IMPROVING RETRIEVAL EMBEDDINGS ON CONJUNCTIVE QUERIES WITH PARTIAL CREDIT

Anonymous authors
Paper under double-blind review

ABSTRACT

As users refine search queries, especially in e-commerce, they tend to add requirements as conjunctions, making each query more specific than the last. One limitation in existing training paradigms is the binary training objective; hard negatives that subtly conflict with the query are considered just as incorrect as easy, obviously irrelevant negatives. To leverage the wasted information, we present the Graded Objective for Learning Distinctions, or GOLD mining, a hard negative mining strategy that generates hard negatives with known violations of query conditions. Additionally, we contribute GOLD reweighting, a loss function modification strategy that assigns partial credit to hard negatives that nearly satisfy the query. We demonstrate that both GOLD mining and GOLD reweighting are compatible with four common existing loss function families. Applying both generally improves win rate and achieves state-of-the-art recall. Our theoretical analysis establishes that under reasonable assumptions, GOLD Margin-MSE gains the guarantee that uniformly low loss on the objective implies approximate affine recoverability of attributes.

1 INTRODUCTION

Embedding-based retrieval aims to recall documents that are semantically relevant to the user's query. Such embeddings are often used in retrieval tasks, such as retrieval augmented generation (RAG), product search, and image search, where fine-grained matching of details between the query and the document determines relevance. As users refine their queries, they tend to increase in specificity over time by filtering out more candidate results. Such user behavior is demonstrated in employee recruitment (Kaya & Bogers, 2023), web search (Jansen et al., 2007), and product search (Papenmeier et al., 2020). Hirsch et al. (2020) find that over 50% of queries on eBay occur during a reformulation session, with queries growing in length over time. These refined queries are often conjunctive: a result must satisfy every condition, including negated conditions (Merra et al., 2023).

Bi-encoders are known to struggle with exactly the operations that such queries require. These include negation (Zhang et al., 2025; Weller et al., 2024), set operations (Malaviya et al., 2023), simultaneous constraint satisfaction (Lu et al., 2025; Killingback & Zamani, 2025), and attribute binding (Yuksekgonul et al., 2022). We hypothesize that user query behavior may suffer from a chicken-and-egg problem; users do not use complex queries because they expect models to fail, while practitioners underprioritize complex queries because they do not appear in user data.

Although datasets such as Amazon ESCI (Reddy et al., 2022) and WANDS (Chen et al., 2022) include simple multi-attribute queries, they offer limited coverage of complex conjunctions. Thus, we concentrate on improving conjunctive queries — those seeking documents that *have* attributes a, b, c ... and *do not have* attributes x, y, z, ... This focus is further grounded in patterns in user queries and the design of pre-existing non-embedding search systems. For example, the conjunctive filtering paradigm already exists in product search in the form of attribute filters on e-commerce sites, such as "screen size $\geq$ 38 cm" for laptops. Unfortunately, attribute ontologies or SQL-style schemas are usually fixed, failing when a user desires a product with or without a novel attribute.

To address this gap, we introduce **G**raded **O**bjective for **L**earning **D**istinctions, or GOLD mining, a strategy for constructing conjunctive queries for existing document pairs, automatically labeling how many query conditions each hard negative violates. We combine it with GOLD reweighting, a

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.

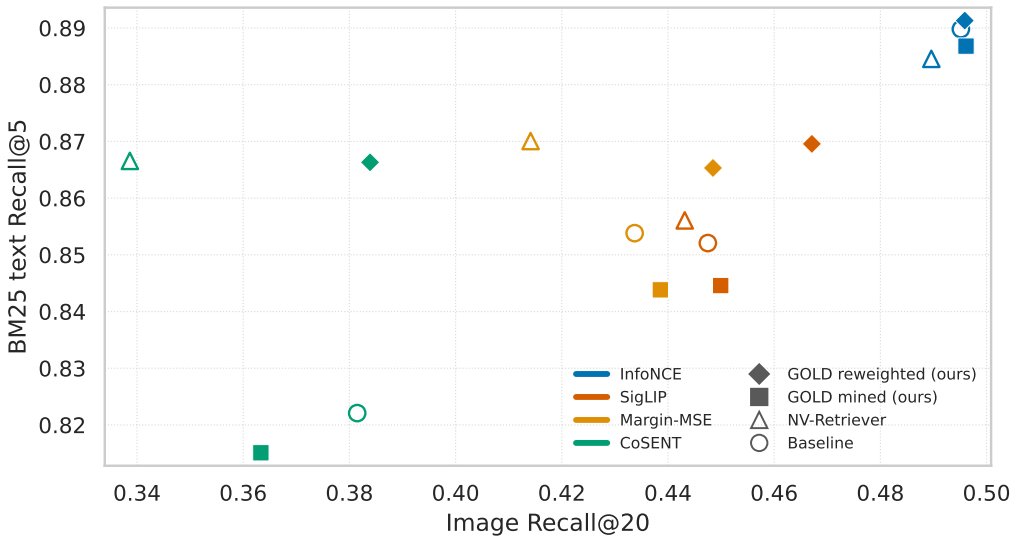


Figure 1: Text versus image retrieval recall. Unfilled and filled shapes indicate baselines and ours, respectively. Color indicates loss function. Our GOLD reweighting strategy achieves best overall performance applied to each loss function in most cases.

broadly applicable loss function modification that assigns partial credit to hard negatives that partially satisfy those conditions. GOLD mining uses a generative LLM to extract attributes of each document in structured form. Then, we construct triplets consisting of a natural language query, positive document, and hard negative document based on the structured attributes (Figure 2). Leveraging the structured attribute lists allows us to calculate the number of query conditions violated by the hard negative, which we use in the GOLD reweighting training objective. This contrasts with baseline loss functions that treat all negatives as equally wrong.

We demonstrate the effectiveness of these strategies on both the text and image modalities. We choose textual product search due to its economic importance, with annual e-commerce sales exceeding $28 trillion USD in 2024 (UN Trade and Development, 2026). We choose apparel image search because it requires objective reasoning over fine-grained visual details, a foundational skill in numerous multimodal tasks. Although we demonstrate the effectiveness of GOLD mining and reweighting on these two domains, we stress that, in principle, the strategies are highly general. GOLD mining can be applied in settings where documents can be reasonably well specified by binary QA (i.e., a list of attributes), such as security video or resume search. GOLD reweighting need not rely on weights derived from GOLD mining specifically, and could instead use, for example, pre-existing structured data. Our results demonstrate state-of-the-art image and text recall using GOLD mining + reweighting.

Our theoretical analysis shows that, under suitable query and comparison coverage assumptions, uniformly low loss on the GOLD Margin-MSE objective excluding easy negatives implies approximate affine recoverability of attributes, a property not guaranteed even by zero binary-target Margin-MSE loss. The converse is also true: affine recoverability of attributes permits low loss on the GOLD Margin-MSE objective under a reasonable scale assumption. Finally, we show that an embedding of fixed length d can represent exponentially many attributes with low loss.

2 Related Work

Dense retrieval and graded supervision Dense bi-encoders enable efficient first-stage retrieval by mapping queries and candidates into a shared embedding space (Reimers & Gurevych, 2019; Karpukhin et al., 2020). Methods such as ANCE and RocketQA improve training by retrieving challenging negatives, sharing negatives across batches, or denoising mined candidates (Xiong et al.,

	Text (Synthetic)	Text (Human)	Image (Synthetic)	Image (Human)
Corpus triples	40,188	133	25,914	83
Training triples	34,890	—	20,592	—
Validation triples	2,630	—	2,566	—
Test triples	2,668	133	2,756	83
Words per document	189.8	177.5	—	—
Words per query	28.8	21.5	27.6	17.5
Attributes stated per query	9.7	7.6	9.8	7.3
Differentiating attributes per query	5.0	—	5.1	—

Table 1: Dataset summary statistics. Main counts are (query, positive, negative) triples; Human counts are (query, positive) pairs.

2020; Qu et al., 2021). Other approaches move beyond uniform positive–negative separation by distilling continuous teacher-score differences, as in Margin-MSE and TAS-Balanced (Hofstätter et al., 2020; 2021), or by using relevance- and attribute-dependent margins in metric-learning objectives (Zhou et al., 2020; Jeong et al., 2021; Zhao et al., 2019). Our work builds on these by using hard negatives with known attribute violations to provide graded supervision for standard dense-retrieval losses.

Compositional, Boolean, and product retrieval Prior product search methods address the problem of compositional requirements through behaviorally informed negative sampling or explicitly representing facets and attribute composition (Mohan et al., 2019; Kong et al., 2022; Choudhary et al., 2022). A related line of work studies logical composition in dense retrieval. QUEST considers natural-language queries representing set operations, while BoolQuestions examines conjunction, disjunction, and exclusion through Boolean-aware training and query decomposition (Malaviya et al., 2023; Zhang et al., 2024). Logical query-embedding methods such as BetaE and ConE instead construct geometric representations that directly support logical operators (Ren & Leskovec, 2020; Zhang et al., 2021). NevIR shows that bi-encoders are often insensitive to negated meaning, while product-search systems have mitigated related failures through explicit detection and filtering of prohibited attributes (Weller et al., 2024). In the multimodal domain, the CIRR benchmark and methods such as CLIP4CIR study retrieval from queries that combine a reference image with a textual modification (Radford et al., 2021; Liu et al., 2021; Baldrati et al., 2022). CoN-CLIP and NegationCLIP improve sensitivity to negated concepts, while CoLLM uses LLM-generated triplets for composed image retrieval (Singh et al., 2025; Park et al., 2025; Huynh et al., 2025). We study the same challenges of compositionality and negation, but address them through attribute-aware supervision for standard multimodal bi-encoders rather than specialized logical representations or query operators.

Synthetic supervision for retrieval Large language models have increasingly been used to generate targeted supervision for dense retrieval. Syntriever and SyNeg generate queries, relevant examples, or challenging negatives that expose distinctions absent from conventional training data (Kim & Baek, 2025; Li et al., 2024). Promptagator (Dai et al., 2022) uses LLMs for few-shot dense retrieval. Wang et al. (2024); Lee et al. (2024) distill embeddings from LLMs. In product retrieval, Agapaki and Gill combine structured product information with LLM-based relevance assessment to identify difficult training examples (Agapaki & Gill, 2026). Our work follows this line of research by using an LLM to construct synthetic conjunctive queries and hard negatives with known violations, enabling targeted supervision for both text and image retrieval.

3 Dataset Construction

We seek to evaluate models on a test dataset containing query-document pairs where the document satisfies all constraints in the query. We target two domains, e-commerce products and images, to demonstrate the multimodal capabilities of our system. For e-commerce, we choose the ESCI corpus of Amazon products (Reddy et al., 2022) due to its large size. For images, we choose the DeepFashion In-Shop dataset because it contains images with subtle but objectively differing attributes. Table 1 summarizes both datasets.

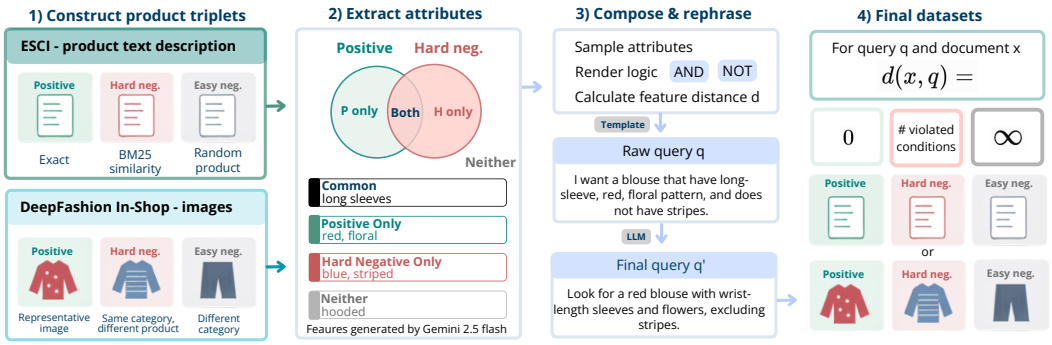


Figure 2: Beginning with documents 1) from the ESCI Amazon products on In-Shop apparel images datasets, we choose a positive product and a similar hard negative product based on dataset metadata (images) or BM25 similarity (text). 2) We extract attributes that apply to one, both, or neither product. 3) We randomly select attributes, using them to construct a query that fully applies to the positive example by negating attributes that do not apply to it. We also calculate the number of said attributes d that do not apply to the hard negative, while easy negatives are assumed to have maximal distance (GOLD mining) or use another similar document as a negative (baselines). We use an LLM to rephrase the query to reduce key phrase overlap between the product and the query, producing 4) the final dataset.

Query Generation We use the dataset-provided positive labeled documents as positives. To generate a contrasting text document, we sample a document uniformly at random from the top five BM25 (Robertson & Zaragoza, 2009) matches for the original ESCI query. For the image dataset, we select a random example from the same category as the positive, such as "blouses" or "graphic tees", excluding all views and colorways of the positive garment. For images, the prompt contains both garments' photographs and metadata, and instructs the model to generate only attributes visible in the photographs.

To generate conjunctive queries, we prompt Gemini 2.5 Flash with a matched positive and negative document pair, using the prompts of Appendix E, one variant for products and one for images. We instruct it to generate up to 5 attributes that apply to only the first example, only the second example, to both examples, or to neither example, a pool of up to 20 attributes. We then select a random subset of these attributes to compose the query, negating attributes that do not apply to the positive example, ensuring that at least one attribute is not satisfied by the hard negative example. Oversampling here is useful to ensure that selected attributes are not merely the most salient. By conditioning attribute generation on the negative document, similar to Mayfield et al. (2023); Zeng et al. (2026); Vedantam et al. (2017) we ensure that forbidden attributes could plausibly apply to a document similar to the positive.

We construct a raw query with the template "I am looking for: {item} that has: {positive} + {common} attributes; and does not have: {negative} + {neither} attributes". We then prompt Gemini 2.5 Flash to rewrite the raw query based on five human-written in-context examples, replacing attribute phrases with synonyms (Appendix E.3). The prompt includes the positive document so synonyms stay consistent with it, requires at least half of the attributes to be reworded, and forbids changing numeric values. Rewriting the query in this way is designed to improve the realism and diversity of queries and reduce surface form overlap between queries and positive documents. This yields a (query, positive) pair.

GOLD mining We note that conditioning query generation on random documents as described above makes these documents, by construction, hard negatives with respect to the query. Specifically, each document will violate exactly distance d conditions in the query, where $1 \leq d \leq \min(k, 10)$, where k is the number of conditions in the query. We term this strategy GOLD mining. Thus, GOLD mining produces (query, positive, negative, distance) tuples, where d is consumed by the GOLD mining objective described in Section 4.2. Although GOLD mining here occurs in parallel with query generation, in principle this strategy is dataset-agnostic. The strategy could be performed in reverse by pairing existing conjunctive queries with negative documents and calcu-

lating d. In contrast, we create a baseline non-mined dataset that selects hard negative examples identically, except that the queries are *not* conditioned on baseline hard negatives.

Human labeled queries Annotators manually annotated 133 and 83 text and image queries, respectively, based on positive documents. We use these query-document pairs to validate that models trained on Main data generalize to human-generated queries.

4 METHODS

Numerous training objectives have been proposed for dense retrieval. We choose Margin-MSE (Hofstätter et al., 2020), CoSENT (Huang et al., 2024), SigLIP (Tschannen et al., 2025), and InfoNCE (van den Oord et al., 2019) as our target loss functions based on their applicability to triple-supervised training and coverage of high-level loss strategy families. Margin-MSE regresses query-document pairs onto a fixed margin; CoSENT ranks pairs ordinally; SigLIP classifies in-batch pairs independently; InfoNCE scores each in-batch candidate against the others using a softmax function. We include their definitions in Appendix C for reference. Because batch size affects CoSENT, SigLIP, and InfoNCE training dynamics through pairwise in-batch comparisons, we hold batch size constant across all runs.

4.1 BASELINE DETAILS

For InfoNCE, SigLIP, and batched Margin-MSE, each of the B queries is scored against all $2B$ documents in the batch, yielding $2B^2$ query–document scores. Documents from other triples are treated as easy negatives; each query's own paired negative may be hard or easy, chosen uniformly at random. InfoNCE uses one softmax over all candidates. Margin-MSE and SigLIP weight the mean own-row and cross-row losses equally: the own-row group contains the paired negative for Margin-MSE, and both the positive and paired negative for SigLIP. CoSENT compares scores of supplied query–document pairs, while pairwise MSE regresses each pair's score independently. We design our experiments such that each model sees each training triple once per epoch. We train every model for one epoch with 32 rows per forward pass and 4 forward passes per optimizer step. Pairwise MSE emits two pairs per triple. To study the NV-Retriever baseline, we perform a search across NV-Retriever mining parameters described in Appendix D.

4.2 GOLD REWEIGHTING

Traditional Margin-MSE loss (Hofstätter et al., 2020) is defined as

$$\mathcal{L}_{\text{MSE}}(q, x^+, x^-) = \Big(\langle f(q), f(x^+) \rangle - \langle f(q), f(x^-) \rangle - y \Big)^2. \tag{1}$$

Here, q is a query, x^+ is a relevant document, x^- is a less relevant document, and y is a scalar representing the margin between their relevance as determined by a teacher model. f is a function mapping queries or documents to their embeddings. For loss functions we choose cosine similarity as our inner product.

Formally, following Lu et al. (2025), let query q_k be a query composed of k conditions $C = \{c_1, c_2, ..., c_k\}$ with $k \in \{1, ..., K\}$. We say a document x is a positive document, x^+, with respect to q_k if $x^+ \models C$, i.e., x^+ satisfies all k conditions. Otherwise, we say the document x^- is negative with respect to q_k. x_h^- and x_e^- denote hard and easy negatives, respectively.

We introduce GOLD reweighting, a strategy for assigning partial credit to hard negative examples. The idea behind GOLD reweighting is to modify a loss function so that the penalty for a hard negative increases with the number of query constraints it violates. Because CoSENT relies on ordering for loss, GOLD reweighting on CoSENT instead treats all hard negatives as rank 1/2. Formally, we define attribute Hamming distance d as

$$d(x, q) = \sum_{i=1}^{k} \mathbb{1}[x \not\models c_i] \tag{2}$$

Losses written for one query or triple are averaged over the batch. We then modify each loss function, with the modification marked by an underbrace:

Margin-MSE: We set the margin term to d/V and apply the loss across all query–document pairs in the batch:

$$\mathcal{L}_{\text{MSE}}(q_i, x^+, x^-) = \Big(\big\langle f(q_i), f(x^+) \big\rangle - \big\langle f(q_i), f(x^-) \big\rangle - \underbrace{\tfrac{d}{V}}_{\text{ours}} \Big)^2 \tag{3}$$

The violation scale hyperparameter V controls the penalty size per violation. Let K^* be the maximum number of constraints violated by any example. In Margin-MSE only, we set $d(x, q) = \eta \geq K^*$ when x is an easy negative. Choosing $V \geq \max(\eta, K^*)$ ensures the d/V term is in $[0, 1]$.

InfoNCE: A batch of B triples contains B positives and B negatives. Each query is scored against all $2B$ of these documents, which form the candidate set X. InfoNCE trains the softmax over X toward a one-hot target on the query's own positive. We keep this target and interpolate between excluding the own hard negative from the denominator and including it, with inclusion weight d/V for its measured $d = d(x_i^-, q_i)$:

$$\mathcal{L}_{\text{INCE}}(q_i) = -\underbrace{(1 - d/V)}_{\text{ours}} \ln \frac{e^{s\langle f(q_i), f(x_i^+)\rangle}}{\sum_{x \in X \setminus \{x_i^-\}} e^{s\langle f(q_i), f(x)\rangle}} - \underbrace{d/V}_{\text{ours}} \ln \frac{e^{s\langle f(q_i), f(x_i^+)\rangle}}{\sum_{x \in X} e^{s\langle f(q_i), f(x)\rangle}}. \tag{4}$$

We fix $s = 20$ per Reimers & Gurevych (2019) default and choose $V \geq K^*$ so that $d/V \in [0, 1]$.

SigLIP: In place of the original binary target, we weight each pair with the graded target $T_i(x)$:

$$T_i(x) = \begin{cases} 1 & x = x_i^+ \\ e^{-s\, d(x_{h,i}^-, q_i)/V} & x = x_{h,i}^- \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

$$\mathcal{L}_{\text{SigLIP}}(q_i) = -\sum_{x \in X} \left[\underbrace{T_i(x)}_{\text{ours}} \ln \frac{1}{1 + e^{-(s\langle f(q_i), f(x)\rangle + b)}} + \underbrace{\big(1 - T_i(x)\big)}_{\text{ours}} \ln \frac{1}{1 + e^{s\langle f(q_i), f(x)\rangle + b}} \right], \tag{6}$$

where s and b are learned parameters.

CoSENT: We assign rank 1 to positives, 1/2 to hard negatives, and 0 to easy negatives. The algorithm otherwise remains unchanged.

The unifying factor among these modifications is to assign partial credit to hard negatives, compared to easy negatives. For Margin-MSE, this regresses towards the condition

$$\langle f(q), f(x_1)\rangle - \langle f(q), f(x_2)\rangle = (d(q, x_2) - d(q, x_1))/V \tag{7}$$

4.3 Relation to Affine Probing

This section develops the intuition behind the connection between our objective and affine probing; formal statements and proofs are provided in Appendix A.

Following Garg et al. (2026), we will assume that inputs come from some set $L = Q \cup D$ consisting of queries or documents composed of text or images. Let $f : L \to \mathbb{R}^d$ give the embeddings of dimension d. We let $z_i : L \to \{0, 1\}$ be an attribute of a query or corresponding product, for

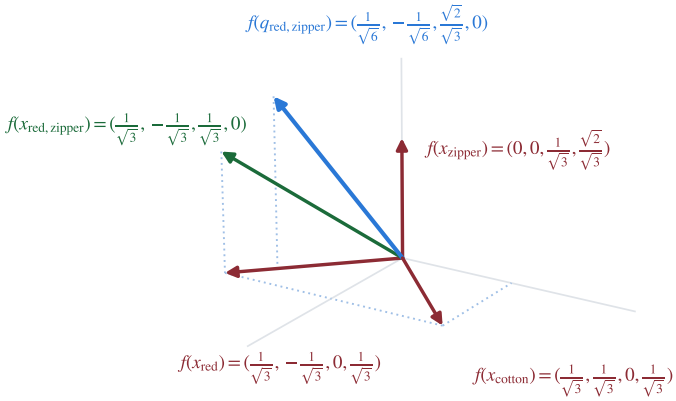


Figure 3: An idealized embedding space in $\mathbb{R}^4$, drawn by dropping the fourth coordinate. Three documents, each with exactly one attribute, have mutually orthogonal visible parts and put the rest of their unit length into a fourth component the query cannot see. With $V = 3/\sqrt{2}$ as the violation scale, the probe sum has norm $\|b_{\text{red}} + b_{\text{zipper}}\| = V$, and $V \langle f(q), f(x) \rangle$ counts the conditions x satisfies: x_{red} and x_{zipper} score $1/V = \sqrt{2}/3$, x_{cotton} scores 0, and the positive $x_{\text{red,zipper}}$, which the probes force to be the visible sum of x_{red} and x_{zipper}, scores $2/V$.

example, whether a query requires a red product, or whether the product in question is red. Note that the value of z_i depends on the underlying input, not its embedding.

For a document set $S \subseteq D$ and tolerance $\epsilon > 0$, we say a feature is (ϵ, S)-affinely recovered by a probe $(b_i, \beta_i) \in \mathbb{R}^d \times \mathbb{R}$ if

$$| \langle b_i, f(\ell) \rangle + \beta_i - z_i(\ell)| < \epsilon \tag{8}$$

for all $\ell \in S$.

We show in Theorem 1, affine recoverability of attributes permits low GOLD reweighting Margin-MSE loss. Conversely, uniformly low loss across all positive–negative comparisons for two queries differing by one condition, each with nonempty positive and negative sets, implies approximate affine recovery of that condition's attribute (Theorem 2). The intuition behind these proofs is that under low loss conditions, the query vector must act as a linear probe for each attribute simultaneously, effectively "counting" the number of satisfied conditions. We illustrate an idealized example of the embedding space in Figure 3. Binary-target Margin-MSE provides no such guarantee: zero loss can coexist with an unrecoverable attribute (Proposition 1). Finally, for documents with at most k' active attributes from a vocabulary of size m, queries with at most k conditions, and $V \geq kk'$, we construct unit document and query embeddings in dimension $d = O\left(\frac{k'^2}{\epsilon^2} \ln m\right)$ that affinely recover every attribute with error below ϵ and achieve loss below $(2k\epsilon/V)^2$ on every valid comparison (Corollary 3).

4.4 IMPLEMENTATION AND EVALUATION DETAILS

We fine-tune `all-mpnet-base-v2` (Song et al., 2020) on text and `clip-ViT-B-32` (Radford et al., 2021) on images, both as distributed by Sentence-Transformers (Reimers & Gurevych, 2019), normalizing every representation to unit length. We optimize with AdamW (Loshchilov & Hutter, 2017) using a learning rate 2×10^{-5}, a cosine schedule warming up over the first 10% of steps, and bf16, training each condition three times and reporting the mean over those trials. For text, we exclude all target products in the combined human annotation set from training and validation, including their use as negative documents.

For GOLD models, we sweep $V \in \{10, 20, 40, 80\}$ for InfoNCE and SigLIP, and $V \in \{20, 40, 80\}$ with $\eta \in \{10, 20, 40\}$ training regimes using η, selecting settings on the validation set. Baseline, NV-Retriever, and GOLD mining models do not use V. Note that V does not carry precisely the

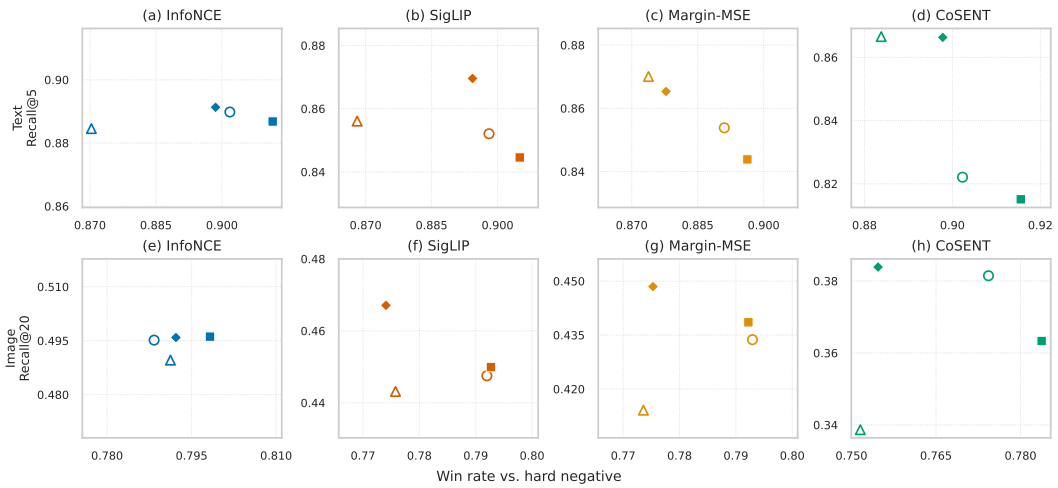


Figure 4: (a–d) Text and (e–h) image recall (vertical axis) versus positive/hard-negative win rate (horizontal axis). Axis scales held constant across plots. Across loss functions, GOLD mining consistently improves win rate while GOLD reweighting improves recall at the expense of win rate.

same meaning across GOLD loss functions. For graded InfoNCE, validation selects $V = 20$ in both modalities. The CoSENT modification introduces no additional hyperparameters.

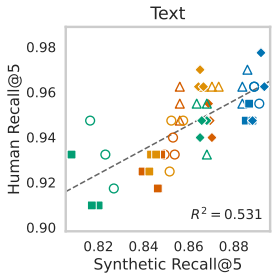
We use Recall@5 for text and Recall@20 for image retrieval for both validation selection and reporting. For text retrieval, we report Recall@5 because this more stringent cutoff provides greater discrimination among high-performing strategies.

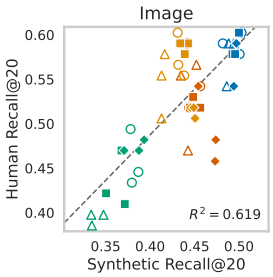
5 RESULTS AND DISCUSSION

We present results in Figure 1. On the larger Main test set, applying both GOLD mining and GOLD reweighting improves mean recall over Baseline for every loss and modality. Overall recall is stratified by loss function, with InfoNCE ranking highest on both modalities, both for our proposed methods and baselines. GOLD mining alone slightly improves recall on images, but reduces recall on text. GOLD mining + reweighting, however, generally performs better than NV-Retriever across modalities, improving recall in the two strongest loss functions on text and all four on images. Overall, the strongest strategy for text recall is InfoNCE with GOLD mining + reweighting, though the improvement is not statistically significant in all cases (Table 3). The strongest strategy for image recall is InfoNCE with GOLD mining only, though adding GOLD reweighting achieves almost identical recall.

We observe that GOLD mining increases hard-negative win rate in seven of eight loss function–modality pairs compared to Baseline (Figure 4). Applying GOLD reweighting improves Main recall in seven of eight loss function–modality pairs at the expense of hard-negative win rate. That is, GOLD mining improves accuracy more on hard negatives while GOLD reweighting trades accuracy on hard negatives for accuracy on easy negatives. This result follows directly from GOLD reweighting penalizing errors in hard negatives less than those in easy negatives in the loss function. Thus, the application of GOLD reweighting, and in particular the V scale, should be considered in the context of the dataset. Datasets containing more hard negatives per query may benefit more from GOLD mining, whereas requiring maximum recall will benefit more from GOLD reweighting.

We see in Figure 6 that InfoNCE accuracy for GOLD mining + reweighting is positively correlated with the number of positive (required) attributes in both modalities. However, accuracy appears to be largely uncorrelated with the number of negative (forbidden) attributes. The correlation with required attributes in conjunctive queries is unsurprising given that keyword or synonym overlap between queries and positive documents is tightly correlated with the number of conditional queries. The lack of correlation with forbidden attributes suggests that overall, embedding models neither induce false positives out of documents containing those attributes nor contribute to magnifying





Measure	Text	Image
Pearson r	0.73	0.79
Spearman ρ	0.72	0.65

Figure 5: Main (synthetic) versus human recall for text (Recall@5, left) and images (Recall@20, center), with correlation statistics. Each plot includes all 48 model seeds; colors and markers follow Figure 1, and dashed lines show linear fits.

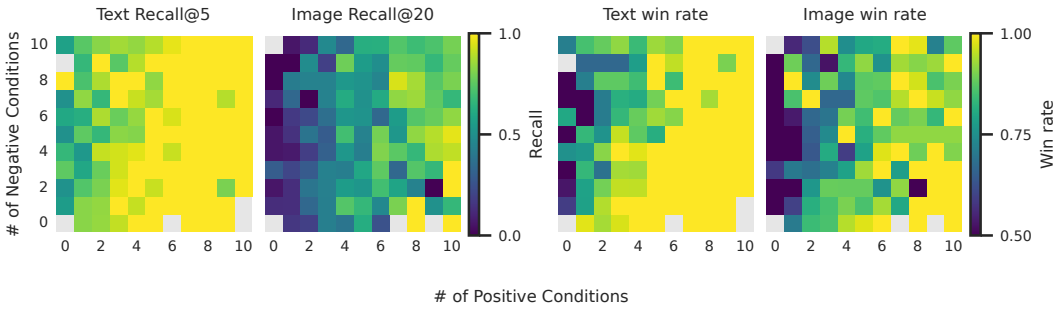


Figure 6: InfoNCE GOLD mining and reweighting recall and win rate by required and forbidden query condition count. Models achieve higher recall and win rate as the number of required conditions increases. The number of forbidden conditions is negatively correlated with win rate but uncorrelated with recall. Other loss functions and strategies show identical trends.

positive documents in the way that required attributes do. Win rate vs. hard negative, however, increases with the number of required conditions but decreases with the number of forbidden conditions. Other loss functions show similar behavior.

To validate the use of synthetic queries, we evaluate the same models, all trained on the Main train split, on both the synthetic (Main) test set and human-written queries (Figure 5). Both text and image models achieve Pearson's r in the 0.7-0.9 range, considered strong linear correlation in the recall metric (Schober et al., 2018). Models achieve Spearman's ρ, measuring ordering correlation, in the same range on text and in the moderate range on images. Thus, we conclude that the Main test set with its greater sample size, is a valid proxy for recall on human-generated queries.

Surprisingly, despite the distribution shift from LLM-generated queries, average model recall on human-generated queries exceeds that on synthetic queries. This cannot be explained simply by the lower average number of conditions in human queries because condition count is in fact positively correlated with recall (Figure 6). Human queries also reuse more wording from their positive text documents than Main test queries (unigram precision 58.6% vs. 38.8%; bigram precision 22.4% vs. 11.2%, respectively). Furthermore, human queries are shorter on average than synthetic queries.

6 CONCLUSION

We present GOLD mining, a multimodal hard negative mining algorithm that improves win rate on conjunctive queries. We also present GOLD reweighting, a loss function modification strategy based on assigning partial credit to negatives. When combined with GOLD mining, reweighting beats the strongest baseline in most cases and vanilla training in all cases.

AI USE STATEMENT

In this work, we used generative AI tools to: generate synthetic datasets, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in writing of proofs, propose or refine hypotheses, design and provide feedback on research methodology and experiments, implement methods, support qualitative and thematic data analysis, create figures, draft parts of this paper, summarize existing literature, discover research topics, search for information, edit this paper, identify relevant literature, and to yell at when frustrated.

We have not used generative AI tools for cleaning and reformatting the dataset except where described in this paper, and translation is not applicable to this work. We have reviewed all AI-assisted work. LLM-written code was proofread and tested with unit and sanity tests. LLM-written text was proofread by 2-4 authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Eva Agapaki and Amritpal Singh Gill. Negative data mining for contrastive learning in dense retrieval at IKEA.com. *arXiv preprint arXiv:2605.00353*, 2026.

Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Effective conditioned and composed image retrieval combining CLIP-based features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21466–21474, June 2022.

Yan Chen, Shujian Liu, Zheng Liu, Weiyi Sun, Linas Baltrunas, and Benjamin Schroeder. Wands: Dataset for product search relevance assessment. In *European Conference on Information Retrieval*, pp. 128–141. Springer, 2022.

Nurendra Choudhary, Nikhil Rao, Sumeet Katariya, Karthik Subbian, and Chandan Reddy. ANTHEM: Attentive hyperbolic entity model for product search. Amazon Science, 2022. URL https://www.amazon.science/publications/anthem-attentive-hyperbolic-entity-model-for-product-search.

Zhuyun Dai, Vincent Y. Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B. Hall, and Ming-Wei Chang. Promptagator: Few-shot dense retrieval from 8 examples, 2022. URL https://arxiv.org/abs/2209.11755.

Nikhil Garg, Jon Kleinberg, and Kenny Peng. How many features can a language model store under the linear representation hypothesis? *arXiv preprint arXiv:2602.11246*, 2026.

Sharon Hirsch, Ido Guy, Alexander Nus, Arnon Dagan, and Oren Kurland. Query reformulation in e-commerce search. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1319–1328, 2020.

Sebastian Hofstätter, Sophia Althammer, Michael Schröder, Mete Sertkan, and Allan Hanbury. Improving efficient neural ranking models with cross-architecture knowledge distillation. *arXiv preprint arXiv:2010.02666*, 2020. URL https://arxiv.org/abs/2010.02666.

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. Efficiently teaching an effective dense retriever with balanced topic aware sampling. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 113–122, 2021. doi: 10.1145/3404835.3462891. URL https://arxiv.org/abs/2104.06967.

Xiang Huang, Hao Peng, Dongcheng Zou, Zhiwei Liu, Jianxin Li, Kay Liu, Jia Wu, Jianlin Su, and Philip S Yu. CoSENT: Consistent sentence embedding via similarity ranking. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2800–2813, 2024.

Chuong Huynh, Jinyu Yang, Ashish Tawari, Mubarak Shah, Son Tran, Raffay Hamid, Trishul Chilimbi, and Abhinav Shrivastava. CoLLM: A large language model for composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3994–4004, June 2025.

Bernard J Jansen, Mimi Zhang, and Amanda Spink. Patterns and transitions of query reformulation during web searching. *International Journal of Web Information Systems*, 3(4):328–340, 2007.

Boseung Jeong, Jicheol Park, and Suha Kwak. ASMR: Learning attribute-based person search with adaptive semantic margin regularizer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12016–12025, October 2021.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.550. URL https://aclanthology.org/2020.emnlp-main.550/.

Mesut Kaya and Toine Bogers. Understanding recruiters' information seeking behavior in talent search. In *Proceedings of the 2023 conference on human information interaction and retrieval*, pp. 14–23, 2023.

Julian Killingback and Hamed Zamani. Benchmarking information retrieval models on complex retrieval tasks. *arXiv preprint arXiv:2509.07253*, 2025.

Minsang Kim and Seung Jun Baek. Syntriever: How to train your retriever with synthetic data from LLMs. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 2523–2539, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.136. URL https://aclanthology.org/2025.findings-naacl.136/.

Weize Kong, Swaraj Khadanga, Cheng Li, Shaleen Gupta, Mingyang Zhang, Wensong Xu, and Mike Bendersky. Multi-aspect dense retrieval. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022.

Jinhyuk Lee, Zhuyun Dai, Xiaoqi Ren, Blair Chen, Daniel Cer, Jeremy R. Cole, Kai Hui, Michael Boratko, Rajvi Kapadia, Wen Ding, Yi Luan, Sai Meher Karthik Duddu, Gustavo Hernandez Abrego, Weiqiang Shi, Nithi Gupta, Aditya Kusupati, Prateek Jain, Siddhartha Reddy Jonnalagadda, Ming-Wei Chang, and Iftekhar Naim. Gecko: Versatile text embeddings distilled from large language models, 2024. URL https://arxiv.org/abs/2403.20327.

Xiaopeng Li, Xiangyang Li, Hao Zhang, Zhaocheng Du, Pengyue Jia, Yichao Wang, Xiangyu Zhao, Huifeng Guo, and Ruiming Tang. SyNeg: LLM-driven synthetic hard-negatives for dense retrieval. *arXiv preprint arXiv:2412.17250*, 2024. URL https://arxiv.org/abs/2412.17250.

Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2125–2134, October 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Xuan Lu, Sifan Liu, Bochao Yin, Yongqi Li, Xinghao Chen, Hui Su, Yaohui Jin, Wenjun Zeng, and Xiaoyu Shen. MultiConIR: Towards multi-condition information retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 13471–13494, 2025.

Chaitanya Malaviya, Peter Shaw, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. QUEST: A retrieval dataset of entity-seeking queries with implicit set operations. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14032–14047, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.784. URL https://aclanthology.org/2023.acl-long.784/.

James Mayfield, Eugene Yang, Dawn Lawrie, Samuel Barham, Orion Weller, Marc Mason, Suraj Nair, and Scott Miller. Synthetic cross-language information retrieval training data, 2023. URL https://arxiv.org/abs/2305.00331.

Felice Antonio Merra, Omar Zaidan, and Fabricio de Sousa Nascimento. Improving the relevance of product search for queries with negations. In *Companion Proceedings of the ACM Web Conference 2023*, pp. 86–89, 2023.

Vijai Mohan, Yiwei Song, Priyanka Nigam, Choon Hui Teo, Weitian Ding, Vihan Lakshman, Ankit Shingavi, Hao Gu, and Bing Yin. Semantic product search, 2019. URL https://www.amazon.science/publications/semantic-product-search.

Andrea Papenmeier, Alfred Sliwa, Dagmar Kern, Daniel Hienert, Ahmet Aker, and Norbert Fuhr. 'a modern up-to-date laptop'-vagueness in natural language queries for product search. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*, pp. 2077–2089, 2020.

Junsung Park, Jungbeom Lee, Jongyoon Song, Sangwon Yu, Dahuin Jung, and Sungroh Yoon. Know "no" better: A data-driven approach for enhancing negation awareness in CLIP. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2825–2835, October 2025.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5835–5847, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.466. URL https://aclanthology.org/2021.naacl-main.466/.

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 Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Chandan K Reddy, Lluís Màrquez, Fran Valero, Nikhil Rao, Hugo Zaragoza, Sambaran Bandyopadhyay, Arnab Biswas, Anlu Xing, and Karthik Subbian. Shopping queries dataset: A large-scale ESCI benchmark for improving product search. *arXiv preprint arXiv:2206.06588*, 2022.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Hongyu Ren and Jure Leskovec. Beta embeddings for multi-hop logical reasoning in knowledge graphs. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 19716–19726. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/e43739bba7cdb577e9e3e4e42447f5a5-Paper.pdf.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and trends® in information retrieval*, 4(1-2):1–174, 2009.

Patrick Schober, Christa Boer, and Lothar A Schwarte. Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*, 126(5):1763–1768, 2018.

Jaisidh Singh, Ishaan Shrivastava, Mayank Vatsa, Richa Singh, and Aparna Bharati. Learning the power of "no": Foundation models with negations. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pp. 7991–8001, February 2025.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding, 2020. URL https://arxiv.org/abs/2004.09297.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

UN Trade and Development. Digital economy and technology. UNCTADstat Data Centre, 2026. URL https://unctadstat.unctad.org/insights/theme/265. Data last updated 21 May 2026.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. URL https://arxiv.org/abs/1807.03748.

Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. Context-aware captions from context-agnostic supervision, 2017. URL https://arxiv.org/abs/1701.02870.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improving text embeddings with large language models, 2024. URL https://arxiv.org/abs/2401.00368.

Orion Weller, Dawn Lawrie, and Benjamin Van Durme. NevIR: Negation in neural information retrieval. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2274–2287, 2024.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*, 2020. URL https://arxiv.org/abs/2007.00808.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022.

Qingcheng Zeng, Puxuan Yu, Aman Mehta, Fuheng Zhao, and Rajhans Samdani. Dual-view training for instruction-following information retrieval, 2026. URL https://arxiv.org/abs/2604.18845.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023.

Wenhao Zhang, Mengqi Zhang, Shiguang Wu, Jiahuan Pei, Zhaochun Ren, Maarten De Rijke, Zhumin Chen, and Pengjie Ren. Excluir: Exclusionary neural information retrieval. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 13295–13303, 2025.

Zhanqiu Zhang, Jie Wang, Jiajun Chen, Shuiwang Ji, and Feng Wu. ConE: Cone embeddings for multi-hop reasoning over knowledge graphs. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 19172–19183. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/a1607097011407045 75d499c997b6ca-Paper.pdf.

Zongmeng Zhang, Jinhua Zhu, Wengang Zhou, Xiang Qi, Peng Zhang, and Houqiang Li. BoolQuestions: Does dense retrieval understand Boolean logic in language? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 2767–2779, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.156. URL https://aclanthology.org/2024.findings-emnlp.156/.

Xiaonan Zhao, Huan Qi, Rui Luo, and Larry Davis. A weakly supervised adaptive triplet loss for deep metric learning. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 3177–3180. IEEE, 2019.

Mo Zhou, Zhenxing Niu, Le Wang, Zhanning Gao, Qilin Zhang, and Gang Hua. Ladder loss for coherent visual-semantic embedding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 13050–13057, 2020.

A Theoretical Analysis

We analyze the hard-negative Margin-MSE objective with measured violation counts, assuming the presence of no easy negatives. Theorem 1 constructs cosine queries from affine attribute probes under explicit norm and scale assumptions. Theorem 2 shows that, under complete comparison coverage, low count loss on a query pair differing in one condition implies affine recovery of the deleted attribute. Proposition 1 compares this result with binary supervision within Margin-MSE, which can fit the same query pair without preserving the deleted attribute. These results describe feasible representations and information required by low-loss embeddings. Corollary 3 gives a unit-norm construction with logarithmic dependence on the attribute vocabulary size under an explicit sufficient bound on V.

A.1 Setup

Let S be a document set, let f map queries and documents to unit vectors in $\mathbb{R}^d$, and fix $V > 0$. For a condition c, define its satisfaction indicator $h_c(x) = \mathbb{1}[x \models c]$, equal to $z_i(x)$ when c requires attribute i and $1 - z_i(x)$ when it forbids that attribute. For a query q_C with $k = |C| \geq 1$ conditions, define

$$H_C(x) = \sum_{c \in C} h_c(x), \qquad y_C(x) = k - H_C(x). \tag{9}$$

Write $P_C = \{x \in S : H_C(x) = k\}$ and $N_C = S \setminus P_C$. For $p \in P_C$ and $n \in N_C$, the target count is $y_C(n) = H_C(p) - H_C(n)$ and the cosine margin is

$$\begin{aligned} M_C(p,n) &= \cos(f(q_C), f(p)) - \cos(f(q_C), f(n)) \\ &= \langle f(q_C), f(p) - f(n) \rangle, \qquad \mathcal{L}(q_C, p, n) = \big(M_C(p,n) - y_C(n)/V\big)^2. \end{aligned} \tag{10}$$

The inner product is Euclidean; the equality uses unit embeddings and does not normalize the difference $f(p) - f(n)$.

An affine probe (b, β) recovers a function h with error $\epsilon > 0$ on S if

$$|\langle b, f(x) \rangle + \beta - h(x)| < \epsilon \qquad \text{for all } x \in S. \tag{11}$$

If (b_i, β_i) recovers z_i, then $(-b_i, 1-\beta_i)$ recovers $1 - z_i$ with the same error, so recovering a condition indicator and recovering its attribute are equivalent.

A.2 Main Results

Theorem 1 (Affine recoverability permits low loss)**.** *Let $U = \operatorname{span}\{f(x) : x \in S\}$ be a proper subspace of $\mathbb{R}^d$. Suppose every h_c, $c \in C$, has an affine probe (b_c, β_c) with error $\epsilon > 0$ on S and $b_c \in U$, and put $u = \sum_{c \in C} b_c$. Using the attribute probes above, u is their signed sum: required attributes contribute b_i and forbidden attributes contribute $-b_i$. If $V \geq \|u\|$, the unit query*

$$f(q_C) = (u + w)/V, \qquad w \perp U, \qquad \|w\|^2 = V^2 - \|u\|^2 \tag{12}$$

attains $\mathcal{L}(q_C, p, n) < (2k\epsilon/V)^2$ for every $(p, n) \in P_C \times N_C$.

Proof. The assumptions ensure that w exists, that $\|f(q_C)\| = 1$, and that w contributes nothing to document scores. Set $r_c(x) = \langle b_c, f(x) \rangle + \beta_c - h_c(x)$. The probe biases cancel in each difference, giving

$$V M_C(p,n) - y_C(n) = \sum_{c \in C} \big(r_c(p) - r_c(n)\big). \tag{13}$$

Each summand has magnitude below 2ϵ, so $|M_C - y_C/V| < 2k\epsilon/V$; squaring proves the claim. □

Corollary 3 below constructs document embeddings and probes satisfying these assumptions when $V \geq kk'$, where k' bounds the number of active attributes per document.

Theorem 2 (Low loss implies affine recoverability)**.** *Let $\delta > 0$, suppose P_C, N_C are nonempty, and assume $\mathcal{L}(q_C, p, n) < \delta^2$ for every $(p, n) \in P_C \times N_C$. Choose $x_0 \in P_C$ and define*

$$a_C = k - V \langle f(q_C), f(x_0) \rangle, \qquad G_C(x) = V \langle f(q_C), f(x) \rangle + a_C. \tag{14}$$

Then $|G_C(x) - H_C(x)| < 2V\delta$ for every $x \in S$. For $c \in C$, if $D = C \setminus \{c\}$ is also queried and satisfies the same hypotheses on S with the same V, δ, define G_D using $|D|$ in place of k and the same reference $x_0 \in P_C \subseteq P_D$. Then $G_C - G_D$ affinely recovers h_c, and hence its underlying attribute, with error $4V\delta$; the probe for h_c is

$$b = V\big(f(q_C) - f(q_D)\big), \qquad \beta = a_C - a_D. \tag{15}$$

Proof. For $x \in N_C$, $G_C(x) - H_C(x) = y_C(x) - VM_C(x_0, x)$ has magnitude below $V\delta$ by the loss assumption. For $x \in P_C$, choose $n \in N_C$ and subtract the controlled margins $M_C(x, n)$ and $M_C(x_0, n)$. Their targets agree, giving $|G_C(x) - H_C(x)| = V|M_C(x, n) - M_C(x_0, n)| < 2V\delta$. Applying the same argument to D and using $H_C - H_D = h_c$ gives

$$|(G_C - G_D)(x) - h_c(x)| \leq |G_C(x) - H_C(x)| + |G_D(x) - H_D(x)| < 4V\delta. \tag{16}$$

□

At fixed δ, the recovery bound $4V\delta$ grows linearly with V; zero loss gives exact recovery.

To compare binary and count supervision within Margin-MSE, we use the same queries and comparisons with binary target 1 for every positive–negative pair.

Proposition 1 (Binary Margin-MSE can discard attributes)**.** *There are unit embeddings and a query pair $C = \{c_1, c_2\}$, $D = \{c_1\}$, both conditions positive, with zero binary Margin-MSE loss*

$$\mathcal{L}_{\text{bin}}(q_C, p, n) = \big(M_C(p, n) - 1\big)^2 \tag{17}$$

on every positive–negative pair for both queries, although every affine probe for z_2 has maximum error at least $1/2$ on the document set.

Proof. Take one document for each $(z_1, z_2) \in \{0, 1\}^2$ and choose $f(q_C) = e_1$, $f(q_D) = e_2$ in $\mathbb{R}^3$, with

(z_1, z_2)	$f(x)$
$(0, 0)$	$(-\frac{1}{2}, -\frac{1}{2}, \frac{1}{\sqrt{2}})$
$(0, 1)$	$(-\frac{1}{2}, -\frac{1}{2}, \frac{1}{\sqrt{2}})$
$(1, 0)$	$(-\frac{1}{2}, \frac{1}{2}, \frac{1}{\sqrt{2}})$
$(1, 1)$	$(\frac{1}{2}, \frac{1}{2}, \frac{1}{\sqrt{2}})$.

(18)

All vectors have unit norm and give margin 1 on every required pair. Any probe assigns the same value to $(0, 0)$ and $(0, 1)$, so one of its errors for z_2 is at least $1/2$. □

Using our GOLD reweighting, on the other hand, $(0, 0)$ and $(0, 1)$ have different targets because they violate two conditions and one condition of C, respectively. Per Theorem 2, sufficiently small loss on these two queries implies that z_2 is approximately recoverable by an affine probe, even without a query for z_2 alone.

A.3 CAPACITY WITH UNIT EMBEDDINGS

We adapt the approximately orthogonal construction of Garg et al. (2026) by applying one common scale and adding two coordinates. One coordinate normalizes the document embeddings; the other normalizes the query embeddings.

Corollary 3 (Capacity for cosine Margin-MSE)**.** *Let $m \geq 2$, $k \geq 1$, and $1 \leq k' \leq m$ be integers, let $0 < \epsilon \leq 1$, and suppose $z(x) \in \{0, 1\}^m$ has at most k' nonzero entries for every $x \in S$. For every $V \geq kk'$, there exist unit document and query embeddings in dimension*

$$d = O\left(\frac{k'^2}{\epsilon^2} \ln m\right) \tag{19}$$

such that every attribute is affinely recovered with error ϵ on S, and every query with $1 \le |C| \le k$, allowing positive and negated conditions, has $\mathcal{L}(q_C, p, n) < (2k\epsilon/V)^2$ for every $(p, n) \in P_C \times N_C$.

Proof. Set $\mu = \epsilon/k'$. Lemma 5 of the cited work constructs m unit vectors $a_i \in \mathbb{R}^{d_0}$ with $|\langle a_i, a_j \rangle| < \mu$ for $i \neq j$, using $d_0 = O(\mu^{-2} \ln m)$ coordinates. Here d_0 is the dimension before the two normalization coordinates are added; put $A = [a_1 \ \cdots \ a_m]$. Since $\|Az(x)\| \le k'$, define document embeddings and attribute probes in $\mathbb{R}^{d_0+2}$ by

$$f(x) = \left(\frac{Az(x)}{k'}, \sqrt{1 - \frac{\|Az(x)\|^2}{k'^2}}, 0 \right), \qquad \widetilde{b}_i = (k' a_i, 0, 0). \tag{20}$$

Every $f(x)$ has unit norm, including when $z(x) = 0$, and the added coordinates contribute nothing to probe outputs. Since each column has unit norm and distinct columns have inner products bounded in magnitude by μ, we have

$$\left| \left\langle \widetilde{b}_i, f(x) \right\rangle - z_i(x) \right| = \left| \sum_{j \neq i} z_j(x) \langle a_i, a_j \rangle \right| < k'\mu = \epsilon. \tag{21}$$

Project $\widetilde{b}_i$ onto the document span U to obtain b_i; this preserves all probe outputs and gives $\|b_i\| \le k'$. The last coordinate of every document is zero, so U is a proper subspace. Positive and negated conditions use $(b_i, 0)$ and $(-b_i, 1)$, respectively, giving

$$\left\| \sum_{c \in C} b_c \right\| \le |C| k' \le k k' \le V. \tag{22}$$

Theorem 1 therefore supplies unit queries with loss below $(2|C|\epsilon/V)^2 \le (2k\epsilon/V)^2$. Substituting $\mu = \epsilon/k'$ into the construction's dimension bound gives $d_0 = O(k'^2 \epsilon^{-2} \ln m)$. The two added coordinates change this by an additive constant, so $d = d_0 + 2 = O(k'^2 \epsilon^{-2} \ln m)$. □

Notation.

S, f, d Document set, unit embedding map, and embedding dimension.

z_i, h_c, H_C Attribute indicator, condition satisfaction indicator, and number of conditions satisfied.

C, k, q_C Condition set, its size $k = |C|$, and its query; in Corollary 3, k bounds the query size.

P_C, N_C Documents satisfying all conditions and documents violating at least one.

b_i, β_i Affine probe for attribute z_i; (b_c, β_c) denotes a probe for condition indicator h_c.

U, u, w Document span, sum of condition-probe vectors, and orthogonal query component in Theorem 1.

M_C, y_C, V Cosine margin, violation count, and fixed positive label scale.

ϵ, δ Uniform probe-error and margin-error tolerances; loss is below δ^2 on each comparison assumed in Theorem 2.

G_C, a_C Affine predictor of H_C and its offset fixed by a reference positive.

m, k' Attribute vocabulary size and document sparsity in Corollary 3.

A, d_0, μ Matrix of unit attribute columns, dimension before padding, and bound on off-diagonal inner products in the capacity construction.

B FULL RESULTS

Loss	Group	Main Text (Recall@5)	Main Image (Recall@20)	Human (Matched) Text (Recall@5)	Human (Matched) Image (Recall@20)
InfoNCE	Baseline	0.890 [.889–.891]	0.495 [.482–.503]	0.960 [.955–.962]	**0.590** [.578–.602]
	NV-Retriever	0.885 [.884–.886]	0.490 [.486–.494]	0.962 [.955–.970]	0.574 [.542–.590]
	GOLD mined	0.887 [.885–.888]	**0.496** [.493–.500]	0.950 [.947–.955]	0.586 [.578–.602]
	GOLD reweighted (ours)	**0.891** [.888–.894]	0.496 [.494–.497]	0.962 [.947–.977]	0.574 [.542–.590]
SigLIP	Baseline	0.852 [.849–.854]	0.448 [.442–.456]	0.935 [.932–.940]	0.550 [.518–.590]
	NV-Retriever	0.856 [.856–.856]	0.443 [.435–.452]	0.955 [.947–.962]	0.530 [.470–.566]
	GOLD mined	0.845 [.840–.848]	0.450 [.445–.456]	0.925 [.917–.932]	0.522 [.518–.530]
	GOLD reweighted (ours)	0.870 [.869–.870]	0.467 [.454–.474]	0.947 [.940–.955]	0.494 [.458–.542]
Margin-MSE	Baseline	0.854 [.852–.858]	0.434 [.431–.438]	0.940 [.925–.947]	0.574 [.554–.602]
	NV-Retriever	0.870 [.867–.873]	0.414 [.413–.417]	0.962 [.962–.962]	0.546 [.506–.578]
	GOLD mined	0.844 [.843–.846]	0.439 [.434–.443]	0.930 [.925–.932]	0.586 [.578–.590]
	GOLD reweighted (ours)	0.865 [.864–.867]	0.448 [.446–.451]	**0.965** [.962–.970]	0.514 [.506–.518]
CoSENT	Baseline	0.822 [.816–.827]	0.381 [.378–.387]	0.932 [.917–.947]	0.458 [.434–.494]
	NV-Retriever	0.867 [.864–.868]	0.339 [.334–.348]	0.947 [.932–.955]	0.394 [.386–.398]
	GOLD mined	0.815 [.808–.820]	0.363 [.351–.372]	0.917 [.910–.932]	0.434 [.410–.470]
	GOLD reweighted (ours)	0.866 [.865–.868]	0.384 [.371–.393]	0.945 [.940–.947]	0.474 [.470–.482]

Table 2: Test-split recall of every loss and negative-selection group, mean of $n = 3$ seeds with the lowest and highest seed in brackets. Bold is the best value in a column; underline is the best within a loss. Synthetic queries are the ones each model trained on; human queries are the human-written set, scored on the same model. Text is Recall@5 and image Recall@20.

Loss	A vs. B	Synthetic text (Recall@5)	Synthetic image (Recall@20)	Human text (Recall@5)	Human image (Recall@20)
InfoNCE	GOLD mined vs. Baseline	-.003 $p = .873$	+.001 $p = .424$	-.010 $p = .935$	-.004 $p = .753$
	GOLD reweighted vs. Baseline	+.001 $p = .343$	+.001 $p = .452$	+.003 $p = .500$	-.016 $p = .885$
	GOLD reweighted vs. GOLD mined	+.004 $p = .090$	+.000 $p = .564$	+.013 $p = .256$	-.012 $p = .911$
	GOLD mined vs. NV-Retriever	+.002 $p = .299$	+.007 $p = .082$	-.013 $p = .790$	+.012 $p = .330$
	GOLD reweighted vs. NV-Retriever	**+.007** $p = .006$	+.006 $p = .093$	+.000 $p = .659$	+.000 $p = .586$
SigLIP	GOLD mined vs. Baseline	-.007 $p = .987$	+.002 $p = .349$	-.010 $p = .937$	-.028 $p = .903$
	GOLD reweighted vs. Baseline	**+.017** $p < .001$	**+.020** $p < .001$	+.013 $p = .227$	-.056 $p = .988$
	GOLD reweighted vs. GOLD mined	**+.025** $p < .001$	**+.017** $p = .003$	+.023 $p = .083$	-.028 $p = .893$
	GOLD mined vs. NV-Retriever	-.011 $p = .935$	+.007 $p = .189$	-.030 $p = .934$	-.008 $p = .610$
	GOLD reweighted vs. NV-Retriever	**+.013** $p = .011$	**+.024** $p < .001$	-.008 $p = .771$	-.036 $p = .844$
Margin-MSE	GOLD mined vs. Baseline	-.010 $p = .999$	+.005 $p = .210$	-.010 $p = .937$	+.012 $p = .354$
	GOLD reweighted vs. Baseline	**+.011** $p = .011$	**+.015** $p = .014$	+.025 $p = .055$	-.060 $p = .971$
	GOLD reweighted vs. GOLD mined	**+.021** $p < .001$	+.010 $p = .065$	**+.035** $p = .024$	-.072 $p = .991$
	GOLD mined vs. NV-Retriever	-.026 $p = 1.000$	**+.024** $p < .001$	-.033 $p = .945$	+.040 $p = .210$
	GOLD reweighted vs. NV-Retriever	-.005 $p = .808$	**+.034** $p < .001$	+.003 $p = .505$	-.032 $p = .794$
CoSENT	GOLD mined vs. Baseline	-.007 $p = .983$	-.018 $p = .998$	-.015 $p = 1.000$	-.024 $p = .917$
	GOLD reweighted vs. Baseline	**+.044** $p < .001$	+.002 $p = .379$	+.013 $p = .128$	+.016 $p = .361$
	GOLD reweighted vs. GOLD mined	**+.051** $p < .001$	**+.021** $p = .002$	**+.028** $p = .033$	+.040 $p = .118$
	GOLD mined vs. NV-Retriever	-.051 $p = 1.000$	**+.025** $p = .001$	-.030 $p = .984$	+.040 $p = .186$
	GOLD reweighted vs. NV-Retriever	+.000 $p = .532$	**+.045** $p < .001$	-.003 $p = .651$	**+.080** $p = .040$

Table 3: Paired permutation tests for the differences Figure 1 argues about: $A - B$ in recall at each modality's cutoff, with each group represented by the mean of $n = 3$ seeds, on the same queries as Table 2. The null distribution flips the sign of each positive product's per-query differences (text test queries share products; image and human queries are one per product); p is the one-sided p-value for the hypothesis that A is better than B, over 10,000 Monte Carlo flips with the add-one correction. Bold is $p < 0.05$.

C Loss Functions

InfoNCE (van den Oord et al., 2019) treats every document in a batch of B triples as a candidate for every query and is defined as

$$\mathcal{L}_{\text{INCE}}(q_i) = -\ln \frac{\exp\left(s\left\langle f(q_i), f(x_i^+)\right\rangle\right)}{\sum_{j=1}^{B} \exp\left(s\left\langle f(q_i), f(x_j^+)\right\rangle\right) + \exp\left(s\left\langle f(q_i), f(x_j^-)\right\rangle\right)}, \tag{23}$$

where (q_i, x_i^+, x_i^-) is the i-th triple and the sum runs over the $2B$ positives and negatives of the batch. The target is one-hot on x_i^+, so the loss constrains only the ordering of the candidates.

SigLIP (Zhai et al., 2023) replaces the softmax over candidates with an independent sigmoid classifier for each (query, candidate) pair and is defined as

$$\mathcal{L}_{\text{SigLIP}}(q_i) = -\sum_{x \in X} \ln \frac{1}{1 + e^{-t_i(x)(s\langle f(q_i), f(x)\rangle + b)}}, \tag{24}$$

where X is the set of $2B$ candidate documents in the batch, $t_i(x) = +1$ if $x = x_i^+$ and -1 otherwise, and s and b are a learnable scale and bias. Each pair is fit to its own target, so the loss constrains the absolute score of every candidate rather than its rank.

CoSENT (Huang et al., 2024) consumes pairs (q_i, x_i) with ordinal relevance labels r_i and penalizes every two pairs whose similarities violate ordering:

$$\mathcal{L}_{\text{CoSENT}} = \ln\Big(1 + \sum_{r_i > r_j} \exp\big(s\big(\langle f(q_j), f(x_j)\rangle - \langle f(q_i), f(x_i)\rangle\big)\big)\Big), \tag{25}$$

where the sum runs over all ordered pairs of pairs in the batch with $r_i > r_j$. The labels enter only through their ordering, so with binary labels $r_i \in \{0, 1\}$ the loss requires every positive pair to outscore every negative pair by a margin set by s.

D NV-Retriever Implementation Details

We mine only the training split, using `e5-mistral-7b-instruct` for text and `clip-ViT-L-14` for images. For each query, we retrieve the top k candidates and select the highest-scoring non-positive candidate below $(1 - m)$ times the positive score. We sweep $k \in \{10, 100\}$ and relative margins m: no filter, 0.05, 0.1, and 0.2 for $k = 10$; 0.05, 0.1, 0.2, and 0.3 for $k = 100$. Validation selects $k = 10$, $m = 0.1$ for text and $k = 100$, $m = 0.3$ for images. If no candidate survives the filter, we use the lowest-scoring retrieved non-positive candidate.

E List of Prompts

Placeholders in braces are substituted at generation time. Attribute counts appear resolved to their configured value of five per bucket.

E.1 Product Attribute Extraction

Applied to each (positive, hard negative) product pair drawn from ESCI, with both product records serialized into the corresponding placeholders.

```
### Product A:

{positive_product}

### Product B:

{substitute_irrelevant_product}

List up to 5 features common to both products ("common_features"), up to 5 features unique to
     product A ("unique_features_a"), up to 5 features unique to product B ("unique_features_b
     "), and up to 5 features that do not apply to either product ("neither_features")). When
     generating common_features, ensure they are present in both products in some way. For
     example, if product A is "low fat" and product B in "low sugar", then "low calories"
     could be in common_features. If there are fewer than 5 common features, generate as many
     as possible. Do not assume anything about products not written in the description. When
     generating neither_features, ensure they are opposite or mutually exclusive with features
      in one or both of the products. For example, if product A is "red" and product B is "
     orange", then "blue" could be in neither_features. Must not generate negated features,
     such as "not red", "no wifi", "without bluetooth", or "doesn't have x".Dont use "not" or
     "without". Use your imagination and create diverse neither_features. All features should
     be objective and no more than 5 words. Return ONLY JSON: {'common_features': ['feature1',
      'feature2', ...], 'unique_features_a': ['feature1', 'feature2', ...], 'unique_features_b
     ': ['feature1', 'feature2', ...], 'neither_features': ['feature1', 'feature2', ...]}
```

E.2 Visual Attribute Extraction

Applied to each (positive, hard negative) garment pair drawn from DeepFashion In-Shop. Both photographs are supplied to the model alongside the prompt.

```
### Product A:
```

```
The first image is Product A. Metadata: {positive_product}

### Product B:

The second image is Product B. Metadata: {hard_negative_product}

List up to 5 visual features common to both products ("common_features"), up to 5 visual
     features unique to product A ("unique_features_a"), up to 5 visual features unique to
     product B ("unique_features_b"), and up to 5 visual features that do not apply to either
     product ("neither_features").

When generating common_features, ensure they are visible in both products in some way. If
     there are fewer than 5 common features, generate as many as possible. Do not assume
     anything not visible in the images or stated in metadata.

When generating neither_features, ensure they are opposite or mutually exclusive with visual
     features in one or both products. For example, if product A is red and product B is
     orange, then "blue" could be in neither_features. Must not generate negated features,
     such as "not red", "no sleeves", "without pattern", or "doesn't have x". Dont use "not"
     or "without". Use your imagination and create diverse neither_features. All features
     should be objective and no more than 5 words.

Return ONLY JSON: {"common_features": ["feature1", "feature2"], "unique_features_a": ["
     feature1", "feature2"], "unique_features_b": ["feature1", "feature2"], "neither_features
     ": ["feature1", "feature2"]}
```

E.3 Query Rephrasing

Applied to the raw query with the positive product's text and five modality-specific human-written example queries to produce the rewritten queries used in the Main dataset.

```
Original query: '{query}'

Context: The user will be happy with the following product:

{product_description}

Task: Rephrase the query so it keeps the same meaning but sounds more natural and uses
     synonyms.
Rephrase it similar to the style of these example queries, but avoid being exactly like them:

1. {example_query_1}
2. {example_query_2}
3. {example_query_3}
4. {example_query_4}
5. {example_query_5}

Rules:
- Preserve intent exactly (positive stays positive, negative stays negative).
- You MUST rephrase at least half of the comma-separated attributes.
- Do NOT force rewording if it sounds unnatural.
- Do NOT change the product/item name unless a very natural alternative exists.
- Do NOT change measurement units or numeric values.
- Keep the query fluent and realistic.

Return ONLY a JSON object: {"rephrased_query": "..."}
```

F Annotator Instructions

Welcome to the Columbia Advanced Search Study. In this study, you will be helping us generate complex search phrasings for product and image search.

- Do not include any personally identifiable information in your responses
- Do not use AI (ChatGPT, Claude, Gemini, etc.) for this task

If you feel discomfort in this task, you can discontinue at any time.

Below is an example product description, followed by the questions for it.

Imagine you are shopping online for an **ipad 6 gen case.** Read the product description and answer the following questions about it.

[Product Description 1]

[Product Description 2]

1. List n_1 attributes of **Product 1** described in the product description that you like about it. These can be quotes from the product description, up to 8 words long. The attributes must be objective (red ✓, cute ⊘). Separate your attributes with commas. If your example contains an image, make sure the features you are describing are visible in the image.

   Example response: Anti-fingerprint coating, easy access to all buttons, fits iPad Air 2
2. List n_2 attributes that **Product 1 does not have** that you **do not** want it to have. These can be attributes of Product 2, but do not have to be. Separate your attributes with commas.

   Example response: Graffiti pattern, Anti-slip grooves, hands-free viewing
3. Imagine you are searching Amazon for a product that has all of the attributes listed in question (1) and none of the attributes listed in question (2). Write a search as if you were searching Amazon. Make sure to include the item itself (e.g. ipad 6 gen case). It is **okay** to include exact quotes from the product

   Example response: 6 gen ipad case with anti-fingerprint coating, easy access to all buttons fitting iPad Air 2. No graffiti pattern, anti-slip grooves, or hands free viewing.
4. Rewrite the query in (3) to avoid exact quotes from either product description, if possible.

   Example response: 6 gen ipad case with non-glossy coating and accessible buttons for iPad Air 2. Does not have any styling, grooves, or stand.