DQ-DETR: Dual Query Detection Transformer for Phrase Extraction and Grounding

Shilong Liu[1,2*], **Shijia Huang**[3], **Feng Li**[2,4], **Hao Zhang**[2,4], **Yaoyuan Liang**[5], **Hang Su**[1], **Jun Zhu**[1†], **Lei Zhang**[2†]

[1] Dept. of CST, BNRist Center, Inst. for AI, Tsinghua-Bosch Joint Center for ML, Tsinghua University.
[2] International Digital Economy Academy (IDEA).
[3] The Chinese University of Hong Kong.
[4] The Hong Kong University of Science and Technology.
[5] Tsinghua-Berkeley Shenzhen Institute, Tsinghua University.
{liusl20,liang-yy21}@mails.tsinghua.edu.cn, {fliay,hzhangcx}@connect.ust.hk, sjhuang@cse.cuhk.edu.hk,
{suhangss,dcszjmail.tsinghua.edu.cn, {leizhang}@idea.edu.cn

Abstract

In this paper, we study the problem of visual grounding by considering both phrase extraction and grounding (PEG). In contrast to the previous phrase-known-at-test setting, PEG requires a model to extract phrases from text and locate objects from image simultaneously, which is a more practical setting in real applications. As phrase extraction can be regarded as a 1D text segmentation problem, we formulate PEG as a dual detection problem and propose a novel DQ-DETR model, which introduces dual queries to probe different features from image and text for object prediction and phrase mask prediction. Each pair of dual queries is designed to have shared positional parts but different content parts. Such a design effectively alleviates the difficulty of modality alignment between image and text (in contrast to a single query design) and empowers Transformer decoder to leverage phrase mask-guided attention to improve the performance. To evaluate the performance of PEG, we also propose a new metric CMAP (cross-modal average precision), analogous to the AP metric in object detection. The new metric overcomes the ambiguity of Recall@1 in many-box-to-one-phrase cases in phrase grounding. As a result, our PEG pre-trained DQ-DETR establishes new state-of-the-art results on all visual grounding benchmarks with a ResNet-101 backbone. For example, it achieves 91.04% and 83.51% in terms of recall rate on RefCOCO testA and testB with a ResNet-101 backbone. Code will be available at https://github.com/IDEA-Research/DQ-DETR.

Introduction

Visual grounding aims to locate objects referred to by language expressions or phrases, which closely relates to object detection (DET, Fig. 1 (a)) in vision. It has received increasing attention for its potential to benefit other multi-modal tasks like visual question answering (VQA) (Fukui et al. 2016) and image retrieval (Karpathy, Joulin, and Fei-Fei 2014; Radenovic, Tolias, and Chum 2016).

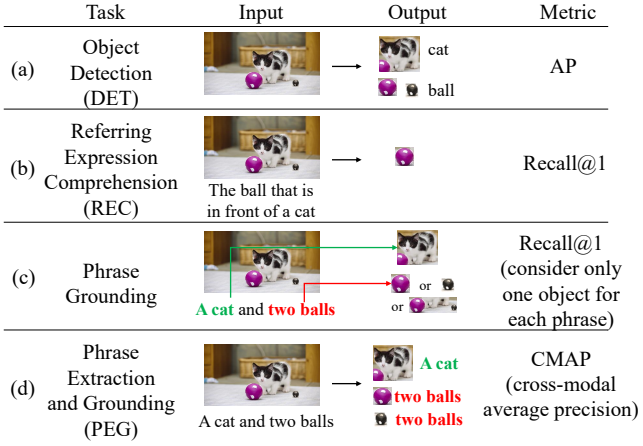


Figure 1: Comparisons of different settings. (a) Given an image, object detection (DET) is to locate objects in predefined categories. The most popular metric for DET is mAP (Lin et al. 2014). (b) Referring expression comprehension (REC) aims at locating objects described by an input text. Its performance is normally evaluated by the recall of the most confident outputs. (c) Phrase grounding needs to ground the spatial regions described by phrases in an image. Most methods treat this task as a ranking problem and evaluate it by recall. The phrases in sentences are usually assumed known during inference. (d) We re-emphasize a phrase extraction and grounding (PEG) setting in this paper. A key difference compared with phrase grounding is that phrases in PEG are unknown during test. We propose a CMAP (cross-modal average precision) metric for the PEG task, analogous to mAP for DET. In this paper, we use the term "visual grounding" for all of three tasks: REC, phrase grounding, and PEG.

Some works (Deng et al. 2021; Huang et al. 2022) treat the terms *visual grounding*, *referring expression comprehension (REC)*, and *phrase grounding* interchangeable. How-

*This work was done when Shilong Liu, Yaoyuan Liang, Feng Li, Shijia Huang, and Hao Zhang were interns at IDEA.
†Corresponding author.

ever, they have subtle differences. Both REC and phrase grounding are sub-tasks of visual grounding. REC locates objects with a free-form guided text, as shown in Fig. 1 (b). It has only one category of objects to detect as requested by a referring expression, while phrase grounding needs to find all objects mentioned in a caption, as shown in Fig. 1 (c). Though they have different definitions, phrase grounding can be reformulated as a REC task by extracting phrases as referring expressions since phrases are assumed known during test (Deng et al. 2021; Du et al. 2022; Huang et al. 2022). Some methods (Mu et al. 2021; Liu et al. 2019b) use non-REC solutions for phrase grounding, while they also treat phrases as known during test.

We argue it is more practical to treat phrases as unknown during test and study the problem of visual grounding in this paper by considering both phrase extraction and grounding (PEG), as shown in Fig. 1 (d). Solving PEG by developing a large-scale image-text-paired training dataset with both phrases and objects annotated is prohibitively costly. A simple way to extend existing REC models (Wang et al. 2022; Zhu et al. 2022) to PEG is to develop a two-stage solution: firstly extracting phrases using an NLP tool like spaCy (Honnibal and Montani 2017) and then applying a REC model. However, such a solution may result in inferior performance (as shown in our Table 2) as there is no interaction between the two stages. For example, an image may have no object or more than one object that corresponds to an extracted phrase. Yet most REC models (Miao et al. 2022; Zhu et al. 2022) predict only one object for each extracted phrase. Let alone inaccurate phrase extraction can mislead a REC model to predict unrelated objects.

We are not the first to propose the PEG setting. Some previous works (Karpathy, Joulin, and Fei-Fei 2014; Karpathy and Fei-Fei 2014) align image regions and phrases for image retrieval. Flickr30k Entities (Plummer et al. 2015) evaluates models under the scenario in which phrases are unknown as well. They extract noun phrases using NLP tools and penalize recall if the tools extract inaccurate phrases. Despite of such early explorations, most successors (Bajaj, Wang, and Sigal 2019; Deng et al. 2021) treat phrase grounding as a retrieval task and use ground truth phrases as inputs. Hence we re-emphasize the PEG setting, where we predict object-phrase pairs given only a pair of image and text as input, without assuming phrases as known input. We can reformulate all three other tasks (DET, REC, and phrase grounding) as PEG tasks.

PEG lifts the importance of phrase extraction, which is often overlooked in previous task formulations and solutions. Some previous works can be used for the PEG task, such as MDETR (Kamath et al. 2021) and GLIP (Li et al. 2021), which use the same query (in a DETR framework) for both object localization and phrase extraction, as shown in Fig. 2 (c). However, their phrase extraction module requires a query to have an extra capability to perform the challenging image-text feature alignment, which can interfere with the bounding box regression branch and result in an inferior performance.

We note that phrase extraction is to localize a noun phrase from an input text, which can be regarded as a 1D text segmentation problem that predicts a 1D text mask for a target phrase. Such a problem is analogous to 2D mask prediction for an object instance in 2D image segmentation. Especially, inspired by recent progress of DETR-like models (e.g., DINO (Zhang et al. 2022), Mask2Former (Cheng et al. 2022)), we develop a more principled solution DQ-DETR, which is a dual query-based[1] DETR-like model for PEG. As shown in Fig. 2 (d), our model uses dual queries to perform object detection and text mask prediction in one DETR framework. The text mask prediction is very similar to instance mask prediction as in Mask2Former, hence we can use masked-attention Transformer decoder layers to improve the performance of text mask prediction. In DQ-DETR, a pair of dual queries is designed to have shared positional parts but different content parts[2]. Such a decoupled query design helps alleviate the difficulty of modality alignment between image and text, yielding faster convergence and better performance.

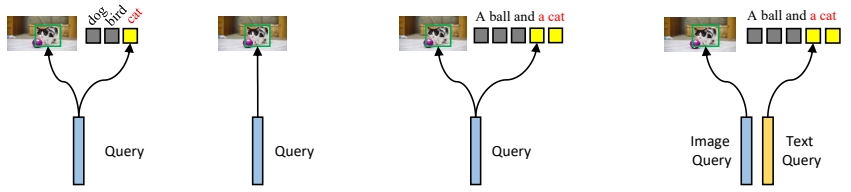
To evaluate models on the PEG setting, we propose a new metric CMAP (cross-modal average precision), which is analogous to the AP metric widely used in object detection. It measures the accuracy of both phrase extraction and object localization, as shown in Fig. 3 (c). The CMAP metric overcomes the ambiguous issue of the previous Recall@1 when multiple objects correspond to one phrase. Recall@1 evaluates the accuracy of boxes with the highest confidence. However, for cases where multiple objects correspond to one phrase, the metric becomes ambiguous. To deal with such cases, previous works (Plummer et al. 2015; Li et al. 2019; Kamath et al. 2021) leveraged two different protocols, which we denote as ANY-BOX (Fig. 3 (a)) and MERGED-BOXES (Fig. 3 (b)) protocols following MDETR (Kamath et al. 2021). The ANY-BOX setting treats a prediction as correct if any of the ground truth boxes is matched. However, it cannot evaluate a model's capability of finding all objects in an image. The other protocol, MERGED-BOXES, combines all objects for one phrase into a big box for evaluation. While being able to capture all objects, this protocol cannot measure the localization accuracy for every object instance.

We summarize our contributions as follows:

1. By comparing three settings in visual grounding: DET, REC, and phrase grounding, we re-emphasize a PEG setting, which is often overlooked in previous works. To take the phrase extraction accuracy into account, we propose a new cross-modal average precision (CMAP) metric for PEG to measure a combined accuracy for both phrase extraction and object localization. The CMAP metric is free of confusion when multiple objects correspond to one phrase.

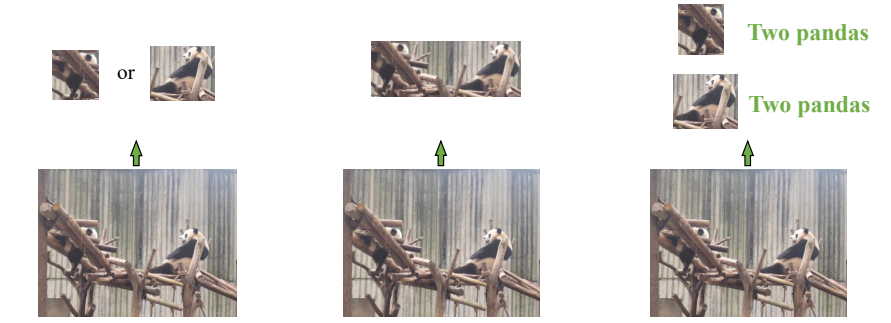
[1]We use the term "query" as the input of the Transformer decoder layers in this paper, following the common practice in the Transformer and DETR-related literature (Vaswani et al. 2017; Meng et al. 2021; Liu et al. 2022). The definition differs from some visual grounding papers, where "query" refers to an input text. A detailed explanation of our dual query is available in the appendix.

[2]A DETR query consists of two parts: a content part and a positional part. More detailed discussions can be referred to (Meng et al. 2021) and (Liu et al. 2022).



(a) DET (b) REC/ Phrase grounding (c) PEG: MDETR/GLIP (d) PEG: DQ-DETR

Figure 2: Comparisons of different models. The "query" here refers to the input of Transformer decoder.



(a) Recall@1, ANY-BOX (b) Recall@1, MERGED-BOXES (c) CMAP(ours)

Figure 3: Comparisons of different metrics. We only plot objects corresponding to the phrases "Two pandas" for a better comparison. (a) and (b) are used for phrase grounding. (a) The ANY-BOX setting treats a prediction as correct if any of the ground truth boxes is matched. (b) MERGED-BOXES combines all objects for one phrase to a big box for evaluation. (c) Our metric CMAP encourages a model to predict all objects and their corresponding phrases.

2. We interpret noun phrase extraction as a 1D text segmentation problem and formulate the PEG problem as predicting both bounding boxes for objects and text masks for phrases. Accordingly, we develop a novel dual query-based DETR-like model DQ-DETR with several techniques to improve the performance of phrase extraction and object localization.
3. We validate our methods on several benchmarks and establish new state-of-the-art results, including Flickr30k, RefCOCO/+/g, and COCO. Our model obtains 76.0% CMAP_{50} and 83.2% Recall@1 at Flickr30k entities (Plummer et al. 2015). Moreover, we achieve 91.04% and 83.51% in terms of recall rate on RefCOCO testA and testB with a ResNet-101 backbone.

PEG & CMAP

We present the PEG (phrase extraction and grounding) problem formulation and the CMAP (cross-modal average precision) definition in this section.

Given an image-text pair as input, PEG requires a model to predict region-phrase pairs from the input image and text pair, as shown in Fig. 3. The PEG task can be viewed as a dual detection problem for image box detection and text mask segmentation, since noun phrase extraction can be interpreted as a 1D text segmentation problem.

To measure both the text phrase extraction accuracy and the image object localization accuracy, we propose a new metric which is similar to the AP metric used in DET. AP is calculated by integrating the area under a P-R curve. The key to plot P-R curves is to decide positive and negative samples. DET benchmarks like COCO (Lin et al. 2014) leverage IOU (intersection over union) between a predicted box and a ground truth box to discriminate positive and negative predictions. As we interpret phrase extraction as a 1D segmentation problem, we use dual IOU to choose positive predictions. The dual IOU is defined as:

$$\text{IOU}_{\text{dual}} = (\text{IOU}_{\text{box}})^{0.5} \times \text{IOU}_{\text{phrase}}, \quad (1)$$

where IOU_{box} is the box IOU and $\text{IOU}_{\text{phrase}}$ is the phrase IOU. We take the square root of IOU_{box} to make IOU_{dual} a two dimensional metric so that its threshold (e.g. 0.5) has a similar meaning to IOU_{box}. Following the common practice in phrase grounding and REC, we use $\text{IOU}_{\text{dual}} >= 0.5$ as positive samples, and vice versa. We use the term "CMAP_{50}" to denote the metric at threshold 0.5.

DQ-DETR

Following DETR (Carion et al. 2020) and MDETR (Kamath et al. 2021), DQ-DETR is a Transformer-based encoder-decoder architecture, which contains an image backbone, a text backbone, a multi-layer Transformer encoder, a multi-layer Transformer decoder, and several prediction heads.

Given a pair of inputs `(Image, Text)`, we extract image features and text features using an image backbone and a text backbone, respectively. The image and text features are flattened, concatenated, and then fed into the Transformer encoder layers. We then use learnable dual queries for the decoder layers to probe desired features from the concatenated multi-modality features. The image queries and text queries will be used for box regressions and phrase localizations, respectively, as shown in Fig. 4 left.

Dual Queries for Dual Detections

We propose to decouple the queries for bounding box regression and phrase localization in DQ-DETR. However, as the dual queries aim to predict paired `(Region, Phrase)` results, both queries need to focus on the same region of an object in the input image and the same position of a phrase in the input text. Hence we propose to share the positional parts and decouple the content parts of the queries. As we formulate the problem as a dual detection problem for image box detection and text phrase segmentation, we introduce two items for the positional queries, i.e., image positional queries and text positional queries. More concretely, the image positional queries are formulated as anchor boxes like DAB-DETR (Liu et al. 2022) and then projected to high dimensions with sine/cosine encoding. The text positional queries are formulated as 1D segmentation masks like Mask2Former (Cheng et al. 2022) and then used for text mask-guided attention. The image positional queries are predicted by the updated image queries, and the text positional queries are generated by performing dot product between the updated image queries and text features from the encoder output, as shown in Fig. 4 right. Both two positional queries will be shared by the dual queries as positional parts for the next layer. Beyond the positional and content parts, we add a learnable modality embedding to the features of different modalities. We list the components of our queries in Table 1.

Text Mask-Guided Attention

The 1D segmentation formulation of phrase localization inspires us to propose a text mask-guided attention to let queries focus on phrase tokens of interest, analogous to the mask attention in Mask2Former (Cheng et al. 2022). Each text query has a text positional query $M_T \in 1^{N_{\text{text}}}$, which is a binary mask with the same length as the text features. We use the encoder output image-text-concatenated features as keys and values for cross-attention. The binary masks will be used as attention masks for the text features in the concatenated features. Text features will be used if their corresponding mask values are ones, otherwise they will be masked out. This operation constrains the attention on the target phrases while the predicted masks are updated layer by layer to to get closer to ground truth masks. We use all-ones masks as inputs for the first decoder layer. Other layers will leverage the predicted masks from their previous layers. The final updated masks are the outputs for phrase localization.

Loss Functions

Following DETR (Carion et al. 2020) and MDETR (Kamath et al. 2021), we use bipartite matching to assign ground truth object boxes and text phrases to dual queries during training. The final loss functions can be grouped into boxes losses for images and phrase losses for texts. We use the L1 loss and the GIOU (Rezatofighi et al. 2019) loss for bounding box regression. For phrase localization, we use a contrastive softmax loss.

For a text query at the output of the decoder $Q^{(\text{out})} \in R^{N_q,D}$, we compute the similarities between this query and the encoder output text features $F_T^{(\text{enc})} \in R^{N_{\text{text}},D}$ to predict a text segmentation mask. We first linearly project the decoder output $Q^{(\text{out})}$ to get $Q = \text{Linear}_Q(Q^{(\text{out})}) \in R^{N_q,D_1}$. Then we linearly project the encoder output text feature $F_T^{(\text{enc})}$ to get $F_T = \text{Linear}_T(F_T^{(\text{enc})}) \in R^{N_{\text{text}},D_1}$. The notations D_1 is the dimension of the projected space, and $\text{Linear}_Q, \text{Linear}_T$ are two linear layers. As some queries may not correspond to any desired phrase, similar to a DETR query not matching with any ground truth object, we set an extra learnable `no_phrase` token $\text{NoPhraseToken} \in R^{D_1}$ for `no_phrase` queries. We then concatenate the projected text feature and the `no_phrase` token to get an extended text feature $F'_T = \text{Concat}(F_T, \text{NoPhraseToken}) \in R^{N_{\text{text}}+1,D_1}$.

The final contrastive softmax loss is performed between the projected query features $Q \in R^{N_q,D_1}$ and the extended text feature $F'_T \in R^{N_{\text{text}}+1,D_1}$. Let S_{q_i} be the set of text token indices of a target phrase for a given query $q_i \in Q$. The phrase localization loss for query q_i is:

$$\mathcal{L}_{\text{phrase},i} = \sum_{j \in S_{q_i}} (-\log \frac{q_i^\top p_j / \tau}{\sum_{k=0}^{N_{\text{text}}+1} q_i^\top p_k / \tau}), \qquad (2)$$

where τ is a temperature parameter which is empirically set to 0.07 in our experiments, and $p_j \in F'_T$ is a text feature or a `no_phrase` token with index j. We down-weight the loss by 0.05 when no objects are assigned to the query q_i to balance the classes.

Experiments

Implementation Details

Models. We use two commonly used image backbones, ResNet-50 and ResNet-101 (He et al. 2016) pre-trained on ImageNet (Deng et al. 2009), for our base setting and pre-training setting, respectively. Both two pre-trained models are provided by PyTorch (Paszke et al. 2017). For the text backbones, we use the pre-trained RoBERTa-base (Liu et al. 2019a) provided by HuggingFace (Wolf et al. 2019) in our experiments. We set $D = 256$ and $D_1 = 64$ in our implementations and use 100 pairs of dual queries. Our models use 6 encoder layers and 6 decoder layers. The learning

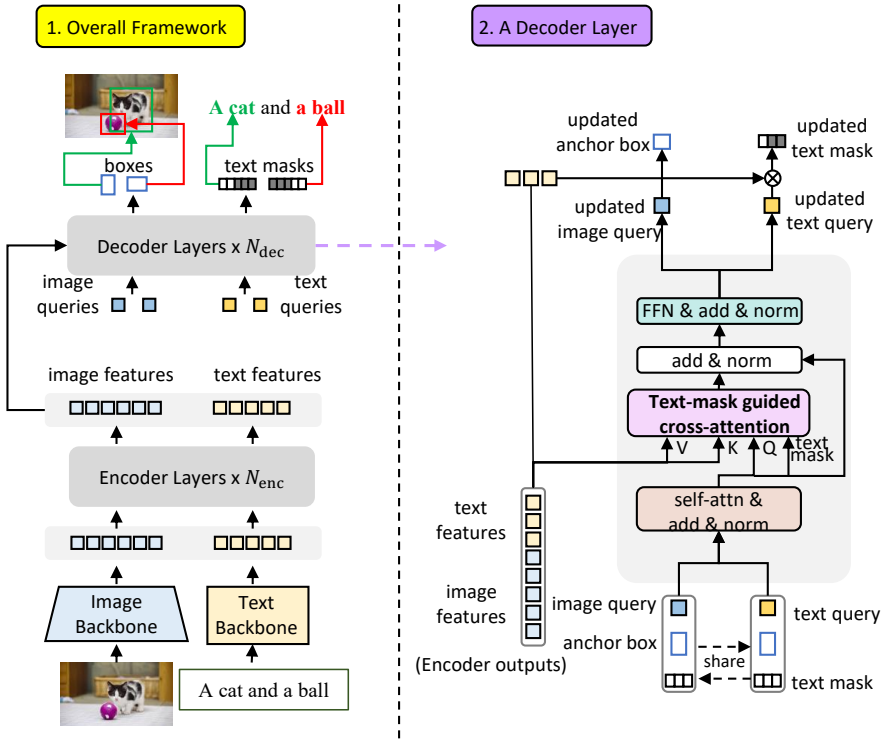


Figure 4: The framework of our proposed DQ-DETR model. The left block is the overall framework. The right block presents the detailed structure of a decoder layer in DQ-DETR.

Queries	Image Queries $Q_I \in R^{N_q,D}$	Text Queries $Q_T \in R^{N_q,D}$
Content Parts	$Q_I^{(C)} \in R^{N_q,D}$	$Q_T^{(C)} \in R^{N_q,D}$
Image Positional Parts	$A_I \in R^{N_q,4}$	$A_T = A_I$
Text Positional Parts	$M_I = M_T$	$M_T \in \{0,1\}^{N_{\text{text}}}$
Modality Embeddings	$\text{ModalityToken}_I \in R^D$	$\text{ModalityToken}_T \in R^D$

Table 1: The table of components of our dual queries.

schedules are different for different settings, which will be described in each subsection. The initial learning rates for the Transformer encoder-decoder and image backbone are $1e^{-4}$ and $1e^{-5}$, respectively. For the text backbone, we use a linear decay schedule from $5e^{-5}$ to 0 with a linear warm-up in the first 1% steps. To stabilize the bipartite graph matching, we use anchor denoising (Li et al. 2022) in our implementations.

The Pre-training Setting

Pre-training Task: PEG Following MDETR (Kamath et al. 2021), we use the combined dataset of Flickr30k, COCO, and Visual Genome for our pre-training. The backbone we used is ResNet-101. We pre-train our model on the combined dataset for 25 epochs and drop the initial learning rate by 10 after the 20-th epoch. The pre-training takes about 100 hours on 16 Nvidia A100 GPUs with 4 images per GPU. We then fine-tune the model on different tasks with 4 GPUs, except for the object detection task, which needs 8 GPUs.

We compare our model with three baselines in Table 2. We use the state-of-the-art REC model OFA-REC (Wang et al. 2022) [3] to demonstrate the necessary of our PEG. OFA is an unified model pre-trained with more than 50M images and can be used for REC tasks. To adapt it to our PEG task, we use spaCy (Honnibal and Montani 2017) to extract noun phrases. The results show that OFA-REC+spaCy is much worse than the other two end-to-end models in terms of the CMAP_{50} metric. One important reason is the failure when multiple objects correspond to one phrase. To decouple the effect of phrase extraction and REC, we design another baseline with spaCy and an ideal REC model named GoldREC. GoldREC outputs the ground-truth object whose corresponding phrase is the most similar to input phrases for any given input phrase. It shows that inaccurate phrase extraction has a large impact on final performance.

[3] We use the OFA_{Base} provided in https://github.com/OFA-Sys/OFA.

Model	Pre-train Data	Epoches	$CMAP_{50}$	R@1
OFA-REC+spaCy	CC, SBU, COCO, VG, OI, O365, YFCC (50M)	-	23.2	58.1
GoldREC+spaCy	-	-	44.4	100.0
MDETR	COCO, VG, Flickr30k (200k)	50	70.2	82.5
DQ-DETR (Ours)	COCO, VG, Flickr30k (200k)	**25**	**76.0** (+5.8)	83.2

Table 2: Pre-training result comparison on Flickr30k Entities with baselines.

Method	Val			Test		
	R@1	R@5	R@10	R@1	R@5	R@10
ANY-BOX Protocol						
BAN	-	-	-	69.7	84.2	86.4
VisualBert	68.1	84.0	86.2	-	-	-
VisualBert	70.4	84.5	86.3	71.3	85.0	86.5
MDETR	82.5	92.9	94.9	83.4	93.5	95.3
DQ-DETR(Ours)	**83.2** (+0.7)	**93.9**	**95.6**	**83.9**	**94.6**	**96.2**
MERGED-BOXES Protocol						
CITE	-	-	-	61.9	-	-
FAOG	-	-	-	68.7	-	-
SimNet-CCA	-	-	-	71.9	-	-
DDPN	72.8	-	-	73.5	-	-
RefTR	-	-	-	81.2	-	-
SeqTR	-	-	-	81.2	-	-
MDETR	82.3	91.8	93.7	83.8	92.7	94.4
DQ-DETR(Ours)	**83.7** (+1.4)	**93.8**	**95.8**	**84.3**	**93.9**	**95.5**

Table 3: Results on the phrase grounding task on Flickr30k Entities.

We also use MDETR (Kamath et al. 2021) as a baseline. Our model outperforms MDETR on Flickr30k Entities with only half of the number of training epochs. It outperforms MDETR by $+5.8$ $CMAP_{50}$, demonstrating the effectiveness of decoupling image and text queries. We provide a visualization of these models' results in Appendix Table 7

Down-stream Task: Phrase Grounding We compare our DQ-DETR with BAN (Kim, Jun, and Zhang 2018), VisualBert (Li et al. 2019), CITE (Plummer et al. 2017), FAOG (Yang et al. 2019), SimNet-CCA (Plummer et al. 2020), DDPN (Yu et al. 2018), RefTR (Li and Sigal 2021), SeqTR (Zhu et al. 2022), and MDETR (Kamath et al. 2021) in Table 3. We fine-tune our pre-trained model on Flickr30k (Plummer et al. 2015) for the phrase grounding task. To compare with previous works in the literature, we follow MDETR (Kamath et al. 2021) and evaluate the models with Recall@k under two different protocols, ANY-BOX and MERGED-BOXES protocols. For the ANY-BOX protocol, we evaluate our pre-trained model on the validation and test splits directly. For the MERGED-BOXES protocol, we fine-tune the pre-trained model for 5 epochs. Our model introduces improvements of $+0.7$ Recall@1 and $+1.4$ Recall@1 on the two validation splits, with only half of the number of pre-training epochs compared with MDETR. We also establish new state-of-the-art results on the two benchmarks with a ResNet-101 backbone.

Down-stream Task: REC We compare our model with state-of-the-art REC methods on RefCOCO/+/g benchmarks after fine-tuning in Table. 4. We evaluate the models with Recall@1. Although our model is not specifically designed for REC tasks, we can convert the REC task to a PEG problem by marking the whole sentence as a phrase corresponding to its referred object. As there are no ground truth phrases labeled in the dataset, we do not use the text mask-guided attention in the fine-tuning process. As we have leveraged all training data of the three REC datasets during pre-training, it is reasonable to fine-tune the models on a combination of these three datasets. To avoid data leakage, we removed all images appeared in the val/test splits of RefCOCO/+/g. This operation removes about 10% of the total images. As a result, our model outperforms all previous works with a ResNet-101 backbone and establishes new state-of-the-art results on the RefCOCO/+/g benchmarks.

Down-stream Task: DET Refer to for more details.

The Base Setting & Ablations

We use MDETR (Kamath et al. 2021) as our baseline and our model is an improvement upon it. As there are only pre-trained models in the MDETR paper, which is not easy for the community to compare, we design the base setting on Flickr30 Entities without pre-training in this section. All models are trained on 4 Nvidia A100 GPUs with a ResNet-50 backbone and each GPU contains 4 images.

Results on Flickr30k Entities and ablations. All models for Flickr30k are trained for 24 epochs with a learning rate drop at the 20th epoch. We compare our DQ-DETR and MDETR in Table 5. Our proposed DQ-DETR outperforms

Method	Pre-training	RefCOCO			RefCOCO+			RefCOCOg	
	image data	val	testA	testB	val	testA	testB	val	test
MAttNet	None	76.65	81.14	69.99	65.33	71.62	56.02	66.58	67.27
VGTR	None	79.20	82.32	73.78	63.91	70.09	56.51	65.73	67.23
TransVG	None	81.02	82.72	78.35	64.82	70.70	56.94	68.67	67.73
ViLBERT	3.3M	-	-	-	72.34	78.52	62.61	-	-
VL-BERT_L	3.3M	-	-	-	72.59	78.57	62.30	-	-
UNITER_L	4.6M	81.41	87.04	74.17	75.90	81.45	66.70	74.86	75.77
VILLA_L	4.6M	82.39	87.48	74.84	76.17	81.54	66.84	76.18	76.71
ERNIE-ViL_L	4.3M	-	-	-	75.95	82.07	66.88	-	-
RefTR	100k	85.65	88.73	81.16	77.55	82.26	68.99	79.25	80.01
SeqTR	174k	87.00	90.15	**83.59**	78.69	84.51	71.87	82.69	83.37
OFA	50M	88.48	90.67	83.30	81.39	87.15	74.29	82.29	82.31
MDETR	200k	86.75	89.58	81.41	79.52	84.09	70.62	81.64	80.89
DQ-DETR(Ours)	200k	**88.63**	**91.04**	83.51	**81.66**	**86.15**	**73.21**	**82.76**	**83.44**

Table 4: Top-1 accuracy comparison on the referring expression comprehension task.

Model	CMAP_{50}	**R@1**	**R@5**
MDETR	61.49	77.46	88.28
MDETR *with 200 queries*	56.89	76.29	86.87
Our baseline for DQ-DETR	66.68	75.44	87.94
+ text mask attention	68.26	76.58	88.46
+ dual queries	69.86	78.87	89.39
+ positional query sharing	**70.63**	**79.16**	**89.84**

Table 5: Ablation results for DQ-DETR and a comparison with MDETR. All models are trained with a ResNet-50 backbone for 24 epochs, with a learning rate drop at the 20-th epoch. We use the ANY-BOX protocol for Recall@k metrics in this table.

MDETR with a large margin, e.g., +13.14% CMAP_{50} and +4.70% Recall@1 on Flickr30k entities.

We provide the ablations in Table 5 as well. The dual query design introduces a gain of 1.60% CMAP_{50} in our experiments, which demonstrates the effectiveness of our dual query design. Moreover, we find the positional query sharing strategy helps improve the results as well, which highlights the necessity of our carefully designed dual queries. The multi-scale design and the text mask-guided attention also help the training of our model, each of which introduces a gain of $1-2\%$ CMAP_{50}. We train a variant of MDETR with 200 queries for a fair comparison with our models. Surprisingly, we find the result drops a lot. We speculate that the data imbalance leads to the result, as more queries will exacerbate the imbalance of classes. In contrast, our model, which outputs 100 results like the original DETR, will not suffer from this problem. We will leave it as a future work to study how to scale up the model with more queries.

Related Work

We present the related work in the appendix.

Conclusion

We have presented an overview of visual grounding tasks and identified an often overlooked phrase extraction step. The observation inspires us to re-emphasize a PEG (phrase extraction and grounding) task and propose a new CMAP (cross-modal average precision) metric. The CMAP overcomes the ambiguity of Recall@1 in many-box-to-one-phrase cases in phrase grounding tasks. Moreover, we propose a new interpretation of the PEG task as a dual detection problem by viewing phrase localization as a 1D text segmentation problem. With this new interpretation, we developed a dual query-based DETR-like model DQ-DETR for phrase grounding. Such a decoupled query design helps alleviate the difficulty of modality alignment between image and text, yielding both faster convergence and better performance. We also proposed a text mask-guided attention to constrain a text query to the masked text tokens in cross-attention modules. We conducted extensive experiments to verify the effectiveness of our model design.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (2020AAA0106302).
We thank Yukai Shi, Linghao Chen, Jianan Wang, Ailing Zeng, and Xianbiao Qi of IDEA CVR groups for their valuable feedbacks. We thank all the reviewers including SPC and AC in AAAI 2023 for their kindly suggestions. We thank the reviewers of our initial version in NeurIPS 2022 for their valuable suggestions, especially the suggestions of the Reviewer btPY, which helped us a lot.

References

Bajaj, M.; Wang, L.; and Sigal, L. 2019. G3raphGround: Graph-Based Language Grounding. *international conference on computer vision.*

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European Conference on Computer Vision*, 213–229. Springer.

Cheng, B.; Choudhuri, A.; Misra, I.; Kirillov, A.; Girdhar, R.; and Schwing, A. G. 2022. Mask2Former for Video Instance Segmentation.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Deng, J.; Yang, Z.; Chen, T.; Zhou, W.; and Li, H. 2021. TransVG: End-to-End Visual Grounding with Transformers. *arXiv: Computer Vision and Pattern Recognition*.

Du, Y.; Fu, Z.; Liu, Q.; and Wang, Y. 2022. Visual Grounding with Transformers. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*.

Fukui, A.; Park, D. H.; Yang, D.; Rohrbach, A.; Darrell, T.; and Rohrbach, M. 2016. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. *empirical methods in natural language processing*.

Fürst, A.; Rumetshofer, E.; Tran, V. H.; Ramsauer, H.; Tang, F.; Lehner, J. M.; Kreil, D. P.; Kopp, M. K.; Klambauer, G.; Bitto-Nemling, A.; and Hochreiter, S. 2021. CLOOB: Modern Hopfield Networks with InfoLOOB Outperform CLIP. *arXiv: Learning*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Honnibal, M.; and Montani, I. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Huang, J.; Qin, Y.; Qi, J.; Sun, Q.; and Zhang, H. 2022. Deconfounded Visual Grounding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(1): 998–1006.

Kamath, A.; Singh, M.; LeCun, Y.; Synnaeve, G.; Misra, I.; and Carion, N. 2021. MDETR-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1780–1790.

Karpathy, A.; and Fei-Fei, L. 2014. Deep Visual-Semantic Alignments for Generating Image Descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Karpathy, A.; Joulin, A.; and Fei-Fei, L. 2014. Deep Fragment Embeddings for Bidirectional Image Sentence Mapping. *neural information processing systems*.

Kim, J.-H.; Jun, J.; and Zhang, B.-T. 2018. Bilinear Attention Networks. *neural information processing systems*.

Kingma, D. P.; and Ba, J. 2014. Adam: A Method for Stochastic Optimization. *arXiv: Learning*.

Li, F.; Zhang, H.; Liu, S.; Guo, J.; Ni, L. M.; and Zhang, L. 2022. DN-DETR: Accelerate DETR Training by Introducing Query DeNoising. In *Computer Vision and Pattern Recognition (CVPR)*.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2019. VisualBERT: A Simple and Performant Baseline for Vision and Language. *arXiv: Computer Vision and Pattern Recognition*.

Li, L. H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.-N.; et al. 2021. Grounded Language-Image Pre-training. *arXiv preprint arXiv:2112.03857*.

Li, M.; and Sigal, L. 2021. Referring Transformer: A One-step Approach to Multi-task Visual Grounding. *arXiv: Computer Vision and Pattern Recognition*.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollar, P. 2020. Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2): 318–327.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Liu, S.; Li, F.; Zhang, H.; Yang, X.; Qi, X.; Su, H.; Zhu, J.; and Zhang, L. 2022. DAB-DETR: Dynamic Anchor Boxes are Better Queries for DETR. In *International Conference on Learning Representations*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019a. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv: Computation and Language*.

Liu, Y.; Wan, B.; Zhu, X.; and He, X. 2019b. Learning Cross-modal Context Graph for Visual Grounding. *arXiv: Computer Vision and Pattern Recognition*.

Loshchilov, I.; and Hutter, F. 2018. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Meng, D.; Chen, X.; Fan, Z.; Zeng, G.; Li, H.; Yuan, Y.; Sun, L.; and Wang, J. 2021. Conditional DETR for Fast Training Convergence. *arXiv preprint arXiv:2108.06152*.

Miao, P.; Su, W.; Wang, L.; Fu, Y.; and Li, X. 2022. Referring Expression Comprehension via Cross-Level Multi-Modal Fusion. *ArXiv*, abs/2204.09957.

Mu, Z.; Tang, S.; Tan, J.; Yu, Q.; and Zhuang, Y. 2021. Disentangled Motif-aware Graph Learning for Phrase Grounding. *national conference on artificial intelligence*.

Nagaraja, V. K.; Morariu, V. I.; and Davis, L. S. 2016. Modeling Context Between Objects for Referring Expression Understanding. *arXiv: Computer Vision and Pattern Recognition*.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in pytorch.

Plummer, B. A.; Kordas, P.; Kiapour, M. H.; Zheng, S.; Piramuthu, R.; and Lazebnik, S. 2017. Conditional Image-Text Embedding Networks. *european conference on computer vision*.

Plummer, B. A.; Shih, K. J.; Li, Y.; Xu, K.; Lazebnik, S.; Sclaroff, S.; and Saenko, K. 2020. Revisiting

Image-Language Networks for Open-ended Phrase Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Plummer, B. A.; Wang, L.; Cervantes, C. M.; Caicedo, J. C.; Hockenmaier, J.; and Lazebnik, S. 2015. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. *International Journal of Computer Vision*.

Radenovic, F.; Tolias, G.; and Chum, O. 2016. CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples. *european conference on computer vision*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. *arXiv: Computer Vision and Pattern Recognition*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6): 1137–1149.

Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; and Savarese, S. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 658–666.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wang, P.; Yang, A.; Men, R.; Lin, J.; Bai, S.; Li, Z.; Ma, J.; Zhou, C.; Zhou, J.; Yang, H.; and Zhou, C. 2022. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Yang, Z.; Gong, B.; Wang, L.; Huang, W.; Yu, D.; and Luo, J. 2019. A Fast and Accurate One-Stage Approach to Visual Grounding. *arXiv: Computer Vision and Pattern Recognition*.

Yu, Z.; Yu, J.; Xiang, C.; Zhao, Z.; Tian, Q.; and Tao, D. 2018. Rethinking Diversified and Discriminative Proposal Generation for Visual Grounding. *international joint conference on artificial intelligence*.

Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L. M.; and Shum, H.-Y. 2022. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. *arXiv preprint arXiv:2203.03605*.

Zhu, C.; Zhou, Y.; Shen, Y.; Luo, G.; Pan, X.; Lin, M.; Chen, C.; Cao, L.; Sun, X.; and Ji, R. 2022. Seqtr: A simple yet universal network for visual grounding. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, 598–615. Springer.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2021. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *ICLR 2021: The Ninth International Conference on Learning Representations*.