A HYPERPARAMETERS

Our hyperparameter selection follows DPR for the text-to-text retrieval while CLIP for the cross-modal retrieval. Details can be found in Table 1 below.

	text-to-text	cross-modal
Batch Size	256	4096
Epoch	20	20
Learning Rate	2e-5	2e-4
Warmup Epoch	1	1
LR Decay	Linear	CosineAnnealing
Normalization	None	L2-norm
Temperature	None	0.07
Hard Negative	1	0
Max Activation Num	768	512
Max Seq Length (Q/P)	256/256	77/49
Transformer Width (Q/P)	768/768	768/768

Table 1: Hyperparameters for training VDR.

B REPRODUCTION COMPARISONS

In Figure 1 and Table 2, we present our reproductions of DPR and CLIP, accompanied by results from other pertinent research papers. This facilitates valid reproduction and fair comparison. Notably, our study consistently showcases the highest levels of performance in relation to these foundational baselines.

Figure 1 showcases our replicated DPR model, which outperforms the versions reported in other studies. Therefore, we present our replicated baselines in main paper.

In Table 2, it's noteworthy that UniCLIP, having undergone pre-training on YFCC15M with a similar configuration, demonstrates superior outcomes. As a result, we have chosen to adopt their outcomes for the cross-modal retrieval aspect, with the exception of ImageNet where we have opted to utilize our own replicated scores.

Model	DPR	DPR†
MS MARCO	17.7	31.7
ArguAna	17.5	**40.8**
Climate-FEVER	14.8	**16.2**
DBPedia	26.3	**30.4**
FEVER	56.2	**63.8**
FiQA	11.2	**23.7**
HotpotQA	39.1	**45.2**
NFCorpus	18.9	**26.1**
NQ	**47.4**	43.2
SCIDOCS	7.7	**10.9**
SciFact	31.8	**47.4**
TREC-COVID	33.2	**60.1**
Touché-2020	13.1	**22.1**
Avg.	26.4	**35.8**
Best on	1	11

Figure 1: Reproduction of DPR from different sources. †: ours.

	ImageNet		MSCOCO						Flickr30k					
			image-to-text			text-to-image			image-to-text			text-to-image		
	Top1	Top5	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Our re-implementation based on DECLIP's checkpoints														
CLIP†	32.80	57.35	16.94	39.50	50.94	10.75	26.19	35.24	34.70	65.00	74.10	23.60	46.90	58.76
SLIP†	33.57	58.60	17.94	40.42	51.82	11.22	26.48	35.36	35.60	65.60	77.30	23.40	47.32	57.96
FILIP†	39.16	64.35	21.64	46.66	59.00	13.72	31.72	41.60	46.30	74.40	83.20	30.66	58.18	68.56
DeCLIP†	43.24	69.40	25.34	51.20	63.44	16.59	35.24	45.41	51.30	80.70	88.50	35.50	63.04	73.02
Results reported by UniCLIP														
CLIP	31.3	-	20.8	43.9	55.7	13.0	31.7	42.7	34.9	63.9	75.9	23.4	47.2	58.9
SLIP	38.3	-	27.7	52.6	63.9	18.2	39.2	51.0	47.8	76.5	85.9	32.3	58.7	68.8
DeCLIP	41.2	-	28.3	53.2	64.5	18.4	39.6	51.4	51.4	80.2	88.9	34.3	60.3	70.7

Table 2: Reproduction of cross-modal retrieval on ImageNet, MS COCO, and Flickr30k from different sources.

C Reliance on Masked Language Model

In this section, we empirically validate the reliance of lexical retriever on the pre-trained masked language models (MLM).

We adhere to the same training pipeline of our approach, while only initializing the linear projection within the DST head of the p encoder and training it from scratch. This configuration is denoted as VDR_{proj}. Additionally, we incorporate BERT-based models as a lexical retrieval baseline, without undergoing further fine-tuning, denoted as BERT_{lex}. We present the training result below.

Model	Epoch	NDCG10@BEIR	MRR10@MARCO
BERT_{lex}	0	20.1	28.9
VDR	1	38.9	28.4
VDR	2	42.3	30.8
VDR	3	42.9	31.7
VDR	4	43.4	32.4
VDR	5	43.7	32.8
VDR_{proj}	5	0.2	0

Table 3: Different setup of lexical retrievers trained in the text-to-text retrieval scenarios.

Our experimental findings show that when we employ the pre-trained MLM projection, which inherently offers a rational weighting distribution from the outset, VDR reliably improve the effectiveness and achieve best results within 5 training epochs. Conversely, when starting from scratch with the projection layer on p side, even with substantial training efforts, the VDR_{proj} setup encounters challenges in attaining effective convergence. This obstacle compromises the final outcomes and makes it even fall behind the performance of the untrained baseline, BERT_{lex}. These findings support and validate the insights presented in Section 3.3.

Moreover, our observations and experiments in cross-modal retrieval suggest that achieving an effective transition from a scratch-initialized distribution to a rational one necessitates a substantial amount of training data, a large batch size, and the inclusion of the contrasting mask.

D Impact of Nonparametric Entry

We emphasize the essential role of incorporating the nonparametric entry during training to achieve disentanglement in our model. Without it, our model tended to assign excessive values to overly common or rare tokens. We conjecture this issue arises from the interdependence between the gating and weighting functions, which amplifies biases rather than mitigating them.

To validate this hypothesis, we examine the embeddings produced by our model with and without the nonparametric entry. In Figure 2, we label our model in cross-modal setting with nonparametric entry as VDR (w/ BoW), without it as VDR (w/o BoW), and a BERT-based model as BERT-text. We take these encoders to embed text and images from the MS COCO test set into lexical representations, calculating average values for each token within these representations. We then visualize the top 100 tokens using word clouds and the distributions of their values using box plots. Our observations reveal that image representations from VDR (w/o BoW) have sharper distributions, characterized

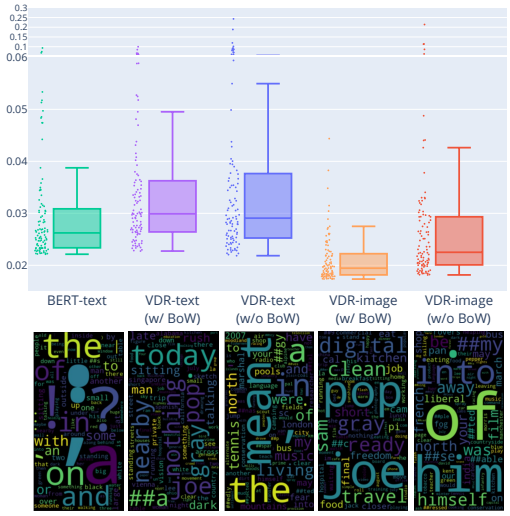

Figure 2: Box plot (top) and word cloud (bottom) of the vocabulary distributions on MS COCO.

by higher upper bounds and mean values in L2 norm space. However, the top 100 tokens in word cloud of VDR (w/o BoW) lacks meaningfulness and fails to convey distinct information. This implies that the omission of the nonparametric input in VDR amplifies biases, causing specific meaningless tokens to be assigned excessive values, thereby consistently dominating the matching outcome.

We term the above phenomenon as "disentanglement laziness". Simply put, the learning process avoids the "hard work" of properly disentangling and instead takes the easy route for optimization, degrading to a entangled fashion. In bi-encoder architecture, we have noticed that relying solely on parametric components causes the model to consistently assign high values to tokens that are either frequently or never encountered, resulting in entangled learning within a subset of the vocabulary space. In doing so, the model seemingly "escapes" from the rigorous work of disentanglement, reducing the problem into an optimization within an entangled representation space. Interestingly, this phenomenon is not exclusive to the research of disentangled representation learning. It is also observed in other research, like the Mixture of Experts (MoE), where it is referred to as "load imbalance". This term alludes to the model's tendency to consistently favor certain experts, thereby causing an unequal distribution of learning and optimization channels. In addressing the observed issue in our experiments, we enhance the disentanglement process by integrating the nonparametric entry, which provides stable and straightforward supervision of the data, independent of any influences from the entangled parametric model.

E SPARSITY V.S. EFFECTIVENESS

We present the effectiveness of VDR with different amounts of activation units k in Table 4 and Table 5.

In the text-to-text scenario, the results demonstrate that the effectiveness of VDR increases as k increases, reaching a peak and then decreasing. This suggests that by properly selecting the number of activation units, VDR is able to achieve considerable improvement.

In the cross-modal scenario, the results demonstrate that the effectiveness of VDR increases consistently as k increases in the majority of cases. This suggests that a higher number of activation units can lead to better performance in cross-modal scenarios.

Model	Word Length		VDR^{α}	VDR					
	Query	Doc		0*	32	64	128	256	768
MS MARCO	-	-	33.8	33.0	34.1	34.4	**34.5**	34.4	34.3
ArguAna	193	167	**48.8**	48.6	27.3	41.7	47.0	47.2	46.5
Climate-FEVER	20	85	**18.1**	17.2	17.1	17.6	17.2	17.2	16.9
DBPedia	5	50	37.6	35.1	38.0	38.6	**39.0**	38.8	38.9
FEVER	8	85	**74.8**	73.7	74.0	73.9	73.9	73.9	73.9
FiQA	11	132	**29.3**	28.1	28.2	28.8	28.8	28.6	28.4
HotpotQA	18	46	**68.4**	64.4	65.0	65.5	65.5	65.4	65.0
NFCorpus	3	232	32.7	32.5	**33.0**	32.9	32.9	32.8	32.5
NQ	9	79	45.8	44.6	45.8	46.4	46.9	47.0	**47.2**
SCIDOCS	9	176	**15.4**	14.8	14.8	15.0	15.1	15.2	15.3
SciFact	12	214	**67.6**	67.3	66.8	67.2	67.1	67.3	66.6
TREC-COVID	11	161	**69.0**	66.5	67.3	67.8	67.6	67.3	66.2
Touché-2020	7	292	27.7	29.1	29.0	29.4	29.5	**29.8**	29.4
average	-	-	44.6	43.5	42.2	43.7	44.2	44.2	43.9

Table 4: Effectiveness of VDR with different amounts of activation number k on MS MARCO and BEIR. **Bold** denotes the overall best result and underline denotes the best query sparsity for VDR.

VDR K	ImageNet		MSCOCO						Flickr30k					
			image-to-text			text-to-image			image-to-text			text-to-image		
	Top1	Top5	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
32	34.8	56.7	18.9	38.9	49.3	15.8	35.7	46.8	34.9	59.2	70.3	31.7	58.0	68.5
64	36.6	59.7	23.3	45.5	56.1	17.1	37.3	48.6	40.9	67.9	78.2	33.3	59.8	71.3
128	38.1	62.0	27.0	49.7	61.1	**17.4**	**38.1**	49.4	44.9	73.9	83.4	**33.3**	60.0	**71.4**
256	38.5	63.3	29.5	53.9	64.4	**17.4**	**38.1**	49.4	49.9	77.2	85.6	32.9	60.0	71.2
512	**38.7**	**63.6**	**30.9**	**54.5**	**65.4**	**17.4**	**38.1**	**49.7**	**51.0**	**79.3**	**86.7**	32.4	**60.1**	70.7

Table 5: Effectiveness of VDR with different amounts of activation number k on ImageNet, MS COCO and Flickr30k. **Bold** denotes the best result.

F Case Study

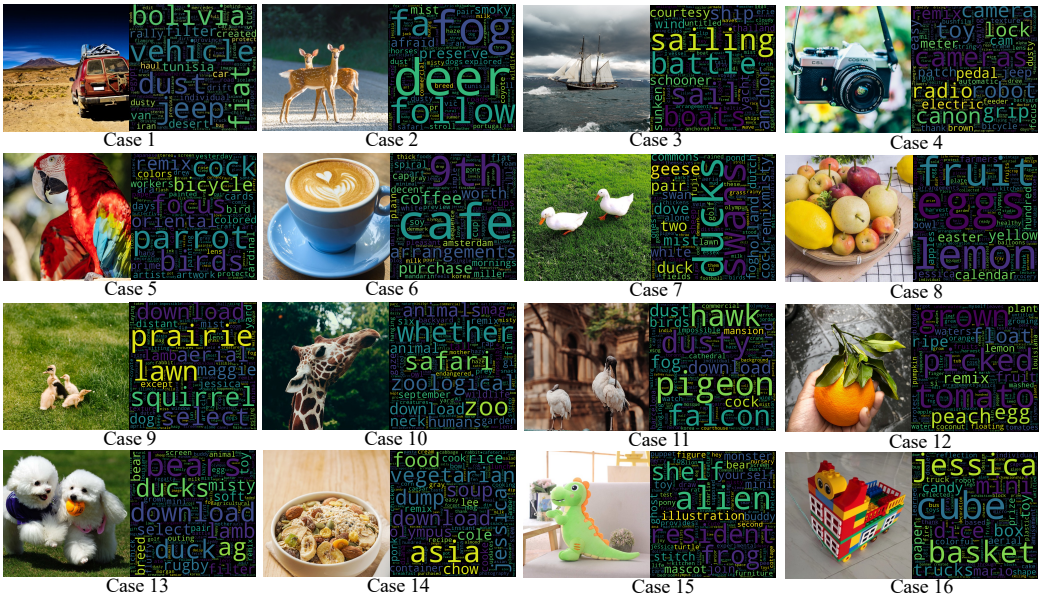

Figure 3: More case study on VDR disentanglement of image.

We provide additional case studies in Figure 3. Cases 1 through 8 represent successful cases as determined by our experts, while cases 9 through 16 illustrate instances where the image encoder did not perform as expected. For those good cases, we can observe that the main concepts present in the images are aptly represented in the word cloud. This indicates that the image encoder of VDR effectively captures the semantic meaning of these images, producing a reasonable and understandable representation within the disentangled vocabulary space. For the unsuccessful cases, we observed some cases of misidentification or misconception. After analysis, we identified two main reasons for these errors. First, there are cases where the encoder fails to correctly identify the object within the image because it resembles another object, thus skewing the results. For example, in cases 9 and 13, the encoder incorrectly identifies ducks as squirrels and dogs as bears, likely due to their similar appearances within the images. Secondly, certain images entail concepts associated with n-gram phrases, which is challenging for internal inspection or word cloud visualization. For instance, in case 10, the term "giraffe" is tokenized into three tokens: "gi", "##raf", and "##fe". While the first token, "gi", appears in the word cloud, the latter two are missing. Such n-gram concepts can be challenging to capture or infer through an internal inspection of the representation. This limitation can be traced back to the choice of tokenizer used prior to training.

G Details in Efficiency Measurement

We perform retrieval using 1k text queries with a pre-embedded corpus consisting of 100k data points. We employed inverted indexes for sparse retrieval. The retrieval experiments are conducted on a single-threaded Linux machine with two 2.20 GHz Intel Xeon Gold 5220R CPUs. The batch size used in the experiments is one and the maximum sequence length for queries is 77. The MS MARCO and MS COCO datasets were utilized for text-to-text and text-to-image retrieval, respectively. The average query length for text-to-text retrieval was 6.8 and for text-to-image retrieval was 11.6. The effectiveness of the retrieval methods was evaluated using the average NDCG@10 scores on the BEIR metric for text-to-text retrieval and the Recall@1 metric for text-to-image retrieval on the MS COCO dataset.

H Details of Human Evaluation

We compare our method with the SOTA captioning model BLIP.

For the VDR, we encoded the image into a disentangled representation and asked human evaluators to select understandable tokens from the top 5 tokens with the largest dimensional values, without access to the original image. For BLIP, evaluators extracted up to 5 tokens from the captions generated by BLIP that best reflected the meaning of the caption, also without access to the image.

To compare the outcomes of our VDR method and the BLIP model, we introduced an additional group of 10 participants. Each participant assessed a total of 20 images. The participants were asked (1) whether the set of tokens effectively captured the key concepts illustrated in the associated image, and (2) which set of tokens better described the image.