Answering Multimodal Exclusion Queries with Lightweight Sparse Disentangled Representations

No Author Given

No Institute Given

Abstract. Multimodal representations that combine data from different modalities —such as text, image, audio, etc.,— in a shared subspace are widely used in tasks such as image and text retrieval. However, these representations often lack interpretability, making it difficult to explain why certain results are retrieved. A typical solution to achieve better interpretability is by learning disentangled sparse multimodal representations by isolating key factors of variation in the data, often using text to guide the process. Unfortunately, this results in very large (as much as the size of the vocabulary) representations —since they treat each word in the vocabulary as a factor contributing to variation— making them slow and expensive to work with. In this paper, we present an approach to address this limitation by generating much smaller fixed-size embeddings that are not only disentangled but also offer fine-grained control for retrieval tasks. Specifically, we demonstrate their utility on a novel multimodal query benchmark consisting of complex queries involving exclusion over MSCOCO and Conceptual Captions datasets. Our experiments show that our approach is superior to traditional dense models such as CLIP and BLIP (gains up to 7% in AP@10), as well as sparse disentangled models like VDR (gains up to 12% in AP@10). We also present qualitative results to further underline the interpretability of disentangled representations.

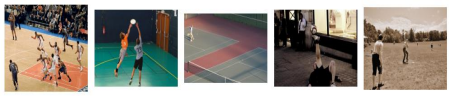
Keywords: Multimodal Retrieval · Disentanglement · Exclusion Queries

1 Introduction

Deep learnt models for learning multimodal representations help to integrate and process information from different types of data, or modalities, such as text, image, audio, video etc. These learnt representations are proven to be invaluable for improving the performance of many downstream tasks including multimodal retrieval. However, while these models can create powerful representations, they often become complex and challenging to interpret. Although some progress has been made, fully understanding and explaining how these models combine different types of information remains an open challenge, limiting our ability to control or generalize the model's behavior in new scenarios. Disentanglement addresses some of these challenges by separating the various underlying factors of variation within the data, thereby enhancing the explainability, interpretability, controllability, and generalizability of the representations [2, 18, 7].

(a) *Sports* using CLIP Embeddings

(c) *Sports* using SR_{clip} Embeddings



(b) *Sports but not Basketball* using CLIP Embeddings

(d) *Sports but not Basketball* using SR_{clip} Embeddings

Fig. 1: **Top retrieved results for simple and exclusion based queries using CLIP and SR_{clip} embeddings**

Although a major challenge in disentanglement is identifying the factors of variation, especially in real-world datasets, where the number of factors is not fixed due to the complexity of the data. One promising approach is to use the text captions associated with images to capture these factors. For instance, works like [21] represent each word or token in the vocabulary by mapping it to a single dimension in the representation. While this approach can capture a large variety of factors, managing such large representations becomes challenging. In our method, we address this issue by developing a model that captures key factors from text captions using significantly smaller representations. Instead of assigning each word or token its own dimension, we use subsets of dimensions to represent similar words and concepts. This approach not only produces disentangled representations by separating concepts based on dimension subsets but also enables more efficient handling of real-world data with lightweight embeddings.

Once we have disentangled embeddings that achieve a degree of factor separation within the datasets, we can leverage this property in retrieval scenarios where such separation is essential. One practical example is a retrieval task where users may not only be looking for specific content but may also want to exclude some other. This ability to control retrieval in such a targeted manner is critical for improving the relevance and accuracy of search results across various domains, such as e-commerce, content moderation, and others where search filtering is required. Disentangled representations, which isolate different concepts and features, make this level of control more achievable. Building on this idea, we introduce a new method to handle the exclusion based retrieval in multimodal scenerio. We explore this approach in greater detail in the following section.

Exclusion Based Retrieval

In typical image retrieval scenarios, the retrieved images generally contain the objects, concepts, things or scenes mentioned in the query. Popular models like BLIP[13] and CLIP have shown great success in handling these straightforward queries, where the goal is simply to retrieve images that match the items described. However, in many real-world situations, queries are more complex, and

users may want to retrieve images containing multiple objects or exclude specific items from the results. One such scenario is exclusion, where the query seeks images of a specific object while explicitly omitting another. This concept has also recently been explored in text retrieval systems, as demonstrated by [20], which addressed exclusion queries in document retrieval by filtering out documents containing specific terms or answers. Similar work has also been done for image web search in [19]

We tackle this issue of exclusion in multimodal datasets like MSCOCO and Conceptual Captions. To better illustrate retrieval for exclusion queries in a multimodal context, we will consider the example shown in Fig. 1. While BLIP and CLIP excel at handling simple queries involving single objects, they struggle with negation-based queries, such as "Images of sports but not basketball." For such complex queries, these models often fail to filter out irrelevant results. Figure 1 illustrates this issue. The first set of images Fig. 1(a) shows the results for the query "sports" using CLIP embeddings. In Fig. 1(b) we see the results for "sports but not basketball" using the same CLIP embeddings, where basketball images are still present. By contrast, Fig. 1(c) displays the results for "sports", and Fig. 1(d) shows the results for "sports but not basketball"using our sparse embeddings. While both CLIP and our representations SR_{clip} retrieve various sports images for the simple query, SR_{clip} embeddings successfully exclude basketball images, yielding more accurate results for the negation query. The key contributions of our paper are:

- **Sparse Disentangled Multimodal Representations:** We develop a method to disentangle factors of variation within multimodal datasets, achieving this with a significantly reduced embedding size compared to traditional approaches.
- **Exclusion Based Retrieval Framework:** We present a new retrieval approach designed to better handle complex queries involving negation and exclusion, where existing models like CLIP and BLIP often face limitations.
- **Exclusion Query Evaluation Dataset:** We release a new dataset for evaluating exclusion queries, by sampling images from the MSCOCO and Conceptual Captions datasets. We quantitatively assess the performance of our model and others on this dataset to benchmark their effectiveness.

2 Related Works

2.1 Disentangled Representations

Disentangled representation learning is inherently linked to interpretability, as these representations facilitate a more understandable data structure by isolating different factors of variation into distinct vectors or dimensions. The primary objective is to make the underlying factors that drive variations in data more transparent and easier to interpret. Early works in image disentanglement, including approaches like Factor VAE ˜[9], β -VAE [8], Relevance Factor VAE [11] predominantly focused on disentanglement in synthetic datasets. These datasets

have fixed and well-defined factors of variation, making it easier to evaluate the success of disentanglement. In these works, VAE-based models are commonly employed, leveraging dense latent embeddings of relatively small size. The dimensions within these latent embeddings are then mapped to the specific factors of variation present in the synthetic datasets. Some studies, such as [15] and [12],[10] explore multimodal disentanglement, extending similar model architectures to scenarios where multiple modalities are involved. In such cases, synthetic datasets with fixed factors of variation and some simpler real world datasets are still favored.[21] These models aim to disentangle the contributing factors across different modalities while maintaining a clear interpretation of the results. More recent research has shifted focus towards real-world datasets, where the factors of variation are not predefined or controlled. For instance, SpLice [3] and VDR [21] have pushed the boundaries of disentangled representation learning by applying these concepts to more complex and less structured real-world data. Unlike earlier approaches, these works utilize representations that match the size of the vocabulary used in the dataset, reflecting a more nuanced and context-dependent understanding of disentanglement. This shift from synthetic to real-world data highlights the evolution of the field towards more practical and scalable applications of disentangled representation learning.

2.2 Sparse Embeddings and Exclusion based Retrieval

Sparse retrieval is a technique that leverages sparse representations for retrieval tasks. Several methods, such as SparTerm [1], SPLADE[6], and SPLADEv2[5], are widely used for text and document retrieval. Recently, this approach has been extended to multimodal retrieval, with methods like VDR[21] and STAIR[4] emerging in this space. A newer paradigm in the retrieval field is exclusion-based retrieval, which focuses on handling exclusion queries. Notable works in this area include Excluir[20] for text retrieval and [19] for image retrieval.

3 Methodology

The architecture of our Sparse Disentangled Model is shown in Fig.3. It follows a binary encoder-decoder framework, where two encoders take pretrained embeddings of images and corresponding texts as input. Our model is compatible with embeddings from multimodal models like CLIP and BLIP, using their pretrained embeddings to generate d-dimensional sparse latent embeddings. To ensure interpretability and disentanglement, we use intrepretable SPINE [17] embeddings of text as a bias. The following sections detail the model's training process.

In **Training step 1**, we generate sparse and interpretable word embeddings for all the words used in the training dataset. For this, we utilize a fully unsupervised model based on [17] which takes the GloVe embeddings of these words and passes them through a sparse autoencoder which imposes a sparsity constraint promoting the learning of more interpretable representations.

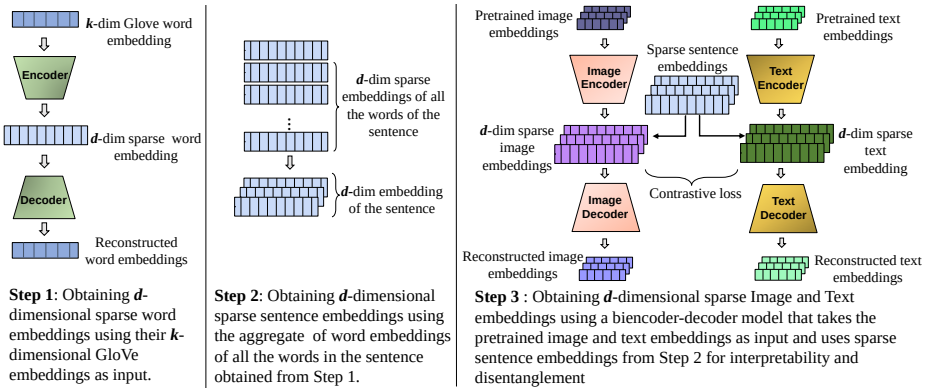


Fig. 2: **Illustration of the three-step training process for generating sparse, disentangled representations. Training step 1 produces interpretable word embeddings for all the words in the vocabulary, which are then used in the second step to create sentence embeddings. In training step 3, a biencoder decoder model is used to create sparse disentangled embeddings of images and texts by using the sentence embeddings created in Step 2 as a bias to activate particular dimensions.**

Let $D = [X_1, X_2, \ldots, X_V] \in \mathbb{R}^{V \times m}$ be the set of m dimensional GloVe embeddings of the vocabulary of size V. We want to project these embeddings from m dimensions to d dimensions, $\mathbb{R}^{V \times m} \rightarrow \mathbb{R}^{V \times d}$ such that the new d dimensional embeddings are sparse. To achieve this, we optimize using three loss functions: a reconstruction loss, along with ASL and PSL losses, which ensure sparsity in the embeddings.

Reconstruction loss : The reconstruction loss (RL) measures the average error in recreating the input representation from the learned embeddings. For an input vector $X \in \mathbb{R}^d$, if the predicted output is $\tilde{X} \in \mathbb{R}^d$, then the loss is computed as:

$$RL(D) = \frac{1}{|D|} \sum_{X \in D} \|X - \tilde{X}\|_2^2$$

Average Sparsity Loss : It penalizes any difference between the observed average activation(ρ) and the target average activation(ρ^*) for each hidden unit across the dataset. The loss is formulated as follows:

$$ASL(D) = \sum_{h \in H} \max\left(0, \rho_{h,D} - \rho^*_{h,D}\right)$$

Partial Sparsity Loss : Partial Sparsity Loss (PSL) is designed to penalize activations that fall between 0 and 1, driving them closer to either extreme. The PSL is formulated as follows:

$$PSL(D) = \frac{1}{|D|} \sum_{X \in D} \sum_{h \in H} Z(X)_h \times (1 - Z(X)_h)$$

The final loss function is given by: $L(D) = RL(D) + \lambda_1 ASL(D) + \lambda_2 PSL(D)$. The sparse embeddings generated by the autoencoder capture the syntactic information of the data, encapsulating the grammatical and structural relationships between words. **Trianing step 2** is used to obtain the sentence embeddings f the sentence captions corresponding to all the images in the training dataset. For this let $S = [w_1, w_2, \ldots, w_n]$ be a sentence consisting of n words, where each word w_i has a corresponding sparse embedding $\mathbf{e}_{w_i}$.

$$\mathbf{e}_S = \frac{1}{n} \sum_{i=1}^{n} \mathbf{e}_{w_i}, \quad \mathbf{e}_S^{\text{norm}} = \frac{\mathbf{e}_S}{\|\mathbf{e}_S\|}$$

where $\mathbf{e}_S^{\text{norm}}$ is the normalized sentence embedding.

These sentence embeddings retain the interpretability of the individual word embeddings, with similar words and features aligned in the same set of dimensions. This ensures the sentence embeddings capture meaningful patterns, making them both syntactically and semantically informative.

In **Training step 3**, we utilize a binary encoder-decoder model, where the inputs for the encoders f_{encoder} are k dimensional pretrained CLIP image embeddings (E_k^{img}) and text embeddings (E_k^{text}). The encoder projects these k-dimensional embeddings into a d-dimensional latent space, and the decoder f_{decoder} projects them back to k-dimensions. The transformations can be represented as:

$$\mathbf{E}_d = f_{\text{encoder}}(\mathbf{E}_k) = \mathbf{E}_k W_e + b_e, \qquad \hat{E}_k = f_{\text{decoder}}(\mathbf{E}_d) = \mathbf{E}_d W_d + b_d$$

To make the latent d-dimensional embeddings sparse and disentangled, we use the sparse disentangled embeddings of the sentence captions as a bias. We apply a mask derived from the sparse embeddings to both the image and text latent d-dimensional embeddings. The masked embeddings are defined as:

$$E_{\text{mask}}^{\text{img}} = e_S^{\text{norm}} \,\text{OR}\, \text{Top}_k(E_d^{\text{img}}), \quad E_{\text{mask}}^{\text{text}} = e_S^{\text{norm}} \,\text{OR}\, \text{Top}_k(E_d^{\text{text}})$$

The sparse representations are then obtained by element-wise multiplication:

$$SR_{\text{img}} = E_{\text{mask}}^{\text{img}} \odot E_d^{\text{img}}, \quad SR_{\text{text}} = E_{\text{mask}}^{\text{text}} \odot E_d^{\text{text}}$$

The loss functions used to optimize the model are:

Reconstruction Loss:

$$RL = \left\| \mathbf{E}_k^{\text{img}} - f_{\text{decoder}}(\mathbf{SR}^{\text{img}}) \right\|_2^2 + \left\| \mathbf{E}_k^{\text{text}} - f_{\text{decoder}}(\mathbf{SR}^{\text{text}}) \right\|_2^2$$

Contrastive Loss:

$$CL = -\frac{1}{2N} \left(\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(\mathbf{SR}_i^{\text{img}}, \mathbf{SR}_i^{\text{text}})/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(\mathbf{SR}_i^{\text{img}}, \mathbf{SR}_j^{\text{text}})/\tau)} \right)$$

The overall loss combines the reconstruction loss and the contrastive loss:

$$L = RL + \lambda \cdot CL$$

The final sparse embedding produced by our model benefits from multiple sources of information. The CLIP embeddings provide rich contextual and semantic content, while the mask incorporates essential semantic and syntactic information from the sparse text embeddings. As a result, the final sparse embedding encapsulates a comprehensive representation of the data, combining contextual, semantic, and syntactic information in a single, interpretable form.

Datasets	Base	Model	MRR@1	MRR@3	MRR@10	NDCG@1	NDCG@3	NDCG@10	AP@1	AP@3	AP@10
MSCOCO		**VDR**	0.7195	0.7797	0.7873	0.7195	0.7259	0.6648	0.7195	0.7315	0.6446
		SpLice	0.0718	0.0964	0.1184	0.0718	0.0602	0.0543	0.0718	0.0591	0.0518
		CBE	0.2960	0.3976	0.4399	0.2960	0.3054	0.3106	0.2960	0.3073	0.3114
	Clip	**SLQ**	$0.6125^{0,1,2}$	$0.6969^{0,1,2}$	$0.7208^{0,1,2}$	$0.6125^{0,1,2}$	$0.5888^{0,1,2}$	$0.5636^{0,1,2}$	$0.6125^{0,1,2}$	$0.5815^{0,1,2}$	$0.5504^{0,1,2}$
		Avg Emb	$0.7981^{0,1,2,3}$	$0.8447^{0,1,2,3}$	$0.8552^{0,1,2,3}$	$0.7981^{0,1,2,3}$	$0.7728^{0,1,2,3}$	$0.7293^{0,1,2,3}$	$0.7981^{0,1,2,3}$	$0.7653^{0,1,2,3}$	$0.7099^{0,1,2,3}$
		SR_{Clip}	$\underline{0.8669}^{0,1,2,3,4}$	$\underline{0.9117}^{0,1,2,3,4}$	$\underline{0.9175}^{0,1,2,3,4}$	$\underline{0.8669}^{0,1,2,3,4}$	$\underline{0.8528}^{0,1,2,3,4}$	$\underline{0.8064}^{0,1,2,3,4}$	$\underline{0.8669}^{0,1,2,3,4}$	$\underline{0.8479}^{0,1,2,3,4}$	$\underline{0.7865}^{0,1,2,3,4}$
	Blip	**SLQ**	$0.7117^{0,1,2}$	$0.8063^{0,1,2}$	$0.8190^{0,1,2}$	$0.7117^{0,1,2}$	$0.7106^{0,1,2}$	$0.6884^{0,1,2}$	$0.7117^{0,1,2}$	$0.7099^{0,1,2}$	$0.6768^{0,1,2}$
		Avg Emb	$0.8376^{0,1,2,5}$	$0.8748^{0,1,2,5}$	$0.8815^{0,1,2,5}$	$0.8376^{0,1,2,5}$	$0.8261^{0,1,2,5}$	$0.7987^{0,1,2,5}$	$0.8376^{0,1,2,5}$	$0.8222^{0,1,2,5}$	$\underline{0.7868}^{0,1,2,5}$
		SR_{Blip}	$\mathbf{0.9226}^{0,1,2,5,6}$	$\mathbf{0.9510}^{0,1,2,5,6}$	$\mathbf{0.9536}^{0,1,2,5,6}$	$\mathbf{0.9226}^{0,1,2,5,6}$	$\mathbf{0.8972}^{0,1,2,5,6}$	$\mathbf{0.8553}^{0,1,2,5,6}$	$\mathbf{0.9226}^{0,1,2,5,6}$	$\mathbf{0.8899}^{0,1,2,5,6}$	$\mathbf{0.8359}^{0,1,2,5,6}$
Conceptual Captions		**VDR**	0.5473	0.6441	0.6687	0.5473	0.5593	0.5536	0.5473	0.5597	0.5512
		SpLice	0.0616	0.0978	0.1271	0.0616	0.0591	0.0564	0.0616	0.0584	0.0553
		CBE	0.2994	0.3870	0.4237	0.2994	0.3027	0.3027	0.2994	0.3035	0.3027
	Clip	**SLQ**	0.5129	0.6066	0.6325	0.5129	0.5001	0.4759	0.5129	0.4966	0.4658
		Avg Emb	$0.6268^{0,1,2,3}$	$0.7182^{0,1,2,3}$	$0.7375^{0,1,2,3}$	$0.6268^{0,1,2,3}$	$0.6250^{0,1,2,3}$	$0.6128^{0,1,2,3}$	$0.6268^{0,1,2,3}$	$0.6243^{0,1,2,3}$	$0.6079^{0,1,2,3}$
		SR_{Clip}	$\underline{0.6749}^{0,1,2,3,4}$	$\underline{0.7530}^{0,1,2,3,4}$	$\mathbf{0.7698}^{0,1,2,3,4}$	$\underline{0.6749}^{0,1,2,3,4}$	$\underline{0.6693}^{0,1,2,3,4}$	$\underline{0.6528}^{0,1,2,3,4}$	$\underline{0.6749}^{0,1,2,3,4}$	$\underline{0.6674}^{0,1,2,3,4}$	$\underline{0.6460}^{0,1,2,3,4}$
	Blip	**SLQ**	$0.5087^{0,1,2}$	$0.6089^{0,1,2}$	$0.6362^{0,1,2}$	$0.5087^{0,1,2}$	$0.5026^{0,1,2}$	$0.4938^{0,1,2}$	$0.5087^{0,1,2}$	$0.5011^{0,1,2}$	$0.4900^{0,1,2}$
		Avg Emb	$\mathbf{0.7028}^{0,1,2,5}$	$\mathbf{0.7537}^{0,1,2,5}$	$\underline{0.7661}^{0,1,2,5}$	$\mathbf{0.7028}^{0,1,2,5}$	$\mathbf{0.6900}^{0,1,2,5}$	$\mathbf{0.6702}^{0,1,2,5}$	$\mathbf{0.7028}^{0,1,2,5}$	$\mathbf{0.6863}^{0,1,2,5}$	$\mathbf{0.6620}^{0,1,2,5}$
		SR_{Blip}	$0.6290^{0,1,2,5,6}$	$0.7149^{0,1,2,5,6}$	$0.7348^{0,1,2,5,6}$	$0.6290^{0,1,2,5,6}$	$0.6104^{0,1,2,5,6}$	$0.5820^{0,1,2,5,6}$	$0.6290^{0,1,2,5,6}$	$0.6048^{0,1,2,5,6}$	$0.5704^{0,1,2,5,6}$

Table 1: **MRR, NDCG and AP scores for MSCOCO and Conceptual Captions Datasets for Controlled Retrieval. Statistically significant improvements over VDR, SpLice, Content-Based Exclusion, CLIP Single line query, CLIP Avg embeddings, BLIP Single line query and BLIP Avg Embeddings are indicated by superscript 0, 1, 2, 3, 4, 5 and 6, respectively(measured by paired t-Test with 99% confidence)**

4 Experiments

4.1 Datasets

We trained our model on the MSCOCO and Conceptual Captions datasets to generate sparse multimodal embeddings and evaluated its performance on the

test set in a standard retrieval setting. To assess exclusion retrieval, we created two new evaluation datasets by sampling images from these datasets. Details of the datasets are provided below.

- **MSCOCO[14]:** The MSCOCO dataset consists of 118K train images and 5K test images, with each image accompanied by 4-5 text captions. These image-text pairs are divided into 80 object categories. For training our model, we have used the image-text pairs from the train set, and for creating the exclusion dataset we have used the images from the test set along with their labels.
- **Conceptual Captions [16]:** Conceptual Captions is a large-scale multimodal dataset sourced from images and their corresponding descriptions found online. Each image in the dataset is paired with a single caption. It contains over 3.3 million image-text pairs, featuring 16,000 unique entity concepts and labels. From this dataset, we selected the 783 most relevant labels and their corresponding image-text pairs, resulting in a train set of 1.4 million pairs and a test set of 20,000 pairs.

- **Exclusion Query Evaluation Dataset**:We created the Exclusion Query evaluation dataset using the test set images and their corresponding labels from the MSCOCO and Conceptual Captions datasets. Queries were constructed using label pairs (A, B), aiming to retrieve images that contain label A while excluding label B. Both A and B are labels from the respective datasets, and we generated label pairs only when relevant images were available. A pair (A, B) was included if the dataset contained images labeled with both A and B. From this process, we identified 3,200 valid label pairs from MSCOCO and 20,000 from Conceptual Captions. For each query (A, B), the ground truth was formed by selecting images from the test sets that contain label A but not label B. This resulted in 3,200 queries and 5,000 images for MSCOCO, and 20,000 queries and 20,000 images for Conceptual Captions. Notably, a single label pair can map to multiple images, and one image may be associated with multiple label pairs.

4.2 Baselines

- **CLIP[19]:** CLIP is a widely used vision-language model that learns joint embeddings for images and text using a contrastive learning approach. It is designed to perform well on a variety of tasks, including image retrieval. For comparison, we used the original CLIP embeddings to retrieve images based on the queries.
- **BLIP [13]:** Similar to CLIP, BLIP is another pre-trained vision-language model designed to produce strong cross-modal embeddings for tasks such as retrieval. We employ the original BLIP embeddings in our baselines.
- **Content-Based Exclusion Queries [19]:** It proposed a method for evaluating exclusion queries in keyword-based image retrieval. We have applied the method described on our dataset using CLIP embeddings of the images for getting the nearest images for the queries.

- **VDR [21]:** VDR is a retrieval-based model designed to learn disentangled representations by mapping both visual data and their linguistic descriptions into a |V|-dimensional lexical space. Each dimension in this space corresponds to a token from the vocabulary V, with the value in each dimension indicating the semantic relevance between the input data and the respective token.
- **SpLiCE [3]:** SpLiCE is a method designed to create representations by decomposing dense CLIP embeddings into sparse combinations of 10000 human-interpretable and semantically meaningful concepts.

4.3 Exclusion based Retrieval

Using our disentangled embeddings, we can perform controlled retrieval for exclusion queries. For instance, as illustrated in Fig-**??**, retrieving images for the query *sports but not basketball* follows these steps:

1. **Initial Retrieval and Dimension Extraction:** First, we retrieve the top-K similar images for a query containing a single label, such as *sports*. From these images, we extract the top active dimensions by applying a threshold t, which selects dimensions contributing to $t\%$ of the embedding's magnitude. We then repeat this process for the second label, *basketball*, resulting in two sets of active dimensions: D_1 for *sports* and D_2 for *basketball*
2. **Dimension Comparison and Final Retrieval:** For the query *sports but not basketball*, we subtract set D_2 from set D_1, isolating the dimensions that differentiate sports from basketball. Finally, we retrieve the top images with the highest magnitude in these remaining dimensions.

4.4 Quantitative Evaluation:

Evaluation on Exclusion Based Queries: We quantitatively evaluate the retrieval process by calculating MRR@k, NDCG@k and Average Precision@k for k values 1, 3 and 10, using label pairs from our Exclusive Query Evaluation Dataset as queries and retrieving images as the results. Our representations are compared against the baselines using the same label pairs across the MSCOCO and Conceptual Captions datasets. Since our model builds on pre-trained CLIP and BLIP embeddings, we present results for both CLIP-based sparse representations (SR_{clip}) and BLIP-based sparse representations (SR_{blip}). For comparison with BLIP and CLIP embeddings, we employed two methods. First, we generated embeddings for the query "images of A without B" for each label pair (A, B), with results shown in the 'SLQ' rows of Table-1. Second, we computed the difference between the average BLIP/CLIP embeddings for A and B to derive embeddings for *A without B*, with results displayed in the 'Avg Emb' rows.

For VDR and SPLICE, we employed a similar multi-step retrieval approach as outlined in Section 4.3, applying a threshold that yields optimal results. Both models aim to disentangle and separate words and concepts across dimensions.

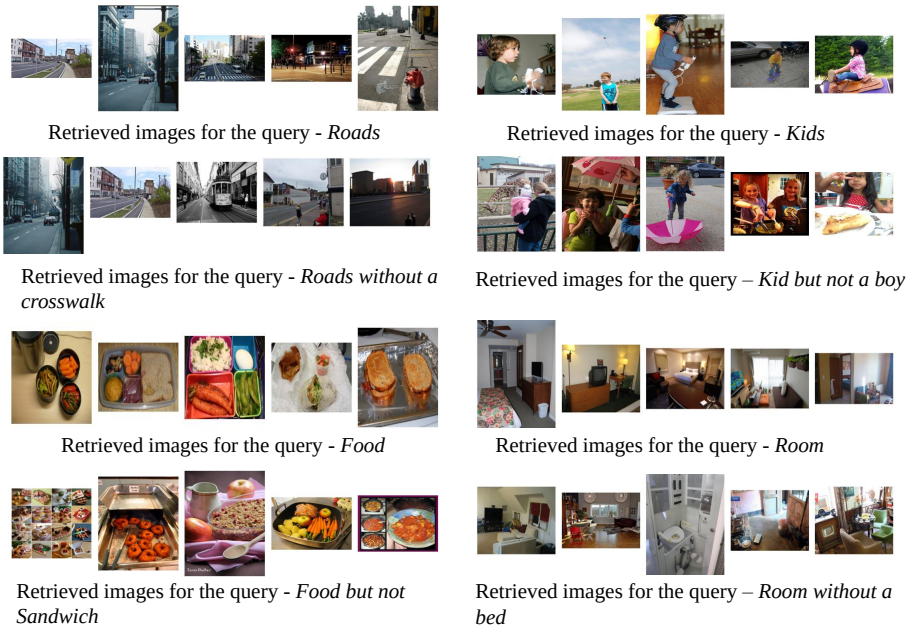


Fig. 3: **Image Retrieval examples from MSCOCO dataset for Exclusion Queries using** SR_{clip}

However, our representation outperforms these, with SPLICE performing notably poorly due to its embeddings being limited to 10,000 predefined concepts and not optimized for retrieval tasks. We also applied the method from [19] on our dataset using CLIP image embeddings, and the results are reported in the CBE row. As shown in Table-1, our methods outperform all baselines on the MSCOCO dataset and rank in the top 2 across all metrics on the Conceptual Captions dataset. Moreover, our CLIP-based representation, SR_{Clip}, surpasses all other CLIP-based models on both datasets.

Classical multimodal Retrieval: We also compare our model's performance on standard image-to-text (I2T) and text-to-image (T2I) retrieval tasks, using Average Precision scores on both the Conceptual Captions and MSCOCO datasets. The results, shown in Table-2, indicate that our BLIP-based BSR model excels in Text-to-Image retrieval for the Conceptual Captions dataset and performs nearly on par with other models in Image-to-Text retrieval, lagging only by a few points. Ore model also outperforms the VDR model, which uses a similar sparse architecture, in both i2t and t2i retrievals. This demonstrates that our approach effectively handles controlled retrieval without significantly compromising on traditional retrieval tasks.

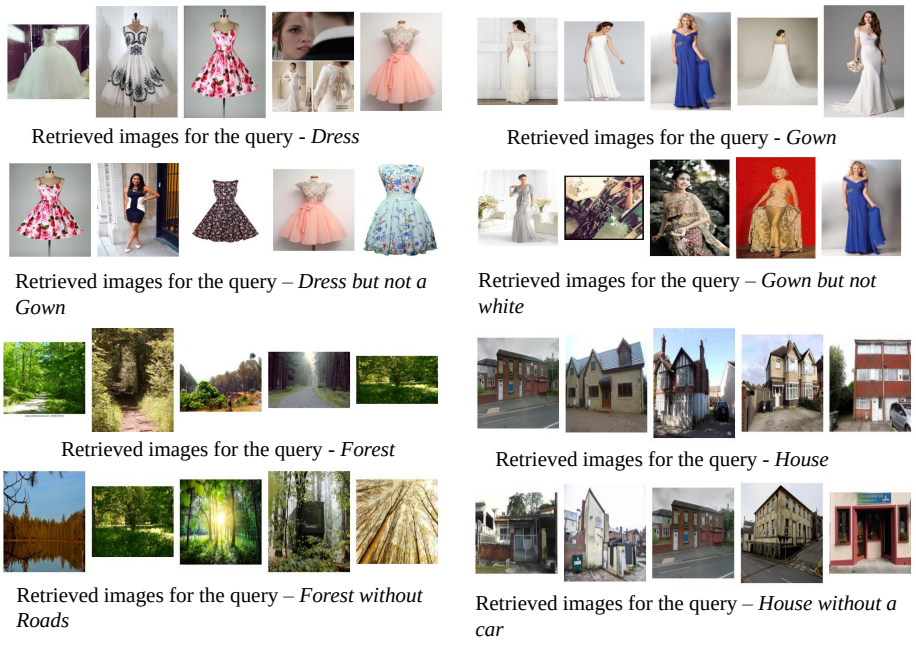


Fig. 4: **Image Retrieval examples from Conceptual Captions dataset for Exclusion Queries using** SR_{clip}

4.5 Qualitative Evaluation:

Exclusion Based Retrieval We next conduct a qualitative evaluation of our representations to examine their effectiveness in exclusion based retrieval scenerio through several examples. Fig-3 present examples from the MSCOCO dataset, while Fig-4 showcase examples from the Conceptual Captions dataset to illustrate controlled retrieval. In the first image of Fig-4, the first row displays the top 5 images retrieved for the query *Road* in which the 3_{rd} and 5_{th} images include crosswalks. For controlled retrieval, the goal is to retrieve images of streets without crosswalks. Using SR_{clip} and our retrieval method, the second row shows the results, which consist of images of roads that do not contain crosswalks.

Disentanglement: Using our method we can demonstrate effective disentanglement of image and text representations. To show this we use few examples from the Conceptual Captions datasets in Fig-5. Since Conceptual Captions is a dataset created by pairing images from the web with their associated captions, the captions often omit key details that could be present in the corresponding images. Consequently, when searching for an image using CLIP embeddings based on a query word, the returned images and the most similar captions fre-

Datasets		Image to Text			Text to Image		
		AP@1	**AP@5**	**AP@10**	**AP@1**	**AP@5**	**AP@10**
MSCOCO	VDR	0.2896	0.1980	0.1405	0.1607	0.0723	0.0471
	CLIP	0.5002	0.3349	0.2226	0.3045	0.1096	0.0662
	SR_{Clip}	0.4834	0.3289	0.2232	0.3469	0.1244	0.0732
	BLIP	**0.7864**	**0.5964**	**0.3721**	**0.6196**	**0.1707**	**0.0914**
	SR_{Blip}	0.7490	0.5548	0.3510	0.5836	0.1662	0.0895
Conceptual Captions	VDR	0.0562	0.0254	0.0170	0.0470	0.0223	0.0153
	CLIP	0.1569	0.0600	0.0369	0.1444	0.0568	0.0356
	SR_{Clip}	0.1212	0.0505	0.0324	0.1409	0.0586	0.0375
	BLIP	**0.2346**	0.0837	0.0507	0.2356	0.0843	0.0511
	SR_{Blip}	0.2243	**0.0842**	**0.0516**	**0.2449**	**0.0900**	**0.0546**

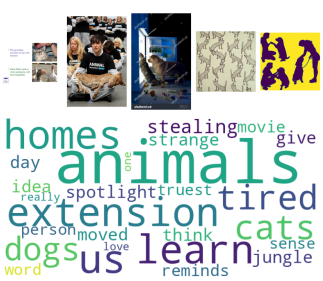
Table 2: **Average Precisions for Image-to-Text and Text-to-Image tasks on MSCOCO and Conceptual Captions datasets**

quently fail to align with the actual content of the images. However, this issue is significantly reduced when using our SR_{clip} embeddings.

We illustrate this in Fig-5 by showing the most similar images for two queries *Animals* and *Food* and a word cloud generated from the most frequent words in the top retrieved captions for these queries. We also compare these results with those retrieved using CLIP embeddings for the queries *images of animals images of food*. Notably, the word cloud generated from our embeddings contains words that are more relevant and closely related to the query. This highlights the disentanglement capability of our embeddings, as they effectively separate and organize data based on similar concepts and words, leading to more accurate and contextually appropriate retrieval results in such datasets.

5 Ablation Studies

In our model, we trained interpretable sentence embeddings with 1000 dimensions and used SPINE embeddings to enforce interpretability and control. To assess the impact of different embedding sizes, we conducted an ablation study by training the model with embeddings of varying sizes—500, 1000, 1500, and 2000 dimensions—on controlled retrieval tasks. As shown in the Table, the results indicate that using SPINE embeddings without modification (1000 dimensions) yields the best performance on both datasets using both CLIP and BLIP based models. When the embedding size was reduced to 500 dimensions, performance dropped slightly, suggesting that while 500 dimensions are somewhat effective, they do not capture all necessary features for optimal performance. Increasing the embedding size to 1500 and 2000 dimensions led to a substantial decline in performance, with AP@1 values of 0.3485 and 0.2557, respectively, likely due to the embeddings becoming too sparse. The performance without SPINE embeddings was lower than with SPINE embeddings but not drastically, indicating that while SPINE embeddings help improve performance, the model still performs



(a) *animals* using CLIP Embeddings

(b) *food* using CLIP Embeddings

(c) *animals* using SR_{clip}

(d) *food* using SR_{clip}

Fig. 5: Retrieved images and wordclouds created using the most occuring words in the retrieved sentences for the queries *animal* and *food* using CLIP and SR_{clip} Embeddings

reasonably well without them. Overall, the results suggest that 1000 dimensions strike the optimal balance for capturing meaningful information.

Datasets		SR_{Clip}			SR_{Blip}		
		AP@1	**AP@3**	**AP@10**	**AP@1**	**AP@3**	**AP@10**
MSCOCO	without SPINE	0.8361	0.8112	0.7575	0.9008	0.8643	0.7990
	500 dim embds	0.8500	0.7117	0.7565	0.8909	0.8606	0.7980
	1000 dim embds	**0.8669**	**0.8479**	**0.7865**	**0.9226**	**0.8899**	**0.8359**
	1500 dim embds	0.5653	0.5470	0.5055	0.5942	0.5566	0.4887
	2000 dim embds	0.3971	0.3902	0.3587	0.4241	0.5295	0.5645
Conceptual Captions	without SPINE	0.6742	0.6643	0.6391	0.5883	0.5756	0.5490
	500 dim embds	0.6805	0.6675	0.6437	0.6007	0.5806	0.5512
	1000 dim embds	**0.6887**	**0.6723**	**0.6463**	**0.6290**	**0.6048**	**0.5704**
	1500 dim embds	0.3585	0.3497	0.3401	0.3485	0.3417	0.3323
	2000 dim embds	0.2684	0.2638	0.2601	0.2557	0.2499	0.2414

Table 3: **Ablation studies on different sized embeddings**

6 Conclusion and Future work

We propose multimodal representations that are not only disentangled but also enable controlled retrieval for complex queries, particularly those involving exclusion. While our approach excels in exclusion-based scenarios, it faces challenges in other contexts where control over retrieval is crucial, such as inclusion queries. Addressing these limitations, along with other scenarios requiring nuanced retrieval control, is an avenue for future work.

Incorporating such controlled retrieval mechanisms would greatly enhance various applications, such as search filtering in e-commerce and content moderation or curation on social media platforms. This would enable users to perform more sophisticated queries, retrieving results tailored to specific needs—for example, finding items that include certain features while excluding others, or searching for products that meet particular criteria in e-commerce.

References

1. Bai, Y., Li, X., Wang, G., Zhang, C., Shang, L., Xu, J., Wang, Z., Wang, F., Liu, Q.: Sparterm: Learning term-based sparse representation for fast text retrieval (2020), https://arxiv.org/abs/2010.00768
2. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives (2014), https://arxiv.org/abs/1206.5538
3. Bhalla, U., Oesterling, A.X., Srinivas, S., du Pin Calmon, F., Lakkaraju, H.: Interpreting clip with sparse linear concept embeddings (splice). ArXiv **abs/2402.10376** (2024), https://api.semanticscholar.org/CorpusID:267740469

4. Chen, C., Zhang, B., Cao, L., Shen, J., Gunter, T., Jose, A.M., Toshev, A., Shlens, J., Pang, R., Yang, Y.: Stair: Learning sparse text and image representation in grounded tokens (2023), https://arxiv.org/abs/2301.13081
5. Formal, T., Lassance, C., Piwowarski, B., Clinchant, S.: Splade v2: Sparse lexical and expansion model for information retrieval (2021), https://arxiv.org/abs/2109.10086
6. Formal, T., Piwowarski, B., Clinchant, S.: Splade: Sparse lexical and expansion model for first stage ranking (2021), https://arxiv.org/abs/2107.05720
7. Greff, K., van Steenkiste, S., Schmidhuber, J.: On the binding problem in artificial neural networks. ArXiv **abs/2012.05208** (2020), https://api.semanticscholar.org/CorpusID:228063925
8. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: beta-VAE: Learning basic visual concepts with a constrained variational framework. In: International Conference on Learning Representations (2017), https://openreview.net/forum?id=Sy2fzU9gl
9. Kim, H., Mnih, A.: Disentangling by factorising (2019), https://arxiv.org/abs/1802.05983
10. Kim, M., Guerrero, R., Pavlovic, V.: 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), https://api.semanticscholar.org/CorpusID:239011496
11. Kim, M., Wang, Y., Sahu, P., Pavlovic, V.: Relevance factor vae: Learning and identifying disentangled factors (2019), https://arxiv.org/abs/1902.01568
12. Lee, M., Pavlovic, V.: Private-shared disentangled multimodal vae for learning of hybrid latent representations (2020), https://arxiv.org/abs/2012.13024
13. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation (2022), https://arxiv.org/abs/2201.12086
14. Lin, T.Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollár, P.: Microsoft coco: Common objects in context (2015), https://arxiv.org/abs/1405.0312
15. Mondal, A., Sailopal, A., Singla, P., Prathosh, A.P.: Ssdmm-vae: variational multimodal disentangled representation learning. Applied Intelligence **53** (07 2022). https://doi.org/10.1007/s10489-022-03936-z
16. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Annual Meeting of the Association for Computational Linguistics (2018), https://api.semanticscholar.org/CorpusID:51876975
17. Subramanian, A., Pruthi, D., Jhamtani, H., Berg-Kirkpatrick, T., Hovy, E.H.: Spine: Sparse interpretable neural embeddings. ArXiv **abs/1711.08792** (2017), https://api.semanticscholar.org/CorpusID:19143983
18. Wang, X., Chen, H., Tang, S., Wu, Z., Zhu, W.: Disentangled representation learning (2024), https://arxiv.org/abs/2211.11695
19. Yoshikawa, E., Tajima, K.: Content-based exclusion queries in keyword-based image retrieval. In: Proceedings of the 2024 International Conference on Multimedia Retrieval. p. 1145–1149. ICMR '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3652583.3657619, https://doi.org/10.1145/3652583.3657619
20. Zhang, W., Zhang, M., Wu, S., Pei, J., Ren, Z., de Rijke, M., Chen, Z., Ren, P.: Excluir: Exclusionary neural information retrieval (2024), https://arxiv.org/abs/2404.17288

21. Zhou, J., Li, X., Shang, L., Jiang, X., Liu, Q., Chen, L.: Retrieval-based disentangled representation learning with natural language supervision (2024), https://arxiv.org/abs/2212.07699