Technical Report of NICE Challenge at CVPR 2023: Retrieval-based Data Discovery and Fusion for Zero-shot Image Captioning

Youngtaek Oh[1] Jae Won Cho[1] Dong-Jin Kim[2] In So Kweon[1] Junmo Kim[1]
[1]KAIST [2]Hanyang University
[1]{youngtaek.oh, chojw, iskweon77, junmo.kim}@kaist.ac.kr [2]djdkim@hanyang.ac.kr

Abstract

This report introduces RETRIEVER*, our solution for the 2023 Challenge on zero-shot image captioning evaluation at the New Frontiers for Zero-Shot Image Captioning Evaluation (NICE) Workshop.* RETRIEVER *efficiently improves image captioners by retrieving from an external memory of image-text pairs in two steps. First, a set of image-text pairs for training is fetched by applying explicit retrieval module to the intended target dataset. In addition, we fuse the knowledge associated with the input sample queried from the retrieval module during training and inference. With this complete framework, specific knowledge in captions can be easily incorporated into the captioner even in the absence of ground-truth captions, and the model can generate better captions conditioned on relevant knowledge from an external data source. Experimentally,* RETRIEVER *improves the base image captioner by the CIDEr score by 229.4 in (held-out) validation data of NICE Challenge 2023 despite its simplicity. On the test data, notably, we ranked 2nd in CIDEr score, and 1st in all the other metrics. Our implementation including codes and checkpoints will be made public at https://github.com/ytaek-oh/retriever.*

1. Introduction

The NICE dataset [1] is a zero-shot evaluation dataset that assesses the robustness of image captioning models. It contains 26k images with captions sourced from Shutterstock [2], providing a far larger range of visual concepts from various domains and image types. Moreover, unlike other typical image captioning datasets such as COCO [4], captions in the NICE dataset often contain specific information like camera angle descriptions and proper nouns like place names. Accordingly, even foundational vision and language models in million or billion scale of parameters pretrained on large-scale image-text corpus [10, 12] struggle to predict these new concepts under zero-shot settings.

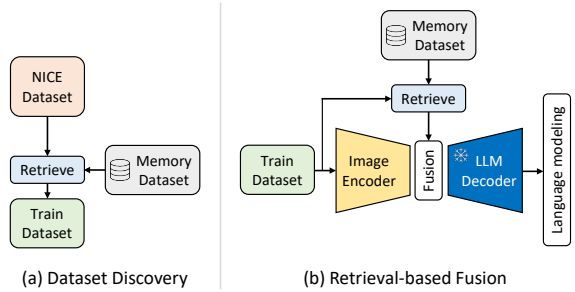


Figure 1. Overview of our RETRIEVER framework. RETRIEVER improves image captioning under zero-shot settings in two stages: retrieval-based dataset discovery for training and retrieval-based fusion conditioned on examples similar to inputs.

To address these challenges, we propose RETRIEVER, a generic framework that consists of an image captioning model and an explicit retrieval module for efficiently expanding the knowledge of the model by referring to an external memory dataset. Our approach draws inspiration from recent success of retrieval-augmented models in image recognition [8, 13] and language models [3, 7]. Extending this to image captioning, our RETRIEVER discovers and combines useful knowledge to the model, which is advantageous under zero-shot settings where direct fine-tuning on massive data is time and computationally prohibitive.

As shown in Fig. 1, the RETRIEVER framework enhances the captioning model in two stages. Firstly, to initiate training a captioning model in the absence of a training dataset, an explicit retrieval module is employed on the NICE dataset. This results in a set of image-text pairs that closely mimic the desired distribution of captions with images. Secondly, during training and inference, we explicitly combine the knowledge associated with the input sample into the model. This helps the model to generate improved captions by conditioning on useful external knowledge. As note, our work differs from previous retrieval-based image captioning approaches [16, 17] in that we focus on fine-tuning of captioning model via retrieval-based relevant data discovery as well as retrieval augmented knowledge fusion.

2. Our Approach

Our RETRIEVER includes a captioning model and a retrieval module, which will be detailed in Secs. 2.1 and 2.2. Then, Sec. 2.3 describes the retrieval process for selecting samples for training our model. Finally, Sec. 2.4 explains the how the retrieved knowledge based on the input query is fused with the model input during training and inference.

2.1. Image Captioning Model

Our RETRIEVER is a generic framework, meaning that our framework can be applied to any captioning model that is composed of an image encoder and a language decoder for generating captions. We choose BLIP-2 [12] as our base captioning model for its scalability and performance. For the image encoder and language decoder architecture, we opt to ViT-g/14 [6] and $\text{OPT}_{2.7\text{B}}$ [18], respectively. We initialize BLIP-2 with official weights after from the phase after the generative pretraining phase.

When we further fine-tune our BLIP-2 model, we keep the language decoder frozen and update the remaining parameters from image encoder and Q-Former, following the fine-tuning procedure for image captioning task [12]. Note that the initial BLIP-2 model here without any fine-tuning already has the captioning ability as a result of the generative pretraining objective.

2.2. Retrieval Module

Given an input query image, the retrieval module retrieves a set of related examples from an image-text pair dataset saved in the external memory. More specifically, it performs a k-nearest neighbor (kNN) search with cosine similarity as a metric, where it compares the query image to all the images in the memory dataset as keys in the embedding space. For efficient querying, we build the index using Hierarchical Navigable Small World (HNSW) approximate k-NN lookup [15] in FAISS library [9], with a hyperparameter M of 32. Note that such assignment is processed offline before training so computational overhead during training is negligible.

For building the memory dataset, we downloaded Shutterstock image-text pairs listed from the metadata[1]. Out of a collection of 15M image-text pairs, we curate a dataset of 1.1M pairs where the image category is specified as *photo* to fit our computational capacity. To obtain embeddings from images, we apply the same image encoder (ViT-g/14) in BLIP-2 initialized as mentioned in Sec. 2.1 to both query and memory dataset. We plan to share the list of images and the FAISS index file publicly for future research.

[1] https://github.com/mlfoundations/clip_quality_not_quantity

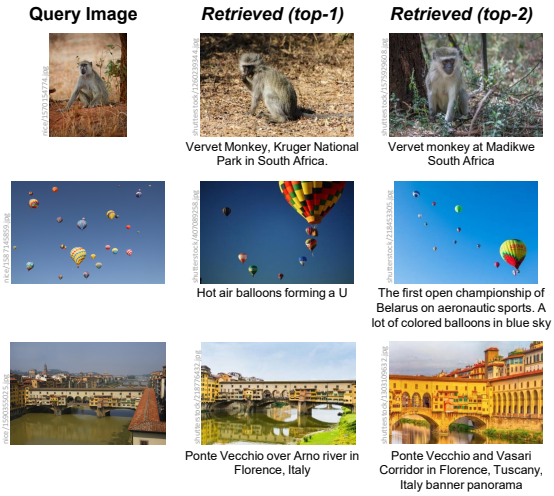


Figure 2. Examples of the discovered images with corresponding captions retrieved from the external memory dataset. These retrieved samples are used to further fine-tune our captioning model.

2.3. Dataset Discovery

As Tab. 1 points out, BLIP-2 model after pretraining on web-scale data and even fine-tuning on COCO exhibits poor performance on NICE dataset. This indicates that simply fine-tuning large models on a large amount of data may not be effective under zero-shot evaluation of NICE task. Therefore, we propose to discover data samples for training that are relevant to the target task using the retrieval module explained in Sec. 2.2. Specifically, we use the NICE dataset of size N as the query and retrieve k image-text pairs per query image from the memory dataset, resulting in kN examples. After a deduplication step, we obtain a training dataset of $\hat{N}$ unique image-text pairs $\mathcal{X} := \{(x_i, c_i)\}_{i=1}^{\hat{N}}$, where $\hat{N} < kN$. Fig. 2 showcases some examples of images with corresponding captions retrieved from an external dataset, using the test set of the NICE dataset as the query.

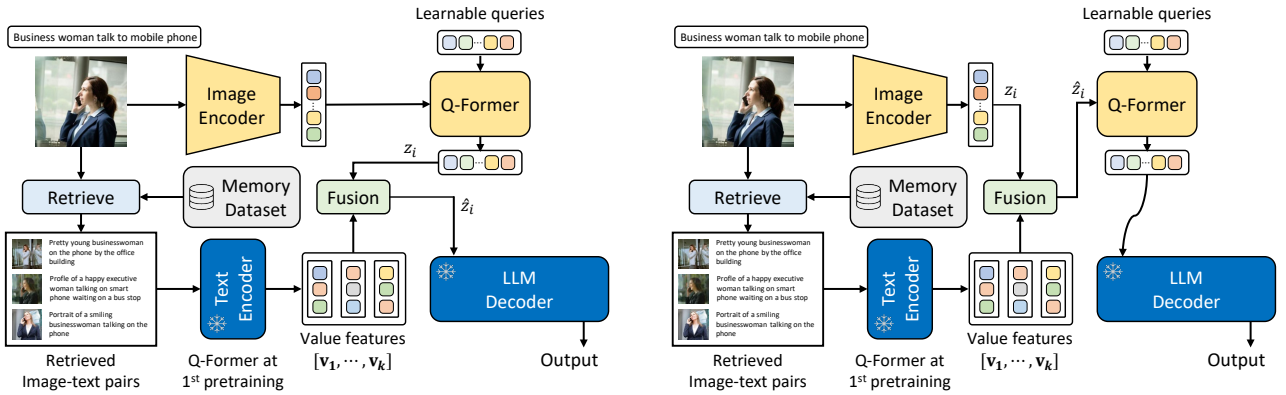
2.4. Retrieval-based Fusion

While fine-tuning BLIP-2 with the data in Sec. 2.3, inspired by retrieval-augmented models [8, 17], we further incorporate the knowledge associated with the input queries into the model. This approach can help address challenging scenarios such as long-tailed or zero-shot examples, where learning becomes difficult for typical parametric models.

In detail, the retrieval module in Sec. 2.2 performs kNN search for an input x_i, and then produces a set of value embeddings $\mathbf{V}_k(x_i)$ using the indices of retrieved images:

$$\mathbf{V}_k(x_i) = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_k], \text{ where } \mathbf{v}_j = \psi(c_j), \quad (1)$$

Here, a separate language encoder $\psi : c \rightarrow \mathbb{R}^{N \times d_v}$ maps the caption c_j corresponding to each retrieved image x_j to the value embeddings $\mathbf{v}_j$ with length of N. These value embeddings provide rich contextual information complemen-



(a) Fusion in output of query representations.

(b) Fusion in visual patch token embeddings.

Figure 3. Illustrations of two potential locations for integrating the retrieved knowledge from the query into the BLIP-2 architecture. **(a)** The fused feature $\hat{z}_i$ is obtained by combining the output query representation z_i with the aggregated knowledge, at the output level of Q-Former. **(b)** $\hat{z}_i$ is produced by fusing the visual patch token embeddings z_i with the aggregated knowledge.

tary to the knowledge that the original model has. We then combine such knowledge with the input query $z_i \in \mathbb{R}^{M \times d_z}$. Note that z_i is determined by the location of the fusion process in the model as described in Fig. 3. Given the token embeddings z_i as the input query, we aggregate the value features by averaging [8] and concatenate it with z_i:

$$\hat{z}_i = \text{Concat}\left(z_i,\ \phi\left(\frac{1}{k}\sum_{j=1}^{k}\mathbf{v}_j\right)\right), \tag{2}$$

where $\phi : \mathbb{R}^{d_v} \rightarrow \mathbb{R}^{d_z}$ is a fully connected layer to match the channel length of $\mathbf{v}_j$ to that of the token embeddings z_i.

For $\psi(\cdot)$, we use the Q-Former to produce the value embeddings, which is initialized from the BLIP-2 after pretraining on the combination of image-text matching (ITM), image-text contrastive learning (ITC), and image-grounded text generation (ITG) objectives. Its architecture is identical to that of $\text{BERT}_{\text{base}}$ [5]. Here, the dimension of the value embeddings d_v is 768, which is then projected to a vision dimension d_z by $\phi(\cdot)$.

The fusion process in Eq. (2) varies based on the choice of z_i. We categorize these cases into two based on the location of fusion within the architecture of BLIP-2 as shown in Fig. 3 and verify the effectiveness of each case in Tab. 2.

Output of query token embedding. During fine-tuning BLIP-2, the query tokens interact with visual embeddings from the image encoder via the cross-attention layers in the Q-Former. This generates the output query representation Z_i which has the same size as a value feature $\mathbf{v} \in \mathbb{R}^{N \times d_v}$, where N is 32 and d_v is 768 as illustrated in [12].

Here, we can augment the LLM decoder input, which is originally Z_i, with the relevant knowledge. As depicted in Fig. 3a, the fused representation $\hat{z}_i \in \mathbb{R}^{2N \times d_v}$ can be obtained by concatenating z_i with the aggregated value features and is further fed into the LLM decoder.

Method	Num. data	BLEU@1	SPICE	CIDEr
pretrained BLIP-2	0	22.6	21.2	79.6
COCO fine-tuned BLIP-2	0	28.9	17.8	68.0
Randomly sampled data	10901	27.8	22.5	83.4
Data discovery (k=1)	10901	**45.8**	**33.0**	**196.4**
Randomly sampled data	19583	27.5	21.7	79.8
Data discovery (k=2)	19583	44.4	32.3	189.1

Table 1. Comparisons of different approaches for utilizing retrieved data relevant to the target task, as explained in Sec. 2.3. Training on the retrieved dataset leads to significantly better performance than simple pretraining or fine-tuning on COCO. 'Num. data' refers to the number of datapoints used for training.

Visual patch token embedding. Alternatively, we can directly combine the visual information with the context from the retrieved captions. As shown in Fig. 3b, we consider the output of the image transformer as z_i and concatenate these visual patch tokens with the aggregated value features, producing a fused representation $\hat{z}_i \in \mathbb{R}^{(M+N) \times d_z}$. This enables better interactions with the query token embeddings in the Q-Former through the cross-attention layers.

3. Experiments

3.1. Implementation details

Fine-tuning settings. We train BLIP-2 on our training dataset specified in Sec. 2.3. The image encoder is ViT-g/14, and the language decoder is $\text{OPT}_{2.7\text{B}}$. We resize images to 364×364 resolution, resulting in the length of patch tokens M 677. During training, we freeze the language decoder. We use an AdamW optimizer [14] with a peak learning rate of 4e-5 and weight decay of 0.05. We decay the learning rate with a cosine schedule and also apply 500 steps of linear warm-up strategy. Our batch size is 16 per GPU, and we train the model using four Quadro RTX 8000

Method	Num.	BLEU@1	SPICE	CIDEr
pretrained BLIP-2	0	22.6	21.2	79.6
w/ data discovery (k=1)	10901	45.8	33.0	196.4
+ query output fusion (k=1)	10901	44.8	31.6	167.6
+ query output fusion (k=2)	10901	36.0	26.4	117.0
+ query output fusion (k=3)	10901	45.8	**34.1**	199.9
+ patch token fusion (k=1)	10901	45.5	32.5	190.1
+ patch token fusion (k=2)	10901	46.0	32.7	192.2
+ patch token fusion (k=3)	10901	**46.1**	33.8	**201.1**

(a) Fine-tuning on a dataset from data discovery (k=1) with $10k$ samples.

Method	Num.	BLEU@1	SPICE	CIDEr
pretrained BLIP-2	0	22.6	21.2	79.6
w/ fine-tuning on NICE val	4000	50.1	36.8	238.3
+ query output fusion (k=1)	4000	52.7	**38.9**	260.6
+ query output fusion (k=2)	4000	51.8	37.4	249.6
+ query output fusion (k=3)	4000	43.4	29.4	153.4
+ patch token fusion (k=1)	4000	52.5	38.3	261.4
+ patch token fusion (k=2)	4000	52.4	38.3	255.8
+ patch token fusion (k=3)	4000	**52.8**	38.6	**263.3**

(b) Fine-tuning on a subset of NICE validation data with $4k$ samples.

Table 2. Comparisons of different locations for retrieval-based fusion, across different types of query dataset for training and varying amount of caption features to be averaged. **(a)** Fine-tuning on a data constructed from the data discovery (k=1). **(b)** Fine-tuning on a subset of NICE validation data. We find that fusing on visual patch tokens produces consistent improvements than fusing on query output level.

Method	Num. data	BLEU@1	SPICE	CIDEr
pretrained BLIP-2	0	22.6	21.2	79.6
fine-tuning on NICE val	4000	50.1	36.8	238.3
+ Data discovery (k=1)	14901	55.8	42.6	305.1
+ patch token fusion (k=2)	14901	56.3	42.7	309.0
Results on NICE Test set	15901	**58.0**	**45.5**	**324.9**

Table 3. Results achieving the top-scoring entry. Our final model produced a CIDEr score of 324.9 on the NICE test set.

GPUs, each with 48GB memory. Our implementation is built on top of the LAVIS library [11], and we generally adopt its training protocols unless specified otherwise.

Evaluation setup. During the decoding step for producing captions, we set the minimum and maximum sequence lengths to 8 and 30, respectively, and use a beam size of 5. To measure the captioning performance, we primarily utilize a 5k validation split of the NICE dataset. We use three commonly used metrics: BLEU, SPICE, and CIDEr scores.

3.2. Main Results

Our experimental design aimed to evaluate the efficacy of dataset discovery, as well as the retrieval-based fusion strategy. Tabs. 1 and 2 present the results of these experiments, respectively. Finally, as shown in Tab. 3, we jointly apply two approaches to create our final model for submitting to the evaluation server.

Effectiveness of data discovery. Tab. 1 compares training datasets for fine-tuning BLIP-2 by evaluating on the NICE evaluation dataset. For the dataset discovery process, we set the query dataset as the test split of NICE dataset, and vary the number of image-text pairs per query sample k to one and two. This resulted in 10,901 and 19,583 distinct image-caption pairs for training, respectively.

Our first finding is that zero-shot approaches that do not explicitly use either Shutterstock or NICE data for training result in unsatisfactory captioning performance. Specifically, both of pretrained BLIP-2 model and COCO fine-tuned one show CIDEr scores of 79.6 and 68.0 on validation split of NICE dataset, respectively.

Furthermore, we observe that when fine-tuning on randomly sampled Shutterstock pairs of the same size from the dataset discovery process, the resulting performance is significantly inferior compared to that achieved through the dataset discovery process itself. This finding highlights the importance of retrieval process that discovers image-text pairs beneficial to the NICE captioning task.

Finally, fine-tuning on the discovered data from retrieval process improved the CIDEr scores to 196.4 and 189.1 when $k = 1$ and $k = 2$ respectively. As note, we found that using lower samples was more effective due to the presence of untrimmed captions crawled from the web, which could introduce noise during training.

Effectiveness of retrieval-based fusion. We investigate the effectiveness of the retrieval-based fusion mechanism with respect to different fusion locations outlined in Sec. 2.4. As shown in Tab. 2, we evaluate on two different types of training dataset, where Tabs. 2a and 2b correspond to the data from the dataset discovery with k of 1 and a subset of NICE validation split in 4k samples, respectively. We vary the number of value features to be averaged in Eq. (2) from one to three. As note, we use the remaining 1k samples of NICE validation data for evaluation.

Fusing the retrieved knowledge on both query output representations and visual patch token embeddings was effective compared to the pretrained BLIP-2 without any training. However, fusing on query output representation shows unstable and poor results at lower k value. When fusing on patch token embeddings, increasing the number of value features (*e.g.*, higher k) led to better performance.

Top-scoring Entry. As presented in Tab. 3, to achieve the top-performing entry, we first fine-tuned the baseline BLIP-2 on 4k samples of the NICE validation data. This resulted in a CIDEr score of 238.3 on the remaining 1k held-out validation data. We then applied the data discovery with k of 1, followed by retrieval-based fusion on visual patch token embeddings with k of 2. This led to CIDEr scores of 305.1 and 309.0, respectively, as in the third and fourth rows of Tab. 3. For our final test model, we included the entire

NICE validation data in training along with the data from dataset discovery, and retrained the model. The final model was then evaluated on the test data and the prediction results are submitted to the evaluation server. As shown in the last row of Tab. 3, our top-scoring model achieved a CIDEr score of 324.9 on the test split of NICE dataset, scoring as the second place on CIDEr while taking the first place in *all the other captioning metrics* on the official leaderboard.

3.3. Captioning Results

We present captioning results on some samples of the validation set of NICE data in Fig. 4. We compare the caption predictions of the pretrained BLIP-2 and our model in CIDEr scores of 79.6 and 309.0 respectively in Tab. 3 along with the ground-truth captions. In addition, we show the top-2 retrieved images with corresponding captions from the query images. Through the retrieval-based fusion mechanism, the captions generated by our model are able to capture specific knowledge and concepts present in the ground-truth captions. For example, our captions include details like camera angle descriptions and proper nouns such as city or country names, as highlighted in green and blue color.

References

[1] NICE: New frontiers for zero-shot image captioning evaluation. https://nice.lgresearch.ai/. Accessed: 2023-05-03. 1

[2] Shutterstock. https://www.shutterstock.com/. Accessed: 2023-05-03. 1

[3] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR, 2022. 1

[4] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 1

[5] 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. 3

[6] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. *arXiv preprint arXiv:2211.07636*, 2022. 2

[7] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR, 2020. 1

[8] Ahmet Iscen, Alireza Fathi, and Cordelia Schmid. Improving image recognition by retrieving from web-scale image-text data. *arXiv preprint arXiv:2304.05173*, 2023. 1, 2, 3

[9] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019. 2

[10] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, et al. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. *arXiv preprint arXiv:2205.12005*, 2022. 1

[11] Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven C. H. Hoi. Lavis: A library for language-vision intelligence, 2022. 4

[12] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 1, 2, 3

[13] Alexander Long, Wei Yin, Thalaiyasingam Ajanthan, Vu Nguyen, Pulak Purkait, Ravi Garg, Alan Blair, Chunhua Shen, and Anton van den Hengel. Retrieval augmented classification for long-tail visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6959–6969, 2022. 1

[14] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 3

[15] Yu A Malkov and Dmitry A Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):824–836, 2018. 2

[16] Rita Ramos, Desmond Elliott, and Bruno Martins. Retrieval-augmented image captioning. *arXiv preprint arXiv:2302.08268*, 2023. 1

[17] Zhuolin Yang, Wei Ping, Zihan Liu, Vijay Korthikanti, Weili Nie, De-An Huang, Linxi Fan, Zhiding Yu, Shiyi Lan, Bo Li, et al. Re-vilm: Retrieval-augmented visual language model for zero and few-shot image captioning. *arXiv preprint arXiv:2302.04858*, 2023. 1, 2

[18] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 2

Figure 4. Comparisons of the caption predictions on the query images taken from the NICE validation set. On the right side, we present the top-2 retrieved images along with their corresponding captions derived from the query. This retrieved knowledge improves the caption predictions of our model by incorporating specific knowledge and concepts found in the ground-truth captions of NICE data.