

Task-aware Cross-modal Feature Refinement Transformer with Large Language Models for Visual Grounding

Wenbo Chen^{1*}, Zhen Xu^{1,2*}, Ruotao Xu², Si Wu^{1,2†}, Hau-San Wong³

¹School of Computer Science and Engineering, South China University of Technology

²Institute for Super Robotics (Huangpu)

³Department of Computer Science, City University of Hong Kong

{cscwb106, csxuzhen}@mail.scut.edu.cn, {xrt, cswusi}@scut.edu.cn, cshswong@cityu.edu.hk

Abstract

The goal of visual grounding is to establish connections between target objects and textual descriptions. Large Language Models (LLMs) have demonstrated strong comprehension abilities across a variety of visual tasks. To establish precise associations between the text and the corresponding visual region, we propose a Task-aware Cross-modal feature Refinement Transformer with LLMs for visual grounding, and our model is referred to as TCRT. To enable the LLM trained solely on text to understand images, we introduce an LLM adaptation module that extracts text-related visual features to bridge the domain discrepancy between the textual and visual modalities. We feed the text and visual features into the LLM to obtain task-aware priors. To enable the priors to guide the feature fusion process, we further incorporate a cross-modal feature fusion module, which allows task-aware embeddings to refine visual features and facilitate information interaction between the Referring Expression Comprehension (REC) and Referring Expression Segmentation (RES) tasks. We have performed extensive experiments to verify the effectiveness of the main components and demonstrate the superior performance of the proposed TCRT over state-of-the-art end-to-end visual grounding methods on RefCOCO, RefCOCOg, RefCOCO+ and ReferItGame.

1. Introduction

Visual grounding aims to predict the locations of the objects specified by textual prompts in images, and is widely applied in the field of human-computer interaction. This task typically involves referring expression comprehension [10, 17, 49, 59] and referring expression segmentation [22, 38]. The main challenge is to comprehensively un-

*Joint first authors.

†Corresponding author.

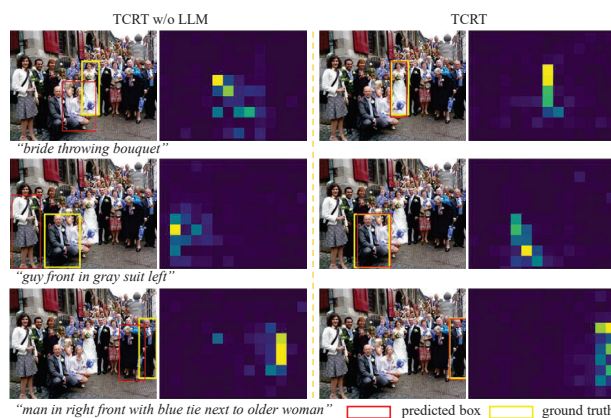


Figure 1. The visualization of the refined visual features from the cross-modal feature fusion module. In challenging environments, the introduction of LLMs can facilitate a better association between visual features and text.

derstand intra-modality information and accurately establish inter-modality correspondences.

Existing visual grounding methods can be categorized into two main types: two-stage and end-to-end methods. The two-stage methods [45] first generate region proposals [36] that may match the text, and then rank the candidate regions according to their relevance to the text. When the proposals generated in the first stage do not contain the target object, the second stage will be unable to correct these proposals. On the other hand, end-to-end grounding methods perform the integration of visual and textual features within an object detector, producing the box coordinates of the target object. The end-to-end methods are typically faster than the two-stage ones. With the advancement of transformer in the vision domain, TransVG [9, 10] and SeqTR [59] employ self-attention or cross-attention mechanisms to fuse textual and visual features. Recently, LLaVA [25] has integrated visual understanding capabilities into large language mod-

els (LLMs). MiniGPT-2 [4] and OMG-LLaVA [58] apply LLMs to visual grounding tasks by integrating textual and visual features. In this work, we make an attempt to adapt LLMs to capture the information of target objects and produce task-aware embeddings for precise grounding results.

More specifically, we propose a task-aware cross-modal feature refinement transformer with large language models for visual grounding, which is referred to as TCRT, leveraging priors from LLMs to guide the multimodal information fusion process and achieve precise alignment between textual and local visual features (see Figure 1). To enable the LLM to more effectively process image data and adapt to visual grounding tasks, we propose an LLM adaptation module that utilizes conditional normalization to adjust the statistical characteristics of visual features based on textual information, while introducing object-text semantic matching regularization to associate local visual features with textual features. Additionally, to utilize the priors provided by the adaptation module, we incorporate a task-aware cross-modal feature fusion module, which continuously refines visual features during multimodal fusion and enables the model to exploit the relationship between the REC and RES tasks. We perform extensive qualitative and quantitative experiments to validate the effectiveness of our model, and demonstrate its advantages on benchmark datasets. The main contributions of this work are summarized as follows:

- Different from the state-of-the-art methods that focus on the application of LLMs in integrating textual and visual features, we further explore the LLM’s capability of learning task-aware embeddings to facilitate visual grounding.
- We incorporate an LLM adaptation module for visual grounding, which learns to capture the information of target objects by modulating the statistical characteristics of visual features conditioned on textual features.
- We further perform task-aware cross-modal feature fusion, allowing the task-aware embeddings derived from LLMs to guide the association of textual and visual information and lead to precise grounding results.

2. Related Work

2.1. Two-stage Visual Grounding

In two-stage methods, a region proposal network [15, 16, 36] is first used to produce a set of candidate regions that may correspond to the target object mentioned in the textual description. These candidate regions are then matched to the textual description, and the predicted result is determined by identifying the region with the highest score. The maximum-margin ranking loss [30] is widely adopted for the second stage. Several researches further enhance the two-stage methods by incorporating context modeling [18, 32, 44, 55] and making use of phrase cooccur-

rences [2, 5, 11]. MAttNet [52] decomposed the expression into three modules: relationship, location, and subject to improve the grounding accuracy. To enhance the model’s capacity for learning detailed alignments between text and visuals, CM-A-E [27] proposed a method capable of generating challenging samples for training the model. Graph learning was incorporated into visual grounding by SGMN [46] to enable the model to better capture the relationships between different objects within the same scene. To increase the recall of target objects, Ref-NMS [6] utilized textual features to direct the non-maximum suppression. To enhance the model’s cross-modality fusion capabilities, GroundingDINO [26] integrated grounded pre-training techniques into detection transformers.

2.2. End-to-end Visual Grounding

End-to-end methods in visual grounding aim to directly predict the bounding boxes in an image from a text in a single pass, resulting in faster inference speeds compared to two-stage methods. As a pioneering work, FAOA [47] concatenated image and textual features, and fed the concatenated features into a prediction layer to achieve target object detection. To reduce the referring ambiguity, ReSC [48] used a sub-query construction to address limitations in processing complex textual descriptions. TransVG [9, 10] is one of the pioneering transformer-based visual grounding models that directly regresses coordinates. To achieve multi-task learning, RefTR [22] used the transformer architecture to integrate REC and RES tasks into a single framework. To address the gap between the visual and textual backbones, QRNet [49] incorporated textual information into the visual backbone. Furthermore, VG-LAW [38] actively extracted text-relevant visual features based on text-adaptive weights and performed multi-task visual grounding. Additionally, PVD [7] introduced a diffusion model into the visual grounding task.

With the development of LLMs, some Multimodal Large Language Models [25] (MLLMs) are gradually being applied to visual grounding tasks. VistaLLM [34] used an instruction-guided image tokenizer to filter and compress image embeddings based on task, extracting visual information relevant to the current task. To efficiently locate and understand multiple target regions in complex scenes, Groma [29] introduced a localized visual tokenization mechanism that decomposes an image into multiple regions of interest, encoding each region as an independent region token. To unleash the potential of large language models, Griffon [53] introduces a foundational language-prompted localization dataset with nearly 6 million pretraining samples.

2.3. Large Language Models

With the expansion of training data and a significant increase in model parameters, large language models (LLMs)

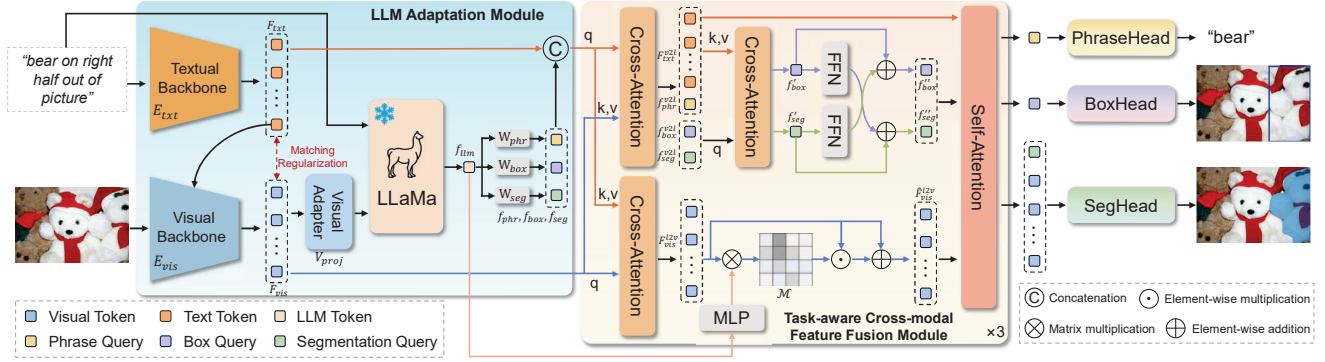


Figure 2. An overview of the proposed method. We first extract textual features F_{txt} using the textual backbone E_{txt} . These features are then injected into the visual backbone E_{vis} , which consists of transformer blocks and conditional normalization, to extract text-related visual features F_{vis} . The object-text semantic matching regularization is employed to associate visual region tokens with textual tokens. We input the visual features processed by the visual adapter V_{proj} , along with the text, into LLaMa. We take the last token f_{llm} output from LLaMa. f_{llm} is then fed into a query projection layer to generate task-relevant tokens f_{box} , f_{seg} , and f_{phr} , which are used to predict the bounding box, segmentation, and phrase, respectively. A task-aware cross-modal feature fusion module is then employed to integrate textual, visual, and task features. Finally, task-relevant tokens are each fed to the corresponding prediction heads to produce the final predictions.

based on the Transformer [42] architecture have developed rapidly. GPT-3 [3], with its 175 billion parameters, demonstrates strong performance in few-shot contextual learning. Following the success of GPT-3, several other LLMs, including PaLM [8], OPT [57], BLOOM [21], and LLaMA [40, 41], have emerged in succession. Mistral [19] introduced a specialized window attention mechanism to better handle extended context modeling. In parallel, Mixtral 8x7B utilized sparse MoE layers [14] to scale up parameters more effectively, enabling it to achieve higher performance with a reduced number of active parameters.

The main differences between the proposed TCRT and existing visual grounding methods are summarized as follows: First, we perform statistical modulation on visual features conditioned on textural features to reduce domain discrepancy before integrating them, while the existing methods [4, 53, 54] only perform dimension adaptation. Second, TCRT leverage the task-aware embeddings derived from LLMs to guide cross-modal feature fusion and association analysis, rather than directly predicting the bounding boxes of target objects as in previous methods [4, 53, 54].

3. Methodology

3.1. Problem Setting

In this section, we begin by defining multi-task visual grounding (MTVG), and then proceed to describe the architecture of the proposed framework. The goal of MTVG is to detect or segment a target object based on a textual description. Formally, given an image $I \in \mathbb{R}^{3 \times H \times W}$ and a corresponding text q , this task is to predict a bounding box $b = \{x, y, w, h\}$ that encompasses the object men-

tioned in the text, as well as a pixel-wise segmentation mask $s \in \mathbb{R}^{1 \times H \times W}$ that provides a more detailed and precise localization. A textual backbone E_{txt} and a visual backbone E_{vis} are used to extract textual and visual features, respectively.

3.2. LLM Adaptation

Pre-trained LLMs possess language understanding capabilities and can handle many complex tasks in natural language processing. However, due to the domain discrepancy between textual data and visual data, directly applying LLMs to grounding tasks may yield suboptimal performance. Therefore, we propose an LLM adaptation module to perform cross-modal adaptation of image information, enabling LLMs to better handle information within visual data. As shown in Figure 2, we first utilize the text encoding model E_{txt} to extract the textual features of q : $F_{txt} = E_{txt}(q)$, where F_{txt} represents the textual features, which include three types of textual features: sentence feature $f_{txt}^{sen} \in \mathbb{R}^{1 \times d}$, individual word features $f_{txt}^{word} \in \mathbb{R}^{L \times d}$, and average pooled word feature $f_{txt}^{mean} \in \mathbb{R}^{1 \times d}$, where d and L represent the dimension of textual features and the number of tokens, respectively.

To enable the visual encoder to comprehend textual information, we designed a text-aware visual encoder E_{vis} . Specifically, a conditioned normalization is inserted after each transformer block in the visual encoder. In the conditioned normalization, we fuse the sentence features f_{txt}^{sen} and the average-pooled word features f_{txt}^{mean} into a textual conditional feature f_{txt}^{cond} via a weighted combination:

$$f_{txt}^{cond} = \alpha^{sen} f_{txt}^{sen} + \alpha^{mean} \phi(f_{txt}^{mean}), \quad (1)$$

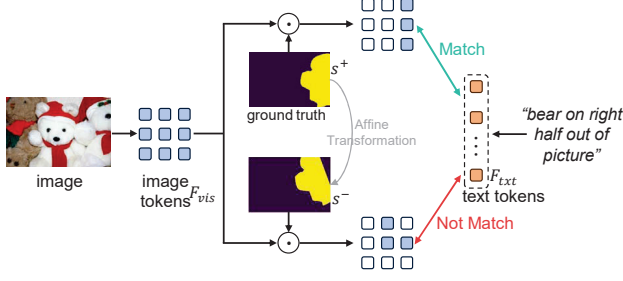


Figure 3. Object-text semantic matching regularization. We constrain the image features using a mask to isolate the object-specific features. For matching text-object pairs, we consider them as positive samples, whereas non-matching pairs are regarded as negative samples.

where α^{sen} and α^{mean} are learnable parameters that control the contribution of f_{txt}^{sen} and f_{txt}^{mean} , respectively. $\phi(\cdot)$ represents the feed-forward network. The scaling γ and shifting β parameters are determined using f_{txt}^{cond} : $\gamma = \phi(f_{txt}^{cond}) + \gamma^{learn}$, $\beta = \phi(f_{txt}^{cond}) + \beta^{learn}$, where γ^{learn} and β^{learn} are learnable parameters. Given the visual features \bar{F}_i^* , which is the output of the i -th layer of transformer block, we calculate the mean $\mu(\bar{F}_i^*)$ and standard deviation $\sigma(\bar{F}_i^*)$. The visual features \bar{F}_i^* are then modulated using the resulting scaling γ and shifting β parameters, as follows:

$$\bar{F}_i^* = \gamma \left(\frac{\bar{F}_i^* - \mu(\bar{F}_i^*)}{\sigma(\bar{F}_i^*) + \epsilon} \right) + \beta, \quad (2)$$

where ϵ is a small constant introduced to avoid division by zero. We obtain the text-aware visual features F_{vis} by flattening the feature map \bar{F}_4^* extracted from the visual backbone. The image features and the input embeddings of LLMs are typically in different dimensional spaces. We use a visual adapter V_{proj} to map the image features to a dimension that matches the LLM input, allowing the LLM to directly process visual information. V_{proj} is a three-layer MLP. We feed the text and the mapped image features into the LLM, and take the last token of the LLM output as the prior embedding f_{llm} :

$$f_{llm} = \Phi(q, V_{proj}(\underbrace{E_{vis}(I, f_{txt}^{cond})}_{F_{vis}})), \quad (3)$$

where Φ represents an LLM model, q denotes the original query text. We pass f_{llm} through three task-aware linear layers (W_{box} , W_{seg} , and W_{phr}) to obtain the task-specific embeddings f_{box} , f_{seg} , and f_{phr} , which are responsible for predicting the bounding box, segmentation, and phrase, respectively.

The text usually describes a specific object within an image, which occupies only a portion of the image. Most visual features do not have a direct relationship with the

text. This makes it challenging to achieve precise semantic alignment between the target object and textual tokens. Thus, we propose the object-text semantic matching regularization to associate the visual region tokens with the textual ones. Object-text matching is a binary classification task, and the model evaluates the multimodal features of an object-text pair to classify it as either positive (matched) or negative (unmatched). As shown in Figure 3, if the textual features and visual features represent the same object, they are considered a positive sample pair; otherwise, they are considered a negative sample pair. The ground truth of object segmentation is represented by s^+ . We apply an affine transformation to s^+ to obtain s^- . s^- represents the inaccurate segmentation result. We leverage the segmentation mask of the target object to extract its features. A learnable token f_{match} is then introduced to determine whether they are matched:

$$\begin{aligned} \{\bar{f}_{match}, \bar{F}_l, \bar{F}_v\} &= \varphi([f_{match}, F_{txt}, F_{vis}]), \\ \{\hat{f}_{match}^j, \hat{F}_l\} &= \theta([\bar{f}_{match}, \bar{F}_l], F_{vis} \odot s^j), \end{aligned} \quad (4)$$

where $j \in \{+, -\}$. $\varphi(q)$ and $\theta(q, kv)$ represent multi-head self-attention and cross-attention, respectively. \odot denotes element-wise multiplication. $[\cdot]$ denotes the concatenation operation. We use the \hat{f}_{match}^j as the joint representation of the object-text pair. A feed-forward network is employed to predict the two-class probability p_{match}^j that indicates whether a textual feature matches a local visual feature. The object-text matching loss is:

$$\mathcal{L}_{match} = \mathbb{E}_{(I, q)} \left[\sum_j H(y_{match}^j, p_{match}^j) \right], \quad (5)$$

where H represents cross-entropy and y_{match}^j is a 2-dimensional one-hot vector that represents the ground truth label. We use the visual features of the target object extracted from the ground truth mask as positive sample, indicating that these features match the textual description. When performing affine transformation (partial overlap may occur), the resulting mask deviates from the target object, and the extracted features are considered as negative sample. Object-text matching loss further associates visual tokens with textual ones by assessing whether they match.

3.3. Task-aware Cross-modal Feature Fusion

When the LLM receives textual and visual information, it can generate more meaningful semantic representations. These representations contain the LLM’s understanding of objects and context, forming prior information that aids downstream tasks. The task-aware cross-modal feature fusion module, denoted as V_F , is incorporated to guide feature the fusion process. Specifically, we first employ a bi-directional multi-head attention module to fuse textual and

visual features. Each modality serves as the query while the other modality acts as the key and value in the attention computation. We concatenate the three tokens used for prediction, f_{box} , f_{seg} , and f_{phr} , with the textual features F_{txt} to obtain a task-relevant textual feature F_{task} . Then, F_{task} and F_{vis} are fed into the bi-directional multi-head attention module to fuse the two types of features. This process can be represented as follows: $F_{task}^{v2t} = \theta(F_{task}, F_{vis})$, $F_{vis}^{t2v} = \theta(F_{vis}, F_{task})$, where F_{task}^{v2t} represents the textual feature enriched with image features, comprising F_{txt}^{v2t} , f_{box}^{v2t} , f_{seg}^{v2t} , and f_{phr}^{v2t} . Meanwhile, F_{vis}^{t2v} denotes the visual feature enriched with textual information.

Target objects are typically located in specific regions. To enable the model to focus on local visual features, we input f_{llm} into an MLP to obtain the refinement prompt f_{prmt} , which guides the feature refinement process. We then assign different weights to image features at different locations by computing the cosine similarity between the image feature F_{vis}^{t2v} and f_{prmt} . This process can be formulated as follows:

$$\hat{F}_v^{t2v} = F_{vis}^{t2v} + F_{vis}^{t2v} \odot \underbrace{(\text{Conv}_{1 \times 1}(F_{vis}^{t2v}) \otimes \text{MLP}(f_{llm}))}_{\mathcal{M}}, \quad (6)$$

where \otimes denotes matrix multiplication and \odot represents element-wise multiplication. The refinement prompt f_{prmt} is dot-multiplied with the image features at each position in F_{vis}^{t2v} to obtain a weight matrix \mathcal{M} . The final refined visual features \hat{F}_v^{t2v} are obtained under the guidance of the LLM. The model can control the strength of the feature representations at different positions in the image by adjusting the weight matrix \mathcal{M} . By assigning lower weights to irrelevant regions, the model can better focus on the features of the target regions.

There is a close relationship between the REC task and the RES task. Both require an understanding and localization of the target object or region. REC tasks generally offer locational information of the target objects, whereas RES tasks offer precise boundary details. To strengthen the relationship between the REC and RES tasks, we perform cross-task interaction to effectively exchange information between the box token f_{box}^{v2t} and the segmentation token f_{seg}^{v2t} . The process is as follows:

$$\begin{aligned} \{f'_{box}, f'_{seg}\} &= \theta([f_{box}^{v2t}, f_{seg}^{v2t}], F_{txt}^{v2t}), \\ f''_{box} &= f'_{box} + \phi(f'_{seg}), f''_{seg} = f'_{seg} + \phi(f'_{box}), \end{aligned} \quad (7)$$

This design allows the model to exploit the correlation between the REC and RES tasks. The updated f_{box} , f_{seg} , and f_{phrase} are each fed into their respective prediction heads to obtain the final bounding box, segmentation, and phrase predictions.

3.4. Model Training

Detection loss. For the REC task, the loss for bounding box regression is calculated using a combination of GIoU loss and L1 loss. We use b to denote the ground truth bounding box and \tilde{b} to represent the predicted bounding box. The following defines the detection loss function:

$$\mathcal{L}_{det} = \mathbb{E}_{(I,q)}[\Phi_{smooth-l1}(b, \tilde{b}) + \Phi_{giou}(b, \tilde{b})], \quad (8)$$

where $\Phi_{smooth-l1}$ and Φ_{giou} denote the smooth L1 loss [15] and Generalized IoU loss [37], respectively.

Segmentation loss. For the RES task, the segmentation loss is computed by combining the focal loss with dice loss. The ground truth mask is denoted as s , while the predicted mask is \hat{s} . The following defines the segmentation loss function:

$$\mathcal{L}_{seg} = \mathbb{E}_{(I,q)}[\Phi_{focal}(s, \hat{s}) + \Phi_{dice}(s, \hat{s})], \quad (9)$$

where Φ_{focal} and Φ_{dice} denote the smooth focal loss [24] and DICE/F-1 loss [31], respectively.

Phrase loss. To enable the model to more accurately capture the entities within the text, we utilize f_{phr} to predict the words in the text related to the target object. We fed f_{phr} into an MLP to predict a word probability p_{phr} . Then, we compute the phrase loss as follows:

$$\mathcal{L}_{phr} = \mathbb{E}_{(I,q)}[H(y_{phr}, p_{phr})], \quad (10)$$

where y_{phr} denotes the ground truth label.

By incorporating these training loss functions, the optimization problem of our model can be formulated as follows:

$$\begin{aligned} \min_{E_{txt}, E_{vis}} \lambda_{match} \mathcal{L}_{match}, \\ \min_{E_{txt}, E_{vis}, V_{proj}, W_{task}, V_F} \mathcal{L}_{det} + \mathcal{L}_{seg} + \lambda_{phr} \mathcal{L}_{phr}, \end{aligned} \quad (11)$$

where $task \in \{box, seg, phr\}$, and λ_{match} and λ_{phr} are used to achieve a balance among the training objectives. \mathcal{L}_{det} and \mathcal{L}_{seg} enable the model to perform basic detection and segmentation tasks. \mathcal{L}_{phr} enables the model to achieve a more detailed understanding of the target objects referred to in the text. \mathcal{L}_{match} is used to associate visual tokens with textual ones and only optimizes the textual backbone and visual backbone. Our LLM adaptation module and task-aware cross-modal feature fusion module are jointly optimized during the training process.

4. Experiments

In this section, we first describe datasets, evaluation metrics, and implementation details. Next, we qualitatively and quantitatively compare our method with state-of-the-art methods. We further perform comprehensive experiments to analyze the main components of TCRT, including conditional normalization, object-text semantic matching regularization, and LLM-guided visual feature refinement.

Methods	Visual Backbone	Multi-task	Venue	RefCOCO			RefCOCOg		RefCOCO+			ReferItGame test
				val	testA	testB	val	test	val	testA	testB	
Two-stage Methods												
MAttNet [52]	RN101	×	CVPR18	76.65	81.14	69.99	66.58	67.27	65.33	71.62	56.02	29.04
CM-A-E [27]	RN101	×	CVPR19	78.35	83.14	71.32	67.99	68.67	68.09	73.65	58.03	-
Ref-NMS [6]	RN101	×	AAAI21	80.70	84.00	76.04	70.55	70.62	68.25	73.68	59.42	-
End-to-end Methods												
MCN [28]	DN53	✓	CVPR20	80.08	82.29	74.98	66.46	66.01	67.16	72.86	57.31	-
RefTR [22]	RN101	✓	NeurIPS21	82.23	85.59	76.57	69.41	69.40	71.58	75.96	62.16	71.42
SeqTR [59]	DN53	✓	ECCV22	81.23	85.00	76.08	71.35	71.58	68.82	75.37	58.78	69.66
VG-LAW [38]	ViT-B	✓	CVPR23	86.62	89.32	83.16	76.90	76.96	76.37	81.04	67.50	77.22
PVD [7]	Swin-B	✓	AAAI24	84.99	88.02	80.03	74.34	74.64	74.27	79.06	65.11	-
Ferret-7B [50]	ViT-L	×	ICLR24	87.49	91.35	82.45	83.93	84.76	80.78	87.38	73.14	-
Vista-7B [34]	ViT-L	✓	CVPR24	88.1	91.5	83.0	86.0	86.4	84.1	90.3	75.8	-
TCRT	Swin-B	✓	-	91.07	92.85	86.24	86.81	87.13	85.75	90.76	77.59	79.09

Table 1. Comparison with existing visual grounding methods on RefCOCO, RefCOCOg, RefCOCO+, and ReferItGame for the REC task.

Methods	Visual Backbone	Multi-task	Venue	RefCOCO			RefCOCOg		RefCOCO+		
				val	testA	testB	val	test	val	testA	testB
MCN [28]	DN53	✓	CVPR20	62.44	64.20	59.71	49.22	49.40	50.62	54.99	44.69
RefTR [22]	RN101	✓	NeurIPS21	70.56	73.49	66.57	58.73	58.51	61.08	64.65	52.73
SeqTR [59]	DN53	✓	ECCV22	67.26	69.79	64.12	55.67	55.64	54.14	58.93	48.19
VG-LAW [38]	ViT-B	✓	CVPR23	75.62	77.51	72.89	65.63	66.08	66.63	70.38	58.89
PVD [7]	Swin-B	✓	AAAI24	75.07	77.29	70.13	63.22	63.89	64.39	69.15	57.19
Vista-7B [34]	ViT-L	✓	CVPR24	74.5	76.0	73.9	69.8	70.9	71.8	74.4	65.6
TCRT	Swin-B	✓	-	77.12	78.12	75.67	70.84	72.08	72.79	74.96	66.53

Table 2. Comparison with existing visual grounding methods on RefCOCO, RefCOCOg, and RefCOCO+ for the RES task.

4.1. Experimental Settings

4.1.1. Datasets

To evaluate the effectiveness of our method, we carry out experiments on the RefCOCO [51], RefCOCOg [33], RefCOCO+ [51], and ReferItGame [20] datasets. RefCOCO, RefCOCOg, and RefCOCO+ datasets are derived from MSCOCO [23]. RefCOCO and RefCOCO+ datasets were partitioned into train, validation (val), testA, and testB subsets. RefCOCOg was partitioned into train, validation (val), and test subsets. In addition, the ReferItGame dataset, which was collected from SAIAPR-12 [13], comprises both train and test subsets.

4.1.2. Evaluation Metrics

For the RSC task, we employ Prec@0.5 as the metric for evaluation, which is consistent with prior works. If the IoU of the ground truth box with the predicted box surpasses 0.5, then the predicted box is deemed accurate. For the RES task, we evaluate performance by calculating MIOU.

4.1.3. Implementation Details

Our visual backbone utilizes a Swin Transformer architecture pre-trained with Mask-RCNN on MSCOCO [23]. The input image size is set to 448×448 . For the textual backbone, we adopt the T5 [35] model and limit the maximum length of text to 40. For the large language model, we use Llama3.1-8B [12], which is frozen during the training

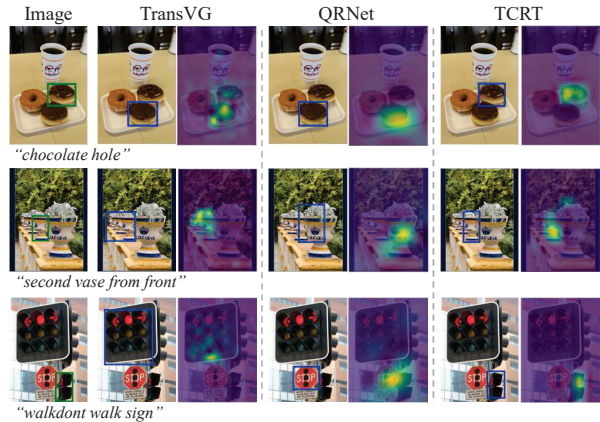


Figure 4. Qualitative comparison with representative methods on REC task. We also visualize the attention scores of each method. Blue and green boxes are predicted regions and the ground truth, respectively.

process. Our model undergoes end-to-end training for 70 epochs. The learning rate begins at 10^{-4} , and then we drop it down to 10^{-5} after 50 epochs. The optimizer used is AdamW. A dropout rate of 0.1 is applied. The batch size is configured to 16. We adopt the same data augmentation strategies as previous works [47, 48]. We implement our TCRT using PyTorch and perform the experiments on NVIDIA A800 GPUs.

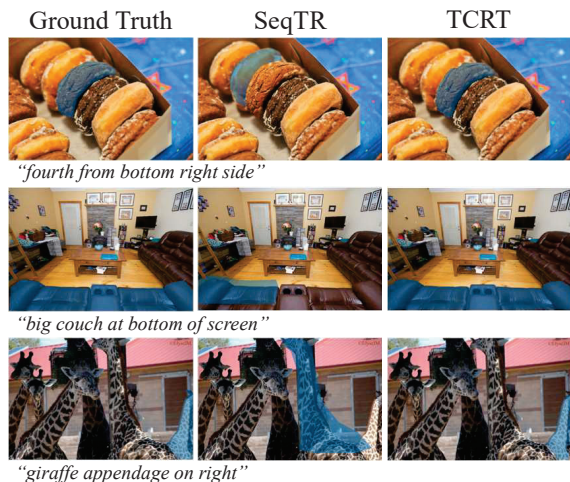


Figure 5. Qualitative comparison with SeqTR on RES task.

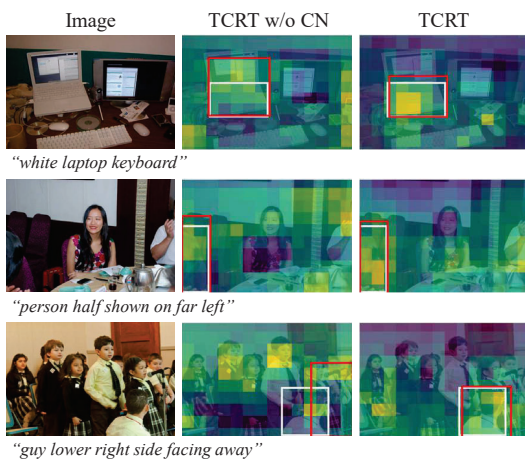


Figure 6. Visualization of features extracted by the visual backbone. The second column shows the results of ‘TCRT w/o CN’, while the third column shows the results of TCRT. Red and white boxes are the predicted regions and the ground truth, respectively.

4.2. Comparisons with State-of-the-arts

To evaluate the effectiveness of the proposed TCRT, we perform both qualitative and quantitative experiments across four datasets: RefCOCO, RefCOCOg, RefCOCO+, and ReferItGame.

4.2.1. REC task

We evaluate our TCRT against existing state-of-the-art models for the REC task. As shown in Table 1, TCRT achieves the best performance across these datasets. Specifically, compared to Vista-7B, TCRT shows absolute improvements of +2.97, +1.35, and +3.24 percentage points on RefCOCO; +1.65, +0.46, and +1.79 percentage points on RefCOCO+; and +0.81 and +0.73 percentage points on RefCOCOg. We also present qualitative results from three

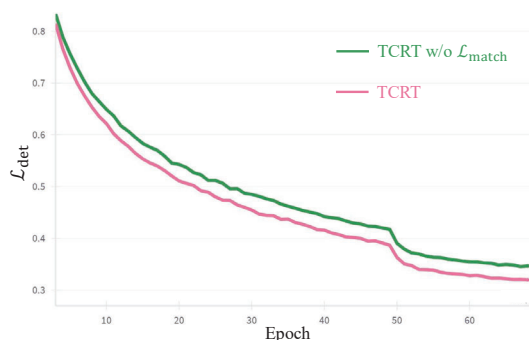


Figure 7. Comparison between the TCRT without object-text matching loss ($\mathcal{L}_{\text{match}}$) and TCRT in terms of \mathcal{L}_{det} .

examples of the RefCOCO testB set for the REC task in Figure 4. We can see that TCRT performs well even when an image contains multiple similar objects, such as “second vase from front”. Furthermore, we visualized the attention scores within the transformer. We observe that TCRT focuses better on the target object compared to other methods.

4.2.2. RES task

We evaluate the effectiveness of our TCRT for the RES task by comparing it with existing state-of-the-art methods. The comparative results are provided in Table 2. Compared with Vista-7B, TCRT achieves higher mIoU scores with absolute improvements of +2.62, +2.12, and +1.77 percentage points on RefCOCO, +0.99, +0.56, and +0.93 percentage points on RefCOCO+, +1.04 and +1.18 percentage points on RefCOCOg. We also present a comparison for the RES task in Figure 5.

4.3. Analysis of Main Components

4.3.1. Conditional Normalization

We first conduct an experiment to demonstrate the effectiveness of conditional normalization (CN). We visualize the features extracted by the visual backbone on RefCOCOg test set. As shown in Figure 6, ‘TCRT w/o CN’ has more dispersed attention, which may lead to inaccurate predictions. Due to the smaller domain discrepancy of TCRT, this allows the LLM to focus more attention on the target objects.

4.3.2. Object-text Semantic Matching Regularization

We use the object-text matching loss ($\mathcal{L}_{\text{match}}$) to associate the visual tokens with the textual ones. To validate the effectiveness of $\mathcal{L}_{\text{match}}$, we compare ‘TCRT w/o $\mathcal{L}_{\text{match}}$ ’ and TCRT in terms of the detection loss (\mathcal{L}_{det}) on RefCOCO val set. In Figure 7, by plotting the loss curves of the models during training, we observe that TCRT achieves lower loss values, which is attributed to its capability to learn semantic-discriminative visual features. This enables the LLM to more easily recognize task-relevant visual features.

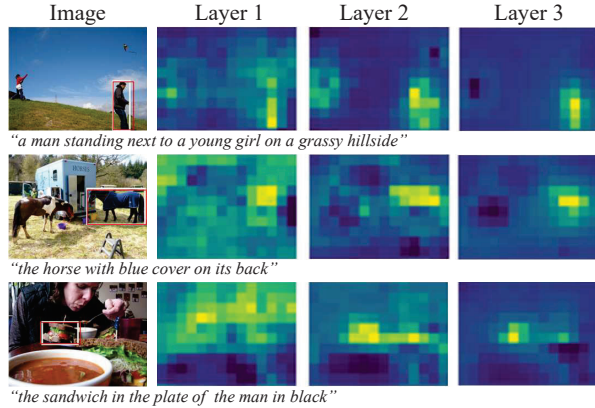


Figure 8. visualization of LLM-guided visual feature refinement. We visualize the weight matrices \mathcal{M} of each layer in the fusion module to validate the role of the LLM in refining visual features. Red and white boxes are the predicted regions and the ground truth, respectively.

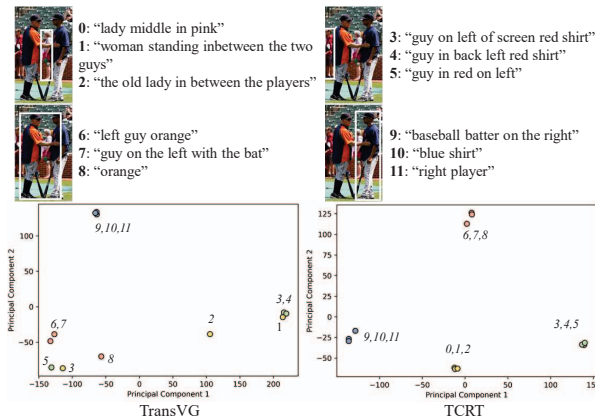


Figure 9. Analysis of Visual Feature Distribution.

4.3.3. LLM-guided Visual Feature Refinement

We visualize the weight matrix \mathcal{M} in Figure 8 to validate the effectiveness of visual feature refinement on RefCOCOg test set. We observe that the attention of the weight matrix is relatively evenly distributed in the earlier layers. In the later layers, the attention of the weight matrix is primarily focused on the target object. This indicates that LLM’s prior knowledge enhances our model’s ability to locate target objects by further refining the visual features.

4.3.4. Analysis of Visual Feature Distribution

To validate the effectiveness of our model in aligning visual and textual features, we extract the visual features after fusion with textual information and study their distribution in the spatial domain. Specifically, we label four target objects in an image and use three different textual descriptions for each object, resulting in twelve data samples. These samples are fed into the grounding model, and the visual

Method	RefCOCO		
	val	testA	testB
TCRT w/o V_F & V_A	87.62	88.60	84.04
TCRT w/o V_F	88.38	90.24	85.57
TCRT	91.07	92.85	86.24

Table 3. Ablation experiments on RefCOCO to evaluate the proposed LLM adaptation module (V_A) and task-aware cross-modal feature fusion module (V_F).

features from the fusion module’s output are subjected to dimensionality reduction and visualization. As shown in Figure 9, we observe that our method causes the visual features describing the same object to cluster together.

4.3.5. Ablation Studies

We validate the design of our proposed TCRT by conducting ablative experiments on major components in the REC task. The results are presented in Table 3. We observe that when the LLM adaptation module and task-aware cross-modal feature fusion module are not used, ‘TCRT w/o V_F & V_A ’ achieves accuracy rates of 87.62%, 88.60%, and 84.04% on the RefCOCO dataset. After incorporating the LLM adaptation module, the accuracy of ‘TCRT w/o V_F ’ increases by 0.76, 1.64, and 1.53 absolute percentage points. When the V_F module is further introduced, TCRT achieves accuracy rates of 91.07%, 92.85%, and 86.24%.

5. Conclusion

In this paper, we propose a task-aware cross-modal feature refinement transformer with large language models to facilitate visual grounding. To adapt LLMs to the grounding task, we design an adaptation module that modulates the statistical characteristics of visual features conditioned on textual features along with object-text semantic matching regularization. An LLM is employed to integrate the aligned information from textual and visual features and produce task-aware embeddings. We further incorporate a task-aware cross-modal feature fusion module, allowing the resulting embeddings to refine visual features and enable information interaction between the REC and RES tasks. Comprehensive experimental results verify the effectiveness of our proposed approach both quantitatively and qualitatively.

6. Acknowledgments

This work was supported in part by TCL Science and Technology Innovation Fund (Project No. 20231752), in part by the Research Grants Council of the Hong Kong Special Administration Region (Project No. CityU 11206622), and in part by the Guangdong Basic and Applied Basic Research Foundation (Project No. 2024A1515011437).

References

- [1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023.
- [2] Mohit Bajaj, Lanjun Wang, and Leonid Sigal. G3graphground: Graph-based language grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4281–4290, 2019. 2
- [3] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 3
- [4] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 2, 3
- [5] Kan Chen, Rama Kovvuri, and Ram Nevatia. Query-guided regression network with context policy for phrase grounding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 824–832, 2017. 2
- [6] 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*, pages 1036–1044, 2021. 2, 6
- [7] 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*, pages 1326–1334, 2024. 2, 6
- [8] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240): 1–113, 2023. 3
- [9] 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. 1, 2
- [10] 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. 1, 2
- [11] Pelin Dogan, Leonid Sigal, and Markus Gross. Neural sequential phrase grounding (seqground). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4175–4184, 2019. 2
- [12] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 6
- [13] Hugo Jair Escalante, Carlos A Hernández, Jesus A Gonzalez, Aurelio López-López, Manuel Montes, Eduardo F Morales, L Enrique Sucar, Luis Villaseñor, and Michael Grubinger. The segmented and annotated iapr tc-12 benchmark. *Computer Vision and Image Understanding*, 114(4): 419–428, 2010. 6
- [14] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022. 3
- [15] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015. 2, 5
- [16] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014. 2
- [17] Chih-Hui Ho, Srikanth Appalaraju, Bhavan Jasani, R Manmatha, and Nuno Vasconcelos. Yoro-lightweight end to end visual grounding. In *European Conference on Computer Vision*, pages 3–23. Springer, 2022. 1
- [18] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. Modeling relationships in referential expressions with compositional modular networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1115–1124, 2017. 2
- [19] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. 3
- [20] 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. 6
- [21] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv e-prints*, pages arXiv–2211, 2022. 3
- [22] 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. 1, 2, 6
- [23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6
- [24] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017. 5
- [25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee.

- Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1, 2
- [26] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 2
- [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. 2, 6
- [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. 6
- [29] Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. Groma: Localized visual tokenization for grounding multimodal large language models. In *European Conference on Computer Vision*, pages 417–435. Springer, 2025. 2
- [30] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11–20, 2016. 2
- [31] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571. Ieee, 2016. 5
- [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. 2
- [33] 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. 6
- [34] Shraman Pramanick, Guangxing Han, Rui Hou, Sayan Nag, Ser-Nam Lim, Nicolas Ballas, Qifan Wang, Rama Chellappa, and Amjad Almahairi. Jack of all tasks master of many: Designing general-purpose coarse-to-fine vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14076–14088, 2024. 2, 6
- [35] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. 6
- [36] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28, 2015. 1, 2
- [37] Hamid Reza Tofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. 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*, pages 658–666, 2019. 5
- [38] 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. 1, 2, 6
- [39] InternLM Team. Internlm: A multilingual language model with progressively enhanced capabilities, 2023.
- [40] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 3
- [41] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shrutu Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 3
- [42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. 3
- [43] Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*, 2023.
- [44] Sibe Yang, Guanbin Li, and Yizhou Yu. Cross-modal relationship inference for grounding referring expressions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4145–4154, 2019. 2
- [45] Sibe Yang, Guanbin Li, and Yizhou Yu. Dynamic graph attention for referring expression comprehension. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4644–4653, 2019. 1
- [46] Sibe Yang, Guanbin Li, and Yizhou Yu. Graph-structured referring expression reasoning in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9952–9961, 2020. 2
- [47] 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. 2, 6
- [48] Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. Improving one-stage visual grounding by recursive subquery construction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 387–404. Springer, 2020. 2, 6
- [49] 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. 1, 2
- [50] 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. 6
- [51] 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. 6
- [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. 2, 6
- [53] Yufei Zhan, Yousong Zhu, Zhiyang Chen, Fan Yang, Ming Tang, and Jinqiao Wang. Griffon: Spelling out all object locations at any granularity with large language models. In *European Conference on Computer Vision*, pages 405–422. Springer, 2025. 2, 3
- [54] Ao Zhang, Liming Zhao, Chen-Wei Xie, Yun Zheng, Wei Ji, and Tat-Seng Chua. Next-chat: An lmm for chat, detection and segmentation. *arXiv preprint arXiv:2311.04498*, 2023. 3
- [55] Hanwang Zhang, Yulei Niu, and Shih-Fu Chang. Grounding referring expressions in images by variational context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4158–4166, 2018. 2
- [56] Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. Tynllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*, 2024.
- [57] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 3
- [58] Tao Zhang, Xiangtai Li, Hao Fei, Haobo Yuan, Shengqiong Wu, Shunping Ji, Chen Change Loy, and Shuicheng Yan. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. *Advances in Neural Information Processing Systems*, 37:71737–71767, 2024. 2
- [59] 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. 1, 6