ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10446291

IMAGE RETRIEVAL WITH COMPOSED QUERY BY MULTI-SCALE MULTI-MODAL FUSION

*Zelong Sun, Guoxing Yang, Zhiwu Lu**

Gaoling School of Artificial Intelligence
Renmin University of China, Beijing, China

Hao Jiang, Guojie Zhu, Zhao Cao

Huawei Possion Lab
Hangzhou, Zhejiang, China

ABSTRACT

Image retrieval with composed query (IR-CQ) is a challenging task since it aims to retrieve the target image according to a hybrid-modality query which consists of a reference image and a text modifier. Previous approaches mainly focus on designing various multi-modal fusion modules to fuse the hybrid-modality query, but these fusion modules are often suboptimal without considering sufficient fusion between the two modalities. In this paper, we propose a general fusion block by taking three fusion strategies: weighted summing, concatenating, and bilinear pooling. Importantly, this general fusion block can be deployed to fuse not only the hybrid-modality query but also the multi-scale features of the reference image. Specifically, we first fuse the multi-scale features of the reference image with the Multi-Scale Fusion (MSF) block and then fuse the features of the reference image and text modifier with the Multi-Modal Fusion (MMF) block, where both MSF and MMF are instantiations of our general fusion block. Extensive experiments on three benchmark datasets show that our proposed model significantly outperforms existing approaches.

***Index Terms*—** Linguistic-Visual Composition, Image Retrieval, Bilinear Pooling Fusion

1. INTRODUCTION

In contrast to the two classical paradigms of image retrieval: image-to-image matching [1] and text-to-image matching [2], whose query is limited to a single modality, image retrieval with composed query (IR-CQ) task [3, 4] incorporates the client's intent into the query in the form of text. Therefore, IR-CQ faces a key challenge in how to better fuse the semantic information of the two modalities, so that the retrieved image can retain most of the attributes of the reference image while satisfying the requirements of the text modifier.

A few recent works have been devoted to the IR-CQ task, and most of them focus on designing various modules to fuse reference images and text modifiers. TIRG [3] proposed to deploy gating and residual modules to compose the visual and textual representations. However, it does not sufficiently consider a wide range of input content and styles. VAL [4] utilized visual multi-level feature maps and fused each of them respectively with the text feature, followed by introducing a hierarchical matching strategy. However, it demands much more computational resources to utilize three-level feature maps. Clip4Cir [5] proposed a combiner module to combine the visual and text features extracted by CLIP [6] and achieved superior results on the IR-CQ task. However, it does not fully explore multi-modal fusion and also fails to utilize the local representation of the image.

In this work, we propose a general fusion block to better learn the joint representation of the hybrid-modality query for IR-CQ. The general fusion block takes three fusion strategies: weighted summing, concatenating, and bilinear pooling. In particular, bilinear pooling [7] can provide more effective feature fusion than linear fusion strategies like summing and concatenating. To fully explore the fine-grained information of the image, we utilize the multi-scale feature maps from the penultimate convolution layer of the image encoder. Different from the hierarchical matching strategy as in previous work [4], we fuse the two-level features of the image in advance with a Multi-Scale Fusion (MSF) block before fusing visual and text features with a Multi-Modality Fusion (MMF) block, which is much more efficient. Note that both MSF and MMF are instantiations of our general fusion block. To further reduce memory consumption, we use average pooling for the feature maps and project feature vectors to a lower-dimensional space where appropriate. These choices enable us to train our model with a much larger batch size and thus include more negative samples for contrastive learning which is needed in model training.

The main contributions of this work are as follows:
(1) We propose a general fusion block, which can be deployed to fuse not only the hybrid-modality query but also the multi-scale features of the reference image.
(2) Instead of a hierarchical matching strategy, we fuse the hybrid-modality query features in a more efficient way, where the multi-scale visual features are first fused together, and then the two-modality features are fused.
(3) Extensive experiments show that our proposed model

*Corresponding author

achieves new state-of-the-art on three benchmark datasets. Moreover, based on our retrieval model, the conditional image generation experiments further demonstrate its effectiveness and generalizability.

2. PROPOSED MODEL

2.1. Architecture Overview

As illustrated in Fig.1(a), our proposed model consists of four major coThe parameters of IE, MSF, and the projection head are shared by the reference image and target image.

2.1.1. Multi-Scale Fusion Block

As convolutional neural networks (CNNs) can learn visual concepts in a composed and hierarchical way [8], we are not able to effectively utilize the visual information at different granularities by only using the final image features from IE. Thus, similar to [4], we extract not only the final feature $x_r^F \in \mathbb{R}^D$ of the reference image I_r from the last layer of IE, but also the middle feature map $x_r^M \in \mathbb{R}^{h*w*c}$ of I_r from the penultimate convolution block of IE. Formally, the multi-scale features of the reference image I_r can be represented as:

$$\{x_r^M, x_r^F\} = IE(I_r) \tag{1}$$

To ensure that x_r^M is compatible with $x_r^F \in \mathbb{R}^D$ for further processing, we further utilize a projection head to map x_r^M to $\nu_r \in \mathbb{R}^D$. The projection head includes a pooling function (e.g. average pooling), a flattening operation, and a linear layer. After that, we feed ν_r together with x^F to the MSF block to fuse the information from different levels, and thus obtain the final image feature $\xi_r \in \mathbb{R}^D$ as follows:

$$\xi_r = MSF(\nu_r, x_r^F) \tag{2}$$

We finally use $\xi_r \in \mathbb{R}^D$ to represent the reference image I_r, which combines the low-level detail information and high-level semantic information of I_r. Note that the target image I_t can be processed in the same way. In this work, the parameters of IE, MSF, and the projection head are shared by the reference image and target image.

2.1.2. Multi-Modal Fusion Block

To learn the joint representation of the reference image I_r and text modifier T_m, we utilize the MMF block to transform the features of the reference image and text modifier into a fused feature. The MMF block takes $\xi_r \in \mathbb{R}^D$ obtained from the MSF block and $x_m \in \mathbb{R}^D$ extracted by TE as inputs, and outputs the fused feature $\psi \in \mathbb{R}^D$. The process of multi-modal fusion can be described as:

$$\psi = MMF(\xi_r, x_m) \tag{3}$$

Note that the final fused feature ψ includes both the multi-modal semantics of the hybrid-modality query and the multi-scale information of the reference image.

2.2. General Fusion Block

In this work, we propose a general fusion block for both multi-scale fusion and multi-modal fusion. Specifically, as shown in Fig.1(b), the general fusion block fuses two original features with three strategies: weighted summing, concatenating, and bilinear pooling. Since the effectiveness of the first two strategies has been shown in [5], we mainly introduce the bilinear pooling strategy below.

Bilinear pooling can create a joint representation space by computing the outer product of two feature vectors. By facilitating multiplicative interactions between all elements of both vectors, it provides more effective feature representation than the linear strategies like summing and concatenating. In this work, we thus introduce a bilinear pooling strategy for learning the joint representation of two feature vectors.

As shown in Fig 1(b), the bilinear component has a triple (f_A, f_B, O), where f_A and f_B are feature projection blocks that are mappings: $\mathbb{R}^D \rightarrow \mathbb{R}^K$, and O is an output block that is also a mapping: $\mathbb{R}^{K^2} \rightarrow \mathbb{R}^D$. Given two original features $A, B \in \mathbb{R}^D$, their bilinear combination is:

$$\eta = bilinear(A, B, f_A, f_B) = f_A(A)^T f_B(B) \tag{4}$$

where A, B can come from different modalities or different scales. We then flatten $\eta \in \mathbb{R}^{K*K}$ into a vector and feed it to the output block O to obtain the final result.

In this work, we utilize the same structure of feature projection block, output block, and scaler block as described in [5]. However, considering the memory overhead of storing the high-dimensional features for bilinear pooling, the feature projection blocks in Bilinear Fusion block are devised to project features to a lower-dimensional space. Although our general fusion block has common elements with the combiner module proposed by [5], there exist two crucial differences between them: (1) In contrast to the combiner, bilinear pooling is additionally considered in our general fusion block, which has been shown to be effective in Section 3.3. (2) The inputs A, B of the combiner are from two modalities (e.g., I_r and T_m). However, the inputs of our general fusion block can be from not only different modalities but also different scales.

2.3. Model Training

The goal of training our proposed model for IR-CQ is to match the representation ψ of the hybrid-modality query (I_r, T_m) with the representation ξ_t of the target image I_t. At each iteration, we have a mini-batch $\{(\psi^{(i)}, \xi_t^{(i)})\}_{i=1}^{N_B}$, where $(\psi^{(i)}, \xi_t^{(i)})$ denotes the i-th pair of (hybrid-modality query, target image), and N_B denotes the size of the mini-batch. Following [3], we define the batch-based classification

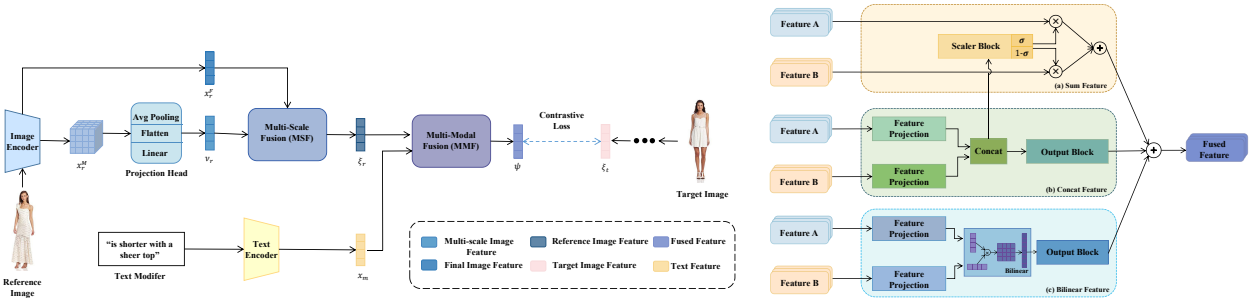

Fig. 1. Illustration of the network architecture of (a) our proposed model, and (b) the general fusion block.

Table 1. Comparison with the state-of-the-art methods on the Fashion-IQ dataset and Shoes dataset. † denotes that ResNet50x4 is used as the image encoder. The second-best result is marked by underline.

Method	Fashion-IQ									Shoes			
	Shirt		Dress		Tops&Tees		Avg						
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50	Rmean	R@10	R@10	R@50	Rmean
TIRG [3]	13.10	30.91	14.13	34.61	14.79	34.37	14.01	33.30	23.66	12.60	45.45	69.39	42.48
VAL (GloVe) [4]	22.38	44.15	22.53	44.00	27.53	51.68	24.15	46.61	35.38	17.18	51.52	75.83	48.18
ARTEMIS [9]	21.78	43.64	27.16	52.40	29.20	54.83	26.05	50.29	38.17	18.72	53.11	79.31	50.38
CoSMo [10]	24.90	49.18	25.64	50.30	29.21	57.46	26.58	52.31	39.45	16.72	48.36	75.64	46.91
DCNet [11]	23.95	47.30	28.95	56.07	30.44	58.29	27.78	53.89	40.84	-	53.82	79.33	-
CLVC-Net [12]	28.75	54.76	29.85	56.47	33.50	64.00	30.70	58.41	44.56	17.64	54.39	79.47	50.50
ContionedIR† [13]	35.76	56.20	27.20	53.57	36.31	61.14	33.09	56.99	45.04	-	-	-	-
Clip4Cir† [5]	39.99	60.45	33.81	59.40	41.41	65.37	38.32	61.74	50.03	21.42	56.69	81.52	53.21
PLIR [14]	39.45	61.78	33.60	58.90	43.96	68.33	39.02	63.00	51.01	22.88	58.83	**84.16**	55.29
Ours† (TE only)	42.29	63.10	35.40	60.23	44.97	66.95	40.89	63.43	52.16	-	-	-	-
Ours† (Both)	**44.85**	**66.29**	**40.45**	**65.29**	**49.26**	**70.98**	**44.85**	**67.52**	**56.19**	**24.94**	**61.08**	83.78	**56.57**

(BBC) loss for model training:

$$L = \frac{1}{N_B}\sum_{i=1}^{N_B} -\log \frac{\exp(\lambda * s(\psi^{(i)}, \xi_t^{(i)}))}{\sum_{j=1}^{N_B} \exp(\lambda * s(\psi^{(j)}, \xi_t^{(j)}))} \quad (5)$$

where $s(\cdot)$ denotes the cosine similarity function, and λ denotes a temperature parameter.

3. EXPERIMENTS

3.1. Datasets and Settings

3.1.1. Datasets

We make evaluations on three benchmark IR-CQ datasets: **(1) Fashion-IQ** [15] is divided into three different categories: *Dress*, *Shirt*, and *Tops&tees*. Following [5, 14], we use $18,000$ triplets for training, and $6,016$ triplets for testing, **(2) Shoes** [16] consists of 10,000 training images for training and 4,658 test images for evaluation and **(3) CIRR** [17] consists of 21,552 real-life images derived from the popular natural language reasoning $NLVR^2$ dataset. Following the standard split, we use 28,225 triplets for training, 2,297 triplets for validating, and 2,315 triplets for testing.

3.1.2. Implementation Details

As in Clip4Cir [5], the image encoder is defined by ResNet50x4, while the text encoder is defined by 12-layers BERT [18]. In this work, these two encoders are initialized with CLIP [6] and then frozen during training our model. The CLIP model used for initialization is obtained by fine-tuning it in two different ways: (1) *TE only* – only the text encoder is fine-tuned, as in Clip4Cir [5]; (2) *Both* – both encoders are fine-tuned. During fine-tuning the CLIP model, we use the AdamW [19] optimizer, and set the learning rate to 1e-6 (with a weight decay coefficient of 1e-2). We choose to fine-tune the CLIP model for 20 epochs (with a batch size of 128).

After CLIP fine-tuning, we freeze the image and text encoders and train the rest of our model with the Adam [20] optimizer (with the learning rate 2e-5). We train our model for a maximum of 300 epochs. The temperature parameter and batch size are respectively set to 70 and 4,096 for Fashion-IQ, 85 and 4096 for CIRR, and 30 and 2,048 for Shoes. We implement our model with PyTorch. We use a single NVIDIA A100-SXM4-80GB for CLIP fine-tuning and model training.

3.2. Comparison with State-of-the-Art Methods

Table 1 presents the comparative results on the Fashion-IQ dataset and the Shoes dataset, respectively. It can be observed that: **(1)** Our proposed model outperforms all the

Table 2. Comparison with the state-of-the-art methods on the CIRR dataset.

Method	Recall@K				R_{sub}@K			$\frac{R@5+R_{sub}@1}{2}$
	K=1	K=5	K=10	K=50	K=1	K=2	K=3	
TIRG [3]	14.61	48.37	64.08	90.03	22.67	44.97	65.14	35.52
MAAF+BERT [21]	10.12	33.10	48.01	80.57	22.04	42.41	62.14	27.57
MAAF-RP [21]	10.22	33.32	48.68	81.84	21.41	42.17	61.60	27.36
ARTEMIS [9]	16.96	46.10	61.31	87.73	39.99	62.20	75.67	43.04
CIRPLANT [17]	19.55	52.55	68.39	92.38	39.20	63.03	79.49	45.87
Clip4Cir [5]	38.53	69.98	81.86	95.93	68.19	85.64	94.17	69.08
Ours (TE only)	39.88	72.75	83.86	96.92	68.65	86.48	94.35	70.70
Ours (Both)	**43.80**	**77.37**	**86.94**	**97.71**	**71.39**	**88.16**	**95.15**	**74.38**

Table 3. Ablation study results on the Fashion-IQ dataset. The second-best result is marked by underline.

Method	Avg		
	R@10	R@50	Rmean
MMF w/o bilinear	38.32	61.74	50.03
MMF w/ bilinear	39.71	62.57	51.14
MMF+MSF w/o bilinear	39.82	62.50	51.16
MSMMF w/ bilinear	39.58	**63.86**	51.72
MMF+MSF w/ bilinear	**40.89**	63.43	**52.16**

representative/state-of-the-art methods by large margins in all cases on both Fashion-IQ and Shoes. This clearly demonstrates the effectiveness of the combination of multi-scale fusion and multi-scale fusion for the IR-CQ task. **(2)** As expected, CLIP-based methods (including ContionedIR [13], Clip4Cir [5], PLIR [14], and our model) achieve significant improvements over those without using CLIP in terms of average results, showing that the large-scale multi-model pretraining models like CLIP play an important role in the IR-CQ task. **(3)** Compared to Clip4Cir [5] (which is the baseline of our model), our proposed model ('Both') leads to an improvement of 6.16% on Rmean on Fashion-IQ dataset and 1.28% on Shoes dataset, thanks to the extra use of bilinear pooling fusion and multi-scale fusion.

Table 2 further shows the results on the CIRR test set. We have the following observations: **(1)** Compared to Clip4Cir [5], our proposed models lead to an improvement of 5.27% on Recall@5, which suggests that our proposed approach has a better fine-grained reasoning ability. **(2)** As expected, our proposed model ('Both') achieves significant improvements over our proposed model ('TE only') on both Fashion-IQ and CIRR, which indicates that the fine-tuning strategy has a significant impact on the IR-CQ task. However, when using the same fine-tuning strategy as Clip4Cir [5], our proposed model ('TE only') still leads to an improvement on every metric over Clip4Cir, which clearly demonstrates that the extra use of bilinear pooling fusion and multi-scale fusion indeed benefits the IR-CQ task.

3.3. Ablation Study

To investigate the contribution of each component of our proposed model, we conduct an ablation study on the Fashion-IQ dataset. Note that we only fine-tuned the text encoder of CLIP (used for initialization) during the ablation study.

Specifically, the ablation study is conducted over the following variants of our proposed model:

(1) MMF w/o bilinear: it is mainly composed of the original multi-modal fusion (MMF) block proposed by [5].

(2) MMF w/ bilinear: it is mainly composed of the MMF block which includes bilinear fusion strategies.

(3) MMF+MSF w/o bilinear: it introduces a multi-scale fusion (MSF) block to utilize the multi-scale information of the image. However, neither the MMF block nor the MSF block includes bilinear fusion.

(4) MSMMF w/ bilinear: It adopts a multi-scale multi-modal fusion (MSMMF) block (based on our general fusion block) to fuse the text feature and the image feature of each scale, and applies the hierarchical matching strategy [4].

(5) MMF+MSF w/ bilinear: it is our full model proposed in this work (see Fig. 1).

We report the ablation study results in Table 3. We have the following observations: **(1)** According to MMF w/o bilinear vs. MMF w/ bilinear, the extra use of bilinear fusion leads to an improvement of 1.11% on Rmean. This demonstrates the effectiveness of the bilinear fusion in the IR-CQ task. Note that we can make the same conclusion for MMF+MSF w/o bilinear (row 3) vs. MMF+MSF w/ bilinear (row 5). **(2)** Compared to MMF w/o bilinear (row 1), MMF+MSF w/o bilinear (row 3) achieves an improvement of 1.13% on Rmean, showing that the MSF block indeed helps to promote the retrieval performance. **(3)** Compared to MSMMF w/ bilinear (row 4), the improvements obtained by our full model (row 5) indicate that our combination of multi-scale fusion and multi-modal fusion is more effective (and also efficient) than the hierarchical matching strategy [4].

4. CONCLUSION

We concentrate on Image Retrieval with Composed Query (IR-CQ) in this study. We present a fusion block encompassing three strategies for enhanced hybrid-modality query representation. The block manifests in two forms: Multi-Scale Fusion (MSF) and Multi-Modal Fusion (MMF), tasked with fusing multi-scale image features and the hybrid-modality query respectively. We prefer pre-fusing multi-scale image features over hierarchical matching. Our model markedly surpasses current methods, as substantiated by extensive testing on three benchmark datasets.

5. ACKNOWLEDGEMENT

This work was supported by National Natural Science Foundation of China (62376274).

6. REFERENCES

[1] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1096–1104.

[2] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler, "Vse++: Improving visual-semantic embeddings with hard negatives," *arXiv preprint arXiv:1707.05612*, 2017.

[3] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays, "Composing text and image for image retrieval-an empirical odyssey," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6439–6448.

[4] Yanbei Chen, Shaogang Gong, and Loris Bazzani, "Image search with text feedback by visiolinguistic attention learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3001–3011.

[5] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo, "Conditioned and composed image retrieval combining and partially fine-tuning clip-based features," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4959–4968.

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, 2021, pp. 8748–8763.

[7] Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng, "Multimodal intelligence: Representation learning, information fusion, and applications," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 478–493, 2020.

[8] Matthew D Zeiler and Rob Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision*, 2014, pp. 818–833.

[9] Ginger Delmas, Rafael Sampaio de Rezende, Gabriela Csurka, and Diane Larlus, "Artemis: Attention-based retrieval with text-explicit matching and implicit similarity," *arXiv preprint arXiv:2203.08101*, 2022.

[10] Seungmin Lee, Dongwan Kim, and Bohyung Han, "Cosmo: Content-style modulation for image retrieval with text feedback," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 802–812.

[11] Jongseok Kim, Youngjae Yu, Hoeseong Kim, and Gunhee Kim, "Dual compositional learning in interactive image retrieval," in *AAAI Conference on Artificial Intelligence*, 2021, pp. 1771–1779.

[12] Haokun Wen, Xuemeng Song, Xin Yang, Yibing Zhan, and Liqiang Nie, "Comprehensive linguistic-visual composition network for image retrieval," in *ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 1369–1378.

[13] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo, "Conditioned image retrieval for fashion using contrastive learning and clip-based features," in *ACM Multimedia Asia*, 2021, pp. 1–5.

[14] Yida Zhao, Yuqing Song, and Qin Jin, "Progressive learning for image retrieval with hybrid-modality queries," *arXiv preprint arXiv:2204.11212*, 2022.

[15] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris, "Fashion iq: A new dataset towards retrieving images by natural language feedback," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11307–11317.

[16] Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Feris, "Dialog-based interactive image retrieval," *Advances in Neural Information Processing Systems*, vol. 31, pp. 676–686, 2018.

[17] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould, "Image retrieval on real-life images with pre-trained vision-and-language models," in *IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2125–2134.

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

[19] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[20] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[21] Eric Dodds, Jack Culpepper, Simao Herdade, Yang Zhang, and Kofi Boakye, "Modality-agnostic attention fusion for visual search with text feedback," *arXiv preprint arXiv:2007.00145*, 2020.