Multimodal Preference Optimization-based Semantic Re-ranking for Improved Vocabulary Tagging of Images

Anonymous ACL submission

Abstract

When scaling open-vocabulary vision-language models (VLMs) to large, realistic label spaces (over 100k objects), the performance of tagging with vision language models (VLMs) sharply decreases. Initial high-confidence matches become saturated with synonymous or taxonomically redundant variations of the primary object, crowding out distinct auxiliary objects from the top-K candidates. This results in both lowered precision and recall when compared to manual tagging. In this paper, we present a novel multimodal re-ranking approach for improving image taggers using dynamic semantic label clustering followed by re-ranking by a regression model trained on visual input and the representative tags from the clusters. A maximum bipartite ground-truth matching-derived objective function is used to capture human preference in tagging. Experiments show the effectiveness of this approach leading to an average 62.4% improvement in NDCG performance in open vocabulary tagging as tested on multiple large vocabulary benchmark collections and against base VLM models with and without re-ranking.

1 Introduction

Automatic image tagging is important in building several commercial applications such as online retail, surveillance systems or organizing photo collections. Open-vocabulary image tagging has advanced dramatically via Vision-Language Models (VLMs)(Zhai et al., 2023a; He et al., 2017; Radford et al., 2021; Gu et al., 2021; Zareian et al., 2021; Hinami and Satoh, 2018; Xuran Pan and Huang, 2022; Yuksekgonul et al., 2023). However, deploying these models against expansive, real-world vocabularies reveals a foundational *recall ceiling problem*. When a scene contains multiple distinct objects, a standard dual-encoder's top-K results are frequently saturated by linguistic variants of the foreground objects as shown in Figure 1(b) for the image shown in Figure 1a. Subsequent expensive cross-encoders cannot help either if this first stage itself filters the auxiliary objects. Choosing a larger K to preserve multi-object recall, on the other hand, leads to lowered precision generating a highly noisy tag distribution.

Ideally, an automated multi-object tagging framework should replicate human annotation as closely as possible: it must comprehensively cover as many visually distinct objects in a scene as possible, while actively avoiding redundant, repetitive, or synonymous descriptions of the same entity, and preventing false positives. Achieving this human-like precision remains exceptionally difficult when scaling open-vocabulary vision-language models (VLMs) (Zhai et al., 2023a) to massive, realistic label spaces (over 100k objects), partly due to the reliance on contextual co-occurrence for modeling semantics in current embedding models.

In this paper, we present a novel approach to achieving multi-object tagging that aims for such human-like precision when scaling open-vocabulary vision-language models (VLMs) to large label vocabularies. Specifically, we first reduce label noise by dynamic semantic label clustering that captures diverse multi-object contexts. The prototypical tags from such clusters are then re-ranked using a new multimodal regression reranker model trained based on an objective function derived from matching such tags to human preferred tags via maximum-weighted bipartite graph matching. Experiments show the effectiveness of this approach leading to an average 62.4% improvement in NDCG performance in open vocabulary tagging as tested on multiple large vocabulary benchmark collections.

2 Related work

Existing approaches to filtering the redundant tags range from Maximal Marginal Relevance (MMR) (Carbonell and Goldstein, 1998), which use dis-

tance in textual embedding spaces to more complex approaches using matrix determinant maximization (Kulesza et al., 2012; Chen et al., 2018). However, because semantic similarity is computed with existing textual embeddings, they could remove contextual co-occurring tags as redundant (e.g. "forest", "tree", and "leaf") thus reducing diversity and recall. In addition, synonymous phrases with distant textual embeddings may still be kept lowering the precision.

Alternative methods that utilize hard-threshold greedy pruning or Reciprocal Rank Fusion (RRF) (Cormack et al., 2009) need multiple ways of ranking and are also incapable of smoothing over the highly fragmented token distributions characteristic of large-scale open vocabularies. Cross-modal re-rankers that leverage deep, patch-to-token cross-attention blocks (Li et al., 2021) and approaches like BLIP-2 (Li et al., 2023) are computationally prohibitive at inference time. Lightweight late-fusion architectures or Multimodal LLM rerankers (Qu et al., 2023; Wu et al., 2025), on the other hand, do suboptimal tag allocation as their loss functions pick the nearest-neighbors to human annotations (Yasumoto et al., 2024) in a greedy way which forces multiple overlapping or synonymous candidate tags to compete for a single ground-truth tag.

3 Overall method

Our overall approach to tag filtering for large vocabularies consists of 3 stages as illustrated in Figure Figure 2a.

Stage 1: Native VLM over-fetching: Given an image, the VLM's native image-to-text ranking capabilities is utilized to extract a broad candidate set from the large vocabulary. Let I be an input image and $\mathcal{V}$ be a large vocabulary of size $|\mathcal{V}|$. Let the ranked list of N tags predicted by VLM models such as CLIP(Radford et al., 2021) using a vocabulary $\mathcal{V}$ for an image I be denoted as $P_I = \{p_{1I}, p_{2I}, ..p_{NI}\}$ where p_{jI} is the jth predicted tag for image I and $N \leq |\mathcal{V}|$. Each of these predicted tags have similarity scores $S_I = \{s_{1I}, s_{2I}, ..s_{NI}\}$ where s_{jI} is the cosine similarity score between the multimodal encoding of the predicted tag p_{ji} and image i in the VLM space. This broad pool N is kept large enough to ensure distinct auxiliary objects are successfully captured in the initial ranked list.

Stage 2: Semantic clustering & tag selection. To condense language-space tag noise, we perform semantic clustering using a language-only embedding representation of the tags. This allows us to capture semantic relationships between terms independent of their image appearance associations.

Dataset	Train	Test	Tags
MS-COCO	118,287	40670	568456
Visual Genome	108,077	7554	83404
SUN	76,128	16637	567
CUB	5,994	5,794	200
AWA2	30337	6985	50

Table 1: Illustration of the datasets used for evaluation.

Each predicted label in the initial list is encoded using a textual embedding such as SBERT(Devlin et al., 2019) denoted as $E_{pI} = \{e_{p1I}, e_{p2I}, ..e_{pNI}\}$ and clustered into $\mathcal{K}$ clusters using K-means clustering. The value of $\mathcal{K}$ is chosen based on the desired NDCG@K(Wang et al., 2013), a metric commonly used in evaluating retrieval algorithms. To eliminate the latency overhead of running online clustering from scratch, we employ an offline vocabulary warm-start by pre-computing global macro-clusters across the entire tag vocabulary. At inference time, the global cluster distribution is used to initialize the local K-means centroids, achieving convergence in just 2-3 iterations. Following convergence, we extract a single "prototype" tag per cluster by selecting the member that achieved the highest initial Stage 1 native visual rank. Thus the visual-textual association ranking is still preserved in the clusters but the number of predicted labels are considerably reduced.

Stage 3: Neural regression reranking: The final ranking scores are computed independently for each candidate prototype tag by pairing it with the image and predicting a distance score using a multimodal reranker model. The image I is passed through a frozen visual backbone to extract a global image embedding $\mathbf{v}_I \in R^d$. Simultaneously, each prototype $y_{jI} \in Y_I$ is encoded using the text embedding to generate its global text representation $e_{y_{jI}} \in R^d$. For each candidate, we construct a joint representation by concatenating the visual and textual features: $\mathbf{x}_I = [\mathbf{v}_I \,||\, e_{y_{jI}}]$. These fused vectors are fed pairwise into the multimodal reranking model (MRM) to obtain a ranking score

$$r_I = \text{MRM}([\mathbf{v}_I \,||\, e_{y_{jI}}]) \tag{1}$$

The reranking model is a 453M parameter regres-

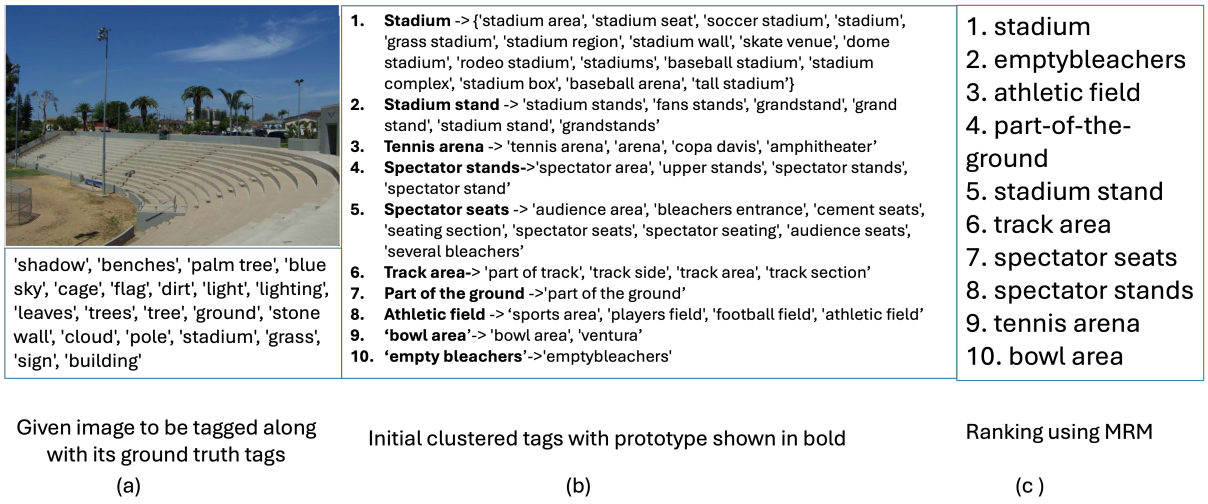

Figure 1: Illustration of semantic clustering with multimodal reranking. (a) Image from Visual Genome collection and its ground truth tags. (b) Result of semantic clustering of top 50 tags into 10 clusters. The prototype chosen from each cluster is listed first in bold. (c) Result of multimodal re-ranking using the MRM model. Note that the redundant tags are after the distinct tags and the false positives are ranked last.

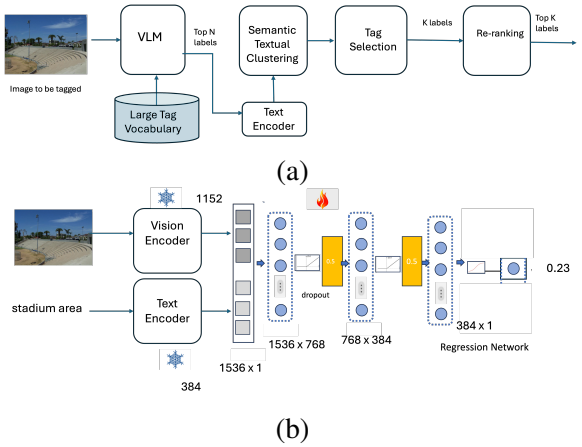

Figure 2: Illustration of overall approach to semantic clustering and tag matching. (a) Overall workflow. (b) Re-ranking model. During training, the distance score is provided. At inference time, it predicts the score.

sion network consisting of two linear layers, two drop out layers with RELU for intermediate layers and a separate sigmoidal function for producing the output regression score as shown in Figure 2b. The prototypes are then re-ranked by r_j to yield the final top-K visual tags. Due to the lightweight reranking model used, the overall execution is in sub-milliseconds.

4 Training the multimodal reranker

While the semantic clustering step significantly condenses the label field, it *reduces but does not eliminate* overlapping and highly similar lexical variations within the remaining candidate pool. In addition, false positives are still present due to the larger pool of initial set of labels. The goal of the multimodal re-ranker is to filter the tags to closely mimic manual tagging by training on a supervision dataset. The loss function reflects continuous preference optimization among the tags such that the tags that most closely match the supervision ground truth labels get ranked above the redundant synonymous tags which in turn get ranked above the false-positives.

Since a cluster tag may match to more than one ground truth tag and vice versa, a natural way to build the preferences for supervision is to form a weighted bipartite graph with the cluster tags and ground truth tags and perform a maximum weight bipartite ground-truth matching to find the best matching pairs. Specifically, let each image I be associated with a set of M ground truth tags $G_I = \{g_{1I}, g_{2I}, ..g_{MI}\}$ where $g_{jI} \in V_g$ is the jth ground truth tag for image I. Let e_{gjI} denote the semantically encoded version of the jth ground truth label for image I. The best pairs are then selected as a maximum matching in the bipartite graph which is the set of the edges chosen in such a way that no two edges share an endpoint. A maximum matching is a matching of maximum size (maximum number of edges). Several algorithms are available for efficient computation of maximum matching in bipartite graphs(Hopcroft and Karp, 1971). Specifically, the standard linear sum assignment formulation to solve this matching (called the Hungarian algorithm) automatically guarantees a

maximum-size (maximum-cardinality) matching when it solves for the maximum weight matching.

4.1 Formulating the Preference Optimization Loss

During each training forward pass, we build a global bipartite graph between $Y_i = \{y_{1i}, y_{2i}, ..y_{Ki}\}$ cluster prototypes and the human ground-truth labels G_i with the cosine distance between label's textual embeddings serving as edge weight to construct distance matrix $C \in R^{K \times M}$. To ensure the matcher aggressively rejects true false positives rather than forcing weak, ungrounded alignments, we apply a strict semantic distance ceiling corresponding to a cosine similarity threshold of 0.5:

$$C_{i,j} = \begin{cases} \mathrm{d}(e_{y_i}, e_{g_j}) & \text{if d} \leq 0.5 \\ 1.0 & \text{otherwise} \end{cases} \tag{2}$$

To account for the discrepancy between the fixed number of prototypes and varying human label counts, the matrix is padded to a square structure $\max(K, M) \times \max(K, M)$ using virtual dummy nodes assigned a maximum distance penalty ($C_{\text{dummy}} = 1.0$). We then solve the linear sum assignment problem using the Hungarian Algorithm to establish a globally optimal, bijective mapping function π^*:

$$\pi^* = \arg\min_{\pi} \sum_{i} C_{i,\pi(i)} \tag{3}$$

By setting the distance of any weakly correlated pair to the maximal penalty of 1.0, we ensure that the Hungarian solver treats visually irrelevant tags identically to empty dummy nodes. Consequently, true false positives are cleanly factored out of the real edge alignment set π^* and into the suppressed target tier, providing a smooth contrastive gradient for the late-fusion reranker model.

4.2 Continuous Preference Optimization

We derive a continuous target vector $\mathbf{t} \in R^K$ for our regression supervision. For each prototype $\hat{y}_i \in \mathcal{Y}_K$, its supervision target t_i is calibrated to reflect its position in the preference ranking:

$$t_i = \begin{cases} 1.0 - C_{i,j} & \text{if } (i,j) \in \pi^+ \\ \gamma \cdot \max_{j \in Y_K}(1.0 - C_{i,j}) & \text{otherwise} \end{cases} \tag{4}$$

where $\pi^+ = \pi^* \backslash \text{dummy}$, and $\gamma \in (0, 1)$ is a hyperparameter dictating the suppression discount factor. We chose this preference ranking because omitting unselected prototypes entirely from the training loop breaks the optimization objective, as the network is never exposed to negative constraints and fails to learn false-positive elimination. Conversely, forcing all unselected tags to an absolute zero target causes gradient conflicts by penalizing valid, unaligned synonyms.

With this configuration, true false-positives are cleanly driven to 0.0, while valid but unaligned synonyms receive a smooth, non-zero floor value. This informs the network that these concepts are visually present but sub-optimal for high-ranking placement. The multimodal re-ranking model (MRM) minimizes the standard Mean Squared Error (MSE) over the entire preference-calibrated landscape:

$$\mathcal{L} = \frac{1}{K} \sum_{i=1}^{K} (\text{MRM}([\mathbf{v}_I, ||, e_{y_i}]) - t_i)^2 \tag{5}$$

By training the model on this globally resolved cost landscape, the MRM inherently internalizes the taxonomic preferences of the ground truth.

The 453M parameter neural reranker model was trained for 100 epochs using the AdamW optimizer on an NVIDIA A100 GPU with 40GB of memory and a batch size of 32. The cosine annealing learning rate scheduler was used with the maximum learning rate of 1e-5 and 50 steps for warm up. The suppression discount hyperparameter γ was anchored to 0.3.

At inference time, even though the re-ranker scores candidates strictly in a standalone, pairwise fashion, it automatically scores the prototype that best fits the global concept landscape higher. With the training, the network packs all correct concepts into the top slots, inherently maximizing the NDCG metric and providing the most preferred non-redundant responses at the top of the tag search. Because our late-fusion re-ranker model optimizes a highly regularized continuous preference landscape, NDCG score variance across different random seeds is negligible (± 0.003.

5 Results

We experimented with several benchmark datasets with varying vocabulary sizes and images shown in Table 1. In each case, we selected the images and their ground truth tags of the training partition to train the re-ranker model. We selected the more powerful SigLIP So400M's vision encoder for the reranker model, and used the same

Dataset	CLIP	OpenCLIP	SigLIP	Ours	MMR	BLIP-2	Clust	GreedyRank
MS-COCO	0.011	0.061	0.091	**0.219**	0.159	0.015	0.112	0.148
Visual Genome	0.014	0.016	0.019	**0.195**	0.18	0.016	0.108	0.171
SUN	0.018	0.021	0.042	**0.154**	0.149	0.020	0.101	0.138
CUB	0.045	0.064	0.091	**0.281**	0.22	0.056	0.121	0.201
AWA2	0.16	0.19	0.21	**0.41**	0.35	0.19	0.32	0.37

Table 2: Illustration of the performance of semantic clustering with re-ranking using semantic ground truth matching in comparison with VLM models and ablation configurations.

textual (SBERT(Reimers and Gurevych, 2019)) encoding throughout for clustering, regression targets, and preference tiers. This ensured closed-loop semantic homogeneity in the pipeline. The tagging performance improvement was tested using CLIP variants including CLIP(Radford et al., 2021), OpenCLIP-H/14 trained specifically on LAION-2B, and SigLIP(Zhai et al., 2023b) as base VLMs.

Semantic reranking example

We first illustrate the quality of our re-ranking using multimodal re-ranker with semantic clustering using an example. Figure 1a shows an image from the Visual Genome collection and its ground truth tags. Figure 1b shows the result of semantic clustering using SBERT encoding of the top N=50 tags using K=10 clusters. Their selected prototypes based on visual similarity ranking of Step-1 are listed first in bold in Figure 1b. As can be seen, although the redundancy has been drastically reduced (from 50 to 10), some redundancy still remains among the cluster prototypes (e.g. *spectator stand* and spectator stand).Finally, Figure 1c shows the result of multimodal re-ranking using the MRM model. Note that the redundant tags are after the distinct tags and the false positives are ranked last using the re-ranker.

Comparing Re-ranking performance

We evaluated the performance of our tagging using NDCG on the test partitions of the datasets in Table 2 in comparison to base VLM models with and without re-ranking. Table 2 Column 1-3 shows the performance of base CLIP model(Radford et al., 2021), OpenCLIP models and SigLIP(Zhai et al., 2023b) without explicit re-rankers. Columns 4 shows the performance of base CLIP model with MMR re-ranking. Column 5 shows the performance of CLIP model with BLIP-2 cross-modal encoder. Finally Column 6 shows the performance of our re-ranking approach with semantic clustering. As can be seen, our approach achieves the best overall performance with an average of 62% improvement in tag prediction as measured through NDCG@50 metric. Further, the overall re-ranking including the semantic clustering is within a few milliseconds per image in comparison to the heavy duty cross-modal encoders such as BLIP-2 which took more than a minute per image for large vocabularies of Table 1.

Ablation studies

We evaluated the relative contribution of semantic clustering and MRM re-ranking to see if both are necessary. Column 7 of Table 2 shows the NDCG@50 performance using semantic clustering alone using base CLIP VLM. While the performance is improved over CLIP VLM, the combination of semantic clustering with MRM re-ranking outperforms semantic clustering alone. Finally, we evaluated the necessity of using the maximum bipartite matching-based tag preference ordering by implementing the reranking using direct MSE regression, i.e. directly predicting the raw SBERT semantic similarity score between every candidate prototype and the closest individual ground-truth tag via greedy nearest neighbors. This result is shown in Column 8. Comparing this performance to Column 6 with our maximum matching-based preference ranking shows that there is a performance jump achieved by using the Hungarian algorithm to untangle synonym clusters globally.

6 Discussion

Our approach was designed to keep the scaling of vocabularies as well as computational limitations in mind. It is able to achieve higher precision as well as recall despite scaling for a few reasons. Retaining the larger initial list N as well as the preference ranking keeps the recall high. Higher precision is due to the semantic clustering followed by preference ranking.

Unlike pointwise and listwise re-ranking methodologies such as MMR and multimodal LLM

rankers, using maximum weight matching ensures that as many unique pairings can be derived without the systemic gradient confusion seen for large vocabularies. Note, however, that while the maximum weight bipartite matching was performed at training time, no such matching is needed at inference time with the reranker model acting as its surrogate instead.

By scaling unselected prototypes down via our suppression discount factor ($\gamma = 0.3$), our model treats the unaligned space as a continuous preference tier rather than a destructive negative boundary. The MRM learns to structure its output landscape as a clean hierarchy: placing preferred visual targets at the top, grouping redundant synonyms next, and true false-positives last.

While multi-billion parameter cross-encoder networks possess exceptional fine-grained region-to-word alignment, they are fundamentally limited by the computational overhead of cross-attention computations per image. In contrast, our approach provides computational efficiencies being a small parameter model with frozen vision and text encoders.

7 Conclusions

In this paper we have presented a simple and effective approach to improving the image tagging performance of vision language models using semantic textual clustering with multimodal re-ranking derived from a continual preference optimization reflecting ground truth labeling. The results across datasets show a significant improvement in tagging performance while keeping it computationally simple.

8 Limitations

Our work relies on the ability of textual encodings to capture synonymous relations in phrases. This is not always guaranteed, particularly for extreme lexical polysemy and homonymy, such as for example the word 'orange' which has multiple senses. This can cause words with multiple senses to be wrongly clustered semantically. Although we have used the maximum bipartite matching formulation to maximize the recall by pairing as many predicted labels with ground truth, there is still a gap between manual tagging and automatic tagging, with the initial draw of N labels still being a limitation in scenes which convey a lot of information and where there are multiple visible objects.

9 Acknowledgments

We acknowledge the providers of datasets and developers of open source models used in this paper. It made our comparative evaluation significantly simpler task. We also gratefully acknowledge the help of LLM tools that were used in editing this manuscript to make it more readable.

References

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–336. ACM.

Laming Chen, Guoxin Zhang, Eric Zhou, Hanning Zhou, and Michael Zheng Shou. 2018. Fast greedy map inference for determinantal point process to improve recommendation diversity. In *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, pages 5625–5636. Curran Associates, Inc.

Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms elegance analysis and alpha-ndcg. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 658–659.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. 2021. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:386–397.

Ryota Hinami and Shin'ichi Satoh. 2018. Discriminative learning of open-vocabulary object retrieval and localization by negative phrase augmentation. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 2605–2615.

John E. Hopcroft and Richard M. Karp. 1971. A $n^5/2$ algorithm for maximum matchings in bipartite graphs. *SIAM J. Comput.*, 2:225–231.

Alex Kulesza, Ben Taskar, and 1 others. 2012. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Bryan Schaffner, Caiming Gondal, and Steven Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34:9694–9705.

Leigang Qu, Meng Liu, Wenjie Wang, Zhedong Zheng, Liqiang Nie, and Tat-Seng Chua. 2023. Learnable pillar-based re-ranking for image-text retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, pages 1252–1261. ACM.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 3982–3992.

Yining Wang, Liwei Wang, Yuanzhi Li, Di He, and Tie-Yan Liu. 2013. A theoretical analysis of ndcg type ranking measures. In *Annual Conference Computational Learning Theory*.

Shangrong Wu, Yanghong Zhou, Yang Chen, Feng Zhang, and P. Y. Mok. 2025. Chain-of-thought re-ranking for image retrieval tasks. *arXiv preprint arXiv:2509.14746*.

Dongchen Han Shiji Song Xuran Pan, Tianzhu Ye and Gao Huang. 2022. Contrastive language-image pre-training with knowledge graphs. In *NeurIPS*, pages 245–345.

Yusuke Yasumoto, Yile Wang, Hao Zhang, and Jiashi Feng. 2024. Open-vocabulary multi-label classification via multi-modal reranking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14023–14033.

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

Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih Fu Chang. 2021. Open-vocabulary object detection using captions. In *Proceedings of the 2021 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14388–14397.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023a. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11975–11986.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023b. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11975–11986.