Probing for Negation: Extracting Compositional Negation Signals from Pretrained Language Representations for Image Retrieval

Animesh Singh
IIT Delhi
bsy257559@iitd.ac.in

Brejesh Lall
IIT Delhi
brejesh@ee.iitd.ac.in

Abstract

Vision-language models such as CLIP exhibit strong zero-shot performance on image-text matching yet consistently fail to handle *negation* compositionally. Queries like "a dog that is not running" are collapsed into co-occurrence statistics, with the negation cue ignored. Critically, this failure is not random: because negation removes the most discriminative concept from a query, negation-blind models are *systematically misled*, they confidently rank the wrong images highest rather than producing an uninformative ranking. Prior work addresses this by retraining with negation-augmented data or learning embedding-level transformations. We propose **vpCLIP** (verypositiveCLIP), a fundamentally different approach: we extract the negation target at the *query level* using a frozen DistilBERT model (66M parameters). vpCLIP probes intermediate attention layers for negation-conditioned attention flow, removes candidate concept directions from hidden representations, and scores the resulting semantic shift to identify the negated concept. The extracted signal drives a retrieve-and-rerank pipeline on top of unmodified CLIP. On three curated negation benchmarks (Visual Genome, MS-COCO, Flickr30k; 800 queries total) where random retrieval averages 47–50%, vpCLIP achieves 68–79% Top-1 accuracy with CLIP ViT-B/32, outperforming CLIP (27–41%), NegCLIP (24–37%), ConCLIP (41–58%), and BLIP-2 (26–34%), all of which score *below* random on average, while requiring no task-specific training. Because the extraction stage is backbone-agnostic, vpCLIP generalizes without modification: paired with SigLIP2 SO400M it reaches 87.5% on VG, 93.2% on COCO, and 85.6% on F30k (88.8% average). We further demonstrate zero-shot transfer to an affirmation-biased medical encoder (PLIP), achieving 96% Top-1 on histology images. Our results show that negation signals are latent in the intermediate representations of an auxiliary pretrained language model and require structured probing to extract, challenging the assumption that compositional negation necessitates large-scale retraining of the vision-language backbone.

1 Introduction

Contrastive vision-language models (VLMs) such as CLIP (Radford et al., 2021) have transformed multimodal understanding, yet exhibit a systematic failure: they cannot process *negation* compositionally. The query "a person not wearing a hat" retrieves images of people wearing hats. The model treats content words as co-occurring concepts and ignores the negation operator. The contrastive objective maximizes alignment between matched image-text pairs, rewarding co-occurrence patterns, but negation requires *suppressing* a concept while preserving the remainder of the query semantics.

Prior approaches address this along two axes. **Data augmentation approaches** (NegCLIP (Yuksekgonul et al., 2023), ConCLIP (Singh et al., 2024)) retrain CLIP with negation-aware pairs, but require large-scale training pipelines and may not transfer across negation types. **Representation-level approaches** (Sammani et al., 2026; Aggarwal et al., 2026) model negation as an embedding-space transformation, conflating negation understanding with embedding geometry. (Jha et al., 2026) address exclusion queries by learning disentangled multimodal representations, but require training a dedicated model.

We propose a third paradigm: **query-level negation extraction**. Rather than modifying the embedding space, we extract the negation signal from the input query using a frozen pretrained language model. Our insight is that negation signals are already encoded in this auxiliary language model's intermediate representations; they are simply not accessible through final-layer pooled

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 525–539
November 6-10, 2026 ©2026 Association for Computational Linguistics

embeddings.

vpCLIP operates in three stages:

1. **Candidate identification**: extract candidate concepts (nouns, adjectives, gerunds) via POS tagging.

2. **Negation target extraction**: probe DistilBERT's intermediate attention layers (1–2), remove each candidate's embedding direction from hidden representations, and select the candidate producing the largest combined signal, them being attention weight, embedding displacement, and positional proximity to the negation cue.

3. **Negation-aware retrieval**: decompose the query into a positive subquery and a negative term; retrieve by positive similarity, then rerank by penalizing similarity to the negated concept.

This modular architecture requires *no additional training*, is *parameter-efficient* (only DistilBERT (Sanh et al., 2019), 66M parameters on top), and is *interpretable*: the system explicitly identifies the negated concept.

Our contributions are:

- We demonstrate that negation signals are latent in the intermediate representations of an auxiliary pretrained language model and can be extracted through structured probing without task-specific training, leaving the vision-language backbone unchanged.

- We introduce a scoring function combining attention-based saliency, embedding-direction removal, and positional proximity to identify negation targets.

- We construct three curated negation benchmarks (Visual Genome (Krishna et al., 2017): 296, MS-COCO (Lin et al., 2014): 220, Flickr30k (Young et al., 2014): 284 queries; 800 total) with controlled hard negatives, showing vpCLIP substantially outperforms all baselines.

- We demonstrate backbone-agnosticism: the same extraction module improves retrieval with NegCLIP, ConCLIP, SigLIP2, CLIP ViT-L/14, and BLIP-2, achieving up to 93.2% Top-1 on COCO with SigLIP2 SO400M (88.8% average), and transfers zero-shot to medical pathology retrieval (96% Top-1 with PLIP).

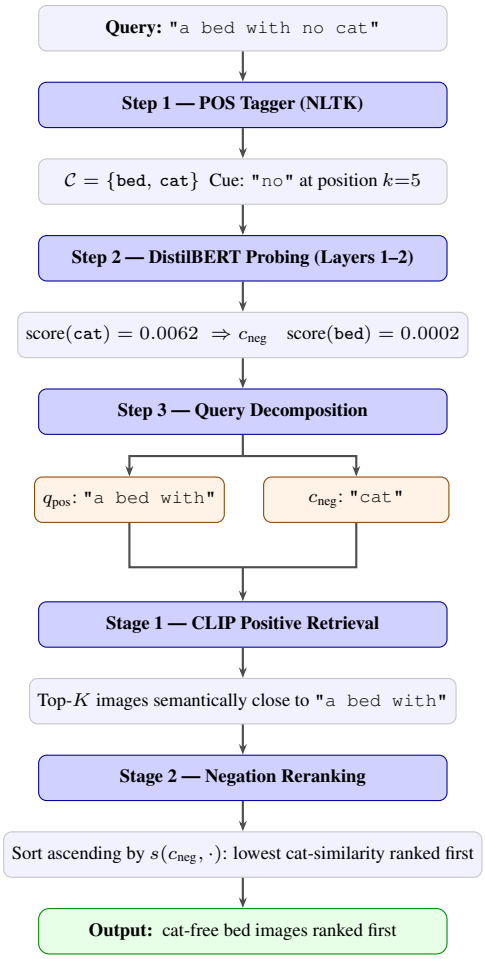


Figure 1: The **vpCLIP** pipeline on "a bed with no cat." A POS tagger extracts candidates; DistilBERT's early layers score each via attention saliency, embedding-direction removal, and proximity to the cue, selecting `cat` as c_{neg}; a two-stage CLIP pipeline retrieves bed-relevant images and reranks by suppressing the negated direction.

2 Related Work

Compositional Understanding in VLMs. Recent work highlights systematic failures of CLIP-style models on compositional reasoning. Yuksekgonul et al. (2023) introduce the ARO benchmark showing CLIP struggles with attribute binding and word order; Thrush et al. (2022) demonstrate similar failures on Winoground. These motivate targeted interventions for specific compositional phenomena, of which negation is among the most challenging.

Negation in Vision-Language Models. NegCLIP (Yuksekgonul et al., 2023) retrains with hard

negatives constructed by swapping objects and attributes; ConCLIP (Singh et al., 2024) fine-tunes with contrastive negation pairs. Both require end-to-end retraining. Alhamoud et al. (2025) introduce NegBench (79k examples across image, video, and medical data), confirming that modern VLMs perform at chance level on negation regardless of phrasing. Sammani et al. (2026) show that a linear classifier on CLIP's intermediate hidden states achieves 99%+ accuracy distinguishing negated from non-negated captions, confirming a linearly separable negation direction exists, supporting our claim that negation signals are latent rather than absent. Their method uses an LLM-generated training set (4,000 captions) and a learned classifier to detect negation; vpCLIP achieves comparable extraction through unsupervised structured probing with no training data. Aggarwal et al. (2026) propose embedding correction for open-corpus retrieval via rule-based negation scope detection and vector arithmetic, requiring dataset-level calibration of a scaling hyperparameter λ (optimal 1.9 on NegBench). Neither constraint is compatible with zero-shot deployment in new domains. vpCLIP addresses a different problem: it is a *reranking module* for hierarchical retrieval where a first stage establishes semantic recall, using structured attention probing with no tunable parameters. In the text-only domain, Hossain et al. (2022) find transformers perform substantially worse on negation instances, and Kassner and Schütze (2020) show language models fail to distinguish negated from affirmed facts. Salin et al. (2022) probe CLIP's intermediate representations for multimodal alignment, and the negation scope detection literature (Morante and Blanco, 2012) provides the linguistic foundation for our proximity-based scoring heuristic.

Probing Transformer Representations. The probing literature (Belinkov et al., 2017; Conneau et al., 2018; Hewitt and Manning, 2019) establishes that transformers encode rich linguistic information in intermediate layers. Clark et al. (2019) identify attention heads specializing in syntactic relations; Vig and Belinkov (2019) find attention aligns with dependency structure in middle layers. vpCLIP operationalizes these findings as a functional retrieval component. We note that our use of "probing" differs from supervised probing classifiers: vpCLIP performs *unsupervised structured extraction* (reading attention weights, removing embedding directions, and measuring semantic displacement) without any learned parameters.

3 Methodology

We present a modular pipeline for negation-aware image retrieval operating on frozen pretrained models, consisting of candidate extraction (§3.3), negation target identification (§3.4), and negation-aware retrieval (§3.5). Figure 1 illustrates the full pipeline; the complete algorithm is in Appendix A.

3.1 Problem Formulation

Let q denote a query containing a negation cue, and $\mathcal{R} = \{I_1, \ldots, I_K\}$ a *candidate set* retrieved by a first-stage system on positive semantic relevance. vpCLIP operates as a **second-stage reranking module**: given $\mathcal{R}$, it reranks candidates to surface those satisfying the negation constraint. This hierarchical formulation reflects production retrieval architectures (e.g., medical image retrieval, e-commerce category-filtered search) where a first stage establishes semantic recall and a second stage applies fine-grained discriminative reasoning. **Scope.** vpCLIP is designed as a *second-stage discriminator* for candidate sets where the visual encoder can meaningfully score images against query concepts; fully open-corpus retrieval is outside its current design scope (see Limitations).

The standard CLIP similarity $s(q, I) = \cos(f_t(q), f_v(I))$ is unreliable for negated queries because $f_t(q)$ does not compositionally represent negation. Our goal is to decompose q into a positive subquery q_{pos} and a negated concept c_{neg}, then retrieve by q_{pos} similarity and rerank by penalizing c_{neg} similarity. The central challenge is automatically identifying c_{neg} in a training-free, domain-agnostic manner.

3.2 Why a Separate Language Model?

CLIP's text encoder is optimized for global semantic alignment, not fine-grained linguistic compositionality. Its contrastive objective provides no signal to distinguish "a dog" from "no dog": both co-occur with similar images, so negation cues become irrelevant noise. In contrast, DistilBERT's masked language modeling objective forces prediction of masked tokens from surrounding context, building token-level sensitivity that makes intermediate attention layers useful for negation scope identification.

3.3 Candidate Extraction

We extract candidates $\mathcal{C}$ via POS tagging, retaining nouns (NN$*$), adjectives (JJ), and gerunds (VBG/VBN), excluding negation cues $\mathcal{N}$ = {"not", "no", "without", "n't", ...}:

$$\mathcal{C} = \{w \mid (w, \tau) \in \text{POS}(q),\ \tau \in \mathcal{T},\ w \notin \mathcal{N}\} \tag{1}$$

If $\mathcal{C} = \emptyset$ (e.g., the query contains only a negation cue and function words), the system falls back to standard CLIP retrieval on the original query q.

3.4 Negation Target Identification

We use a frozen DistilBERT model g with 6 layers, accessing both attention weights and hidden states. Let the query be tokenized into $[t_1, \ldots, t_T]$, with k the position of the negation cue and j the position of candidate c.

3.4.1 Attention-Based Saliency

We aggregate attention from the negation cue across intermediate layers. Let $A^{(l)} \in \mathbb{R}^{H \times T \times T}$ be the attention matrix at layer l:

$$\alpha_i = \sum_{l \in \mathcal{L}} \frac{1}{H} \sum_{h=1}^{H} A^{(l)}_{h,k,i} \tag{2}$$

where $\mathcal{L} = \{1, 2\}$ (selected empirically on a held-out set of images and queries disjoint from all NegHard benchmarks; see Table 3 for layer ablations). The normalized weight for candidate c at position j is $\hat{\alpha}_j = \alpha_j / (\sum_i \alpha_i + \epsilon)$. Early-to-intermediate layers are chosen because they preserve token-level identity while integrating contextual information, whereas deeper layers smooth individual token contributions.

3.4.2 Embedding Direction Removal

We measure each candidate's semantic importance by quantifying the sentence representation change when its embedding direction is removed. Let $\mathbf{h} \in \mathbb{R}^{T \times D}$ denote hidden states from the last layer and $\mathbf{e} \in \mathbb{R}^{T \times D}$ the token embeddings. For candidate c at position j, the embedding direction is $\mathbf{d}_j = \mathbf{e}_j / (\|\mathbf{e}_j\| + \epsilon)$. We remove this direction via orthogonal projection:

$$\tilde{\mathbf{h}}_i = \mathbf{h}_i - (\mathbf{h}_i \cdot \mathbf{d}_j)\, \mathbf{d}_j \tag{3}$$

The original and modified sentence embeddings are:

$$\mathbf{E}_{\text{orig}} = \text{Norm}\left(\tfrac{1}{T} \textstyle\sum_i \mathbf{h}_i\right) \qquad \mathbf{E}^{(j)}_{\text{mod}} = \text{Norm}\left(\tfrac{1}{T} \textstyle\sum_i \tilde{\mathbf{h}}_i\right) \tag{4}$$

The displacement $d_j = \|\mathbf{E}_{\text{orig}} - \mathbf{E}^{(j)}_{\text{mod}}\|_2$ and importance $\delta_j = 1/(d_j + \epsilon)$.

Intuition. The *subject* (e.g., "bed") is typically the semantic core: removing its direction causes large displacement. The *negation target* ("cat") is subordinate, producing smaller displacement. Taking $\delta_j = 1/d_j$ down-weights dominant candidates and favours the scoped target. In multi-concept queries (e.g., "a bed with a dog but no cat"), displacement alone cannot distinguish "dog" from "cat"; however, the multiplicative combination with attention flow and proximity to the negation cue (Eq. 5) disambiguates, since "no" attends strongly to "cat" and is positionally adjacent to it. We use embedding-layer directions rather than hidden states because embeddings provide clean, context-independent lexical directions. This step is related to representation erasure (Ravfogel et al., 2020), but repurposed as a diagnostic probe rather than a debiasing operation.

3.4.3 Composite Scoring

We combine attention saliency, displacement, and positional proximity multiplicatively:

$$\text{score}(c_j) = \delta_j \cdot \hat{\alpha}_j \cdot \frac{1}{|j - k| + 1} \tag{5}$$

The negation target is $c_{\text{neg}} = \arg\max_{c_j \in \mathcal{C}} \text{score}(c_j)$. Multiplication enforces a strict conjunction: a near-zero value on any factor collapses the score, so a candidate must satisfy all three criteria simultaneously. Each signal is necessary but insufficient alone (see Appendix D for component ablations): displacement alone selects semantically dominant candidates (typically the grammatical subject rather than the negation target); attention from the cue is sometimes maximal on the cue token itself (a self-loop artefact), so attention alone mislabels the cue as the target; proximity alone fails on long-distance constructions such as "a person that is not wearing a hat".

3.5 Negation-Aware Retrieval

Given c_{neg}, we construct $q_{\text{pos}} = q \setminus (\mathcal{N} \cup \{c_{\text{neg}}\})$. Note that q_{pos} may be grammatically incomplete (e.g., "a bed with" from "a bed with no cat"); this is by design, as CLIP's text encoder is known to behave as a bag-of-words (Yuksekgonul et al., 2023), so residual function words have negligible effect on the retrieved embedding. Retrieval proceeds in two stages:

Stage 1: Positive Retrieval. Retrieve top-K images by $s(q_{\text{pos}}, I_i)$. K is a user-configurable parameter; on NegHard's small controlled pools, setting K to the full pool size passes all candidates to Stage 2 so retrieval performance is unaffected by the choice of K. For larger candidate sets, K should be set proportional to the pool size to ensure Stage 1 recall includes the correct positives.

Stage 2: Negation Reranking. Rerank by sorting $\mathcal{R}_K$ ascending by $s(c_{\text{neg}}, I)$: images least similar to the negated concept are ranked first. This two-stage formulation avoids tuning a penalty weight, which we found sensitive in preliminary experiments.

4 Experimental Setup

4.1 Benchmark Construction

Existing benchmarks do not adequately evaluate negation in retrieval. VALSE (Parcalabescu et al., 2022) includes negation but evaluates caption-image matching, not retrieval, and uses only binary pools. We introduce three domain-diverse negation retrieval benchmarks using images from MS-COCO val2017 (Lin et al., 2014), Visual Genome (Krishna et al., 2017), and Flickr30k (Young et al., 2014).

Each benchmark was constructed in two independent phases and combined to yield 220–296 queries per corpus (800 total), with zero query overlap between phases. Queries cover object negation ("a bed with no cat"), attribute/state negation ("an airplane that is not white"), relational negation ("a bike not on a road"), and phrasing variants ("without" alongside "no"/"not") to validate robustness across negation surface forms. For each query, we construct a pool of **positives** (images satisfying the full query including the negation constraint, verified against annotations) and **distractors** (images matching the positive concepts but *violating* the negation). Pool sizes vary across queries; the expected random Top-1 equals $n_{\text{pos}}/(n_{\text{pos}} + n_{\text{dis}})$ per query, averaging 47–50% across the three benchmarks.

The controlled pool is a deliberate methodological choice: by holding the context object constant, the only variable is the negated concept, isolating negation handling from general retrieval quality (following Winoground (Thrush et al., 2022); our multi-image pools are strictly more informative than binary pools). Construction used only caption text and bounding-box annotations; no query was selected, modified, or excluded based on model outputs. Crucially, NegHard's difficulty is *perceptual, not linguistic*: because every pool contains distractors that match the query if negation is ignored, a negation-blind model does not produce a random ranking, it confidently ranks the wrong images highest. This is what drives the large gap between baselines and vpCLIP: CLIP, NegCLIP, ConCLIP, and even BLIP-2 (3B parameters) are not confused by NegHard; they are *systematically misled* by it, consistently surfacing the negated concept as the top result. We evaluate retrieval on NegHard rather than NegBench (Alhamoud et al., 2025) because large-scale synthetic benchmarks suffer from two pathologies: (1) false-negative labeling, where correctly retrieved images are marked incorrect because only a single image per caption is treated as positive (Sammani et al., 2026), and (2) query underspecification, where removing the negated concept leaves an underspecified signal matching the majority of corpus images, so a negation-blind model may accidentally score correctly while a correct model is penalised. Empirically, vpCLIP achieves 70.2% extraction accuracy on NegBench yet retrieval scores remain poor, whereas on NegHard both extraction (91.2%) and retrieval (74.2% average) are strong, consistent with evaluation-side noise rather than an extraction failure. NegHard eliminates both pathologies: every positive and distractor is manually verified.

4.2 Baselines

We compare against: **CLIP ViT-B/32** and **CLIP ViT-L/14** (Radford et al., 2021) (zero-shot); **NegCLIP** (Yuksekgonul et al., 2023) (ViT-B/32 fine-tuned with hard negatives); **ConCLIP** (Singh et al., 2024) (ViT-B/32 fine-tuned with contrastive negation pairs); **SigLIP2 ViT-L-16** (Tschannen et al., 2025) (zero-shot); **SigLIP2 SO400M** (Tschannen et al., 2025) (zero-shot); **BLIP-2** (Li et al., 2023) (ITC retrieval head, zero-shot). Because vpCLIP's extraction is backbone-agnostic, we also evaluate with **NegCLIP**, **ConCLIP**, **CLIP ViT-L/14**, **SigLIP2 ViT-L-16**, **SigLIP2 SO400M**, and **BLIP-2** as backbones.

4.3 Metrics

Top-1 Accuracy: fraction of queries where the top-ranked image is a positive. **MRR@3**: Mean Reciprocal Rank over the top 3. **Recall@3**: fraction with at least one positive in the top 3.



Figure 2: Example retrieval pools for two NegHard queries. **Positives** satisfy the query including negation; **Distractors** contain the context object but violate the negation.

		Visual Genome (296q)			**MS-COCO** (220q)			**Flickr30k** (284q)		
Model	**Latency (ms)**	Top-1	MRR@3	R@3	Top-1	MRR@3	R@3	Top-1	MRR@3	R@3
Random	—	47.4	.690	92.0	50.0	.713	95.0	50.1	.713	95.0
CLIP-B/32	10.9	27.7	.542	79.7	40.5	.617	78.6	27.1	.533	75.7
CLIP-L/14	19.5	31.1	.564	79.4	35.0	.584	78.2	27.8	.544	76.8
NegCLIP	10.8	28.7	.545	78.0	37.3	.597	76.8	23.9	.503	69.7
ConCLIP	11.0	49.0	.701	92.2	57.7	.756	91.4	41.2	.651	90.1
SigLIP2 ViT-L-16	40.9	29.4	.552	78.7	34.1	.581	73.6	25.7	.513	71.1
SigLIP2 SO400M	47.1	32.8	.571	76.0	30.9	.564	73.6	26.1	.515	69.7
BLIP-2	25.7	29.7	.543	70.9	33.6	.574	72.7	26.4	.522	71.8
vpCLIP (Ours)	**30.1**	**67.6**	**.822**	**97.6**	**76.4**	**.866**	**98.6**	**78.5**	**.878**	**98.2**

Table 1: Negation retrieval on all three NegHard benchmarks (800 queries total; controlled pools of verified positives and hard-negative distractors; random Top-1 averages 47–50% per dataset). All baselines except ConCLIP fall *below* random on VG and F30k, confirming systematic negation-blindness; ConCLIP edges above random on VG (49.0% vs. 47.4%) and COCO but still trails vpCLIP by +19–37 pp. BLIP-2's QFormer text encoder fares no better than CLIP despite its cross-modal pretraining. vpCLIP outperforms the strongest baseline (ConCLIP) by +19–37 pp Top-1 using the same CLIP-B/32 backbone with no task-specific training. 95% bootstrap CIs for vpCLIP Top-1: ±5.4 pp (VG), ±5.7 pp (COCO), ±4.8 pp (F30k); all gaps over baselines are significant.

5 Results

5.1 Main Results

Table 1 presents the combined results across all three benchmarks.

Key Observations. (1) **Scale, architecture, and fine-tuning do not solve negation.** CLIP-L/14 (∼3× larger) shows only marginal improvement on VG and *decreases* on COCO. NegCLIP performs on par with or below vanilla CLIP-B/32 on all three benchmarks, suggesting hard-negative fine-tuning is insufficient for diverse compositional negation. BLIP-2, despite its cross-modal QFormer architecture, scores below CLIP-B/32 on COCO (33.6% vs. 40.5%) and F30k (26.4% vs. 27.1%), and only marginally above on VG (29.7% vs. 27.7%), confirming that richer cross-attention does not inherently resolve negation.

(2) **vpCLIP dominates across all datasets:** 67.6% on VG (+18.6 pp over ConCLIP), 76.4% on COCO (+18.7 pp), and 78.5% on F30k (+37.3 pp), with no task-specific training.

(3) **Negation scope vs. global absence.** Baselines collapse compositional queries ("an animal not running on grass") into global absence constraints ("animal not on grass"). vpCLIP correctly identifies "running" as the negation target via the composite score, decomposing into a positive subquery ("an animal on grass") and a penalty ("running").

Backbone	VG	COCO	F30k	Avg
CLIP ViT-B/32	67.6	76.4	78.5	74.2
CLIP ViT-L/14	68.2	79.5	79.2	75.6
NegCLIP ViT-B/32	78.7	84.5	82.0	81.7
ConCLIP ViT-B/32	68.9	74.5	79.9	74.4
SigLIP2 ViT-L-16	85.1	93.2	83.8	87.4
SigLIP2 SO400M	**87.5**	**93.2**	**85.6**	**88.8**
BLIP-2	70.3	77.3	78.2	75.3

Table 2: vpCLIP Top-1 (%) across backbones on NegHard (800 queries). The negation extraction module is identical in all rows; only the similarity backbone changes. NegCLIP as standalone scores 24–37%; as a vpCLIP backbone it reaches 81.7% average. BLIP-2 as standalone scores 26–34%; with vpCLIP it jumps to 75.3% (+45.4 pp), the largest absolute gain of any backbone.

Layers	Top-1	MRR@3
Layer 0 (embedding)	42.0	0.583
Layers 1–2	**73.0**	**0.835**
Layers 1–3	69.0	0.801
Layers 3–5	58.0	0.712
Layer 5 (final)	51.0	0.665

Table 3: Layer selection ablation (CLIP-B/32) on a held-out set (25 queries, 60 images) disjoint from all NegHard benchmarks. Intermediate layers (1–2) yield the best negation extraction; the embedding layer lacks contextual scope and the final layer smooths individual token contributions, consistent with negation scope being encoded intermediately and washed out by final-layer pooling.

5.2 Backbone Generalization

Table 2 verifies that vpCLIP's negation extraction is backbone-agnostic.

SigLIP2 SO400M yields the largest relative gains (+14.6 pp average over CLIP-B/32), with improvements concentrated on COCO (+16.8 pp) where the stronger encoder better discriminates fine-grained co-occurrence patterns. NegCLIP as backbone reaches 84.5% on COCO, dramatically higher than standalone NegCLIP (37.3%), illustrating that its fine-tuning provides useful visual grounding but requires explicit query-level negation handling to be effective. BLIP-2 shows the largest absolute improvement: from 29.9% standalone average (VG: 29.7%, COCO: 33.6%, F30k: 26.4%) to 75.3% with vpCLIP (+45.4 pp average), demonstrating that even cross-modal architectures with QFormer attention benefit from explicit negation extraction. ConCLIP as backbone yields similar average to CLIP-B/32 (74.4% vs. 74.2%), as expected: vpCLIP decomposes the query *before* encoding, so the backbone only sees q_{pos} and c_{neg}, neither of which contains negation. *No component of the extraction pipeline was modified* across any variant.

Negation extraction accuracy. We evaluate extraction accuracy using ground-truth labels from NegHard query structure and NegBench `negative_objects` annotations (full results in Appendix C). On NegHard, accuracy is 87.5% (VG), 90.0% (COCO), and 96.1% (F30k), with an overall average of 91.2%; on NegBench (24,077 queries), 70.2%. Extraction accuracy remains above 87% even on VG, whose prepositional constructions ("not *on* a road") are the hardest for the proximity and displacement signals; F30k is highest (96.1%) owing to its majority "with no X" patterns. The strong correlation between extraction and retrieval accuracy confirms that retrieval quality is primarily bounded by extraction quality. The lower NegBench figure (70.2% vs. 91.2% on NegHard) reflects two additive factors: (1) the greater syntactic diversity of NegBench's 24k machine-generated queries, which genuinely stresses the heuristic scoring beyond simple "no X" patterns, and (2) the evaluation pathologies described in §4.1, namely under-specification and false-negative labeling, which depress measured retrieval performance independently of extraction quality.

6 Discussion

Negation is not absent - it is latent. Our results challenge the assumption that pretrained models fail to encode negation. The strong performance of vpCLIP demonstrates that negation signals are present in DistilBERT's intermediate representations but not accessible through standard last-layer pooled embeddings. Table 3 localizes this directly: extracting from DistilBERT's intermediate layers (1–2) yields 73% downstream Top-1, against 42% from the embedding layer and 51% from the final layer, indicating that negation scope is encoded intermediately and smoothed away by final-layer pooling. This aligns with the probing literature (Belinkov et al., 2017; Hewitt and Manning, 2019) and is confirmed by Sammani et al. (2026), who show a linear classifier on CLIP's intermediate hidden states achieves 99%+ accuracy distinguishing negated from non-negated captions.

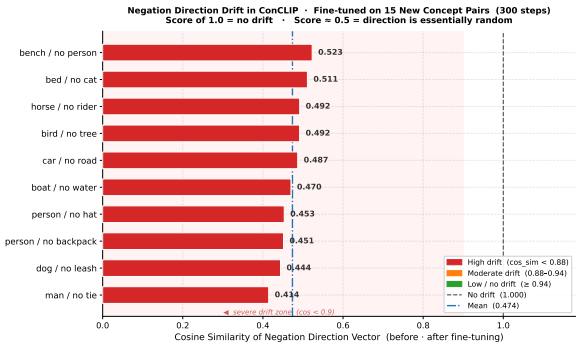


Figure 3: Per-concept direction drift in ConCLIP after 300 gradient steps on 15 new concept-negation pairs. Each bar is the cosine similarity between the negation direction vector *before* and *after* fine-tuning; 1.0 = no drift, ≈0.5 = random. The dash-dot line marks the mean (0.474). All original pairs fall in the severe-drift zone after minimal fine-tuning on unrelated concepts.

Computational efficiency. vpCLIP adds 19.2 ms to bare CLIP-B/32 for a total of 30.1 ms per query (2.8×; Table 1), well within interactive latency budgets (<100 ms). The additional 19 ms buys +18.6–37.3 pp Top-1 over the strongest baseline.

Robustness to concept-space expansion. NegCLIP and ConCLIP encode negation as a geometric transformation from finite contrastive pairs, creating structural fragility: extending to new concepts via fine-tuning risks catastrophic forgetting of the learned negation subspace. We verify this empirically by fine-tuning ConCLIP's text encoder on 15 new concept-negation pairs for 300 steps. The mean cosine similarity between negation direction vectors before and after fine-tuning is **0.47** (effectively random alignment) across 10 original concept pairs (Figure 3). PCA visualization confirms geometric mode collapse: all direction vectors collapse to a single point (Appendix E). This is consistent with the embedding-space analyses of Aggarwal et al. (2026) and Sammani et al. (2026). vpCLIP is structurally resilient: DistilBERT identifies any token without concept-specific training, and CLIP's frozen encoder scores any image–concept pair it was pretrained on, so introducing new vocabulary requires no retraining of either component.

Domain generalization: medical pathology retrieval. vpCLIP's extraction stage enables direct transfer to specialized domains. We demonstrate this on **Pathology_NegHard**, a 25-query proof-of-concept set from NCT-CRC-HE-100K colorectal histology (Kather et al., 2019), with 5 positives and 5 distractors per query across four tissue classes (TUM, LYM, STR, NORM). This is a *generalization demonstration*, not a benchmark; its purpose is to show that the extraction module transfers to a domain unseen during development. All negation targets are single-token pathology concepts with confirmed positive concept-direction margins in PLIP (Huang et al., 2023) embedding space. PLIP's training corpus (PathCap, Twitter pathology posts) consists of purely *positive* descriptions of visible findings, so it exhibits the same affirmation bias as general-domain CLIP. Unlike radiology-trained models such as MedCLIP (Wang et al., 2022) and BiomedCLIP (Zhang et al., 2023), whose training corpora are dominated by negated findings (e.g., "no evidence of metastasis," "unremarkable lymph nodes"), causing the encoder to associate disease concepts with their *absence* rather than their presence. This reverses concept-direction embeddings: ascending rerank by "carcinoma" would paradoxically *promote* images containing carcinoma, since the encoder's "carcinoma" direction points toward negated instances.

Method	Top-1	MRR@3
Random baseline	50.0%	—
PLIP	56.0%	0.688
vpCLIP + PLIP	**96.0%**	**0.980**

Table 4: Negation retrieval on **Pathology_NegHard** (25 queries, NCT-CRC-HE-100K; 5+5 pool; random = 50%). vpCLIP + PLIP achieves 96% Top-1, a 40 pp gain over PLIP alone. Representative pools are in Appendix F.

vpCLIP + PLIP achieves **96% Top-1** (Table 4), a 40 pp gain over PLIP alone, confirming the extraction module transfers directly without domain-specific modification.

7 Conclusion

vpCLIP achieves 68–79% Top-1 across three negation benchmarks (800 queries) with no task-specific training, outperforming all baselines by 19–37 pp using a frozen 66M-parameter DistilBERT probe. The backbone-agnostic extraction reaches 88.8% average with SigLIP2 SO400M and transfers zero-shot to medical pathology (96% Top-1 with PLIP). Our key finding is that negation signals are latent in the intermediate represen-

tations of an auxiliary pretrained language model and extractable through structured probing, challenging the assumption that compositional negation requires retraining the vision-language backbone.

Future Work. We plan to extend to phrase-level negation targets, multiple negation handling, and open-corpus retrieval via soft blended penalties. An open question is whether improving fine-grained visual grounding in foundation models would yield the dominant gains, making the bottleneck perceptual rather than linguistic.

Limitations

1. **Single-token assumption.** The current method assumes the negation target is a single token. Multi-token targets (e.g., "teddy bear") cause the scoring to select an incomplete token, accounting for the majority of COCO extraction errors. Extensions to phrase-level direction removal would address this.

2. **Multiple negations.** The pipeline handles single negation. Extending to multiple negation cues ("not a dog and not running") is a direct generalization: extraction collects one target per cue, and reranking scores each candidate by its *maximum* similarity to any negated concept (ascending), so that only images far from every negated concept rank first, reducing to the single-cue case when one cue is present. A benchmark for multi-cue negation to validate this quantitatively does not yet exist.

3. **Fine-grained visual concepts.** Reranking quality is bounded by the visual encoder's grounding capability. For queries involving small or occluded objects that the encoder does not reliably distinguish, the negation penalty becomes weak. Switching to SigLIP2 SO400M improves average accuracy from 74.2% to 88.8%, confirming stronger backbones directly improve retrieval.

4. **Evaluation scale.** NegHard benchmarks contain 220–296 queries each (800 total). Construction relied solely on captions and bounding-box annotations, with no model outputs informing query selection. As partial external validation, we evaluate extraction accuracy on the independently constructed NegBench (Alhamoud et al., 2025) (24,077 queries), achieving 70.2% (§5). We report extraction accuracy rather than a retrieval score on NegBench by design: its open-corpus setting falls outside vpCLIP's curated-pool scope, where Stage 1 reduces to standard CLIP retrieval and the Stage 2 rerank is calibrated for small pools (§3.5); combined with the false-negative labeling and saturation pathologies of §4.1, an open-corpus retrieval score on this benchmark does not isolate negation handling for any method.

5. **Open-corpus retrieval.** vpCLIP's ascending rerank relies on the visual encoder producing meaningful similarity scores for the negated concept across the candidate pool. In fully open-corpus settings, many candidate images have no semantic relationship to the query, so the encoder produces noisy scores that degrade the rerank signal. Adapting to open-corpus retrieval via a soft blended penalty that down-weights the negation term for low-relevance candidates is left as future work.

Broader Impact

vpCLIP targets negation-aware retrieval, a broadly beneficial capability for search and content filtering. The principal risk lies in high-stakes settings such as clinical decision support, where a silently mis-ranked result can mislead; the medical results here are a proof of concept, and any such deployment should retain human oversight.

References

Bhuvan Aggarwal, Amit More, Mudit Soni, and S Divakar Bhat. 2026. Seeing what's not there: Negation understanding needs more than training. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Kumail Alhamoud, Shaden Alshammari, Yonglong Tian, Guohao Li, Philip HS Torr, Yoon Kim, and Marzyeh Ghassemi. 2025. Vision-language models do not understand negation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP*.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Md Mosharaf Hossain, Dhivya Chinnappa, and Eduardo Blanco. 2022. An analysis of negation in natural language understanding corpora. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 716–723.

Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. 2023. A visual–language foundation model for pathology image analysis using medical twitter. *Nature Medicine*, 29:2307–2316.

Prachi Jha, Sumit Bhatia, and Srikanta Bedathur. 2026. Being positive about negative queries: Exclusion aware multimodal retrieval using disentangled representations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Jakob Nikolas Kather, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Enken Herpel, Cleo-Aron Weis, Timo Gaiser, Alexander Marx, Nektarios A Valous, Dyke Ferber, and 1 others. 2019. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLOS Medicine*, 16(1):e1002730.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael S Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Roser Morante and Eduardo Blanco. 2012. *SEM 2012 shared task: Resolving the scope and focus of negation. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 265–274.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2022. 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)*, pages 8253–8280.

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. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Emmanuelle Salin, Badreddine Farah, Stéphane Ayache, and Benoit Favre. 2022. Are vision-language transformers learning multimodal representations? a probing perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Fawaz Sammani, Tzoulio Chamiti, Paul Gavrikov, and Nikos Deligiannis. 2026. When negation is a geometry problem in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Multimodal Algorithmic Reasoning Workshop*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Jaisidh Singh, Ishaan Shrivastava, Mayank Vatsa, Richa Singh, and Aparna Bharati. 2024. Learn "no" to say "yes" better: Improving vision-language models via negations. *arXiv preprint arXiv:2403.20312*.

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

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Andreas Steiner, Daniel Keysers, Neil Houlsby, and Xiaohua Zhai. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP*.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022. MedCLIP: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

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

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, and 1 others. 2023. BiomedCLIP: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*.

A Algorithm

Algorithm 1 vpCLIP: Negation-Aware Image Retrieval

Require: Query q, image corpus $\mathcal{I}$, CLIP encoders f_t, f_v, DistilBERT g

Extract candidates: $\mathcal{C} \leftarrow \{w \mid (w, \tau) \in \text{POS}(q),\ \tau \in \mathcal{T},\ w \notin \mathcal{N}\}$ (Eq. 1)

Locate negation cue: $k \leftarrow \text{FindNeg}(q)$

Run DistilBERT g on q to obtain attention matrices $A^{(l)}$ and hidden states $\mathbf{h}$

Compute attention scores: $\alpha_i \leftarrow \sum_{l \in \mathcal{L}} \frac{1}{H} \sum_h A^{(l)}_{h,k,i}$ (Eq. 2)

Normalize: $\hat{\alpha}_j \leftarrow \alpha_j / (\sum_i \alpha_i + \epsilon)$; $\mathbf{E}_{\text{orig}} \leftarrow \text{Norm}\big(\frac{1}{T} \sum_i \mathbf{h}_i\big)$

for each $c_j \in \mathcal{C}$ **do**

  Direction: $\mathbf{d}_j \leftarrow \mathbf{e}_j / (\|\mathbf{e}_j\| + \epsilon)$

  Project out: $\tilde{\mathbf{h}}_i \leftarrow \mathbf{h}_i - (\mathbf{h}_i \cdot \mathbf{d}_j)\, \mathbf{d}_j$ for all i (Eq. 3)

  $\mathbf{E}^{(j)}_{\text{mod}} \leftarrow \text{Norm}\Big(\frac{1}{T} \sum_i \tilde{\mathbf{h}}_i\Big)$

  $d_j \leftarrow \|\mathbf{E}_{\text{orig}} - \mathbf{E}^{(j)}_{\text{mod}}\|_2$; $\delta_j \leftarrow 1/(d_j + \epsilon)$

  $\text{score}(c_j) \leftarrow \delta_j \cdot \hat{\alpha}_j \cdot \frac{1}{|j-k|+1}$ (Eq. 5)

end for

Select: $c_{\text{neg}} \leftarrow \arg\max_{c_j \in \mathcal{C}} \text{score}(c_j)$

Build: $q_{\text{pos}} \leftarrow q \setminus (\mathcal{N} \cup \{c_{\text{neg}}\})$

Stage 1: Retrieve $\mathcal{R}_K \leftarrow \text{TopK}\big(s(f_t(q_{\text{pos}}),\ f_v(I_i))\big)_{i=1}^{N}$

Stage 2: Re-rank Sort $\mathcal{R}_K$ ascending by $s(f_t(c_{\text{neg}}),\ f_v(I))$

return Reranked image list

B Pipeline Illustration

Stage	Details
Input Query	``a bed with no cat''
Step 1: Candidates	$\mathcal{C} = \{\text{bed}, \text{cat}\}$
Step 2: Negation-Target	cue "no" at $k = 5$; scores: `cat`: 0.0062, `bed`: 0.0002 $\Rightarrow$ $c_{\text{neg}} =$ `cat`
Step 3: Decompose	q_{pos}: ``a bed with'' $c_{\text{neg}}$: ``cat''
Step 4: Retrieve & Rerank	Top-K beds retrieved; reranked ascending by $s(\texttt{cat}, \cdot) \Rightarrow$ cat-free beds ranked 1st

Table 5: Step-by-step pipeline walkthrough for "a bed with no cat." The composite scoring correctly identifies `cat` as the negation target (31× higher score than `bed`).

C Negation Extraction Accuracy

Dataset	Queries	Correct	Accuracy
NegHard – VG	296	259	87.5%
NegHard – COCO	220	198	90.0%
NegHard – F30k	284	273	96.1%
NegHard (avg)	800	730	91.2%
NegBench	24,077	16,897	70.2%

Table 6: Negation target extraction accuracy. Ground truth for NegHard from query text; for NegBench from `negative_objects` annotations. Two systematic failure patterns account for nearly all errors: (1) prepositional constructions ("a bike not *on* a road") where the intervening preposition causes proximity and displacement to favour the subject; (2) multi-word compound nouns ("teddy bear") where the single-token assumption selects an incomplete token.

D Component and Positional Ablation

Configuration	Top-1	MRR@3
CLIP only (no negation)	27.0	0.467
CLIP + POS rule (first noun after cue)	41.0	0.583
CLIP + Attention only	55.0	0.698
CLIP + Displacement only	48.0	0.641
CLIP + Full scoring (vpCLIP)	**73.0**	**0.835**

Table 7: Component ablation (CLIP-B/32) on a held-out set of 25 queries (simple to complex, including rephrased variants) over 60 curated images, disjoint from all NegHard evaluation benchmarks. Individual signals alone provide partial improvement, but the multiplicative combination of all three substantially outperforms any single component, confirming that the full pipeline is necessary.

Method (full pipeline, 800 queries)	Top-1	MRR@3
CLIP only (no negation)	31.0	0.559
Proximity only	72.5	0.844
POS rule (first noun after cue)	73.5	0.852
Full composite (vpCLIP)	**72.5**	**0.845**

Table 8: Positional baselines within the full vpCLIP pipeline, evaluated on all 800 NegHard queries. On NegHard's syntax, where the negated target is typically adjacent to the negation cue, positional heuristics match the full composite; the decisive gain over CLIP-only ($31.0 \rightarrow 72.5$ Top-1) comes from query decomposition rather than the choice of extractor. The composite's advantage over positional scoring emerges on the syntactically varied held-out set (Table 7).

E Drift PCA Visualization

Figure 4 visualizes the principal components of the negation direction vectors before and after fine-tuning.

F Pathology Retrieval Pools

Figure 5 provides examples of the retrieved positive and distractor pools from the NCT-CRC-HE-100K histology dataset.

G Qualitative Retrieval Comparison

Figures 6 and 7 show the top-1 retrieved image for six representative queries across all five evaluated models.

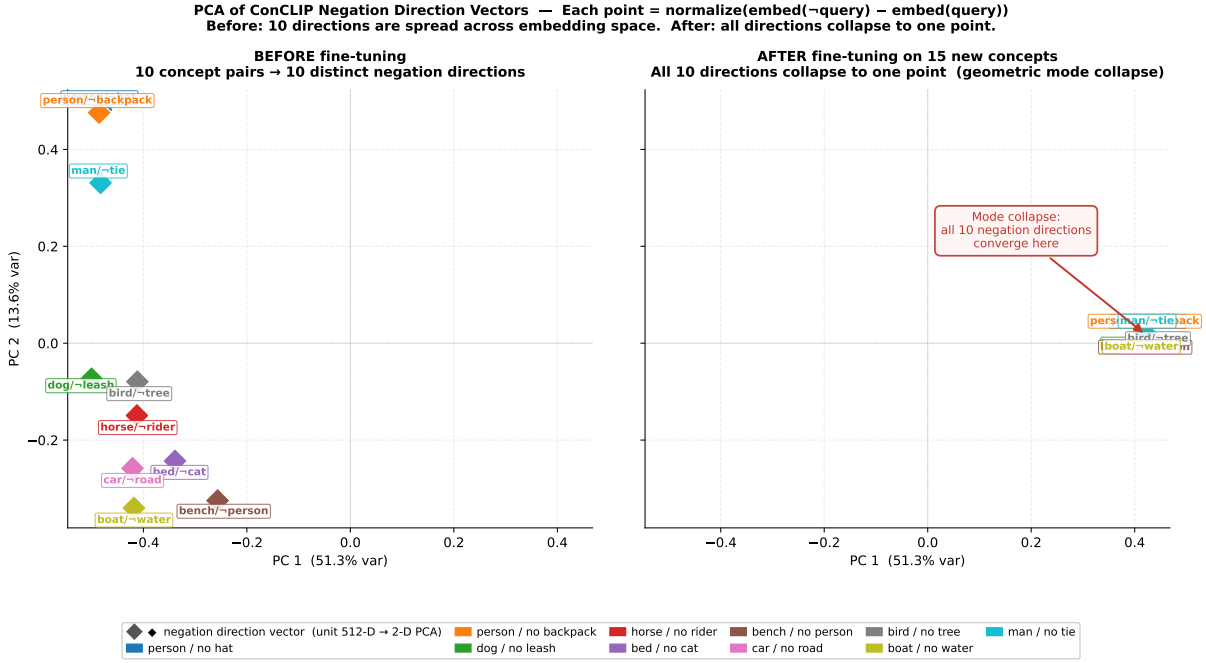


Figure 4: PCA of ConCLIP *negation direction vectors*; each point is normalize(embed(negated) − embed(positive)) projected to 2D. **Left (before):** 10 direction vectors spread across distinct regions, each encoding unique negation geometry. **Right (after):** after 300 fine-tuning steps on 15 new concept pairs, all vectors collapse to a single point, indicating geometric mode collapse that destroys all concept-specific structure.

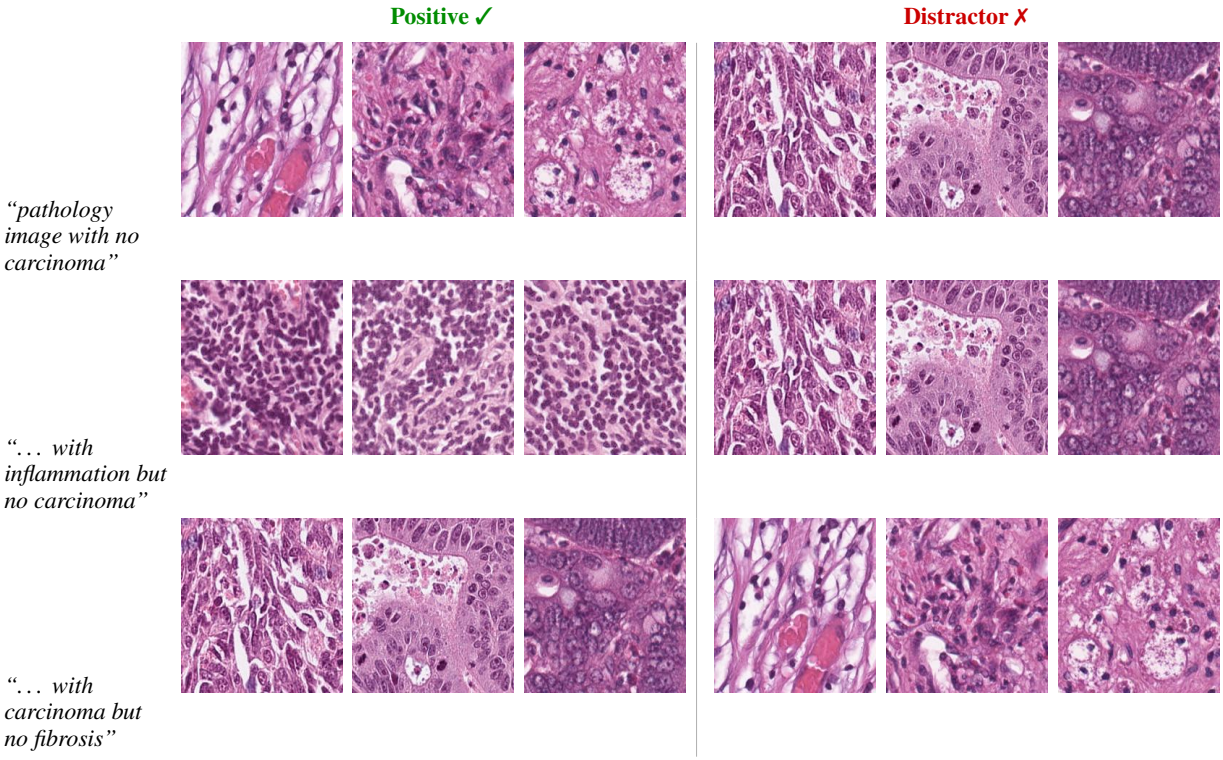


Figure 5: Representative retrieval pools from **Pathology_NegHard** (NCT-CRC-HE-100K). **Positives** satisfy the query including negation; **Distractors** contain the context tissue but violate the negation. *Top:* simple negation; *Middle:* contextual negation (inflammation present, carcinoma absent); *Bottom:* contextual negation (carcinoma present, fibrosis absent). vpCLIP + PLIP correctly retrieves positives for all three; PLIP alone fails.

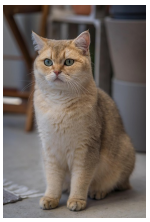
Query	CLIP-B/32	CLIP-L/14	NegCLIP	ConCLIP	vpCLIP
					
"a fridge without fruits"	✗	✗	✓	✗	✓
					
"people playing in a park that are not adults"	✗	✗	✗	✗	✓
					
"a garden with no flowers in it"	✗	✗	✗	✗	✓
					
"a red car that is not moving"	✗	✗	✗	✓	✓
					
"an animal that is not on grass at all"	✗	✗	✗	✗	✓

Figure 6: Top-1 retrieved image per model for five unambiguous negation queries. ✓ = correct (negated concept absent); ✗ = incorrect (negated concept present). vpCLIP correctly suppresses the negated concept in all five cases.

Query	CLIP-B/32	CLIP-L/14	NegCLIP	ConCLIP	vpCLIP
					
"a person not wearing a red cap"	✗	✗	✗	✗	✗

Failure case: semantic ambiguity in negation scope. *The query admits multiple valid interpretations: no cap at all, a cap of another colour, or holding a red cap without wearing it. No retrieval model can resolve this ambiguity without additional context. This class of failure is inherent to compound-attribute negation and not specific to vpCLIP.*

Figure 7: Failure case: compound-attribute negation query where all models fail due to semantic ambiguity.