

# Adapting Pre-Trained Visual and Language Models for Medical Image Question Answering

Notebook for the Baidu Intelligent Health Unit and Peng Cheng Laboratory Joint Team at CLEF 2023

Siqi Wang<sup>1</sup>, Wenshuo Zhou<sup>1</sup>, Yehui Yang<sup>1</sup>, Haifeng Huang<sup>1</sup>, Zhiyu Ye<sup>2</sup>,  
Tong Zhang<sup>2,\*</sup> and Dalu Yang<sup>1,\*</sup>

<sup>1</sup>Baidu Intelligent Health Unit, Beijing 100085, China

<sup>2</sup>Peng Cheng Laboratory, Shenzhen 518055, China

## Abstract

This paper presents the work carried out by the "wsq4747" team in the ImageCLEFmedical2023 title for the visual Question Answering subtask. Medical image question answering presents unique challenges due to the specialized nature of the medical field. Not only does it require the model to generate accurate and coherent answers through the image and the question, but it also needs to capture the basic medical information conveyed by the image. In order to leverage the capabilities of pre-trained large image models, we utilized the state-of-the-art BLIP-2 combined with a giant visual transformer (vit-g) and an open pre-training transformer language model (GLM-6B) as the foundation for our title prediction subtask. To adapt this model to the medical field, we employed a two-stage fine-tuning process. During the entire training process, the pre-trained GLM-6B was kept fixed, and step-by-step fine-tuning was applied to the vit-g and Q-Former modules to better align with the features of medical data. Our team's approach produced promising results with an accuracy(ACC) of 0.7396, as our method achieved an ACC of over 0.8 on 6 questions, an ACC of over 0.7 on 10 questions, and an ACC of around 0.1 on two questions (due to our oversight)

## Keywords

ImageCLEF, Visual Question Answering, Blip-2

## 1. Introduction

ImageCLEF[1], short for Image Retrieval Evaluation Campaign, is a component of CLEF (Cross-Language Evaluation Forum), a European project in the fields of computer science and information retrieval that organizes various research challenges annually to evaluate the performance of multilingual information retrieval. The objective of ImageCLEF is to propel the advancement of computer vision and multimedia information retrieval technologies. The task encompasses

---

*CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece*

\*Corresponding author.

†These authors contributed equally.

✉ wsq\_47@126.com (S. Wang); ws.zhou@foxmail.com (W. Zhou); yangyehuisw@126.com (Y. Yang);

huanghaifeng@baidu.com (H. Huang); yezhy@pcl.ac.cn (Z. Ye); zhangt02@pcl.ac.cn (T. Zhang);

albertyoung@live.cn (D. Yang)

ORCID 0000-0002-8838-4963 (T. Zhang)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

both natural language processing and image recognition. It provides a query, which could be a question, a statement, or a description in another form, and then requires the system to find images relevant to the query from a vast image library.

The challenges presented in ImageCLEF comprise of multiple subtasks, including Visual Question Answering (VQA)[2] and medical image description. VQA is a multimodal task in the field of artificial intelligence, with the objective of developing models that can interpret visual content such as images or videos, and provide responses to corresponding natural language queries. This task becomes significantly more complex when applied to intricate scenarios in medical imaging. In the medical image question-answering task, we focused on colonoscopy images. Colonoscopy images are inherently complex and high-dimensional, with intricate relationships between visual features and medical semantics. Effectively modeling these relationships presents a significant challenge. Additionally, the language used in medical queries is often laden with complex medical jargon, necessitating a deep understanding of medical concepts that may not be encapsulated in general language models. Furthermore, procuring large-scale annotated data for training such models poses a difficulty due to privacy concerns and the requirement for expert annotations. Nevertheless, despite these challenges, the potential of medical VQA is considerable. Ongoing research continues to push the boundaries of our capabilities in this crucial area. In this work, our team primarily focuses on the task of medical image question-answering.

Some methods freeze the image encoder, including the early work which adopts a frozen object detector to extract visual features [3, 4, 5], and the recent LiT [6] which uses a frozen pre-trained image encoder for CLIP[7] pre-training. Some methods freeze the language model to use the knowledge from LLMs for vision-to-language generation tasks [8, 9]. The key challenge in using a frozen LLM is to align visual features to the text space. To achieve this, Frozen[8] finetunes an image encoder whose outputs are directly used as soft prompts for the LLM

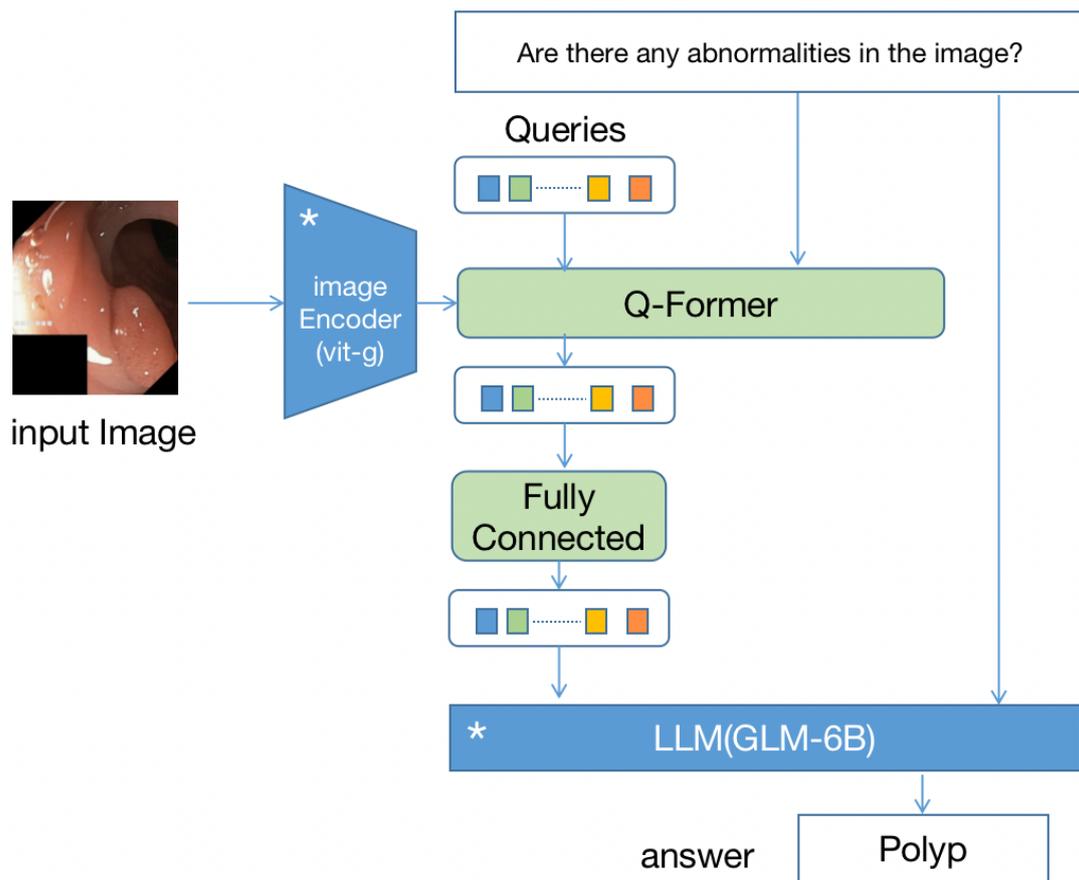
In this task, we employed BLIP-2[10]. BLIP-2 is a recently proposed vision-language pre-training method by Li et al[10]. Blip-2 is a generic and efficient pretraining strategy that bootstraps vision-language pre-training from off-the-shelf frozen pre-trained image encoders and frozen large language models(LLM). BLIP-2 bridges the modality gap with a lightweight Querying Transformer, which is pretrained in two stages.building upon their previous work of BLIP[10], and it has demonstrated superior performance compared to various other vision-language

pre-training methods, including Flamingo[11], across a range of vision-language tasks such as visual question answering, image captioning, and image-text retrieval. In this paper, our method is specifically introduced in Section 2, the experiments, data, and results are demonstrated in Section 3 and a brief summary is given in Section 4.

## 2. Method

### 2.1. Architecture

BLIP-2[10] is a sophisticated framework designed for vision-to-language tasks, comprised of three main components: an image encoder, a Query Transformer (Q-Former), and a LLM.As



**Figure 1:** In our method, the first stage of BLIP-2 is a bootstrapping process that involves freezing the image encoder and LLM and BLIP-2’s second-stage vision-to-language generative pre-training, which bootstraps from frozen LLM.

shown in Figure 1. The Q-Former as the trainable module to bridge the gap between a frozen image encoder and a frozen LLM.

During the pre-training generation phase, we connected the Q-Former, equipped with a frozen image encoder, to the frozen Language Model (LLM) to leverage its language generation capacity. As depicted in Figure 1, we employed a fully connected (FC) layer to linearly project the output query embedding into the same dimensionality as the LLM’s text embeddings. This projected query embedding was then added prior to the input text embeddings. Acting as soft visual prompts, they placed the LLM upon the visual representation extracted by the Q-Former. Given that the Q-Former had been pre-trained to extract visual representations carrying language information, it effectively played the role of an information bottleneck, offering the most useful details to the LLM while discarding irrelevant visual data. This mitigated the burden of learning visual-language alignment on the LLM, thereby alleviating the issue of catastrophic forgetting.

In response to our task, we employed the BLIP-2 model for visual question answering and selected vite-g/14 from EVA-CLIP[12] as our image encoder. For the LLM, we selected GLM-

**Table 1**

Distribution of image sizes in our dataset. The table shows the number of images for three different size categories: less than 500, between 500 and 1000, and greater than 1000.

image size	amount
<500	21
500<image<1000	1441
>1000	538

6b[13], a prefix-based, decoder-only model, to serve as our language language model (LLM).

## 2.2. Dataset

The dataset encompasses images spanning the entirety of the gastrointestinal tract, from the mouth to the anus. It encapsulates various instances, including abnormalities, surgical instruments, and normal findings. The images are procured from different procedures such as gastroscopy, colonoscopy, and capsule endoscopy. The distribution of image sizes within our dataset is illustrated in Table 1, exhibiting a broad array of dimensions. A minimal fraction of the images, exactly 21, have dimensions less than 500 pixels. The bulk of our images, amounting to 1441, belong to the medium size category with pixel dimensions ranging from 500 to 1000. Lastly, a significant subset of our data, constituting 538 images, features dimensions exceeding 1000 pixels. This heterogeneity in image size amplifies the diversity and intricacy of our dataset, thereby increasing the challenge and comprehensiveness of the Visual Question Answering task at hand. For both Task 1 (VQA) and Task 2 (Visual Question Generation, VQG), a minimum of 2000 image samples have been provided, each accompanied by eighteen question-and-answer pairs. It should be noted that not all questions are pertinent to the corresponding image.

In Task1, since the data did not divide the training set and the validation set, we randomly selected 10% of the image-text question-answer pairs as the validation set.

## 2.3. Training Strategy

The training protocol for BLIP-2 is carried out in two distinct stages. In the first stage, a process known as vision-language representation learning, the image encoder and language models are frozen, allowing the model to tap into its inherent image understanding capabilities. The second stage involves vision-to-language generative learning, where the LLM is frozen, maintaining its existing text generation capabilities. When applying the BLIP-2 model to downstream tasks, such as visual question answering, the LLM is kept frozen during the fine-tuning phase. Meanwhile, the parameters of the image encoder and Q-Former are updated.

Throughout these two stages, the language models are kept frozen to preserve their initial functionalities. In contrast, the Q-Former is exclusively trained during this pre-training phase. The role of the Q-Former is to effectively extract visual representations that align with the corresponding textual information and to relay this information to the LLM. This focused training approach allows BLIP-2 to achieve a higher level of correspondence between visual

**Table 2**

In our model ablation study on the validation set, due to time constraints, we only employed the version with an accuracy of 0.9105 for comparison. \* denote instances where model parameters were frozen during training.

Models	model parameters	accuracy
vit-l*+Q-former+t5-base*	715M	0.5227
vit-l+Q-former+t5-base*	715M	0.6048
vit-g*+Q-former+t5-base*	1.4B	0.6017
vit-g+Q-former+t5-base*	1.4B	0.7577
vit-g*+Q-former+t5-3b*	4.2B	0.7427
vit-g+Q-former+t5-3b*	4.2B	0.8324
vit-g*+Q-former+GLM-6b*	7.2B	0.9105(submitted)
vit-g+Q-former+GLM-6b*	7.2B	0.9317(uns submitted)

and textual data.

In our pre-training phase, we initialized our large-scale visual transformer (vit-g) and Query Transformer (Q-Former) with weights from BLIP-2, which had been previously pre-trained on the ImageNet[14] and COCO[15] datasets. However, our specific task focused on medical imaging (endoscopic images) for the visual question answering task. It’s worth noting that there is a significant domain shift between natural images and medical imaging data.

In order to address this issue and to allow the visual encoder to extract image features more delicately, we adopted a fine-tuning strategy. During this fine-tuning process, the parameters of the LLM(GLM-6B) were kept frozen, while the vit-g and Q-Former were trained concurrently. This strategy was designed to leverage the powerful visual representation capabilities of vit-g and Q-Former, while also accommodating the specific characteristics and challenges of the medical imaging domain.

### 3. Experiments

#### 3.1. Implementation Details

Our framework was developed using PaddlePaddle1 version 2.4.2 and trained on 8 Ascend 910 NPUs. The adapter plug-in PaddleCustomDevice2 was utilized in order to be compatible with the Ascend NPU. The entire process, encompassing two training stages, spanned a total duration of four days. The input image size was set to  $224 \times 224$ , and the batch size was fixed at 16 for both fine-tuning stages. The model underwent fine-tuning for 100 epochs in the first stage, and 50 epochs in the second stage. Optimization was performed using an AdamW optimizer with a weight decay of  $10^{-4}$ . The initial learning rate was set to  $10^{-4}$  and was incrementally adjusted through a 1000-step warm-up phase. Additionally, we set the maximum output length of our model to 10, as the majority of answers from the data clearly fell below this threshold.

**Table 3**

Performance of our final model on the validation set, showing the accuracy for each question. Accuracy improvement due to use of separate classification models are indicated for 'What is the size of the polyp?' and 'What type of polyp is present?' questions.

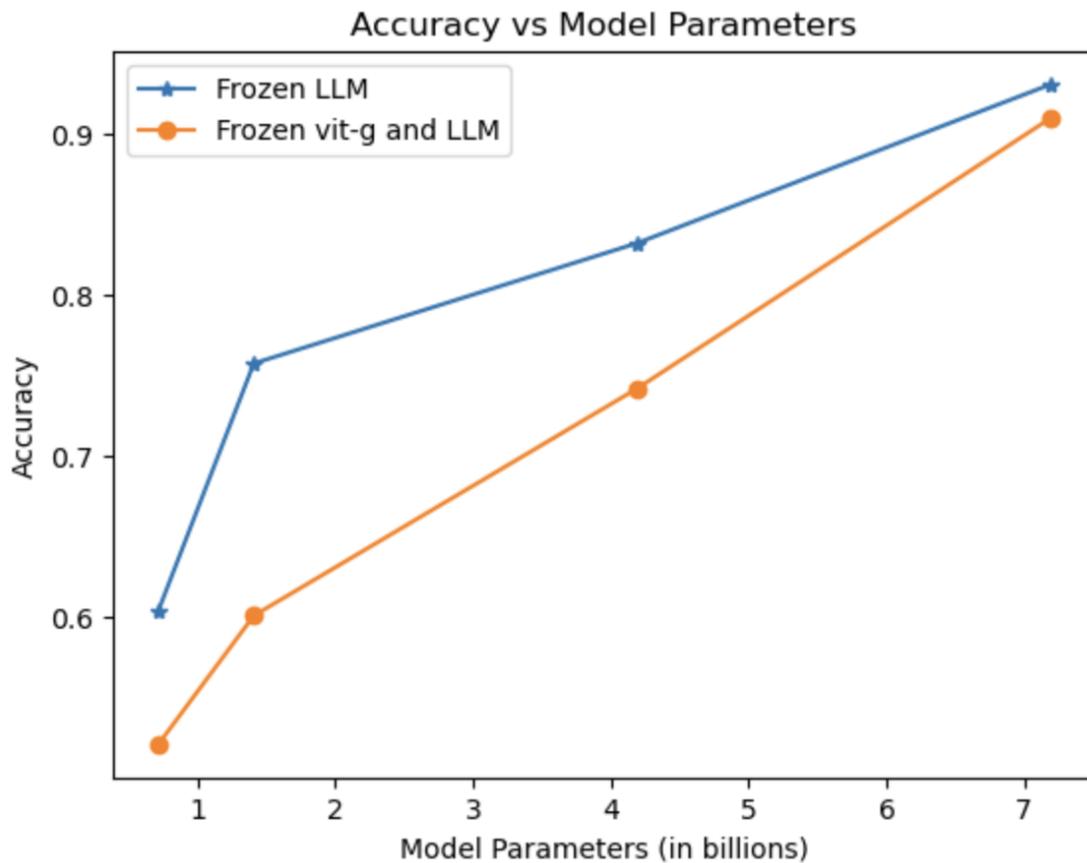
question	accuracy
Are there any abnormalities in the image?	0.989
Are there any anatomical landmarks in the image?	0.940
Are there any instruments in the image?	0.921
Have all polyps been removed?	0.989
How many findings are present?	0.862
How many instruments are in the image?	0.965
How many polyps are in the image?	0.970
Is there a green/black box artefact?	0.868
Is there text?	0.842
Is this finding easy to detect?	0.862
What color is the abnormality?	0.869
What color is the anatomical landmark?	0.956
What is the size of the polyp?	0.784 -> 0.99
What type of polyp is present?	0.852 -> 0.99
What type of procedure is the image taken from?	0.995
Where in the image is the abnormality?	0.939
Where in the image is the anatomical landmark?	0.915
Where in the image is the instrument?	0.873

### 3.2. Experimental Settings

In the endoscopic dataset, since there is no predefined division between training and validation sets, we conducted a manual split to validate the performance of our model. Specifically, 10% of the data, equating to 200 images with corresponding 3200 questions, was earmarked as the validation set. Beyond this, we employed accuracy as our primary evaluation metric to gauge the model's effectiveness in predicting the correct responses. As demonstrated in Table 3, we have evaluated our model performance on the validation set that was manually partitioned by us. For the final submission, however, we leveraged the entire dataset for fine-tuning the model.

### 3.3. Results on Validation Data

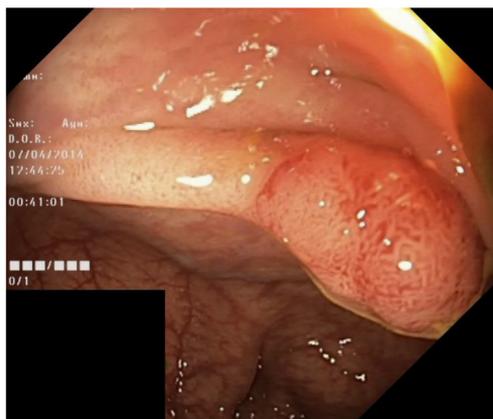
Table 2 presents a comparative overview of the results from the ablation study conducted on BLIP-2. It is evident that as the model parameters progressively increase, the performance also improves. Moreover, fine-tuning only the LLM (GLM-6b) yields better results compared to freezing both the Image Encoder (vit-g) and the LLM simultaneously. For the LLM, we compared FlanT5-3B [16], based on an encoder-decoder structure, and GLM-6b [13], based on a prefix decoder-only structure. The results showed that GLM-6b demonstrated superior performance in zero-shot images, thus we adopted GLM-6B as our LLM. However, due to time constraints, we ultimately opted for the model with an accuracy of 0.9105 as the final choice for our backbone model.



**Figure 2:** Accuracy vs Model Parameters for both frozen and unfrozen models. This figure demonstrates the impact of the number of model parameters on the accuracy of the VQA task. The plot compares the accuracy of models with different parameter sizes, under conditions where the models were kept frozen (blue line) and where the models were allowed to unfreeze (orange line). The trend illustrates that larger models tend to perform better, and unfreezing the models generally leads to higher accuracy.

In Figure 2, we provide an illustrative comparison between the performances of models under different parameter sizes, and the conditions whether the models were kept frozen or allowed to unfreeze. As shown in figure 2, there is an observable trend that larger models generally outperform the smaller ones, suggesting that the number of parameters plays a significant role in the accuracy of the VQA task. Furthermore, it is also evident that when the models are unfrozen, allowing the parameters to adjust during training, the accuracy increases across all model sizes. This underlines the importance of parameter fine-tuning in optimizing model performance in the context of medical VQA tasks.

Additionally, we analyzed the accuracy rates for individual questions, as shown in Table 3. From the validation set, it was apparent that the model had poor performance in predicting the size and type of polyps. We summarized the answers to these two questions in Table 4. To address these shortcomings, for the question *What is the size of the polyp?*, we trained a



Q: Are there any abnormalities in the image?  
A: Polyp  
Q: Are there any anatomical landmarks in the image?  
A: No  
Q: How many findings are present?  
A: 1  
Q: How many polyps are in the image?  
A: 1  
Q: How many instruments are in the image?  
A: 0

**Figure 3:** One example of predicted results in the validation set of answer prediction task.

**Table 4**

Distribution of answers for questions related to polyp size and type

question	answer statistic
What is the size of the polyp?	Not relevant, 11-20mm, 5-10mm, >20mm, < 5mm
What type of polyp is present?	Not relevant, Paris ip, Paris iia, Paris is

five-class classification model (ResNet34), and for the question *What type of polyp is present?*, we trained a four-class classification model (ResNet34). Both models achieved an accuracy rate of over 0.99 in the validation set, thereby outperforming the BLIP-2 model's output. Consequently, the overall accuracy on the validation set rose from 0.9105 to 0.9305. This was the result we submitted in the end.

### 3.4. Results on Test Data

Our submission results, as depicted in Table 5, were not as impressive as anticipated on the test set, yielding an accuracy of 0.7396. On six questions, the accuracy exceeded 0.8, while on ten questions, it surpassed 0.7. However, the accuracy was around 0.1 for the questions *What color is the abnormality?* and *Where in the image is the abnormality?*, significantly dragging down the overall average. We speculate that this may be due to the visual encoder's inability to extract detailed features related to color and position. As illustrated in Figure 3, we present one example of the predicted results obtained from the validation set for the answer prediction task. This figure visually demonstrates how our model performs in terms of predicting answers, offering an insight into the capabilities of our approach.

## 4. Summary

This paper has presented the work of the "wsq4747" team in the Visual Question Answering task for imageCLEFmedical VQA 2023. The model we utilized, which is a variant of BLIP-2

**Table 5**

Per-question accuracy of our model on the test set. The last row presents the global accuracy across all questions.

question	accuracy
Are there any abnormalities in the image?	0.7962
Are there any anatomical landmarks in the image?	0.7255
Are there any instruments in the image?	0.7270
Have all polyps been removed?	0.9324
How many findings are present?	0.7544
How many instruments are in the image?	0.8803
How many polyps are in the image?	0.8922
Is there a green/black box artefact?	0.7833
Is there text?	0.7312
Is this finding easy to detect?	0.8055
What color is the abnormality?	0.1393
What color is the anatomical landmark?	1.0
What is the size of the polyp?	0.7828
What type of polyp is present?	0.7993
What type of procedure is the image taken from?	0.9912
Where in the image is the abnormality?	0.1146
Where in the image is the anatomical landmark?	0.7219
Where in the image is the instrument?	0.7363
<b>global metrics</b>	<b>0.7396</b>

vit-g GLM-6b, underwent two stages of fine-tuning. Our team’s final accuracy was 0.7396, demonstrating the effectiveness of our approach in generating high-quality question-answering for medical images.

## Acknowledgments

The computing resources of Pengcheng Laboratory Cloudbrain II are used in this research. We acknowledge the support provided by OpenI Community (<https://git.openi.org.cn>).

## References

- [1] B. Ionescu, H. Müller, A. Drăgulinescu, W. Yim, A. Ben Abacha, N. Snider, G. Adams, M. Yetisgen, J. Rückert, A. Garcia Seco de Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, S. A. Hicks, M. A. Riegler, V. Thambawita, A. Storås, P. Halvorsen, D. J. A. A. R. I. C. V. K. A. S. G. I. Nikolaos Papachrysos, Johanna Schöler, H. Manguinhas, L. Ştefan, M. G. Constantin, M. Dogariu, J. Deshayes, A. Popescu, Overview of ImageCLEF 2023: Multimedia retrieval in medical, socialmedia and recommender systems applications (2023).
- [2] P. H. T. d. L. M. A. R. V. T. Steven A. Hicks, Andrea Storås, Overview of imageclefmedical 2023 – medical visual question answering for gastrointestinal tract (2023).

- [3] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, J. Liu, Uniter: Universal image-text representation learning (2020). URL: [http://dx.doi.org/10.1007/978-3-030-58577-8\\_7](http://dx.doi.org/10.1007/978-3-030-58577-8_7). doi:10.1007/978-3-030-58577-8\_7.
- [4] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, X. Li, Y. Choi, J. Gao, M. Corporation, W. Embeddings, Oscar: Object-semantics aligned pre-training for vision-language tasks (2020).
- [5] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, J. Gao, Vinvl: Making visual representations matter in vision-language models (2021).
- [6] X. Zhai, X. Wang, B. Mustafa, A. Steiner, D. Keysers, A. Kolesnikov, L. Beyer, Lit: Zero-shot transfer with locked-image text tuning (2022). URL: <http://dx.doi.org/10.1109/cvpr52688.2022.01759>. doi:10.1109/cvpr52688.2022.01759.
- [7] A. Radford, J. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, Cornell University - arXiv (2021).
- [8] M. Tsimpoukelli, J. Menick, S. Cabi, S. Eslami, O. Vinyals, F. Hill, Multimodal few-shot learning with frozen language models, Neural Information Processing Systems (2021).
- [9] A. Tiong, J. Li, B. Li, S. Savarese, S. Hoi, Plug-and-play vqa: Zero-shot vqa by conjoining large pretrained models with zero training (2022).
- [10] J. Li, D. Li, S. Savarese, S. Hoi, Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models (2023).
- [11] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, K. Simonyan, Flamingo: a visual language model for few-shot learning (2022).
- [12] Y. Fang, W. Wang, B. Xie, Q. Sun, L. Wu, X. Wang, T. Huang, X. Wang, Y. Cao, Eva: Exploring the limits of masked visual representation learning at scale (2022).
- [13] A. Zeng, X. Liu, Z. Du, Z. Wang, H. Lai, M. Ding, Z. Yang, Y. Xu, W. Zheng, X. Xia, et al., Glm-130b: An open bilingual pre-trained model, arXiv preprint arXiv:2210.02414 (2022).
- [14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database (2009). URL: <http://dx.doi.org/10.1109/cvpr.2009.5206848>. doi:10.1109/cvpr.2009.5206848.
- [15] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context (2014) 740–755. URL: [http://dx.doi.org/10.1007/978-3-319-10602-1\\_48](http://dx.doi.org/10.1007/978-3-319-10602-1_48). doi:10.1007/978-3-319-10602-1\_48.
- [16] H. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. Le, J. Wei, Scaling instruction-finetuned language models (2022).