BEYOND AGREEMENT: CONTRASTIVE EVIDENCE CALIBRATION FOR COMPOSED IMAGE RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Composed image retrieval requires preserving relevant reference-image attributes while satisfying a textual modification. Multiple query branches provide complementary retrieval signals, but their agreement may reflect shared appearance cues rather than support for the requested change. We propose Contrastive Evidence Calibration (CEC), a score-level approach that distinguishes branch disagreement, modification matching, reference resemblance, and composed-query compatibility. Starting from a fixed contrastive offset, CEC learns five bounded coefficients over frozen retrieval signals. A regularized full-gallery objective controls the adjustment, while an optional group objective emphasizes distinctions among competing candidates. We characterize the objective in effective coefficient space and derive a sufficient condition for preserving pairwise rankings. With independent training on each dataset, three-seed score ensembles achieve mean recall of 68.82 on the complete FashionIQ validation galleries and 82.43 on the official CIRR test. Matched comparisons show that the fixed offset accounts for the FashionIQ gain, while the complete calibrator improves its CIRR parent by 2.21 points. Applying CIRR's group objective to automatically constructed FashionIQ groups reduces recall, motivating separate choices for the score representation and its local supervision.

1 INTRODUCTION

Composed image retrieval (CIR) retrieves a target image using a reference image and a textual modification (Vo et al., 2019; Wu et al., 2021; Liu et al., 2021). The reference supplies visual context, while the text specifies the desired change. For a black short-sleeved shirt queried with "has longer sleeves," retrieval should preserve the relevant appearance while changing sleeve length. The central challenge is to determine which reference attributes remain relevant after applying the modification.

Existing methods address this challenge through learned image–text fusion, sentence-level prompting, and dual-query representations (Vo et al., 2019; Bai et al., 2024; Wen et al., 2024). Other approaches explicitly resolve conflicts between reference attributes and the modification (Tian et al., 2025; Chen et al., 2025). These methods emphasize query composition. In systems with multiple query branches, score fusion introduces a further ambiguity. Several branches can favor a candidate because of the same reference appearance, even when they provide different support for the requested change. A high combined score alone does not distinguish these sources of agreement.

We examine whether a branch adds modification support beyond the appearance already favored by the retriever. The intuition resembles testimony in the social deduction game Werewolf: two witnesses repeating one observation need not supply two independent reasons for a decision. Similarly, an image-based branch and a description-based branch can both favor a black garment while differing in their support for longer sleeves. Figure 1 illustrates this distinction. The analogy motivates contrasting branch scores; the method itself ranks candidates using continuous evidence channels.

We introduce Contrastive Evidence Calibration (CEC), a compact adjustment to a pretrained, task-adapted retrieval system, which we refer to as the parent. Five channels represent a contrast between parent scores, disagreement between image-based and description-based queries, direct modification matching, reference resemblance, and composed-query matching. A fixed offset adjusts the relative contribution of the query paths and the modification signal. Five bounded coefficients then refine

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.



Figure 1: **Agreement can repeat the wrong reason.** A longer-sleeve query illustrates the difference between reference resemblance and modification support. The witness statements are schematic, not generated model explanations. The reference and target are a FashionIQ validation pair; the target ranks, 11 before calibration and 8 afterwards, are measured over the complete shirt gallery.

this score, allowing reference resemblance and modification support to receive different weights. All backbone and parent-head parameters remain frozen during calibration.

The calibration objective determines which candidate comparisons drive this adjustment. Full-gallery supervision compares the annotated target with the entire retrieval pool. An optional group objective emphasizes a small set of alternatives, while a coefficient penalty limits departures from the initial score. CIRR supplies annotated candidate groups; FashionIQ requires constructing them automatically. A shared scoring architecture allows these supervision choices to be evaluated separately. For fixed features and groups, nonnegative loss weights, and positive regularization, the objective is strongly convex in its effective coefficients. A margin bound further characterizes when calibration preserves a pairwise order.

Experiments use independently trained FashionIQ and CIRR systems, complete validation galleries on FashionIQ, and the official CIRR test server. The resulting score ensembles achieve Means of 68.82 and 82.43, respectively. Matched controls identify different sources of improvement: the fixed offset already reaches the FashionIQ result, while the complete calibrator improves the CIRR parent by 2.21 points. Directly transferring CIRR's group objective to automatically constructed FashionIQ groups reduces recall. This asymmetry motivates treating local supervision as a dataset-dependent choice within the shared score model.

Our contributions are:

- An explicit contrastive score representation for calibrating branch support, modification matching, and reference resemblance with five bounded coefficients.
- A regularized calibration objective combining global retrieval with optional local competition, together with an analysis of its coefficient-space geometry and ranking margins.
- A two-dataset evaluation with matched FashionIQ controls that separate fixed offsets, learned calibration, auxiliary representations, and group supervision.

2 Related Work

Representing a composed query. TIRG combines image and text features through gating and residual composition (Vo et al., 2019). SPRC constructs sentence-level prompts (Bai et al., 2024), while DQU-CIR performs multimodal fusion at the raw-data level (Wen et al., 2024). CCIN resolves conflicts between reference appearance and the requested modification (Tian et al., 2025); OFFSET revises visual focus through segmentation (Chen et al., 2025). IMAGINE enriches composition with schema imagery (Huang et al., 2026). These approaches primarily develop query representations. CEC calibrates the candidate scores produced by an existing retrieval system, holding its representations fixed during the final training stage.

Learning from difficult or ambiguous candidates. QuRe studies query-relevant hard-negative sampling (Kwak et al., 2025), and TME addresses noisy triplet correspondence (Li et al., 2025). ReCALL generates corrections from retrieval failures (Yang et al., 2026); FIRE-CIR uses attribute questions for fine-grained reranking (Gardères et al., 2026). These methods address ambiguity in the relationship between a query and its candidate images. In particular, an unannotated candidate can satisfy an instruction even when the benchmark specifies a single target. Our group objective uses the available target labels, and the FashionIQ comparisons evaluate the effect of constructing competing groups automatically.

Score fusion and controlled adaptation. Reciprocal rank fusion combines rankings without relying on raw score scales (Cormack et al., 2009); learned score combinations also provide strong retrieval baselines (Bruch et al., 2024). Orthogonal projection has been explored in CIR (Cheng & Liu, 2026). CEC uses constrained linear fusion with an explicit contrastive representation, initialization, and regularization. The parent additionally contains a residual update constrained by KL divergence, an instance of controlled distributional adaptation (Schulman et al., 2015). This inherited component remains fixed in all matched calibration comparisons.

3 CONTRASTIVE EVIDENCE CALIBRATION

3.1 TASK AND FROZEN RETRIEVAL SIGNALS

Let $q = (I_r, t)$ denote a reference image and a modification instruction. For a dataset gallery $\mathcal{G}$, define $\mathcal{G}_q$ by excluding the reference image. Training provides an annotated target $y \in \mathcal{G}_q$. At inference, the system scores each candidate using the query and candidate representations, without access to the test target label.

Our primary encoder is a task-adapted Qwen2.5-VL-7B model (Bai et al., 2025). One query path receives the reference image with the instruction, and the other receives a reference-only description with the instruction. Separate residual adapters produce v_q and t_q in a 3,584-dimensional retrieval space. For candidate feature g_j, define

$$V_{qj} = v_q^\top g_j, \qquad T_{qj} = t_q^\top g_j, \qquad D_{qj} = T_{qj} - V_{qj}. \tag{1}$$

Both query paths share the same 7B backbone. Their initial mixture assigns weights 0.9 and 0.1 to the image-based and description-based paths, respectively. A positive D_{qj} indicates that the description-based path assigns candidate j a higher score; it does not establish whether that candidate satisfies the modification.

An auxiliary SigLIP2 SO400M-384 encoder (Tschannen et al., 2025) provides reference-image and modification-text features r_q and e_q, and candidate features h_j. The two encoder towers remain frozen. A previously trained composition head H combines r_q and e_q; the corresponding scores are

$$R_{qj} = r_q^\top \bar{h}_j, \qquad E_{qj} = e_q^\top \bar{h}_j, \qquad C_{qj} = H(r_q, e_q)^\top \bar{h}_j, \tag{2}$$

where $\bar{h}_j = h_j - |\mathcal{G}|^{-1} \sum_{k \in \mathcal{G}} h_k$ centers candidate features using the split gallery mean. Centering subtracts a query-dependent constant from each score channel and therefore preserves candidate order within that channel. R measures reference resemblance, E measures modification-text matching, and C measures the learned composition of reference and text.

The parent score A combines these representations with the previously trained query, gallery, fusion, and candidate heads. The auxiliary score B comes from an earlier configuration of the same task-specific system, before the later heads and SigLIP2 branch. Both scores remain fixed during calibration. Their construction is detailed in Appendix A. The contrast $B - A$ permits calibration to adjust the contribution of the additional parent components.

3.2 A CONTRASTIVE SCORE REPRESENTATION

We initialize calibration with a fixed offset,

$$s_{qj}^0 = A_{qj} + \eta D_{qj} + \xi E_{qj}, \qquad \eta = \xi = 0.2, \tag{3}$$

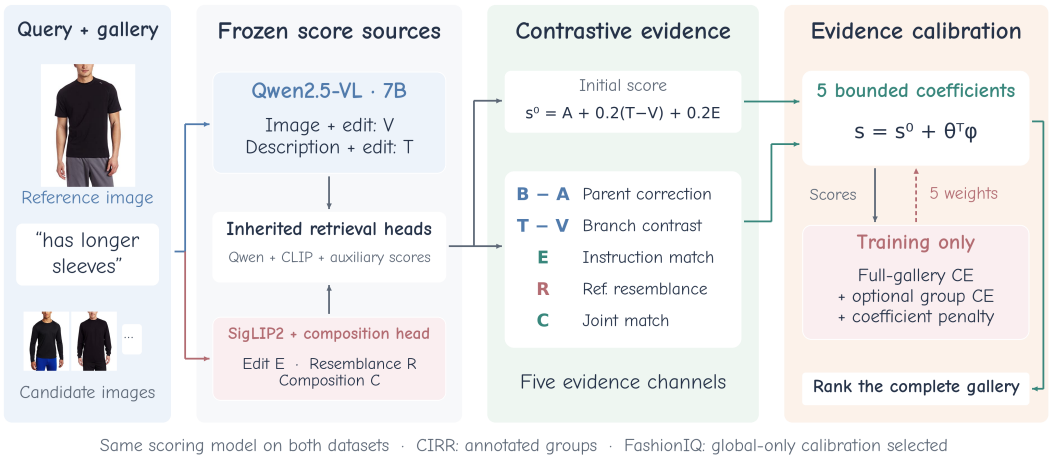

Figure 2: **A shared score model with dataset-specific supervision.** Task-adapted Qwen paths and frozen auxiliary representations supply candidate scores. CEC learns five bounded coefficients over their contrasts. Full-gallery and optional group losses update only these coefficients in the final stage; all inherited weights are frozen. At inference, a single calibrated score ranks the complete gallery. CIRR's subset output filters that same ranking to the official group.

and collect five evidence channels,

$$\phi_{qj} = \left[B_{qj} - A_{qj},\ D_{qj},\ E_{qj},\ R_{qj},\ C_{qj}\right]^\top. \tag{4}$$

The final score is

$$s_{qj} = s^0_{qj} + \theta^\top \phi_{qj}, \qquad \theta = \ell \odot \tanh z, \qquad \ell = (1, 0.6, 0.5, 0.3, 0.5)^\top. \tag{5}$$

Only $z \in \mathbb{R}^5$ is optimized during calibration. The resulting coefficients are shared across queries and, on FashionIQ, across all three categories. Equation 5 uses candidate scores alone; target labels enter the training loss, and group membership is not an input to the scoring function.

Each channel controls a different part of the score. D adjusts the balance between the image-based and description-based queries, while E provides direct modification matching. A negative coefficient for R reduces the preference for reference resemblance conditional on the remaining scores. C retains joint visual and textual matching. Regularization favors small corrections when these adjustments provide little training benefit.

Equation 5 is equivalently a constrained linear combination,

$$s = (1-\theta_1)A + \theta_1 B + (0.2+\theta_2)(T-V) + (0.2+\theta_3)E + \theta_4 R + \theta_5 C. \tag{6}$$

Equation 6 expresses CEC as constrained linear fusion. Its channels need not be statistically independent: the contrastive representation specifies the initial scoring rule and the coordinates in which departures are penalized. The coefficients consequently describe conditional score adjustments, rather than causal effects of individual image attributes.

3.3 Global retrieval and local competition

For a candidate set $\mathcal{S}$ containing y, define

$$\mathcal{L}_{\mathcal{S}}(q, y) = -\log \frac{\exp(s_{qy}/\tau)}{\sum_{j\in\mathcal{S}} \exp(s_{qj}/\tau)}, \qquad \tau = 0.035. \tag{7}$$

The training objective combines the entire gallery with an optional five-candidate group $\mathcal{H}_q$:

$$\mathcal{L}(\theta) = \frac{1}{n}\sum_{q=1}^{n} \left[\mathcal{L}_{\mathcal{G}_q}(q, y_q) + \lambda \mathcal{L}_{\mathcal{H}_q}(q, y_q)\right] + \gamma \|\theta\|_2^2. \tag{8}$$

The global term maintains competition with the complete gallery. The group term increases the weight of comparisons within $\mathcal{H}_q$, and the penalty shrinks the effective coefficients toward zero. Together, these terms adjust the initial scoring rule in Equation 3 while retaining the full retrieval objective.

After removing the reference image, each CIRR group contains five candidates, including the annotated target. FashionIQ provides no equivalent groups. We construct two training alternatives: the target together with the four closest non-target images to the reference in SigLIP2 space, or the target together with the four highest-scoring non-target images under A. Both constructions exclude the reference image and use only the training split. Evaluation always uses the complete validation galleries.

The selected FashionIQ model uses $\lambda = 0$ and $\gamma = 200$, corresponding to global calibration without artificial-group supervision. CIRR uses $\lambda = 16$ and $\gamma = 0.2$. The score representation, coefficient limits, temperature, learning rate, and optimizer schedule are shared across datasets. Section 5 compares alternative supervision strengths; Appendix B describes configuration selection.

3.4 Computation and parameter sharing

Given the initial score and five evidence channels, the calibration correction requires five multiplications and a short sum per candidate, or $O(5|\mathcal{G}_q|)$ work per query. Candidate representations can be cached, and calibration introduces no image-pair cross-encoder. Training reuses these scores for the full-gallery loss; the group loss applies a second softmax to five selected entries. Representation extraction and the inherited heads remain part of the overall inference cost.

The coefficients are global, so every query receives the same calibration rule. This limits parameter count but prevents query-specific weighting of heterogeneous instructions. Our three-seed ensemble averages candidate scores from three independently fitted sets of downstream heads sharing one task-adapted backbone. Parent and calibrated ensembles contain the same three downstream configurations, allowing their difference to measure the final score adjustment under a matched ensemble budget.

4 Understanding the calibration objective

Gradient interpretation. Let $p_{q,\mathcal{S}}$ be the softmax distribution over $\mathcal{S}$. Differentiating a retrieval loss with respect to the effective coefficients gives

$$\nabla_\theta \mathcal{L}_\mathcal{S} = \frac{1}{\tau}\left(\mathbb{E}_{j\sim p_{q,\mathcal{S}}}[\phi_{qj}] - \phi_{qy}\right). \tag{9}$$

For this loss term, gradient descent increases a coefficient when the target's corresponding feature exceeds its expectation under the competing-candidate distribution, and decreases it in the reverse case. The global and group losses compare the same target against different distributions. A visually similar group may emphasize modification-sensitive distinctions, but automatically mined groups can also contain plausible unannotated matches. The group objective therefore requires evaluation separately from the score parameterization.

Proposition 1 (Geometry in effective coefficient space)**.** *For fixed score features, fixed training groups, and $\lambda \geq 0$, Equation 8 is 2γ-strongly convex in θ whenever $\gamma > 0$. Its restriction to the closed box $|\theta_k| \leq \ell_k$ has a unique minimizer.*

Proof. The Hessian of each candidate-set loss is

$$\nabla^2_\theta \mathcal{L}_\mathcal{S} = \tau^{-2}\,\mathrm{Cov}_{p_{q,\mathcal{S}}}(\phi) \succeq 0. \tag{10}$$

The nonnegative weighted sum remains positive semidefinite, and the penalty adds $2\gamma I$. Continuity on the compact box ensures existence; strong convexity ensures uniqueness. □

The proposition applies to effective coefficients θ. The implemented AdamW optimization uses $\ell \odot \tanh z$, so convexity in z and exact attainment of the coefficient-space minimizer do not follow. Equation 10 nevertheless identifies the source of curvature: variation in evidence features among score-weighted competing candidates, together with the coefficient penalty.

Proposition 2 (A sufficient condition for preserving a pairwise order)**.** *For candidates i, j with initial margin $m^0_{ij} = s^0_{qi} - s^0_{qj} > 0$, calibration preserves the order $s_{qi} > s_{qj}$ if*

$$m^0_{ij} > \sum_{k=1}^{5} |\theta_k| \, |\phi_{qi,k} - \phi_{qj,k}|. \tag{11}$$

Proof. The score-margin change is $\theta^\top(\phi_{qi} - \phi_{qj})$. The triangle inequality bounds its magnitude by the right-hand side, so it cannot reverse a larger positive initial margin. □

The bound relates ranking stability to the initial margin and the magnitude of the evidence correction. Close candidates may change order under updates that leave larger margins intact. Group supervision changes the relative weight of candidate comparisons, whereas regularization penalizes the resulting coefficient adjustment. Uniform, query-specific offsets of the evidence channels leave rankings and candidate-set losses unchanged because they add the same constant to every candidate score.

5 Experiments

5.1 Datasets, training, and evaluation

FashionIQ. We evaluate the dress, shirt, and top/tee categories of FashionIQ (Wu et al., 2021). Calibration uses 18,000 training queries. Evaluation includes all 6,016 validation queries and the complete category galleries of 3,817, 6,346, and 5,373 images, respectively. We report R@10 and R@50 for each category. Mean denotes their unweighted six-value average, giving equal weight to categories and recall cutoffs.

CIRR. All task-specific weights are trained using the 28,225 CIRR training queries (Liu et al., 2021). The backbone is initialized from generic Qwen2.5-VL-7B-Instruct, independently of FashionIQ. Configurations are selected on the official validation split and evaluated on the hidden test server, which covers 4,148 queries and 2,315 gallery images. Mean is $(\mathrm{R@5} + \mathrm{R_s@1})/2$. Full-gallery and subset rankings use the same candidate scores. We use the M-BEIR processed JPEG images (Wei et al., 2024) with the original CIRR identifiers.

Training scope and model selection. Calibration freezes the task-adapted 7B model, CLIP, SigLIP2, and all parent heads. Each five-parameter calibrator is trained for 12 epochs with batch size 128, learning rate 0.001, AdamW weight decay 0.01, and gradient-norm clipping at 1. The fixed offset and CIRR configuration are selected on CIRR validation. FashionIQ selection compares seven initial configurations and six settings with stronger regularization, yielding the no-group model reported here. Appendix B provides the complete search and selection procedure.

We use downstream training seeds 42, 43, and 44, with one shared task-adapted backbone per dataset. The single-model summary averages recall across these three runs. The score ensemble instead averages candidate scores before ranking. Validation is used for model selection, so the reported uncertainty intervals are conditional on the fitted models and selected configurations.

5.2 FashionIQ retrieval accuracy

CEC obtains a three-seed single-model mean of 68.72 and a score-ensemble Mean of 68.82 (Table 1). The parent ensemble obtains 68.51. Ensemble R@10 increases from 58.88 to 59.10, and R@50 from 78.15 to 78.53. The Mean improvement is 0.306 percentage points, with a conditional 95% bootstrap interval of $[0.055, 0.551]$. The interval uses 5,000 resamples grouped by reference image within each category.

Table 1 places these results alongside selected representative approaches to query composition, hard-negative sampling, correspondence learning, and reranking. DQU-CIR provides a dual-query fusion comparison using its Original-Split evaluation. Our system combines a 7B backbone with CLIP and SigLIP2 encoders and a downstream score ensemble. The following matched-parent study isolates calibration while holding this representation and ensemble budget fixed.

Table 1: **FashionIQ validation recall (%).** Literature rows show selected representative methods. DQU-CIR uses its Original-Split results, matching our complete category galleries. Single denotes mean recall across three downstream seeds. The ensemble shares one backbone but retains three sets of downstream heads. Mean averages the six category recalls; values are rounded for display.

	Dress		Shirt		Top/tee		
Method	R@10	R@50	R@10	R@50	R@10	R@50	Mean
SPRC (Bai et al., 2024)	49.18	72.43	55.64	73.89	59.35	78.58	64.85
DQU-CIR (Wen et al., 2024)	51.90	74.37	53.57	73.21	58.48	79.23	65.13
QuRe (Kwak et al., 2025)	46.80	69.81	53.53	72.87	57.47	77.77	63.04
TME (Li et al., 2025)	49.73	71.69	56.43	74.44	59.31	78.94	65.09
CCIN (Tian et al., 2025)	49.38	72.58	55.93	74.14	57.93	77.56	64.59
ReCALL (Yang et al., 2026)	51.81	73.48	58.49	76.59	60.83	79.19	66.73
FIRE-CIR (Gardères et al., 2026)	54.09	75.41	58.68	75.76	61.70	80.88	67.75
CEC (single, mean of 3 seeds)	53.79	76.10	59.34	77.09	63.76	82.25	68.72
CEC (3-seed score ensemble)	53.74	76.20	59.52	76.89	64.05	82.51	68.82

Table 2: **Matched FashionIQ calibration study.** All rows share the same frozen parent representations and complete galleries. The first three numeric columns average three individual models; the last column evaluates their score ensemble. Reference and retrieval groups are constructed from training data only.

Configuration	R@10	R@50	Mean	Ensemble
Frozen parent	58.89	78.10	68.49	68.51
Fixed contrastive offset	58.95	78.49	68.72	68.82
Calibration, $\lambda = 0, \gamma = 0.2$	58.64	77.99	68.32	68.41
Reference groups, $\lambda = 16, \gamma = 0.2$	51.69	72.50	62.10	62.19
Retrieval groups, $\lambda = 1, \gamma = 0.2$	57.64	77.22	67.43	67.46
Reference groups, $\lambda = 16, \gamma = 200$	58.62	78.34	68.48	68.59
Retrieval groups, $\lambda = 16, \gamma = 200$	58.77	78.36	68.56	68.66
CEC: $\lambda = 0, \gamma = 200$	58.96	78.48	68.72	68.82

5.3 Ablation of calibration and supervision

Fixed contrasts explain the observed improvement. The fixed offset $0.2(T - V) + 0.2E$ yields an ensemble Mean of 68.82 without fitting new coefficients (Table 2). Strongly regularized global calibration matches this Mean, with small differences in individual rankings and R@10/R@50. Its coefficients remain close to zero. Thus the fixed scoring rule accounts for the observed FashionIQ improvement; coefficient fitting provides no additional aggregate gain at the selected setting.

Aggressive artificial-group supervision does not transfer. Transferring the CIRR setting $\lambda = 16, \gamma = 0.2$ to FashionIQ reference-neighbor groups reduces ensemble Mean to 62.19. Retrieval-score groups with $\lambda = 1, \gamma = 0.2$ obtain 67.46. At $\lambda = 16$, increasing γ to 200 recovers Means of 68.59 and 68.66 for reference and retrieval groups, respectively. The latter remains 0.159 points below the matched no-group model, with a conditional 95% difference interval of $[-0.273, -0.044]$. Thus neither selected group construction improves on global-only calibration in this comparison.

A five-candidate objective emphasizes local discrimination, whereas FashionIQ's metrics measure top-10 and top-50 retrieval over larger galleries. Visually similar non-targets may also satisfy the modification despite lacking a target annotation. These differences offer plausible explanations for negative transfer, although aggregate recalls do not isolate their individual effects.

Controlling for auxiliary representations. Adding the frozen SigLIP2 branch to the pre-calibration system raises the three-seed single-model Mean from 67.75 to 68.49. With the same encoder and training schedule, a simpler composition head obtains 68.45, only 0.041 points below the interaction head; the conditional interval for that difference crosses zero. The encoder-level gain is therefore distinct from the head-level comparison. All calibration ablations hold the selected composition head fixed, and their five-parameter count covers only the final calibration stage.

Table 3: **FashionIQ changes at the query level.** Parent and CEC use matched three-seed ensembles. A recovery enters the corresponding top-K set, and a regression leaves it. These counts include every validation query and expose changes hidden by the macro average.

	Mean		R@10 changes		R@50 changes	
Category	Parent	CEC	Recovered	Lost	Recovered	Lost
Dress	64.72	64.97	17	21	32	18
Shirt	68.35	68.20	22	24	11	15
Top/tee	72.46	73.28	34	15	18	5

Table 4: **CIRR official test recall (%).** Literature rows show selected representative methods. Our two rows are successive versions of the same three-seed ensemble, independently trained on CIRR. Mean is $(\mathrm{R@5} + \mathrm{R_s@1})/2$. Published configurations and their computational budgets differ.

	Full gallery				Subset			
Method	R@1	R@5	R@10	R@50	R_s@1	R_s@2	R_s@3	Mean
SPRC (Bai et al., 2024)	51.96	82.12	89.74	97.69	80.65	92.31	96.60	81.39
DQU-CIR (Wen et al., 2024)	46.22	78.17	87.64	97.81	70.92	87.69	94.68	74.55
QuRe (Kwak et al., 2025)	52.22	82.53	90.31	98.17	78.51	91.28	96.48	80.52
TME (Li et al., 2025)	53.42	82.99	90.24	98.15	81.04	92.58	96.94	82.02
OFFSET (Chen et al., 2025)	52.19	82.60	90.07	98.07	81.37	93.08	97.54	81.99
Frozen parent (3-seed ensemble)	53.93	83.61	91.66	98.51	76.84	90.75	96.48	80.23
CEC (3-seed score ensemble)	55.35	84.53	91.78	98.53	80.34	92.82	97.21	82.43

5.4 Category and query-level changes

The category-level effect is largest for top/tee, whose Mean increases by 0.816 points. Dress improves by 0.248, while shirt decreases by 0.147 (Table 3). Across categories, 73 targets enter the top ten and 60 leave it; at rank 50, the corresponding counts are 61 and 38. The aggregate improvement therefore combines recoveries and regressions, with different net effects across categories.

These counts describe the net effect of calibration on retrieval decisions; they do not measure the score margins in Proposition 2. The example in Figure 1 is a top-ten recovery within the shirt category, whose aggregate Mean decreases. It illustrates one corrected ranking, while Table 3 reports the balance of changes across all validation queries.

5.5 Official CIRR testing

On the official CIRR test, CEC achieves full-gallery R@1 of 55.349, R@5 of 84.530, and subset R@1 of 80.337, yielding Mean 82.4335. Relative to the matched parent ensemble, Mean increases by 2.205 points, R@5 by 0.916, and subset R@1 by 3.494. The complete calibration therefore improves both full-gallery recall and the top-ranked choice within the official candidate group.

The parent comparison holds representations and ensemble size fixed, measuring the combined effect of the initial offset and learned coefficients. It does not isolate the group-loss contribution. Table 4 also reports selected published configurations for context; their encoder and training budgets differ. Reporting full-gallery and subset recalls separately distinguishes retrieval coverage from discrimination among the supplied alternatives.

5.6 Qualitative example and coefficient analysis

The FashionIQ query in Figure 1 pairs a black short-sleeved shirt with the annotations "is long sleeved" and "has longer sleeves." Its target is a black long-sleeved garment. Under the reported score ensemble, calibration moves the target from rank 11 to rank 8 in the complete 6,346-image shirt gallery, converting a miss into an R@10 hit. The example was selected for visual clarity from queries with an observed R@10 recovery.

The learned coefficients differ consistently across datasets. CIRR assigns reference resemblance a coefficient of approximately -0.263 in all three runs and gives the additional modification-matching term a positive coefficient. FashionIQ's five corrections all have magnitude below 0.002 under strong regularization, explaining its similarity to the fixed-offset model. The CIRR coefficients reduce reference-similarity support conditional on the remaining channels, but do not identify the semantic cause of individual retrieval errors.

6 DISCUSSION

Shared scoring and dataset-specific supervision. The score representation and inference rule are shared across FashionIQ and CIRR, but their local supervision differs. CIRR's annotated groups and FashionIQ's mined neighbors define different candidate comparisons. The transfer results show that adopting the same score model does not make these comparisons interchangeable. Treating group supervision as optional allows its effect to be evaluated within a fixed architecture.

Scope and limitations. CEC calibrates a retrieval system containing multiple pretrained encoders and learned heads; its five coefficients account only for the final adjustment. Learned calibration provides no additional aggregate FashionIQ gain over the fixed offset at the selected setting. The CIRR parent comparison evaluates the complete adjustment and does not separately identify the benefit of group supervision. These findings limit claims about the incremental contribution of coefficient learning and local competition.

The theoretical results concern the score objective and pairwise margins, not semantic correctness or robustness to arbitrary input changes. Each dataset uses one task-adapted backbone, so downstream seeds do not quantify backbone-training uncertainty. Repeated validation-based development also means that bootstrap intervals are conditional on the selected models. Finally, comparisons with published methods do not control for encoder, supervision, or ensemble budgets.

7 CONCLUSION

We presented Contrastive Evidence Calibration, which adjusts branch support, modification matching, and reference resemblance through five bounded coefficients over a frozen retriever. The objective combines full-gallery supervision, optional group comparisons, and a coefficient penalty. FashionIQ and CIRR evaluations show gains over matched parent ensembles, with different sources of improvement: a fixed contrastive offset suffices on FashionIQ, while the complete calibrator improves CIRR test performance. The negative transfer of artificial-group supervision further supports evaluating candidate selection separately from score calibration.

AI USE DISCLOSURE

In this work, we used generative AI tools to polish the manuscript, including correcting grammatical errors and improving clarity and readability.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, et al. Qwen2.5-VL Technical Report. *arXiv:2502.13923*, 2025. URL https://arxiv.org/abs/2502.13923.

Yang Bai, Xinxing Xu, Yong Liu, Salman Khan, Fahad Khan, Wangmeng Zuo, Rick Siow Mong Goh, and Chun-Mei Feng. Sentence-level Prompts Benefit Composed Image Retrieval. In *ICLR*, 2024.

Sebastian Bruch, Siyu Gai, and Amir Ingber. An Analysis of Fusion Functions for Hybrid Retrieval. *ACM Transactions on Information Systems, 42(1), Article 20*, 2024. URL https://doi.org/10.1145/3596512.

Zhiwei Chen, Yupeng Hu, Zixu Li, Zhiheng Fu, Xuemeng Song, and Liqiang Nie. OFFSET: Segmentation-based Focus Shift Revision for Composed Image Retrieval. In *ACM Multimedia*, 2025. URL https://arxiv.org/abs/2507.05631.

Su Cheng and Guoyang Liu. SOP: Selective Orthogonal Projection for Composed Image Retrieval. *Sensors, 26(5), 1621*, 2026. URL https://www.mdpi.com/1424-8220/26/5/1621.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. Reciprocal Rank Fusion outperforms Condorcet and individual Rank Learning Methods. In *SIGIR*, 2009. URL https://cormack.uwaterloo.ca/cormacksigir09-rrf.pdf.

Luyu Gao, Yunyi Zhang, Jiawei Han, and Jamie Callan. Scaling Deep Contrastive Learning Batch Size under Memory Limited Setup. In *RepL4NLP*, 2021. URL https://aclanthology.org/2021.repl4nlp-1.31.pdf.

François Gardères, Camille-Sovanneary Gauthier, Jean Ponce, and Shizhe Chen. FIRE-CIR: Fine-grained Reasoning for Composed Fashion Image Retrieval. *arXiv:2604.09114v1*, 2026. URL https://arxiv.org/pdf/2604.09114v1.

Jiale Huang, Zixu Li, Zhiwei Chen, Zhiheng Fu, Chunxiao Wang, and Yupeng Hu. IMAGINE: Adaptive Schema-Imagery Enhanced Composition for Composed Video Retrieval. In *ICMR*, 2026. URL https://arxiv.org/abs/2606.08144v1.

Jaehyun Kwak, Ramahdani Muhammad Izaaz Inhar, Se-Young Yun, and Sung-Ju Lee. QuRe: Query-Relevant Retrieval through Hard Negative Sampling in Composed Image Retrieval. In *ICML*, 2025.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *ICML*, 2023. URL https://proceedings.mlr.press/v202/li23q/li23q.pdf.

Shuxian Li, Changhao He, Xiting Liu, Joey Tianyi Zhou, Xi Peng, and Peng Hu. Learning with Noisy Triplet Correspondence for Composed Image Retrieval. In *CVPR*, 2025.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. DoRA: Weight-Decomposed Low-Rank Adaptation. In *ICML*, 2024. URL https://proceedings.mlr.press/v235/liu24bn.html.

Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image Retrieval on Real-Life Images With Pre-Trained Vision-and-Language Models. In *ICCV*, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. In *NeurIPS Datasets and Benchmarks*, 2022. URL https://arxiv.org/abs/2210.08402.

John Schulman, Sergey Levine, Philipp Moritz, Michael Jordan, and Pieter Abbeel. Trust Region Policy Optimization. In *ICML*, 2015. URL https://proceedings.mlr.press/v37/schulman15.pdf.

Likai Tian, Jian Zhao, Zechao Hu, Zhengwei Yang, Hao Li, Lei Jin, Zheng Wang, and Xuelong Li. CCIN: Compositional Conflict Identification and Neutralization for Composed Image Retrieval. In *CVPR*, 2025.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features. *arXiv:2502.14786*, 2025. URL https://arxiv.org/abs/2502.14786.

Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing Text and Image for Image Retrieval - an Empirical Odyssey. In *CVPR*, 2019.

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. In *ECCV*, 2024. URL https://github.com/TIGER-AI-Lab/UniIR.

Haokun Wen, Xuemeng Song, Xiaolin Chen, Yinwei Wei, Liqiang Nie, and Tat-Seng Chua. Simple but Effective Raw-Data Level Multimodal Fusion for Composed Image Retrieval. In *SIGIR*, 2024. URL https://arxiv.org/abs/2404.15875.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *ICML*, 2022. URL https://proceedings.mlr.press/v162/wortsman22a/wortsman22a.pdf.

Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion IQ: A New Dataset Towards Retrieving Images by Natural Language Feedback. In *CVPR*, 2021.

Tianyu Yang, Chenwei He, Xiangzhao Hao, Tianyue Wang, Jiarui Guo, Haiyun Guo, Leigang Qu, Tat-Seng Chua, and Jinqiao Wang. ReCALL: Recalibrating Capability Degradation for MLLM-based Composed Image Retrieval. *arXiv:2602.01639v2*, 2026. URL https://arxiv.org/abs/2602.01639v2.

A PARENT RETRIEVER AND TRAINING DETAILS

A.1 INDEPENDENT TASK ADAPTATION

FashionIQ and CIRR use separate task-specific weights. The FashionIQ backbone is a previously adapted Qwen2.5-VL-7B model whose adapter is the equal-weight parameter average of two compatible rank-16 DoRA adapters (Liu et al., 2024; Wortsman et al., 2022), with 22,962,176 adapter parameters. Both are trained on FashionIQ: one continuation uses bounded model-generated query corrections, and the other uses a counterfactual curriculum. The former runs for 100 updates with effective batch size 256, learning rate 4×10^{-5}, and temperature 0.03. The latter runs for 100 updates with learning rate 2×10^{-5} and the same temperature. These backbone adaptations precede the calibration experiments.

The CIRR backbone is initialized from generic Qwen2.5-VL-7B-Instruct and adapted using CIRR training data only. Rank-16 DoRA uses scaling 64, dropout 0.1, and 300 updates with learning rate 2×10^{-5}, a 10% warmup, temperature 0.02, and effective batch size 512. Gradient caching (Gao et al., 2021) uses microbatches of 16. AdamW weight decay is 0.01, and the gradient norm is clipped at 1. The final backbone checkpoint is shared across the three downstream runs; no FashionIQ-trained weights enter the CIRR pipeline.

FashionIQ reference descriptions are generated by Qwen from reference images. CIRR uses the public DQU-CIR descriptions generated with BLIP-2 (Li et al., 2023; Wen et al., 2024). Neither procedure uses the target image to form a query description. These texts feed the description-based Qwen path, while the SigLIP2 modification feature encodes the instruction alone.

A.2 THE EARLY SCORE AND THE FROZEN PARENT

Each Qwen query adapter is a rank-32 residual map,

$$A_q(x) = \text{normalize}\left(x + W_{\text{up}}\,\text{GELU}(W_{\text{down}}x)\right). \tag{12}$$

The image-based and description-based adapters have separate weights and contain 458,752 parameters in total. Their outputs are mixed with weights 0.9 and 0.1, respectively. Initial training runs for ten epochs with learning rate 10^{-4}, temperature 0.07, and squared-drift penalty weight 10. A five-epoch continuation supervises both branches using temperature 0.035 and drift weight 1.

The CLIP ViT-B/32 auxiliary branch (Radford et al., 2021; Schuhmann et al., 2022) produces 512-dimensional image and text features. Rank-128 residual adapters and a gate combine reference-image and modification-text features. The head is trained for ten epochs at learning rate 10^{-4}. A five-epoch continuation also trains rank-8 low-rank updates in the final image and text MLP blocks, using learning rate 10^{-5} for those updates. Auxiliary scores enter with coefficient 0.2. A 513-parameter linear reranker then uses fixed 64-dimensional CLIP projections to adjust the retrieved top 128 candidates through a bounded $0.01\tanh(\cdot)$ output. Reranker training uses ten epochs, learning rate 10^{-4}, and penalty weight 1,000.

Let a denote the CLIP score contribution and b the remaining early score. The auxiliary score B is constructed using $p_0 = \text{softmax}(b/0.035)$. First, a and b are centered under p_0; the weighted projection of the centered a onto the centered b is removed, and the residual is normalized to obtain u. The update is $B = b + 0.035\alpha u$, where $p_\alpha \propto p_0 e^{\alpha u}$ and $\alpha \in [0, 1]$ satisfies $D_{\text{KL}}(p_\alpha \| p_0) \leq 0.005$. Sixteen bisection steps approximate the largest feasible α using the feasible lower endpoint. Anchor and residual variance floors are 10^{-12} and 10^{-14}. This inherited update remains fixed during calibration.

The parent score A additionally includes query and gallery corrections, a score gate, candidate-density correction, a spectral score adjustment, a candidate feature matcher, and the SigLIP2 composition head. Table 5 lists the corresponding parameter counts and training settings. All components remain fixed in the calibration comparisons. CIRR uses separately trained instances of the architecture developed on FashionIQ.

We use the public SigLIP2 SO400M checkpoint with patch size 14 and input resolution 384. The complete image and text towers contain approximately 1.136 billion frozen parameters and produce 1,152-dimensional features. The composition head adapts reference and text features and includes a rank-64 multiplicative interaction. It is trained for 30 epochs at temperature 0.035 with weight decay

Table 5: **Later parent heads, before the five-coefficient stage.** Counts describe these heads only; backbone adapters and earlier heads are additional. All listed training uses the corresponding dataset's training split.

Head	Trainable parameters	Epochs	Learning rate
Query correction	786,432	6	2×10^{-4}
Joint query/gallery correction	1,245,184	10	2×10^{-4}
Wide score gate	161	5	10^{-3}
Candidate-density correction	3	8	2×10^{-4}
Spectral adjustment	16	8	10^{-2}
Candidate feature matcher	898	8	10^{-3}
SigLIP2 composition head	884,801	30	10^{-4}

0.01, and its score enters A with coefficient 0.2. The encoder and composition head are shared by the parent and its matched calibration controls.

B CALIBRATION, MODEL SELECTION, AND INFERENCE

Calibration initializes $z = 0$, giving zero effective coefficients in Equation 5. AdamW updates z with optimizer weight decay, while the explicit loss penalizes θ. Each batch contains 128 queries. FashionIQ batches are formed within categories and then shuffled across categories, updating one shared coefficient vector. Scores use the full training gallery of the corresponding category. CIRR uses its full training gallery of 16,939 images.

Algorithm 1 Training and using the five-coefficient calibrator

Require: Fixed parent scores and evidence features; training targets; λ, γ
Set $z \leftarrow 0$; construct training groups if $\lambda > 0$
for 12 epochs **do**
  for training minibatch $\mathcal{B}$ **do**
    Set $\theta \leftarrow \ell \odot \tanh z$
    Compute $s = s^0 + \theta^\top \phi$ over the complete training gallery
    Exclude reference images and evaluate Equation 8
    Update z with AdamW and gradient clipping
  end for
end for
Freeze z
At inference, compute the same scores without targets or training groups
return Complete-gallery rankings; restrict the same scores for CIRR subset output

FashionIQ model selection. FashionIQ model selection uses two stages. The first evaluates seven settings at $\gamma = 0.2$: no group loss, or reference-neighbor and retrieval-score groups at $\lambda \in \{1, 4, 16\}$. The seed-42 results motivate six additional settings: both group constructions at $\lambda = 16$ with $\gamma \in \{20, 200\}$, and the corresponding no-group controls. In the first stage, the directly transferred reference-group model, no-group control, and selected grouped model are evaluated with three seeds. The second stage similarly evaluates its selected grouped models and controls with three seeds. The procedure comprises 25 calibration fits. Table 6 reports every seed-42 setting. The final FashionIQ configuration is the no-group model with $\gamma = 200$, selected after these validation comparisons.

CIRR model selection. CIRR validation is used to select the initial offset $0.2(T - V) + 0.2E$. Candidate extensions include five-coefficient calibration, query-vector corrections, and candidate-caption features. A subsequent coefficient and group-weight comparison selects $\lambda = 16, \gamma = 0.2$. The reported model uses the five-coefficient calibrator without a candidate-caption extension or an additional query network at this stage. The three-seed score ensemble is fixed before hidden-test predictions are generated, and Table 4 reports the resulting official test metrics.

Table 6: **Complete FashionIQ calibration search, seed 42.** Names encode group construction, group-loss weight g, and coefficient penalty r. Initial settings use $\gamma = 0.2$. Parent and fixed-offset rows require no new training. These are validation-selection results, not independent test estimates.

Configuration	R@10	R@50	Mean
Initial: `parent`	58.87	78.10	68.48
Initial: `fixed_offset`	58.90	78.51	68.71
Initial: `g0`	58.64	78.03	68.34
Initial: `reference_g1`	57.41	77.15	67.28
Initial: `reference_g4`	55.67	75.28	65.48
Initial: `reference_g16`	51.51	72.49	62.00
Initial: `retrieval_g1`	57.41	77.27	67.34
Initial: `retrieval_g4`	56.09	75.99	66.04
Initial: `retrieval_g16`	54.39	74.63	64.51
Regularized: `reference_g16_r20`	56.51	76.57	66.54
Regularized: `reference_g16_r200`	58.55	78.43	68.49
Regularized: `retrieval_g16_r20`	57.44	77.45	67.45
Regularized: `retrieval_g16_r200`	58.82	78.47	68.64
Regularized: `reference_g0_r20`	58.92	78.55	68.73
Regularized: `reference_g0_r200`	58.90	78.50	68.70

Reproducibility checks. Reloading the reported FashionIQ calibration checkpoints reproduces their complete validation rankings. Parent scores and their ensemble also reproduce the pre-calibration results. Rankings are independently checked using stable sorting with ties resolved by gallery index. The scoring function is evaluated from cached query and candidate features without target labels or group membership. All 4,148 CIRR submission lists exclude the reference and contain the required number of unique candidates. Subset lists restrict the same score ordering to the supplied group.

C Additional numerical details

C.1 Effective coefficients

Table 7: **Learned coefficients of the reported models.** Columns correspond to $(B - A, T - V, E, R, C)$. FashionIQ uses $\lambda = 0, \gamma = 200$; CIRR uses $\lambda = 16, \gamma = 0.2$.

Dataset	Seed	θ_1	θ_2	θ_3	θ_4	θ_5
FashionIQ	42	-0.00053	-0.00077	-0.00018	0.00020	0.00029
FashionIQ	43	-0.00050	-0.00065	-0.00042	-0.00026	0.00009
FashionIQ	44	-0.00106	-0.00058	-0.00009	0.00061	-0.00011
CIRR	42	0.06087	0.10095	0.11345	-0.26396	0.06090
CIRR	43	0.06630	0.08606	0.13676	-0.26280	0.04575
CIRR	44	0.04716	0.09408	0.13444	-0.26285	0.04241

CIRR assigns a negative reference coefficient in all three runs, whereas FashionIQ's coefficients remain near zero. The selected FashionIQ objective combines stronger regularization with no group term. The coefficients therefore reflect both choices; this table alone cannot separate their individual effects.

C.2 Metric aggregation and uncertainty

FashionIQ Mean averages the six category recalls with equal weight. The selected model's per-seed Means are 68.6996, 68.8325, and 68.6351, giving a mean of 68.7224 and a sample standard deviation of approximately 0.101. The score ensemble obtains 68.8186 after averaging candidate scores and recomputing rankings. This ensemble metric is distinct from the mean of the per-seed metrics.

The CIRR server values R@5= 84.530 and R_s@1= 80.337 give Mean 82.4335; the additional decimal in the mean results from averaging the rounded server values. Full-gallery R@2 is 69.614, and the other test recalls appear in Table 4. For DQU-CIR, the displayed components 78.17 and 70.92 yield 74.545, rounded to 74.55. TME's components yield 82.015, rounded to 82.02. The same aggregation rule is applied without changing the reported component measurements.

FashionIQ difference intervals use 5,000 bootstrap resamples. Queries sharing a reference image are resampled as a group within each category, and category metrics retain equal weight in every resample. The 2.5th and 97.5th percentiles define the conditional 95% interval. This procedure accounts for queries sharing references but conditions on the fitted backbones and validation-based model choices. It does not adjust for the full history of configuration selection.

C.3 Uniform offsets and changes of score scale

Let c_q be a query-specific vector independent of candidate j. For fixed coefficients, replacing ϕ_{qj} by $\phi_{qj} + c_q$ adds the constant $\theta^\top c_q$ to every candidate score, leaving rankings and log-softmax losses unchanged. Gallery centering in Equation 2 likewise contributes only candidate-independent offsets. Feature rescaling, however, changes the effective regularization unless the coefficient bounds and penalty are transformed consistently. The stated coefficients and limits therefore apply to the fixed feature representations used in our experiments.

C.4 Example provenance and interpretation

The illustrated query uses FashionIQ reference `B0068VM5T4` and target `B003Y74AOI`, with captions "is long sleeved" and "has longer sleeves." The parent score ensemble ranks the target 11th, and the reported calibrator ranks it 8th. Both use the complete shirt gallery with the reference excluded. Selection was restricted to R@10 recoveries and based on visual clarity. Table 3 complements this example with recovery and regression counts over all validation queries.