NSFL: A Post-Training Neuro-Symbolic Fuzzy Logic (NSFL) $\triangle$ Framework for Boolean Operators in Neural Embeddings

Anonymous authors
Paper under double-blind review

Abstract

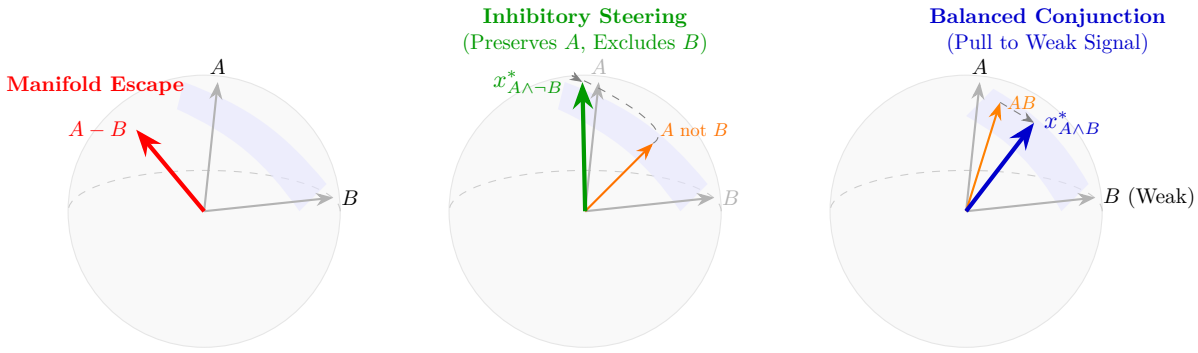
Standard dense retrievers lack a native calculus for multi-atom logical constraints. We introduce **Neuro-Symbolic Fuzzy Logic (NSFL)**, a framework that adapts formal t-norms and t-conorms to neural embedding spaces without requiring retraining. NSFL operates as a **first-order hybrid calculus**: it anchors logical operations on isolated zero-order similarity scores while actively steering representations using **Neuro-Symbolic Deltas (NS-Δ)**—the first-order marginal differences derived from contextual fusion. This preserves pure atomic meaning while capturing domain reliance, preventing the representation collapse and manifold escape endemic to traditional geometric baselines. For scalable real-time retrieval, **Spherical Query Optimization (SQO)** leverages Riemannian optimization to project these fuzzy formulas into manifold-stable query vectors. Validated across six distinct encoder configurations and two modalities (including zero-shot and SOTA fine-tuned models), NSFL yields mAP improvements up to **+81%**. Notably, NSFL provides an additive **20% average and up to 47% boost** even when applied to encoders explicitly fine-tuned for logical reasoning. By establishing a training-free, order-aware calculus for high-dimensional spaces, this framework lays the foundation for future dynamic scaling and learned manifold logic.

1 Introduction

Neural embeddings transform raw data into high-dimensional dense vectors, capturing semantic relationships for tasks like retrieval (Reimers & Gurevych, 2019; Radford et al., 2021). However, these embeddings struggle with Boolean logic queries (e.g., "A AND NOT B"), as fusions often yield additive effects rather than crisp set operations. While many researchers focus on in-training Neuro-Symbolic methods (Badreddine et al., 2022), the Neuro-Symbolic Fuzzy Logic (NSFL) framework operates post-training. This paper motivates NSFL through empirical analyses of embedding behaviors, introducing an empirico-conjectural methodology to bridge the gap between continuous vector similarity and symbolic reasoning.

To this end, we summarize the primary contributions of our work as follows:

- **A First-Order Semantic Calculus:** We introduce the **Neuro-Symbolic Delta (Δ)**, a training-free decomposition that shifts dense retrieval from zero-order similarity points to first-order semantic gradients. This enables the isolation of individual atomic contributions within fused representations, avoiding the geometric trap of convex aggregation.
- **Manifold-Aware Logical Operators:** We formally define asymmetric Conjunctive (AND), Inhibitory (NOT), and Disjunctive (OR) operators that safely steer queries without signal collapse. By resolving *semantic bottlenecks* and treating negation as semantic noise to be pruned, the framework prevents manifold escape. Furthermore, grounded in the *NSFL Delta Directionality* conjecture, we introduce a **boundary stability** method that benefits all logical operations. For the OR operator specifically, this approach extends the classical Zadeh t-conorm to incorporate fused and monolithic representations, elegantly accommodating the superadditivity frequently observed in modern neural retrievers.



(a) Negation by Linear Subtraction (b) NSFL: A NOT B (c) NSFL: A AND B

Figure 1: Comparison of embedding manifold operations: (a) Euclidean operations cause **Manifold Escape** and signal loss; (b) NSFL Inhibitory steering preserves context A while excluding B; (c) NSFL Conjunctive steering resolves the **Semantic Bottleneck** by pulling the query toward the weaker atom B.

- **Scalable Deployment via Spherical Query Optimization (SQO):** To support latency-constrained environments, we introduce SQO, which projects multi-atom logical constraints into a single manifold-stable query vector via Riemannian optimization. While our reranking operators maximize logical accuracy (requiring K=1000 candidate rescoring), SQO offers a complementary deployment mode: complex Boolean queries can be approximated through standard Approximate Nearest Neighbor Search (ANNS) with no post-retrieval computation. This establishes a practical accuracy–latency trade-off, allowing practitioners to choose between optimal reranking and fast single-vector approximation based on system constraints.
- **Robust Empirical Generalization:** We validate our framework across six encoder configurations spanning vision and text modalities (BGE-Large v1.5, Nomic-v2, `intfloat_E5-base-v2`, and `LogiCoL-E5-v2`). NSFL consistently outperforms monolithic querying and geometric baselines, with mAP improvements up to **+81%**. Notably, NSFL is complementary to training-time methods: applied to LogiCoL (Shen et al., 2025), it yields a further **+20% average improvement**, demonstrating additive gains on top of SOTA fine-tuning.

2 Related Work

The integration of symbolic logic with neural representations has evolved from early fuzzy systems to modern high-dimensional dense retrievers. Our work sits at the intersection of fuzzy set theory, vector-symbolic architectures, and modern representation learning.

2.1 Classical Fuzzy Logic and Multi-Valued Logic

Modern fuzzy logic traces its lineage to Zadeh (1965; 1975), who introduced fuzzy sets and their subsequent logical formalizations. Classical operators rely on t-norms and t-conorms, such as those utilized in the Mamdani (Mamdani & Assilian, 1975) and Sugeno (Sugeno, 1974) inference systems. These foundations were later expanded into continuous relaxations of Boolean operators to handle uncertainty in symbolic domains (Hájek, 2001).

Limitations: A primary bottleneck when applying classical fuzzy logic to neural latent spaces is the assumption of calibrated probabilistic scores. As noted by Guo et al. (2020), neural ranking models are typically optimized for relative ranking via contrastive losses rather than absolute truth values. This leads to uncalibrated similarity distributions where standard t-norm axioms, such as monotonicity, are frequently violated.

2.2 Vector Symbolic Architectures and Compositionality

Structural compositionality was explored extensively via Vector Symbolic Architectures (VSA) before the deep learning era. Notable frameworks include Tensor Product Representations (Smolensky, 1990) and Sparse Distributed Memory (Kanerva, 1988). Contributions such as Holographic Reduced Representations (Plate, 1995; 2003) and Multiplicative Binding (Gaylerm, 1998) provided algebraic pathways for logic-like operations in vector spaces.

Limitations: VSAs require explicit binding operations (e.g., circular convolution) that are not natively supported by standard contrastive training objectives, such as CLIP (Radford et al., 2021) or SigLIP (Zhai et al., 2023), which do not inherently support the specific binding operations required by VSA.

2.3 In-Training Neuro-Symbolic Models

Significant progress has been made in differentiable logic and end-to-end neuro-symbolic integration. Logic Tensor Networks (LTN) (Badreddine et al., 2022; Serafini et al., 2017) and DeepProbLog (Manhaeve et al., 2018) represent the state-of-the-art in learnable logical constraints. Other frameworks, such as the Neuro-Symbolic Concept Learner (Mao et al., 2019) and differentiable proving systems (Rocktäschel & Riedel, 2017), focus on grounding symbolic reasoning within neural architectures.

Limitations: These methods typically require full optimization or specialized training regimes. As discussed by d'Avila Garcez & Lamb (2022), training-based Neuro-Symbolic AI is powerful but computationally intensive. In contrast, post-training interventions are often more practical for large-scale, pre-trained retrieval systems.

2.4 Logic in Modern Dense Retrievers

Recent benchmarks such as QUEST (Malaviya et al., 2023), NevIR (Weller et al., 2024) have exposed the logical deficiencies of standard encoders like SBERT (Reimers & Gurevych, 2019). Research has responded with query decomposition (Geva et al., 2021), geometric region-based approaches such as Box (Hamilton et al., 2018; Chen et al., 2021) and Beta Embeddings (Ren & Leskovec, 2020), and logic-aware contrastive learning frameworks like LogiCoL (Shen et al., 2025). As empirically demonstrated by Shen et al. (2025), dense retrievers frequently suffer from representation collapse, treating logically contradictory queries (e.g., "A AND B" versus "A NOT B") as highly correlated semantic equivalents ($r \approx 0.88$). However, correcting this via LogiCoL requires computationally expensive in-training t-norm regularizations on specifically curated logical mini-batches. Furthermore, even with advanced training regimes, modern retrievers face fundamental geometric hurdles: Weller et al. (2025) established that embedding dimensionality imposes a strict theoretical ceiling on the number of document subsets a single query vector can retrieve, while most existing multi-atom fusion methods rely on convex aggregation, which Grabisch (1996) identifies as being strictly limited by the range of the constituent inputs.

Set-Based and Geometric Query Embeddings. A parallel line of research represents logical queries as geometric regions rather than point vectors. Query2Box (Ren et al., 2020) embeds queries as axis-aligned hyperrectangles, enabling intersection via box overlap. Subsequent extensions include BetaE (Ren & Leskovec, 2020), which models queries as Beta distributions to handle negation probabilistically, and ConE (Zhang et al., 2021), which uses cone embeddings for improved expressiveness. While these methods achieve strong results on knowledge graph completion, they require specialized training objectives and custom index structures incompatible with standard ANNS pipelines. NSFL, by contrast, operates post-hoc on arbitrary pre-trained encoders using standard cosine similarity, requiring no architectural modification or retraining.

Set-Based and Geometric Query Embeddings. A parallel line of research represents logical queries as geometric regions rather than point vectors. Query2Box (Ren et al., 2020) embeds queries as axis-aligned hyperrectangles, enabling intersection via box overlap. Subsequent extensions include BetaE (Ren

& Leskovec, 2020), which models queries as Beta distributions to handle negation probabilistically, and ConE (Zhang et al., 2021), which uses cone embeddings for improved expressiveness. While these methods achieve strong results on knowledge graph completion, they require specialized training objectives and custom index structures incompatible with standard ANNS pipelines. NSFL, by contrast, operates post-hoc on arbitrary pre-trained encoders using standard cosine similarity, requiring no architectural modification or retraining.

Comparison with NSFL NSFL departs from existing logical retrieval frameworks by transitioning from zero-order similarity to a training-free, first-order hybrid calculus. While prior methods focus on embedding-space reorganization, NSFL distinguishes itself through three core architectural advantages:

- **Post-hoc Conflict Resolution:** Unlike in-training regularizations (e.g., *LogiCoL*) that mandate global latent space reorganization to prevent representation collapse, NSFL structurally resolves logical conflicts post-hoc. This allows for complex logical steering without the risk of degrading the underlying encoder's general retrieval performance.
- **Non-convex Steering via the NS-Delta (Δ):** Standard convex aggregation often diminishes the identity of individual constituents within a joint embedding. NSFL preserves these identities by processing atoms both in isolation and in context. By extracting the **Neuro-Symbolic Delta** (Δ), the system enables non-convex steering that bypasses the capacity constraints identified by Weller et al. (2025), ensuring atomic logical contributions are neither diluted nor lost.
- **Orthogonality to SOTA:** NSFL is inherently complementary to training-time optimizations. While a fine-tuned model like `LogiCoL-E5` significantly outperforms a zero-shot base model, applying NSFL as a reranking layer on top of `LogiCoL` yields a further **20% average mAP improvement**. This demonstrates that algebraic reranking captures nuances that training-time signals alone may miss, particularly in complex disjunctions ($A \vee B \vee C$) where we observe gains up to **37%**.

3 Empirical Observations on Embedding Behaviors

Our framework is grounded in observed geometric behaviors across six modern contrastive encoder configurations, including BLIP (vision–language transformers), SigLIP (sigmoid-loss models), BGE-Large v1.5 (BERT-style text encoders), Nomic Embed Text v2 (mixture-of-experts), `e5-base-v2`, and the logic-fine-tuned `LogiCoL-E5v2`. Despite their heterogeneous architectures and training objectives, these models exhibit consistent geometric patterns that highlight the fundamental divergence between dense vector similarity and formal Boolean logic.

- **Non-probabilistic and Uncalibrated Scores:** Similarity scores in dense retrieval are typically uncalibrated and do not represent absolute truth values. As noted by Guo et al. (2020), these scores are optimized for relative ranking rather than probability estimation. Furthermore, Shen et al. (2025) demonstrated that standard retrievers often collapse disparate logical operators into a single "union-like" representation, failing to distinguish between intersection and negation.
- **The State Space of Semantic Aggregation:** Unlike classical fuzzy logic which enforces strict monotonicity ($a \wedge b \leq \min(a, b)$), concept fusion in neural latent spaces is highly non-linear. When atomic concepts (A and B) are fused into a monolithic query ($A \cup B$), the resulting joint similarity score (S_{AB}) occupies a complete geometric state space characterized by three distinct behaviors:
  1. *Superadditivity (Synergistic Amplification):* $S_{AB} > S_A + S_B$. The fusion yields a "co-occurrence reward" (Yuksekgonul et al., 2023), creating a semantic reasoning bias where the joint signal is stronger than the sum of its independent parts.
  2. *Subadditivity:* $S_{AB} < S_A + S_B$ (while typically remaining $S_{AB} \geq \min(S_A, S_B)$). The latent representation balances the concepts but fails to preserve their maximum independent magnitudes.

3. *Strong Decrease (Reductive Attenuation):* $S_{AB} < \min(S_A, S_B)$. Incongruent tokens severely dilute the semantic vector (Weller et al., 2024), pulling the query centroid entirely away from the relevant document manifold.

While Shen et al. (2025) attempt to manage these violations via in-training t-norm regularization, NSFL leverages this exact state space post-hoc, utilizing first-order delta analysis to mathematically map these fluctuations to precise logical operations.

- **Negation Stability (Negation as Noise):** Probing tasks reveal that negation tokens (e.g., "not", "without") are often treated as semi-meaningful noise that attenuates signal magnitude without inducing semantic reversal (Weller et al., 2024; Kassner & Schütze, 2020). This lack of directional inversion necessitates a "positive-only" operator approach to avoid instability in the negative regime of the manifold (Thrush et al., 2022).

4 NSFL Empirico-Conjectural Framework

Following the principles of representation learning (Bengio et al., 2013; LeCun et al., 2015), we propose an **Empirico-Conjectural framework** grounded in the geometric behaviors observed in Section 3.

4.1 Notation and Preliminaries

To provide a formal grounding for the **Neuro-Symbolic Fuzzy Logic (NSFL)** framework, we define the following primitives:

Table 1: Summary of Notation and Definitions.

Symbol	Description
$\mathcal{D} = \{D_1, \ldots, D_m\}$	A corpus of data items embedded as vectors $\mathbf{v}_{D_i} \in \mathbb{R}^d$.
L_A, L_B, L_C	Semantic query atoms (e.g., text labels or raw concepts).
$\cup$	The Fusion Operator (FO): $L_{AB} = L_A \cup L_B$. Merges atoms into a joint label.
L_M	Monolithic Query: The full boolean expression as a single natural language string.
$\mathbf{v}_X$	Embedding vector of query component $X \in \{A, B, AB, M\}$.
S_X^i	Similarity score between data item i and query X, normalized to $[0, 1]$.
Δ_X	**Neuro-Symbolic Delta**: The marginal change $S_{AX}^i - S_A^i$ (or $S_{AB}^i - S_B^i$).
S^{sm}	Smoothed variant of an operator utilizing confidence-based gating.
S^{stable}	Final operator score after applying boundary stability (noise floor) logic.

4.2 Realizing the Fusion Operator ($\cup$)

The fusion operator $\cup$ instantiates a logical query ϕ as a single string passed to the encoder. We compare two surface realizations to assess whether NSFL is robust to the syntactic form of the fused query:

- **Simple Fusion:** direct logical templates, e.g., "L_A AND L_B" or "L_A AND NOT L_B".
- **Contextual Phrasing:** natural-language formulations following the QUEST style (Shen et al., 2025), e.g., "L_A that are also L_B" or "L_A that are not L_B".

The two realizations induce different absolute similarity scores S_{AB}^i, but as the $\Delta_\cup$ column of Table 4 shows, NSFL yields positive gains under *both* configurations across all logic types. The *Neuro-Symbolic Delta* therefore captures the semantic effect of the logical operator independently of the surface form used to express it.

4.3 Neuro-Symbolic Delta (Δ)

The **Neuro-Symbolic Delta (NS-Δ)** quantifies the marginal contribution of an individual atom to a fused semantic representation. For a binary fusion, the score S^i_{AB} is modeled as a first-order decomposition:

$$S^i_{AB} = S^i_A + \Delta^i_B = S^i_B + \Delta^i_A \tag{1}$$

To generalize this to a set of n atoms $\mathcal{A} = \{A_1, \ldots, A_n\}$, we define a permutation σ^i that orders atoms by their atomic similarity scores such that $S^i_{A_{\sigma^i(1)}} \leq S^i_{A_{\sigma^i(2)}} \leq \cdots \leq S^i_{A_{\sigma^i(n)}}$. We define the j-th rank-ordered selection function:

$$\text{minchoice}^i_j(\mathcal{A}) = A_{\sigma^i(j)} \tag{2}$$

This allows us to isolate atoms by their relative impact. For instance, minchoice^i_1 identifies the atom with the smallest similarity (the semantic bottleneck), while minchoice^i_n identifies the dominant atom. Using this rank-ordered selection, the fused score is expressed as the sum of the dominant zero-order signal and the marginal deltas of the remaining atoms:

$$S^i_{\text{fused}} = S^i_{\text{minchoice}^i_n(\mathcal{A})} + \sum_{k=1}^{n-1} \Delta^i_{\text{minchoice}^i_k(\mathcal{A})} \tag{3}$$

Methodological Framework of the NS-Delta. We define the **Neuro-Symbolic Delta (Δ)** not as an empirical discovery of a latent property, but as a purposeful **first-order analytic decomposition**. By isolating the difference between an atom's isolated similarity (S^i_A) and its contextualized similarity within a fused query (S^i_{AB}), we create an analytical lens through which the semantic shift induced by concept fusion can be quantified. This operationalization is critical: it allows us to treat the delta as a steerable correction signal, enabling the enforcement of non-convex logical operators (e.g., inhibitory NOT) that are structurally inaccessible via standard zero-order similarity metrics alone.

4.4 Conjectures

We present the following as **empirical conjectures**: testable claims about the behavior of modern contrastive encoders, supported by the observations in Section 3 and by the robust performance of the operators derived from them (Sections 6 and 8.3). A formal geometric proof across arbitrary latent topologies remains open; we offer these conjectures as targets for future verification, and as the design rationale motivating the algebraic choices in Section 6.

Conjecture 1: NS-Delta Directionality (Reinforcement and Departure). The sign of the logical delta (Δ_X) serves as a latent semantic classifier. We posit that a positive delta indicates *manifold reinforcement*, while a negative delta signifies *semantic departure*. This builds on the "Search-with-Explanation" framework by Malaviya et al. (2023), which suggests that the direction of query shifts in vector space correlates with the presence or absence of grounding evidence in the retrieval corpus.

Conjecture 2: Asymmetric Delta-Sensitivity (The Semantic Bottleneck). We posit that atomic constituents contribute unequally to fused representations. This is measurable via the Neuro-Symbolic Delta (Δ_X), where the "weaker" atom—yielding the smallest $|\Delta_X|$—serves as the primary constraint on logical validity. In vector space, this creates a *semantic bottleneck* where constituents with low semantic projection limit the joint vector's alignment with target documents (Shen et al., 2025). This behavior, supported by the *Winoground* benchmark (Thrush et al., 2022), suggests that correct atomic identification does not guarantee logical resolution. Such *semantic dilution* or *query drift* (Mitra & Craswell, 2018; Yuksekgonul et al., 2023) occurs when incongruent components pull the query centroid away from the relevant manifold. Consequently, NSFL weights fusions by this minimal "semantic pull" to prevent high-confidence atoms from masking logical inconsistencies.

Conjecture 3: Negation as Signal Attenuation (Semi-Meaningful Perturbation). We conjecture that semantic negations in dense vector embeddings (e.g., "not B") do not induce a semantic reversal, as expected in classical Boolean logic ($\neg P \equiv 1 - P$). Instead, they manifest as *semi-meaningful perturbations* that attenuate the magnitude of the Neuro-Symbolic Delta $|\Delta_X|$, while preserving its directional sign. This lack of directional inversion necessitates a dedicated logical operator to manually enforce inhibitory behavior.

Table 2 empirically demonstrates **Conjecture 3**: the monolithic score for "dog but not giraffe" (0.26) is lower than the logically incorrect "dog and giraffe" (0.27) and the base "dog" score (0.35), showing the encoder treats negation as semantic noise. Furthermore, the unequal deltas for "dog" (+0.14) and "man" (+0.09) in the congruent case support **Conjecture 2**, confirming that atomic constituents exert varying semantic pull even when both are correctly grounded.

Table 2: Empirical Δ-Extraction and Monolithic Negation Failure (BLIP-Base).

Context	Atom (L_x)	S_x	Query (L_{AB}/L_{Mono})	S_x	NS-Δ_x
	dog	0.35	dog and giraffe	0.27	+0.07 (**Pos.**)
	giraffe	0.20			−0.08 (**Neg.**)
			dog **but not** giraffe	0.26	similar
	dog	0.35	dog and man	0.44	+0.14 (**Pos.**)
	man	0.30			+0.09 (**Pos.**)

5 Formalizing Logic for Latent Spaces

While logical operators have been formalized for symbolic and probabilistic domains, a native calculus for high-dimensional neural manifolds has remained elusive. We propose that the optimal logical score S_L for a data item i can be estimated via a **functional expansion of m-order Neuro-Symbolic differences**:

$$S_L^i = f(S_n^i) + \sum_{k=1}^{m} g_k\left(\Delta_n^{(k),i}\right) + \mathcal{O}\left(\Delta_n^{(m+1),i}\right) \tag{4}$$

Where $f(S_n^i)$ represents the **zero-order similarity** (the base atom-to-item score), g_k are functions of the **Neuro-Symbolic differences** of order k, and the terminal term represents the **residual error** of the m-order approximation. Also where n indexes the dominant atom (i.e., minchoice$_n$), $f(S_n^i)$ represents the zero-order similarity, and g_k are functions of the k-th order Neuro-Symbolic differences. This formulation allows NSFL to adapt mathematically grounded **t-norms** and **t-conorms** into the neuro-symbolic environment as specific functional realizations of these semantic differences.

5.1 Zero and First-Order Semantic Signals

We define the **Neuro-Symbolic Delta (NS-Δ)** as the first-order semantic difference between query atoms. For a query involving a primary context A and an auxiliary atom B, the first-order signal is:

$$\Delta_{B|A}^i = S_{AB}^i - S_A^i \tag{5}$$

As illustrated in Figure 2, while the zero-order signal defines the item's baseline congruence, the $NS - \Delta$ introduces a first-order correction. This enables the NSFL framework to "steer" the query along the manifold toward a logically congruent intersection, rather than just shifting the vector linearly into a semantic void.

6 Logic Operators and Scoring Methods

The **NSFL-Δ Scoring Method** is a re-ranking framework designed to maximize the score gap between items satisfying a Boolean query and those that do not. Following the functional expansion proposed in

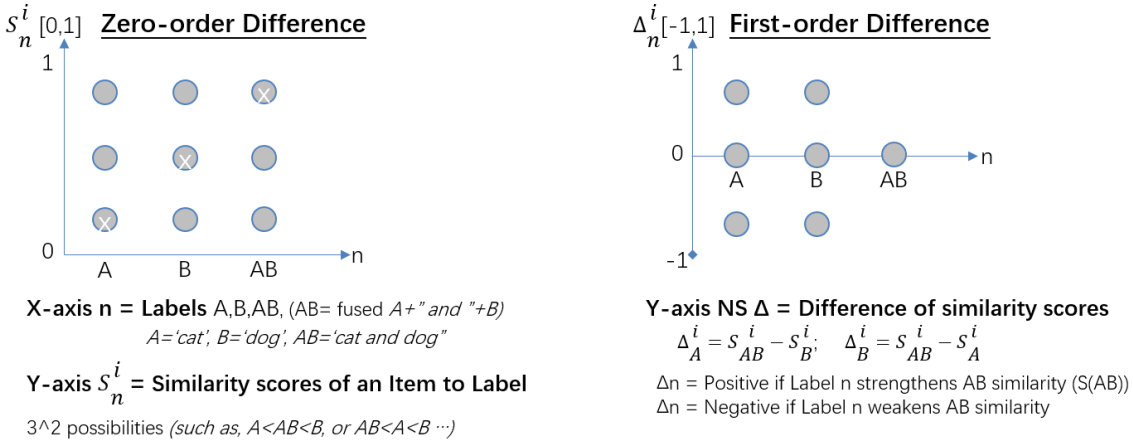


Figure 2: Comparison of zero-order similarity scores and first-order Neuro-Symbolic Deltas across retrieval candidates.

Equation 4, our current implementation focuses on the **first-order approximation** ($m = 1$). It functions by strategically adding or subtracting **NS-Deltas**—the first-order differences—to the zero-order similarity scores to perform controlled amplification of the logical margin. The general logical score S_L for an item i is formulated as:

$$S_L(A, B) := S^L_{\text{base}}(A, B, A \cup B) + \alpha_L \cdot \delta^L_{\text{relevant}}(A, B) \tag{6}$$

Where S^L_{base} provides the manifold-anchored position and $\delta^L_{\text{relevant}}$ serves as the first-order corrective signal. In the following subsections, we define specific operators by mapping these differences to the uncalibrated geometry of neural latent spaces.

6.1 The AND Operator ($A \wedge B$)

The AND operator requires finding the intersection of two concepts on the manifold. Classical fuzzy logic relies on the Gödel t-norm ($T(a, b) = \min(a, b)$), but in a deep learning context this creates a non-differentiable ridge that clips gradients and prevents smooth traversal between semantic states. Instead, NSFL leverages the encoder's native fused representation (S_{AB}) and evaluates its first-order difference against the dominant atomic signal, yielding a differentiable steering mechanism grounded in the parametric Hamacher t-norm and related non-compensatory operators (Klement et al., 2000), adapted here for uncalibrated vector spaces.

Goal: Re-rank conjunctive queries by amplifying the marginal impact of the "weaker" atom (**Conjecture 2**) when the fused representation lacks synergistic reinforcement. When $S_{AB} > S_A + S_B$, the encoder's native co-occurrence reward already satisfies the conjunction; otherwise, delta-amplification ensures the semantic bottleneck drives the final score. If the weaker delta is positive, the rank is boosted; if negative, the item is penalized, directly reflecting a failure to satisfy the strict conjunctive intersection.

$$\begin{aligned} \Delta^i_{minchoice^i_1(A,B)} &= S^i_{AB} - \max(S^i_A, S^i_B) \\ S^i_{A \wedge B} = S^i_{AB} + \Delta^i_{minchoice^i_1(A,B)} &= 2S^i_{AB} - \max(S^i_A, S^i_B) \end{aligned} \tag{7}$$

Boundary Stability (Noise Floor): Following **Conjecture 1 (NS-Delta Directionality)**, if $S^i_{AB} > (S^i_A + S^i_B)$, we retain the inherent "co-occurrence reward" observed in models that exhibit bag-of-words behaviors (Yuksekgonul et al., 2023), gracefully defaulting to the raw fused score to avoid degrading naturally cohesive concepts. Otherwise, delta-amplification enforces the logical bottleneck to prune incongruent results:

$$S^i_{A \wedge B} = \begin{cases} S^i_{AB} & \text{if } S^i_{AB} > S^i_A + S^i_B \\ 2S^i_{AB} - \max(S^i_A, S^i_B) & \text{otherwise} \end{cases} \tag{8}$$

Score Range. The AND formula $2S^i_{AB} - \max(S^i_A, S^i_B)$ can produce values outside $[0, 1]$ when S_{AB} is low relative to atomic scores. In practice, we do not clip these values, as the relative ranking order is preserved and out-of-range scores occur rarely ($< 2\%$ of candidates in our experiments). For applications requiring bounded scores, post-hoc min-max normalization across candidates is straightforward.

6.2 The NOT Operator ($A \wedge \neg B$)

Goal: Subtract the marginal contribution of B from the base score of A, manually enforcing inhibitory behavior that dense encoders fail to induce through native negation tokens (Weller et al., 2024).

$$S^i_{A \wedge \neg B} = S^i_A - \Delta^i_B = 2S^i_A - S^i_{AB} \tag{9}$$

The term $(S^i_{AB} - S^i_A)$ represents the Neuro-Symbolic Delta (Δ_B)—the degree to which B's presence reinforces or interferes with the signal of A. As posited in **Conjecture 3**, encoders treat negation as a noisy modifier that preserves the sign of Δ_B while merely attenuating its magnitude. By explicitly subtracting this delta from the baseline S^i_A, we perform a manual logical inversion of the encoder's directional hints, transforming a non-directional signal into a strictly inhibitory one.

Smoothed Variant with Local Confidence Normalization. To mitigate abrupt ranking shifts, we introduce a confidence-gated variant. Because raw scores are uncalibrated (Section 3), we define $\tilde{S}^i_B$ as a max-normalized score:

$$\tilde{S}^i_B = \frac{S^i_B}{\hat{S}^{\max}_B + \epsilon} \tag{10}$$

where $\hat{S}^{\max}_B$ is either the maximum S_B observed in the current candidate pool, or a pre-computed corpus-level upper bound (typically < 1 for uncalibrated encoders). We omit min-normalization, as filtered candidate pools exhibit selection bias that inflates the observed minimum beyond the true similarity floor.

The smoothed operator interpolates between baseline and full penalty based on local confidence:

$$S^{\text{sm},i}_{A \wedge \neg B} = S^i_A - \tilde{S}^i_B (S^i_{AB} - S^i_A) \tag{11}$$

When $\tilde{S}^i_B \to 1$ (strongest B-match in pool), full negation applies; when $\tilde{S}^i_B \to 0$, the score reverts to S^i_A.

The smoothing gate $\tilde{S}^i_B$ modulates penalty strength; ablations (not shown) confirm it improves stability on low-confidence queries but has minimal impact on aggregate mAP. We retain it for robustness.

Boundary Stability (Noise Floor): Following **Conjecture 1**, when the encoder signals manifold departure ($S^i_{AB} < S^i_A$ and $S^i_{AB} < S^i_B$), we default to S^i_{AB} to prevent over-rotation in regions lacking semantic grounding:

$$S^{\text{stable},i}_{A \wedge \neg B} = \begin{cases} S^i_{AB} & \text{if } S^i_{AB} < S^i_A \text{ and } S^i_{AB} < S^i_B \\ S^i_A - \tilde{S}^i_B (S^i_{AB} - S^i_A) & \text{otherwise} \end{cases} \tag{12}$$

6.3 The OR Operator ($A \vee B$)

Goal: Extend the Zadeh t-conorm to include fused and monolithic scores (S^i_M, the similarity to the full natural-language query L_M), allowing for the superadditivity often observed in retrieval models (Zhai et al., 2023).

$$S^i_{A \vee B} = \max(S^i_A, S^i_B, S^i_{AB}, S^i_M) \tag{13}$$

Boundary Stability (Noise Floor): Following **Conjecture 1 (NS-Delta Directionality)**, when the encoder signals a *manifold departure* ($\Delta_A < 0, \Delta_B < 0$), we introduce a "noise floor" to respect the model's inherent rejection. This prevents the operator from "hallucinating" a positive result from the union of two irrelevant components, ensuring that the logical output remains grounded in regions of the embedding space where the model lacks semantic evidence.

$$S_{A\vee B}^{\text{stable},i} = \begin{cases} \min(S_{AB}^i, S_M^i) & \text{if } S_{AB}^i < S_A^i \text{ and } S_{AB}^i < S_B^i \\ \max(S_A^i, S_B^i, S_{AB}^i, S_M^i) & \text{otherwise} \end{cases} \tag{14}$$

Note: This boundary condition is a *practical stability mechanism*, not a formal t-conorm extension. It does not preserve classical axioms (e.g., associativity, monotonicity) but empirically prevents score inflation in regions where the encoder lacks semantic grounding. The primary OR formulation (max) remains the operative t-conorm; the noise floor serves as a safeguard for edge cases.

Development: This preserves the maximal semantic signal available from atomic, fused, or monolithic (S_M) components. By enforcing a strict "noise floor" via the minimum in the rejection regime, we ensure the logical output does not over-rotate where atomic semantic grounding is absent.

6.4 Compositional Operator AND + AND ($A \wedge B \wedge C$)

Goal: Re-rank conjunctive queries by amplifying the marginal impact of the "weaker" atoms, extending the logic defined in 6.1:

$$\begin{aligned} \sum_{k=1}^{2} \Delta^i_{\text{minchoice}_k^i(\mathcal{A})} &= S_{ABC}^i - \max(S_A^i, S_B^i, S_C^i) \\ S_{A\wedge B\wedge C}^i = S_{ABC}^i + \sum_{k=1}^{2} \Delta^i_{\text{minchoice}_k^i(\mathcal{A})} &= 2S_{ABC}^i - \max(S_A^i, S_B^i, S_C^i) \end{aligned} \tag{15}$$

Boundary Stability (Noise Floor): Following **Conjecture 1**, we respect the synergistic alignment of the triple-fused state when it exceeds the additive baseline. Otherwise, we enforce the logical bottleneck:

$$S_{A\wedge B\wedge C}^i = \begin{cases} S_{ABC}^i & \text{if } S_{ABC}^i > S_A^i + S_B^i + S_C^i \\ 2S_{ABC}^i - \max(S_A^i, S_B^i, S_C^i) & \text{otherwise} \end{cases} \tag{16}$$

6.5 Compositional Operator AND + AND NOT ($A \wedge B \wedge \neg C$)

Goal: Re-rank conjunctive queries by amplifying the marginal impact of the "weaker" atoms between A and B, and as defined in 6.1, then subtract the normalized, marginal contribution of C:

$$S_{A\wedge B\wedge\neg C}^i = S_{A\wedge B}^i - \tilde{S}_C^i(S_{ABC}^i - S_{AB}^i) \tag{17}$$

6.6 Operators Summary

The proposed **Neuro-Symbolic Fuzzy Logic (NSFL)** framework is summarized in Table 3. Each operator is designed as a mathematical remedy for specific geometric failure modes identified in current dense encoders, such as *semantic dilution* and *negation blindness*. By grounding these operations in the **Neuro-Symbolic Delta** (Δ), we ensure that the logical output remains sensitive to the marginal semantic contributions of each query constituent.

Table 3: Comprehensive Summary of NSFL Operators: Fuzzy Logic Formulations and Boundary Stability.

Operator	**Fuzzy Logic (S^{sm})**	**Stability Trigger**	**Boundary Value (S^{stable})**
AND ($A \wedge B$)	$2S_{AB}^i - \max(S_A^i, S_B^i)$	$S_{AB}^i > S_A^i + S_B^i$	S_{AB}^i (Reinforcement)
NOT ($A \wedge \neg B$)	$S_A^i - \tilde{S}_B^i(S_{AB}^i - S_A^i)$	$S_{AB}^i < S_A^i, S_B^i$	S_{AB}^i (Departure)
OR ($A \vee B$)	$\max(S_A^i, S_B^i, S_{AB}^i, S_M^i)$	$S_{AB}^i < S_A^i, S_B^i$	$\min(S_{AB}^i, S_M^i)$ (Departure)

Computational Overhead. NSFL reranking adds negligible latency: rescoring K=1000 candidates requires only 2–3 additional dot-product operations per atom (matrix multiplications over pre-computed embeddings), completing in microseconds on CPU. The dominant cost remains the initial ANNS retrieval.

Notes on Table 3:

- **Confidence Gating:** The **NOT** operator utilizes S_B^i as a smoothing gate to interpolate between the baseline and the negation penalty, preventing rank hallucinations in low-confidence regions.
- **Trigger Asymmetry:** While **NOT** and **OR** boundary conditions are triggered by *manifold departure* (rejection), the **AND** stability check protects *manifold reinforcement*, where the encoder provides a synergistic reward that exceeds atomic sums.

Boundary Case Analysis: Asymmetry in Stability Triggers. As shown in Table 3, the triggers for boundary stability reflect a fundamental divergence in how dense encoders handle synergistic versus inhibitory relationships. For the **AND** operator, the stability check is *upward-facing*: it protects regions of extreme manifold reinforcement where the encoder provides a synergistic co-occurrence reward ($S_{AB}^i > S_A^i + S_B^i$) that naturally exceeds the requirements of the logical intersection. Conversely, for the **NOT** and **OR** operators, the stability checks are *downward-facing*: they respond to *manifold departure* ($\Delta < 0$). In these regimes, where the encoder signals a global rejection of the query constituents, we enforce a strict noise floor to prevent the mathematical rotation of the operator from creating "hallucinated" rankings in regions of the embedding space where the model lacks semantic grounding.

7 Spherical Query Optimization (SQO)

Standard ANNS indices are natively optimized for single-objective distance minimization, making them structurally incompatible with the non-linear requirements of fuzzy logical constraints. To bridge this gap, we propose **Spherical Query Optimization (SQO)**, a pre-search query transformation that translates a complex logical formula into a single optimal theoretical query vector.

7.1 Optimization on the Unit Hypersphere

We assume a corpus $\mathcal{D}$ where item vectors are normalized to the unit hypersphere S^{d-1}. Given a logical formula, we define a continuous and differentiable scoring function $f(x)$ using our proposed **NSFL** operators. We then solve for the optimal theoretical solution:

$$x^* = \arg\max_{x \in S^{d-1}} f(x) \tag{18}$$

This optimization is performed via **Riemannian Stochastic Gradient Descent (RSGD)**. To maintain $x \in S^{d-1}$ without numerical drift, we utilize a formal **retraction map** $R_x : T_x S^{d-1} \to S^{d-1}$ (Absil et al., 2009). For the unit hypersphere, this is computed as the L_2 normalization of the tangent update:

$$R_x(\eta) = \frac{x + \eta}{\|x + \eta\|}, \quad \text{where } \eta = -\alpha \nabla_R f(x) \tag{19}$$

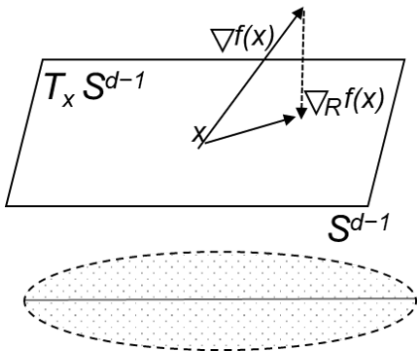


Figure 3: Projection of Euclidean gradient $\nabla f(x)$ onto $T_x S^{d-1}$ during SQO.

As illustrated in Figure 3, RSGD updates the candidate vector x by projecting the Euclidean gradient $\nabla f(x)$ onto the tangent space $T_x S^{d-1}$, followed by the retraction R_x back to the manifold surface. The complete implementation logic and algorithmic steps are detailed in **Algorithm 1** (see Appendix 12.3).

Since our scoring functions $f(x)$ are compositions of N atomic vectors in d-dimensional space, each RSGD iteration maintains a linear computational complexity of $O(N \cdot d)$ (Bonnabel, 2013). In practice, convergence is rapid: empirical seed-variance analysis (n=100 queries, 5 seeds, d=768) confirms mean per-query

runtime under 1ms on CPU with negligible score variance across initializations ($< 0.1\%$). Hyperparameter configurations are provided in Table 6.

7.2 Hybrid Retrieval Strategy

Because our fuzzy logic scoring functions are continuous, vectors in the index that are geometrically proximal to x^* are highly likely to be strong logical matches. We employ a two-step hybrid process to bridge the gap between theoretical optima and practical index constraints:

1. **ANNS Candidate Retrieval:** Perform an ANNS search using x^* to retrieve the top K neighbors $\{v_{i_1}, \ldots, v_{i_K}\}$.
2. **Local Rescoring:** Re-evaluate these K candidates using the full $f(x)$ formulation to identify the best practical match: $\hat{v} = \arg\max_{1 \le j \le K} f(v_{i_j})$.

SQO allows legacy ANNS indices to support complex logical constraints without requiring structural modification or expensive data-dependent pre-processing.

8 Experimental Setup

To evaluate the generalizability of the **NSFL** framework, we designed a cross-modal benchmarking suite covering text-to-text and text-to-image retrieval.

Datasets: 1) **QUEST** (Malaviya et al., 2023): A text-to-text benchmark using a 325k Wikipedia corpus. We utilize the 323 validation queries for modular ablation and the **1,727 test queries** for primary evaluation across six logical templates. 2) **COCO-LOGIC**: A text-to-image benchmark derived from the **COCO-2017** validation set (5k images) (Lin et al., 2014). We generated 600 queries across six templates (100 each) to evaluate visio-linguistic compositionality. Ground truth is strictly mapped to COCO multi-label annotations; for instance, $A \wedge \neg B$ is satisfied only if category A is present and B is absent. The dataset will be released upon publication.

We focus on QUEST and COCO-LOGIC as they provide multi-template logical queries with explicit ground truth. We additionally evaluate on NevIR (Weller et al., 2024) (Appendix 15), a pairwise negation benchmark; NSFL's inhibitory operator yields consistent improvements of **+7% to +10.5%** in pairwise accuracy across encoders.

Zero-Shot Evaluation Protocol: Crucially, our framework requires **no hyperparameter tuning** or training on validation splits; the stability triggers and inhibitory logic derived in Section 3 are applied directly to raw encoder outputs. We evaluate this "zero-shot" logical correction against a standard neural baseline using two pipelines:

- **Baseline Retrieval:** Performs top-k retrieval ($k = 100$) using *Monolithic Query* (L_M) score S^i_M.
- **NSFL Reranking:** Performs an initial "over-sampling" ($K = 1000$) based on L_M, followed by **NSFL** rescoring using atomic (Δ) and stability triggers. Final top-k ($k = 100$) items are re-selected from this corrected pool.

Evaluation Metric: We utilize *Mean Average Precision* (mAP) as our primary metric. Unlike Recall@k, mAP is sensitive to the precise rank-order, capturing the model's ability to prioritize semantically grounded results over logical hallucinations.

Computational Overhead. NSFL reranking adds negligible latency: rescoring K=1000 candidates requires only 2–3 additional dot-product operations per atom (matrix multiplications over pre-computed embeddings), completing in microseconds on CPU. The dominant cost remains the initial ANNS retrieval, making NSFL practical for production deployments.

8.1 Baselines for Comparison

To isolate the specific contributions of the **NSFL** framework, we compare our results against three distinct baseline strategies, ranging from standard neural outputs to naive linear geometric modifications.

1. Monolithic Baseline (Standard): The primary point of comparison is the raw output of the encoder using the fused or natural language query (L_M) without any post-hoc intervention. This represents the current state-of-the-art approach in neural retrieval, where logical constraints are handled implicitly by the model's pre-training on complex phrases.

2. Linear Geometric Baselines: We evaluate two linear strategies to determine if logical intent can be recovered through simple vector arithmetic on the unit hypersphere, a common heuristic in distributional semantics (Mitchell & Lapata, 2008; Kanerva, 2009):

- **Orthogonal Projection (Gram-Schmidt):** For inhibitory constraints (e.g., L_A NOT L_B), we apply Gram-Schmidt orthogonalization (Strang, 2006) to maximize the representation of the primary atom $\mathbf{v}_A$ by removing its projection onto the negated atom $\mathbf{v}_B$:

$$S_{A,B} = \mathbf{v}_A^\top \mathbf{v}_B, \quad \mathbf{v}_{temp} = \mathbf{v}_A - S_{A,B}\mathbf{v}_B, \quad \mathbf{v}^* = \frac{\mathbf{v}_{temp}}{\|\mathbf{v}_{temp}\|_2} \tag{20}$$

- **Normalized Vector Addition:** For conjunctive constraints (e.g., L_A AND L_B), we evaluate the linear summation of atomic embeddings followed by re-normalization (Mitchell & Lapata, 2008):

$$\mathbf{v}_{temp} = \mathbf{v}_A + \mathbf{v}_B, \quad \mathbf{v}^* = \frac{\mathbf{v}_{temp}}{\|\mathbf{v}_{temp}\|_2} \tag{21}$$

Why Not Set Intersection? A natural symbolic baseline is to retrieve independently for each atom and intersect result sets. However, this approach is ill-suited to dense retrieval: (1) similarity scores are continuous, not Boolean—there is no natural intersection threshold; (2) independent top-K retrievals for each atom yield $O(K^n)$ candidate intersections for n atoms, creating prohibitive scaling; (3) intersection discards the relative ranking information that NSFL explicitly leverages via delta signals.

Why Not Sparse or Hybrid Systems? Sparse retrieval (BM25), learned sparse methods (SPLADE), and late-interaction models (ColBERT) are known to handle conjunction-like queries through explicit term matching. However, NSFL targets a different deployment scenario: augmenting *existing* dense retrieval pipelines without retraining or index restructuring. Sparse systems require separate indices and infrastructure; hybrid approaches add complexity. We position NSFL as complementary to—not competing with—these alternatives, offering a lightweight post-hoc solution for systems already committed to dense single-vector retrieval.

Comparison and Failure Analysis: As detailed in Section 8.3, these linear methods demonstrate a fundamental inability to preserve logical intent within the embedding space. While orthogonal projection provides marginal gains in inhibitory scenarios, it is significantly outperformed by our proposed neuro-symbolic operators. More critically, normalized vector addition leads to a **catastrophic degradation of performance** across all conjunctive templates. These failures empirically validate our **Conjecture 2 (Semantic Bottleneck)**, first observed in Section 3, suggesting that naive Euclidean transformations cannot navigate the non-linear semantic densities of modern neural manifolds.

8.2 Correction Tiers and Ablations

Our protocol distinguishes between two operational stages to isolate the impact of logical intervention:

1. **Pre-ANNS Query Optimization:** Evaluates query-side adjustments—specifically *NSFL Query Optimization on Sphere* (SQO)—applied prior to the initial vector search.

2. **Post-ANNS Reranking:** Our primary intervention, applying fuzzy logic operators to the top $K = 1000$ candidates. We compare this against *Geometric Baselines* in our ablation studies (Table 5) to evaluate the necessity of non-linear intervention.

8.3 Results

Table 4: Mean Average Precision (mAP@100) across logical templates. We compare the *Monolithic Baseline* (raw encoder) against *NSFL Reranking*. **Rel.** Δ indicates the percentage improvement over the baseline.

Dataset	Model	$A \wedge B$	$A \wedge B \wedge C$	$A \wedge \neg B$	$A \wedge B \wedge \neg C$	$A \vee B$	$A \vee B \vee C$	**Avg.**	$\Delta_\cup$
Text-to-Text (QUEST)									
`BGE-Large-v1.5`	Baseline	0.038	0.042	0.050	0.015	0.135	0.107	0.065	
	NSFL (Ours)	**0.053**	**0.044**	**0.082**	**0.025**	**0.144**	**0.135**	**0.080**	+0.006
	Rel. Δ	*+37%*	*+5%*	*+63%*	*+61%*	*+7%*	*+25%*	*+24%*	
`Nomic-Embed-v2`	Baseline	0.033	0.032	0.054	0.019	0.158	0.126	0.070	
	NSFL (Ours)	**0.038**	**0.036**	**0.094**	**0.034**	**0.160**	**0.141**	**0.084**	+0.001
	Rel. Δ	*+16%*	*+11%*	*+75%*	*+81%*	*+1%*	*+12%*	*+19%*	
`E5-base-v2`	Baseline	0.045	0.048	0.057	0.020	0.141	0.104	0.069	
	NSFL (Ours)	**0.052**	**0.048**	**0.086**	**0.032**	**0.156**	**0.143**	**0.086**	−0.006
	Rel. Δ	*+16%*	*+1%*	*+51%*	*+60%*	*+11%*	*+37%*	*+25%*	
`LogiCoL-E5-v2`	Baseline	0.083	**0.095**	0.113	0.052	0.215	0.153	0.118	
	NSFL (Ours)	**0.085**	0.093	**0.136**	**0.076**	**0.250**	**0.211**	**0.142**	+0.004
	Rel. Δ	*+3%*	*-1%*	*+20%*	*+47%*	*+16%*	*+38%*	*+20%*	
Image-to-Text (COCO)									
`BLIP-Large`	Baseline	0.142	0.107	0.323	0.154	0.622	0.622	0.328	
	NSFL (Ours)	**0.179**	**0.127**	**0.542**	**0.262**	**0.730**	**0.818**	**0.443**	+0.112
	Rel. Δ	*+26%*	*+19%*	*+68%*	*+70%*	*+17%*	*+31%*	*+35%*	
`SigLIP`	Baseline	**0.166**	**0.093**	0.282	0.137	0.734	0.765	0.363	
	NSFL (Ours)	0.155	0.086	**0.461**	**0.167**	**0.753**	**0.826**	**0.408**	+0.012
	Rel. Δ	*-7%*	*-8%*	*+64%*	*+22%*	*+3%*	*+8%*	*+12%*	

Note: NSFL (Ours) reports the better of the two fusion operators (Simple Fusion vs. Contextual Phrasing, see §4.2) per cell. $\Delta_\cup$ *reports the largest signed gap (Contextual − Simple) across logic types; per-column gaps are in the supplementary material.*

All reported improvements are statistically significant (Wilcoxon signed-rank test, $p < 0.01$; see Appendix 14).

Inhibitory vs. Conjunctive Logic. Disjunctions consistently outperform conjunctions: on BLIP, $A \vee B \vee C$ reaches 0.818 mAP while $A \wedge B \wedge C$ achieves 0.127 mAP. This gap reflects the `AND` operator's sensitivity to atomic signal weighting, formalized as the "Conjunction Gap" in Section 9.

Additive Gains on SOTA. Applying NSFL to `LogiCoL-E5-v2` yields a further **+20% average improvement** over the already-enhanced baseline (0.118 → 0.142 mAP), with gains up to **+38%** on $A \vee B \vee C$. This confirms NSFL is complementary to training-time optimizations.

Recall Comparison. To facilitate comparison with Shen et al. (2025), Table 7 reports Recall@K for LogiCoL-E5-v2, confirming gains are not metric-dependent.

Pairwise Negation (NevIR). On the NevIR benchmark (Weller et al., 2024), NSFL's delta amplification improves pairwise negation accuracy by **+7% to +10.5%** absolute across three encoders, and by **+19% to +25%** under the more stringent "both documents correct" criterion. Full results in Appendix 15.

Table 5: Systematic ablation of modular tiers: mAP@100 averaged over the six logical query types, comparing Pre-ANNS Optimization and Post-ANNS Reranking on BGE/QUEST (text) and BLIP-Large/COCO-Logic (cross-modal). Best non-baseline value per dataset in **bold**. See Table 8 in Appendix 13 for the per-logic-type breakdown.

		Configuration		Avg. mAP@100	
Method	**Variant**	**Pre-Opt.**	**Post-Rerank**	**QUEST**	**COCO-L**
Baseline	Monolithic	—	—	0.065	0.328
NSFL	Rerank only	—	✓	**0.079**	**0.438**
	Opt. only	✓	—	0.069	0.378
	Hybrid	✓	✓	**0.079**	**0.438**
Geometric	Rerank only	—	✓	0.068	0.418
	Opt. only	✓	—	0.059	0.366
	Hybrid	✓	✓	0.065	0.417

Role of SQO. Table 5 confirms that reranking provides the primary accuracy gains; SQO alone yields modest improvements and adds nothing when combined with reranking. However, SQO serves a distinct purpose: it enables logical query constraints through *standard ANNS pipelines* without post-retrieval computation. In latency-critical deployments where reranking K=1000 candidates is prohibitive, SQO offers a single-query approximation compatible with existing indices and sub-millisecond search. We position SQO as a **deployment alternative**, not an accuracy maximizer—users choose between SQO (fast, approximate) and reranking (slower, optimal) based on latency budgets.

Modular Tiers. Table 5 shows NSFL consistently outperforms geometric baselines. While geometric orthogonalization yields modest gains on simple inhibition, it degrades performance on complex compositions ($A \wedge B \wedge \neg C$: -58%), validating our fuzzy-logic approach.

Encoder-Specific Behavior. `SigLIP`'s sigmoid objective causes regression on binary conjunction (-7%), but inhibitory templates recover strongly ($A \wedge B \wedge \neg C$: $+22\%$), demonstrating NS-Delta steering remains effective even with suppressed base signals.

9 Limitations and Broader Impacts

While our framework demonstrates robust empirical gains, we identify two primary boundaries—one mathematical and one societal—that define the critical path for future research.

The Conjunction Gap. Conjunctive operators remain the weakest performers in absolute terms, even after NSFL improvement. Across all encoders, $A \wedge B$ achieves 0.038–0.179 mAP and $A \wedge B \wedge C$ achieves 0.044–0.127 mAP, compared to 0.082–0.542 for inhibitory ($A \wedge \neg B$) and 0.135–0.826 for disjunctive ($A \vee B \vee C$) templates. While NSFL provides consistent relative gains on binary conjunctions ($+3\%$ to $+26\%$), ternary conjunctions show smaller or mixed results ($+5\%$ to $+19\%$ on most encoders, with marginal regression on LogiCoL and SigLIP).

We attribute this to the inherent strictness of conjunction: *all* atomic concepts must be simultaneously satisfied, leaving no room for partial matches. In contrast, disjunction succeeds when *any* atom is present, and inhibition requires only selective exclusion. This asymmetry manifests in encoder behavior—fused representations S_{ABC} for conjunctions are often weaker than their atomic constituents, limiting the corrective power of delta amplification. Additionally, our static coefficient c compounds across atoms:

$$S^i_{A \wedge B \wedge C} = S^i_{ABC} + c \cdot \Delta^i_{\text{bottleneck}} \tag{22}$$

Encoders that already capture conjunctive semantics may be over-corrected, while those with weak fused signals cannot be fully salvaged.

Toward Adaptive Scaling. Our preliminary grid search ($c^* \in [0, 2]$) on COCO-LOGIC yielded +10% improvement for binary AND. We propose two concrete directions for future work:

1. **Geometric Closure Scaling:** Derive c from the angular relationship between atoms: $c = 1 - \cos(\mathbf{v}_A, \mathbf{v}_B)$. When atoms are semantically distant (low cosine), stronger amplification is needed; when aligned, the encoder's fused S_{AB} is more reliable.
2. **Per-Query Learned Coefficients:** Train a lightweight regression model to predict optimal c from query features (atom similarity, delta magnitudes) using a small validation set, preserving the zero-shot spirit on test queries.

Importantly, this limitation affects conjunction only; inhibitory and disjunctive templates show consistently strong improvements across all encoders.

Ethical Considerations and Bias Amplification. Because NSFL operates post-hoc on pre-trained encoders, it inherits representational biases in the latent space. A critical risk arises with the `NOT` operator: if an encoder harbors spurious correlations between an underrepresented demographic and concept B, excluding B may disproportionately suppress results featuring marginalized groups. Deployment should therefore be accompanied by bias-auditing mechanisms. **Empirical validation of demographic fairness across protected attributes remains important future work.**

10 Conclusion

In this work, we introduced **Neuro-Symbolic Fuzzy Logic (NSFL)**, a training-free framework to bridge high-dimensional embedding manifolds with discrete logical constraints through a **first-order hybrid calculus**. By anchoring on isolated zero-order similarities and steering via first-order semantic gradients (Δ), our approach prevents representation collapse and preserves semantic integrity where standard geometric baselines fail. Our experiments across six encoders and two modalities demonstrate that NSFL provides substantial and consistent gains—yielding mAP improvements up to **+81%**—particularly in inhibitory and disjunctive scenarios, with ternary conjunctions representing an isolated limitation addressable via adaptive scaling.

To ensure the practical scalability of this framework, we proposed **Spherical Query Optimization (SQO)**. This method serves as a vital operational bridge, projecting multi-step logical formulas back into manifold-stable query vectors. Consequently, NSFL can be deployed on standard ANNS indices with constant-time query latency (after a one-shot per-query SQO step), requiring no structural modifications or custom distance metrics.

Finally, while our empirical analyses reveal a persistent "Conjunction Gap" in modern encoders, this limitation defines a clear and exciting roadmap for future research. To fully harmonize strict symbolic logic with the uncalibrated, continuous nature of neural vector similarities—which, under semantic aggregation, may exhibit superadditivity, subadditivity, or a strong decrease in signal strength—we identify two primary paths: (1) **Dynamic Scaling**, utilizing formulaic coefficients derived from geometric closure to exceed the constraints of traditional smoothing; and (2) **Learned Coefficients**, which offer the potential for encoder-specific optimization. By evolving from static operators to these adaptive, manifold-aware scaling strategies, we lay the foundation for a truly unified retrieval architecture that respects both continuous semantic nuance and strict logical rigor.

Although our current Hybrid configuration shows no aggregate gain over Rerank-only (Table 5), future work may explore adaptive hybrid strategies where SQO pre-filtering operates on smaller candidate pools tuned per query, potentially combining the speed of approximate search with the precision of algebraic rescoring.

References

P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2009. Foundational text for retraction maps and Riemannian convergence.

Samy Badreddine, Artur d'Avila Garcez, Luciano Serafini, and Michael Spranger. Logic tensor networks. *Artificial Intelligence*, 303(C):103649, February 2022. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2021.103649. URL https://www.sciencedirect.com/science/article/pii/S0004370221002009.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Silvere Bonnabel. Stochastic gradient descent on riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013. Provides the formal proofs for RSGD convergence on compact manifolds.

Xuelu Chen, Michael Boratko, Muhao Chen, Shib Sankar Dasgupta, Xiang Lorraine Li, and Andrew McCallum. Probabilistic box embeddings for uncertain knowledge graph reasoning. 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. 882–893, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.68. URL https://aclanthology.org/2021.naacl-main.68/.

Artur d'Avila Garcez and Luis C. Lamb. Neurosymbolic ai: The 3rd wave. *Artificial Intelligence Review*, pp. 1–21, 2022.

Ross W. Gaylerm. Multiplicative binding, representation operators and analogy, 1998. URL http://cogprints.org/502/.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 2021.

Michel Grabisch. The application of fuzzy integrals in multicriteria decision making. *European Journal of Operational Research*, 89:438–456, 1996.

Jiafeng Guo, Yixing Fan, Liang Pang, Liu Yang, Qingyao Ai, Hamed Zamani, Chen Wu, W Bruce Croft, and Xueqi Cheng. A deep look into neural ranking models for information retrieval. *Information Processing & Management*, 57(6):102067, 2020.

Petr Hájek. *Metamathematics of fuzzy logic*, volume 4. Springer Science & Business Media, 2001.

William L. Hamilton, Payal Bajaj, Marinka Zitnik, Dan Jurafsky, and Jure Leskovec. Embedding logical queries on knowledge graphs. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 2030–2041, Red Hook, NY, USA, 2018. Curran Associates Inc.

Pentti Kanerva. *Sparse distributed memory*. MIT Press, 1988.

Pentti Kanerva. Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors. *Cognitive Computation*, 1(2):139–159, 2009. Foundational for vector-based logic and superposition.

Nora Kassner and Hinrich Schütze. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7811–7818, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.698. URL https://aclanthology.org/2020.acl-main.698/.

Erich Peter Klement, Radko Mesiar, and Endre Pap. *Triangular Norms*. Kluwer Academic, 2000.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, volume 8695 of *Lecture Notes in Computer Science*, pp. 740–755. Springer, 2014.

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/.

Ebrahim H. Mamdani and Setrak Assilian. An experiment in linguistic synthesis with a fuzzy logic controller. *International Journal of Man-Machine Studies*, 7:1–13, 1975.

Robin Manhaeve, Sebastijan Dumancic, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. Deepproblog: Neural probabilistic logic programming. In *Advances in Neural Information Processing Systems*, 2018.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *International Conference on Learning Representations*, 2019.

Jeff Mitchell and Mirella Lapata. Vector-based models of semantic composition. In *Proceedings of ACL-08: HLT*, pp. 236–244, 2008. Discusses the limitations of simple vector addition for semantic composition.

Bhaskar Mitra and Nick Craswell. An introduction to neural information retrieval. *Foundations and Trends $\hat{W}$ in Information Retrieval*, 13(1):1–126, 2018. doi: 10.1561/1500000061.

Tony A Plate. Holographic reduced representations. *IEEE Transactions on Neural networks*, 6(3):623–641, 1995.

Tony A. Plate. *Holographic reduced representation: Distributed representation for cognitive structures*, volume 150. CSLI Publications, 2003.

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.

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. *Advances in Neural Information Processing Systems*, 33:19716–19726, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/e43739bba7cdb577e9e3e4e42447f5a5-Paper.pdf.

Hongyu Ren, Weihua Hu, and Jure Leskovec. Query2box: Reasoning over knowledge graphs in vector space using box embeddings. In *ICLR*, 2020.

Tim Rocktäschel and Sebastian Riedel. End-to-end differentiable proving. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/b2ab001909a8a6f04b51920306046ce5-Paper.pdf.

Luciano Serafini, Ivan Donadello, and Artur d'Avila Garcez. Learning and reasoning in logic tensor networks: theory and application to semantic image interpretation. In *Proceedings of the Symposium on Applied Computing*, SAC '17, pp. 125–130, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450344869. doi: 10.1145/3019612.3019642. URL https://doi.org/10.1145/3019612.3019642.

Yanzhen Shen, Sihao Chen, Xueqiang Xu, Yunyi Zhang, Chaitanya Malaviya, and Dan Roth. LogiCoL: Logically-informed contrastive learning for set-based dense retrieval. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 12114–12125, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.608. URL https://aclanthology.org/2025.emnlp-main.608/.

Paul Smolensky. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, 46(1):159–216, 1990. ISSN 0004-3702. doi: https://doi.org/10.1016/0004-3702(90)90007-M. URL https://www.sciencedirect.com/science/article/pii/000437029090007M.

Gilbert Strang. *Linear Algebra and Its Applications*. Thomson Brooks/Cole, Belmont, CA, 4 edition, 2006. ISBN 0-03-010567-6.

Michio Sugeno. *Theory of fuzzy integrals and its applications*. PhD thesis, Tokyo Institute of Technology, 1974.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5238–5248, June 2022.

Orion Weller, Dawn Lawrie, and Benjamin Van Durme. NevIR: Negation in neural information retrieval. In Yvette Graham and Matthew Purver (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2274–2287, St. Julian's, Malta, March 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.eacl-long.139. URL https://aclanthology.org/2024.eacl-long.139/.

Orion Weller, Michael Boratko, Iftekhar Naim, and Jinhyuk Lee. On the theoretical limitations of embedding-based retrieval. *arXiv preprint arXiv:2508.21038*, 2025.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words models, and what to do about it. In *International Conference on Learning Representations (ICLR)*, 2023. URL arXivpreprintarXiv:2210.01936.

L. A. Zadeh. Fuzzy sets. *Information and Control*, 8:338–353, 1965.

L. A. Zadeh. Fuzzy logic and approximate reasoning. *Synthese*, 30(3-4):407–428, 1975. doi: 10.1007/bf00485052.

Xiaohua Zhai, Basil Mustafa, Alexander Kolevatov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11975–11986, October 2023.

Zhanqiu Zhang, Jie Wang, Jiajun Chen, Shuiwang Ji, and Feng Wu. Cone: Cone embeddings for multi-hop reasoning over knowledge graphs. In *NeurIPS*, 2021.

11 Appendix

You may include other additional sections here.

12 Geometric Justification of Asymmetric Delta-Sensitivity

To formalize the intuition behind Conjecture 2, we consider the neural retrieval process as a projection within a high-dimensional Hilbert space $\mathcal{H} \cong \mathbb{R}^d$. We represent atomic query constituents as unit vectors $\mathbf{a}, \mathbf{b} \in \mathcal{H}$ and a target document as $\mathbf{d} \in \mathcal{H}$.

The similarity score S is typically computed as the cosine similarity, which simplifies to the dot product for normalized vectors:

$$S(\mathbf{x}, \mathbf{d}) = \langle \mathbf{x}, \mathbf{d} \rangle = \cos(\theta_{\mathbf{x},\mathbf{d}}) \tag{23}$$

12.1 The Delta as a Displacement Gradient

When fusing two atoms via a normalized centroid (the standard mechanism for multi-concept queries in encoders such as CLIP or SBERT), the joint query vector $\mathbf{q}_{A \cup B}$ is defined as:

$$\mathbf{q}_{A \cup B} = \frac{\mathbf{a} + \mathbf{b}}{\|\mathbf{a} + \mathbf{b}\|} \tag{24}$$

The Neuro-Symbolic Delta for atom B, denoted Δ_B, measures the marginal shift in the alignment profile when B is introduced to the fusion:

$$\Delta_B = S(\mathbf{q}_{A \cup B}, \mathbf{d}) - S(\mathbf{a}, \mathbf{d}) \tag{25}$$

12.2 Asymmetry and the Bottleneck Principle

Geometrically, if $|\Delta_B| \approx 0$, it implies that the addition of $\mathbf{b}$ failed to significantly alter the projection of the query onto the document manifold. This occurs under two conditions:

1. **Orthogonality:** $\mathbf{b}$ is semantically irrelevant to $\mathbf{d}$ ($\langle \mathbf{b}, \mathbf{d} \rangle \approx 0$).
2. **Directional Dominance:** The magnitude of $\mathbf{a}$'s alignment with $\mathbf{d}$ is so high that $\mathbf{b}$ cannot provide additional constructive interference.

In classical fuzzy logic, the truth of a conjunction is constrained by the Gödel t-norm: $T(A \wedge B) = \min(T(A), T(B))$. In the neural regime, we posit that Δ_X serves as a continuous proxy for this constraint. A small $|\Delta_X|$ signifies a *semantic bottleneck*, where an incongruent or weak constituent prevents the joint vector from achieving the necessary "pull" toward the target.

This justifies the NSFL requirement to weight logical fusions by the minimum marginal contribution, ensuring that high-similarity atoms (e.g., a "strong" A) do not mask the logical invalidity of a "weak" or mismatched B.

12.3 Reproducibility and Hyperparameters

We provide the formal pseudocode for the Spherical Query Optimization (SQO) process. This implementation maps a logical formula to an objective function $f(x)$ using our proposed Neuro-Symbolic Fuzzy Logic (NSFL) operators, followed by an optimization on the unit hypersphere S^{d-1}.

Initialization Stability. SQO uses Gaussian random initialization. To verify robustness, we ran 5 random seeds across 100 synthetic queries at d=768 for both AND and NOT operators. The mean score range across seeds was 0.00058 for AND and $< 10^{-6}$ for NOT, with mean wall-clock time of 0.51ms and 0.32ms per query respectively. These results confirm that the NSFL objective landscape is smooth and unimodal: SQO converges to effectively identical optima regardless of initialization, validating single-seed reporting throughout the paper.

To ensure the reproducibility of the **Spherical Query Optimization (SQO)** results, we provide the exact configuration used for the Riemannian Stochastic Gradient Descent (RSGD) throughout our experiments.

Table 6: Hyperparameter Configuration for SQO Implementation

Parameter	Implementation Detail	Value
α	Learning Rate (Step Size)	0.2
Steps	Maximum Iterations	100
Patience	Early Stopping (No improvement)	10
tol	Convergence Tolerance	10^{-6}
init	Initialization Strategy	Gaussian ($\mathcal{N}(0,1)$)
$Manifold$	Geometry	$\mathcal{S}^{d-1}$ (Unit Sphere)

Algorithm 1 Spherical Query Optimization (SQO)

Require: Formula ϕ, Constituents $\{v_k\}$, Monolithic vector v_M, Steps S, rate η, tolerance τ, patience P
Ensure: Optimized query vector x^*

// Part 1: Objective Function Construction
if $\phi = A \wedge B$ **then**
  function $f(x)$
    $S_A \leftarrow x \cdot v_A, \quad S_B \leftarrow x \cdot v_B$ ▷ Atomic Similarity Scores
    $S_{AB} \leftarrow x \cdot v_{AB}$ ▷ Fused Representation Similarity
    return $2S_{AB} - \max(S_A, S_B)$ ▷ AND T-norm
  end function
end if
// Part 2: Riemannian Optimization on S^{d-1}
$x \leftarrow x_0/\|x_0\|$ ▷ Initial Projection to Unit Sphere
$best\_loss \leftarrow \infty$, counter $\leftarrow 0$
for $step = 0$ **to** S **do**
  $g \leftarrow \nabla f(x)$ ▷ Automatic Differentiation
  $g_R \leftarrow g - (x \cdot g)x$ ▷ Riemannian Gradient Projection
  $x \leftarrow x - \eta g_R$ ▷ Gradient Update
  $x \leftarrow x/\|x\|$ ▷ Retraction (L_2 Normalization)
  $loss \leftarrow -f(x)$ ▷ Maximize objective by minimizing negative loss
  if $loss + \tau < best\_loss$ **then**
    $best\_loss \leftarrow loss$, counter $\leftarrow 0$
  else
    counter $\leftarrow$ counter $+ 1$
  end if
  if counter $\geq P$ **then**
    break ▷ Early Stopping
  end if
end for
return $x^* \leftarrow x$

Note: Algorithm 1 hereby introduces only the AND case, Other operators are obtained by substituting the corresponding $f(x)$ from Section 6; full implementations are in the published code..

12.4 Comparison with LogiCoL under Recall Metrics.

To facilitate direct comparison with Shen et al. (2025), we report Recall@K for the LogiCoL-E5-v2 encoder in Table 7. NSFL reranking yields consistent improvements across all cutoffs, confirming that our mAP@100 gains (Table 4) are not an artifact of metric choice.

Table 7: Recall@K Comparison on QUEST using LogiCoL-e5-v2. Baseline denotes monolithic retrieval; NSFL denotes reranking with our proposed operators.

Metric	Method	$A \wedge B$	$A \wedge B \wedge C$	$A \wedge \neg B$	$A \wedge B \wedge \neg C$	$A \vee B$	$A \vee B \vee C$
R@20	Baseline	0.178	0.198	0.172	0.089	0.298	0.225
	NSFL	**0.185**	**0.206**	**0.200**	**0.122**	**0.338**	**0.283**
R@100	Baseline	0.338	0.406	0.397	0.237	0.504	0.423
	NSFL	**0.350**	**0.415**	**0.423**	**0.284**	**0.550**	**0.491**

13 Per-Logic-Type Ablation Breakdown

Table 5 in the main paper reports the average mAP@100 for each variant of our framework. To complement that view, Table 8 expands the same experiments across the six logical query types, on both BGE/QUEST (text retrieval) and BLIP-Large/COCO-Logic (cross-modal retrieval). Two observations that are obscured by the averaged view are worth noting. First, **GEO Rerank-only collapses on positive conjunctions** ($A \wedge B$, $A \wedge B \wedge C$) while remaining competitive on queries involving negation: on COCO-Logic it actually achieves the best score on $A \wedge \neg B$ (0.639 vs. 0.514 for NSFL Rerank-only). Second, **NSFL is uniformly strong across all logic types and both modalities**, which is what makes its averaged numbers consistently lead the table. The asymmetry suggests that the geometric operators encode useful structure for set-difference queries but lose signal on intersections, a direction we leave to future work.

Table 8: Per-logic-type ablation of NSFL and Geometric variants on BGE/QUEST and BLIP-Large/COCO-Logic. **Bold** marks the best non-baseline value per column within each encoder block.

Encoder	Method	$A \wedge B$	$A \wedge B \wedge C$	$A \wedge \neg B$	$A \wedge B \wedge \neg C$	$A \vee B$	$A \vee B \vee C$	**Avg.**
BGE / QUEST								
BGE	Baseline	0.038	0.042	0.050	0.015	0.135	0.107	0.065
	NSFL — Rerank only	**0.053**	**0.044**	0.075	**0.025**	0.144	**0.135**	**0.079**
	NSFL — Opt only	0.040	0.039	0.073	0.021	0.135	0.107	0.069
	NSFL — Hybrid	0.052	**0.044**	**0.076**	0.024	**0.147**	0.129	**0.079**
	GEO — Rerank only	0.021	0.023	0.067	0.018	0.144	**0.135**	0.068
	GEO — Opt only	0.022	0.023	0.054	0.015	0.135	0.107	0.059
	GEO — Hybrid	0.022	0.023	0.054	0.015	0.144	**0.135**	0.065
BLIP-Large / COCO-Logic								
BLIP-L	Baseline	0.142	0.107	0.323	0.154	0.622	0.622	0.328
	NSFL — Rerank only	**0.179**	0.127	0.514	**0.262**	**0.730**	**0.818**	**0.438**
	NSFL — Opt only	0.160	0.123	0.495	0.243	0.622	0.622	0.378
	NSFL — Hybrid	0.178	**0.134**	0.507	0.260	**0.730**	**0.818**	**0.438**
	GEO — Rerank only	0.049	0.022	**0.639**	0.251	**0.730**	**0.818**	0.418
	GEO — Opt only	0.048	0.021	0.632	0.250	0.622	0.622	0.366
	GEO — Hybrid	0.048	0.021	0.632	0.250	**0.730**	**0.818**	0.417

14 Statistical Significance Analysis

We performed Wilcoxon signed-rank tests comparing per-query Average Precision between Baseline and NSFL across all encoder–template configurations. Tables 9 and 10 show detailed results for BGE-Large-v1.5 on QUEST and BLIP-Large on COCO-Logic as representative examples.

Summary. Across all 72 tested configurations (encoder × fusion operator × logic template), **43** (60%) achieved $p < 0.01$ and an additional 8 achieved $p < 0.05$. Given that the majority of significant results exhibit $p < 10^{-10}$, these findings remain robust under conservative multiple-comparison corrections (e.g.,

Holm-Bonferroni). The non-significant cases were concentrated on the $A \wedge B$, $A \wedge B \wedge C$ templates, where per-query gains are smaller in absolute terms (see Section 9). Full per-configuration numerical results are available in the supplementary material.

Table 9: Statistical significance analysis for BGE-Large-v1.5 on QUEST (per-query Wilcoxon signed-rank test).

Template	Baseline	NSFL	Δ	95% CI	p-value
$A \wedge B$	0.038	0.053	+0.014	[+0.004, +0.025]	0.044
$A \wedge B \wedge C$	0.042	0.044	+0.002	[-0.005, +0.009]	0.018
$A \wedge \neg B$	0.050	0.075	**+0.025**	[+0.017, +0.035]	$< 10^{-11}$
$A \wedge B \wedge \neg C$	0.015	0.025	**+0.009**	[+0.005, +0.014]	$< 10^{-4}$
$A \vee B$	0.135	0.144	+0.009	[-0.001, +0.019]	0.128
$A \vee B \vee C$	0.107	0.135	**+0.027**	[+0.019, +0.036]	$< 10^{-10}$

Table 10: Statistical significance analysis for BLIP-Large on COCO-Logic (per-query Wilcoxon signed-rank test).

Template	Baseline	NSFL	Δ	95% CI	p-value
$A \wedge B$	0.140	0.177	**+0.038**	[+0.018, +0.061]	$< 10^{-4}$
$A \wedge B \wedge C$	0.107	0.127	**+0.020**	[+0.010, +0.032]	$< 10^{-5}$
$A \wedge \neg B$	0.244	0.386	**+0.142**	[+0.117, +0.168]	$< 10^{-15}$
$A \wedge B \wedge \neg C$	0.149	0.256	**+0.106**	[+0.074, +0.143]	$< 10^{-10}$
$A \vee B$	0.240	0.285	**+0.045**	[+0.035, +0.055]	$< 10^{-16}$
$A \vee B \vee C$	0.155	0.206	**+0.051**	[+0.043, +0.058]	$< 10^{-17}$

15 Evaluation on NevIR Pairwise Negation Benchmark

NevIR (Weller et al., 2024) evaluates negation understanding via paired queries: q_1 (positive) and q_2 (negated, differing by a single negation token), alongside two documents where d_1 is correct for q_1 and d_2 for q_2. The metric is pairwise accuracy: does the model rank the correct document higher for each query?

NSFL Application. For positive queries (q_1) we use standard cosine similarity. For negated queries (q_2), we apply delta amplification following Conjecture 3 (negation as attenuation): $S_{\text{NSFL}}(q_2, d) = 2S(q_2, d) - S(q_1, d)$. This amplifies the subtle score differences the encoder assigns due to negation, transforming an attenuated signal into a discriminative one.

Table 11: Pairwise accuracy on NevIR negation benchmark.

	Baseline		NSFL	
Encoder	**Overall**	**Both correct**	**Overall**	**Both correct**
BGE-Large-v1.5	0.6114	0.2560	**0.7169**	**0.5083**
E5-base-v2	0.5766	0.1909	**0.6764**	**0.4382**
LogiCoL-E5-v2	0.5437	0.1171	**0.6139**	**0.3095**

These results confirm that NSFL's inhibitory operator generalizes beyond set-based retrieval to pairwise ranking scenarios, the primary evaluation paradigm in the negation literature. **Across encoders, NSFL yields consistent improvements of approximately +7% to +10.5% absolute in overall pairwise accuracy, and larger gains of +19% to +25% under the more stringent "both documents correct" setting.**