i \alpha_i\rho_i U_i(y_i) + \lambda \sum_{e=(i,j)} b_e\rho_i\rho_j C_e(y_i, y_j)\right]. \tag{13}$$

While a fully factored graph does not require the product of all M ratios, updating an edge penalty strictly requires the likelihood ratios of both endpoints ($\rho_i\rho_j$). In contrast, standard clipped PPO surrogates typically apply token-level clipping independently of neighbor trajectory probabilities. A PPO-style surrogate with fixed sampled advantages matches the first-order gradient at $\theta = \theta_{\text{old}}$ only when its likelihood-ratio derivative matches the sequence score function. Away from that point, omitting the neighbor ratio generally yields a biased surrogate for the coupled objective. This theoretical limitation motivates our implementation choice to use fresh rollout batches (on-policy) for each optimizer step, thereby avoiding complex cross-case importance weighting.

C Dataset Composition and Rubric Taxonomy

C.1 Industry Composition

Figure 2 summarizes the behavioral dataset of 7,044 same-product image pairs. The 92 original product categories are systematically grouped into 16 mutually exclusive display industries to provide a macroscopic view of the dataset's diversity. The inclusion rule requires at least 100 exact-image sampled exposures for each image and an absolute CTR difference of at least 0.1 percentage points. These thresholds define the behavioral contrast used to retain pairs, consistent with Section 2.

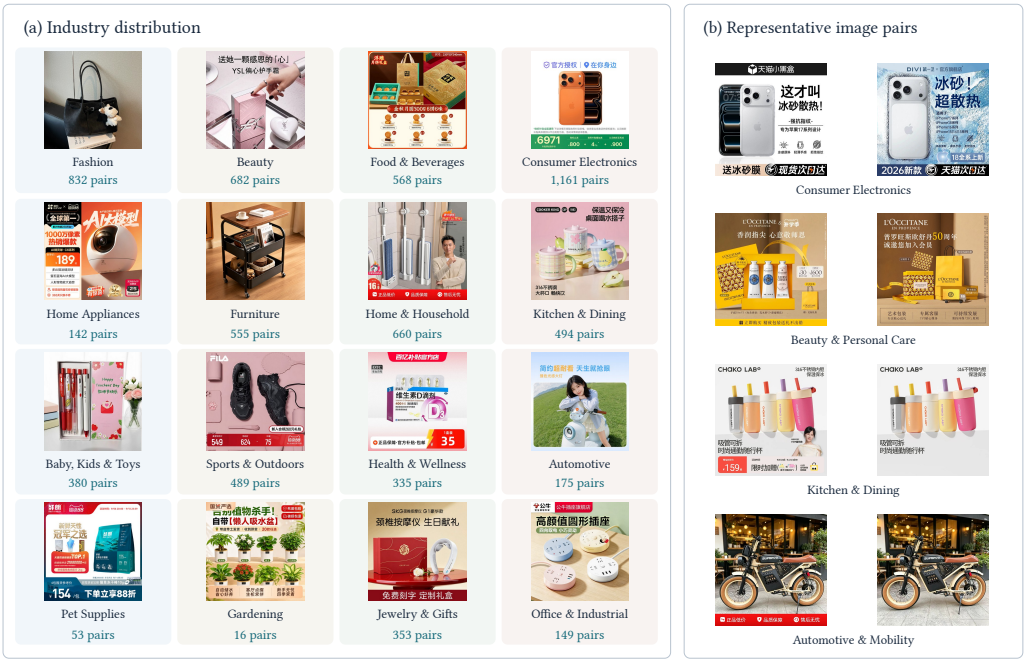

Figure 2: Industry composition and representative A/B image pairs in the 7,044-pair dataset. The right panel illustrates alternative images for the same product, retaining original in-image text and promotional cues.

C.2 Expert Rubric Coverage and Distribution

To anchor the evaluation framework, we utilized a highly curated 723-pair annotated subset. The comprehensive expert taxonomy consists of 54 distinct visual identifiers (R01–R54), categorized into ten overarching visual-content families (Figure 3). The "Marketing and Price" family is the most frequently selected family in the currently tabulated expert references (335/723, or 46.3% of the full annotated subset), followed by "Specifications and Purchase Form" and "Brand, Origin, and Campaigns". This distribution summarizes the criteria emphasized in expert annotations: informational and promotional clarity appear alongside aesthetic considerations. It describes expert judgments rather than separately estimated effects of each criterion on CTR.

C.3 Illustrative Image-Pair Cases

Figure 4 provides concrete examples of the dimension-level scoring process. The rationales decouple the holistic preference into specific signed scores (e.g., scoring the visibility of product details independently from celebrity familiarity), illustrating the fine-grained visual evidence required by our cross-case relational graph.

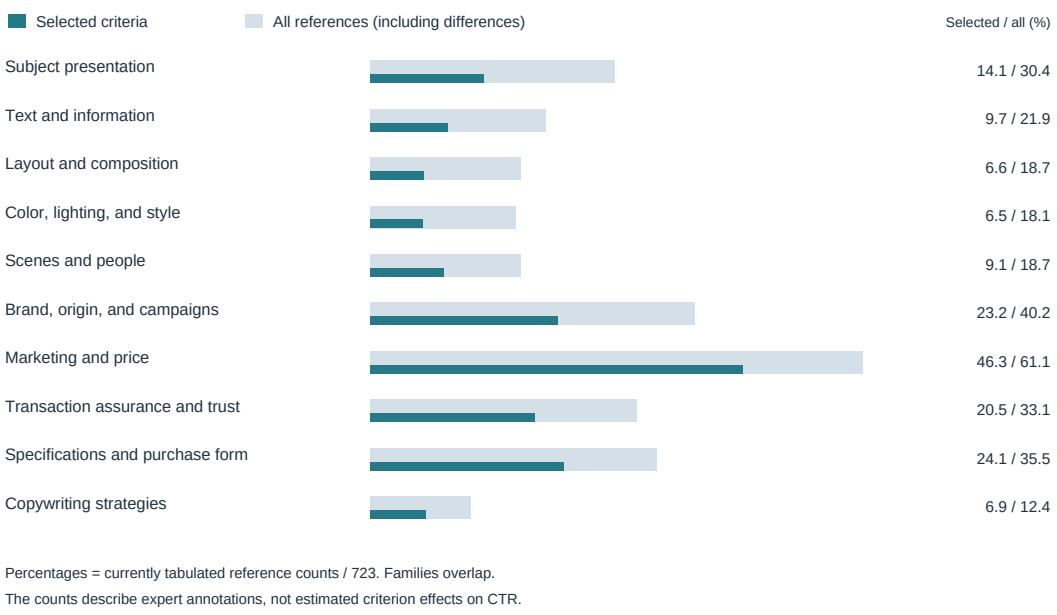

Figure 3: Rubric-family coverage normalized by the 723-pair annotated subset. Dark bars show expert-selected criteria used in the annotated judgments; light bars show all criterion references, including those in visual difference descriptions. Percentages use the currently tabulated reference counts, keeping the denominator of 723 constant across all families.

A



B



(a) Beauty: color and contrast.

R13: Color harmony and visual contrast (-3). The review favors A's background color scheme over B's black-and-white presentation.

Observed winner: **A**.

A



B



(b) Kitchen supplies: details and use.

R04: Visibility of product details (+2). B's close-up makes the lid, latch, and dial easier to inspect.

R21: Celebrity or IP familiarity (+2). The review identifies B's endorser portrait as an attention and trust cue.

R18: Usage and function demonstrations (+2). B shows a hand opening the lid, making its operation more explicit.

Observed winner: **B**.

A



B



(c) Mobility: specifications and function.

R41: Model, configuration, and capacity (-2). A explicitly states an 800 W motor; B omits this specification.

R07: Function and mechanism information (-2). A highlights an anti-tip feature; B does not state an equivalent functional claim.

Observed winner: **A**.

Figure 4: Expert-annotated rubric selections, signed dimension-level scores, and translated rationales across three distinct industries. Scores are expert-provided B-versus-A criterion judgments, the negatives of the A-versus-B contributions Δ_{ir} defined in Section 3. Swapping A/B preserves their magnitudes and reverses their signs; negating their sum gives the A-versus-B decision margin.