Negation Matters: Training-Free Negation-Aware Image Retrieval

Aashish Pokhrel and Shivanand Venkanna Sheshappanavar
Geometric Intelligence Research Lab., Dept. of Electrical Engineering and Computer Science
University of Wyoming, USA
{apokhrel, ssheshap}@uwyo.edu

Figure 1. Top-5 retrieval results for the negated query *"An image of a dog not on a beach."* **Red borders** indicate incorrect retrievals (images containing a beach or sand); **green borders** indicate correct retrievals (non-beach scenes). The CLIP baseline [24] ranks beach images at all five positions, failing to respect the negation constraint. Our SpaceVLM-DRC correctly excludes the negated concept across all top-5 ranks without any additional training.

Abstract

Negation—the linguistic ability to assert the absence of a concept—is a critical bottleneck in vision-language understanding. Vision-language models have demonstrated remarkable success in aligning visual and textual representation across domains such as content moderation, medical image retrieval, and natural language-guided search. Yet, these models consistently fail to handle negation robustly, often retrieving images that contain the very concept that was explicitly excluded. Existing methods address this limitation through fine-tuning on synthetic negation corpora—an approach that is computationally expensive, dataset-dependent, and prone to compromising generalization to unseen distributions. We propose ***SpaceVLM-DRC****, which introduces* ***D****ynamic* ***R****epulsion with* ***C****ontext Anchoring (DRC) into the SpaceVLM framework as a training-free inference time negation resolution mechanism over a frozen CLIP backbone. This framework first decomposes queries into affirmative, negated, and counterfactual components, then applies dynamic repulsion to push negated concepts away in the embedding space, and finally anchors retrieval within the full-caption context to preserve semantic coherence. SpaceVLM-DRC surpasses state-of-the-art results on the MSRVTT negation retrieval benchmark and achieves performance comparable to fully fine-tuned approaches on the COCO negated retrieval dataset. Crucially, it requires no model retraining while preserving zero-shot generalization on non-negated queries. Our code is available at https://github.com/aashishpokhrel27/spacevlm-drc.*

1. Introduction

Vision-language models (VLMs) have become central to modern multimodal applications, including image re-

trieval [6, 10, 24], visual question answering [2, 6], and natural-language-guided decision-making in domains such as healthcare [8], robotics [18], and content management [6]. By jointly encoding visual and textual input, these VLMs enable users to query large visual collections using natural language, supporting context-aware responses that go far beyond traditional vision pipelines [4, 16, 17, 21]. As VLMs are deployed in increasingly safety-critical and user-facing systems [1, 15, 20], their ability to correctly follow nuanced linguistic instructions becomes essential. Within this landscape, negation stands out as a particularly important yet underexplored capability [3, 5].

Many realistic queries depend not only on what should be present in a scene but also on what must be explicitly absent, such as *"retrieve scans showing a tumor but without calcification," "show street scenes with cars but no pedestrians,"* or *"find images of a beach without people."* These are natural, well-formed requests—yet current VLMs routinely fail them. Such failures produce qualitatively incorrect retrievals even when the models appear strong on standard benchmarks, directly undermining trust in safety-critical applications such as medical imaging [8], autonomous systems [18], and content filtering [6].

The root cause lies in the training of VLMs [22]. Models such as CLIP are optimized on large-scale collections of positive image-text pairs, implicitly reinforcing associations between visual content and descriptive language [24]. Negation words such as "no," "not," and "without" appear infrequently in the pretraining corpora ($\leq 0.7\%$ of the captions) and are never explicitly modeled [3, 22, 23]. Consequently, at retrieval time, models tend to latch onto dominant positive nouns and attributes while effectively ignoring the negation operator [3, 22]. Empirical studies confirm this: when queried with negated descriptions, strong VLMs frequently return images containing the very object that was supposed to be excluded [3]. This failure is not subtle. As illustrated in Figure 1, given the query *"An image of a dog not on a beach,"* CLIP retrieves beach scenes of every rank [24].

Previous works [3, 22] have largely addressed negation through additional training or fine-tuning on synthetic datasets, where captions are programmatically edited to introduce constructs such as "no X," "without Y," or "A but not B." While such approaches improve performance on targeted benchmarks [22], they have significant drawbacks: high computational cost, dependence on large, curated corpora, and overfitting to the synthetic patterns observed during training. As a result, generalization to natural queries degrades, and negation handling becomes inseparable from model weights [7, 25].

Evaluation progress has also been constrained by the scarcity of negation-focused benchmarks, with early efforts like MSR-VTT and COCO-based ones providing valuable but limited insights into model unreliability [3, 28]. These benchmarks reveal the unreliability of current VLMs under negation, but their limited scale and diversity hinder the development of systematic methods and fair comparisons [3, 22].

In this paper, we introduce **SpaceVLM-DRC**, which incorporates **D**ynamic **R**epulsion with **C**ontext Anchoring into the SpaceVLM framework as a training-free, inference-time negation resolution mechanism over a frozen CLIP backbone. A lightweight language model first decomposes each query into its affirmative, negated, and counterfactual components. Dynamic repulsion then adaptively pushes negated concepts away in the image embedding space, with exclusion scaled to negation intensity. Context anchoring finally enriches the affirmative signal with full-caption context, preserving semantic coherence throughout retrieval. Our main contributions are as follows:

- **Training-free negation resolution:** We propose SpaceVLM-DRC, a frozen-backbone framework that resolves negation purely at inference time, requiring no gradient updates, synthetic corpora, or weight modifications.
- **Inference-time decomposition and dynamic control:** We introduce a query decomposition pipeline that separates affirmative, negated, and counterfactual components and dynamically scales repulsion and exclusion regions based on negation strength.
- **Strong empirical performance with zero-shot preservation:** SpaceVLM-DRC surpasses state-of-the-art negation-aware baselines on MSR-VTT negation retrieval, matches fine-tuned models on COCO negated retrieval, and retains zero-shot generalization on non-negated queries, all without modifying model parameters.

2. Related Work

2.1. Vision-Language Pretraining and CLIP

Large-scale contrastive vision-language pretraining has made CLIP-style models a standard backbone for image-text alignment in tasks, including retrieval [24], captioning [16], depth estimation [13], and multimodal reasoning [2]. These models learn a joint embedding space from internet-scale image-text pairs [2, 24] and support zero-shot transfer via simple text prompts [24]. However, pretraining corpora are typically uncurated and do not capture challenging linguistic phenomena such as negation and counterfactuals [3, 9]. As subsequent VLMs continue to rely on CLIP-like encoders as scoring or guidance components [16], their inability to handle negation descriptions is largely inherited from the original contrastive pretraining objective [3, 24].

2.2. Compositional and Training-Free Image-Text Retrieval

A parallel line of work improves compositional generalization in image-text retrieval without fine-tuning the underlying encoder. CIReVL demonstrates that training-free composition of text embeddings, combined with lightweight re-ranking, can significantly improve retrieval for complex multi-attribute queries [12]. SpaceVLM extends this idea by treating CLIP's embedding space geometrically, introducing angular composition operators that construct query directions from multiple semantic anchors while enforcing separation constraints [25]. These methods demonstrate that meaningful gains in compositional retrieval can be achieved using geometric operations on frozen CLIP embeddings alone [12, 25]. However, neither approach explicitly models the semantics of negation or the asymmetry between concepts that are present and those that are explicitly absent in a scene—a gap our work directly addresses.

2.3. Negation-Aware Vision-Language Models

Recent work has shown that standard VLMs, including CLIP [24], often fail in captions containing explicit negation, such as "no dog on the couch" or description of absent objects [3, 11, 26, 32]. This failure stems from an affirmative bias where models behave like "bags of words," prioritizing nouns while effectively ignoring negation operators [3, 11, 26]. Benchmarks such as CC-Neg [26], NegBench [3], and NegRefCOCOg [22] have been developed to expose these vulnerabilities, revealing that CLIP's performance drops sharply when distinguishing "X" from "without X". Model-side approaches such as Con-CLIP [26] and NegationCLIP [22] incorporate negation-aware training objectives or utilize LLM-generated synthetic data to improve the recognition of absent concepts, although often at the cost of additional fine-tuning.

In specialized domains such as medical imaging [14], paired image-text data and techniques such as dynamic soft labels or graph embeddings have been used to train models to distinguish normal findings from negative pathologies. More recently, Vu and Sheshappanavar [27] showed that contrastive fine-tuning of the text encoder alone can improve negation understanding in vision-language models, increasing negation retrieval accuracy by up to 15% in chest radiograph data. Although many methods require access to training data and gradients, newer alternatives offer more flexible solutions: SpaceVLM [25] models negation as a semantic subspace rather than a single point, NEAT [7] employs parameter-efficient test-time adaptation of normalization layers, and NEGTOME [11] addresses the structural loss of negation cues through semantic token merging. Finally, CLIPGLASSES [29] provides a non-intrusive, plug-and-play framework that uses a Lens module to disentangle negated semantics and a Frame module to apply context-aware repulsion, penalizing alignment with negated content without modifying the base model's weight.

2.4. Positioning of Our Approach

Our work sits at the intersection of training-free compositional retrieval and negation-aware vision-language. Unlike fine-tuning-based methods, we introduce no new learnable parameters and do not require training data specific to negation. Unlike prior geometric methods, we explicitly model the asymmetry between affirmative and negated concepts by conditioning dynamic exclusion regions on the LLM-predicted negation strength. Unlike heuristic plug-and-play approaches, we apply a principled image-side repulsion penalty that preserves the frozen CLIP encoder's zero-shot generalization. Together, these components form a unified test-time scoring framework that narrows the gap between fully trained negation-aware models and purely geometric training-free baselines.

3. Methodology

Our method is based on the geometric subspace modeling of SpaceVLM [25] and the negation-sensitive scoring of CLIPGLASSES [29]. We propose *SpaceVLM-DRC* (**D**ynamic **R**epulsion with **C**ontext anchoring), a training-free inference-time framework for negation-aware image retrieval. It extends CLIP-based similarity scoring without modifying any pretrained model parameters. As shown in Fig. 2, an input caption is simultaneously processed by two LLM modules-a decomposer and a negation-strength estimator-whose outputs are combined with CLIP encodings to construct negation-aware query directions. These directions are matched against image embeddings via cosine similarity, and an image-space repulsion penalty is subtracted to yield the final score.

3.1. Embedding Space and Base Retrieval

Let $\mathcal{I} = \{I_j\}_{j=1}^{N}$ denote the image corpus and $\mathcal{C} = \{c_i\}_{i=1}^{M}$ the set of textual captions. We use a frozen CLIP model [24] with image and text encoders $(\Phi_{\text{img}}, \Phi_{\text{txt}})$ to project both modalities into a shared D-dimensional ℓ_2-normalized embedding space:

$$e_I(I_j) = \frac{\Phi_{\text{img}}(I_j)}{\|\Phi_{\text{img}}(I_j)\|_2}, \qquad e_T(c_i) = \frac{\Phi_{\text{txt}}(c_i)}{\|\Phi_{\text{txt}}(c_i)\|_2} \tag{1}$$

Since both embeddings are ℓ_2-normalized, as shown in equation 1, the cosine similarity reduces to a dot product, giving the standard CLIP retrieval score:

$$S_{\text{CLIP}}(I_j, c_i) = e_I(I_j)^\top e_T(c_i) \tag{2}$$

For captions in which the LLM detects no negation, S_{CLIP} is used directly as the final score, and the remainder of the

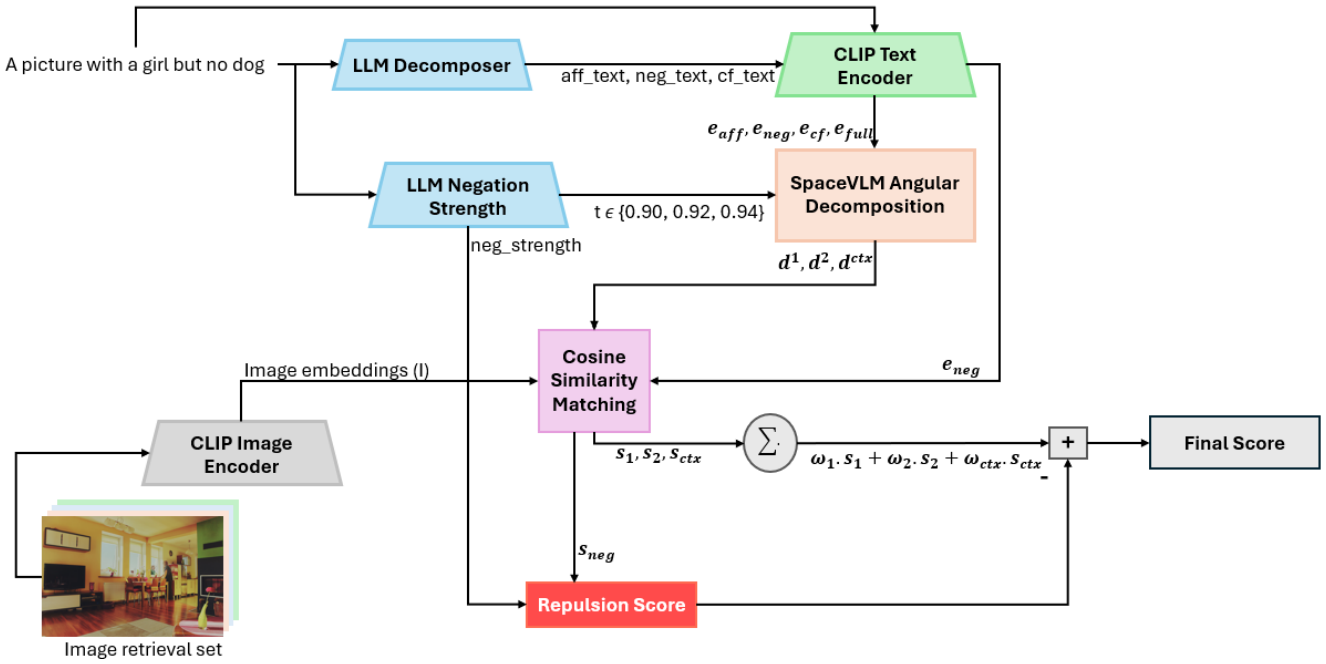

Figure 2. Architecture of SpaceVLM-DRC. An input caption is processed by a parallel LLM decomposer and negation-strength estimator; their outputs are encoded by CLIP and composed into negation-aware query directions via SpaceVLM angular decomposition. Cosine similarity with image embeddings and an image-space repulsion penalty are combined to produce the final retrieval score.

pipeline is skipped. Only captions with a confirmed negation structure proceed through the stages below.

3.2. LLM-Based Query Decomposition and Negation-Strength Estimation

Negation tokens such as "no", "not", and "without" are routinely underweighed by contrastive vision-language models. To expose this structure explicitly, we process each caption c with two independent prompts issued to the frozen instruction-tuned model `Qwen2.5-14B-Instruct` [31].

Query Decomposition: The first prompt decomposes c into a structured triplet:

$$(\text{aff},\ \text{neg},\ \text{cf}) = c \tag{3}$$

where aff is a short phrase for affirmative that describes what is present in the scene, neg is negation or the explicitly absent concept, and cf is a counterfactual phrase describing the scene as if the negations were removed (i.e. aff and neg co-occurring). For the caption:

"A picture with a girl but no dog"

The decomposition yields:

$$\begin{aligned}\text{aff} &= \text{"a picture with a girl"}\\ \text{neg} &= \text{"dog"}\\ \text{cf} &= \text{"a picture with a girl and a dog"}\end{aligned}$$

If the LLM returns a plain string rather than a structured triplet (aff, neg, cf) and the caption proceeds through the remaining pipeline. Otherwise, the caption is scored with S_{CLIP} directly as defined in equation 2.

Negation-Strength Estimation: The second prompt classifies the linguistic strength of the negation:

$$\sigma(c) \in \{\text{strong},\ \text{moderate},\ \text{weak}\} \tag{4}$$

where *strong* covers explicit markers ("no," "without"), *moderate* covers hedged constructions ("doesn't have," "lacks"), *weak* covers implicit absences ("appears absent").

Each of these classes is assigned a cosine threshold t and a base repulsion weight ω_{base}:

$$(t,\ \omega_{\text{base}}) = \begin{cases}(0.90,\ 0.30) & \sigma = \text{strong}\\ (0.92,\ 0.20) & \sigma = \text{moderate}\\ (0.94,\ 0.10) & \sigma = \text{weak}\end{cases} \tag{5}$$

A lower t widens the angular exclusion zone (described in SpaceVLM [25]), applying greater geometric pressure away

from the negated concept. The per-sample effective repulsion weight is $\omega_{\text{rep}} = \omega_{\text{base}} \times \rho$, where ρ is a repulsion scale hyperparameter (default $\rho = 1.5$). This dynamic mapping extends the fixed threshold of SpaceVLM [25] and approximates the trained adaptive λ of CLIPGLASSES [29] without any fine-tuning.

3.3. CLIP Text Encoding

Once decomposition is confirmed, all four text components — aff, neg, cf, and the original caption c — are prefixed with *"A photo of"* and independently encoded by the frozen CLIP text encoder (from equation 1) to yield four unit-normalized embeddings (one for each component):

$$e_{\text{aff}}, e_{\text{neg}}, e_{\text{cf}}, e_{\text{full}} = e_T(\text{aff}, \text{neg}, \text{cf}, \text{c}) \tag{6}$$

Here, e_{full} encodes the complete original caption and serves as a **context anchor** that retains richer scene semantics beyond the short decomposed phrase. These four embeddings, together with the threshold t from the negation-strength estimator, are passed to the angular decomposition module.

3.4. Angular Decomposition (SpaceVLM)

To construct negation-aware query directions, we apply the angular composition operator of SpaceVLM [25]. For two unit vectors a and b and the target cosine threshold t, we define:

$$\alpha = \arccos(t), \qquad \theta = \arccos\big(a^\top b\big) \tag{7}$$

where α is the angular exclusion radius determined by the negation strength threshold t, and θ is the angle between unit vectors. The composed direction is:

$$\hat{d}(a, b, t) = \frac{\sin\big(\alpha + \frac{\theta}{2}\big)}{\sin\theta}\, a + \frac{\sin\big(\alpha - \frac{\theta}{2}\big)}{\sin\theta}\, b \tag{8}$$

The second coefficient is naturally negative when $\alpha < \theta/2$, geometrically repelling the composed vector away from b. The result is ℓ_2 normalized. Using the four CLIP embeddings and threshold t, three complementary query directions are derived:

$$\begin{aligned} \hat{d}_1 &= \hat{d}(e_{\text{aff}}, e_{\text{neg}}, t) \\ \hat{d}_2 &= \hat{d}(e_{\text{aff}}, e_{\text{cf}}, t) \\ \hat{d}_{\text{ctx}} &= \hat{d}(e_{\text{full}}, e_{\text{neg}}, t) \end{aligned} \tag{9}$$

$\hat{d}_1$ is the core query direction, pulled toward the affirmative concept and pushed away from the negated one. $\hat{d}_2$ adds a contrastive signal by repelling the query from the counterfactual, penalizing scenes where both concepts co-occur. $\hat{d}_{\text{ctx}}$ uses the full caption as the affirmative anchor instead of the short decomposed phrase, preserving richer semantics—a strategy inspired by the lens component of CLIPGLASSES [29].

3.5. Image Encoding, Cosine Similarity Matching, and Final Scoring

Image Encoding: In parallel with the text processing pipeline, candidate images from the retrieval set are encoded once by the frozen CLIP image encoder and ℓ_2 normalized to produce image embeddings $\{e_I(I_j)\}$.

Cosine Similarity Matching: Each image embedding is matched against the three composed query directions and against e_{neg}:

$$s_k(j) = e_I(I_j)^\top \hat{d}_k, \quad k \in \{1,\ 2,\ \text{ctx}\} \tag{10}$$

where $s_k(j)$ measures how well the image I_j aligns with each composed query direction from Equation 9. In parallel, the image is also matched directly against the negation concept embedding:

$$s_{\text{neg}}(j) = e_I(I_j)^\top e_{\text{neg}} \tag{11}$$

where $s_{\text{neg}}(j)$ measures the visual similarity between image I_j and the negated concept. A high $s_{\text{neg}}(j)$ indicates the image visually contains the absent concept and should be penalized.

Repulsion Score: While the angular directions reshape the query in text space, s_{neg} from equation 11, directly measures whether a candidate image visually resembles the negated concept. Images with high s_{neg} should be suppressed regardless of their alignment with the composed directions. We therefore define the repulsion penalty as follows:

$$\text{Repulsion}(j) = -\,\omega_{\text{rep}} \cdot \max(s_{\text{neg}}(j),\ 0) \tag{12}$$

where $\omega_{\text{rep}} = \omega_{\text{base}} \times \rho$ is the effective repulsion weight, determined by the negation strength class σ and the repulsion scale hyperparameter ρ. The $\max(\cdot, 0)$ term ensures that only images with positive similarity to the negated concept are penalized.

Final Score: The sum of weighted directional scores representing the angular alignment and the repulsion penalty representing the image-space penalty are combined in the summation block as shown in equation 13:

$$\begin{aligned} S_{\text{DRC}}(I_j, c) = \underbrace{w_1\, s_1(j) + w_2\, s_2(j) + w_{\text{ctx}}\, s_{\text{ctx}}(j)}_{\text{angular alignment}} \\ + \underbrace{\text{Repulsion}(j)}_{\text{image-space penalty}} \end{aligned} \tag{13}$$

where

$$(w_1,\ w_2,\ w_{\text{ctx}}) = (0.35,\ 0.35,\ 0.30) \tag{14}$$

The angular alignment term governs *query geometry* – directing the search toward the intended scene – while the repulsion term governs *image geometry* – suppressing candidates that visually contain the absent concept. The images are ranked in descending order of the S_{DRC} score.

3.6. Inference Pipeline

Our SpaceVLM-DRC is training-free, and all components remain frozen at inference. Our three-step pipeline is:

- **Image feature extraction:** The CLIP image embeddings are computed from 1, and ℓ_2 normalized for all images in the corpus once, prior to any query processing.
- **Caption processing:** Each caption is simultaneously passed to the LLM Decomposer and the LLM Negation Strength estimator. If no negation is detected, the caption is scored with standard S_{CLIP}. Otherwise, the decomposed triplet $(\text{aff},\ \text{neg},\ \text{cf})$ and the original caption c are encoded by the CLIP text encoder as in equation 6. The dynamic threshold t and the repulsion weight ω_{rep} are retrieved from the strength mapping (from equation 5).
- **Negation-aware scoring:** The SpaceVLM angular decomposition module computes $\hat{d}_1$, $\hat{d}_2$, $\hat{d}_{\text{ctx}}$ from the CLIP text embeddings and the dynamic threshold t (equation 9). The cosine similarity scores s_1, s_2, s_{ctx} (equation 10), and s_{neg} (equation 11) are computed against all image embeddings. The repulsion penalty is subtracted from the weighted sum to obtain S_{DRC} (equation 13), which is then used to rank all candidate images.

4. Experimental Results

Datasets: We evaluated our SpaceVLM-DRC on two negation-focused retrieval benchmarks. **COCO Retrieval-Neg** [3] is derived from the MS-COCO validation set [19]. Its captions are modified to include explicit negation constraints of the form "A but not B," requiring models to retrieve images that satisfy the affirmative condition while excluding the negated concept. **MSRVTT Retrieval-Neg** [3] is a negation-augmented variant of the MSR-VTT text-to-video retrieval benchmark [30], where queries are negated descriptions of video content. Both benchmarks assess whether a model can handle presence-absence asymmetry in natural-language queries.

Baselines: We compare SpaceVLM-DRC with the following baselines:

- **CLIP Baseline** [24]**:** Standard frozen CLIP with cosine similarity scoring, representing the unmodified contrastive baseline, which often exhibits an affirmative bias.
- **NEAT** [7]**:** A negation-aware test-time adaptation method that efficiently adjusts distribution-related parameters (specifically normalization layers) during inference to tackle dual-concept shifts.
- **SpaceVLM** [25]**:** A training-free geometric framework that models negation as a semantic subspace rather than a single point, deriving a representative direction from intersecting concept regions in a frozen CLIP encoder.
- **CLIP + SpaceVLM-DRC (Ours):** Our methodology incorporates the subspace modeling of SpaceVLM [25] and the dual-module architecture of CLIPGLASSES [29], which utilizes a Lens module for semantic disentanglement and a Frame module for context-aware repulsion.

Metrics: We report Recall@K (R@1, R@5, R@10) and median rank (medR) metrics for COCO Retrieval-Neg, and R@1, R@5, R@10 metrics for MSRVTT Retrieval-Neg. For Recall@K, higher values indicate better retrieval performance (↑); for medR, lower values are better (↓).

4.1. Implementation Details

All experiments use a frozen ViT-B/32 CLIP backbone. Caption decomposition and negation strength estimation are performed using `Qwen2.5-14B-Instruct` [31], which remains frozen throughout. Gradient updates are not performed at any stage. Angular alignment weights are set to $(w_1, w_2, w_{\text{ctx}}) = (0.35, 0.35, 0.30)$ with $\rho = 1.5$ as the default. All experiments are conducted on 8 NVIDIA H100 GPUs using data-parallel distributed evaluation, with image embeddings broadcast once across all workers and per-caption metrics aggregated at the end.

Although SpaceVLM-DRC does not require model training, the two LLM modules—the query decomposer and the negation-strength estimator—introduce additional per-query inference latency compared to CLIP [24] and SpaceVLM [25], which employs only a single LLM module for query decomposition. This overhead is inherent in the dual-module pipeline and is most pronounced when using larger instruction-tuned models such as `Qwen2.5-14B-Instruct` [31]. As demonstrated in SpaceVLM [25], substituting a lighter model such as TinyLlama-1B achieves a favorable accuracy-latency trade-off, and our pipeline is fully compatible with such alternatives. Alternatively, merging the two LLM prompts into a single call is a straightforward path to reducing the gap. For the retrieval settings targeted in this work (medical image search and content moderation), batch offline processing makes this latency acceptable in practice. Reducing inference overhead through lighter decomposition or unified prompting remains a practical direction for future work.

Table 1. COCO Retrieval-Neg Performance. For R@K, higher values (↑) and for medR, lower values (↓) are better. Best result in bold; parentheses show improvements over the CLIP baseline.

Method	R@1 ↑	R@5 ↑	R@10 ↑	medR ↓
CLIP Baseline [24]	25.0	47.9	59.1	6.0
CLIP + NEAT [7]	30.0 (↑5.0)	54.6 (↑6.7)	65.6 (↑6.5)	–
CLIP + SpaceVLM [25]	29.9 (↑4.3)	**55.1** (↑**7.2**)	**66.4** (↑**7.3**)	**4.0**
CLIP + Ours	**30.2** (↑**5.2**)	54.9 (↑7.0)	65.5 (↑6.4)	**4.0**

COCO Retrieval-Neg: Table 1 reports the results of the COCO negation retrieval benchmark. The CLIP baseline achieves an R@1 of 25.0, confirming the well-documented inability of standard contrastive models to handle negated queries. NEAT, a test-time adaptation method that adjusts

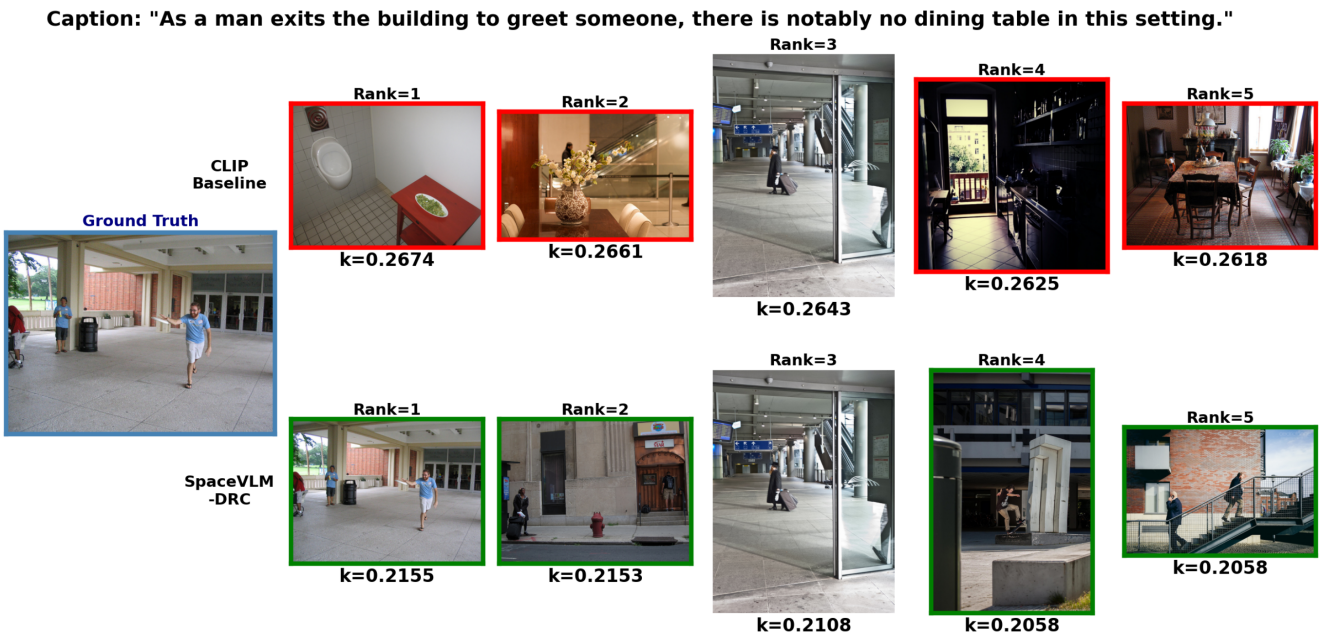

Figure 3. **Qualitative retrieval results on a strong-negation caption.** The query caption reads: *"As a man exits the building to greet someone, there is notably no dining table in this setting."* The ground truth image (blue border, left) depicts a man walking outside a building entrance. **Row 1 – CLIP Baseline** (red borders): CLIP retrieves indoor scenes prominently featuring dining tables, failing to suppress the negated concept. **Row 2 – SpaceVLM-DRC** (green borders): SpaceVLM-DRC successfully retrieves the ground truth at Rank 1, with the remaining results depicting outdoor building and entrance scenes, confirming correct negation handling. Rank 3 (no border) is shared by both methods, the only result in which both models agree. Overall, SpaceVLM-DRC demonstrates a clear improvement over CLIP by replacing semantically incorrect retrievals with contextually appropriate results.

Table 2. MSRVTT Retrieval-Neg Performance. For R@K, higher values (↑) are better. Best results are in bold; parentheses show improvements over the CLIP baseline.

Method	**R@1** ↑	**R@5** ↑	**R@10** ↑
CLIP Baseline [24]	23.8	45.9	56.6
CLIP + NEAT [7]	24.8 (↑1.0)	47.6 (↑1.7)	58.1 (↑1.5)
CLIP + SpaceVLM [25]	26.1 (↑2.3)	49.4 (↑3.5)	63.1 (↑6.5)
CLIP + Ours	**28.9** (↑5.1)	**53.0** (↑7.1)	**63.2** (↑6.6)

model behavior at inference time without modifying pre-trained weights, improves R@1 to 30.0 (+5.0). SpaceVLM, operating purely geometrically on frozen CLIP embeddings, reaches an R@1 of 29.9 (+4.3), demonstrating that the training-free geometric composition is competitive with adaptation-based approaches. Our method, **SpaceVLM-DRC**, achieves an R@1 of **30.2** (+5.2), matching NEAT's performance while also attaining a median rank of **4.0**—on par with the best-performing baselines—and doing so without any parameter updates or test-time adaptation overhead.

MSRVTT Retrieval-Neg: Table 2 reports the results of the MSRVTT negation retrieval benchmark. The CLIP baseline scores an R@1 of 23.8, with NEAT providing a modest improvement to 24.8 (+1.0). SpaceVLM improves further to 26.1 (+2.3). **SpaceVLM-DRC** achieves the strongest results across all metrics, with an R@1 of **28.9** (+5.1), R@5 of **53.0** (+7.1), and R@10 of **63.2** (+6.6)— surpassing all baselines, including the test-time adaptation method NEAT, and doing so entirely through inference-time geometric and linguistic operations on a frozen CLIP encoder.

Summary: Across both benchmarks, SpaceVLM-DRC consistently outperforms all baselines for training-free and test-time adaptation. The gains are particularly pronounced on MSRVTT, suggesting that our dynamic repulsion with context-anchored mechanisms generalizes well to video-language negation settings. In particular, SpaceVLM-DRC achieves these results without training data, gradient updates, test-time adaptation, or synthetic corpus construction.

4.2. Qualitative Analysis

Figure 3 presents a qualitative retrieval example for a caption containing strong negation. The CLIP baseline does not retrieve the correct image within the top 5; instead, it returns indoor scenes (except Rank 3) that prominently feature dining tables, which is negated. In contrast, SpaceVLM-DRC successfully retrieves the ground-truth image at Rank 1, with all five results depicting outdoor entrance settings consistent with the caption's affirmative content. In particular,

the Rank 3 result is shared by both methods, representing the only point of agreement between CLIP and SpaceVLM-DRC. This improvement is attributed to the interplay of three components: (i) the *context anchor*, which uses the full caption embedding as an affirmative anchor to capture richer scene-level semantics; (ii) the *dynamic threshold*, which calibrates the exclusion zone according to the detected strong-negation strength, and (iii) the *direct repulsion* term, which further penalizes images whose embeddings align with the negated concept.

Table 3. Component-wise ablation on COCO Retrieval-Neg. One novel component is added to each row. Best results in bold.

Method	R@1 ↑	R@5 ↑	R@10 ↑	medR ↓
CLIP Baseline	25.0	47.9	59.1	6.0
$+\hat{d}_1$ only	27.2	50.7	61.7	5.0
$+\hat{d}_1 + \hat{d}_2$	27.5	51.1	62.1	5.0
$+\hat{d}_1 + \hat{d}_2 + \hat{d}_{ctx}$	29.1	53.2	64.3	5.0
$+\hat{d}_1 + \hat{d}_2 + \hat{d}_{ctx}$+Rep.	**30.2**	**54.9**	**65.5**	**4.0**

4.3. Ablation Study

We conducted two ablation studies on the COCO Retrieval-Neg benchmark to analyze the contribution of each component of SpaceVLM-DRC: (1) a component-wise ablation that progressively adds each module and (2) a sweep over the repulsion scale hyperparameter ρ.

Table 4. Ablation study on the repulsion scale hyperparameter ρ on COCO Retrieval-Neg. Best results in bold.

Method	R@1 ↑	R@5 ↑	R@10 ↑	medR ↓
$\rho = 0.0$ (no repulsion)	27.7	51.7	62.7	5.0
$\rho = 0.5$	29.0	53.2	64.3	5.0
$\rho = 1.0$	29.9	54.4	65.2	4.0
$\rho = 1.5$ (default)	**30.2**	**54.9**	**65.5**	**4.0**
$\rho = 2.0$	29.7	54.4	65.3	4.0

Table 3 reports the effect of progressively adding each component of SpaceVLM-DRC on top of the CLIP baseline. Starting from vanilla CLIP, we first introduce the standard SpaceVLM direction $\hat{d}_1$ (affirmative vs. negated), then add the counterfactual direction $\hat{d}_2$ (affirmative vs. counterfactual), then add the context anchor direction $\hat{d}_{\text{ctx}}$ (full caption vs. negated), and finally add the image-space repulsion penalty. All configurations use the dynamic threshold unless otherwise stated.

The effective repulsion weight per-sample is $\omega_{rep} = \omega_{base} \times \rho$, where ω_{base} is determined by the negation-strength estimator, and ρ is a repulsion scale hyperparameter. Table 4 sweeps ρ over $\{0.0, 0.5, 0.1, 1.5, 2.0\}$, where $\rho = 0.0$ completely disables the repulsion penalty and reduces the three angular alignment scores.

5. Conclusion and Future Work

In this paper, we present SpaceVLM-DRC, a training-free, negation-aware retrieval framework that operates entirely at inference time on a frozen CLIP backbone. Our novel method combines LLM-based query decomposition, dynamic control of negation strength, and a geometric image-space repulsion mechanism to reshape similarity scores without gradient updates or synthetic corpus construction. Empirically, this design narrows the gap to fine-tuned negation-aware models on COCO Retrieval-Neg and outperforms both geometric and test-time adaptation baselines on MSRVTT Retrieval-Neg, while preserving zero-shot performance on non-negated queries—showing that robust negation handling can be achieved without retraining the underlying vision–language encoder.

Future work will extend this inference-time framework to stronger VLM backbones and video–text encoders. We also plan to incorporate additional linguistic operators, such as conjunctions, disjunctions, and quantifiers, alongside negation. On the efficiency side, we aim to reduce the per-query latency introduced by the dual LLM modules—either by unifying the decomposer and negation-strength estimator into a single prompt, or by replacing the larger `Qwen2.5-14B-Instruct` [31] model with a lighter alternative such as TinyLlama-1B, which SpaceVLM [25] has shown to offer a strong accuracy-latency tradeoff. Finally, future efforts will focus on developing larger and more diverse benchmarks that capture open-world, long-tail visual concepts and multi-sentence instructions.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 2

[3] Kumail Alhamoud, Shaden Alshammari, Yonglong Tian, Guohao Li, Philip HS Torr, Yoon Kim, and Marzyeh Ghassemi. Vision-language models do not understand negation. In *CVPR*, pages 29612–29622, 2025. 2, 3, 6

[4] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *Euro-*

pean conference on computer vision, pages 1–21. Springer, 2022. 2

[5] Abraham Michael Fowler. *Negation in natural language processing*. The University of Texas at Dallas, 2006. 2

[6] Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, and Jianfeng Gao. Vision-language pre-training: Basics, recent advances, and future trends. *arXiv preprint arXiv:2210.09263*, 2022. 2

[7] Haochen Han, Alex Jinpeng Wang, Fangming Liu, and Jun Zhu. Negation-aware test-time adaptation for vision-language models. *arXiv preprint arXiv:2507.19064*, 2025. 2, 3, 6, 7

[8] Iryna Hartsock and Ghulam Rasool. Vision-language models for medical report generation and visual question answering: a review. *Frontiers in Artificial Intelligence*, Volume 7 - 2024, 2024. 2

[9] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in neural information processing systems*, 36:31096–31116, 2023. 2

[10] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916. PMLR, 2021. 2

[11] Inha Kang, Youngsun Lim, Seonho Lee, Jiho Choi, Junsuk Choe, and Hyunjung Shim. What" not" to detect: Negation-aware vlms via structured reasoning and token merging. *arXiv preprint arXiv:2510.13232*, 2025. 3

[12] Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. Vision-by-language for training-free compositional image retrieval. *arXiv preprint arXiv:2310.09291*, 2023. 3

[13] Nischal Khanal and Shivanand Venkanna Sheshappanavar. Edadepth: Enhanced data augmentation for monocular depth estimation. In *2024 International Conference on Machine Learning and Applications (ICMLA)*, pages 620–627. IEEE, 2024. 2

[14] Hanbin Ko and Chang-Min Park. Bringing clip to the clinic: Dynamic soft labels and negation-aware learning for medical analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25897–25906, 2025. 3

[15] Tony Lee, Haoqin Tu, Chi H Wong, Wenhao Zheng, Yiyang Zhou, Yifan Mai, Josselin S Roberts, Michihiro Yasunaga, Huaxiu Yao, Cihang Xie, et al. Vhelm: A holistic evaluation of vision language models. *Advances in Neural Information Processing Systems*, 37:140632–140666, 2024. 2

[16] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900. PMLR, 2022. 2

[17] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pages 19730–19742. PMLR, 2023. 2

[18] Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. A survey of state of the art large vision language models: Benchmark evaluations and challenges. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1587–1606, 2025. 2

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 6

[20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 2

[21] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 2

[22] Junsung Park, Jungbeom Lee, Jongyoon Song, Sangwon Yu, Dahuin Jung, and Sungroh Yoon. Know" no" better: A data-driven approach for enhancing negation awareness in clip. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2825–2835, 2025. 2, 3

[23] Vincent Quantmeyer, Pablo Mosteiro, and Albert Gatt. How and where does clip process negation? In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 59–72, 2024. 2

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 6, 7

[25] Sepehr Kazemi Ranjbar, Kumail Alhamoud, and Marzyeh Ghassemi. Spacevlm: Sub-space modeling of negation in vision-language models. *arXiv preprint arXiv:2511.12331*, 2025. 2, 3, 4, 5, 6, 7, 8

[26] Jaisidh Singh, Ishaan Shrivastava, Mayank Vatsa, Richa Singh, and Aparna Bharati. Learn" no" to say" yes" better: Improving vision-language models via negations. *arXiv preprint arXiv:2403.20312*, 2024. 3

[27] Jasmine Vu and Shivanand Venkanna Sheshappanavar. Improving negation understanding in medical vision-language models via contrastive fine-tuning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 395–404, 2026. 3

[28] Ziyue Wang, Aozhu Chen, Fan Hu, and Xirong Li. Learn to understand negation in video retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 434–443, 2022. 2

[29] Junhao Xiao, Zhiyu Wu, Hao Lin, Yi Chen, Yahui Liu, Xiaoran Zhao, Zixu Wang, and Zejiang He. Not just what's there: Enabling clip to comprehend negated visual descriptions without fine-tuning. *arXiv preprint arXiv:2602.21035*, 2026. 3, 5, 6

[30] Jun Xu, Tao Mei, Ting Yao, and Yong 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*, pages 5288–5296, 2016. 6

[31] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng

Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 4, 6, 8

[32] Yuhui Zhang, Yuchang Su, Yiming Liu, and Serena Yeung-Levy. Negvqa: Can vision language models understand negation? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3707–3716, 2025. 3