CMAC: CROSS-MODAL ATTRIBUTE CONSISTENCY AS LABEL-FREE SUPERVISION FOR MULTIMODAL PRODUCT RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal representation is crucial for e-commerce tasks such as identical product retrieval. Attribute-enhanced models define the state of the art on this task, yet nothing inside the pipeline checks whether the extracted attributes are faithful: our bad-case analysis of 1,918 retrieval queries finds 48.7% of failures involve missing, conflicting, or hallucinated entities. To address this, we propose CMAC, which defines attribute faithfulness as a three-view consistency problem. It runs one extractor under independent image-only, text-only, and multimodal input masks, then uses TRIAD, a closed-form reliability score whose conflict penalty is adjusted by multimodal arbitration and whose complementarity term retains view-specific information. CMAC uses TRIAD for consistency-weighted contrastive learning, conflict-driven adversarial supervision, and a dense, policy-dependent GRPO reward—whose pure form needs no retriever, oracle, or extra annotation in the reward loop. This paper evaluates the contrastive and RL stages; the adversarial stage is specified and its calibrated evaluation is deferred. After reinforcement learning, the generator improves on the fixed evaluation sample: attribute consistency rises from 0.6265 to 0.6916, parse failures drop from 3.41% to 0.36%, and response length falls from 697 to 339 tokens. Under the strongest published baseline's *own* evaluation protocol, the CMAC encoder consuming identical entity inputs outperforms it by $+5.98$ Recall@1 (95% paired-bootstrap CI $[+4.91, +7.08]$, $\approx 5\times$ the protocol's 2σ resampling-noise floor), following the first faithful external reproduction of that baseline. These results establish three-view consistency as an evidence-grounded source of label-free supervision—first for the retrieval encoder that consumes attributes, then for the generator that produces them.

1 INTRODUCTION

Multimodal representation is crucial for e-commerce tasks such as identical product retrieval: a shopper who photographs a serum bottle on one platform must find the same listing in a catalog of billions (Zheng et al., 2023; Shin et al., 2022), under a different seller, photo, and often language, and the only reliable tie is a handful of structured facts—brand, series, model number, capacity. Matching at this granularity drives duplicate detection, price comparison, and catalog organization (Dong et al., 2022), and is hardest where generic similarity fails: two listings agreeing on everything except one token. General-purpose multimodal embedders such as VLM2Vec (Jiang et al., 2024) lack a mechanism for token-level distinctions—a model number is just another subword sequence. Attribute-enhanced methods close this gap by extracting structured entities and injecting them into the embedding input; the strongest such system, AFMRL (Zhang et al., 2026a), pairs a GRPO-trained entity generator (Shao et al., 2024) with an entity-augmented contrastive encoder and defines the state of the art—but inherits every defect of its extractor, and nothing inside the pipeline notices. Our evaluation uses one frozen public protocol—the M5Product fine-grained benchmark under the strongest baseline's own released setup (§4.1).

The attribute-quality gap. We quantified that inheritance. A bad-case analysis of 1,918 fine-grained M5Product (Dong et al., 2022) queries shows 48.7% of failures involve at least one

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.

entity defect—a clearly-present attribute missed (25.0%), the image- and text-derived values of one attribute contradicting (23.0%), or an attribute fabricated with support from neither modality (16.9%)—and against a 72B oracle the compact extractor recovers only 49.1% of attributes on average. Every remedy pushes the cost elsewhere: annotation does not scale, a larger oracle reintroduces the dependency the compact extractor removed, uniform confidence heuristics stay blind to whether an attribute *should* be extractable, and the RL stage is no help—rewarded with sparse Recall@k, AFMRL's RL phase adds only $+1.86$ R@1 over its contrastive phase (Zhang et al., 2026a) because most rollouts change no ranking. What is missing is a *faithfulness* signal computed from the pipeline's own outputs, without oracle, annotation, or external judge, firing per attribute rather than per sample. This gap inspires our core research question: *can the pipeline's own extraction passes supply a per-attribute faithfulness signal—without oracle, annotation, or a retriever in the reward loop?*

Our insight: three views decompose cross-modal information. Run the extractor under three *independent* input masks—image-only, text-only, and multimodal—and the attributes they emit reveal not just agreement or disagreement, but *what kind of information each modality carries*: *shared* attributes (consistent across unimodal views), *unique* attributes (present in only one view), and *synergistic* attributes (emerging only from the multimodal view or from its arbitration of a conflict). Two views detect agreement and conflict but cannot *arbitrate* (image says A, text says B: which is right?) or detect *synergy*; the third, multimodal view is the minimal addition that resolves both. Three observations turn the decomposition into supervision: *conflict, not absence, is the diagnostic* (a modality may legitimately fail to see an attribute, so demanding coverage would require a ground-truth schema, whereas a conflict is unambiguous evidence that at least one view is wrong); *conflicts are not equally informative* (weighted by modality reliability: a contradiction on category or material is alarming, one on color or brand expected noise); and *unique and synergistic attributes are information, not noise* (preserved as complementary signal, not penalized as missing evidence). Population-level evidence backs the rules: among 1,800 hard queries (self-retrieval failures), *series* conflicts enrich $7.0\times$ over a 13,307-item random baseline (43.3% vs. 6.2%) and *brand* conflicts $2.3\times$ (40.6% vs. 17.8%), while generic keys show no enrichment. The resulting per-product measure of *supervision reliability* costs nothing beyond the inference the pipeline already performs and is used only to assess the reliability of attribute-derived supervision.

Figure 1 makes the distance concrete before we formalize it: it is not generic image–text similarity, but the normalized pattern of agreement, contradiction, and complementary attributes emitted for the same product. For example, the mango record has an image-supported brand (*Youguo Xiansheng*) and a title-supported brand (*Huantai*); the joint view retains the former while recovering the title-supported 500g size. TRIAD prices these attribute-level disagreements by their expected modality reliability, while preserving unique information rather than treating an absent field as an error.

The CMAC framework. To address this, we propose CMAC, which defines attribute faithfulness as a three-view consistency problem. It leverages MLLM generative extraction under the three independent input masks above and distills their agreement into **TRIAD** (Three-view Reliability from Independent Attribute Decomposition), a closed-form, training-free score in $[0, 1]$: Tier 1 prices conflicts on *shared* attributes by attribute-level modality reliability (balanced keys such as category/material cost more than modality-dominated ones such as color/brand); Tier 2 rewards *complementarity*—attributes unique to one view—as information rather than treating absence as missing evidence; Tier 3 arbitrates surviving conflicts by a three-way vote with the multimodal view as arbiter, also surfacing *synergistic* attributes that neither unimodal view extracts alone (exact weights and score in §3; Figure 2). The same score then acts at the two ends of the attribute pipeline. It first supervises the *retrieval encoder*, which consumes attributes: *TRIAD-weighted contrastive learning* (Mode I) applies it as a sample weight and injection-confidence gate that down-weights samples whose shared attributes conflict, and *conflict-driven adversarial supervision* (Mode II, CDARL) turns cross-view conflicts into margin-armed adversarial negatives. It then trains the *attribute generator*, which produces them: *TRIAD-GRPO* (Mode III) is a dense, policy-dependent GRPO reward computed on the policy's own rollouts so within-group advantage stays non-zero—whose pure form needs no retriever, oracle, or extra annotation in the loop. Because the generator's output is the encoder's input, the two stages are designed to close a loop: the signal first shows its worth where attributes are *used*, then improves the attributes where they are *made*.

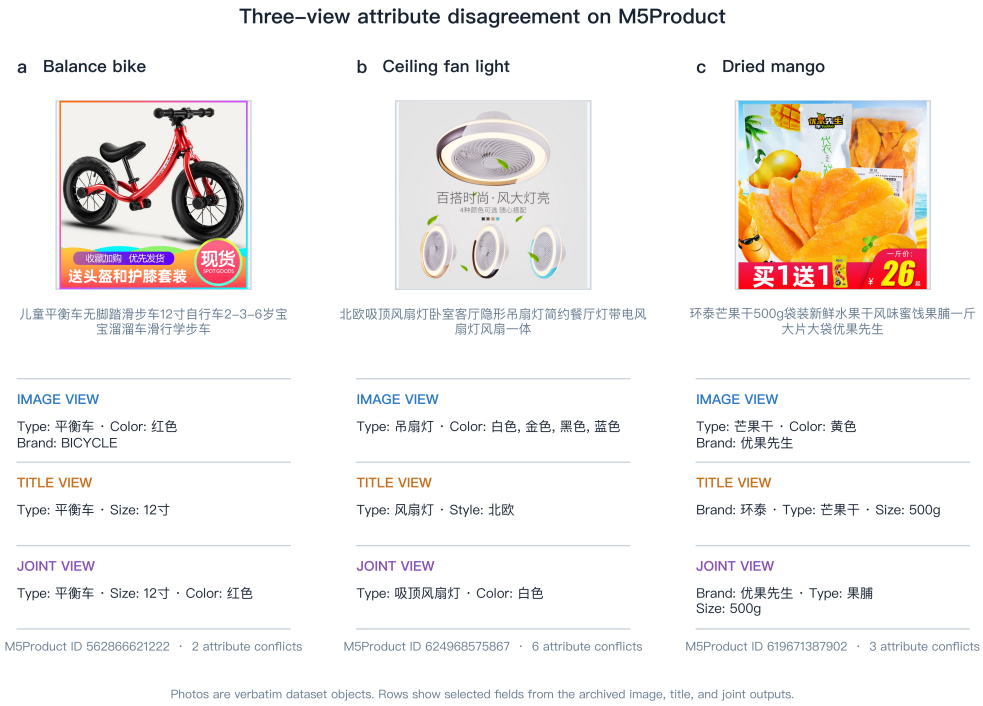

Figure 1: **Three-view attribute disagreement on M5Product.** For each traceable record, image-only, title-only, and joint outputs expose which normalized attributes agree, conflict, or complement one another. In (c), the image-derived brand is *Youguo Xiansheng*, whereas the title-derived brand is *Huantai*; the joint output retains the image-supported brand while recovering the 500g size. Counts summarize conflicts over the complete archived output; rows show selected fields. This is qualitative evidence explaining TRIAD's attribute-level distance, not a retrieval result. Photos are verbatim dataset objects without crop or local visual adjustment.

Protocol discipline. All relative claims use a frozen protocol. Our headline comparison uses the authors' released M5Product evaluation protocol and identical entity inputs for both encoders; implementation details are provided in the appendix.

After reinforcement learning. The RL stage is where the dense reward pays off: unlike Recall@k rewards that vanish for most rollouts, the TRIAD reward varies with what the policy writes—within-group reward variance stays positive at every one of the 350 training steps (minimum 0.0003)—and the training reward rises from 0.661 to 0.746 (peak 0.756). After reinforcement learning, the generator improves on every quality axis: attribute consistency $0.6265 \rightarrow 0.6916$ against the SFT warm start, parse failures $3.41\% \rightarrow 0.36\%$, injection coverage 98.4%, and response length halved ($697 \rightarrow 339$ tokens), lowering the encoder's input cost. The same signal that supervises the encoder is what Mode III uses here to train the generator whose outputs are that encoder's input. We scope these strictly as generator-quality and optimization-health evidence, not a downstream retrieval claim (§3.5; Appendix F).

Findings.

(i) Under the strongest baseline's *own* protocol, the CMAC encoder consuming identical entity inputs outperforms it by $+5.98$ R@1 (paired-bootstrap 95% CI $[+4.91, +7.08]$; $\approx 5\times$ the 2σ resampling-noise floor) after a faithful reproduction (GT 1,918/1,918 exact; Table 1).

(ii) After reinforcement learning with the TRIAD reward, the generator improves on every measured quality axis (consistency $+0.065$, parse failures $3.41\% \rightarrow 0.36\%$, response length reduced); we claim generator-level effects only (§3.5).

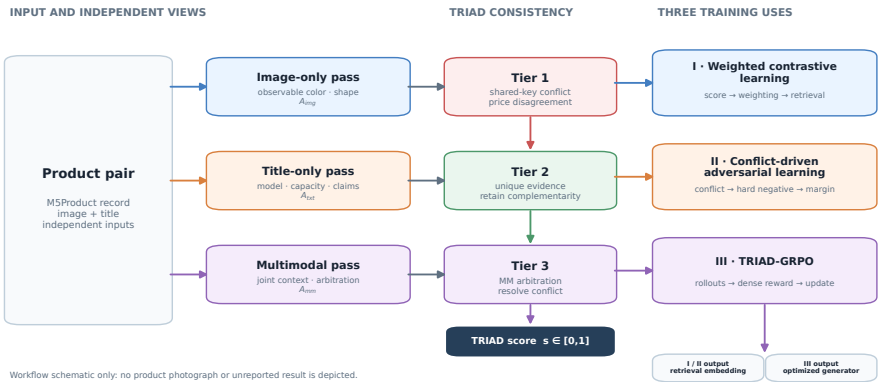

Figure 2: **CMAC workflow.** A product image and title are processed in independent image-only, title-only, and multimodal passes. The tiered TRIAD consistency reading prices shared-key conflict, preserves complementary evidence, and arbitrates supported disagreement; the resulting label-free score first supervises the encoder through weighted contrastive learning (I) and conflict-driven adversarial learning (II), then trains the attribute generator through a dense GRPO reward (III); the generator's improved entities are the encoder's input. This is a schematic: it depicts no product photograph or unreported numerical result.

Read as one arc: the signal first proves itself on the encoder, where attributes are consumed—as a training signal (Mode I), it lifts retrieval (i); it then reaches the generator, where attributes are produced—as a reward (Mode III), it makes the attribute source more faithful (ii). The arc thus returns where the paper opened: on the quality of the attributes the encoder consumes. Whether the improved attributes in turn lift retrieval—closing the loop experimentally—is the next evaluation, not a result claimed here.

Contributions. In summary, this work makes the following contributions:

1. **A systematic quantification of the attribute-faithfulness gap.** A bad-case analysis of 1,918 fine-grained queries showing 48.7% of failures involve missing, conflicting, or hallucinated entities, and a three-view decomposition of 1,330,685 products into shared, unique, and synergistic attributes (§3.3).
2. **A novel label-free supervision framework.** CMAC defines attribute faithfulness as a three-view consistency problem; its closed-form TRIAD score—requiring no oracle, extra annotation, or external evaluator—first supervises the retrieval encoder through contrastive weighting (Mode I) and conflict-driven adversarial supervision (Mode II), then trains the attribute generator through a dense, policy-dependent GRPO reward (Mode III) that markedly improves generator quality while halving response length; these improved outputs are what the encoder consumes (§3).
3. **Evidence-controlled evaluation.** The first faithful external reproduction of the strongest published attribute-enhanced baseline (ground truth 1,918/1,918 exact), a strict same-input comparison ($+5.98$ R@1, $\approx 5\times$ the 2σ resampling-noise floor), tiered entity-injection control, and a fixed-sample evaluation of TRIAD-GRPO generator quality (§4.2, §3.5).

2 Related Work

2.1 Multimodal product retrieval, attribute enhancement, and protocol

Domain-specific contrastive pretraining came early to e-commerce (e-CLIP (Shin et al., 2022); a large-scale e-commerce search system (Zheng et al., 2023)); a second wave replaces the dual tower with a multimodal LLM—VLM2Vec (Jiang et al., 2024) and its successors UniIR (Wei et al., 2023) and GRIT (Muennighoff et al., 2024), MOON (Zhang et al., 2026b) for products, and the concurrent MOON3.0 (Wu et al., 2026), which makes the attributes themselves the object of chain-of-thought optimization. Qwen3-VL-Embedding-2B (Li et al., 2026) is a strong open embedder on MMEB-V2. Closest to us, the attribute-enhanced line—M5Product (Dong et al., 2022) and especially AFMRL (Zhang et al., 2026a), which injects GRPO-generated entities into a contrastive

encoder—defines the strongest published numbers of this family; as of September 2026 no third-party reproduction or comparison existed, so we play a dual role toward AFMRL: first faithful external reproduction, then same-protocol transcendence (§4.2). All these systems engineer the encoder side; none carries a signal for whether the attributes are faithful—our bad-case analysis attributes 48.7% of fine-grained failures to missing, conflicting, or hallucinated entities.

Protocol fragmentation. The field is split by protocol: the AFMRL line reports fine-tuned fine-grained R@1 on M5Product, the MOON line zero-shot R@k on its own benchmarks—in-domain discrimination versus zero-shot generalization, incomparable, with no system reporting under both (even AFMRL and MOON3.0, contemporaneous and overlapping, neither cites nor benchmarks the other). We therefore treat protocol transparency as a contribution: every comparison runs under a frozen, fingerprinted protocol (§4.1), attribute injection is explicitly controlled (identical entity inputs in the head-to-head, §4.2; tiered confidence gating in Mode I, §3.4), and paired bootstrap CIs are published wherever per-query data exists, even when the interval covers zero (§5). CMAC is orthogonal to this axis: it scores faithfulness from the pipeline's own extraction passes, plugs into any attribute-augmented encoder, and—available before training—shapes training itself, not only the inference-time input.

2.2 Consistency signals, RL rewards, and extraction background

Using agreement between views as supervision is long-standing in contrastive learning—multi-view coding (Tian et al., 2020), debiasing (Chuang et al., 2020), hard-negative mixing (Kalantidis et al., 2020), supervised contrastive objectives (Khosla et al., 2020)—but there, agreement shapes the *objective*; none uses it to price the quality of an intermediate structured prediction such as extracted attributes. Down-weighting unreliable supervision is likewise well-studied as label-noise learning: confident learning (Northcutt et al., 2021), co-teaching (Han et al., 2018), dataset cartography (Swayamdipta et al., 2020), and web-scale caption filtering (CapFilt in BLIP (Li et al., 2022); DataComp (Gadre et al., 2023))—these estimate noise from gold labels, a peer network, training dynamics, or an extra scorer, whereas TRIAD prices the pipeline's own structured predictions and is available before any training begins. Two 2026 works move consistency toward our role: R-C2 (Zhang et al., 2026c) turns cross-modal inconsistency into a dense label-free RL reward, and MOON3.0 (Wu et al., 2026) rewards generated attributes via a distilled MLLM evaluator reading reference label, image, and attributes. TRIAD needs neither R-C2's extra generation cycles nor MOON3.0's evaluator/label/CoT machinery, and pays no inference-time reasoning overhead: a closed-form score over extraction passes the pipeline already runs, pricing each conflict by attribute-level modality reliability, and weighting contrastive samples, gating injection, and supervising adversarial negatives. On the RL side, GRPO (Shao et al., 2024) replaced PPO's value model with a group-relative advantage, DeepSeek-R1 (DeepSeek-AI et al., 2025) showed RL with verifiable rewards elicits reasoning at scale, and stacks such as verl (Sheng et al., 2024) make such runs efficient; but the reward, not the optimizer, is the bottleneck for multimodal generation—a generated attribute has no cheap oracle, so systems fall back on retrieval-rank rewards (zero for most rollouts) or on evaluator models (Wu et al., 2026) and cycle-consistency proxies (Zhang et al., 2026c). TRIAD-GRPO (Mode III, §3.5) instead rewards each rollout by the consistency of its own generated attributes against two reference views—closed-form, dense, and policy-dependent, so within-group advantage stays non-zero; an early constant-per-prompt reward produced exactly the zero-advantage degeneracy GRPO cannot tolerate. Finally, *attribute extraction* measures quality against gold labels or oracles—OpenTag (Zheng et al., 2018), Shopping MMLU (Jin et al., 2024), and the GRPO/CoT generators of AFMRL (Zhang et al., 2026a) and MOON3.0 (Wu et al., 2026)—but provides no in-pipeline, training-time faithfulness estimate, the gap TRIAD fills.

3 Method

This section formalizes CMAC as implemented. The starting point is a self-supervised three-view decomposition of attributes; the TRIAD score is its closed-form distillation; the three usage modes wire that score into contrastive weighting, adversarial supervision, and a GRPO reward (notation in Table 3, Appendix E).

3.1 PROBLEM FORMULATION AND NOTATION

Given a product $p = (I, T)$ with image I and title T, learn an encoder $f_\theta : (I, T) \mapsto \mathbf{z} \in \mathbb{R}^d$ (ℓ_2-normalized) mapping same-identity products to proximal embeddings; retrieval is nearest-neighbor search of a query against a gallery $\mathcal{G}$ (d and training configuration in Table 2, §B). Attribute-augmented systems additionally extract structured attributes $\mathcal{A} = \{(k, v)\}$ and feed them to the encoder. CMAC's premise is that these attributes can be *decomposed* by running the extractor under three independent input masks:

$$\mathcal{A}^{\text{img}} = \text{Ext}(I), \quad \mathcal{A}^{\text{txt}} = \text{Ext}(T), \quad \mathcal{A}^{\text{mm}} = \text{Ext}(I, T), \tag{1}$$

where Ext is a single attribute extractor and the two unimodal views are generated with *no access* to the other modality. The three views partition the emitted attributes by their information source:

- **Co-observed** ($\mathcal{K}^\cap$): keys present in both unimodal views. They partition into agreeing keys $\mathcal{S}$ and conflicting keys $\mathcal{C}$.
- **Unique** ($\mathcal{U}^{\text{img}}, \mathcal{U}^{\text{txt}}$): keys present in only one unimodal view—information legitimately carried by one modality and not the other.
- **Synergistic** ($\mathcal{Y}$): keys surfaced only by the multimodal view or resolved from a conflict by that view—information requiring joint context.

The unimodal views provide the conflict and complementarity evidence; the multimodal view arbitrates conflicts (Definition 2), attenuates their score penalty when it supports a unimodal value, and supplies entities injected into encoder prompts (Mode I).

3.2 TRIAD: THREE-VIEW ATTRIBUTE RELIABILITY

Definition 1 (Normalization, agreement, conflict, and decomposition)**.** *Values are compared after normalization ν: lowercasing, punctuation stripping, whitespace collapsing, and a small synonym table (e.g.,* navy→blue*). For each co-observed key $k \in \mathcal{K}^\cap$, the values* agree *when $\nu(v^{\text{img}}) = \nu(v^{\text{txt}})$ or either normalized value contains the other as a substring; otherwise k conflicts and enters $\mathcal{C}$. Thus $\mathcal{K}^\cap = \mathcal{S} \dot\cup \mathcal{C}$. A key present in only one view is unique and is not penalized, because absence can be a legitimate modality limitation. Keys present only in the multimodal view are synergistic.*

The substring rule absorbs formatting granularity (*mango* ⊂ *mango-flavored*) at the cost of judging near-miss pairs strictly; one normalization asymmetry (ASCII comma vs. CJK U+3001) survives as a flagged edge case (worked examples on real test items in Appendix C). Each key k carries a reliability class $\rho(k)$ (Table 4, Appendix E)—*balanced* (both modalities should perceive it: category, subcategory, material, style, gender, season, occasion, product_type), *text-dominant* (brand, model, specification, size, function, series, ingredients), or *image-dominant* (color, pattern, shape, texture, packaging); unknown keys default to balanced. A conflict on a balanced key is priced $\pi(k) = 1.0$ and one on a modality-dominant key 0.3, reflecting whether the two modalities are expected to corroborate that attribute. Population-level conflict enrichment supports this asymmetry (§1).

Definition 2 (TRIAD score)**.** *Let $\mathcal{K}^\cap = \mathcal{K}^{\text{img}} \cap \mathcal{K}^{\text{txt}}$ and $\mathcal{C} \subseteq \mathcal{K}^\cap$ be the conflicting keys (Definition 1). For a conflicting key, $a_k \in [0, 1]$ is the three-way arbitration agreement: $a_k = \frac{1}{3}[\mathbf{1}(v_k^{\text{img}} = v_k^{\text{mm}}) + \mathbf{1}(v_k^{\text{txt}} = v_k^{\text{mm}}) + \mathbf{1}(v_k^{\text{img}} = v_k^{\text{txt}})]$ when the multimodal value is available, and $a_k = 0$ otherwise. The arbitration-adjusted penalty, conflict-free rate, and complementarity are*

$$\tilde\pi(k) = \pi(k)(1 - a_k), \qquad \phi = 1 - \frac{\sum_{k \in \mathcal{C}} \tilde\pi(k)}{|\mathcal{K}^\cap|}, \qquad \gamma = \frac{|\mathcal{U}^{\text{img}}| + |\mathcal{U}^{\text{txt}}|}{|\mathcal{K}^{\text{img}} \cup \mathcal{K}^{\text{txt}}|}. \tag{2}$$

The score is

$$c = \begin{cases} \alpha\phi + \beta\beta_b\gamma, & |\mathcal{K}^\cap| > 0, \\ 0.5 + \beta\beta_b\gamma, & |\mathcal{K}^\cap| = 0, \end{cases} \qquad \alpha = 0.8,\ \beta = 0.2,\ \beta_b = 0.2. \tag{3}$$

If both views are empty, $c = 0.5$.

Three remarks make the definition precise. **(i) Range.** Because $0 \le \tilde\pi(k) \le 1$ and $0 \le \gamma \le 1$, the shared-key branch lies in $[0, 0.84]$ and the no-co-observation branch in $[0.5, 0.54]$; no numerical clamp is required for the stated implementation. **(ii) Three-view scalar.** The multimodal view enters

c only by reducing a conflict penalty when it agrees with one of the unimodal values. It also records the majority value for prompt construction. Thus the third view has an explicit numerical role in arbitration as well as a semantic role in surfacing synergy. **(iii) Contradiction, not sparsity.** TRIAD does not penalize a key that only one modality can express; only incompatible co-observed values receive a penalty. The complete scoring procedure is Algorithm 1 (Appendix E); computational overhead is zero GPU passes (§E.1).

3.3 WHY THREE VIEWS? POPULATION-LEVEL DECOMPOSITION

On 1,330,685 M5Product test items (Qwen3.5-27B 3-view extraction), the attribute dictionaries contain 44.7% co-observed (shared) attributes, 51.8% unique attributes, and 3.5% synergistic attributes; 59.2% of co-observed keys conflict, and the multimodal view resolves 69.3% of those conflicts (Table 6, Appendix F.5). Three observations motivate the design. **(1) Co-observed attributes are the supervision anchor:** they expose direct agreement or contradiction. **(2) Unique attributes are information, not noise:** more than half appear in only one view; penalizing absence would require a ground-truth schema and discard complementary signal. **(3) The third view is necessary:** it arbitrates a conflict and surfaces synergy. The score is therefore a reliability measure for attribute-derived supervision, not a surrogate retrieval metric.

One score, three usages. The same scalar is consumed through specified maps: Mode I uses an affine weight $w = \lambda_w c + (1 - \lambda_w)$ (§3.4); Mode III scores generated attributes against frozen reference views before group-relative advantage estimation (§3.5); and a Mode II gate $\mathbf{1}[c \geq \tau]$ selects an alignment or conflict term (specification in Appendix F). This paper evaluates Modes I and III only: Mode II's current gate calibration does not support a performance claim, so it is deferred to future work (held-results disclosure, Appendix G). The score's semantics are held fixed across these uses: it measures the reliability of attribute-derived supervision.

3.4 MODE I: CONSISTENCY-WEIGHTED CONTRASTIVE TRAINING

Mode I is *label-noise down-weighting*: a low-TRIAD sample has incompatible co-observed attributes, so supervision derived from its multimodal entities receives proportionally less gradient weight. The score is interpreted only as a reliability prior on the supervision. Each sample's contribution to a multi-positive InfoNCE loss with additive margin is scaled by $w_i = \lambda_w c_i + (1 - \lambda_w)$, $\lambda_w = 0.3$, so $w_i \in [0.7, 1.0]$: a fully conflicting sample keeps 70% of its gradient rather than being discarded—down-weighting, not censoring—and the counterfactual extraction ladder verifies the reliability ordering the weight assumes (induced weight distribution and per-key conflict enrichment in Appendix H, Figure 5; worked weight examples in Appendix C; loss equations in Appendix F). A finer per-attribute variant assigns each candidate entity a title-verified confidence c_a and injects on three tiers (≥ 0.7 verbatim, 0.4–0.7 with an explicit "(?)" marker, < 0.4 omitted), intercepting the dominant hallucination mode at prompt assembly. Values of τ and m per benchmark are in Table 2 (§B).

3.5 MODE III: CONSISTENCY AS DENSE GRPO REWARD (TRIAD-GRPO)

For the entity-generation policy, retrieval-based rewards are sparse: most rollouts change no ranking, so within-group rewards are equal and group-relative advantages vanish (the RL stage of the strongest attribute-enhanced baseline gains only $+1.86$ R@1 over its contrastive stage). TRIAD-GRPO replaces the sparse term with the same consistency score on each rollout's own output.

For prompt i the policy π_θ samples $G = 8$ rollouts, each parsed into an attribute dictionary $\mathcal{A}^g$; against frozen reference views extracted once per prompt, the reward is the mean TRIAD of $\mathcal{A}^g$ against both, $r_g^{\text{triad}} = \frac{1}{2}\big[c(\mathcal{A}^g, \mathcal{A}_{\text{ref}}^{\text{img}}) + c(\mathcal{A}^g, \mathcal{A}_{\text{ref}}^{\text{txt}})\big]$, mixed with any retrieval term as $r_g = \omega_r\, r_g^{\text{retr}} + \omega_c\, r_g^{\text{triad}}$ (ω_r=0.7, ω_c=0.3 in the mixed h73 arm; ω_r=0 in the pure arm reported below) and standardized within group before the clipped GRPO objective (Shao et al., 2024); an unparseable rollout gets $r_g = -0.1$, unscaled, before the mixture (Definition 3 and the reward ladder in Table 5, Appendix F). Policy-dependence is required: an early content-ignoring slot gave every rollout of a prompt the same constant, collapsing within-group variance to zero (no GRPO gradient); scoring against both reference views keeps variance positive across all seven smoke steps (0.0012–0.0038; Figure 3). The same recipe then trains the extractor itself: a *single* view-conditioned

Table 1: **AFMRL Table 2 reproduction and CMAC head-to-head** under the authors' evaluation protocol (single-positive resampling, 50 iterations, full 113,105-item gallery; percentages). Top: reproduction ladder (Δ vs. published). Bottom: the CMAC encoder with entity inputs identical to the AFMRL row (*same-input*; features extracted via the authors' pipeline) and with its own training-time entity inputs (*native-input*). The $+5.98$ R@1 gap ($\approx 5\times$ the sampling-noise floor $2\sigma \approx 1.2$) is the same-input strict comparison; native-input adds $+0.51$ (inter-arm difference $\ll$ noise).

Arm	R@1	Pub. R@1	Δ	R@5	R@10	NDCG@5	mAP@10
Reproduction ladder (authors' released artifacts)							
VLM2Vec	48.04	48.05	-0.01	66.76	72.73	—	—
+AGCL	49.98	51.06	-1.08	66.89	72.61	—	—
+Distill	51.64	52.42	-0.78	69.80	75.72	—	—
AFMRL	53.04	54.28	-1.24	70.83	76.62	62.79	60.87
Anchors (published baseline features)							
CLIP	16.03	14.98	$+1.05$	25.23	30.72	—	—
BLIP	21.17	19.42	$+1.75$	32.23	38.25	—	—
ALBEF	22.11	20.59	$+1.52$	33.24	39.28	—	—
Head-to-head (CMAC encoder; features extracted via the authors' pipeline)							
CMAC (ours, same-input)	**59.02**	—	—	**75.31**	**80.68**	**67.73**	**65.93**
CMAC (ours, native-input)	59.53	—	—	75.57	80.71	68.08	66.29

policy—same 3B scale as the baseline's RL generator—produces all three views of §3.1 through input-level modality blanking, each view scored against its own modality's frozen reference (Definition 4, Appendix F).

Outcome after reinforcement learning. On the full 350-step run the reward behaves as designed (Figure 4): it rises from 0.661 (first-50-step mean) to 0.746 (last-50; peak 0.756 at step 335) with within-group reward variance positive at every step (min 0.0003), and the generator improves on every quality axis measured on a fixed eval sample—attribute consistency $0.6265 \rightarrow 0.6916$ against the SFT-10k warm start (this one axis is measured in the reward's own units; the others are reward-independent), parse failures $3.41\% \rightarrow 0.36\%$, injection coverage $89.2\% \rightarrow 98.4\%$, and response length halved ($697 \rightarrow 339$ tokens; entity region $133 \rightarrow 98$ characters), lowering the encoder's input cost. We scope these as generator-quality and optimization-health evidence, not a downstream retrieval claim (full statistics, the h73 mixed-reward arm, and honest caveats in Appendix F).

4 Experiments

4.1 Benchmarks and protocols

M5Product in-domain fine-grained retrieval. The fine-grained benchmark builds on Zhang et al. (2026a)'s evaluation data under the authors' own protocol: a 113,105-product gallery over 267 fine labels with 1,918 dedicated queries. We first reproduce AFMRL's published Table 2 from the authors' released artifacts, then run the CMAC encoder on identical entity inputs (§4.2; split counts and metrics in Appendix A).

4.2 AFMRL Table 2 reproduction and head-to-head comparison

Our headline result is a same-input head-to-head under the strongest published baseline's *own* protocol: we first reproduced AFMRL's Table 2 from the authors' released artifacts—the first faithful external reproduction of this system—then ran the CMAC encoder on *identical* entity inputs. GT reconstruction is exact (1,918/1,918 at set level), three anchors align within sampling noise, and the key AFMRL row reproduces within tolerance; the CMAC encoder reaches 59.02 R@1, $+5.98$ over the reproduced AFMRL arm and $+4.74$ over the *published* value ($\approx 5\times$ and $\approx 4\times$ the 2σ resampling-noise floor, respectively; full ladder and head-to-head in Table 1).

The reproduction confirms three facts (Table 1): (i) the released artifacts identify the correct checkpoint generation for the published AGCL-family features, resolving a version confound; (ii) sampling noise (± 1.2 at 2σ) explains the -1.24 gap; and (iii) under identical inputs the CMAC encoder

beats the reproduced AFMRL arm far above the noise floor. We quote the gap as a multiple of that floor rather than a σ level: the three anchor offsets share one sign and two individually exceed 2σ, indicating a small systematic component the resampling estimate does not capture. A paired percentile bootstrap over the 1,918 queries (10,000 resamples, within-query pairing) puts the same-input gap at $+5.98$ R@1 with 95% CI $[+4.91, +7.08]$, excluding zero. Among the encoder variants trained under this protocol, we report, per a pre-registered rule, the arm that minimizes same-input confounds relative to the reproduced AFMRL arm (configuration in Appendix B); arms that trail are not silently dropped—they are held out under our disclosure policy with artifacts retained, re-entering only alongside a completed improved run (Appendix G).

Coarse cross-modal retrieval (Table 1 of Zhang et al. 2026a). On the full 1.33M-item M5Product test set (6,422 labels), a 12-variant scan (3 directions $\times$ 2 rank-masking $\times$ 2 self-inclusion modes) under the AP-hits formula reproduces both checkpoints' published category-level mAP@K within $\leq$1.74 pp, confirming the AP-hits metric family and the category-level relevance definition; instance-level R@K is not reproducible from the public materials (13–30 pp self-match gaps; no disclosed evaluation subset or recall definition).

5 Discussion and Limitations

The evidence supports two bounded claims. First, under the authors' released M5Product protocol and identical entity inputs, CMAC improves the reproduced AFMRL encoder by $+5.98$ R@1 (paired-bootstrap 95% CI $[+4.91, +7.08]$, excluding zero; Table 1). Second, TRIAD-GRPO improves extraction quality on a fixed evaluation sample; this is generator-quality evidence rather than a retrieval result. The evaluation is limited to the released M5Product protocol, and the current paper does not claim cross-domain retrieval generalization. Mode II is specified as a conflict-supervision mechanism (Appendix F) but requires a dedicated calibrated evaluation before it can support a performance claim (held-results disclosure, Appendix G). Future work should test the same scorer and arbitration rule under independently released retrieval protocols and image-query settings.

6 Conclusion

CMAC turns three independently extracted attribute views into label-free supervision. TRIAD separates conflict penalties, complementarity, and multimodal arbitration in one bounded score, then instantiates that score as contrastive weighting, conflict supervision, and a dense GRPO reward. Under identical entity inputs and the reproduced AFMRL evaluation protocol, the CMAC encoder improves R@1 by $+5.98$ ($\approx 5\times$ the 2σ resampling-noise floor). TRIAD-GRPO further improves measured generator consistency, parse validity, and output compactness. These results support three-view attribute consistency as an evidence-grounded training signal while leaving cross-domain retrieval generalization for future evaluation.

Reproducibility Statement

All code, extracted three-view attributes, evaluation splits, and training/evaluation configurations supporting reported claims will be released upon publication, together with versioned artifacts and protocol details. Results without a demonstrated improvement are excluded from this manuscript rather than treated as pending evidence.

AI Use Statement

In this work, we used generative AI tools (LLM-based coding and writing assistants) for the following subtasks whose disclosure is *required* by the ICLR 2027 AI policy: implementing methods (training, evaluation, and analysis code was drafted with AI assistance and then executed, reviewed, and verified by the authors against frozen protocol fingerprints and a 31-case unit-test suite), cleaning and reformatting dataset artifacts, and interpreting results (all statistics are computed by deterministic scripts from versioned run artifacts; AI assistance was used to draft the accompanying prose, which the authors checked against those artifacts). We did not use generative AI tools to generate

synthetic datasets, to formulate or prove mathematical claims, or to run experiments autonomously. Additionally, we used generative AI tools for subtasks whose disclosure is *recommended*: creating and editing scientific figures (each figure is regenerated from, and asserted against, the underlying result JSONs), drafting and editing parts of the research paper, and formatting references. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Ching-Yao Chuang, Jialin Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

DeepSeek-AI, Daya Guo, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Xiao Dong, Xunlin Zhan, Yangxin Wu, Yunchao Wei, Michael C. Kampffmeyer, Xiaoyong Wei, Minlong Lu, Yaowei Wang, and Xiaodan Liang. M5Product: Self-harmonized contrastive learning for e-commercial multi-modal pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, et al. DataComp: In search of the next generation of multimodal datasets. In *Advances in Neural Information Processing Systems 36, Datasets and Benchmarks Track*, 2023.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W. Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems 31*, 2018.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. VLM2Vec: Training vision-language models for massive multimodal embedding tasks. *arXiv preprint arXiv:2410.05160*, 2024.

Yilun Jin, Zheng Li, Chenwei Zhang, Tianyu Cao, Yifan Gao, Pratik Jayarao, Mao Li, Xin Liu, Ritesh Sarkhel, Xianfeng Tang, Haodong Wang, Zhengyang Wang, Wenju Xu, Jingfeng Yang, Qingyu Yin, Xian Li, Priyanka Nigam, Yi Xu, Kai Chen, Qiang Yang, Meng Jiang, and Bing Yin. Shopping MMLU: A massive multi-task online shopping benchmark for large language models. In *Advances in Neural Information Processing Systems 37, Datasets and Benchmarks Track*, 2024.

Yannis Kalantidis, Mert Bulent Sariyildiz, Noë Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Prannay Khosla, Piotr Teterwak, Chen Wang, Ajaya Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-VL-embedding and qwen3-VL-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026.

Niklas Muennighoff, Hao Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. Generative representational instruction tuning. *arXiv preprint arXiv:2402.09906*, 2024.

Curtis G. Northcutt, Lu Jiang, and Isaac L. Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, 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.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. HybridFlow: A flexible and efficient RLHF framework. *arXiv preprint arXiv:2409.19256*, 2024.

Wonyoung Shin, Jonghun Park, Taekang Woo, Yongwoo Cho, Kwangjin Oh, and Hwanjun Song. e-CLIP: Large-scale vision-language representation learning in e-commerce. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM)*, 2022. doi: 10.1145/3511808.3557067.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9275–9293, 2020.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European Conference on Computer Vision (ECCV)*, 2020.

Peng Wang, Shuai Bai, Sinan Tan, Shitong Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengyuan Du, Xu Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. UniIR: Training and benchmarking universal multimodal information retrievers. *arXiv preprint arXiv:2311.17136*, 2023.

Junxian Wu, Chenghan Fu, Zhanheng Nie, Daoze Zhang, Bowen Wan, Wanxian Guan, Chuan Yu, Jian Xu, and Bo Zheng. MOON3.0: Reasoning-aware multimodal representation learning for e-commerce product understanding. In *Proceedings of the 34th ACM International Conference on Multimedia (ACM MM)*, 2026.

Biao Zhang et al. AFMRL: Attribute-enhanced fusion multimodal representation learning. In *Findings of the Association for Computational Linguistics: ACL 2026*, 2026a. arXiv:2604.20135.

Daoze Zhang, Chenghan Fu, Zhanheng Nie, Jianyu Liu, Wanxian Guan, Yuan Gao, Jun Song, Pengjie Wang, Jian Xu, and Bo Zheng. MOON: Generative MLLM-based multimodal representation learning for e-commerce product understanding. In *Proceedings of the 19th ACM International Conference on Web Search and Data Mining (WSDM)*, 2026b.

Zirui Zhang, Haoyu Dong, Kexin Pei, and Chengzhi Mao. R-C2: Cycle-consistent reinforcement learning improves multimodal reasoning. *arXiv preprint arXiv:2603.25720*, 2026c.

Guineng Zheng, Subhabrata Mukherjee, Xin Luna Dong, and Feifei Li. OpenTag: Open attribute value extraction from product profiles. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2018. doi: 10.1145/3219819.3219839.

Xiaoyang Zheng, Fuyu Lv, Zilong Wang, Qingwen Liu, and Xiaoyi Zeng. Delving into e-commerce product retrieval with vision-language pre-training. In *Proceedings of the 6th ACM SIGIR Symposium on Information Retrieval in Practice (SIRIP)*, 2023. doi: 10.1145/3539618.3591859.

Appendix roadmap. This appendix records public-benchmark protocol details, implementation configuration, reproducibility artifacts, formal definitions, and generator optimization evidence supporting the main paper.

A Dataset and protocol details

A.1 M5Product fine-grained protocol

M5Product (Dong et al., 2022) is the public training corpus: after filtering, 5,760,482 products with images and titles (6,000 coarse categories). The fine-grained evaluation builds on Zhang et al. (2026a)'s fine-grained data and follows the authors' own protocol: a 113,105-product gallery over 267 fine labels with 1,918 dedicated queries. We first reproduce AFMRL's published Table 2 from the authors' released artifacts (§4.2), then run the CMAC encoder on identical entity inputs for a strict same-input comparison. No in-domain M5Product claim in this paper rests on any other protocol. Metrics: R@1/5/10, NDCG@10, mAP@10, kNN, NMI, ARI, computed with FAISS over cosine similarities.

B Implementation details

Table 2: Implementation configuration (reference design). Values matched to Zhang et al., 2026a are marked †. For the RL stage, "matched" means the same training budget—350 steps, G=8 rollouts, lr 10^{-6}—and nothing broader; in particular, our KL coefficient is 10^{-3} against their 10^{-2}.

Component	Configuration
Base encoder	Qwen2-VL-2B-Instruct (Wang et al., 2024), VLM2Vec (Jiang et al., 2024) init (checkpoint-2500: BS 2048†, IB 64†, GradCache 16/16, lr 5e-5, 8 GPUs)
Embedding dimension	768
LoRA	rank 8, α 16, dropout 0.05
Attribute views	image-only / text-only / multimodal prompts (Appendix prompts in code)
Training-side extractor	Qwen2.5-VL-3B-Instruct
Test-side / oracle extractor	Qwen2.5-VL-72B-Instruct
Decomposition / RL-reference extractor	Qwen3.5-27B (frozen; §3.3 three-view decomposition, TRIAD-GRPO reference views)
TRIAD scorer	α=0.8, β=0.2, penalties 1.0/0.3, bonus weight 0.2; CPU string ops
Mode I	InfoNCE τ=0.08, margin m=0.2; $w = 0.3\,c + 0.7$; entity tiers: HIGH keys 1.0/0.5, name 0.3, func/type 0.5, default 0.6; $\geq$0.7 write, 0.4–0.7 marker "(?)", $<$0.4 omit
Mode II (CDARL, deferred)	gate τ=0.7; w_c=0.5, margin 0.3; w_a=0.2 (Appendix F)
Mode III (TRIAD-GRPO)	lr 1e-6†, 350 steps†, KL β 0.001, clip ε 0.2, G=8 rollouts†, recall@k k=50; weights 0.7/0.3 (smoke arm 0/1); invalid JSON -0.1 unscaled before weighting

C Walkthrough examples

The two examples below are real test-set items; attribute values are translated from the original Chinese listings for readability, and every number is the scorer's output. They illustrate the two behaviors a reviewer most needs to see: tolerance for explainable single-key conflicts, and severity toward mutual contradiction.

Example 1: dual-brand title (TRIAD 0.6971, Mode I weight 0.9091). A dried-mango listing whose title contains *two* brand tokens—the producer brand and the flagship-store brand. The image view reads the store brand; the text view reads the producer brand. Of four shared keys, two conflict (brand: text-dominant, priced 0.3; shape "sliced" vs. "large pieces": image-dominant, priced 0.3) while the flavor pair agrees fuzzily (*mango* $\subset$ *mango-flavored*). The score is $0.8 \times 0.85 + 0.2 \times (0.4286 \times 0.2) = 0.6971$ (complementarity: 3 unique keys over the 7-key unimodal union), so the sample keeps almost full contrastive weight ($w = 0.9091$). The entity-confidence tiering then writes only the title-verified brand at full confidence and marks the remaining keys with a "(?)" marker—noting honestly that title verification cannot adjudicate *which* of the two brands is correct (the title itself is the dual-brand source); it only guarantees that what is written agrees with the title.

Example 2: all-conflict card holder (TRIAD 0.1162, weight 0.7349). A business-card holder whose two views disagree on *all six* shared keys: material (leather vs. genuine leather, balanced, 1.0), style (business vs. high-end, 1.0), capacity ($\sim$25 cards vs. multi-slot, 1.0), target (unisex vs. men, 1.0), product type (card case vs. business-card holder, 1.0), and function (storage vs. card-holder, text-dominant, 0.3). Total penalty 5.3 over 6 shared keys gives $\phi = 0.1167$; complementarity over the full dictionaries is 0.5714; the score is 0.1162 and the contrastive weight 0.7349, near the 0.7 floor—this sample's loss contribution is cut by about 27%. The instructive counterfactual (actually

Table 3: Notation.

Symbol	Meaning
$p = (I, T)$	product: image I, title T
$f_\theta,\ \mathbf{z}$	encoder and its ℓ_2-normalized output
$\mathcal{G}$	retrieval gallery
$\mathrm{Ext},\ \mathcal{A}^v$	attribute extractor; view $v \in \{\mathrm{img}, \mathrm{txt}, \mathrm{mm}\}$
$\mathcal{K}^v$	key set of view v; $\mathcal{K}^\cap = \mathcal{K}^{\mathrm{img}} \cap \mathcal{K}^{\mathrm{txt}}$ (co-observed); $\mathcal{S}$ agreeing subset
$\nu(\cdot)$	value normalization (Definition 1)
$\rho(k) \in \{\mathrm{image}, \mathrm{text}, \mathrm{balanced}\}$	modality reliability of key k
$\pi(k)$	conflict penalty of key k: 1.0 balanced, 0.3 dominant
$\phi,\ \gamma$	conflict-free rate; complementarity
α, β, β_b	score weights: 0.8, 0.2, 0.2
$c(p)$ or $c(X, Y)$	TRIAD score (Definition 2); two-dictionary form
$w_i,\ g_i,\ \tau$	Mode I sample weight; Mode II gate and threshold (0.7)
$\mathbf{z}_i^{\mathrm{mm}},\ \mathbf{z}_i^{\mathrm{alt}}$	Mode II embeddings of the multimodal / a sampled unimodal view
$r_g,\ r_g^{\mathrm{triad}},\ \omega_r, \omega_c$	Mode III reward, its TRIAD term, and mixture weights
$A_g,\ G$	group-relative advantage; rollouts per prompt (8)

Table 4: Attribute classification and conflict pricing. The table is static domain knowledge—the only external input TRIAD receives. Unknown keys are conservatively priced as balanced (full penalty).

Reliability class	Keys	$\pi(k)$	Rationale
balanced	category, subcategory, material, style, gender, season, occasion, product_type	1.0	both modalities reliable; disagreement is strong evidence of an extraction error
text-dominant	brand, model, specification, size, function, series, ingredients	0.3	the weaker modality is expected to miss or misread
image-dominant	color, pattern, shape, texture, packaging	0.3	(symmetric rationale)
unknown	anything else	1.0	no reliability knowledge ⇒ assume informative conflict

computed, not estimated): delete the six conflicting keys from both views and the sample falls into the no-shared-key branch, scoring 0.54 with weight 0.862. TRIAD penalizes *mutual contradiction*, not information poverty—the opposite of what a coverage-style metric would do, and the property that lets genuinely unimodal attributes pass unpunished.

Nine further walkthroughs—including a six-key conflict containing a punctuation-induced pseudo-conflict, a fully consistent sample at 0.8229, a same-label near-duplicate confuser pair that is inseparable at the attribute level (TRIAD 0.80 with zero conflicts), and the GRPO reward-ladder group of §3.5—are archived with full raw values in the mechanism-analysis report released with the code.

D Data licenses and ethics

M5Product is a public research benchmark used under its released terms (Dong et al., 2022). Product listings contain no personal information beyond standard catalog content. All base models are used under their respective open licenses.

E Method notation, attribute pricing, and scoring algorithm

The notation table, attribute-pricing table, scoring algorithm, reward ladder, and computational-overhead accounting referenced in §3 are collected here.

Table 5 shows the ladder. Three properties matter. The score decreases strictly monotonically with conflict count. The empty dictionary out-scores three-plus-conflict rollouts ($0.5400 > 0.4920$): with no shared keys it lands in the neutral branch, a direct consequence of punishing contradiction rather than sparsity. We disclose this *lazy-rollout loophole* openly: it is bounded by the unscaled -0.1

Algorithm 1 TRIAD scoring (as implemented).

Require: dictionaries $\mathcal{A}^{\text{img}}, \mathcal{A}^{\text{txt}}$; optional arbiter $\mathcal{A}^{\text{mm}}$; penalties π_{bal}=1.0, π_{dom}=0.3; weights α=0.8, β=0.2, β_b=0.2
Ensure: score $c \in [0, 1]$; per-conflict verdicts

```
normalize both dictionaries with ν                ▷ lowercase, strip punctuation, synonyms
K∩ ← K^img ∩ K^txt; C ← ∅
for k ∈ K∩ do
    if v̄_k^img = v̄_k^txt or v̄_k^img ⊂ v̄_k^txt or v̄_k^txt ⊂ v̄_k^img then
        continue                                  ▷ agreement: equality or substring containment
    else
        C ← C ∪ {k}
    end if
end for
γ ← |K^img △ K^txt| / |K^img ∪ K^txt|
if K∩ = ∅ then
    return c = 0.5 + ββ_bγ                        ▷ no-shared-keys branch
end if
φ ← 1 − (Σ_{k∈C} π(k)) / |K∩|                     ▷ π(k) by Table 4
return c = αφ + ββ_bγ
```

Table 5: **Reward ladder** for one real prompt (DERMAFIRM toner; reference views extracted from a real catalog row, which itself carries a cross-lingual brand/series conflict—a constant cost shared by every rollout in the group). Rollouts are constructed counterfactuals. Group statistics: mean 0.5916, variance 0.016137, std 0.1270.

Rollout	**Keys conflicting vs. refs**	r^{triad}
all five keys agree with references	0	0.7560
brand wrong (Lancôme)	1	0.7260
+ product type wrong (lotion)	2	0.5460
+ function wrong	3	0.4920
all five keys conflict	5	0.4380
empty dictionary {}	—	0.5400
invalid JSON	—	−0.1 (unscaled)

invalid penalty below and, in the mixed arm ($\omega_r = 0.7,\ \omega_c = 0.3$), by the retrieval leg that an empty dictionary cannot satisfy; the pure-TRIAD smoke arm ($\omega_c = 1$) is the stress test that exposes the property cleanly.

E.1 Computational overhead

TRIAD adds no model. Scoring is $O(|\mathcal{K}^{\cap}|)$ string comparisons on CPU—no forward pass, no retrieval-index query. The extraction passes are the honest cost: beyond the multimodal pass that attribute-augmented pipelines run anyway (it is the entity source for Mode I prompts), TRIAD requires the two unimodal passes, once per product at training-data-preparation time, not per training step. In Mode III, the reward for each rollout is two dictionary comparisons with no retrieval call and no evaluator forward pass—against the extra inference cycles of cycle-consistency rewards (Zhang et al., 2026c) and the distilled-evaluator forward pass of composite attribute rewards (Wu et al., 2026). The only piece of injected domain knowledge is Table 4 itself: three reliability classes over ∼20 keys, static, and auditable.

F Relocated method equations, result tables, and analyses

The nine-page body limit relocates the following verbatim from §3–§5. Nothing here is new; each block is the complete, unabridged material the body summarizes and points to.

F.1 USAGE-MODE EQUATIONS AND ENTITY GATING (FROM §3.4–§3.5)

Mode I: consistency-weighted contrastive training. Given a multi-positive InfoNCE loss with additive margin m on positives,

$$\mathcal{L}_i^{\text{InfoNCE+AM}} = -\log \frac{\sum_{j\in\mathcal{P}_i} \exp\big(\mathbf{z}_i^\top \mathbf{z}_j/\tau - m\big)}{\sum_{j\neq i} \exp\big(\mathbf{z}_i^\top \mathbf{z}_j/\tau\big)}, \tag{4}$$

each sample's contribution is scaled by

$$w_i = \lambda_w\, c_i + (1-\lambda_w), \qquad \lambda_w = 0.3 \;\Rightarrow\; w_i \in [0.7, 1.0], \tag{5}$$

$$\mathcal{L}^{\text{ModeI}} = \frac{1}{N}\sum_{i=1}^{N} w_i\, \mathcal{L}_i^{\text{InfoNCE+AM}}. \tag{6}$$

The affine map is deliberate: a fully conflicting sample keeps 70% of its gradient rather than being discarded—down-weighting, not censoring—and the counterfactual extraction ladder verifies the reliability ordering the weight assumes. Values of τ and m are in §B.

Entity-level confidence gating. A finer, per-attribute use assigns each candidate entity $a = (k, v)$ from the multimodal view a confidence c_a (title-verified high-trust keys brand/specification/series: 1.0 if the value occurs in the normalized title, else 0.5; product name 0.3; function/product_type 0.5; other keys 0.6) and injects on three tiers—$c_a \geq 0.7$ verbatim, $0.4 \leq c_a < 0.7$ with an explicit "(?)" marker, $c_a < 0.4$ omitted—intercepting the dominant hallucination mode (fabricated product names) at prompt assembly, the cheapest interception point.

Mode II: conflict as supervision (CDARL). Down-weighting wastes the information in a conflict. CDARL (Conflict-Driven Adversarial Representation Learning) instead routes by the score: for sample i, encode the multimodal view and one unimodal view, $\mathbf{z}_i^{\text{mm}} = f_\theta(T_i \oplus \mathcal{A}_i^{\text{mm}})$ and $\mathbf{z}_i^{\text{alt}} = f_\theta(T_i \oplus \mathcal{A}_i^{v})$, $v \in \{\text{img}, \text{txt}\}$ sampled, where $\oplus$ denotes prompt assembly, and let $s(\cdot,\cdot)$ be cosine similarity. With gate $\tau = 0.7$,

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{retrieval}} + w_c\, \mathcal{L}_{\text{conflict}} + w_a\, \mathcal{L}_{\text{align}}, \tag{7}$$

$$\mathcal{L}_{\text{conflict}} = \frac{1}{|\mathcal{M}_c|}\sum_{i\in\mathcal{M}_c} \max\big(0,\; s(\mathbf{z}_i^{\text{mm}}, \mathbf{z}_i^{\text{alt}}) + m_c\big), \qquad \mathcal{L}_{\text{align}} = \frac{1}{|\mathcal{M}_a|}\sum_{i\in\mathcal{M}_a} \big(1 - s(\mathbf{z}_i^{\text{mm}}, \mathbf{z}_i^{\text{alt}})\big), \tag{8}$$

with $\mathcal{M}_c = \{i : c_i < \tau\}$, $\mathcal{M}_a = \{i : c_i \geq \tau\}$, $w_c = 0.5$, $w_a = 0.2$, $m_c = 0.3$. The hinge's gradient is $+1$ in s whenever $s > -m_c$, actively pushing the conflicting views of the *same product* apart (target $s < -0.3$)—turning the product's own noisy view into a margin-armed adversarial negative—while high-consistency views are pulled together. Mode II remains a mechanism specification pending a calibrated performance evaluation (held-results disclosure, Appendix G).

Mode III: consistency as dense GRPO reward (TRIAD-GRPO).

Definition 3 (Policy-dependent TRIAD reward)**.** *For prompt (product) i, the policy π_θ samples $G = 8$ rollouts $o_1, \ldots, o_G$, each parsed into an attribute dictionary $\mathcal{A}^g$. Against frozen reference views extracted once per prompt,*

$$r_g^{\text{triad}} = \tfrac{1}{2}\Big[c\big(\mathcal{A}^g, \mathcal{A}_{\text{ref}}^{\text{img}}\big) + c\big(\mathcal{A}^g, \mathcal{A}_{\text{ref}}^{\text{txt}}\big)\Big], \qquad r_g = \omega_r\, r_g^{\text{retr}} + \omega_c\, r_g^{\text{triad}}, \tag{9}$$

with $A_g = (r_g - \mu_{\text{group}})/\sigma_{\text{group}}$ entering the standard clipped GRPO objective (Shao et al., 2024). A rollout that fails to parse is assigned $r_g = -0.1$, unscaled, before the mixture.

Policy-dependence is a design requirement, not a detail. An early version scored rollouts through a slot that ignored their content, so every rollout of a prompt got the same constant (the no-shared value $0.5 + \beta\beta_b\gamma$), within-group variance collapsed to zero, and GRPO had no gradient. Scoring the generated dictionary against both reference views makes r_g^{triad} a function of what the policy wrote—within-group reward variance stays positive across all seven smoke steps (0.0012–0.0038; Figure 3), whereas a retrieval reward is zero for most rollouts.

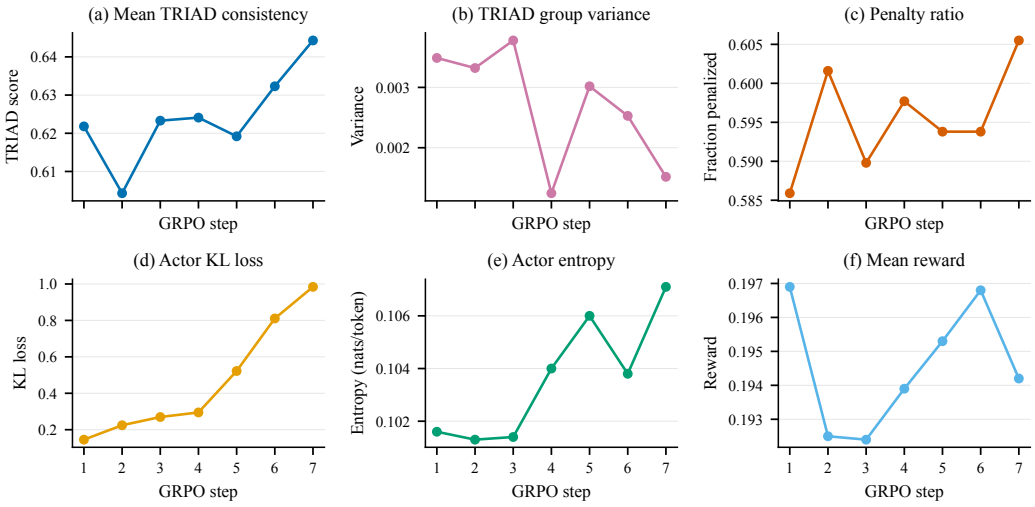

Figure 3: **TRIAD-GRPO 7-step smoke run** (Mode III signal-health check; pure-TRIAD reward ω_c=1, G=8 rollouts per prompt). (a) Mean TRIAD rises $0.622 \rightarrow 0.644$ (endpoints; non-monotone over seven steps). (b) Within-group TRIAD variance stays positive (0.0012–0.0038)—the property GRPO requires and a sparse Recall@k reward lacks. (c–f) Penalty ratio, actor KL (rising $0.145 \rightarrow 0.984$), entropy, and mean reward.

Mode III extension: view-conditioned extraction policy.

Definition 4 (View-conditioned extraction policy)**.** *Fix one instruction template with a text slot and an image slot. The view prompt x^v fills the slots as $x^{\text{img}} = (I, \varnothing)$, $x^{\text{txt}} = (\varnothing, T)$, $x^{\text{mm}} = (I, T)$, where $\varnothing$ denotes the blanked slot (title replaced by the empty string; image absent). A single policy π_θ—the same 3B-parameter VLM scale as the RL-stage generator of the strongest attribute-enhanced baseline—is trained across all three views: each product contributes three prompt rows, each row forming its own GRPO group of $G = 8$ rollouts. Each view is scored against its* own *modality's frozen reference,*

$$r_g^{\text{triad}}(v) = \begin{cases} c\big(\mathcal{A}^g, \mathcal{A}_{\text{ref}}^{\text{img}}\big), & v = \text{img}, \\ c\big(\mathcal{A}^g, \mathcal{A}_{\text{ref}}^{\text{txt}}\big), & v = \text{txt}, \\ \frac{1}{2}\big[c\big(\mathcal{A}^g, \mathcal{A}_{\text{ref}}^{\text{img}}\big) + c\big(\mathcal{A}^g, \mathcal{A}_{\text{ref}}^{\text{txt}}\big)\big], & v = \text{mm}, \end{cases} \tag{10}$$

entering $r_g = \omega_r\, r_g^{\text{retr}} + \omega_c\, r_g^{\text{triad}}(v)$ with advantages standardized within each view-row's group; the multimodal row is exactly Definition 3. Design commitments: one policy (not three specialists) serves all three extractions at one third the serving cost with cross-view parameter transfer; view-matched references avoid rewarding inference of title-only facts from pixels (a hallucination incentive); blanking enters through the data, requiring no model surgery—only a per-sample tolerant multi-modal-input concatenation for mixed-modality micro-batches. Disclosed limits: the unimodal references are the frozen 27B teacher's same-modality outputs (distillation, not verification), and the lazy-rollout loophole of Table 5 persists per view (empty dictionary $\Rightarrow 0.5 + \beta\beta_b\gamma = 0.54$ against a non-empty reference), so the 3-view run is evaluated first in the pure-TRIAD configuration; the retrieval leg's MMEB prompt embeds image and title regardless of view, leaving view-consistent retrieval rewards open.

F.2 TRIAD-GRPO SMOKE AND GENERATOR-ABLATION DETAIL

A 7-step smoke run of TRIAD-GRPO (Mode III) establishes that the dense reward is *non-degenerate*: mean TRIAD rises from 0.622 to 0.644 with positive within-group variance throughout (Figure 3), in contrast to Recall@k rewards that are zero for most rollouts. This smoke result is signal-health evidence only—too few steps to measure retrieval transfer.

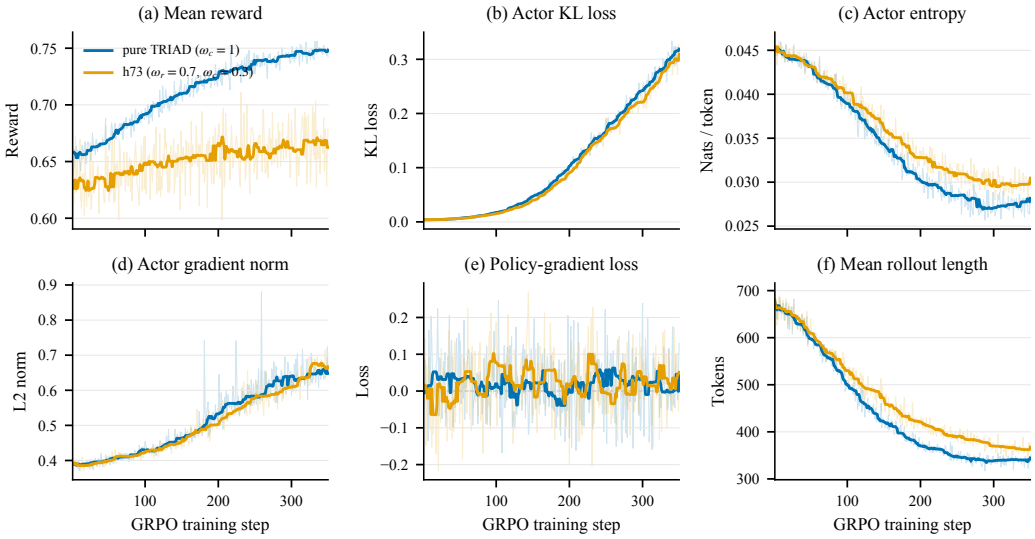

Figure 4: **TRIAD-GRPO 350-step training dynamics** (raw per-step values faint, solid curves are centered 15-step rolling medians). (a) Mean reward: pure TRIAD rises $0.661 \rightarrow 0.746$ (first/last 50-step means; peak 0.756 at step 335), while the h73 composite (ω_r=0.7 retrieval, ω_c=0.3 TRIAD) moves $0.632 \rightarrow 0.665$ with larger fluctuations. (b) Actor KL rises smoothly; (c) entropy declines; (d) gradient norm trends upward; (e) the clipped policy-gradient contribution fluctuates around zero throughout; (f) mean rollout length compresses over the run, the training-side counterpart of the eval-sample compression in §3.5. Optimization-health evidence only, not a downstream-retrieval claim; AFMRL's historical RL run has no archived stepwise training trace, so the panels cover the two TRIAD-GRPO reward arms.

Generator quality improves substantially: mean TRIAD rises to 0.6916 (pure) / 0.6837 (h73) from the SFT-10k warm start's 0.6265 ($+0.065$ / $+0.057$); parse failures drop from 3.41% to 0.36% / 0.53%; injection coverage reaches 98.4% / 97.2% (vs. 89.2% for SFT-10k)—these three are measured on the GRPO generator-quality eval sample. The pure arm's reward rises from 0.661 to 0.746 over the first/final 50-step windows (peak 0.756 at step 335; Figure 4). Its KL rises from approximately 0.003 to 0.31, entropy declines from 0.045 to 0.027 nats/token, and gradient norm increases smoothly from 0.39 to 0.65, while the clipped policy-gradient contribution continues to fluctuate around zero. Within-group TRIAD variance stays positive at every step of the pure run (min 0.000285, max 0.003416). The h73 composite reward moves from 0.632 to 0.665 with larger fluctuations, although its TRIAD component rises from 0.665 to 0.738. These are optimization-health observations only: they support stable, non-stalled policy updates, but do not establish a downstream retrieval gain or rule out all forms of reward misspecification. AFMRL's historical run has no archived stepwise KL, policy-gradient, entropy, or gradient-norm trace, so direct loss-curve comparison is unavailable.

F.3 Deployment cost (from §5)

The evidence in this paper distinguishes demonstrated encoder and generator improvements from mechanisms that remain unevaluated. This section specifies the resulting deployment cost and evidence scope.

Deployment cost. The online hot path is structurally identical to AFMRL: one generator forward pass producing entities, then one encoder forward pass consuming title + injected entities. Only the *multimodal* view's attributes are injected at inference; the text-only and image-only 27B reference views and the TRIAD scorer are training-side assets—pre-extracted once, stored statically, zero online latency. The query-side representation is cacheable and gallery re-indexing amortizes over the catalog's lifetime. RL training additionally *compresses* the generator's output—response length halves from 697 to 339 tokens over the 350-step pure run ($697 \rightarrow 376$ for h73), with a

shorter injected entity region ($98/93$ vs. 133 characters for the SFT warm start)—so the encoder's input sequence, and its forward cost, are *lower* than those of a naive long-injection baseline. The representation stays multimodal (VLM2Vec encodes image + text + injected attributes), so CMAC adds no modality or infrastructure beyond the base system.

F.4 Future work

Future work should evaluate calibrated Mode II conflict supervision and test the same three-view scoring rule under independently released image-query retrieval protocols.

F.5 Three-view decomposition statistics

Table 6 gives the per-item breakdown underlying the M5Product population statistics reported in §3.3. On 1,330,685 test items the extractor emits 8.73 image attributes, 5.68 text attributes, and 8.67 multimodal attributes on average. The unimodal union plus synergistic attributes partitions into 44.7% shared, 51.8% unique, and 3.5% synergistic; 59.2% of shared attributes conflict, and the multimodal view resolves 69.3% of those conflicts.

Table 6: **Three-view attribute decomposition on M5Product** (1,330,685 test items, Qwen3.5-27B extraction). Columns report per-item averages and fractions of the unimodal union plus synergistic attributes.

View	Attrs/item	Fraction	Role in TRIAD
Image-only	8.73	—	Provides $\mathcal{U}^{\mathrm{img}}$ and half of $\mathcal{K}^{\cap}$
Text-only	5.68	—	Provides $\mathcal{U}^{\mathrm{txt}}$ and half of $\mathcal{K}^{\cap}$
Shared (co-observed)	4.50	44.7%	Conflict-free rate ϕ measures quality
Unique (one view only)	5.41	51.8%	Complementarity γ is rewarded, not penalized
Synergistic (mm-only)	0.36	3.5%	Surfaces information neither view extracts alone
Conflicts among shared	2.67	59.2% of shared	Penalized by attribute reliability
Resolved by mm view	1.85	69.3% of conflicts	Arbitration verdict consumed downstream

G Backfill registry (pending experiments)

Held-results disclosure policy (2026-09-22): underperforming results—arms, curves, and the analysis built on them—enter this paper only alongside a completed improved run, and each exclusion is disclosed here rather than silent. Currently held: (i) four title-only recipe arms for same-product matching on an e-commerce platform (margin $m{=}0.1/0.2/0.3$), evaluated against the AFMRL reference under a derived split fingerprint, each trailing it with a paired bootstrap CI excluding zero (archived per-query sidecars; paths available in the released artifacts), to re-enter only with a completed improved recipe; (ii) the miscalibrated-gate CDARL Mode II evaluation, entering together with the quantile-calibrated retrain.

H MODE-I WEIGHTING MECHANICS

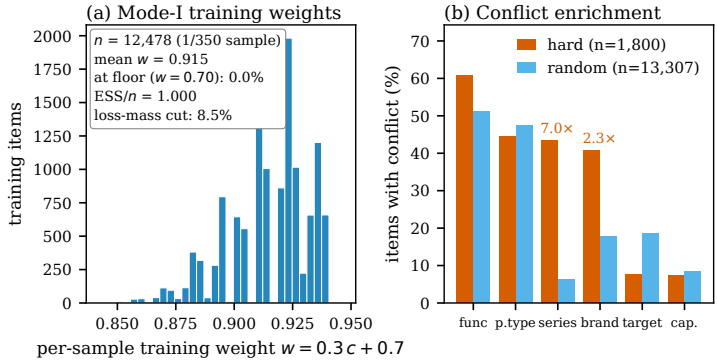

Figure 5: **Mode-I weighting mechanics.** (a) Distribution of the per-sample Mode-I training weight $w = 0.3c + 0.7$ (c = TRIAD reliability score) on a $1/350$ systematic sample of the $4{,}367{,}101$-item training pool ($n = 12{,}478$). The weight concentrates near the cap (mean 0.915, range $[0.839, 0.940]$, 0% at the 0.70 floor, ESS/$n = 0.9996$): TRIAD weighting is a *gentle reweighting* reshaping 8.5% of loss mass, not a hard filter. (b) Per-key image↔text conflict rates on $1{,}800$ retrieval-hard queries vs. a $13{,}307$-item random baseline: discriminative keys *series* (43.3% vs. 6.2%, $7.0\times$) and *brand* (40.6% vs. 17.8%, $2.3\times$) enrich sharply in failures; generic keys do not.

Two readings. **(1) The weighting is gentle by construction.** Because the TRIAD reliability score rarely binds below its cap on this training pool, the induced weights occupy a narrow band near 0.94 and discard only 8.5% of loss mass—the mechanism regularizes emphasis rather than excluding samples, so it cannot explain large gains (or harms) by itself. **(2) Failure enrichment is key-specific.** The conflict signal concentrates on exactly the discriminative keys that separate near-duplicate listings (series, brand), corroborating that TRIAD's conflict channel measures attribute defects where they matter for retrieval, not generic extraction noise.