Being Positive about Negative Queries:
Exclusion Aware Multimodal Retrieval using Disentangled Representations

Prachi Jha[1], Sumit Bhatia[2], and Srikanta Bedathur[1]
[1]Indian Institute of Technology Delhi, India
[2]Media and Data Science Research Lab, Adobe Systems, India
prachi@cse.iitd.ac.in, sumit.bhatia@adobe.com, srikanta@cse.iitd.ac.in

Abstract

The handling of exclusion in multimodal retrieval remains an underexplored challenge with significant implications for the accuracy and reliability of information retrieval systems. Although existing approaches have advanced multimodal understanding, they typically lack mechanisms to explicitly process exclusion. To address this, we propose a novel model ExclMM (pronounced as "exclaim") that leverages disentangled representations to effectively handle exclusion in multimodal retrieval. Our approach enables precise differentiation between the presence and absence of specific elements in an image, outperforming existing methods. To evaluate our model, we construct a dataset, ExcluCOCO that pairs exclusion-based queries with ground-truth images sourced from MSCOCO. This dataset serves as a robust benchmark for assessing exclusion comprehension in multimodal contexts. By explicitly incorporating exclusion, our work advances multimodal retrieval by introducing both a model tailored for exclusion-aware retrieval and a benchmark to facilitate future research.

1. Introduction

Exclusion is a fundamental aspect of human language and cognition, allowing us to specify not just what is present in a scene but also what is absent. It plays a crucial role in how we communicate and reason about the world, enabling us to express exclusion, contradiction, and prohibition. In natural language processing (NLP), exclusion based query handling is increasingly explored due to its utility in tasks such as sentiment analysis, question answering, and textual entailment [28, 31]. Although multimodal retrieval models aim to retrieve relevant images given a text query, they inherently involve challenges in aligning textual and visual semantics. Exclusion adds an additional layer of complexity, as it requires models not only to identify relevant content but also to recognize and explicitly account for the absence of certain elements in an image. For instance, a query such as "*Find the images of a room without a TV*" demands that the retrieval system not only recognize the concept of a *room* but also ensure the explicit absence of a *TV*. This is fundamentally different from standard retrieval tasks, where models primarily focus on detecting the presence of objects rather than reasoning about their exclusion. Most existing retrieval approaches, particularly those based on deep embeddings, struggle with such queries because they rely on holistic representations in which semantic attributes are entangled. This entanglement makes it challenging to explicitly distinguish, suppress, or exclude specific concepts [1]. Addressing this limitation is essential for developing models capable of nuanced, exclusion-aware reasoning.

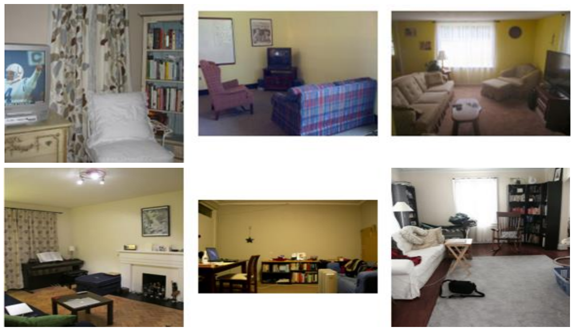


Figure 1. Top-3 retrieved images for the query "room without a TV" using CLIP (top row) and ExclMM (bottom row) embeddings. CLIP fails to retrieve images that do not contain "TV" object, while ExclMM is able to (detailed evaluation in Section 5).

To address this challenge, we propose a disentanglement-based multimodal retrieval model, ExclMM, that enables explicit reasoning about the presence and absence of objects in images. Disentangled representations aim to separate factors of variation in data, providing greater control and interpretability [26]. We use textual captions as weak supervision signal to guide the disentanglement of corresponding image representations, ensuring that different sets of dimensions capture distinct semantic components shared between the image and text. By structuring the representation in this way, we achieve fine-grained control over individual attributes, allowing targeted modifications. We leverage this property for exclusion-aware retrieval by training the model to suppress or attenuate the dimensions corresponding to excluded features, ensuring that retrieval aligns with the intended query semantics. For example, for the query "room without a TV", traditional models like CLIP [22] tend to retrieve images containing both a room and a TV, as they mainly focus on the presence of objects rather than exclusion (Fig. 1, top row). In contrast, our approach explicitly suppresses the dimensions associated with the TV while preserving the representation of the room, ensuring that the retrieved images align with the intended exclusion constraint (Fig. 1, bottom row).

ExclMM learns sparse, disentangled multimodal embeddings by enforcing structured factorization of information across different latent dimensions. Unlike traditional models [14, 22], with tightly coupled textual and visual features in a dense embedding space, our approach explicitly separates them, allowing selective manipulation of query semantics. Specifically, for a given query, the model decomposes its representation into distinct feature components, making it possible to selectively suppress dimensions corresponding to excluded concepts while preserving relevant ones.

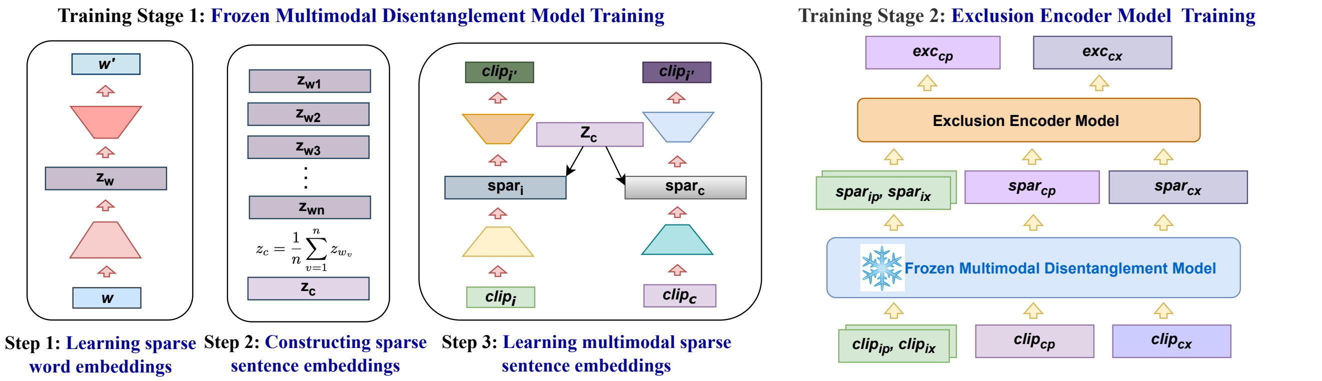
To evaluate our model, we built ExcluCOCO, a benchmark dataset specifically designed for exclusion-aware multimodal retrieval. This dataset comprises exclusion-based queries paired with corresponding ground truth images, leveraging existing labeled images from MSCOCO [16]. Unlike traditional datasets such as MSCOCO, which mainly focus on retrieving images based on the presence of objects, our dataset ensures that models are explicitly assessed on their ability to exclude specified elements, making it a valuable resource for studying exclusion in retrieval tasks. While the recently proposed CC-Neg dataset [23] provides exclusion-based queries, it lacks ground truth images, making it unsuitable for evaluating retrieval performance. In summary, our work makes the following key contributions:

- A disentanglement-based retrieval model ExclMM, that explicitly separates visual and textual features, allowing fine-grained control over embeddings to effectively handle exclusion queries. By structuring representations in a way that enables selective suppression of negated attributes, our approach improves retrieval accuracy for exclusion-aware queries.
- A benchmark dataset ExcluCOCO, to support rigorous evaluation of exclusion-aware multimodal retrieval models. This dataset provides ground truth for exclusion-based retrieval tasks along with labels for included and excluded items, offering a valuable resource for future research in this underexplored area.

2. Related Work

Exclusion handling is a relatively new research area, with numerous emerging studies spanning natural language processing (NLP), computer vision, video retrieval, and multimodal systems. In NLP, several works have focused on exclusion-aware textual retrieval and language modeling. The NevIR benchmark [28] evaluates neural retrieval models' ability to process negation by using contrastive document pairs that differ only in negation. Similarly, ExcluIR [31] introduces training and evaluation datasets designed to enhance retrieval models' performance on exclusionary queries. Early studies [5, 17, 18] explored the challenges associated with negation, often treating user queries as logical expressions of Boolean operations to model exclusion. In image retrieval, [29] discuss the difficulties posed by exclusion and introduces a framework for evaluating exclusion-based queries in keyword-driven image retrieval. Recent efforts in video retrieval have also tackled the complexities of exclusion and negative queries. For example, [2] proposed the NA-VMR framework to filter out irrelevant queries, while [27] leveraged soft negative captions to enhance CLIP's ability to handle negated queries. Multimodal foundation models have also been assessed for their effectiveness in processing exclusion. Research by [25] has shown that while instruction tuning and scaling model size provide some improvements, negation comprehension remains a persistent challenge. Addressing this issue, [34] systematically evaluates the susceptibility of state-of-the-art multimodal large language models to negation-based gaslighting. Furthermore, [23] introduced a dataset consisting of positive image-caption pairs along with their negated captions and proposes CoN-CLIP, a framework that modifies the negated caption embeddings to be distinct from both the positive caption embeddings and the corresponding image embeddings.

While some progress has been made in exclusion-aware retrieval, it remains an emerging area, particularly in multimodal contexts. Many existing methods struggle with explicitly handling exclusion, highlighting the need for structured approaches to improve exclusion comprehension. Additionally, the lack of well-curated datasets with explicit ground truth for exclusion queries poses a major challenge, underscoring the importance of developing better benchmarks for training and evaluation. One promising direction for improving exclusion-aware retrieval is the use of disentangled representations, which have been widely studied for their ability to separate underlying factors of variation in data. Unlike traditional deep embeddings [14, 22], where features are entangled and difficult to manipulate, disentangled representations allow for explicit control over individual attributes [1, 4, 26]. This property makes them particularly suited for exclusion-aware retrieval, where the goal is to ensure that retrieval models not only capture rel-


Training Stage 1: Frozen Multimodal Disentanglement Model Training
w'
z_w
w
z_w1
z_w2
z_w3
z_wn
$z_c = \frac{1}{n}\sum_{v=1}^{n} z_{w_v}$
z_c
clip_i'
clip_i'
Z_c
spar_i
spar_c
clip_i
clip_c
Step 1: Learning sparse word embeddings
Step 2: Constructing sparse sentence embeddings
Step 3: Learning multimodal sparse sentence embeddings
Training Stage 2: Exclusion Encoder Model Training
exc_cp
exc_cx
Exclusion Encoder Model
spar_ip, spar_ix
spar_cp
spar_cx
Frozen Multimodal Disentanglement Model
clip_ip, clip_ix
clip_cp
clip_cx

(a) Three step process to create a multimodal disentangled representation model. In step 1, d-dimensional sparse word embeddings are generated from their m-dimensional pretrained embeddings for all the words in the vocabulary. In step 2, these sparse word embeddings are aggregated to obtain d-dimensional sentence embeddings for each query c. In step 3, a biencoder-decoder model is trained to produce d-dimensional sparse image and text embeddings by taking their pretrained embeddings as input and using the sparse sentence embeddings from step 2 as mask to enforce dimension level disentanglement.

(b) The multimodal disentangled representation model from stage 1 is used as a frozen model to generate sparse embeddings. The pretrained embeddings of all images, positive queries, and exclusion queries are first passed through the frozen model to obtain their sparse embeddings. These embeddings are then processed by the Exclusion Encoder model to generate refined representations for exclusion queries.

Figure 2. Illustration of the two-stage training process of ExclMM. In Stage 1, a multimodal disentangled model is trained on CLIP embeddings of images and their text captions. In Stage 2, this trained model is used as a frozen model to generate sparse, disentangled embeddings for positive and exclusion queries along with their ground-truth images. These embeddings are then processed by an Exclusion Encoder to create refined embeddings for exclusion queries.

evant attributes but also systematically suppress excluded ones. Early works such as β-VAE [6], FactorVAE [9], and Relevance FactorVAE [10] primarily focused on disentangling representations in images. Some studies have extended these techniques to multimodal settings [11, 13], creating disentangled representations for both images and text. However, these approaches are often restricted to synthetic or relatively simple datasets with a limited number of factors of variation. More recent work has explored disentanglement in real-world multimodal datasets for retrieval [32]. Additionally, disentangled representations have been leveraged for attribute-conditioned retrieval and conditional similarity search [7], recommendation systems [15], and image-to-image translation [12], demonstrating their broader applicability.

3. Methodology

We propose a two-stage training approach for generating representations tailored to exclusion-based queries, as illustrated in Fig. 2. In the first stage, we develop a bi-encoder-decoder model that produces sparse, disentangled embeddings for both images and text. These embeddings are structured such that similar dimensions are activated with high values for semantically related concepts. In the second stage, we freeze the disentanglement model and train an encoder that operates on the sparse text query embeddings generated by the disentanglement model. This encoder learns to construct representations for exclusion queries, enabling the retrieval of images that explicitly satisfy the exclusion constraint. We now describe these steps in detail.

Stage 1: Multimodal Disentangled Model training

This training stage follows a three-step pipeline to learn sparse and disentangled multimodal embeddings.

Step 1: Learning Sparse Word Embeddings We first obtain sparse and disentangled word embeddings for all the words in the train vocabulary $\mathcal{V}$ by transforming pre-trained m-dimensional word embeddings, such as GloVe [21] (which captures semantic relationships between words), into a higher-dimensional space ($d > m$) using a sparse autoencoder, following the same approach as followed by [24]. Given a word embedding $\mathbf{w} \in \mathbb{R}^m$, we pass it through a sparse autoencoder [19], defined as:

$$\mathbf{z}_w = f_{\text{enc}}(\mathbf{w}; \theta_{\text{enc}}), \quad \hat{\mathbf{w}} = f_{\text{dec}}(\mathbf{z}_w; \theta_{\text{dec}}) \tag{1}$$

where $\mathbf{z}_w \in \mathbb{R}^d$ is the latent representation of the word, that is encouraged to become sparse by the training objective (i.e., sparsity is induced via optimization rather than an *a priori* property), and $\hat{\mathbf{w}}$ is the reconstructed word embedding. The model is trained using following loss functions.

1. **Reconstruction loss(RL)** is the average loss in reconstructing the input representation from learned representation and reconstructed word embeddings.

$$RL = \frac{1}{\mathcal{V}} \sum_{v \in \mathcal{V}} \|\hat{\mathbf{w}}_v - \mathbf{w}_v\|_2^2 \tag{2}$$

We adopt an L2 loss as it provides stable gradients and better preserves the geometry of the original embeddings. In contrast, an L1 loss made the representations excessively sparse and caused loss of important semantic information.

2. **Average sparsity Loss(ASL)** penalizes deviations of the observed average activation value from the target activation

value for a given hidden unit across a dataset.

$$ASL = \sum_{h \in H} \max(0, (\rho_{h,\mathcal{V}} - \rho^*_{h,\mathcal{V}}))^2 \tag{3}$$

where $\rho^*_{h,\mathcal{V}}$ denotes the *desired* sparsity level for hidden unit h, and $\rho_{h,\mathcal{V}}$ is the *actual* sparsity, computed as the average activation of unit h in all words of the vocabulary $\mathcal{V}$.
3. **Partial Sparsity Loss(PSL)** penalizes the values that are neither close to 0 nor 1 and pushes them close to 0 and 1, adding more sparsity to the embeddings.

$$PSL = \frac{1}{\mathcal{V}} \sum_{v \in \mathcal{V}} \sum_{h \in H} (z^h_{w_v} * (1 - z^h_{w_v})) \tag{4}$$

where $\mathcal{H}$ refers to the set of hidden units in the latent layer. Finally, the d-dimensional latent representations $\mathbf{z}_w$ serve as the sparse disentangled word embeddings. The representations obtained using this sparse autoencoder method exhibit inherent interpretability and disentanglement at the dimension level. Similar approach has also been explored in several recent works [8, 20].
Step 2: Constructing Sentence Embeddings Once we get the d-dimensional embeddings for all words in the vocabulary, we construct sentence embeddings for captions by computing the weighted average of the word embeddings for a given image caption $c = (w_1, w_2, \dots, w_n)$:

$$z_c = \frac{1}{n} \sum_{v=1}^{n} z_{w_v} \tag{5}$$

This ensures that the sentence embeddings retain the interpretability and disentangled properties of the sparse word embeddings.
Step 3: Learning Multimodal Sparse Embeddings We employ a bi-encoder-decoder architecture to project both image and text representations into a shared sparse latent space. Given an image i and its corresponding caption c, we first obtain their k-dimensional CLIP embeddings, denoted as $clip_i$ and $clip_c$. These embeddings are then projected into a d-dimensional space using separate encoders(f^{enc}_{text} and f^{enc}_{img}) and subsequently reconstructed back to k dimensions using decoders(f^{dec}_{text} and f^{dec}_{img}).

$$E^d_i = f^{enc}_{\text{img}}\left(clip_i\right), \quad clip_{i'} = f^{dec}_{\text{img}}\left(E^d_i\right) \tag{6}$$

$$E^d_c = f^{enc}_{\text{text}}\left(clip_c\right), \quad clip_{c'} = f^{dec}_{\text{text}}\left(E^d_c\right) \tag{7}$$

To enforce sparsity and disentanglement, we create d-dimensional masks($mask_i$ and $mask_c$) (similar to that used in [3, 32]) that combines the top t active dimensions of the image/text embeddings(E^d_i and E^d_c) using the active dimensions of the sentence embedding z_c. This ensures that the disentangled structure from the sentence embeddings is transferred to the multimodal representations

$$mask_i = z_c \,\text{OR}\, \text{Top}_t\left(E^{\text{d}}_i\right) \tag{8}$$

$$mask_c = z_c \,\text{OR}\, \text{Top}_t\left(E^{\text{d}}_c\right) \tag{9}$$

$spar_c, spar_i \in \mathbb{R}^d$ are the final sparse multimodal representations obtained by element-wise multiplication:

$$spar_i = mask_i \odot E^d_i \tag{10}$$

$$spar_c = mask_c \odot E^d_c \tag{11}$$

To further align image and text embeddings, we optimize a softmax-based contrastive loss.:

$$\mathcal{L}_{\text{contrast}} = -\frac{1}{2N} \left(\sum_{a=1}^{N} \log \frac{\exp(spar^a_c \cdot spar^a_i/\tau)}{\sum_{b=1}^{N} \exp(spar^a_c \cdot spar^b_i/\tau)} + \sum_{a=1}^{N} \log \frac{\exp(spar^a_i \cdot spar^a_c/\tau)}{\sum_{b=1}^{N} \exp(spar^a_i \cdot spar^q_b/\tau)} \right) \tag{12}$$

where $spar^a_i$ and $spar^a_c$ denote the ℓ_2-normalized sparse image and text embeddings respectively for the a^{th} sample in a batch of size N, and τ is a temperature scaling parameter (taken as 1 in our case).

The final training objective combines this contrastive loss with pair of reconstruction losses between initial and reconstructed clip embeddings for image and text:

$$\mathcal{L}^c_{\text{rec}} = \|clip_{c'} - clip_c\|^2_2, \quad \mathcal{L}^i_{\text{rec}} = \|clip_{i'} - clip_i\|^2_2 \tag{13}$$

The final loss is given by:

$$\mathcal{L} = \mathcal{L}^c_{\text{rec}} + \mathcal{L}^i_{\text{rec}} + \mathcal{L}_{\text{contrast}}$$

This formulation ensures that the learned embeddings are not only sparse and disentangled but also well-aligned across modalities, leading to interpretable multimodal representations (see Section 5 for examples of dimension-level disentanglement). Sparse embeddings provide the disentangled structure needed for reliable exclusion reasoning. By expanding the dimensionality and enforcing sparsity, Stage 1 assigns different semantic concepts to distinct, minimally overlapping axes, unlike dense CLIP embeddings where concepts remain entangled. This factorization allows exclusion to be implemented by simply suppressing the coordinates associated with the excluded concept. Our ablations confirm that removing sparsity or relying solely on dense embeddings leads to mixed activations and unreliable exclusion performance.

Stage 2: Training the Exclusion Query Encoder

After obtaining sparse, disentangled multimodal embeddings from the frozen model, we introduce an exclusion-aware encoder to refine representations for exclusion queries. Specifically, it takes the embeddings of positive queries, exclusion queries, and their corresponding ground-truth images and optimizes them to generate distinct embeddings for positive and exclusion based queries.

Input Notation. Using the frozen model, we generate the following embeddings:
$spar_{cp}$: Disentangled Embedding of the **positive query**.
$spar_{cx}$: Disentangled Embedding of the **exclusion query**.
$spar_{ip}$: Disentangled Embedding of the **positive image**.
$spar_{ix}$: Disentangled Embedding of the **exclusion image**.

Encoder Training: The encoder $\mathcal{F}_\theta$ takes the disentangled text embeddings generated from the frozen model embeddings as input and generates refined embeddings for the positive and exclusion queries as:

$$exc_{cp} = \mathcal{F}_\theta(spar_{cp}), \quad exc_{cx} = \mathcal{F}_\theta(spar_{cx}) \tag{14}$$

Loss Function: The training proceeds with the following two components of loss function,
1. **Softmax-based Contrastive Loss:** The model learns to align exclusion queries with their corresponding exclusion images while ensuring separation from other images. The contrastive loss similar to that used in the Stage 1 training is used between embeddings of exclusion queries generated from the exclusion encoder(exc_{cx}) and sparse embeddings of their corresponding ground truth images generated from the frozen disentanglement model($spar_{ix}$):

$$\begin{aligned}\mathcal{L}_{\text{c}} = -\frac{1}{2Q}\Bigg(&\sum_{a=1}^{Q} \log \frac{\exp(exc_{cx}^{a} \cdot spar_{ix}^{a}/\tau)}{\sum_{b=1}^{Q} \exp(exc_{cx}^{a} \cdot spar_{ix}^{b}/\tau)} \\ &+ \sum_{a=1}^{Q} \log \frac{\exp(spar_{ix}^{a} \cdot exc_{cx}^{a}/\tau)}{\sum_{b=1}^{Q} \exp(spar_{ix}^{a} \cdot exc_{cx}^{b}/\tau)}\Bigg)\end{aligned} \tag{15}$$

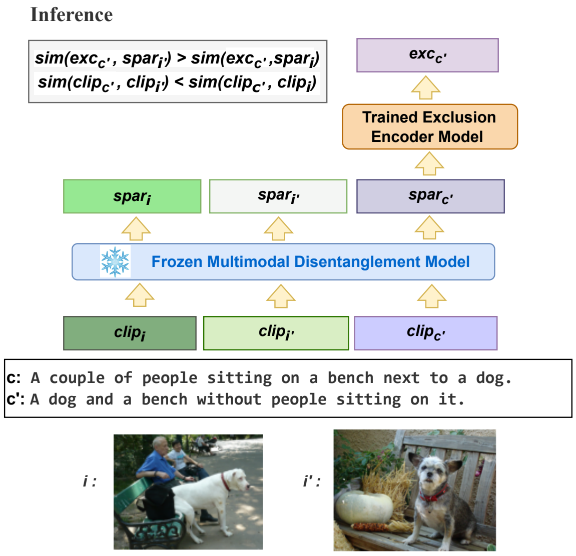
where exc_{cx}^{a} and $spar_{ix}^{a}$ denote the ℓ_2-normalized exclusion query and image embeddings respectively for the a^{th} query in a batch of size Q, and τ is the temperature parameter (set to 1 in this case).
2. **Triplet Loss for Positive and Exclusion Queries:** This loss ensures that a positive image embeddings $spar_{ip}$ is distant to exclusion query embedding exc_{cx} compared to the positive query embedding exc_{cp} and is given by:

$$\begin{aligned}\mathcal{L}_{\text{t}} = \max(0, &\, d(spar_{ip}, exc_{cp}) \\ &- d(spar_{ip}, exc_{cx}) + m)\end{aligned} \tag{16}$$

where $d(a, b)$ denotes the distance between a and b in the embedding space, computed using cosine distance. The margin m defines a minimum separation between the distances that encourage meaningful separation while allowing gradient flow. The overall loss function is a weighted sum of the contrastive and triplet losses:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{c}} + \lambda_2 \mathcal{L}_{\text{t}} \tag{17}$$

where λ_1, λ_2, are hyperparameters balancing the loss terms. The final loss ensures that there is proper alignment between the exclusion query-image pairs and the exclusion queries are also further from the positive images in the embedding space.


Inference
sim(exc_c', spar_i') > sim(exc_c',spar_i)
sim(clip_c', clip_i') < sim(clip_c', clip_i)
exc_c'
Trained Exclusion Encoder Model
spar_i
spar_i'
spar_c'
Frozen Multimodal Disentanglement Model
clip_i
clip_i'
clip_c'
c: A couple of people sitting on a bench next to a dog.
c': A dog and a bench without people sitting on it.
i :
i' :

During inference, the CLIP embeddings of the exclusion query c', positive image i, and exclusion image i'(ground truth for c') are first passed through the frozen model to obtain their disentangled embeddings $\text{spar}_{c'}$, spar_{i}, and $\text{spar}_{i'}$. $\text{spar}_{c'}$ is then processed by the exclusion-based encoder to generate the final exclusion embedding $\text{exc}_{c'}$.

Figure 3. Illustration of the inference process of ExclMM, where the CLIP embeddings of the exclusion query($clip_{c'}$) and the positive and exclusion ground-truth images($clip_i$,$clip_{i'}$) are passed to the frozen model to obtain their sparse disentangled embeddings ($spar_{c'}$,$spar_i$ and $spar_{i'}$ respectively). The exclusion query embedding($spar_{c'}$) is then passed through the trained Exclusion Encoder to generate a refined exclusion embedding($exc_{c'}$).Results show that while $clip_{c'}$ is initially closer to $clip_i$ than to $clip_{i'}$, after processing through the model, $exc_{c'}$ becomes closer to $spar_{i'}$ than to $spar_i$ demonstrating that the exclusion query embedding effectively captures the intended representation.

Inference

As illustrated in Fig. 3, the inference process begins with an exclusion query, which is first converted into its CLIP embedding. This embedding is then passed through the frozen model to obtain a sparse representation. To refine it further, the trained exclusion encoder processes the sparse embedding, generating the final query representation:

$$exc_{query} = \mathcal{F}_\theta(spar_{query}) \tag{18}$$

This embedding can then be used for retrieval, ensuring proper alignment between exclusion queries and their corresponding images.

4. Dataset Construction

We introduce **ExcluCOCO**, a new multimodal dataset designed to support exclusion-based retrieval. This dataset is derived from the well-established multimodal benchmark **MSCOCO**[16]. Our dataset construction methodology ensures that each *exclusion-based query* is paired with a set of

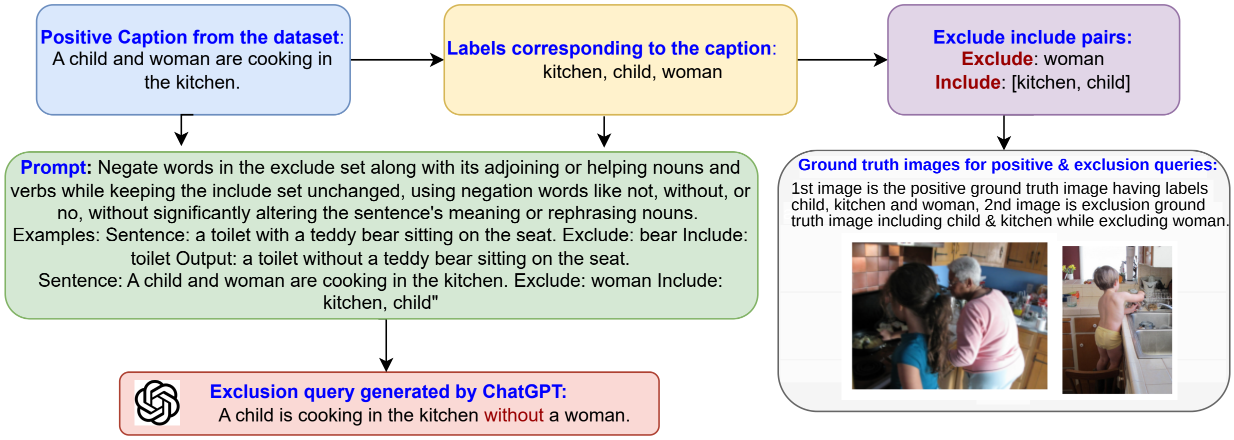

Positive Caption from the dataset: A child and woman are cooking in the kitchen.
Labels corresponding to the caption: kitchen, child, woman
Exclude include pairs:
Exclude: woman
Include: [kitchen, child]
Prompt: Negate words in the exclude set along with its adjoining or helping nouns and verbs while keeping the include set unchanged, using negation words like not, without, or no, without significantly altering the sentence's meaning or rephrasing nouns.
Examples: Sentence: a toilet with a teddy bear sitting on the seat. Exclude: bear Include: toilet Output: a toilet without a teddy bear sitting on the seat.
Sentence: A child and woman are cooking in the kitchen. Exclude: woman Include: kitchen, child"
Ground truth images for positive & exclusion queries:
1st image is the positive ground truth image having labels child, kitchen and woman, 2nd image is exclusion ground truth image including child & kitchen while excluding woman.
Exclusion query generated by ChatGPT:
A child is cooking in the kitchen without a woman.

Figure 4. Overview of ExcluCOCO dataset construction process, where captions from MSCOCO are used to form inclusion-exclusion pairs, modified via ChatGPT, and paired with ground truth images that satisfy the exclusion constraints.

ground truth images that adhere to the specified inclusion and exclusion constraints.

4.1. Generating Exclusion Queries

For **MSCOCO**, we leverage the text captions associated with each image to construct *exclusion-based queries*. The process involves the following steps:

1. **Label Extraction:** Each image in MSCOCO dataset is associated with a set of labels corresponding to the objects present in the image. We extend this label set by extracting frequently occurring words from the captions, ensuring a comprehensive representation of image content. A dictionary is created by mapping each image to its corresponding label set.

2. **Candidate Label Selection:** From each caption, we identify words present in the label set. For instance, given the caption:*"A child and woman are cooking in the kitchen"*, the extracted labels are *women*, *kitchen*, and *child*. From this list, we select one label to exclude and the remaining labels to include, forming an *inclusion-exclusion pair*. Multiple such combinations are generated for each caption.

3. **Ground Truth Image Sampling:** For each inclusion-exclusion pair, we retrieve candidate images from the dataset. An image is considered a valid ground truth for a given query if it contains all the labels in the inclusion set and it does not contain the label in the exclusion set. We retain only the pairs for which such ground truth images are available in the dataset. Since the sampling of the label pairs is conditioned only on the availability of ground-truth images that contain the included object(s) and exclude the specified object(s), the class distribution in ExcluCOCO naturally mirrors that of MSCOCO.

4.2. Constructing Exclusion Queries

Once we have the positive query (caption), the inclusion-exclusion label pairs, and the corresponding ground truth images, we generate the *exclusion queries* using ChatGPT. The overall dataset construction process is illustrated with an example in Figure 4. We follow a prompt engineering based approach for query generation in which the original caption is provided to ChatGPT along with the corresponding *inclusion-exclusion labels*. A structured prompt instructs ChatGPT to generate a exclusion query that retains the same structure as the original caption, includes the words from the inclusion list and excludes the word from the exclusion list, using appropriate negation constructs (e.g., 'without', 'missing', 'except for','but not'). Few-shot prompting is employed, providing multiple handcrafted examples to improve the quality of generated queries.

This systematic approach ensures that ExcluCOCO effectively captures the semantics of exclusion in a multimodal setting while maintaining high-quality ground truth annotations. As a result, we obtain 26,932 exclusion query-image pairs from the MSCOCO training set and 3,009 pairs from the test set. On average, queries involve 1–2 included objects and 1 excluded object, and query lengths are comparable to MSCOCO captions (on average 10 words). Figure 5 presents a few examples from the constructed dataset. In the first example, the positive query is "A person feeding a cat with a banana". The first image in the row (highlighted in a green box) is the positive ground truth image, depicting a person's hand feeding a banana to a cat. The corresponding exclusion query is "A cat with a banana but not a person feeding it". The ground truth images for this exclusion query are shown in the purple box above it, where each image contains a cat and a banana but no person, demonstrating that the person is the excluded concept while cat and banana are included. Similarly, in the other example, the included-excluded pairs are surfboard (included) and person (excluded) respectively. The ground truth images conform to these inclusion-exclusion constraints, ensuring that

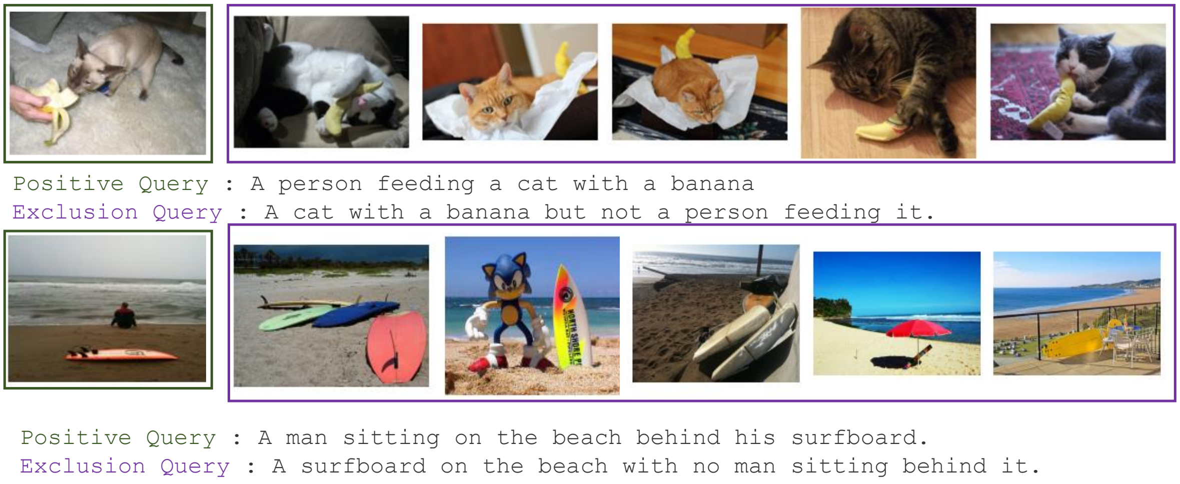

Positive Query : A person feeding a cat with a banana
Exclusion Query : A cat with a banana but not a person feeding it.
Positive Query : A man sitting on the beach behind his surfboard.
Exclusion Query : A surfboard on the beach with no man sitting behind it.

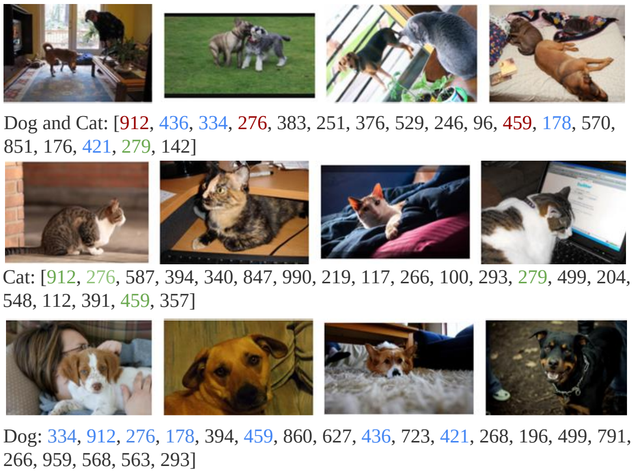
Figure 5. Examples from ExcluCOCO dataset, the first image of each row corresponds to the image for the Positive query while the rest of the images are the ground truth images for the Exclusion query. The Positive & Exclusion queries are given below the images in each row

the excluded concept is absent while the included concepts remain.

5. Experiments

5.1. Quantitative Evaluation

We evaluate our model's performance on exclusion-based retrieval tasks using exclusion queries from the ExcluCOCO dataset. We compare it against CoN-CLIP [23], an exclusion-based retrieval model that creates separate embeddings for exclusion queries, as well as widely used vision-language representation models such as CLIP [22], BLIP [14], SIGLIP [30] and the universal multimodal retrieval model, VISTA [33]. The results in Table 1 show that our model outperforms all baselines, with CoN-CLIP performing particularly poorly. This is because CoN-CLIP is trained to generate separate embeddings for positive and exclusion queries, positioning exclusion query embeddings farther from their corresponding positive query ground truth than from the positive queries of these images. However, it does not explicitly align exclusion query embeddings with the correct exclusion-ground-truth images, leading to suboptimal retrieval performance. Additionally, the results demonstrate that despite their effectiveness in traditional text-image retrieval, multimodal models such as CLIP, BLIP, SiIGLIP and VISTA struggle with exclusion queries. Although the dataset contains only one exclusion per query, our model is capable of handling more than one exclusion as well.


Dog and Cat: [912, 436, 334, 276, 383, 251, 376, 529, 246, 96, 459, 178, 570, 851, 176, 421, 279, 142]

Cat: [912, 276, 587, 394, 340, 847, 990, 219, 117, 266, 100, 293, 279, 499, 204, 548, 112, 391, 459, 357]

Dog: 334, 912, 276, 178, 394, 459, 860, 627, 436, 723, 421, 268, 196, 499, 791, 266, 959, 568, 563, 293]

Figure 6. Top-4 images retrieved for each query using disentangled image and text embeddings, considering only the top 10% active dimensions. Active dimension lists are shown alongside queries: common dimensions between "Dog and Cat" and "Dog" are in green, those shared with "Cat" are in blue, and dimensions active in all three are marked in red. This illustrates dimension-level disentanglement across modalities and concepts.

5.2. Ablation Study

We conduct an ablation study on both loss functions and Stage 1 design choices, with results in Table 2. Using only the triplet loss leads to near-collapse, since queries are pushed away from positives but not aligned with exclusion targets. Using only contrastive loss also hurts performance, showing the two losses are complementary: contrastive aligns queries to exclusion images, while triplet sharpens separation from exclusion-relevant negatives. For architecture, removing the zero-concept mask (z_c) reduces performance, confirming its role in preserving included concepts while suppressing excluded ones. Skipping Stage 1 and applying Stage 2 directly to CLIP embeddings corresponds to our CLIP-based ablations: frozen CLIP embeddings perform worse than our full method, while fine-tuning CLIP on exclusion queries gives AP@1 ≈ 0.017 and joint fine-tuning collapses further. These results show that naive fine-

Table 1. Results for exclusion query to image retrieval. We report numbers for CoN-CLIP, VISTA, BLIP, CLIP, and ExclMM. Statistically significant improvements over CoN-CLIP, VISTA, BLIP, CLIP, SIGLIP are indicated by superscripts 0, 1, 2, 3 and 4, respectively (measured by paired t-Test with 99% confidence)

Method	**AP@1**	**AP@5**	**MRR@5**	**MRR@10**	**NDCG@5**	**NDCG@10**	**Hits@5**	**Hits@10**
CoN-CLIP	0.0588	0.0580	0.0919	0.1032	0.0586	0.0603	0.1517	0.2380
VISTA	0.2346^{0}	0.2784^{0}	0.3889^{0}	$0.4080^{0,2}$	$0.2760^{0,2}$	$0.2895^{0,2,3}$	$0.6543^{0,3}$	$0.7940^{0,2,3}$
BLIP	0.1896^{0}	$0.2795^{0,1}$	0.3695^{0}	0.3868^{0}	0.2677^{0}	$0.2815^{0,3}$	$0.6609^{0,1,3}$	$0.7920^{0,3}$
CLIP	$0.2633^{0,1,2}$	$0.2800^{0,1,2}$	$0.4097^{0.1,2}$	$0.4278^{0,1,2}$	$0.2819^{0,1,2}$	0.2814^{0}	0.6500^{0}	0.7814^{0}
SIGLIP	$0.2033^{0,2}$	0.2710^{0}	$0.3721^{0,2}$	$0.3904^{0,2}$	0.2639^{0}	0.2788^{0}	$0.6507^{0,3}$	$0.7853^{0,3}$
ExclMM	$\mathbf{0.5040}^{0,1,2,3,4}$	$\mathbf{0.4372}^{0,1,2,3,4}$	$\mathbf{0.6865}^{0,1,2,3,4}$	$\mathbf{0.6866}^{0,1,2,3,4}$	$\mathbf{0.4651}^{0,1,2,3,4}$	$\mathbf{0.4320}^{0,1,2,3,4}$	$\mathbf{0.9991}^{0,1,2,3,4}$	$\mathbf{1.0}^{0,1,2,3,4}$

Table 2. Ablation with different combination of losses and architectural design. ExclMM with both pair of losses and disentanglement based frozen model shows the best performance.

Method	**AP@1**	**AP@5**	**MRR@5**	**MRR@10**	**NDCG@5**	**NDCG@10**	**Hits@5**	**Hits@10**
ExclMM	**0.5040**	**0.4372**	**0.6865**	**0.6866**	**0.4651**	**0.4320**	**0.9991**	**1.0**
w/o $\mathcal{L}_t$	0.4650	0.4076	0.6205	0.6312	0.4321	0.4153	0.8763	0.9529
w/o $\mathcal{L}_c$	0.0163	0.0170	0.0299	0.0354	0.0169	0.0177	0.0565	0.0983
with CLIP base	0.2633	0.2800	0.4097	0.4278	0.2819	0.2814	0.6500	0.7814
w/o $z_c mask$	0.4313	0.3885	0.5822	0.5956	0.4108	0.4004	0.8245	0.9219
CLIP finetuned on exclusion queries	0.0322	0.0142	0.0434	0.0510	0.0170	0.0161	0.0707	0.1602
CLIP finetuned on exclusion and positive queries	0.0079	0.0118	0.0255	0.0312	0.0115	0.0110	0.0594	0.1023

tuning mixes inclusion and exclusion, while Stage 1 disentanglement creates sparse, concept-aligned dimensions that enable reliable masking and strong performance.

5.3. Disentanglement

The multimodal disentanglement model trained in the first stage generates sparse embeddings for images and text, where disentanglement occurs at the dimension level. This means that specific subsets of dimensions capture similar concepts in image-text data, with these dimensions exhibiting higher activation values. We illustrate this in Figure 6 by taking three queries "Dog", "Cat", and "Dog and Cat" and generating their embeddings using the trained multimodal disentanglement model. We identify the top-10% most active dimensions in each embedding and display them alongside the queries. Next, we retrieve images based on their disentangled embeddings, selecting those that share the same active dimensions. As shown in the figure, the same set of dimensions remains active and highly valued across text and image embeddings, demonstrating alignment between the two modalities at both the overall embedding and individual dimension levels. Furthermore, the embedding of "Dog and Cat" shares active dimensions with the embeddings of "Dog" and "cat". In the figure, the dimensions common to "Cat" and "Dog and Cat" are highlighted in green, those shared between "Dog" and "Dog and Cat" in blue, and dimensions present in all three embeddings in red within the "Dog and Cat" dimension list. This confirms that similar concepts activate similar dimensions across embeddings. Using this property, exclusion queries can be effectively handled by removing the active dimensions associated with the concept to be excluded, demonstrating the power of such disentangled embeddings.

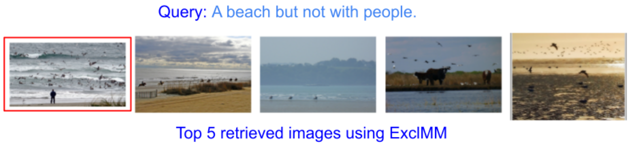

Query: A beach but not with people.
Top 5 retrieved images using ExclMM

Figure 7. Failure case example for ExclMM. It retrieved the image inside the red box, although query excludes people appearing in the image.

6. Conclusion

In this work, we propose a disentanglement-based multimodal retrieval model that explicitly handles exclusion by suppressing features of excluded concepts, allowing fine-grained control over embeddings for inclusion–exclusion queries. We also introduce a new benchmark dataset tailored for exclusion-aware retrieval. Our method advances the state-of-the-art in this setting and highlights exclusion as a crucial step toward human-like multimodal reasoning. Unfortunately, our work is not without limitations – for instance, Figure 7 illustrates a failure mode for ExclMM. Future work can explore extending disentanglement techniques to more complex negation scenarios and generalizing the approach to broader multimodal reasoning tasks.

Acknowledgements

This work was partially supported by the CSE Research Acceleration Fund of IIT Delhi. Srikanta Bedathur acknowledges his DS Chair Professor of AI fellowship at IIT Delhi.

References

[1] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE*

Trans. Pattern Anal. Mach. Intell., 35(8):1798–1828, 2013. 1, 2

[2] Kevin Flanagan, Dima Damen, and Michael Wray. Moment of untruth: Dealing with negative queries in video moment retrieval, 2025. 2

[3] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. Splade: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 2288–2292, New York, NY, USA, 2021. Association for Computing Machinery. 4

[4] Klaus Greff, Sjoerd van Steenkiste, and Jürgen Schmidhuber. On the binding problem in artificial neural networks. *ArXiv*, abs/2012.05208, 2020. 2

[5] Valerie J. Harvey, Jeanne M. Baugh, Bruce Johnston, Constance M. Ruzich, and Arthur J. Grant. The challenge of negation in searches and queries. In *Business Information Systems*, 2011. 2

[6] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017. 3

[7] Yuxin Hou, Eleonora Vig, Michael Donoser, and Loris Bazzani. Learning attribute-driven disentangled representations for interactive fashion retrieval. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12127–12137, 2021. 3

[8] Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024. 4

[9] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *Proceedings of the 35th International Conference on Machine Learning*, pages 2649–2658. PMLR, 2018. 3

[10] Minyoung Kim, Yuting Wang, Pritish Sahu, and Vladimir Pavlovic. Relevance factor vae: Learning and identifying disentangled factors, 2019. 3

[11] Minyoung Kim, Ricardo Guerrero, and Vladimir Pavlovic. Learning disentangled factors from paired data in cross-modal retrieval: An implicit identifiable vae approach. *Proceedings of the 29th ACM International Conference on Multimedia*, 2021. 3

[12] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Computer Vision – ECCV 2018*, pages 36–52, Cham, 2018. Springer International Publishing. 3

[13] Mihee Lee and Vladimir Pavlovic. Private-shared disentangled multimodal vae for learning of latent representations. pages 1692–1700, 2021. 3

[14] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. 2, 7

[15] Zhenyang Li, Fan Liu, Yinwei Wei, Zhiyong Cheng, Liqiang Nie, and Mohan Kankanhalli. Attribute-driven disentangled representation learning for multimodal recommendation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, page 9660–9669, New York, NY, USA, 2024. Association for Computing Machinery. 3

[16] 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 *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. 2, 5

[17] April R. McQuire and Caroline M. Eastman. The ambiguity of negation in natural language queries to information retrieval systems. *J. Am. Soc. Inf. Sci.*, 49(8):686–692, 1998. 2

[18] April R. McQuire and Caroline M. Eastman. The ambiguity of negation in natural language queries to information retrieval systems. *J. Am. Soc. Inf. Sci.*, 49:686–692, 1998. 2

[19] Andrew Ng. Sparse autoencoder. In *Stanford University CS294A Lecture Notes*, 2011. 3

[20] Charles O'Neill, Christine Ye, Kartheik G. Iyer, and John F Wu. Towards interpretable scientific foundation models: Sparse autoencoders for disentangling dense embeddings of scientific concepts. In *Neurips 2024 Workshop Foundation Models for Science: Progress, Opportunities, and Challenges*, 2024. 4

[21] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, 2014. Association for Computational Linguistics. 3

[22] 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, 2021. 2, 7

[23] Jaisidh Singh, Ishaan Shrivastava, Mayank Vatsa, Richa Singh, and Aparna Bharati. Learn ”no” to say ”yes” better: Improving vision-language models via negations, 2024. 2, 7

[24] Anant Subramanian, Danish Pruthi, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Eduard H. Hovy. Spine: Sparse interpretable neural embeddings. *ArXiv*, abs/1711.08792, 2017. 3

[25] Mayank Vatsa, Aparna Bharati, Surbhi Mittal, and Richa Singh. From no to know: Taxonomy, challenges, and opportunities for negation understanding in multimodal foundation models, 2025. 2

[26] Xin Wang, Hong Chen, Si'ao Tang, Zihao Wu, and Wenwu Zhu. Disentangled representation learning, 2024. 1, 2

[27] 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*, page 434–443, New York, NY, USA, 2022. Association for Computing Machinery. 2

[28] Orion Weller, Dawn Lawrie, and Benjamin Van Durme. NevIR: Negation in neural information retrieval. In *Proceedings of the 18th Conference of the European Chapter of the*

Association for Computational Linguistics (Volume 1: Long Papers), pages 2274–2287, St. Julian's, Malta, 2024. Association for Computational Linguistics. 1, 2

[29] Eisaku Yoshikawa and Keishi Tajima. Content-based exclusion queries in keyword-based image retrieval. In *Proceedings of the 2024 International Conference on Multimedia Retrieval, ICMR 2024, Phuket, Thailand, June 10-14, 2024*, pages 1145–1149. ACM, 2024. 2

[30] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023. 7

[31] Wenhao Zhang, Mengqi Zhang, Shiguang Wu, Jiahuan Pei, Zhaochun Ren, Maarten de Rijke, Zhumin Chen, and Pengjie Ren. Excluir: Exclusionary neural information retrieval, 2024. 1, 2

[32] Jiawei Zhou, Xiaoguang Li, Lifeng Shang, Xin Jiang, Qun Liu, and Lei Chen. Retrieval-based disentangled representation learning with natural language supervision, 2024. 3, 4

[33] Junjie Zhou, Zheng Liu, Shitao Xiao, Bo Zhao, and Yongping Xiong. VISTA: Visualized text embedding for universal multi-modal retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3185–3200, Bangkok, Thailand, 2024. Association for Computational Linguistics. 7

[34] Bin Zhu, Hui yan Qi, Yinxuan Gui, Jingjing Chen, Chong-Wah Ngo, and Ee Peng Lim. Calling a spade a heart: Gaslighting multimodal large language models via negation, 2025. 2