

Know-Thy-Not: Training-Free Targeted Negation in Dual-Encoder Vision-Language Models

DoYoung Kim¹, Changkyu Choi², Adín Ramírez Rivera², YunSeok Choi^{1†}, Jee-Hyong Lee^{1†}

¹Sungkyunkwan University, South Korea

²University of Oslo, Norway

Abstract

Dual-encoder vision-language models systematically fail at negation, scoring “a cat without a dog” nearly identically to “a cat with a dog.” We show that this failure is inherent to the single-embedding framework. Under contrastive training, single-embedding methods can at best ignore the negated concept (dilution) but never penalize its presence (rejection). We propose **Know-Thy-Not (KTN)**, a training-free method that introduces an image-side rejection dimension to penalize the negated concept directly in the visual representation. We also introduce COCO-Confusion, the first benchmark requiring genuine rejection to succeed. Across 9 backbones, KTN improves COCO-Confusion mAP by up to +25.6 over vanilla VLMs and +9.6 over the strongest existing method.

1 Introduction

Dual-encoder vision-language models (VLMs) [1, 2] have become foundational backbones for multimodal systems, gaining structural flexibility by aligning images and text in a shared embedding space. This flexibility has made VLMs the default starting point for nearly every modern vision-language application, ranging from zero-shot image classification [3, 4] and text-image retrieval [5, 6] to image generation without task-specific finetuning [7–9].

However, a growing body of work has revealed that these models struggle with compositional reasoning: they fail at attribute binding [10] and spatial understanding [11], with dedicated benchmarks [12–14] consistently exposing near-chance performance on structured linguistic tests. Asking a VLM to retrieve “a car *without* a traffic light” returns images dominated by traffic lights (Figure 1, top), as the model treats the query as if the negation were absent. This failure is consistent and severe: On VOC2007, CLIP scores 82% on affirmative queries but only 3% on their negated counterparts [15]. Negation is not a peripheral linguistic phenomenon. It is essential to expressing what users do not want, and its absence undermines VLMs as reliable retrieval and grounding systems.

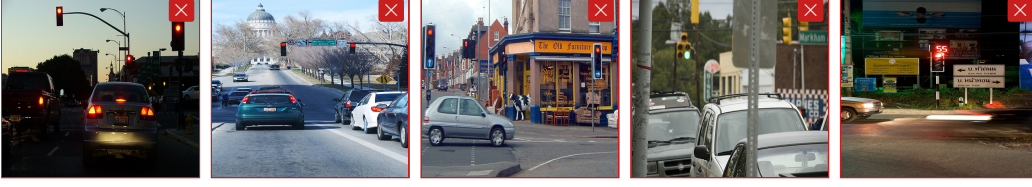
Recent work has attempted to close this gap from two directions. Retraining-based methods such as ConCLIP [16] and NegationCLIP [17] augment training data with negation captions and retrain the model, but show modest improvements despite substantial cost. Alternatively, the *training-free* method [18] extracts the negated concept from the query and suppresses it by projecting out its direction from the text embedding, achieving notable gains without retraining. Both, however, still represent the negated query within a single embedding vector, which we show is structurally insufficient for negation.

We introduce a formal analysis which reveals that the optimal text coefficient (6) on the negated concept is driven to zero by the non-negative nature of natural visual data. This implies that approaches operating solely on the text embedding are inherently bounded. The conventional embedding correction approach, which cancels the direction of the negated concept from the text embedding,

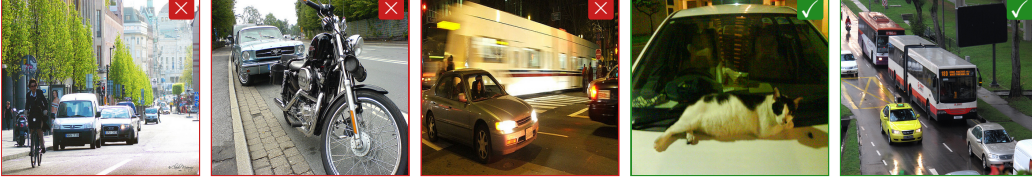
[†]Corresponding authors.

Query: “a car is present *but no* traffic light” (COCO-Confusion, SigLIP2 SO400M)

Vanilla



Embedding correction [18]



Proposed method

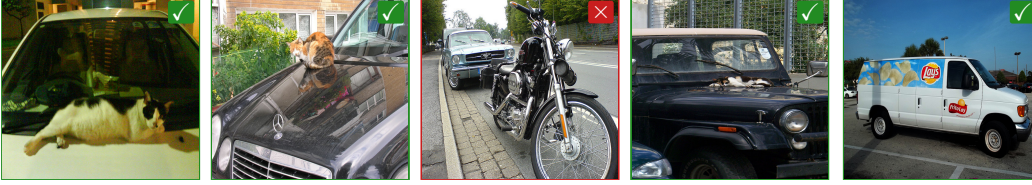


Figure 1: **Top-5 retrieval on our COCO-Confusion benchmark** (leftmost is rank 1). A vanilla VLM ignores the negation, retrieving primarily confusing images that contain a **traffic light**. The embedding correction approach [18] removes the negated direction from the text embedding; while this filters out some confusing images (ranks 4–5), it still places **traffic-light** scenes at the top. In contrast, our proposed *Know-Thy-Not* (KTN) introduces an explicit image-grounding step, successfully retrieving correct images. **Green borders** indicate images containing only the affirmed concept (2), while **red borders** indicate confusing images containing the additional negated concept (3).

is one such case. While this limits the model’s tendency to reward the negated concept, it never verifies whether the negated concept is actually present in the image. We argue that this is *dilution*, not *rejection*: VLMs need not merely ignore the negated concept (dilution) but actively penalize its presence (rejection). This stems from a representational asymmetry: visual concepts manifest as presence or absence, never as anti-presence, leaving no dimension on which a single text vector can push back against the negated concept. We attribute this dilution to a structural limitation of contrastive training: under the in-kind alignment component of contrastive training [1, 2], the model learns to treat the negated concepts as irrelevant noise to be ignored, rather than as signals to be actively rejected. Prior data-centric retraining attempts yield only marginal gains in practice.

To escape this limitation, we propose ***Know-Thy-Not* (KTN)**, a training-free method that scores a query along two independent dimensions. A text-side dimension affirms c^+ via a corrected text embedding, while an image-side dimension detects c^- in the image and penalizes its presence. We further introduce COCO-Confusion, a benchmark whose galleries contain images with the negated concept, making it the first to measure rejection rather than dilution. Existing benchmarks place the negated concept in unfiltered galleries where it rarely co-occurs with the affirmed concept, so methods that merely ignore the negation already score well. COCO-Confusion enforces co-occurrence in every gallery, making genuine rejection (not dilution) necessary to succeed. Across 9 backbones spanning CLIP, EVA-02, SigLIP, and SigLIP2, KTN consistently improves negation handling, with the largest gains on COCO-Confusion (+9.6 mAP over Embedding Correction [18]), where prior methods plateau.

2 Preliminaries

Vision-language retrieval. We consider a dual-encoder vision-language model [1, 2, 19], consisting of a text encoder $f_T: \mathcal{T} \rightarrow \mathbb{R}^d$ and an image encoder $f_I: \mathcal{I} \rightarrow \mathbb{R}^d$, both producing ℓ_2 -normalized embeddings in a shared d -dimensional space. Given a text query $q \in \mathcal{T}$ and a gallery $\{i_j\}_{j=1}^M \subset \mathcal{I}$,

image retrieval ranks M candidate images by their inner product to the query

$$s(q, i) = f_{\mathcal{T}}(q)^{\top} f_{\mathcal{I}}(i). \quad (1)$$

Negation queries. We define a negation query as a compositional semantic constraint $q = (c^+, c^-)$, typically formulated as “ c^+ without c^- ” or equivalent exclusionary construction. Here, c^+ denotes the *affirmed* concept and c^- denotes the *negated* concept. To evaluate the faithfulness of the retrieval system, we define a binary indicator function $\phi(i, c) \in \{0, 1\}$ that denotes the presence of concept c in image i . For a given $q = (c^+, c^-)$, we partition the gallery $\mathcal{I}$ into three mutually exclusive sets:

- *Correct set* $\mathcal{I}^+$: Images satisfying the affirmed concept while strictly excluding the negated concept

$$\mathcal{I}^+ = \{i \in \mathcal{I} \mid \phi(i, c^+) = 1, \phi(i, c^-) = 0\}. \quad (2)$$

- *Confusing set* $\mathcal{I}^{\pm}$: Images containing the affirmed concept but failing the negation constraint

$$\mathcal{I}^{\pm} = \{i \in \mathcal{I} \mid \phi(i, c^+) = 1, \phi(i, c^-) = 1\}. \quad (3)$$

- *Irrelevant set* $\mathcal{I}^-$: Images that fail to satisfy the primary affirmed concept

$$\mathcal{I}^- = \{i \in \mathcal{I} \mid \phi(i, c^+) = 0\}. \quad (4)$$

In Figure 1, the affirmed concept c^+ is “car” and the negated concept c^- is “traffic light”. Faithful negation retrieval ranks $\mathcal{I}^+$ above $\mathcal{I}^{\pm}$, where $\mathcal{I}^{\pm}$ is the most challenging distractor.

3 Why single text embeddings cannot reject negation

Existing approaches only dilute the negated concept rather than reject it. We formalize this limitation by proving that contrastive training provides no gradient toward the negative coefficient that rejection requires. We confirm this limitation empirically on COCO-Confusion.

3.1 Existing approaches only dilute, not reject

Data-centric strategies typically augment training sets with negated captions and retrain the model to enhance its grasp of negation. However, these approaches often yield marginal improvements. Despite the substantial computational overhead of retraining, the model’s ability to distinguish between a caption and its negation remains far from satisfactory [15, 17].

To address this bottleneck, Aggarwal et al. [18] introduced a *training-free* embedding correction method that isolates the negated concept c^- from the query q and projects out its direction from the original query embedding

$$\tilde{\mathbf{t}} \propto f_{\mathcal{T}}(q) - \lambda \cdot \text{PROJECT}_{c^-} f_{\mathcal{T}}(q), \quad (5)$$

where λ is a scalar hyperparameter.

While this approach yields notable empirical gains, we observe that the resulting embedding $\tilde{\mathbf{t}} \in \mathbb{R}^d$ remains *agnostic* to the visual grounding of c^- . The subtraction operation modifies the text embedding in isolation, failing to verify the absence of c^- against its actual features grounded in the visual manifold. Consequently, such embedding correction merely *dilutes* the semantic influence of c^- without penalizing its material presence in the image (see Figure 2).

We posit that faithful negation handling requires an *explicit mechanism of rejection*—a capability that is structurally precluded by the conventional single embedding format. This observation forms the crux of our argument: the conventional single-embedding VLMs lack the expressive bandwidth to encode exclusionary logic within the same representation used for semantic matching.

3.2 A structural barrier in single-embedding negation handling

Under standard ℓ_2 -normalized contrastive training, the contrastive objective provides no learning signal toward the negative coefficient that rejection requires. Moreover, query-side post-hoc correction is also bounded by dilution. We formalize both limits below.

Setup. For each concept c , let $\mathbf{e}_c^* \in \mathbb{R}^d$ denote a unit-norm proxy vector, ideally representing the concept c in the space shared by both encoders. We define

$$\alpha_c(i) := f_{\mathcal{I}}(i)^{\top} \mathbf{e}_c^* \quad (\text{image activation}), \quad \beta_c(q) := f_{\mathcal{T}}(q)^{\top} \mathbf{e}_c^* \quad (\text{text coefficient}). \quad (6)$$

For analytical tractability we adopt a linear concept model: $f_{\mathcal{I}}(i) = \sum_c \alpha_c(i) \mathbf{e}_c^*$ on an orthonormal concept basis. We focus on negation queries $q = “c^+ \text{ without } c^-”$.

Query: “a photo of a **cat** without a **dog**”



Figure 2: Given the query “A photo of a **cat** without a **dog**”, a vanilla VLM ignores the negation and activates on both concepts (left). Embedding correction approach [18] removes the dog direction from the text embedding but never verifies the visual features, which amounts to passively diluting the concept (center). Faithful negation requires explicitly locating the dog in the image and penalizing its presence (right), which forms the crux of our method *Know-Thy-Not* (KTN).

Assumption 1 (Concept Cone, idealized). *In the idealized model, concept activations are non-negative: $\alpha_c(i) \geq 0$, with $\alpha_c(i) > 0$ iff $c \in i$. This captures the intuition that semantic similarity measures presence, not anti-presence.*

Lemma 1 (Rejection criterion). *For $q = “c^+ \text{ without } c^-”$, consider images $i^+ \in \mathcal{I}^+$ and $i^\pm \in \mathcal{I}^\pm$ that differ only along the c^- axis. Under the linear concept model,*

$$s(i^+, q) - s(i^\pm, q) = -\alpha_{c^-}(i^\pm) \beta_{c^-}(q). \quad (7)$$

Faithful retrieval demands $s(i^+, q) > s(i^\pm, q)$, and since $\alpha_{c^-}(i^\pm) > 0$, this is equivalent to

$$\boxed{\beta_{c^-}(q) < 0.} \quad (8)$$

Theorem 1 (In-kind alignment carries no gradient toward $\beta_{c^-}(q) < 0$). *Consider the positive-alignment objective*

$$\mathcal{A}(q) = f_{\mathcal{T}}(q)^\top \mu(q), \quad \mu(q) := \mathbb{E}_{i \in \mathcal{I}_q} [f_{\mathcal{I}}(i)], \quad (9)$$

the positive-pair component of standard contrastive losses. For $q = “c^+ \text{ without } c^-”$, every $i \in \mathcal{I}_q$ satisfies $c^- \notin i$, so $\mu_{c^-}(q) = 0$, and decomposing $\mathcal{A}(q) = \sum_k \beta_k(q) \mu_k(q)$ gives

$$\boxed{\frac{\partial \mathcal{A}(q)}{\partial \beta_{c^-}(q)} \equiv 0.} \quad (10)$$

The objective is invariant in $\beta_{c^-}(q)$: positive-alignment training provides no gradient that could drive $\beta_{c^-}(q)$ below zero.

Corollary 1 (Optimal text coefficient under unit-norm constraint). *Under the unit-norm constraint $\|f_{\mathcal{T}}(q)\| = 1$ standard for ℓ_2 -normalized encoders, the maximizer of $\mathcal{A}(q) = \beta(q)^\top \mu(q)$ is, by Cauchy-Schwarz, $\beta^*(q) = \mu(q)/\|\mu(q)\|$. Since $\mu_{c^-}(q) = 0$ for negation queries, $\beta_{c^-}^*(q) = 0$. Combined with Lemma 1, this implies $s(i^+, q) = s(i^\pm, q)$ at the optimum: even an optimally-trained encoder under in-kind alignment cannot rank correct above confusing. (Details in Appendix A.)*

Beyond in-kind alignment. Hard-negative supervision lies outside the positive-alignment regime but, under token frequency asymmetry, collapses to a generic negation offset rather than concept-specific rejection (Appendix A). Post-hoc embedding correction still only reaches $\tilde{\beta}_{c^-}(q) = 0$ and fails to overcome the Lemma 1 barrier. $\tilde{\beta}_{c^-}(q) = 0$ is the ceiling of all existing approaches, training-based or post-hoc, yet remains insufficient for rejection.

3.3 The limit of dilution-based approaches

We demonstrate these limitations with an example query on COCO-Confusion (Figure 3).

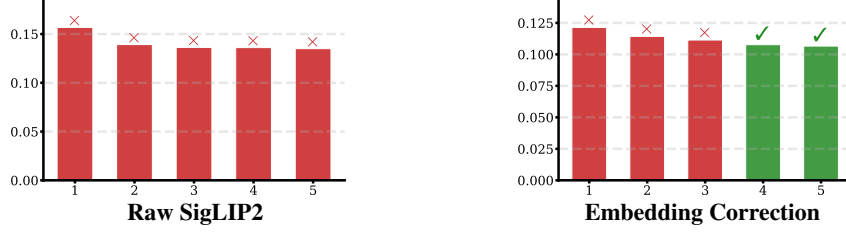


Figure 3: **Dilution cannot separate correct from confusing images.** Top-5 similarity scores for “a car is present but no traffic light” on COCO-Confusion, SigLIP2 SO400M. **Raw SigLIP2** (left): all top-5 are confusing (red) with uniform scores (≈ 0.15). **Embedding Correction** (right): scores drop overall (≈ 0.12), but red (confusing) and green (correct) bars remain interleaved, failing to separate correct from confusing.

Vanilla VLM. Vanilla VLM does not process the negation word “no,” so the query is effectively treated as “a car is present and traffic light.” As a result, images containing both car and traffic light receive the highest similarity score—all top-5 results are confusing images (red bar) with nearly uniform scores (≈ 0.15).

Embedding correction. Embedding Correction dilutes the traffic light concept by removing its direction from the text embedding. As a result, the overall scores of confusing images decrease (≈ 0.12). However, since Embedding Correction *ignores* traffic light rather than *rejecting* it, images containing both car and traffic light still receive high similarity scores. The score bars remain nearly uniform across confusing (red) and correct (green) images, failing to separate the two groups.

This confirms the dilution ceiling: single-embedding scoring cannot reject c^- . **Negation needs multi-dimensional scoring.** One dimension to affirm c^+ , another to penalize the image-side activation of c^- . KTN realizes this decomposition.

4 Proposed method: *Know-Thy-Not*

Section 3 established that single-embedding scoring cannot achieve rejection: contrastive training provides no gradient toward $\beta_{c^-}(q) < 0$ (Theorem 1), and post-hoc correction only reaches $\hat{\beta}_{c^-}(q) = 0$, still short of the rejection condition required by Lemma 1. All single-embedding approaches share this limitation.

We propose ***Know-Thy-Not* (KTN)**, a training-free method that extends scoring from one dimension to two: (i) a *text-side affirmation dimension* that handles c^+ via a corrected text embedding $\tilde{t}$, and (ii) an *image-side rejection dimension* that handles c^- by detecting and penalizing its visual presence in i . The full scoring is

$$s(q, i) = \underbrace{\tilde{t}^\top f_{\mathcal{I}}(i)}_{\text{Dim 1: Affirm } c^+} - w_n \cdot \underbrace{p_{c^-}(i)}_{\text{Dim 2: Reject } c^-}, \quad (11)$$

$p_{c^-}(i)$ is the relative grounding score defined in Eq. (13), and w_n is a scaling parameter. Dimension 1 operates on the text side and is constrained by Theorem 1—by itself, it can only achieve dilution. Dimension 2 operates on the image side using a separate computation that does not pass through $f_{\mathcal{T}}(\cdot)$, hence is not bound by Theorem 1’s unit-norm budget. The two dimensions share no parameter budget, allowing $p_{c^-}(i)$ to provide the rejection signal that Lemma 1 requires and Corollary 1 forbids in any single-dimension formulation.

4.1 Text-side affirmation

Given a natural-language query q such as “a cat without a dog”, we first extract the negated concept c^- and the affirmed concepts $\mathcal{C}^+$ via spaCy’s dependency parser [20]. Negation cues—*no*, *not*, *without*, *absent*, *missing*, *free of*—are identified by dependency relation (neg) or lemma matching, and the head noun of the negation span yields c^- ; remaining noun phrases form $\mathcal{C}^+$.

We then compute the corrected text embedding by projecting out the c^- direction:

$$\tilde{t} = \text{NORMALIZE}(f_{\mathcal{T}}(q) - \text{PROJECT}_{c^-} f_{\mathcal{T}}(q)) \quad (12)$$

where $\text{proj}_{c^-}(\cdot)$ denotes orthogonal projection onto the c^- direction, making $\tilde{t}$ orthogonal to e_{c^-} . This achieves $\beta_{c^-}(\tilde{t}) = 0$, matching the constrained optimum of Corollary 1 — yet by Lemma 1 still insufficient for rejection.

4.2 Image-side rejection

The second scoring dimension produces an image-conditioned penalty that grows when c^- is grounded in image i and vanishes when it is not. A naive thresholding on $f_{\mathcal{I}}(i)^\top e_{c^-}$ fails because raw similarity scales vary substantially across images: some images yield globally high similarity to all concepts, producing false positives.

We then compute a relative detection score $p_{c^-}(i)$ using a per-image reference $\tau(i)$

$$p_{c^-}(i) = \sigma(\kappa(f_{\mathcal{I}}(i)^\top e_{c^-} - \tau(i))) - \frac{1}{2}\sigma(\kappa(f_{\mathcal{I}}(i)^\top e_{c^+} - \tau(i))), \quad (13)$$

$$\tau(i) = \frac{1}{|\mathcal{C}^+|} \sum_{c \in \mathcal{C}^+} f_{\mathcal{I}}(i)^\top e_c. \quad (14)$$

where $\sigma(\cdot)$ is the sigmoid acting as a soft thresholding operator — penalizing c^- only when $\alpha_{c^-}(i)$ exceeds $\tau(i)$ — and κ controls the sharpness of the transition. Real VLMs deviate from the ideal encoder of Assumption 1: $\alpha_{c^-}(i)$ remains non-zero even when c^- is absent. We therefore anchor the threshold to the image’s own affirmed-concept similarity, $\tau(i) = \frac{1}{|\mathcal{C}^+|} \sum_{c \in \mathcal{C}^+} \alpha_c(i)$, providing a query-relevant, image-specific baseline. When c^- ’s similarity exceeds this baseline, the first term drives $p_{c^-}(i) \rightarrow 1$, and otherwise it falls to 0. The second term further suppresses false positives that arise when c^+ and c^- are semantically correlated (e.g., *cat* and *dog*). In such cases $\alpha_{c^-}(i)$ may exceed $\tau(i)$ even when c^- is absent, and discounting by the c^+ -side activation prevents correct images from spurious penalty. Crucially, $p_{c^-}(i)$ is computed entirely from $f_{\mathcal{I}}(i)$ and concept directions—it does **not** pass through the text-query embedding $f_{\mathcal{T}}(q)$, which is what allows the rejection dimension to penalize c^- regardless of the alignment-induced ceiling on $\beta_{c^-}^*$.

4.2.1 Why KTN avoids the constraint

Corollary 1 bounds the contribution of any single text-side coordinate: under the unit-norm constraint, $\beta_{c^-}^* = 0$ exactly. Dimension 1 respects this bound— $\tilde{t}^\top f_{\mathcal{I}}(i)$ is a single-vector dot product and inherits the $\beta_{c^-}^* = 0$ ceiling.

Dimension 2 sidesteps the bound by introducing a separate scoring axis computed outside the text-side unit-sphere budget. The total score is no longer a single dot product; it is a two-dimensional composition where each dimension operates in its own budget. The matched-pair test of Lemma 1—which requires a negative ranking signal—is satisfied by Dimension 2.

The asymmetry of natural visual data—*presence learns; absence does not*—is resolved by decomposing the problem into two dimensions: Dimension 1 handles affirmation, where the data measure naturally provides positive signal ($\mu_{c^+} > 0$); Dimension 2 handles rejection, where the data measure provides no signal ($\mu_{c^-} = 0$) but direct image-side detection circumvents this gap entirely.

5 Experiments

5.1 Introducing COCO-Confusion: A benchmark for negation rejection

Existing negation retrieval benchmarks such as NegBench evaluate retrieval over a large, unfiltered gallery (e.g., 5,000 COCO [21] images), where the negated concept rarely co-occurs with the affirmed concept in the same image. In such settings, even simple embedding correction methods perform well because there are few confusing images to distinguish from correct ones. This means existing benchmarks cannot distinguish between *ignoring* and *rejecting* the negated concept. To expose this critical gap, we construct COCO-Confusion, a retrieval benchmark specifically designed to test whether a method can reject confusing images that contain the negated concept.

Construction. We build COCO-Confusion from COCO val2017 using bounding box annotations. For each concept pair (c^+, c^-) , we partition the annotated images into two groups: *correct* images containing c^+ but not c^- , and *confusing* images containing both c^+ and c^- . We consider two settings based on query complexity: COCO-Confusion-2C uses one affirmed and one negated concept (e.g., “A car is present but no traffic light”; 950 pairs), and COCO-Confusion-3C uses two affirmed concepts

Table 1: Negation performance on CLIP ViT-B/32 [2]. We compare against training-based negation methods [17, 27] and training-free embedding correction method (EC) [18] across MCQ, Retrieval, and our COCO-Confusion benchmark. (*) indicates models fine-tuned on CC12M-NegFull [15]. Best in **bold**; second-best underlined.

	MCQ			Retrieval					
	COCO	VOC2007	MSRvtt	COCO		MSRvtt		Conf-2C	Conf-3C
Model				R@5	R-Neg@5	R@5	R-Neg@5	mAP	mAP
CLIP [2]	24.4	27.2	30.1	<u>65.3</u>	57.9	<u>55.7</u>	48.8	63.2	60.5
NegCLIP [27]	28.5	30.4	28.4	69.5	65.2	57.0	51.6	64.4	61.4
CLIP* (NegBench) [15]	54.4	54.9	45.3	54.9	52.3	48.8	43.1	68.8	65.0
NegationCLIP [17]	35.9	44.2	37.4	62.8	<u>62.0</u>	52.5	49.7	72.9	68.4
EC [18]	70.0	75.5	51.6	65.2	58.3	<u>55.7</u>	51.0	<u>73.2</u>	<u>70.8</u>
KTN (Ours)	79.5	92.1	66.9	65.2	<u>62.0</u>	<u>55.7</u>	51.9	82.8	75.2

and one negated concept (e.g., “An apple and banana are present but no bowl”; 1,740 pairs). Each group is capped at 50 images per pair to ensure balanced galleries. We verify all images using OWLv2 [22], an open-vocabulary object detector, to filter annotation noise where the negated concept is heavily occluded, truncated, or visually ambiguous in the bounding-box annotations, removing pairs where the detector cannot confirm concept presence. (details in Appendix D) The final benchmark contains 2,690 verified pairs across 4,708 unique images.

Evaluation metric. For each pair, the gallery consists of its correct and confusing images. We report mean Average Precision (mAP), which measures how well a method ranks correct images above confusing ones. High mAP indicates successful rejection of confusing images beyond dilution.

Why COCO-Confusion matters. On existing retrieval benchmarks, Embedding Correction method achieves near-optimal performance because confusing images are rare in the gallery. On COCO-Confusion, every gallery is designed to contain confusing images, making it the first benchmark that explicitly measures negation *rejection* rather than *dilution*.

5.2 Experimental setup

Backbones. We evaluate on 9 pretrained VLMs spanning three model families and diverse scales: CLIP (ViT-B/32, L/14, H/14 [23]; LAION-2B [24]), EVA-02 [25] (ViT-B/16, L/14; Merged-2B), and SigLIP/SigLIP2 (ViT-B/16, B/32, L/16, SO400M; WebLI with sigmoid loss). All weights are public and frozen.

Evaluation tasks. We evaluate on three complementary tasks. (1) **MCQ:** Given an image and four caption options, select the correct one. We report accuracy on COCO MCQ and VOC2007 MCQ from NegBench. This measures negation *comprehension*. (2) **Standard negation retrieval:** Given a negated query, retrieve matching images from a large gallery. We report Recall@5 on COCO and MSRvtt [26] negated retrieval from NegBench. This measures negation handling on natural galleries where confusing images are rare. (3) **COCO-Confusion retrieval:** Given a negated query, retrieve correct images from a gallery containing both correct and confusing images. We report mAP on COCO-Confusion-2C and COCO-Confusion-3C setting. This measures negation *rejection*. Qualitative retrieval examples are provided in Appendix.

Baselines. We compare against Raw VLM, which uses standard cosine similarity with no negation processing, and Embedding Correction, a training-free method that removes the negated concept direction from the text embedding. Since Embedding Correction reports results only for CLIP ViT-B/32 and does not release code, we use their reported numbers for this backbone. For CLIP ViT-B/32, we additionally compare against NegCLIP and NegationCLIP.

5.3 Main results

MCQ. Our method substantially improves negation comprehension across all MCQ benchmarks. On CLIP ViT-B/32 (Table 1), we achieve 79.5% on COCO, 92.1% on VOC2007, and 66.9% on MSRvtt, improving over Embedding Correction by +9.5, +16.6, and +15.3 points respectively. Both training-based baselines fall far behind: NegCLIP and NegationCLIP show only marginal gains over the raw baseline despite requiring retraining. These improvements generalize consistently across all

Table 2: Negation performance across multiple VLM backbones [1, 2, 19, 25]. Each backbone shows three rows: Vanilla VLM, embedding correction (EC) [18], and our KTN. **Bold** indicates best-per-metric across the three methods. KTN consistently outperforms both vanilla and embedding correction on MCQ comprehension and Confusion retrieval, while preserving standard retrieval R@5.

Model	Backbone	MCQ			Retrieval					
		COCO	VOC2007	MSRvtt	COCO		MSRvtt		Conf-2C	Conf-3C
		R@5	R-Neg@5	R@5	R@5	R-Neg@5	R@5	R-Neg@5	mAP	mAP
Vanilla	CLIP [2]	25.4	25.3	30.9	70.8	64.1	60.4	55.0	62.1	59.4
EC [18]	ViT-L/14	48.4	45.3	41.9	70.8	66.6	60.4	55.9	73.8	67.4
KTN (Ours)		79.6	92.4	69.1	70.8	67.8	60.5	56.5	82.4	74.8
Vanilla	CLIP [2]	30.2	31.9	29.9	73.2	66.8	62.6	55.0	62.0	59.3
EC [18]	ViT-H/14	53.9	49.9	43.3	73.2	68.5	62.6	57.8	72.2	66.1
KTN (Ours)		79.4	92.2	68.1	73.2	69.9	62.6	59.2	81.6	74.0
Vanilla	EVA-02 [25]	30.2	27.8	29.2	66.6	61.2	58.7	51.1	62.3	58.7
EC [18]	ViT-B/16	44.3	40.0	43.5	66.6	63.5	58.7	53.8	76.1	68.5
KTN (Ours)		77.1	91.2	63.8	66.6	64.4	58.7	53.5	84.4	76.2
Vanilla	EVA-02 [25]	31.5	31.9	30.4	70.9	65.5	62.8	56.9	61.1	57.9
EC [18]	ViT-L/14	38.3	25.4	39.7	70.9	67.7	62.8	58.2	77.4	68.1
KTN (Ours)		79.7	92.4	65.9	70.9	68.5	62.8	58.4	84.6	76.2
Vanilla	SigLIP [1]	28.8	31.1	32.2	72.3	65.0	56.1	49.7	62.3	60.2
EC [18]	ViT-B/16	56.3	77.1	51.6	72.3	68.6	56.1	50.5	79.2	72.2
KTN (Ours)		71.1	82.9	59.2	72.3	69.6	56.1	51.5	85.5	78.1
Vanilla	SigLIP2 [19]	27.1	27.3	28.2	73.2	64.8	56.9	51.3	61.8	60.2
EC [18]	ViT-B/32	63.3	78.2	51.2	73.2	68.4	56.9	49.2	81.4	74.4
KTN (Ours)		78.5	93.4	61.4	73.2	69.5	56.9	50.5	86.7	79.2
Vanilla	SigLIP2 [19]	25.2	24.1	28.8	77.1	71.0	64.1	56.8	61.1	59.1
EC [18]	ViT-L/16	57.7	71.8	53.9	77.1	74.0	64.1	58.0	78.6	71.4
KTN (Ours)		80.5	94.6	64.6	77.1	75.3	64.1	58.4	87.0	79.6
Vanilla	SigLIP2 [19]	23.9	23.8	28.9	78.0	71.5	64.9	58.0	61.5	59.8
EC [18]	So400m	55.1	71.1	54.0	78.0	74.6	64.9	59.5	78.6	69.8
KTN (Ours)		81.0	94.6	65.7	77.9	75.9	64.9	59.6	87.1	79.3

backbones (Table 2): MCQ accuracy improves by +42 to +57 points on COCO, +51 to +71 points on VOC2007, and +31 to +37 points on MSRvtt relative to the raw baseline. The gains are consistent regardless of model family (CLIP, EVA-02, SigLIP, SigLIP2) or scale (ViT-B to SO400M).

Retrieval. On standard negation retrieval, where confusing images are rare in the gallery, our method preserves R@5 and consistently improves R-Neg@5. On CLIP ViT-B/32, R@5 remains at 65.2 while R-Neg@5 improves from 58.3 to 62.0 (+3.7). Across all backbones, R-Neg@5 gains range from +1.5 to +4.7 on COCO and +1.5 to +4.2 on MSRvtt, with no degradation in R@5.

COCO-Confusion. The largest gains appear on COCO-Confusion, the benchmark designed to test negation rejection rather than dilution. On CLIP ViT-B/32, our method achieves 82.8% mAP on Conf-2C and 75.2% on Conf-3C, improving over Embedding Correction by +9.6 and +4.4 points. Across all backbones, the pattern is striking: raw baselines cluster around 61–65% mAP regardless of model scale, confirming that larger models do not resolve negation failure. Embedding Correction provides only modest improvement. Our method consistently reaches 81–87% on Conf-2C and 74–79% on Conf-3C, with the best result of 87.1% mAP on SigLIP2 SO400M. This confirms that the image-side rejection dimension successfully separates correct from confusing images, precisely the gap that single-embedding scoring cannot close.

5.4 Ablation study

Table 3 isolates the contribution of each scoring dimension on SigLIP2 SO400M. Dimension 1 alone (Proj) already yields a large gain over Raw on MCQ (+31.2pp) and COCO-Confusion (+17.1pp), confirming that dilution is effective when the task requires distinguishing affirmative from negated captions. However, adding Dimension 2 KTN produces a further +25.9pp on MCQ and +8.5pp on COCO-Confusion, gains that are inaccessible to any single-embedding method by Theorem 1. On

Table 3: **Ablation study** (SigLIP2 SO400M). Dimension 1 alone (Proj) achieves dilution. Dimension 2 provides rejection.

Method	COCO MCQ	R-Neg@5	Conf-2C
SigLIP2	23.9	71.5	61.5
+ Proj (Dim 1)	55.1	74.6	78.6
+ Ours (Dim 1+2)	81.0	75.9	87.1

standard negation retrieval (R-Neg@5), the gap is small (+1.3pp), consistent with the observation that confusing images are rare in unfiltered galleries and dilution alone suffices. The ablation confirms that the dominant source of improvement is the image-side rejection dimension, precisely the signal that the dilution ceiling forbids.

6 Related work

Negation failure in VLMs. A growing body of work documents that VLMs systematically fail on negation. The ARO benchmark [27] showed that CLIP behaves as a bag-of-words and largely ignores word order and negation. NegBench [15] evaluated multiple VLMs across 18 negation tasks and found near-chance performance, with CLIP scoring 82% on affirmative queries but only 3% on negated ones for VOC2007. NegationCLIP [17] traced this failure to a training data shortfall and reported that only 0.7% of LAION-400M captions contain any negation word. DCSM [28] provides a theoretical account and argues that no CLIP-like joint embedding space can support both basic content representation and negation simultaneously.

Approaches for improving negations awareness. Data-centric methods augment training data with negation captions and retrain the model. ConCLIP [16] constructed 228K negation triples and fine-tuned with negated-caption distractors. NegationCLIP [17] generated visually grounded negation captions through LLM and MLLM pipelines. However, these approaches require retraining and show limited generalization. On the training-free side, the embedding correction method [18] showed that CLIP text embeddings follow compositional arithmetic and proposed a rule-based projection that extracts negated concepts and removes their semantic direction from the embedding. This is effective at reducing negation bias but limited to dilution rather than rejection. We go beyond this limitation by adding a separate dimension that actively penalizes the negated concept’s visual presence.

7 Conclusion

Representing negation within a single embedding vector is fundamentally limited. Even with perfect training data and infinite capacity, the non-negative nature of visual concept activations drives the text coefficient on negated concepts to zero, making dilution the theoretical ceiling of single-embedding scoring. We proposed *Know-Thy-Not* (KTN), a training-free multi-dimensional scoring method that breaks this ceiling by separating affirmation and rejection into independent dimensions. Combined with COCO-Confusion, a benchmark designed to measure rejection beyond dilution, experiments across nine backbones confirm consistent improvements in MCQ, negation retrieval, and confusion rejection. Beyond performance gains, COCO-Confusion makes the distinction between dilution and rejection measurable for the first time, revealing a gap that prior benchmarks could not expose.

Limitations

KTN’s image-side rejection relies on cosine similarity, which can yield high scores for c^- even when it is absent from the image. We mitigate this with adaptive thresholding using a per-image reference $\tau(i)$ (Eq. 14), but it does not fully eliminate false positives.

Our method targets dual-encoder VLMs (CLIP, SigLIP) that score via a single dot product. Extending to generative VLMs like LLaVA [29, 30] is a natural next step. Their scoring is conditioned on autoregressive decoding rather than embedding similarity, so the text coefficient β_{c^-} analyzed in our impossibility result does not directly apply. Still, the broader principle that rejection requires a scoring signal beyond a single embedding axis likely transfers, and we expect LLaVA-style architectures to also benefit from a multi-dimensional formulation that rejects the negated concept rather than merely diluting it.

References

- [1] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, “Sigmoid loss for language image pre-training,” in *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, 2023, pp. 11 975–11 986.
- [2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *International conference on machine learning (ICML)*, 2021, pp. 8748–8763.
- [3] M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, and J. Jitsev, “Reproducible scaling laws for contrastive language-image learning,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2023, pp. 2818–2829.
- [4] X. Zhai, X. Wang, B. Mustafa, A. Steiner, D. Keysers, A. Kolesnikov, and L. Beyer, “Lit: Zero-shot transfer with locked-image text tuning,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2022, pp. 18 123–18 133.
- [5] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, “Scaling up visual and vision-language representation learning with noisy text supervision,” in *International conference on machine learning (ICML)*, 2021, pp. 4904–4916.
- [6] J. Li, D. Li, C. Xiong, and S. Hoi, “Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation,” in *International conference on machine learning (ICML)*, 2022, pp. 12 888–12 900.
- [7] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2022, pp. 10 684–10 695.
- [8] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, “Hierarchical text-conditional image generation with clip latents,” *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.
- [9] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, “Photorealistic text-to-image diffusion models with deep language understanding,” *Advances in neural information processing systems (NeurIPS)*, vol. 35, pp. 36 479–36 494, 2022.
- [10] T. Thrush, R. Jiang, M. Bartolo, A. Singh, A. Williams, D. Kiela, and C. 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)*, 2022, pp. 5238–5248.
- [11] A. Kamath, J. Hessel, and K.-W. Chang, “What’s “up” with vision-language models? investigating their struggle with spatial reasoning,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023, pp. 9161–9175.
- [12] Z. Ma, J. Hong, M. O. Gul, M. Gandhi, I. Gao, and R. Krishna, “Crepe: Can vision-language foundation models reason compositionally?” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 10 910–10 921.
- [13] C.-Y. Hsieh, J. Zhang, Z. Ma, A. Kembhavi, and R. Krishna, “Sugarcrape: Fixing hackable benchmarks for vision-language compositionality,” *Advances in neural information processing systems (NeurIPS)*, vol. 36, pp. 31 096–31 116, 2023.
- [14] L. Parcalabescu, M. Cafagna, L. Muradjan, A. Frank, I. Calixto, and A. Gatt, “Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, vol. 1, 2022, pp. 8253–8280.
- [15] K. Alhamoud, S. Alshammari, Y. Tian, G. Li, P. H. Torr, Y. Kim, and M. Ghassemi, “Vision-language models do not understand negation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 29 612–29 622.

- [16] J. Singh, I. Shrivastava, M. Vatsa, R. Singh, and A. Bharati, “Learn no to say yes better: Improving vision-language models via negations,” *arXiv preprint arXiv:2403.20312*, 2024.
- [17] J. Park, J. Lee, J. Song, S. Yu, D. Jung, and S. 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)*, 2025, pp. 2825–2835.
- [18] B. Aggarwal, A. More, M. Soni, and S. D. Bhat, “Seeing what’s not there: Negation understanding needs more than training,” in *International conference on learning representations (ICLR)*, 2026.
- [19] M. Tschannen, A. Gritsenko, X. Wang, M. F. Naeem, I. Alabdulmohsin, N. Parthasarathy, T. Evans, L. Beyer, Y. Xia, B. Mustafa *et al.*, “Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features,” *arXiv preprint arXiv:2502.14786*, 2025.
- [20] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, “spaCy: Industrial-strength Natural Language Processing in Python,” 2020.
- [21] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *European conference on computer vision (ECCV)*. Springer, 2014, pp. 740–755.
- [22] M. Minderer, A. Gritsenko, and N. Houlsby, “Scaling open-vocabulary object detection,” *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, pp. 72 983–73 007, 2023.
- [23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” *International conference on learning representations (ICLR)*, 2021.
- [24] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman *et al.*, “Laion-5b: An open large-scale dataset for training next generation image-text models,” *Advances in neural information processing systems (NeurIPS)*, vol. 35, pp. 25 278–25 294, 2022.
- [25] Y. Fang, Q. Sun, X. Wang, T. Huang, X. Wang, and Y. Cao, “Eva-02: A visual representation for neon genesis,” *Image and Vision Computing*, vol. 149, p. 105171, 2024.
- [26] J. Xu, T. Mei, T. Yao, and Y. Rui, “Msr-vtt: A large video description dataset for bridging video and language,” in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2016, pp. 5288–5296.
- [27] M. Yuksekgonul, F. Bianchi, P. Kalluri, D. Jurafsky, and J. Zou, “When and why vision-language models behave like bags-of-words, and what to do about it?” *International conference on learning representations (ICLR)*, 2023.
- [28] R. Kang, Y. Song, G. Gkioxari, and P. Perona, “Is CLIP ideal? no. can we fix it? yes!” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025, pp. 22 436–22 446.
- [29] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual instruction tuning,” *Advances in neural information processing systems (NeurIPS)*, vol. 36, pp. 34 892–34 916, 2023.
- [30] H. Liu, C. Li, Y. Li, and Y. J. Lee, “Improved baselines with visual instruction tuning,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2024, pp. 26 296–26 306.

A Theoretical details

A.1 Per-image budget compression

For each image i , let $\mathcal{P}(i) := \{c : \alpha_c(i) > 0\}$ and $\mathcal{N}(i) := \{c : \alpha_c(i) = 0\}$, with $K(i) := |\mathcal{P}(i)| \ll |\mathcal{C}|$. That is, $\mathcal{P}(i)$ is the set of concepts visually present in image i , $\mathcal{N}(i)$ is the set of concepts absent from it, and $K(i)$ is the number of present concepts, which is typically much smaller than the total concept vocabulary $|\mathcal{C}|$.

Recall that for each concept c , we define the text coefficient $\beta_c(q) := f_{\mathcal{T}}(q)^\top e_c$, which measures how much the text embedding of query q activates the direction of concept c . Suppose the text query assigns uniform affirmed-concept margin $\gamma_+ > 0$ across present concepts and uniform negated-concept margin $\gamma_- > 0$ across absent concepts:

$$\beta_{c^+}(q) \geq \gamma_+ \quad \forall c \in \mathcal{P}(i), \quad \beta_{c^-}(q) \leq -\gamma_- \quad \forall c \in \mathcal{N}(i). \quad (15)$$

The unit-norm constraint $\sum_c \beta_c(q)^2 \leq 1$ then implies

$$K(i)\gamma_+^2 + (|\mathcal{C}| - K(i))\gamma_-^2 \leq 1, \quad (16)$$

and consequently

$$|\gamma_-| \leq \sqrt{\frac{1 - K(i)\gamma_+^2}{|\mathcal{C}| - K(i)}} = \mathcal{O}\left(\frac{1}{\sqrt{|\mathcal{C}|}}\right). \quad (17)$$

Equation (17) complements Corollary 1: the conclusion $\beta_{c^-}^*(q) = 0$ arises both from optimization (the unit-norm maximizer of in-kind alignment) and from representation geometry (no unit vector can assign uniform negated concept margin across the dense set of absent concepts). Per-concept negated concept margin is compressed at rate $\mathcal{O}(1/\sqrt{|\mathcal{C}|})$ regardless of the training objective.

A.2 Hard-negative supervision and frequency collapse

We approximate the text encoder as token-additive: $f_{\mathcal{T}}(q) \approx \sum_{t \in \text{tokens}(q)} W_t$. For negation queries $q = "c^+ \text{ without } c^-"$ sampled uniformly over $(c^+, c^-) \in \mathcal{C}^2$, the per-step participation rates are

- “without” cue token: rate 1 (appears in every negation example);
- concept token c : rate $1/|\mathcal{C}|$ in the negated position, dominantly trained by affirmation contexts elsewhere.

The learned embedding is therefore biased toward

$$f_{\mathcal{T}}("c^+ \text{ without } c^-") \approx \underbrace{e_{c^+}}_{\text{affirm}} + \underbrace{n_{\text{generic}}}_{\text{query-indep. offset}} + \underbrace{\mathcal{O}(1/|\mathcal{C}|)}_{\text{concept-specific}}. \quad (18)$$

Generic offset is not concept-specific rejection. Although n_{generic} may yield $\beta_{c^-}(q) < 0$, it fails to distinguish the negated concept c^- from any unmentioned-but-allowed concept r . Faithful retrieval requires $\beta_{c^-}(q) < \beta_r(q)$ for all such r , but a query-independent offset penalizes c^- and r identically.

This is empirically confirmed by NegBench’s PCA visualization of ConCLIP, where ‘not(c)’ embeddings collapse onto a single point regardless of c . [15] Consequently, the conventional dual-encoder VLMs exhibit global conservatism, diluting the negated concepts rather than specifically rejecting them.

B Implementation details

B.1 Hyperparameter settings

KTN uses one hyperparameters, all fixed across all backbones and task families. We do not tune any hyperparameter per backbone or per dataset.

Table 4: KTN hyperparameters. All values fixed across backbones and tasks. N denotes the candidate-pool size for the task ($N=4$ for MCQ, $N=|\mathcal{G}|$ for retrieval).

Symbol	Value	Description
κ	50	Sigmoid sharpness in $p_{c-}(i)$
w_n	$1/\sqrt{N}$	Rejection weight, scaled by candidate-pool size

B.2 Concept embedding

Each concept c in the vocabulary $\mathcal{C}$ is encoded as the L2-normalized average of four prompt templates:

$$\mathbf{e}_c = \text{NORMALIZE} \left(\frac{1}{4} \sum_{j=1}^4 \text{normalize}(f_{\mathcal{T}}(T_j(c))) \right), \quad (19)$$

with templates $T_1(c)$ ="a photo of a {c}", $T_2(c)$ ="a picture of a {c}", $T_3(c)$ ="an image showing a {c}", and $T_4(c)$ ="a {c} in a photo".

B.3 Negation parsing

We use spaCy’s `en_core_web_sm` for dependency parsing. A token is marked as a *negation cue* if its dependency arc is `neg` or its lemma is in the cue set:

NEG_CUES = {"no", "not", "without", "nothing", "nowhere",
"none", "never", "absent", "missing", "free of"}

The head noun of the dependency subtree rooted at the cue is taken as the negated concept c^- . Remaining noun phrases attached to the sentence root populate the affirmed concept set $\mathcal{C}^+$. The negation-stripped text e_{strip} used in $\tilde{t}$ ’s multi-view averaging is constructed by removing all tokens in the negation subtree and re-joining with normalized whitespace and trailing punctuation.

Example. For the query “A car is present but no traffic light”:

- Negation cue: `no` $\Rightarrow$ subtree covers “no traffic light”
- c^- = “traffic light”
- $\mathcal{C}^+$ = {“car”}
- e_{strip} encodes “A car is present”

B.4 Image and video preprocessing

We follow the standard CLIP preprocessing pipeline (resize, center-crop, normalize) at the backbone-specific input resolution (224 for B/32, 336 for L/14, 384 for SigLIP2 SO400M, etc.). For each image x , we encode both the original and its horizontal mirror, and use the L2-normalized mean of the two feature vectors:

$$f_{\mathcal{I}}(x) = \text{NORMALIZE} \left(\frac{1}{2} (f_{\mathcal{I}}^{\text{enc}}(x) + f_{\mathcal{I}}^{\text{enc}}(\text{FLIP}(x))) \right). \quad (20)$$

For MSR-VTT video retrieval, we extract four evenly-spaced frames per video (at quartile time positions) using `decord.VideoReader`. Each frame is encoded, and we use the L2-normalized mean of the four normalized frame features:

$$f_{\mathcal{I}}(v) = \text{NORMALIZE} \left(\frac{1}{4} \sum_{k=1}^4 \text{NORMALIZE} (f_{\mathcal{I}}^{\text{enc}}(v_k)) \right). \quad (21)$$

C Embedding correction reimplementation

Aggarwal et al. [18] report results only for CLIP ViT-B/32 and do not release code. We re-implement their method following the procedure described in their paper: the negated concept direction is projected out from the query embedding with $\lambda = 1.9$, anchored to a neutral text embedding as described in their work. Table 2 compares Embedding Correction and our method across all backbones.

D COCO-Confusion: Construction details

This appendix details the construction of COCO-Confusion, our benchmark designed to evaluate whether a VLM can *reject* a negated concept rather than merely dilute it. While Section 5 introduces the benchmark, here we provide the full pipeline, statistics, and qualitative examples.

D.1 Motivation: Why a new benchmark is needed

Existing negation retrieval benchmarks (e.g., NegBench) evaluate retrieval over a large unfiltered gallery (e.g., 5,000 COCO images). In such pools, the negated concept rarely co-occurs with the affirmed concept in the same image. As a result, even simple embedding correction-based methods perform competitively because there are few *confusing* images to distinguish from correct ones. This makes existing benchmarks inadequate for measuring *rejection*: without confusing distractors, dilution suffices.

COCO-Confusion is constructed precisely to expose this gap. For every query, the gallery is composed of images that all contain the affirmed concept, partitioned into:

- **Correct images:** contain c^+ but *not* c^-
- **Confusing images:** contain c^+ *and* c^-

A method that merely dilutes c^- scores both groups similarly. Only a method that genuinely *rejects* c^- separates them.

D.2 Concept pair selection

We use the 80 object categories of COCO from the val2017 split, which contains 5,000 images with bounding-box annotations. For each image we extract its set of present categories from the official annotations.

Two-object pairs. For every ordered pair (c^+, c^-) with $c^+ \neq c^-$, we partition images that contain c^+ into:

$$\begin{aligned}\mathcal{I}^+(c^+, c^-) &= \{i : c^+ \in i \wedge c^- \notin i\}, \\ \mathcal{I}^\pm(c^+, c^-) &= \{i : c^+ \in i \wedge c^- \in i\}.\end{aligned}$$

We retain a pair only if $|\mathcal{I}^+| \geq M_{\min}$ and $|\mathcal{I}^\pm| \geq M_{\min}$ with $M_{\min} = 10$. From each subset we sample up to $M_{\max} = 50$ images uniformly at random (seed 42). The query template is:

“A c^+ is present but no c^- .”

Three-object pairs. For every triple of distinct categories (a, b, c) with at least $M_{\min}$ images containing all three, we generate three query partitions: $((a, b), c)$, $((a, c), b)$, and $((b, c), a)$, each treating two as affirmed and one as negated. Selection criteria are identical to the 2-object case applied to the partitioned subsets. The query template is:

“A c_1^+ and c_2^+ are present but no c^- .”

Coverage. The selection procedure covers 76 of 80 COCO categories in 2-object pairs and 62 of 80 in 3-object pairs (Table 5). Categories absent from the benchmark are those with insufficient co-occurrence frequency to satisfy $M_{\min}=10$ in matched-pair construction (e.g., *toaster*, *hair drier*).

Table 5: Concept coverage in COCO-Confusion.

	2-object	3-object
Total pairs (post-verification)	950	1,740
Unique affirmed concepts	75	62
Unique negated concepts	76	62
COCO categories used	76/80	62/80

D.3 OWLv2 verification

COCO bounding-box annotations are accurate but may have boundary cases (heavy occlusion, partial visibility). To ensure clean labels, we re-verify every image–concept pair using OWLv2, an open-vocabulary object detector. We use the public checkpoint `google/owlv2-base-patch16-ensemble` with text queries of the form “*a photo of a c.*”

For each image i and concept set $\mathcal{C}_q = \{c^+\} \cup \{c^-\}$, OWLv2 produces a maximum detection confidence $\text{score}_c(i) \in [0, 1]$ per concept. We retain images according to the following rules at threshold $\tau_{\text{OWL}} = 0.1$:

- **Correct image kept** iff $\max_{c \in \mathcal{C}^+} \text{score}_c(i) \geq \tau_{\text{OWL}}$ and $\text{score}_{c^-}(i) < \tau_{\text{OWL}}$.
- **Confusing image kept** iff $\max_{c \in \mathcal{C}^+} \text{score}_c(i) \geq \tau_{\text{OWL}}$ and $\text{score}_{c^-}(i) \geq \tau_{\text{OWL}}$.
- **Pair kept** iff ≥ 5 verified images remain in both subsets.

The low threshold $\tau_{\text{OWL}}=0.1$ favors recall, reducing the risk of falsely classifying a partially-visible c^- object as “absent.”

Retention rates. Pair-level retention is high: 99.9% (2-object) and 97.1% (3-object), consistent with COCO annotations being already accurate. Filtering primarily refines the per-pair image lists rather than discarding pairs.

Table 6: COCO-Confusion final statistics.

	2-object	3-object
Total pairs	950	1,740

D.4 Evaluation protocol

For each query, the retrieval gallery is the union of its verified correct and confusing images. We compute mean Average Precision (mAP) over the gallery, with correct images as ground-truth positives. The benchmark-level metric is the macro-average of per-pair mAP across the 950 (Conf-2C) or 1,740 (Conf-3C) pairs.

D.5 Qualitative examples

Figure 4 shows a 2-object pair (*cat / dog*). Figure 5 shows a 3-object pair (*apple + banana / bowl*). In each figure, the top row shows verified correct images (green border, contain the affirmed concept(s) but not the negated concept) and the bottom row shows verified confusing images (red border, contain both the affirmed and negated concepts).

D.6 What makes COCO-Confusion different

Compared to NegBench standard retrieval (1 query against a 5,000-image gallery dominated by irrelevant images):

1. **Curated galleries with high confusing density.** Each pair gallery contains 50–100 images, of which 20–50% contain the negated concept. Standard retrieval has $< 1\%$ confusing images per query.
2. **OWLv2 cross-validation.** Every image label is double-verified by an external open-vocabulary detector, eliminating bounding-box edge cases.

Together these properties make COCO-Confusion a stress test for *rejection*, while standard retrieval benchmarks remain useful for measuring *dilution-level* negation handling.

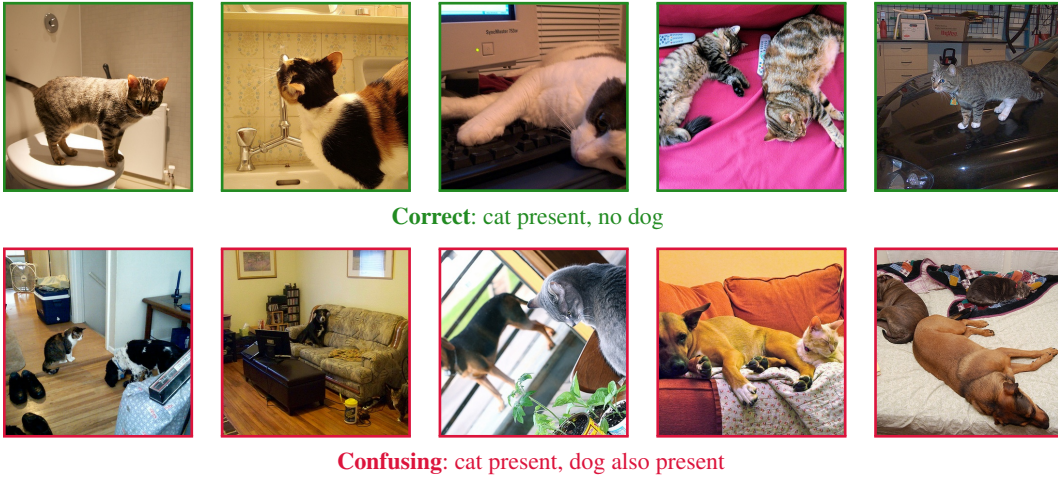


Figure 4: 2-object COCO-Confusion example. *Query: “a cat is present but no dog.”* Top row: correct images, containing a cat but no dog. Bottom row: confusing images, containing both a cat and a dog. All ten images contain a cat—only the negated concept differs across rows. A method that merely dilutes the “dog” direction in the text embedding scores all ten images similarly. Only a method that genuinely rejects c^- separates correct from confusing.



Figure 5: 3-object COCO-Confusion example. *Query: “an apple and banana are present but no bowl.”* The query requires affirming both apple and banana while rejecting bowl. All shown images contain an apple and a banana. Only the bottom row also contains a bowl.