Enhancing Vision-Language Models for Global Cultural Understanding through Semantic Expansion and Diversity Reranking

Zirui Hou Xiangzhe Yin Guangyu Gao*
School of Computer Science, Beijing Institute of Technology
guangyu.ryan@gmail.com

Abstract

Current Vision-Language Models (VLMs) often lack sufficient understanding and representation of cultural diversity in globalized scenarios. To bridge this gap, we propose a novel approach, named Semantic Expansion and Diversity Optimization (SEDO), an innovative method that leverages external knowledge bases for semantic enrichment, employs diversity-aware reranking, and uses Segment Anything Model (SAM) for precise localization refinement. Using the GlobalRG benchmark, SEDO significantly improves retrieval relevance (88%*) and cultural diversity (*79.75%*), achieving an Intersection over Union (IoU) of* 0.7012 *in visual cultural grounding tasks. Comprehensive experiments confirm the effectiveness of each proposed component, underscoring robust performance and generalization across diverse cultural contexts. Our work provides valuable guidance toward more inclusive and culturally aware vision-language models.*

1. Introduction

Recently, Multimodal Visual Language models (VLMs) have shown great potential in tasks such as image retrieval, generation, and captioning. However, their understanding of cultural diversity is still insufficient in the context of globalization. To this end, the GlobalRG Challenge [2] launched by the VLMs-4-All Workshop aims to comprehensively examine the performance of VLMs in terms of cultural inclusion and diversity. The GlobalRG benchmark introduces two tasks: (i) Retrieval Across Universals and (ii) Cultural Visual Grounding. The Retrieval task not only focuses on the model's understanding of universal concepts but also examines its ability to retrieve diverse images in a cross-cultural context. While in the Grounding task, the evaluation focuses on the model's ability to incorporate culturally specific objects and concepts into parsed images.

*Corresponding Author. This work was supported by the National Natural Science Foundation of China under No. 62472033.

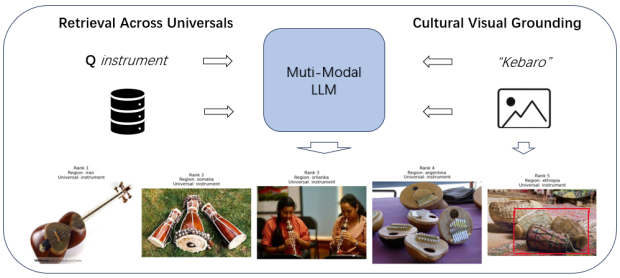

Figure 1. GlobalRG tasks overview: (i) Retrieval Across Universals tests a VLM's ability to retrieve culturally diverse images; (ii) Cultural Visual Grounding evaluates its ability to recognize and ground culturally specific concepts.

Current multimodal models have achieved substantial progress, notably through frameworks such as CLIP [22], which excels in cross-modal image-text alignment and retrieval. In visual grounding tasks, models like Grounding DINO [17] effectively localize textual descriptions within images. Additionally, generalist multimodal models such as Qwen-VL [1] and MiniGPT [5] have shown strong versatility, including the ability to predict bounding box coordinates. However, these models still encounter difficulties in recognizing culturally-specific objects and customs, often misinterpreting their significance or failing to prioritize culturally relevant retrieval results.

Inspired by human cognitive processes in learning and comprehending unfamiliar concepts, we propose an approach named Semantic Expansion and Diversity Optimization (SEDO). SEDO integrates external knowledge bases to enhance semantic understanding through prior knowledge embedding and semantic expansion. Additionally, for the retrieval task, we introduce a diversity-aware reranking algorithm to improve the cultural diversity of retrieved images. For the grounding task, we employ a secondary refinement step that uses the Segment Anything Model (SAM) to improve the localization precision. Experimental results validate the effectiveness of SEDO, achieving a retrieval relevance of 88%, a regional diversity of 79.75%, and an IoU

of 0.7012 in the grounding task, significantly outperforming the baseline and securing the second ranking in the challenge. Through this work, our objective is to establish robust methods for VLMs to better understand and represent cultural nuances, ultimately contributing to the global applicability and inclusiveness of multimodal technologies.

2. Related work

2.1. Text-to-image retrieval

Text-to-image retrieval is to retrieve semantically relevant images from a database given a text query. Early methods primarily relied on two-stream architectures [8, 12], which learned separate representations for text and images and mapped them into a shared semantic space for similarity matching. Recently, the introduction of Transformer-based architectures [6, 18] has significantly advanced this task. Notably, CLIP [21] established a new paradigm for cross-modal representation learning by leveraging large-scale contrastive training on image-text pairs. This approach aligns visual and textual representations within a unified embedding space, achieving strong zero-shot performance on image-text retrieval. Building on CLIP, subsequent methods have further improved cross-modal retrieval by scaling up training datasets or refining model architectures [11, 13, 28]. These CLIP-style models now represent the mainstream in vision-language retrieval, characterized by large-scale pretraining, contrastive learning, and strong generalization capabilities across diverse tasks.

2.2. Visual Grounding

Visual grounding [19, 20] seeks to align natural language descriptions with specific regions within an image. Early approaches relied on object detectors [9, 24], which generated candidate regions and matched them to language embeddings to achieve localization. These methods were later enhanced by the introduction of attention mechanisms [25], enabling finer-grained interactions between textual and visual modalities. More recently, the emergence of large-scale pre-trained cross-modal models has enabled end-to-end training [14], significantly improving semantic understanding and nuanced localization. Grounding DINO, for instance, integrates the Transformer-based detector DINO [30] with large-scale semantic pretraining. By combining feature enhancement, language-guided query formulation, and cross-modal decoding, it effectively links arbitrary objects with textual descriptions and has achieved strong performance on benchmarks such as COCO [16] and LVIS [10]. With the advancement of vision-language models (VLMs), visual grounding capabilities have further improved. Generalist models like QwenVL are trained end-to-end with high-resolution visual inputs and multilingual textual data, enabling precise localization, robust text understanding, and strong zero-shot generalization across diverse cultural and linguistic settings.

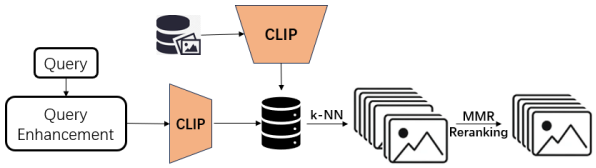

Figure 2. Overview of the Retrieval Across Universals framework: image features are encoded by CLIP and indexed.

2.3. Open Vocabulary Grounding

In the task, we want to improve the understanding and recognition of culture-specific objects and concepts, that is, our task is in the Open Vocabulary [27] setting. Open Vocabulary is a special setting in zero-shot learning. Its concept was first proposed in OVR-CNN [29] and became a popular setting in the detection field with the introduction of CLIP. Ov-vg [26] proposes to use CLIP as a text encoder in a zero-shot framework, while using additional training data to enhance the generalization ability of new categories. In addition, using external knowledge bases or prior knowledge is also an important zero-shot learning strategy [7]. This type of method uses pre-built semantic knowledge graphs or information extracted from large knowledge bases such as Wikipedia to enrich the semantic expression of categories, thereby making up for the problem of insufficient training data. For example, VisPRE [15] verifies that the performance of VLM is positively correlated with the prior knowledge of the visual encoder by introducing a new metric, while end-to-end fine-tuning VLM is not effective in improving performance. Wiki-LLaVA [3] uses external document knowledge sources to improve the effectiveness of VLM in dealing with questions and dialogues, and it is able to maintain the proficiency of VLM in different tasks. Inspired by the above, our method refers to an external knowledge base, so that the concept is expanded and enhanced in the preprocessing stage.

3. Proposed Method

The lightweight approach SEDO is proposed to enhance VLMs' cultural understanding. It integrates semantic expansion, diversity-aware reranking, and segmentation refinement to improve retrieval and grounding performance.

3.1. Retrieval Across Universals

Figure 2 illustrates the overall framework of Retrieval Across Universals. Before querying, all images will be encoded by the CLIP model and then saved to the vector repository for index construction. When querying, the

query will be processed by semantic expansion and enhancement, encoded by clip, and the cosine similarity with the encoded vector of the image will be calculated. It will be sorted by the k-NN algorithm to obtain a set of candidate images with high relevance, and finally re-sorted by the diversity algorithm to obtain a set of images with high relevance and guaranteed diversity.

Query Enhancement We use query semantic enhancement in the Retrieval task, mainly because there are many semantic overlaps and confusions in the given queries. For example, *dinner* and *lunch*, *music* and *instrument*, etc. The image content corresponding to these words is very similar in most cultures, which will cause great difficulties for the accuracy of the query. For example, we expanded the interpretation of *music* to *"A group of people playing different instruments."*, emphasizing the scene of playing different instruments, and thus distinguishing it from *instrument*. We use the Query Enhancement method, mainly to improve the model's retrieval relevance capabilities.

Reranking To improve the diversity of Retrieval output, we borrowed and introduced the Maximal Marginal Relevance(MMR) algorithm[4]. The MMR algorithm aims to reduce the redundancy of sorting results while ensuring the relevance of the results. It was first applied to fields such as text summary extraction and information retrieval. Under multimodal retrieval, we give the formula as follows:

$$\begin{aligned} \text{MMR}(Q, D, C) = \arg\max_{I_i \in D} \Big[& \lambda \text{sim}(Q, I_i) \\ & - (1-\lambda) \max_{I_j \in k} (\text{sim}(I_i, I_j)) \Big] \end{aligned} \tag{1}$$

where Q is the query text, D is the image dataset, and C is the candidate set initially retrieved based on relevance. The first term in the formula is the similarity between the query and the candidate image, and the second term is the similarity between the candidate images. The weights of the two are controlled by the parameter λ. Therefore, while considering and ensuring the relevance of the retrieval, the Formula 1 increases the distance between similar images, making the retrieval results more diverse. When we calculate the similarity between images, we consider that images of the same regional culture will be closer in the semantic space, and therefore have a higher similarity. Therefore, formula 1 has a great effect on the model's ability to retrieve diverse images.

3.2. Cultural Visual Grounding

Figure 3 illustrates the overall framework of Cultural Visual Grounding. First, the concept will be semantically expanded based on the external knowledge base, so that unfamiliar words are expanded into sentences that are easy to understand and describe. Then the concept and image are passed into the VLM to form a prompt, which will answer and return a preliminary prediction box. In post-processing, we selected the SAM segmentation model, using the preliminary box as the prompt to perform precise segmentation on the image, and then obtain the final precise prediction box through the segmentation mask.

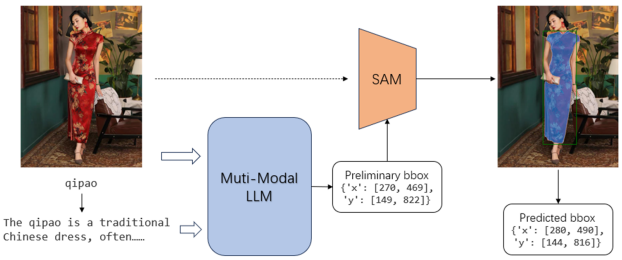

Figure 3. Overall framework of Cultural Visual Grounding. Concepts are first expanded using external knowledge; the VLM generates initial boxes, which SAM refines using them as spatial cues for accurate prediction.

Semantic Extensions It is impossible for the model to understand and recognize completely unfamiliar concepts, so we believe that using external knowledge bases is a necessary and effective method. Therefore, for the given concept, we will use Wikipedia to obtain its explanation to build a knowledge base for it. In addition, considering the model's logic of understanding the text, we use GPT to adjust the order and number of words in the extended explanation. In detail, we set the concept to give priority to "what it is", then "what it looks like/what it is made of", and finally "what it is used for", and control the number of words to keep the semantics concise. Through the above setting method, the model will better capture the semantic information of the concept.

Models Currently, the VLMs suitable for Visual Grounding are divided into Specialist Model and Generalist Model. Specialist Model refers to a model specifically suitable for this task, with Grounding DINO as the main one. This model inputs text and images and returns an accurate prediction box directly. While Generalist Model is a general multimodal large model, suitable for many multimodal tasks such as VQA and Captioning, etc. This model needs to build specific prompts and restrict the model output format. The model will give a prediction box in the corresponding format in a dialogue. Compared with Generalist Models, Grounding DINO has limited ability to understand sentences, and will automatically capture all nouns that appear in the text, which will affect the recognition of the target concept. Therefore, based on the semantic extension method and practical experimental comparison, we choose the universal model QwenVL as the target base model.

However, the disadvantage of the Generalist Model is that the prediction box coordinates given by the model are

Method	Relevance		Diversity(Country)		Diversity(Region)	
	pre@5	pre@10	div@5	div@10	div@5	div@10
openclip(Baseline)	78.00	76.50	93.11	91.95	72.53	61.43
+query enhancement	90.00	86.50	95.69	92.25	74.26	62.75
+Reranking	88.00(-)	88.00	95.69	93.75	79.75	66.36

Table 1. Results & Ablation Study on the retrieval across universals task, in terms of Relevance and Diversity

often not accurate enough. After inspection, it is found that the prediction box always has a certain offset in position or size, as shown in Figure 3. Therefore, we use SAM [23], a general image segmentation model which can output accurate segmentation results through certain prompts. We use the preliminary prediction box answered by QwenVL as the prompt, and after obtaining the segmented mask, we calculate the coordinates to get the final accurate prediction box.

4. Experiments

4.1. Implementation Detail

Datasets The datasets are provided by the official challenge project. The Retrieval task contains 3000 images from 50 countries on average and 20 common concepts as queries, thus avoiding imbalance in diversity metrics. The Grounding task contains 590 test images from 15 countries. The only concept corresponding to each image is the target that the task needs to locate.

Evaluation Setup The evaluation criteria are also set by the official challenge project. The first task needs to evaluate the relevance and diversity of the retrieval. The relevance is measured by the standard *precision@k*, which is the proportion of correctly retrieved images among k images; and for diversity, the official proposes the *diversity@k* indicator, which uses entropy to measure the cultural diversity of the first k images retrieved:

$$\text{diversity@k} = -\frac{1}{\log m}\sum_{i=1}^{m} p_i \log(p_i) \tag{2}$$

Where p_i is the proportion of images from the i_{th} country in the first k images retrieved, and m is the total number of countries that appear in the images. For Grounding task, the IoU metric commonly used in object detection is used for evaluation.

4.2. Results & Ablation Study

The results of the Retrieval Across Universals task are shown in Table 1. Considering the universality of our method on all models, we only selected OpenClip's ViT-B-16 as the baseline. It can be seen that query enhancement has a huge effect on improving retrieval relevance. Considering that the reranking algorithm will sacrifice a certain degree of accuracy, the relevance *prec@5* has decreased, but the diversity, especially the region diversity output, has been significantly improved. The above analysis shows that our method is effective in improving the model's understanding ability and diversity output performance.

Method	checkpoints	IoU
specialist model		
Grounding DINO	GroundingDINO-B	0.4754
+Sematic Extensions		0.635
generalist model		
QwenVL	Qwen2.5-7B	0.4695
+Sematic Extensions		0.6129
+SAM	sam2.1_large	0.7012

Table 2. Results & Ablation Study Cultural Visual Grounding task.

Rank	id	Public Score(IoU)
1	GroundingCulture	0.7148
2	Excelsior7	0.7012
3	spearhead	0.601
4	huruhao	0.5891
5	lastee1e12	0.4601
6	Baseline (Qwen VL 7B)	0.4499

Table 3. Results Comparison. Our solution achieves 0.7012 IoU on the testset, ranking 2nd in the challenge.

The results of the Cultural Visual Grounding task are shown in Table 2. We tested both the Specialist Model and the Generalist Model. After adding Semantic Extensions, the results of both models were greatly improved, which shows that this method is directly effective in improving the model's understanding ability. However, the score of the QwenVL is not as good as the Grounding DINO at this time, so we used SAM in the post-processing stage to help with accurate segmentation, and then obtained the optimal IoU score. Table 3 shows the performance of our solution compared with other teams. Our approach ranks second on the Private Leaderboard.

5. Conclusion

In this work, we present a novel approach to the GlobalRG Challenge, addressing the limitations of Vision-Language Models (VLMs) in capturing cultural diversity and unfamiliar concepts. Our method, Semantic Expansion and Diversity Optimization (SEDO), is a simple yet effective approach for enhancing VLMs' cultural understanding in the GlobalRG Challenge. By integrating semantic expansion, diversity-aware reranking, and SAM-based refinement, our method achieves fairly good performance in both retrieval and grounding tasks. The results demonstrate the value of external knowledge and diversity optimization in building more culturally inclusive vision-language models.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1

[2] Mehar Bhatia, Sahithya Ravi, Aditya Chinchure, Eunjeong Hwang, and Vered Shwartz. From local concepts to universals: Evaluating the multicultural understanding of vision-language models. *arXiv preprint arXiv:2407.00263*, 2024. 1

[3] Davide Caffagni, Federico Cocchi, Nicholas Moratelli, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Wiki-llava: Hierarchical retrieval-augmented generation for multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1818–1826, 2024. 2

[4] Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336, 1998. 3

[5] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 1

[6] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120, 2020. 2

[7] Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *CoRR*, 2023. 2

[8] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems*, 26, 2013. 2

[9] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015. 2

[10] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 2

[11] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916, 2021. 2

[12] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014. 2

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

[14] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 2

[15] Qiao Liang, Yanjiang Liu, Ben He, Yaojie Lu, Hongyu Lin, Jia Zheng, Xianpei Han, Le Sun, and Yingfei Sun. Expanding the boundaries of vision prior knowledge in multi-modal large language models. *arXiv preprint arXiv:2503.18034*, 2025. 2

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European conference on Computer vision*, pages 740–755, 2014. 2

[17] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55, 2024. 1

[18] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 2

[19] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 2

[20] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *Proceedings of the European Conference on Computer Vision*, pages 792–807, 2016. 2

[21] 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*, pages 8748–8763, 2021. 2

[22] 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*, pages 8748–8763, 2021. 1

[23] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 4

[24] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2016. 2

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[26] Chunlei Wang, Wenquan Feng, Xiangtai Li, Guangliang Cheng, Shuchang Lyu, Binghao Liu, Lijiang Chen, and Qi Zhao. Ov-vg: A benchmark for open-vocabulary visual grounding. *Neurocomputing*, 591:127738, 2024. 2

[27] Jianzong Wu, Xiangtai Li, Shilin Xu, Haobo Yuan, Henghui Ding, Yibo Yang, Xia Li, Jiangning Zhang, Yunhai Tong, Xudong Jiang, et al. Towards open vocabulary learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7):5092–5113, 2024. 2

[28] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 2

[29] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14393–14402, 2021. 2

[30] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 2