Multimodal Learned Sparse Retrieval for Image Suggestion Task

Thong Nguyen Mariya Hendriksen Andrew Yates
University of Amsterdam
Amsterdam, The Netherlands
{t.nguyen2,m.hendriksen,a.c.yates}@uva.nl

ABSTRACT

Learned Sparse Retrieval (LSR) is a group of neural methods designed to encode queries and documents into sparse lexical vectors. These vectors can be efficiently indexed and retrieved using an inverted index. While LSR has shown promise in text retrieval, its potential in multi-modal retrieval remains largely unexplored. Motivated by this, in this work we explore the application of LSR in the multi-modal domain, i.e., we focus on Multi-Modal Learned Sparse Retrieval (MLSR). We conduct experiments using several MLSR model configurations and evaluate the performance on the image suggestion task. We find that solving the task solely based on the image content is challenging. Enriching the image content with its caption improves the model's performance significantly, implying the importance of image captions to provide fine-grained concepts and context information of images. Our approach presents a practical and effective solution for training LSR retrieval models in multi-modal settings.

ACM Reference Format:
Nguyen, Hendriksen, Yates. 2023. Multimodal Learned Sparse Retrieval for Image Suggestion Task. In *Proceedings of (TREC 2023).* ACM, New York, NY, USA, 5 pages.

1 INTRODUCTION

Learned Sparse Retrieval (LSR) is a neural retrieval method that encodes queries and documents to bags of tokens, which could be indexed and retrieved efficiently by an inverted index.

Image Suggestion Task. The task is defined as follows: given a query and a set of candidates, we rank all candidates w.r.t. their relevance to the query. The query is a text, whereas the set of candidate items are images. Hence, we aim to retrieve relevant images that accurately describe the textual query.

Vision-Language Learned Sparse Retriever. We propose to address the task using a Vision-Language Sparse Retriever approach. The architecture is a bi-encoder, including a query encoder and document encoder. Both query and document encoders are based on transformer encoder architecture with either a Multi-Layer Perceptron (MLP) or Masked Language Model (MLM) sparse projection head on top. We experiment with the following model configurations:

- $M_{\mathcal{T}\rightarrow\mathcal{I}}$: the model uses image information to build document representations.
- $M_{\mathcal{T}\rightarrow\mathcal{T}}$: the model uses textual information to build document representations.
- $M_{\mathcal{T}\rightarrow\mathcal{T}+\mathcal{I}}$: the model uses both textual and image information to build document representations.

2 APPROACH

We follow the same terminology and notation as in [14, 32]. The input dataset can be represented as image-text pairs. For the image suggestion task, we use textual data as a query and we aim to retrieve a ranked list of top-k images that describe the textual query.

Vision-Language Learned Sparse Retrieval. To address the task, we explored the application of learned sparse retrieval that leverages visual and textual information. The model comprises a query and document encoder. Depending on the modality of the query and document, the encoder could either be an MLP and MLM encoder. The MLP encoder can only used with text modality, while the MLM is applicable to both textual and visual data. Note that in this task, the document could be either an image, caption, or both. Figure 1 illustrates the architecture of the MLP and MLM encoder.

MLP Encoder. The MLP encoder takes a text as input and produces an important weight for each token of the input text. For example, for the input text "*text image retrieval*", an MLP encoder outputs weights, such as *{"text": 10 , "image": 20 , "retrieval": 50 }.*

An MLP encoder is a network that takes a sequence of contextualized embeddings h_j produced by the dense encoder for each input term to generate the term's score:

$$w_i(t) = \sum_{j=1...L} log\left(\mathbb{1}(v_i = t_j)\left(ReLU(h_jW + b)\right) + 1\right) \qquad (1)$$

where w_i is the *i-th* token in the vocabulary $\mathcal{V}$.

As described in Figure 1, an MLP is comprised of a Linear layer on top of a transformer encoder. The Linear layer takes the last hidden states of the transformer's encoder as input and projects each state to a positive scalar representing the weight of the corresponding input token.

An MLP encoder requires the input to be tokenized into a sequence of vocabulary words; therefore, it can only encode textual data and not images. In addition, MLP encoder does not have the capability to expand the input to relevant terms.

MLM Encoder. Unlike the MLP encoder, the MLM encoder can be applied to both text and image and has the freedom to expand the input to any relevant terms in the vocabulary. As described in Figure 1 (right), the MLM encoder consists of a transformer encoder and a sparse projection layer on top. The transformer encoder takes either a textual or visual input and outputs a single intermediate dense vector which is then passed into a sparse MLM projection:

TREC 2023, 14-18 November, TBA, MD, USA

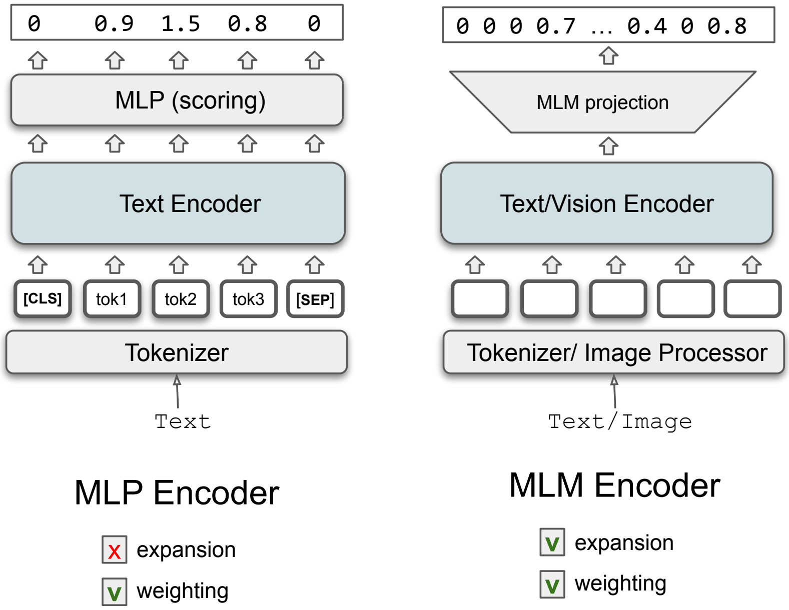


Figure 1: MLP and MLM encoder

$$w_i(t) = ReLU(h_0^{\intercal} e_i + b_i) \tag{2}$$

where w_i is the *i-th* item (token) in the vocabulary $\mathcal{V}$.

2.1 Full Bi-Encoder Configuration

The full bi-encoder configuration includes a query encoder, f_θ^Q, and a document encoder, $f_\phi^{\mathcal{D}}$. Each query and document encoder is either a MLP or MLM as described in the previous section. In this paper, we experimented with different configurations as listed in Table 1. We did not explore all possible configurations due to resource constraints. Among all of the abovementioned configurations, the first model (M1) is the multi-modal version of Splade [8]. M2 is the multi-modal version of EPIC[27]. In M3 and M4, we re-use the MLP query encoder from M2 to encode the caption associated with each image.

Model	Query encoder f_θ^Q	Document Encoder $f_\phi^{\mathcal{D}}$	
		Caption	**Image**
M1	MLM	-	MLM
M2	MLP	-	MLM
M3	MLP	MLP (*)	-
M4	MLP	MLP (*)	MLM

Table 1: Sparse Bi-encoder Variants. (*) The MLP encoder is re-used from the query side for encoding caption.

3 EXPERIMENTS

Dataset. We conducted our experiments on the AToMiC dataset[46] which has around 11 million images and more than 10 million text queries collected from Wikipedia. Each text query contains four different fields, including *page_title*, *section_title*, *context_page_description*, *context_section_description*. We exclude the *context_page_description* and concatenate the remaining fields with space between them to form a single text query. Each image in the dataset also comes with a multilingual caption (*caption_reference_description*), we tested different configurations where we use either or both of them to represent the document. Due to computing resource limitations, our experiments were only conducted on images with English captions. For training, we only use the 4.4 text-image pairs provided by the dataset's authors and train for 5 epochs with InforNCE loss and only in-batch negatives.

Metrics. To evaluate model performance, we report NDCG@k, MAP@K, and R@k where $k = \{5, 10, 100, 500, 1000\}$.

4 DISCUSSION

One of our criteria for model selection is that the model should produce meaningful, interpretable sparse vectors (bags of terms). However, we found that model M1 does not meet this requirement. As demonstrated in Table 5, M1 generates terms that do not reflect the content of the image and are difficult for humans to interpret. The reason is that M1 uses the MLM encoder both on query and document sides, allowing the input text and image to be projected into any latent dimensions as long as these dimensions are co-activated (having non-zero values) in both query and document

Table 2: Performance of the submitted models on the image suggestion task, evaluated on NDCG@k scores.

Run	NDCG@5	NDCG@10	NDCG@100	NDCG@500	NDCG@1000
$M_{\mathcal{T}\rightarrow\mathcal{I}}$	0.40	0.39	0.28	0.50	0.58
$M_{\mathcal{T}\rightarrow\mathcal{T}}$	9.39	**10.79**	15.59	18.67	19.59
$M_{\mathcal{T}\rightarrow\mathcal{T}+\mathcal{I}}$	**9.77**	10.74	**15.63**	**18.79**	**19.68**

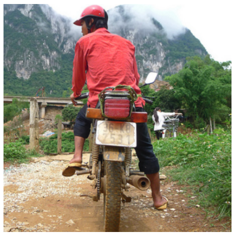
Table 3: Performance of the submitted models on the image suggestion task, evaluated on MAP@k scores.

Run	MAP@5	MAP@10	MAP@100	MAP@500	MAP@1000
$M_{\mathcal{T}\rightarrow\mathcal{I}}$	0.02	0.04	0.04	0.05	0.05
$M_{\mathcal{T}\rightarrow\mathcal{T}}$	3.51	**4.48**	5.82	6.04	6.07
$M_{\mathcal{T}\rightarrow\mathcal{T}+\mathcal{I}}$	**3.58**	**4.48**	**5.84**	**6.07**	**6.10**

Table 4: Performance of the submitted models on the image suggestion task, evaluated on Recall@k scores.

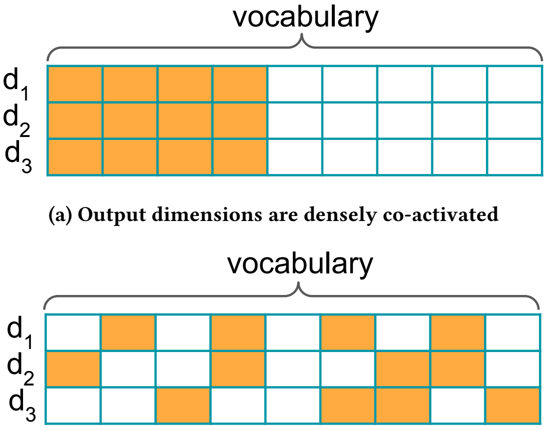
Run	R@20	R@100	R@500	R@1000
$M_{\mathcal{T}\rightarrow\mathcal{I}}$	0.37	0.43	1.31	1.61
$M_{\mathcal{T}\rightarrow\mathcal{T}}$	**15.54**	24.54	37.00	41.24
$M_{\mathcal{T}\rightarrow\mathcal{T}+\mathcal{I}}$	15.50	**24.57**	**37.48**	**41.37**

Table 5: Demonstration of not interpretable output. Top-10 highest-scored terms are shown for both M1 and M2 models.

Image	M2 (MLP, MLM)	M1 (MLM, MLM)
	mountain mountains bike bee dirt mo red path ##oot person	accent ship natural de crown yourself " ra now wild

representations. The projection freedom of MLM leads to the issue of high co-activation, forming a sub-dense space inside the vocabulary space as described in Figure 2. This high co-activation could also harm the retrieval efficiency of the inverted index employed in LSR. Because of the above issues, we did not submit the run from M1 for evaluation.

One solution to the above issues is to constrain the projection to semantically relevant terms in the vocabulary. This constraint could be achieved by using the MLP encoder on the query side. Indeed, the design of the MLP encoder only allows it to score the impact of the input tokens but does not allow it to expand to other tokens. MLP only produces positive weights for tokens in the input query and keeps the remaining tokens in the vocabulary to be zero. In order to match an image to a relevant query, the image encoder (MLM) needs to project the image into terms appearing in the relevant query. This constraint forces the model to produce more interpretable and sparsely co-activated image representation. However, we observe that using the MLP query encoder still does not prevent the problems entirely as it could still rely on stop words and punctuation marks for encoding latent senses. The stop words and punctuation marks are especially popular in long texts, as in the Atomic queries, which are taken from Wikipedia articles. For this reason, we resorted to using the short caption and image pairs for training our M2, M3, and M4 models. An example output of M2's document encoder-trained caption, image pairs is shown in Table 5.



(a) Output dimensions are densely co-activated

(b) Output dimensions are sparsely co-activated

Figure 2: Dense activation vs. sparse co-activation. All documents have 4 vocab terms activated (yellow colors).

Regarding the effectiveness, for each model (M2, M3, M4), we submitted one run for evaluation. The results are shown in Table 4, Table 3, and Table 2 for Recall, MAP, and NDCG respectively. We observe that M2 performs poorly across different metrics, implying that the task of image suggestion for writing assistants solely based on the images' content is a challenging task. We hypothesize that the context information of an image, which is critical for the task, could not be encoded in the images, but in text captions. Given a picture of World War 2 (WW2), for example, it is generally very difficult for a model to predict that this picture is about WW2 because any war picture has similar concepts (e.g., "soldiers", "weapon"). Similarly, given a picture of a less-popular street in Amsterdam, it is difficult to infer any terms relevant to Amsterdam based on the image content only. For this reason, we argue that to solve the image suggestion task well, the image caption should be used to provide more fine-grained and contextual information that is challenging to infer from the visual data. This argument is supported by the result of our second run produced by my M3 model. By using only image caption, the M3 model could outperform M1 significantly and by a large margin. The result of M4 also shows that using both images and captions could slightly improve the overall performance of the task, but the improvement is not consistent across different metrics.

5 RELATED WORK

Learned sparse retrieval (LSR). LSR is a neural retrieval method encoding queries and documents into sparse lexical vectors, efficiently indexed and searched with an inverted index. Various LSR approaches exist, using MLP or MLM encoders [8, 31, 48]. MLP encoders predict term importance without expansion, while MLM encoders use masked language model logits for weighting and expansion. Splade is a recent text-oriented LSR approach employing MLM encoders [7, 8], while other methods use MLP encoders [4, 22, 27]. Recent research suggests a cancellation effect between query and document expansion [32].

Cross-Modal Retrieval (CMR). CMR methods create a multimodal representation space, measuring concept similarity across modalities [2]. Early CMR approaches used canonical correlation analysis [11, 15], followed by RNN-CNN encoders with hinge loss [9, 44], hard-negative mining [6] and attention mechanisms such as dual attention, stacked cross-attention, and bidirectional focal attention [17, 25, 29, 37]. Other approaches include modality-specific graphs [43] and image-text generation modules [12]. Domain-specific research targets CMR in fashion [10, 16], e-commerce [13], conversational systems [35, 36, 38], and music video recommendations [42].

Recent methods use transformer-based dual encoders trained on extensive data. ALBEF [19] aligns unimodal representations before fusion, X-VLM [49] adds a cross-modal encoder for fine-grained VL representations. Florence [47] uses adaptation models for object-level representations, and CLIP [33] predicts image-caption pairs. ALIGN [19] uses a dual encoder on image alt-text pairs.

Another line of work adopts transformer encoders [41] for the CMR task [28], adapting BERT-like models [5]. ViLBERT [26] and LXMERT [40] introduce a two-stream architecture, while B2T2 [1], VisualBERT [20], Unicoder-VL [18], VL-BERT [39], and UNITER [3] feature single-stream architectures. Oscar [21] incorporates caption object tags with region features, and BEIT-3 [45] adapts multiway transformers.

Our focus. Unlike prior work that focuses on sparse to dense conversion [23, 24], we focus on dense to sparse conversion in the multi-modal domain. Challenges include dimension co-activation and semantic deviation [34].

6 CONCLUSION

In this work, we explored the application of learned sparse retrieval (LSR) for the image suggestion task to support multimedia content creation. We identify the challenges that arise when transferring state-of-the-art LSR techniques from the text domain to the multi-modal domain and propose a simple solution to mitigate the problems. We analyze the effectiveness of our trained models with various configurations and conclude that using image captions is critical for the task as image captions provide fine-grained concepts and context information that are difficult to encode in the visual content itself. We address this problem in the follow-up work [30].

REFERENCES

[1] Chris Alberti, Jeffrey Ling, Michael Collins, and David Reitter. 2019. Fusion of Detected Objects in Text for Visual Question Answering. In *EMNLP*. 2131–2140.

[2] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 2 (2019), 423–443.

[3] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: Universal image-text representation learning. In *ECCV*. 104–120.

[4] Zhuyun Dai and Jamie Callan. 2019. Context-aware sentence/passage term importance estimation for first stage retrieval. *arXiv preprint arXiv:1910.10687* (2019).

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[6] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*. BMVA Press, 12.

[7] Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2022. From Distillation to Hard Negative Sampling: Making Sparse Neural IR Models More Effective. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 2353–2359.

[8] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. 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*. 2288–2292.

[9] Andrea Frome, Greg Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. Devise: A deep visual-semantic embedding model. (2013).

[10] Kenneth Goei, Mariya Hendriksen, and Maarten de Rijke. 2021. Tackling Attribute Fine-grainedness in Cross-modal Fashion Search with Multi-level Features. In *SIGIR 2021 Workshop on eCommerce*. ACM.

[11] Yunchao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. 2014. Improving image-sentence embeddings using large weakly annotated photo collections. In *European Conference on Computer Vision*. Springer, 529–545.

[12] Jiuxiang Gu, Jianfei Cai, Shafiq R Joty, Li Niu, and Gang Wang. 2018. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7181–7189.

[13] Mariya Hendriksen, Maurits Bleeker, Svitlana Vakulenko, Nanne van Noord, Ernst Kuiper, and Maarten de Rijke. 2022. Extending CLIP for Category-to-image Retrieval in E-commerce. In *European Conference on Information Retrieval*. Springer, 289–303.

[14] Mariya Hendriksen, Svitlana Vakulenko, Ernst Kuiper, and Maarten de Rijke. 2023. Scene-Centric vs. Object-Centric Image-Text Cross-Modal Retrieval: A Reproducibility Study. In *European Conference on Information Retrieval*. Springer, 68–85.

[15] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. 2014. Fisher vectors derived from hybrid gaussian-laplacian mixture models for image annotation. *arXiv preprint arXiv:1411.7399* (2014).

[16] Katrien Laenen, Susana Zoghbi, and Marie-Francine Moens. 2018. Web search of fashion items with multimodal querying. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. 342–350.

[17] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 201–216.

[18] Gen Li, Nan Duan, Yuejian Fang, Daxin Jiang, and Ming Zhou. 2020. Unicoder-VL: A Universal Encoder for Vision and Language by Cross-modal Pre-training. In *AAAI*.

[19] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems* 34 (2021), 9694–9705.

[20] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557* (2019).

[21] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*. 121–137.

[22] Jimmy Lin and Xueguang Ma. 2021. A few brief notes on deepimpact, coil, and a conceptual framework for information retrieval techniques. *arXiv preprint arXiv:2106.14807* (2021).

[23] Sheng-Chieh Lin and Jimmy Lin. 2021. Densifying sparse representations for passage retrieval by representational slicing. *arXiv preprint arXiv:2112.04666*

(2021).

[24] Sheng-Chieh Lin and Jimmy Lin. 2023. A dense representation framework for lexical and semantic matching. *ACM Transactions on Information Systems* 41, 4 (2023), 1–29.

[25] Chunxiao Liu, Zhendong Mao, An-An Liu, Tianzhu Zhang, Bin Wang, and Yongdong Zhang. 2019. Focus your attention: A bidirectional focal attention network for image-text matching. In *Proceedings of the 27th ACM International Conference on Multimedia*. 3–11.

[26] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *NeurIPS*. 13–23.

[27] Sean MacAvaney, Franco Maria Nardini, Raffaele Perego, Nicola Tonellotto, Nazli Goharian, and Ophir Frieder. 2020. Expansion via prediction of importance with contextualization. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*. 1573–1576.

[28] Nicola Messina, Giuseppe Amato, Andrea Esuli, Fabrizio Falchi, Claudio Gennaro, and Stéphane Marchand-Maillet. 2021. Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 17, 4 (2021), 1–23.

[29] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. 2017. Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 299–307.

[30] Thong Nguyen, Mariya Hendriksen, Andrew Yates, and Maarten de Rijke. 2024. Multimodal Learned Sparse Retrieval with Probabilistic Expansion Control. In *ECIR 2024: 46th European Conference on Information Retrieval*. Springer.

[31] Thong Nguyen, Sean MacAvaney, and Andrew Yates. 2023. Adapting Learned Sparse Retrieval for Long Documents. *arXiv preprint arXiv:2305.18494* (2023).

[32] Thong Nguyen, Sean MacAvaney, and Andrew Yates. 2023. A Unified Framework for Learned Sparse Retrieval. In *European Conference on Information Retrieval*. Springer, 101–116.

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[34] Ori Ram, Liat Bezalel, Adi Zicher, Yonatan Belinkov, Jonathan Berant, and Amir Globerson. 2023. What Are You Token About? Dense Retrieval as Distributions Over the Vocabulary. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 2481–2498.

[35] Jerome Ramos, To Eun Kim, Zhengxiang Shi, Xiao Fu, Fanghua Ye, Yue Feng, and Aldo Lipani. 2022. Condita: A state machine like architecture for multimodal task bots. *Alexa Prize TaskBot Challenge Proceedings* (2022).

[36] Zhengxiang Shi and Aldo Lipani. 2023. DePT: Decomposed Prompt Tuning for Parameter-Efficient Fine-tuning. *arXiv preprint arXiv:2309.05173* (2023).

[37] Zhengxiang Shi, Pin Ni, Meihui Wang, To Eun Kim, and Aldo Lipani. 2022. Attention-based ingredient phrase parser. *arXiv preprint arXiv:2210.02535* (2022).

[38] Zhengxiang Shi, Procheta Sen, and Aldo Lipani. 2023. Lexical Entrainment for Conversational Systems. *arXiv preprint arXiv:2310.09651* (2023).

[39] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. In *ICLR*.

[40] Hao Tan and Mohit Bansal. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *EMNLP-IJCNLP*. 5099–5110.

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* 30 (2017).

[42] Karel Veldkamp, Mariya Hendriksen, Zoltán Szlávik, and Alexander Keijser. 2023. Towards contrastive learning in music video domain. *arXiv preprint arXiv:2309.00347* (2023).

[43] Hao Wang, Doyen Sahoo, Chenghao Liu, Ke Shu, Palakorn Achananuparp, Ee-peng Lim, and Steven CH Hoi. 2021. Cross-modal food retrieval: learning a joint embedding of food images and recipes with semantic consistency and attention mechanism. *IEEE Transactions on Multimedia* 24 (2021), 2515–2525.

[44] Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5005–5013.

[45] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. 2022. Image as a foreign language: BEiT pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442* (2022).

[46] Jheng-Hong Yang, Carlos Lassance, Rafael Sampaio De Rezende, Krishna Srinivasan, Miriam Redi, Stéphane Clinchant, and Jimmy Lin. 2023. AToMiC: An Image/Text Retrieval Test Collection to Support Multimedia Content Creation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2975–2984.

[47] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. 2021. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432* (2021).

[48] Hamed Zamani, Mostafa Dehghani, W Bruce Croft, Erik Learned-Miller, and Jaap Kamps. 2018. From neural re-ranking to neural ranking: Learning a sparse representation for inverted indexing. In *Proceedings of the 27th ACM international conference on information and knowledge management*. 497–506.

[49] Yan Zeng, Xinsong Zhang, and Hang Li. 2022. Multi-Grained Vision Language Pre-Training: Aligning Texts with Visual Concepts. In *ICML*. 25994–26009.