Referencing Where to Focus: Improving Visual Grounding with Referential Query

Yabing Wang [1], **Zhuotao Tian** [2], **Qingpei Guo** [3],
Zheng Qin [1], **Sanping Zhou** [1], **Ming Yang** [3], **Le Wang** [1*]
[1] National Key Laboratory of Human-Machine Hybrid Augmented Intelligence,
National Engineering Research Center for Visual Information and Applications,
and Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University
[2] Harbin Institute of Technology, Shenzhen, China
[3] AntGroup
{wyb7wyb7,qinzheng}@stu.xjtu.edu.cn
tianzhuotao@link.cuhk.edu.hk
{spzhou, lewang}@xjtu.edu.cn
{qingpei.gqp, m.yang}@antgroup.com

Abstract

Visual Grounding aims to localize the referring object in an image given a natural language expression. Recent advancements in DETR-based visual grounding methods have attracted considerable attention, as they directly predict the coordinates of the target object without relying on additional efforts, such as pre-generated proposal candidates or pre-defined anchor boxes. However, existing research primarily focuses on designing stronger multi-modal decoder, which typically generates learnable queries by random initialization or by using linguistic embeddings. This vanilla query generation approach inevitably increases the learning difficulty for the model, as it does not involve any target-related information at the beginning of decoding. Furthermore, they only use the deepest image feature during the query learning process, overlooking the importance of features from other levels. To address these issues, we propose a novel approach, called RefFormer. It consists of the query adaption module that can be seamlessly integrated into CLIP and generate the referential query to provide the prior context for decoder, along with a task-specific decoder. By incorporating the referential query into the decoder, we can effectively mitigate the learning difficulty of the decoder, and accurately concentrate on the target object. Additionally, our proposed query adaption module can also act as an adapter, preserving the rich knowledge within CLIP without the need to tune the parameters of the backbone network. Extensive experiments demonstrate the effectiveness and efficiency of our proposed method, outperforming state-of-the-art approaches on five visual grounding benchmarks.

1 Introduction

Visual grounding is a challenging multi-modal task that involves localizing a specific object based on a given natural language description. This task requires algorithms to comprehend fine-grained human language expressions and accurately establish correspondences with the target objects. In recent years, it has gained significant attention in research due to its potential for advancing vision-language understanding, such as cross-modal retrieval [42, 30, 44, 9, 41, 43] and image captioning [14, 29].

*Corresponding author

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-1502

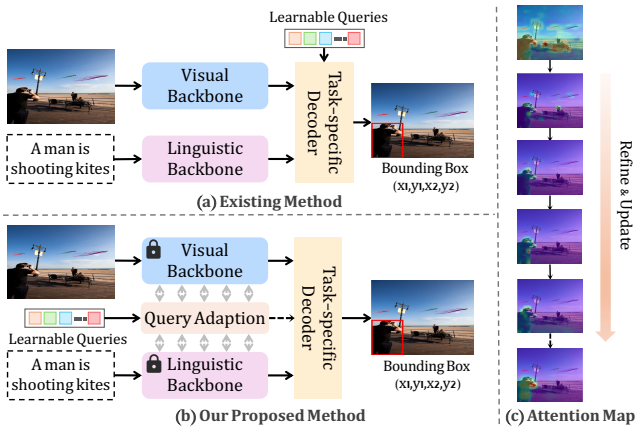


Figure 1: Comparison of DETR-like method and our proposed method for visual grounding. (a) The existing method typically adopts the random initialization queries directly into the decoder to predict the target object. (b) We introduce the query adaption module (QA) to learn target-related context progressively, providing valuable prior knowledge for the decoder. (c) The attention map of the last layer in every QA module and decoder (bottom), respectively.

Existing works in this field typically follow the object detection framework and incorporate multi-modal fusion to tackle this task. Earlier studies [52, 12, 27, 3, 58] mainly focus on a two-stage pipeline, which first generates a set of region proposals using object detectors, and then finds the best-matched region by interacting these regions with linguistic expressions. However, the performance of this method is limited by the quality of the generated region candidates. To address this issue, some studies [48, 47, 4] adopt the one-stage pipeline, which removes the proposal generation stage. Unfortunately, these methods make dense predictions with a sliding window over pre-defined anchor boxes, resulting in sub-optimal performance due to the failure to capture object relations effectively. Recently, some methods [17, 7, 10, 21, 8, 36] inspired by the DETR [1] structure, which adopt a standard multi-modal transformer framework to establish the multi-modal correspondence (as shown in Figure 1 (a)). These methods predict bounding boxes of target objects directly from learnable queries, eliminating the need for extra efforts to obtain candidates, such as region proposals or predefined anchor boxes.

While these methods have shown promising results, their primary focus remains on designing stronger multi-modal decoders. By contrast, much less work has been done to improve the learnable queries, which have been gained extensive attention in the object detection field. The queries that are inputted to the decoder in these methods are typically generated through random initialization or by utilizing linguistic embeddings. We argued that this vanilla approach has two critical issues: **i)** this target-agnostic query inevitably increases the learning difficulty of the decoder. **ii)** During the query learning process, these methods tend to focus solely on the deepest visual features of the backbone, overlooking the texture information that is crucial for the grounding task and present in low and mid-level features, as emphasized by [15, 38].

Drawing from these discussions, this paper seeks to address two critical research questions: **i)** *Can we produce the target-related referential queries for the decoder to alleviate the learning difficulty that the decoder faces?* and **ii)** *How can we effectively incorporate the multi-level visual context information into the query learning process?* We believe that tackling these issues together would promote the learnable query to more comprehensively and accurately learn the corresponding target object information in the image for the visual grounding task.

Considering CLIP [34] carries rich visual-language alignment knowledge, thus we adopt it as the backbone of our approach. Existing methods typically apply CLIP on the visual grounding by fine-tuning its parameters, as CLIP's training objective is to match entire images with text descriptions, rather than capturing fine-grained alignment between regions and textual elements. This may risk losing the general knowledge of CLIP and require significant computational resources. To tackle the challenges mentioned above, we propose a novel approach called RefFormer. Our approach incorporates a query adaptation (QA) module to generate referential queries, which provide the decoder with target-related context (as illustrated in Figure 1 (b)). By strategically inserting QA

module into different layers of CLIP, the query adaptively learns target-related information from multi-level image feature maps, and iteratively refines the acquired information layer by layer. Furthermore, our proposed Reformer can also act as an adapter, enabling CLIP to keep frozen and preserve the original rich knowledge. It adopts the bi-directional interaction scheme, performs the multi-modal fusion by incorporating a small number of trainable parameters, and residually injects new task-specific knowledge into CLIP throughout the entire feature extraction process. Extensive experiments conducted on five popular visual grounding benchmarks (i.e., RefCOCO/+/g [53, 31], Flickr30K [33], and ReferItGame [18]) demonstrate the superior performance of our proposed method.

Our contributions can be summarized as follows: (1) Unlike the previous methods that focus on designing sophisticated multi-modal decoders, we further improve the learning process of the learnable queries, a crucial aspect that has been overlooked in existing work. (2) We propose a query adaption module (QA), which can adaptively capture the target-related context, providing valuable referential knowledge for the decoder. (3) We conduct extensive experiments on five visual grounding benchmarks, demonstrating the effectiveness and potential of our method.

2 Related Work

2.1 Visual Grounding

Visual grounding aims to ground the target objects based on natural language descriptions by understanding the given images and expressions. Early work [52, 12, 27, 3, 58] primarily focuses on two-stage methods, which formulates the grounding task as a matching task. These methods employ object detectors to generate proposal candidates and then identify the best-matched candidate based on the matching score computed between each proposal and the referring expression. For example, MAttNet [52] proposes to decompose the language expression into three phrase embeddings, which are used to trigger three separate visual modules. While achieving successful performance, two-stage methods heavily rely on the quality of the generated proposals. Based on this, some studies [48, 47, 4] have been dedicated to one-stage methods to remove the proposal generation stage. These methods typically fuse visual features and language features first and then densely regress the bounding box on each position of the feature map grid. For instance, FAOA [48] incorporates linguistic embedding into the YOLOv3 detector to establish a one-stage pipeline, balancing between accuracy and speed.

Recently, transformer-based visual grounding methods [17, 7, 21, 23, 57, 40, 50, 8, 36, 10] have emerged, which leverages the self-attention mechanism to effectively capture intra- and inter-modality relationships and achieve improved performance. Among these methods, the mainstream approach [17, 7, 10, 21, 8, 36] adopts DETR-like structures to decode bounding boxes from learnable queries. For example, Transvg [7] and Transvg++ [8] employ a standard multimodal transformer framework, along with the REG token, to establish multi-modal correspondence and predict the coordinates of the referring object. Notably, the performance improvement of these methods primarily arises from the design of stronger backbones or multi-modal decoders. In this work, we focus on the design of learnable queries, which have received considerable attention in object detection field.

2.2 Learnable Queries in DETR and Its Variants

In the object detection field, DETR presents an end-to-end object detection model that is built in an encoder-decoder transformer architecture. However, it suffers from slow training convergence. To address this issue, some follow-up works [55, 19, 49, 45, 20, 26, 24, 54] solve this issue by optimizing the learnable queries in DETR. For instance, Anchor DETR [45] directly treats 2D reference points as queries, while DAB-DETR [24] further investigates the role of queries in DETR and proposes the use of 4D anchor boxes as queries. In contrast to these model-level improvements, DN-DETR [20] introduces query denoising training to mitigate the instability of bipartite graph matching, which is further enhanced by DINO [54].

Additionally, similar research have been explored in other tasks [22, 13, 37]. For example, EaTR [13] formulates a video as a set of event units and treats video-specific event units as dynamic moment queries in video grounding tasks. MTR++ [37] introduces distinct learnable intention queries generated by the k-means clustering algorithm to handle trajectory prediction across different motion modes in motion prediction tasks.

 https://doi.org/10.52202/079017-1502

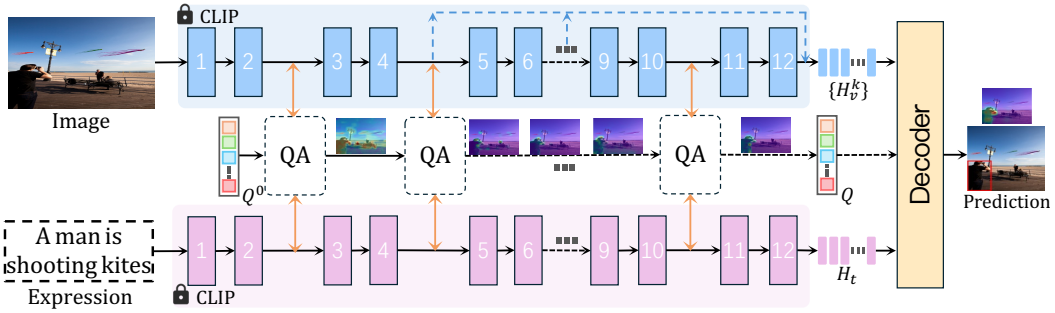


Figure 2: Overview of RefFormer. It adopts a DETR-like structure, consisting of a query adaptation (QA) module that seamlessly integrates into various layers of CLIP, along with a task-specific decoder. By incorporating the QA module, RefFormer can iteratively refine the target-related context and generate referential queries, which provide the decoder with prior context.

3 Preliminary

Considering the impressive vision-language alignment capability of CLIP, we take it as the backbone of our method to extract image and text representations, and keep the parameters frozen during training. The feature extraction process can be represented as follows:

Image Encoder. For an input image $V \in \mathbb{R}^{H\times W\times 3}$, it is divided into N non-overlapping patches of size $P \times P$, where $N_v = \frac{H\times W}{P^2}$. These patches are then flattened into a set of vectors, represented as $\{\mathbf{x}_v^i \in \mathbb{R}^{3P^2}\}_{i=1}^{N}$. Next, these vectors are transformed into token embeddings using a linear projection layer $\phi_e(\cdot)$. Furthermore, a classification token $\mathbf{x}_{cls} \in \mathbb{R}^D$ is added at the beginning of the token embeddings. Subsequently, the positional embeddings $\mathbf{E}_v$ are incorporated, and a layer normalization (LN) is applied. This process can be expressed as follows:

$$\mathbf{Z}_v^0 = LN([\mathbf{x}_{cls}; \phi_e(\mathbf{X}_v)] + \mathbf{E}^v) \tag{1}$$

where [;] denotes the concatenate operation. The sequence of tokens $\mathbf{Z}_v^0$ is then passed through L transformer layers. Each transformer layer comprises two submodules: the multi-head self-attention (MHSA) and the multilayer perceptron (MLP), with each submodule preceded by layer normalization.

$$\bar{\mathbf{Z}}_v^i = MHSA(LN(\mathbf{Z}_v^{i-1})) + \mathbf{Z}_v^{i-1},\ i = 1, ..., L \tag{2}$$

$$\mathbf{Z}_v^i = MLP(LN(\bar{\mathbf{Z}}_v^i)) + \bar{\mathbf{Z}}_v^i \tag{3}$$

where $\mathbf{Z}_v^i \in \mathbb{R}^{N\times D}$ denote the output of i-th transformer layer.

Text Encoder. Given an referring expression T, it is first transformed into a sequence of word embeddings using lower-cased byte pair encoding representations $\mathbf{X}_t$. The word embeddings are bracketed with the [SOS] and [EOS] tokens, producing a sequence of length N_t. Similar to the image encoder, these tokens are summed with positional embeddings $\mathbf{E}_t$ and passed through the L transformer layers to extract the text representations:

$$\bar{\mathbf{Z}}_t^i = MHSA(LN(\mathbf{Z}_t^{i-1})) + \mathbf{Z}_t^{i-1},\ i = 1, ..., L \tag{4}$$

$$\mathbf{Z}_t^i = MLP(LN(\bar{\mathbf{Z}}_t^i)) + \bar{\mathbf{Z}}_t^i \tag{5}$$

where $\mathbf{Z}_t^0 = [\mathbf{x}_{sos}; \mathbf{X}_t; \mathbf{x}_{eos}] + \mathbf{E}_t$, representing the word embedding layer in text encoder.

4 Method

The framework is shown in Figure 2. In the following, we first describe our query adaptation module in Section 4.1. We then introduce our decoder that decodes with referential query and training objectives in Section 4.2 and Section 4.3. Furthermore, we extend RefFormer to dense grounding task in Section 4.4. Finally, we provide a discussion in Section 4.5.

https://doi.org/10.52202/079017-1502

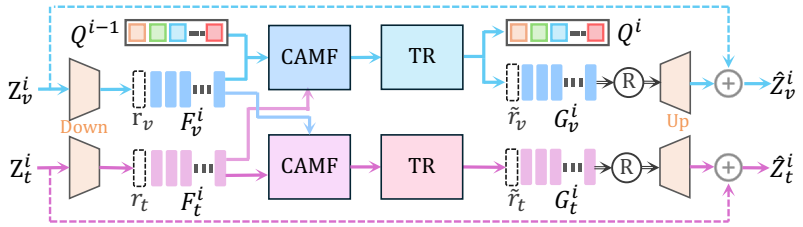


Figure 3: Illustration of our proposed Query Adaption Module, which mainly consists of CAMF and TR modules to generate the referential queries and promote the multi-modal features interaction. "R" represents the feature modulation.

4.1 Query Adaptation Module (QA)

In this section, we propose a QA module (as shown in Figure 3) that can generate the referential query to provide the decoder with the target-related context, thereby enhancing the decoder's grounding capabilities. *Importantly, our approach incorporates multi-level features into the query learning process, enabling the queries to capture more comprehensive target object information and can be refined layer by layer.* Furthermore, *QA can also act as an adapter*, eliminating the need to fine-tune the entire parameters of the backbone.

Down-projection. Considering the image and language representations $\mathbf{Z}_v^i$ and $\mathbf{Z}_t^i$ obtained from the i-th layer of the backbone, we initially use the MLP layers $\phi_{vd}^i(\cdot)$ and $\phi_{td}^i(\cdot)$ to project them to lower-dimensional features to reduce the computation memory:

$$\mathbf{F}_v^i = \phi_{vd}^i(\mathbf{Z}_v^i),\ \mathbf{F}_t^i = \phi_{td}^i(\mathbf{Z}_t^i) \tag{6}$$

Condition Aggregation and Multi-modal Fusion (CAMF). We randomly initialize N_q learnable queries $\mathbf{Q} \in \mathbb{R}^{N_q \times D_l}$, where D_l denotes the dimension after projected. These queries are specifically designed to capture potential target object context. Next, we concatenate these queries with the image features and input them, along with the language features into the CAMF block. Specifically, the CAMF block mainly consists of a cross-attention layer, which takes the image and query features $[\mathbf{Q}; \mathbf{F}_v]$ and language features $\mathbf{F}_t$ as the query respectively. This approach enables us to not only incorporate the expression condition into the learnable queries $\mathbf{Q}$ but also to extract relevant information from other modalities, thereby facilitating the fusion of target-related cross-modal features. Besides, we incorporate two learnable regulation tokens $\mathbf{r}_v, \mathbf{r}_t \in \mathbb{R}^{D_l}$ to modulate the final output of each QA. This process can be formalized as follows:

$$\bar{\mathbf{r}}_v, \bar{\mathbf{Q}}_c^i, \bar{\mathbf{F}}_v^i = MHCA([\mathbf{r}_v; \mathbf{Q}^{i-1}; \mathbf{F}_v^i], \mathbf{F}_t^i, \mathbf{F}_t^i) \tag{7}$$

$$\hat{\mathbf{Q}}_c^i = LN(\bar{\mathbf{Q}}_c^i) + \mathbf{Q}^{i-1},\ \hat{\mathbf{F}}_v^i = LN(\bar{\mathbf{F}}_v^i) + \mathbf{F}_v^i \tag{8}$$

$$\bar{\mathbf{r}}_t, \bar{\mathbf{F}}_t^i = MHCA([\mathbf{r}_t; \mathbf{F}_t^i], \mathbf{F}_v^i, \mathbf{F}_v^i),\ \hat{\mathbf{F}}_t^i = LN(\bar{\mathbf{F}}_t^i) + \mathbf{F}_t^i \tag{9}$$

where $\mathbf{Q}^{i-1}$ represents learnable queries that output from the previous QA, while $\mathbf{Q}^0$ are randomly initialized. The symbol $[;]$ indicates the concatenate operation, and $MHCA(,,)$ and $LN(\cdot)$ denote the multi-head cross-attention layers and layer normalization, respectively.

Target-related Context Refinement (TR). Following this, we feed the queries $\hat{\mathbf{Q}}_c$ and multi-modal ehanced feature maps $\hat{\mathbf{F}}_v^i$ and $\hat{\mathbf{F}}_t^i$ into the TR block. First, we use the queries $\hat{\mathbf{Q}}_c$ that have aggregated conditions to interact with the multi-modal enhanced image feature maps $\hat{\mathbf{F}}_v^i$, refining the target-related visual context within them.

$$\mathbf{Q}_v^i = MHCA(\hat{\mathbf{Q}}_c^i, \hat{\mathbf{F}}_v^i, \hat{\mathbf{F}}_v^i),\ \mathbf{Q}^i = LN(MLP(\mathbf{Q}_v^i)) + \hat{\mathbf{Q}}_c^i \tag{10}$$

Moreover, for feature maps $\hat{\mathbf{F}}_v^i$ and $\hat{\mathbf{F}}_t^i$ that have aggreaged other modality information, we use the self-attention to further enhance their target-related contextual semantics:

$$\widetilde{\mathbf{r}}_v, \widetilde{\mathbf{F}}_v^i = MHSA([\bar{\mathbf{r}}_v; \hat{\mathbf{F}}_v^i], \hat{\mathbf{F}}_v^i, \hat{\mathbf{F}}_v^i),\ \mathbf{G}_v^i = LN(MLP(\widetilde{\mathbf{F}}_v^i)) + \hat{\mathbf{F}}_v^i \tag{11}$$

$$\widetilde{\mathbf{r}}_t, \widetilde{\mathbf{F}}_t^i = MHSA([\bar{\mathbf{r}}_v; \hat{\mathbf{F}}_t^i], \hat{\mathbf{F}}_t^i, \hat{\mathbf{F}}_t^i),\ \mathbf{G}_t^i = LN(MLP(\widetilde{\mathbf{F}}_t^i)) + \hat{\mathbf{F}}_t^i \tag{12}$$

Up-projection. Finally, we utilize MLP to restore the channel dimension of the image and language features back to their original sizes. These features are then passed as inputs to the next layer of the

https://doi.org/10.52202/079017-1502

backbone in a residual manner. Prior to this, we utilize the regulation token to modulate the features $\mathbf{G}_v$ and $\mathbf{G}_t$, which helps prevent the multi-modal signal from overpowering the original signal.

$$\hat{\mathbf{Z}}_v^i = \phi_{vu}^i(\mathbf{G}_v^i \times \sigma(\widetilde{\mathbf{r}}_v)) + \mathbf{Z}_v^i,\ \hat{\mathbf{Z}}_t^i = \phi_{tu}^i(\mathbf{G}_t^i \times \sigma(\widetilde{\mathbf{r}}_t)) + \mathbf{Z}_t^i \tag{13}$$

where $\phi_{vu}(\cdot)$ and $\phi_{tu}(\cdot)$ denote the MLP layer, and $\sigma(\cdot)$ denotes the sigmoid function.

Finally, by iteratively performing the above process, the queries Q can progressively focus on the target-related context, and generate the referential queries to provide the prior context for the decoder.

4.2 Decoding with Referential Query.

Language-guided Multi-level Fusion. By inserting the QA at different layers of CLIP, the referential queries can be adaptively updated using the multi-level image feature maps. Additionally, to enhance the image features in decoder, we aggregate the multi-level visual features under the language guidance to yield language-aware multi-level image features. Specifically, given a multi-level image feature set $\{\hat{\mathbf{Z}}_v^k\}$ (including low, mid, and high levels), where $k \in \mathcal{K}$ represents selected layer index, we inject the language features $\mathbf{Z}_t^{last}$ (the final output of the text encoder) into each level of image features using MHCA:

$$\mathbf{H}_{t_{sos}} = \phi_{mt}(\mathbf{Z}_t^{last}),\ \mathbf{H}_v^k = \phi_{mv}^k(\hat{\mathbf{Z}}_v^k) \tag{14}$$

$$\hat{\mathbf{H}}_v^k = MHCA(\mathbf{H}_v^k, \mathbf{H}_{t_{sos}}, \mathbf{H}_{t_{sos}}) + \mathbf{H}_v^k,\ k \in \mathcal{K} \tag{15}$$

where $\phi_{mt}(\cdot)$ and $\phi_{mv}(\cdot)$ denote the linear project function used to map features to the same dimension. Besides, $\mathbf{H}_{t_{sos}}$ represents the [SOS] token in $\mathbf{H}_t$, which extracts the global information of the text. Subsequently, the multi-level language-aware image features are produced by simple concatenation, followed by a linear projection function $\phi_{vml}(\cdot)$ to map to the original dimension:

$$\bar{\mathbf{H}}_{vml} = Concat(\{\hat{\mathbf{H}}_v^k\}), k \in \mathcal{K} \tag{16}$$

$$\mathbf{H}_{vml} = \phi_{vml}(\bar{\mathbf{H}}_{vml}) \tag{17}$$

Decoding. Following, we first initialize the queries $\mathbf{Q}'$ with the same size as the referential query $\mathbf{Q}$, and add them together to utilize the prior context in $\mathbf{Q}$. Note that, to avoid interference from $\mathbf{Q}'$ during the initial stage, we initialize $\mathbf{Q}'$ as an all-zero matrix. Then, we concatenate the queries with the image features to interact with the language features to aggregate the condition information and produce the multi-modal feature map H_{mm}. This can be represented as:

$$\bar{\mathbf{O}}_c, \bar{\mathbf{H}}_{mm} = MHCA([\phi_q(\mathbf{Q}) + \mathbf{Q}'; \mathbf{H}_{vml}], \mathbf{H}_t, \mathbf{H}_t) \tag{18}$$

$$\mathbf{O}_c = LN(\bar{\mathbf{O}}_c) + \bar{\mathbf{O}}_c,\ \mathbf{H}_{mm} = LN(\bar{\mathbf{H}}_{mm}) + \bar{\mathbf{H}}_{mm} \tag{19}$$

where $\phi_q(\cdot)$ is the MLP layer, which regulates the significance of the query $\mathbf{Q}$. As the importance approaches zero, the query degenerate into a vanilla query. Then, we feed the queries $\mathbf{O_c}$ and multi-modal feature map $\mathbf{H}_{mm}$ into the MHCA layer to extract target embddings $\mathbf{O} \in \mathbb{R}^{N_q \times D}$. It can be formulated as:

$$\bar{\mathbf{O}} = MHCA(\mathbf{O}_c, \mathbf{H}_{mm}, \mathbf{H}_{mm}) \tag{20}$$

$$\mathbf{O} = LN(\phi_r(\bar{\mathbf{O}})) + \bar{\mathbf{O}} \tag{21}$$

where $\phi_r(\cdot)$ represents the linear projection function.

Grounding Head. We built the two MLPs ($\phi_{box}(\cdot)$ and $\phi_{cls}(\cdot)$) over the target embeddings $\mathbf{O}$. The final outputs consist of the predicted center coordinates of the target object, denoted as $b = (x, y, h, w) \in \mathbb{R}^4$, and the predicted confidence score $y \in \mathbb{R}^2$ that encompass the target object:

$$b = \phi_{box}(\mathbf{O}),\ y = \phi_{cls}(\mathbf{O}) \tag{22}$$

4.3 Training Objectives

Similar to DETR, we employ bipartite matching to find the best match between the predictions $\{b, y\}$ and the ground-truth targets $\{b_{tgt}, y_{tgt}\}$. In our case, the class prediction is confidence prediction aims to estimate the confidence of a query containing a target object. To supervise the training, we use the box prediction losses (L1 and GIoU), and a cross-entropy loss after matching.

$$\mathcal{L}_{det} = \lambda_{iou}\mathcal{L}_{iou}(b_{gt}, b) + \lambda_{L1}||b_{gt} - b|| + \lambda_{ce}\mathcal{L}_{ce}(y_{gt}, y) \tag{23}$$

where λ denotes the corresponding loss weight. Additionally, to encourage the referential queries in every QA module to effectively focus on the target-related context, we also introduce the auxiliary loss $\mathcal{L}_{aux}$ that is similar to the above objective function to provide supervision for them. The final training objective can be defined as:

$$\mathcal{L}_{final} = \mathcal{L}_{det} + \lambda_{aux}\mathcal{L}_{aux} \tag{24}$$

where λ_{aux} denotes the weight of the auxiliary loss.

4.4 Extend to Dense Grounding

In addition to object-level grounding, our method can easily extend to the dense grounding task by incorporating a segmentation head. Specifically, similar to the MaskFormer [5], we utilize the MLP to transform the target embeddings $\mathbf{O}$ into mask embeddings $\mathbf{M} \in \mathbb{R}^{N_q \times D}$. The binary mask prediction $s_i = [0, 1] \in \mathbb{R}^{H \times W}$ is then computed by performing a dot product between the mask embeddings $\mathbf{M}$ and the multi-modal feature map $\mathbf{H}_{mm}$ and followed by a sigmoid activation. During training, we use the mask prediction losses (Focal and Dice), which can be defined as follows:

$$\mathcal{L}_{seg} = \lambda_{focal}\mathcal{L}_{focal}(s_{gt}, s) + \lambda_{dice}\mathcal{L}_{dice}(s_{gt}, s) \tag{25}$$

where s_{gt} denotes the ground-truth mask.

4.5 Discussion

In this work, we aim to explore how to further optimize the learning process of queries. To reduce the learning difficulties posed by vanilla query, we introduce a simple query adaption module to adaptively capture target-related context and iteratively refine it. As illustrated in Figure 5, the attention maps produced by each query adaption module consistently align with our objective: to progressively focus on the target-related context and provide prior context for the decoder. It is worth noting that while "multi-level", "adapter", and "self-attention" may be extensively applied in other research fields, our approach aims to integrate them to address the challenges in visual grounding tasks, instead of designing a specific module to achieve the mentioned functions individually.

5 Experiment

5.1 Datasets and Evaluation Metric

RefCOCO/RefCOCO+/RefCOCOg. RefCOCO [53] comprises 19,994 images featuring 50,000 referred objects, divided into train, val, testA, and testB sets. Similarly, RefCOCO+ [53] contains 19,992 images with 49,856 referred objects and 141,564 referring expressions. It contains more attributes than absolute locations compared to RefCOCO, and has the same split. RefCOCOg [31] has 25,799 images with 49,856 referred objects and expressions. Following a common version of split [32], i.e., train, val, and test sets.

Flickr30K. Flickr30k Entities [33] contains 31,783 images and 158k caption sentences with 427k annotated phrases. We follow [7] to split the images into 29,783 for training, 1000 for validation, and 1000 for testing, and report the performance on the test set.

ReferItGame. ReferItGame [18] includes 20,000 images with 120,072 referring expressions for 19,987 referred objects. We follow [7] to split the dataset into train, validation and test sets, and report the performance on the test set.

Evaluation Metric. For referring expression comprehension (REC), we use Prec@0.5 evaluation protocol to evaluate the accuracy, which is consistent with prior works. In this evaluation, a predicted region is considered correct if its intersection-over-union (IoU) with the ground-truth bounding box is greater than 0.5. For referring expression segmentation (RES), we report the Mean IoU (MIoU) between the predicted segmentation mask and ground truth mask.

5.2 Implementation Details

Following [7, 36], the resolution of the input image is resized to 640×640. We employ the pre-trained CLIP as our backbone to extract both image and language features, and we freeze its parameters

 https://doi.org/10.52202/079017-1502

Table 1: Comparisons with the state-of-the-art approaches on three benchmarks, i.e., RefCOCO, RefCOCO+, RefCOCOg. * indicates models that use additionally data beyond RefCOCO series. † indicates that models simply combine RefCOCO, RefCOCO+, and RefCOCOg.

Methods	Visual Backbone	RefCOCO			RefCOCO+			RefCOCOg	
		val	testA	testB	val	testA	testB	val	test
Single Dataset:									
Two-stage:									
MAttNet CVPR2018 [52]	ResNet101	76.65	81.14	69.99	69.99	71.62	56.02	66.58	67.27
CM-A-E CVPR2019 [27]	ResNet101	78.35	83.14	71.32	68.09	73.65	58.03	67.99	68.67
Ref-NMS AAAI2021 [3]	ResNet101	80.70	84.00	76.04	68.25	73.68	59.42	70.55	70.62
PBREC AAAI2024 [56]	ResNet101	82.20	85.26	79.21	68.25	72.63	78.96	73.92	73.18
One-stage:									
FAOA ICCV2019 [48]	DarkNet53	72.54	74.53	68.50	56.81	60.23	49.60	61.33	60.36
ReSC-Large ECCV2020 [47]	DarkNet53	77.63	80.45	72.30	63.59	68.36	56.81	67.30	67.20
MCN CVPR2020 [28]	DarkNet53	80.08	82.29	74.98	67.16	72.86	57.31	66.46	66.01
PLV-FPN TIP2022 [23]	ResNet101	81.93	84.99	76.25	71.20	77.40	61.08	70.45	71.08
Transformer-based:									
TransVG ICCV2021 [7]	ResNet101	81.02	82.72	78.35	64.82	70.70	56.94	68.67	67.73
RefTR NIPS2021 [21]	ResNet101	82.23	85.59	76.57	71.58	75.96	62.16	69.41	69.40
SeqTR ECCV2022 [57]	DarkNet53	81.23	85.59	76.08	68.82	75.37	58.78	71.35	71.58
QRNeT CVPR2022 [50]	Swin-small	84.01	85.85	82.34	72.94	76.17	63.81	71.89	73.03
LADS AAAI2023 [39]	ResNet50	82.85	86.67	78.57	71.16	77.64	59.82	71.56	71.66
TransVG++ TPAMI2023 [40]	ViT-Base/16	86.28	88.37	80.97	75.39	80.45	66.28	76.18	76.30
Dynamic MDETR TPAMI2023 [40]	ViT-Base/16	85.97	88.82	80.12	74.83	81.70	63.44	74.14	74.49
VG-LAW CVPR2023 [40]	ViT-Base/16	86.06	88.56	82.87	75.74	80.32	66.69	75.31	75.95
PVD AAAI2024 [6]	Swin-Base	84.52	86.19	76.81	73.89	78.41	64.25	74.13	71.51
Ours	ViT-Base/32	83.97	87.80	77.45	73.55	81.09	62.24	76.33	75.33
Ours	ViT-Base/16	**86.52**	**90.24**	81.42	**76.58**	**83.69**	**67.38**	**77.80**	**77.60**
Multiple/Extra Datasets:									
VILLA_L * NIPS2020 [11]	ResNet101	82.39	87.48	74.84	76.17	81.54	66.84	76.18	76.71
RefTR * NIPS2021 [21]	ResNet101	85.43	87.48	79.86	76.40	81.35	66.59	78.43	77.86
MDETR * ICCV2021 [17]	ResNet101	86.75	89.58	81.41	79.52	84.09	70.62	81.64	80.89
ShiKra-7B * ARXIV2023 [2]	ViT-Large	87.01	90.61	80.24	81.60	87.36	72.12	82.27	82.19
Ferret-7B * ARXIV2023 [51]	ViT-Large	87.49	91.35	82.45	80.78	87.38	73.14	83.93	84.76
APE †CVPR2024 [35]	ViT-Large	85.50	89.10	81.30	73.40	80.70	64.40	83.00	78.00
Pink * CVPR2024 [46]	ViT-Large	88.30	91.70	84.00	81.80	88.20	73.90	83.90	84.30
Ours †	ViT-Base/16	88.82	92.52	84.87	80.91	86.64	73.35	82.29	83.15
Ours †	ViT-Large	**90.91**	**93.69**	**86.56**	**83.33**	**89.00**	**75.78**	**84.97**	**84.88**

Table 2: Comparison with state-of-the-art approaches on the Flickr30K Entities and ReferItGame.

Methods	Flickr30K	ReferItGame
	test	test
RefTR [21]	78.66	71.42
TransVG [7]	79.10	70.73
QRNet [50]	81.95	74.61
TransVG++ [8]	81.49	74.70
Dynamic MDETR [36]	81.89	70.37
VG-LAW [40]	-	76.60
Ours	**82.10**	**82.18**

Table 3: Comparisons with the state-of-the-art dense grounding approaches on three benchmarks for RES task, i.e., RefCOCO, RefCOCO+, and RefCOCOg.

Methods	RefCOCO			RefCOCO+			RefCOCOg	
	val	testA	testB	val	testA	testB	val	test
MAttNet [28]	56.51	62.37	51.70	46.67	52.39	40.08	47.64	48.61
MCN [28]	62.44	64.20	59.71	50.62	54.99	44.69	49.22	49.40
LTS [16]	65.43	67.76	63.08	54.21	58.32	48.02	54.40	54.25
RefTR [21]	70.56	73.49	66.57	61.08	64.59	52.73	58.73	58.51
SeqTR [57]	67.26	69.79	64.12	54.14	58.93	48.19	55.67	55.64
VG-LAW [40]	75.05	77.36	71.69	66.61	70.30	58.14	65.36	65.13
PVD [6]	74.82	77.11	69.52	63.38	68.60	56.92	63.13	63.62
Ours	74.47	**77.92**	**72.30**	**66.70**	**72.28**	**60.43**	**65.51**	**66.26**

during training. The model is optimized end-to-end using AdamW for 40 epochs, with a batch size of 32. We set the learning rate to 1e-4 and the weight decay to 1e-2. The experiments are conducted on V100 GPUs. The loss weight $\lambda_{iou}, \lambda_{L1}, \lambda_{ce}$, and λ_{aux}, we set to 3.0, 1.0, 1.0, and 0.1. For dense grounding, we set the parameters λ_{focal}, and λ_{dice} to 5.0, and 1.0.

5.3 Comparisons with State-of-the-art Methods

REC Task. For REC task, we compare the performance with the state-of-the-art REC methods, including the two-stage methods, one-stage methods, and transformer-based methods. As reported in Table 1 and Table 2, our proposed method achieves the best performance. In particular, when comparing to the transformer-based method Dynamic MDETR, which adopts the DETR-like structure and uses the same backbone as ours, we can see that our method performs better with $+0.53\%, +1.90\%, +1.62\%$ on RefCOCO, $+1.11\%, +2.44\%, +6.21\%$ on RefCOCO+, and $+4.94\%, +4.18\%$ on RefCOCOg. Additionally, under multiple/extra datasets setting, our method also surpasses recent state-of-the-art methods that incorporate large language models or utilize more training data.

RES Task. Following RefTR [21] and VG-LAW [40], we also conduct the dense grounding experiments and report the results in Table 3 in terms of mIoU. It can be seen that our model

Table 4: Ablation study of the generation method of learnable queries on RefCOCOg.

Fusion layer ($\mathcal{K}$)	val	test
{4}	65.42 (-9.31)	65.12 (-9.02)
{8}	73.31 (-1.42)	72.47 (-1.69)
{12}	74.73 (-1.59)	74.16 (-1.17)
{4,8}	72.71 (-2.02)	73.03 (-1.13)
{4,12}	75.33 (+0.60)	74.51 (+0.35)
{8,12}	75.61 (+0.88)	74.82 (+0.66)
{4,8,12}	**76.33 (+1.60)**	**75.33 (+1.17)**

Table 5: Ablation study of the QA position on RefCOCOg.

RefFormer layer	val	test
None	65.50	65.54
{4,8,12}	74.08 (+8.58)	73.82 (+8.28)
{4,6,8,10,12}	**76.33 (+10.83)**	**75.33 (+9.79)**
{2,4,6,8,10,12}	75.84 (+10.34)	75.32 (+9.78)

achieves superior performance without extra deliberate design to dense grounding, demonstrating the generalization of our method.

5.4 Ablation Studies

In this section, we conduct the ablation studies to verify the effectiveness the each part of our proposed method on RefCOCOg. Following previous work [7, 17, 8], the visual backbone we apply the ViT-Base/32.

Effect on the position of QA. As presented in Table 5, firstly, we can observe that removing the QA would lead to a Sharp decline in performance, highlighting the effectiveness of QA. We then explored the impact of QA's position in the CLIP to determine where QA should be added. We chose three groups for the ablation study: $\{4, 8, 12\}$, $\{4, 6, 8, 10, 12\}$, and $\{2, 4, 6, 8, 10, 12\}$. The results indicate that performance is best when we use the $\{4, 6, 8, 10, 12\}$ configuration. Therefore, we default to this position in our experiments.

Effect on the layers of multi-level fusion. In Table 4, we analyze the impact of the fusion layers in the decoder. We first conduct experiments using single-level image features, and then proceed with multi-level features. The results show that utilizing multi-level features significantly improves performance, demonstrating that low- and mid-level features complement high-level features. Additionally, using the $\{4, 8, 12\}$ achieves the best performance, which we adopt for our experiments.

Effect on the different backbone. In the first line of Table 6, we apply our method to single-modal encoders, i.e., Swin-Transformer + Bert. The results demonstrate that our method is not only applicable to multi-modal encoders but is also compatible with single-modal encoders.

Effect on the auxiliary loss. In the second line of Table 6, We experiment with auxiliary loss, and the results demonstrate the effectiveness of auxiliary loss. By employing auxiliary loss, the reference query can capture the target-related visual contexts more effectively.

Effect on learnable queries. In the third line of Table 6, we validate the effectiveness of the learnable queries. Specifically, we replace the prior queries generated by the QA module with randomly initialized queries or linguistic embeddings input to decoder while keeping other modules unchanged. We can observe that introducing the prior queries can bring significant performance improvement. This result demonstrates that prior queries aid the decoder in more accurately locating the target object. Additionally, we investigate the accuracy of our referential queries, which are designed to provide prior information to the decoder. Since the channel dimension in the QA module is lower, the reference query may not accurately predict the coordinates of the targets.

Convergence curves. In Figure 4 we illustrate the convergence curve of our proposed method compared to other open-source DETR-like visual grounding methods. Notably, our method demonstrates accelerated training convergence, reducing the training time by half compared to the TransVG, while also outperforming other existing methods.

5.5 Qualitative Results

As shown in Figure 5, the attention maps in the QA module illustrate the refined process of how the referential query captures the target-related context. Initially, the attention map appears noisy but gradually focuses on the target-related context, such as the couch in (a). By incorporating the referential query, the attention map in the decoder accurately concentrates on the target object.

https://doi.org/10.52202/079017-1502

Table 6: Ablation studies of backbone, auxiliary loss, and learnable queries on RefCOCOg.

Method	val	test
Backbone:		
Swin+Bert	75.25 (-0.64)	**75.61** (+0.29)
Auxiliary loss:		
W/o $\mathcal{L}_{aux}$	74.24 (-1.60)	73.82 (-1.50)
Learnable queries:		
Referential query	52.92 (-22.92)	51.87 (-23.45)
Linguistic embeddings	71.36 (-4.48)	71.07 (-4.25)
Random initialization	73.40 (-2.44)	73.12 (-2.21)
Ours	**75.84**	75.32

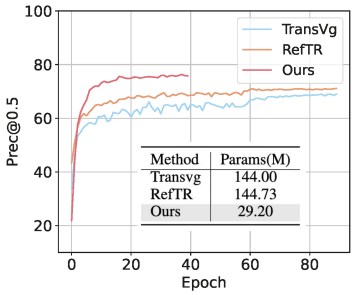


Figure 4: Convergence curves. Our method achieves better results with fewer training epochs on RefCOCOg.

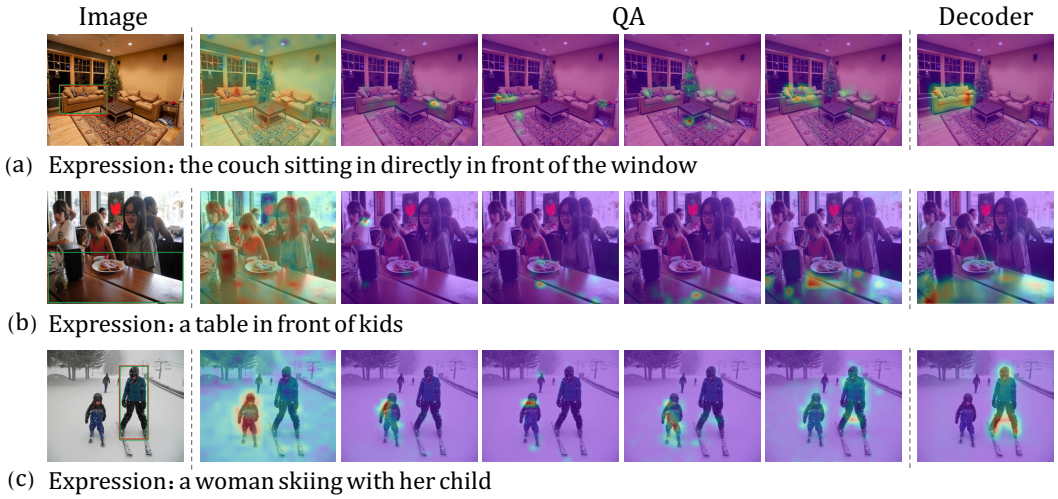


Figure 5: Qualitative results on RefCOCOg. The bounding boxes in green and red correspond to predictions of our model and the ground truth. Columns 2-6 showcase the attention maps generated by each QA module, while the last column represents the attention map from the decoder.

Besides, it is important to note that the referential query may not precisely focus on the target object due to the lower feature dimension in the QA module, but it still captures target-related information.

6 Concluding and Remarks

In this paper, we propose a novel approach, called RefFormer that can be seamlessly integrated into CLIP. The RefFormer can not only generate the referential query to provide the target-related context for decoder, but also act as the adaptor to preserve the original knowledge of CLIP and reduce the training cost. Extensive experiments demonstrate the effectiveness of our method, and visualization results illustrate the refined process of our proposed RefFormer.

Limitations : Although our method is specifically designed for the REC task and surpasses existing SOTA methods in REC, there is still significant room for improvement in the RES task. This is because we have not yet optimized our approach specifically for the RES task.

7 Acknowledgments

This work was supported in part by National Science and Technology Major Project under Grant 2023ZD0121300, National Natural Science Foundation of China under Grants 62088102, 12326608 and 62106192, Natural Science Foundation of Shaanxi Province under Grant 2022JC-41, and Fundamental Research Funds for the Central Universities under Grant XTR042021005.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[2] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

[3] Long Chen, Wenbo Ma, Jun Xiao, Hanwang Zhang, and Shih-Fu Chang. Ref-nms: Breaking proposal bottlenecks in two-stage referring expression grounding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 1036–1044, 2021.

[4] Xinpeng Chen, Lin Ma, Jingyuan Chen, Zequn Jie, Wei Liu, and Jiebo Luo. Real-time referring expression comprehension by single-stage grounding network. *arXiv preprint arXiv:1812.03426*, 2018.

[5] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems*, 34:17864–17875, 2021.

[6] Zesen Cheng, Kehan Li, Peng Jin, Siheng Li, Xiangyang Ji, Li Yuan, Chang Liu, and Jie Chen. Parallel vertex diffusion for unified visual grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1326–1334, 2024.

[7] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1769–1779, 2021.

[8] Jiajun Deng, Zhengyuan Yang, Daqing Liu, Tianlang Chen, Wengang Zhou, Yanyong Zhang, Houqiang Li, and Wanli Ouyang. Transvg++: End-to-end visual grounding with language conditioned vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 2023.

[9] Jianfeng Dong, Yabing Wang, Xianke Chen, Xiaoye Qu, Xirong Li, Yuan He, and Xun Wang. Reading-strategy inspired visual representation learning for text-to-video retrieval. *IEEE transactions on circuits and systems for video technology*, 32(8):5680–5694, 2022.

[10] Ye Du, Zehua Fu, Qingjie Liu, and Yunhong Wang. Visual grounding with transformers. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, Jul 2022.

[11] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. *Advances in Neural Information Processing Systems*, 33:6616–6628, 2020.

[12] Richang Hong, Daqing Liu, Xiaoyu Mo, Xiangnan He, and Hanwang Zhang. Learning to compose and reason with language tree structures for visual grounding. *IEEE transactions on pattern analysis and machine intelligence*, 44(2):684–696, 2019.

[13] Jinhyun Jang, Jungin Park, Jin Kim, Hyeongjun Kwon, and Kwanghoon Sohn. Knowing where to focus: Event-aware transformer for video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13846–13856, 2023.

[14] Jiayi Ji, Yiwei Ma, Xiaoshuai Sun, Yiyi Zhou, Yongjian Wu, and Rongrong Ji. Knowing what to learn: a metric-oriented focal mechanism for image captioning. *IEEE Transactions on Image Processing*, 31:4321–4335, 2022.

[15] Dongsheng Jiang, Yuchen Liu, Songlin Liu, Xiaopeng Zhang, Jin Li, Hongkai Xiong, and Qi Tian. From clip to dino: Visual encoders shout in multi-modal large language models. 2023.

[16] Ya Jing, Tao Kong, Wei Wang, Liang Wang, Lei Li, and Tieniu Tan. Locate then segment: A strong pipeline for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9858–9867, 2021.

 https://doi.org/10.52202/079017-1502

[17] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021.

[18] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014.

[19] Feng Li, Ailing Zeng, Shilong Liu, Hao Zhang, Hongyang Li, Lei Zhang, and Lionel M Ni. Lite detr: An interleaved multi-scale encoder for efficient detr. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18558–18567, 2023.

[20] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13619–13627, 2022.

[21] Muchen Li and Leonid Sigal. Referring transformer: A one-step approach to multi-task visual grounding. *Advances in neural information processing systems*, 34:19652–19664, 2021.

[22] Pandeng Li, Chen-Wei Xie, Hongtao Xie, Liming Zhao, Lei Zhang, Yun Zheng, Deli Zhao, and Yongdong Zhang. Momentdiff: Generative video moment retrieval from random to real. *Advances in neural information processing systems*, 36, 2024.

[23] Yue Liao, Aixi Zhang, Zhiyuan Chen, Tianrui Hui, and Si Liu. Progressive language-customized visual feature learning for one-stage visual grounding. *IEEE Transactions on Image Processing*, 31:4266–4277, 2022.

[24] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv preprint arXiv:2201.12329*, 2022.

[25] Sun-Ao Liu, Yiheng Zhang, Zhaofan Qiu, Hongtao Xie, Yongdong Zhang, and Ting Yao. Caris: Context-aware referring image segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 779–788, 2023.

[26] Wenze Liu, Hao Lu, Yuliang Liu, and Zhiguo Cao. Box-detr: Understanding and boxing conditional spatial queries. *arXiv preprint arXiv:2307.08353*, 2023.

[27] Xihui Liu, Zihao Wang, Jing Shao, Xiaogang Wang, and Hongsheng Li. Improving referring expression grounding with cross-modal attention-guided erasing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1950–1959, 2019.

[28] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 10034–10043, 2020.

[29] Yiwei Ma, Jiayi Ji, Xiaoshuai Sun, Yiyi Zhou, and Rongrong Ji. Towards local visual modeling for image captioning. *Pattern Recognition*, 138:109420, 2023.

[30] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 638–647, 2022.

[31] 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.

[32] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 792–807. Springer, 2016.

[33] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

[34] 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. PMLR, 2021.

[35] Yunhang Shen, Chaoyou Fu, Peixian Chen, Mengdan Zhang, Ke Li, Xing Sun, Yunsheng Wu, Shaohui Lin, and Rongrong Ji. Aligning and prompting everything all at once for universal visual perception. 2024.

[36] Fengyuan Shi, Ruopeng Gao, Weilin Huang, and Limin Wang. Dynamic mdetr: A dynamic multimodal transformer decoder for visual grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[37] Shaoshuai Shi, Li Jiang, Dengxin Dai, and Bernt Schiele. Mtr++: Multi-agent motion prediction with symmetric scene modeling and guided intention querying. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[38] Mainak Singha, Ankit Jha, Bhupendra Solanki, Shirsha Bose, and Biplab Banerjee. Applenet: Visual attention parameterized prompt learning for few-shot remote sensing image generalization using clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[39] Wei Su, Peihan Miao, Huanzhang Dou, Yongjian Fu, and Xi Li. Referring expression comprehension using language adaptive inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2357–2365, 2023.

[40] Wei Su, Peihan Miao, Huanzhang Dou, Gaoang Wang, Liang Qiao, Zheyang Li, and Xi Li. Language adaptive weight generation for multi-task visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10857–10866, 2023.

[41] Yabing Wang, Jianfeng Dong, Tianxiang Liang, Minsong Zhang, Rui Cai, and Xun Wang. Cross-lingual cross-modal retrieval with noise-robust learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 422–433, 2022.

[42] Yabing Wang, Fan Wang, Jianfeng Dong, and Hao Luo. Cl2cm: Improving cross-lingual cross-modal retrieval via cross-lingual knowledge transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5651–5659, 2024.

[43] Yabing Wang, Le Wang, Qiang Zhou, Zhibin Wang, Hao Li, Gang Hua, and Wei Tang. Multimodal llm enhanced cross-lingual cross-modal retrieval. *arXiv preprint arXiv:2409.19961*, 2024.

[44] Yabing Wang, Shuhui Wang, Hao Luo, Jianfeng Dong, Fan Wang, Meng Han, Xun Wang, and Meng Wang. Dual-view curricular optimal transport for cross-lingual cross-modal retrieval. *IEEE Transactions on Image Processing*, 33:1522–1533, 2024.

[45] Yingming Wang, Xiangyu Zhang, Tong Yang, and Jian Sun. Anchor detr: Query design for transformer-based detector. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 2567–2575, 2022.

[46] Shiyu Xuan, Qingpei Guo, Ming Yang, and Shiliang Zhang. Pink: Unveiling the power of referential comprehension for multi-modal llms. 2024.

[47] Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. Improving one-stage visual grounding by recursive sub-query construction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 387–404. Springer, 2020.

 https://doi.org/10.52202/079017-1502

[48] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4683–4693, 2019.

[49] Zhuyu Yao, Jiangbo Ai, Boxun Li, and Chi Zhang. Efficient detr: improving end-to-end object detector with dense prior. *arXiv preprint arXiv:2104.01318*, 2021.

[50] Jiabo Ye, Junfeng Tian, Ming Yan, Xiaoshan Yang, Xuwu Wang, Ji Zhang, Liang He, and Xin Lin. Shifting more attention to visual backbone: Query-modulated refinement networks for end-to-end visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15502–15512, 2022.

[51] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023.

[52] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1307–1315, 2018.

[53] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016.

[54] 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.

[55] Chuyang Zhao, Yifan Sun, Wenhao Wang, Qiang Chen, Errui Ding, Yi Yang, and Jingdong Wang. Ms-detr: Efficient detr training with mixed supervision. *arXiv preprint arXiv:2401.03989*, 2024.

[56] Peizhi Zhao, Shiyi Zheng, Wenye Zhao, Dongsheng Xu, Pijian Li, Yi Cai, and Qingbao Huang. Rethinking two-stage referring expression comprehension: A novel grounding and segmentation method modulated by point. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7487–7495, 2024.

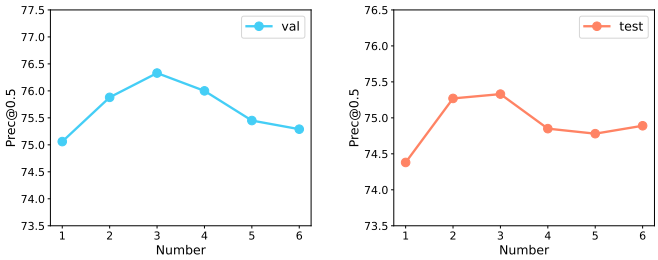
[57] Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, Xiaoshuai Sun, and Rongrong Ji. Seqtr: A simple yet universal network for visual grounding. In *European Conference on Computer Vision*, pages 598–615. Springer, 2022.

[58] Bohan Zhuang, Qi Wu, Chunhua Shen, Ian Reid, and Anton Van Den Hengel. Parallel attention: A unified framework for visual object discovery through dialogs and queries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4252–4261, 2018.

A Appendix

Table 7: Ablation study of the direction of the features flow from the QA module on RefCOCOg.

RefFormer direction	val	test
None	65.50	65.54
Only text	70.00 (+4.50)	69.24 (+3.70)
Only image	72.57 (+7.07)	72.56 (+7.02)
Image & Text	**76.33 (+10.83)**	**75.33 (+9.79)**



(a) val (b) test

Figure 6: The performance of different numbers of learnable queries on RefCOCOg.

A.1 Effect on the RefFormer's direction.

In RefFormer, the QA module can serve as an adapter, injecting specific knowledge into the frozen CLIP model. In Table 7, we investigate the direction of feature flow from QA module. We find that using a dual-direction approach achieves the best performance. Through QA module, language features have aggregated relevant visual context information. As pointed to [25], incorporating rich visual context into linguistic features aids in achieving strong vision-language alignment and better indicating target objects.

A.2 Effect on the number of learnable queries.

We depict the performance in terms of Prec@0.5 according to the number of learnable queries N_q in Figure 6. When we adopt the $N_q = 3$, the performance is best. However, further increases yield only slight improvements in metrics, as a large number of N_q increases the difficulty of the model. Therefore, we default set the $N_q = 3$ in our experiments.

A.3 Visualization

Due to space limitations, we present additional visualization results here. As shown in Figure 7, the referential queries gradually focus on the target object and effectively provide target-related context for the decoder. These results demonstrate the effectiveness of our proposed methods.

https://doi.org/10.52202/079017-1502

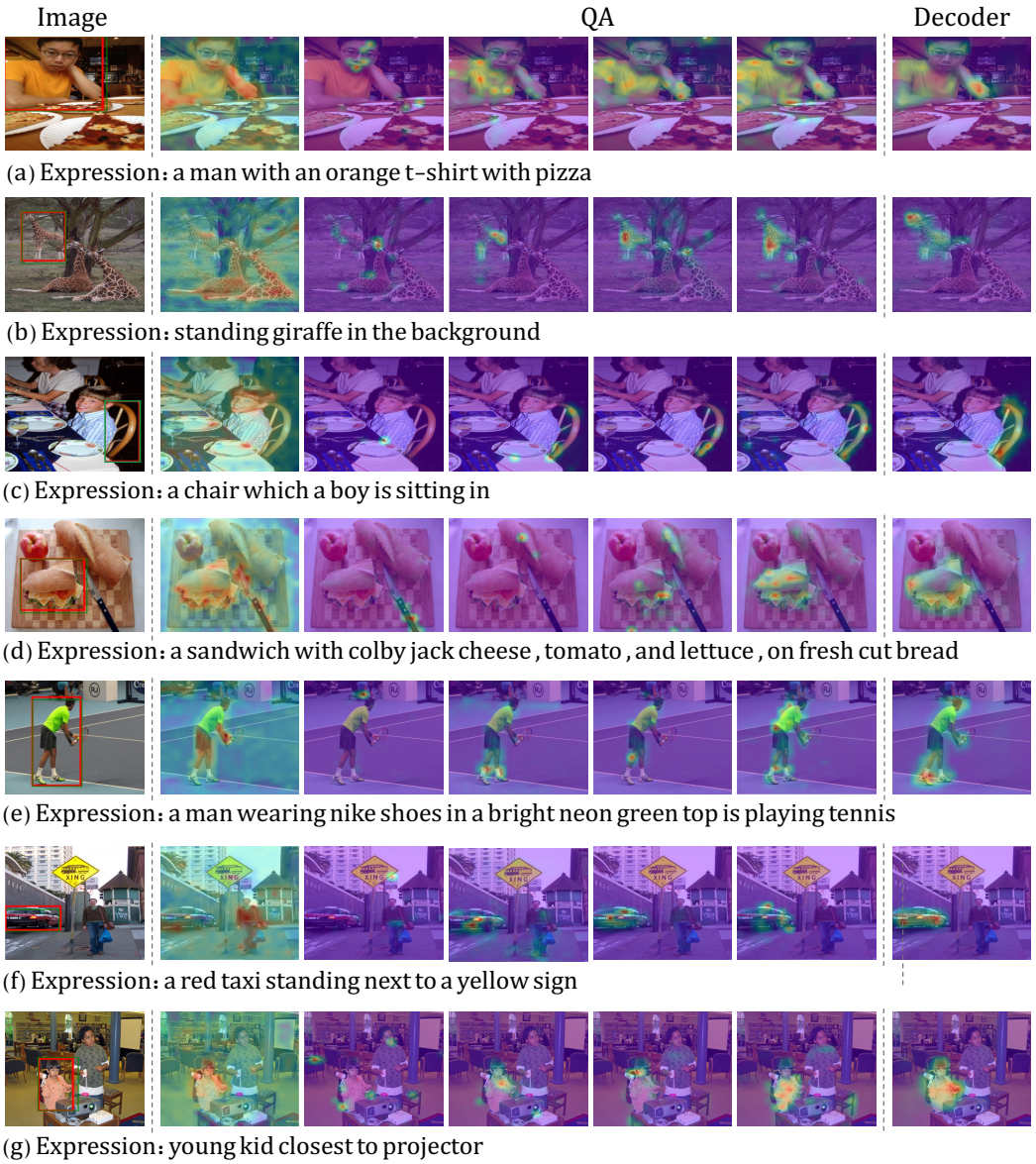


Figure 7: Qualitative results on RefCOCOg. The bounding boxes in green and red correspond to outputs from our model and the ground truth. Columns 2-6 showcase the attention maps generated by each QA module, while the last column represents the attention map from the decoder.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction (Section 1) accurately reflect our paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.

- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of the work in the paper.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: Our paper does not include theoretical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.

https://doi.org/10.52202/079017-1502

- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We fully describe the proposed model and implementation details in Section 4 and Section 5.2, and we submit our main code in the form of a zipped file in additional supplementary materials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We submit our main code in the form of a zipped file in additional supplementary materials, and we will release the complete code after review.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We describe the datasets, metrics and implementation details in Sec. **??** and supplemental material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Error bars are not reported because it would be too computationally expensive.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).

https://doi.org/10.52202/079017-1502

- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: We only provide the GPU type in the supplemental material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms with the NeurIPS Code of Ethics in every respect.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA] .

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA] .

    Justification: This paper does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA] .

    Justification: This paper does not release new assets.

https://doi.org/10.52202/079017-1502

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.