Referring to Any Person

Qing Jiang[1,2] , Lin Wu[1,2] , Zhaoyang Zeng[2] , Tianhe Ren[2] , Yuda Xiong[2]
Yihao Chen[2] , Liu Qin[2] , Lei Zhang[1,2†]
[1]South China University of Technology
[2]International Digital Economy Academy (IDEA)
mountchicken@outlook.com , leizhang@idea.edu.cn

Figure 1. We introduce referring to any person, a task that requires detecting all individuals in an image which match a given natural language description, and a new model RexSeek designed for this task with strong perception and understanding capabilities that effectively captures attributes, spatial relations, interactions, reasoning, celebrity recognition, etc.

Abstract

Humans are undoubtedly the most important participants in computer vision, and the ability to detect any individual given a natural language description, a task we define as referring to any person, holds substantial practical value. However, we find that existing models generally fail to achieve real-world usability, and current benchmarks are limited by their focus on one-to-one referring, which hinders progress in this area. In this work, we revisit this task from three critical perspectives: task definition, dataset design, and model architecture. We first identify five aspects of referable entities and three distinctive characteristics of this task. Next, we introduce HumanRef, a novel dataset designed to tackle these challenges and better reflect real-world applications. From a model design perspective, we integrate a multimodal large language model with an object detection framework, constructing a robust referring model named RexSeek. Experimental results reveal that state-of-the-art models, which perform well on commonly used benchmarks like RefCOCO/+/g, struggle with HumanRef due to their inability to detect multiple individuals. In con-

This work was done when Qing Jiang and Lin Wu were interns at IDEA.

† *Corresponding author.*

trast, RexSeek not only excels in human referring but also generalizes effectively to common object referring, making it broadly applicable across various perception tasks. Code is available at `https://github.com/IDEA-Research/RexSeek`.

1. Introduction

Humans are central to computer vision [4, 10–12, 16, 21, 27, 28, 30, 45, 46, 63, 65, 84, 85], and the ability to identify and detect specific individuals based on natural language descriptions, a task we define as referring to any person, is crucial for numerous applications, including human-robot interaction, industrial automation, healthcare, etc.

However, we argue that progress in this area has been hindered by unclear task definitions and a lack of high-quality data. Our findings show that despite achieving state-of-the-art performance on referring benchmarks RefCOCO/+/g [50, 75], most models remain impractical for real-world applications, as illustrated in Figure 2. To address this challenge, we revisit this task from three perspectives: task definition, dataset construction, and model design.

We begin by formally defining the task of referring to any person: *given a natural language description and an input image, the model needs to detect all individuals in the image who match the description*. To comprehensively capture the scope of this task, we identify five key aspects that define how humans can be referred to: **i) Attributes:** Encompassing intrinsic characteristics such as gender, age, action, clothing, etc. **ii) Position:** Describing spatial relationships both among individuals and between individuals and their surroundings. **iii) Interaction:** Accounting for human-to-human, human-to-object, and human-to-environment interactions. **iv) Reasoning:** Involving multi-step inference that considers multiple objects to resolve complex expressions. **v) Celebrity Recognition:** Identifying specific individuals, whether by their real names or characters names.

Next, we identify three crucial characteristics that define this task: **i) Multi-Instance Referring**: A referring expression can correspond to multiple individuals. While mainstream referring datasets RefCOCO/+/g [50, 75] typically assume that each expression refers to a single object, this does not align with real-world scenarios. We find through experiments that most models experience significant performance degradation when tasked with identifying more than one individual. **ii) Multi-Instance Discrimination**: The image should contain multiple individuals in addition to the target person. This setting ensures that the model fully comprehends the referring expression to identify the correct individual rather than simply detecting all people in the image. **iii) Rejection of Non-existence**: If the referred person is not present in the image, the model should refuse to generate a result rather than produce a hallucinated output.

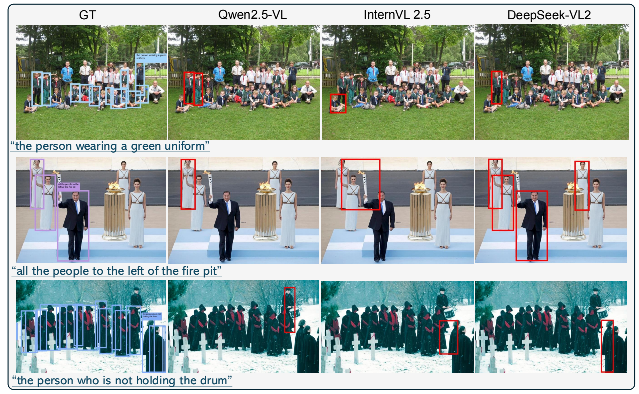


Figure 2. Visualization results of Qwen2.5-VL [3], InternVL-2.5 [14], and DeepSeek-VL2 [70] on the human referring task. Despite achieving strong results on referring benchmarks RefCOCO/+/g [50, 75], state-of-the-art models struggle when tasked with identifying multiple individuals as they output an insufficient number of bounding boxes.

Based on the task definition, we manually constructed a novel dataset for human referring, named HumanRef. Unlike the traditional ReferItGame [29] annotation approach, where one annotator describes an object and another finds it based on the description, we adopt a different annotation methodology. Our process begins with annotators listing the key properties of individuals in an image according to the predefined referable entities. Next, for each person, they determine whether these properties apply and result in a property dictionary. Finally, a large language model [71] composes these properties into referring expressions. HumanRef comprises 103,028 referring statements, with each expression referring to an average of 2.2 instances. We also split a benchmark from HumanRef with 6,000 referring expressions spanning six subsets, ensuring comprehensive coverage across all referable properties.

From the model design perspective, we argue that a robust referring model should possess two key characteristics: **i) Robust Perception Ability**: The model should be capable of detecting all individuals in an image. **ii) Strong Language Comprehension**: The model should effectively interpret complex language descriptions of people. To address these requirements, we introduce RexSeek, a detection-oriented multimodal large language model specifically designed for this task. Inspired by ChatRex [25], we formulate referring as a retrieval-based task. RexSeek integrates a person detector [60] as its box input, ensuring strong perception capabilities while incorporating Qwen2.5 [71] as the LLM to enhance language comprehension. We adopt a multi-stage training approach that progressively refines both detection and comprehension skills, equipping RexSeek with strong referring capabilities.

Experimental results indicate that most state-of-the-art models [3, 9, 14, 25, 49, 60, 70, 74] exhibit performance degradation on the HumanRef benchmark, despite achieving strong results on RefCOCO/+/g. The primary limitation is that these models typically detect only a single instance, as they are trained on datasets that assume one-to-one referring. In contrast, RexSeek, trained on HumanRef, exhibits strong referring capabilities. Additionally, benefiting from the multi-stage training approach, RexSeek also emerges with the ability to refer to generalized objects, extending its applicability beyond human-centric tasks. To summarize, our contributions are threefold:

- We introduce referring to any person with a clear definition by identifying five aspects of referable entities and three key characteristics that distinguish this task.
- We introduce HumanRef, a novel referring dataset, and establish a challenging benchmark to drive progress in human-centric referring expression research.
- We propose RexSeek, a detection-oriented multimodal large language model trained through a multi-stage process, demonstrating strong referring capabilities for both humans and general objects.

2. Related Work

Referring Expression Comprehension Task. Referring Expression Comprehension (REC) [29, 36, 48, 50, 56, 72, 75, 76, 76, 81] involves interpreting a natural language expression to localize specific objects within an image. Unlike open-vocabulary object detection [15, 26, 34, 42, 52, 60, 61, 69, 73, 78] or phrase grounding [18, 23, 31, 54, 68], which identify objects based on brief category names or short phrases, REC requires understanding complex, free-form descriptions. This task necessitates not only recognizing object attributes and relationships but also comprehending spatial configurations and interactions, making it inherently more challenging. In this work, we systematically analyze the referable entities and the critical characteristics that define this task.

REC Datasets and Benchmarks. The first large-scale Referring Expression Comprehension (REC) dataset, ReferItGame [29], was created through a two-player game in which one annotator describes an object, and another selects it. This was later followed by more sophisticated datasets [7, 13, 17, 19, 54, 55], such as RefCOCO [75], RefCOCO+ [75], and RefCOCOg [50], which leverage MSCOCO [37] images to provide more complex referring expressions. Beyond these general datasets, others address specific challenges. CLEVR-Ref+ [40] focuses on geometric object referring. RefCrowd [57] targets person detection in crowded scenes. Ref-L4 [8] handles longer and more detailed descriptions. GRES [77] introduces multi-target referring expression segmentation. However, existing datasets typically assume a one-to-one correspondence between a referring expression and a single instance, which fails to reflect real-world scenarios. To address this gap, we refine the referring task and introduce HumanRef, a dataset specifically designed to support multi-instance referring and advance research in this domain.

domain	sub-domains	examples
attribute	gender, age, race, profession, posture, appearance, clothing and accessories, action	*male, female, white man, the police officer, person with a shocked expression, person wearing a mask, person standing*
position	inner position (human to human), outer position (human to environment)	*the second person from left to right, person at the right, person closest to the microphone, person sitting in the chair*
interaction	inner interaction (human with human), outer interaction (human with environment)	*two people holding hands, people locked in each other's gaze, the person holding a gun, person holding the certificate in hand*
reasoning	inner position reasoning, outer position reasoning, attribute reasoning	*all the people to the right of the person closest to the glass, person wearing a lab coat but not putting their hand on the board*
celebrity recognition	actor, character, athlete, entrepreneur, scientist, politician, singer	*Brad Pitt, Bruce Wayne, Cristiano Ronaldo, Rihanna, Elon Musk, Albert Einstein, Donald Trump*
rejection	attribute, position, interaction, reasoning	*a man in red hat, three women in a circle*

Table 1. The primary annotation domains and their corresponding sub-domains within HumanRef.

MLLM-based REC Methods Multimodal Large Language Models (MLLMs) [1–3, 14, 20, 32, 33, 35, 44, 47, 53, 64, 66, 70, 82] have demonstrated strong capabilities in both text and image comprehension, motivating efforts to integrate referring expression understanding into these models. A common approach involves outputting bounding box coordinates as tokens [3, 9, 14, 51, 67, 70, 74, 79, 80, 83]. Alternatively, methods like Groma [49] and ChatRex [25] frame detection as a retrieval task, where a proposal model generates bounding boxes, and the LLM selects the index of the relevant box based on the referring expression. While these MLLM-based methods achieve high performance on RefCOCO/+/g, our experiments reveal that they remain inadequate for practical applications due to low recall rate on multi-instance referrings.

3. HumanRef Dataset

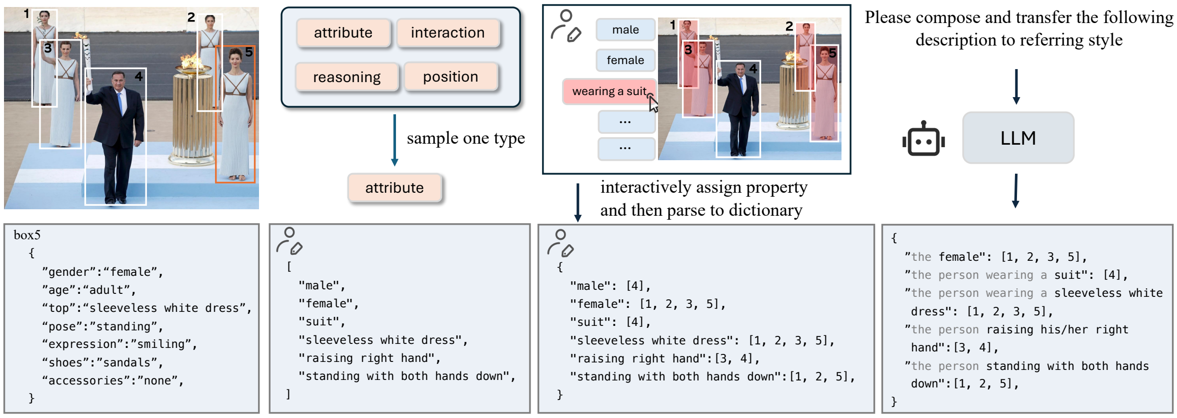
In this section, we present the design philosophy, data acquisition process, annotation pipeline, and dataset statistics of the proposed HumanRef dataset.

3.1. Data Design Philosophy

We define five key aspects that determine how humans can be referred to using natural language, including attribute, position, interaction, reasoning, and celebrity recognition. These categories are further elaborated with definitions and examples in Table 1. A key distinction between HumanRef and existing referring datasets is its focus on multi-instance referring rather than one-to-one object referring. Our dataset ensures that a single referring expression can correspond to multiple individuals, providing a more realistic and practical reflection of real-world scenarios.

3.2. Data Acquisition

The HumanRef dataset is designed to capture human presence across diverse contexts, including natural environments, industrial settings, healthcare, sports, films, animations, etc. To ensure dataset diversity, we sourced images



a) pseudo labeling b) write property list c) assign property to each person d) transfer to referring style with LLM

Figure 3. Overview of the manual annotation pipeline of the HumanRef dataset.

containing humans from the web image dataset [5]. To filter candidate images, we first retained those with a resolution larger than 1000 × 1000 pixels to ensure high-quality content. Next, we use an open-set object detector DINO-X [60] to detect human instances. To align with the multi-instance discrimination requirement, we retain only images containing at least four individuals.

To assist the annotator in writing properties, we prompt the QwenVL-2.5 [3] model to create a structured property dictionary for each person in the image, capturing details such as gender, clothing, actions, etc. Ultimately, this phase produced image, person box, and person description triples used for further annotation.

3.3. Manual Annotation

For attribute, position, interaction, and reasoning subsets, we adopt manual annotation. This annotation process consists of three main steps: property listing, property assignment, and referring style rewriting. Given an image, along with the corresponding person boxes and pre-labeled property dictionary, the annotation system will randomly select one annotation type from attribute, location, interaction, and reasoning to assign to the annotator. The following annotation process is then carried out:

Property Listing: The annotator examines all individuals in the image, considering both their visual appearance, action, position, interaction, and the pre-labeled property dictionary. Based on these observations, the annotator compiles a list of properties. To enhance dataset richness, annotators are encouraged to label attributes shared by multiple individuals while avoiding those common to all. Additionally, we monitor the word frequency of labeled referring expressions and restrict the use of high-frequency words to improve data diversity.

Property Assignment: Once the properties are listed, annotators systematically assign them to the corresponding individuals. This interactive process involves selecting a property value and clicking on the associated bounding boxes to link it to the correct person. The final output is a structured dictionary, where keys represent property names and values contain lists of bounding box indices corresponding to the individuals possessing each property.

Referring Style Rewriting: In the final step, we prompt Qwen2.5-14B [71] to reformulate the structured attribute dictionary into short, natural language referring expressions. The final annotated data also undergoes a thorough review process to ensure its quality.

3.4. Automatic Annotation

For celebrity recognition and rejection referring, we employ two efficient and effective automatic annotation pipelines.

Celebrity Recognition: We first categorize celebrities into seven distinct fields: actors, film characters, athletes, singers, entrepreneurs, scientists, and politicians. For each field, we identify the most well-known individuals, compiling a final list of 636 names, which we then used as prompts to retrieve images via the Bing Search API. The collected images include both individual and group photos, necessitating a method to accurately associate each celebrity name with the correct person in the image. To achieve this, we first use the DINO-X [60] model to detect all human faces and persons, linking each detected face to its corresponding person box based on overlap measurements. If an image contains only one person, we assume this individual is the target celebrity. For images featuring multiple individuals, we use a Python face recognition library, leveraging a single-person image as a recognition template to match and identify the same person in such images.

Rejection Referring: The objective of this sub-dataset is to ensure that when a referring description targets a per-

https://www.microsoft.com/en-us/bing/apis
https://github.com/ageitgey/facerecognition

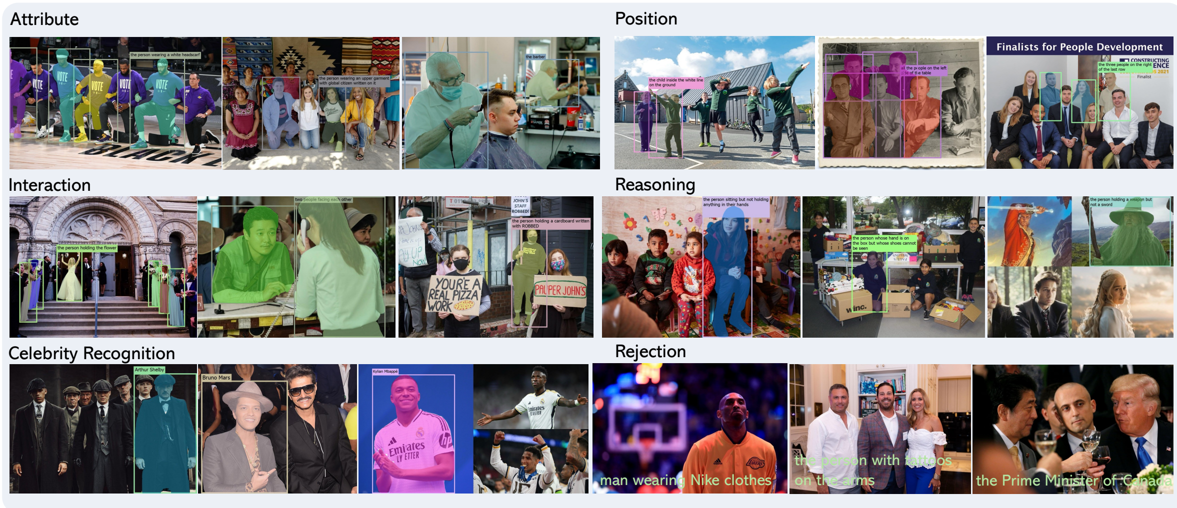


Figure 4. Visualization of the six subsets in the HumanRef Benchmark.

son who does not exist in the input image, the model rejects the referring request instead of hallucinating and outputting an incorrect bounding box. To construct this dataset, we first extract referring expressions from the attribute, position, interaction, and reasoning subsets. We then prompt Qwen2.5 [71] to modify these descriptions, transforming them into similar but semantically altered versions. For instance, a description such as *"the person wearing a blue hat"* may be changed to *"the person wearing a red hat"*. To validate the generated descriptions, we prompt Molmo [20] to detect the modified referring expression. If no matching object is found in the output, the data is retained.

3.5. HumanRef Benchmark

To construct the HumanRef Benchmark, we sample 1,000 referring expressions from each of the four manually annotated subsets. Additionally, for the celebrity and rejection subsets, we conduct a separate manual annotation process to create 1,000 new referring expressions for each category, ensuring high-quality and challenging evaluation data. To further support advancements in referring expression segmentation, we utilize SAM2 [59] to generate masks for each ground truth bounding box. Figure 4 presents example cases from the HumanRef Benchmark, illustrating the diversity and complexity of the dataset.

3.6. Statistics

We first present the basic statistics of the HumanRef dataset and its subsets in Table 2, and then illustrate the characteristics of multi-instance referring and multi-instance discrimination in HumanRef in Figure 5. Additionally, Table 3 compares the HumanRef Benchmark with widely used referring benchmarks, including RefCOCO, RefCOCO+, and RefCOCOg. A key distinction of HumanRef is its higher image resolution and larger number of individuals per image, requiring models to precisely identify all correct individuals among multiple people. Unlike traditional benchmarks, where each referring expression corresponds to a

HumanRef Train							
type	attribute	position	interaction	reasoning	celebrity	rejection	total
images	8,614	7,577	1,632	4,474	4,990	7,519	34,806
referrings	52,513	22,496	2,911	6,808	4,990	13,310	103,028
avg. boxes/ref	2.9	1.9	3.1	3.0	1.0	0	2.2
HumanRef Benchmark							
type	attribute	position	interaction	reasoning	celebrity	rejection	total
images	838	972	940	982	1,000	1,000	5,732
referrings	1,000	1,000	1,000	1,000	1,000	1,000	6,000
avg. boxes/ref	2.8	2.1	2.1	2.7	1.1	0	2.2

Table 2. Main statistics of the HumanRef dataset, including the number of images, the number of referring expressions, the average word count per referring expression, and the average number of instances associated with each referring expression.

Datasets	images	refs	vocabs	avg. size	avg. person/image	avg. words/ref	avg. boxes/ref
RefCOCO [75]	1,519	10,771	1,874	593x484	5.72	3.43	1
RefCOCO+ [75]	1,519	10,908	2,288	592x484	5.72	3.34	1
RefCOCOg [50]	1,521	5,253	2,479	585x480	2.73	9.07	1
HumanRef	5,732	6,000	2,714	1432x1074	8.60	6.69	2.2

Table 3. Comparison of the HumanRef Benchmark with RefCOCO/+/g. For a fair comparison, we present only the statistics related to human referring in RefCOCO/+/g.

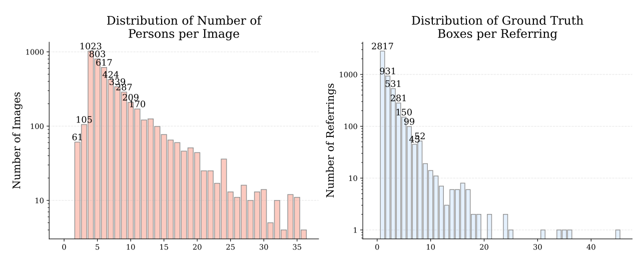


Figure 5. Distribution of the number of individuals per image and the number of individuals referenced by each referring expression.

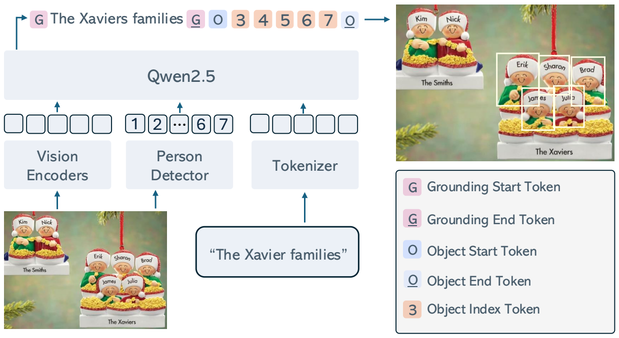


Figure 6. Overview of the RexSeek model. RexSeek is a retrieval-based model built upon ChatRex [25]. By integrating a person detection model, RexSeek transforms the referring task from predicting box coordinates to retrieving the index of input boxes.

single person, HumanRef supports multi-instance referring, offering a more realistic and challenging evaluation setting for referring expression comprehension.

4. RexSeek Model

4.1. Model Design Philosophy

From a model design perspective, we argue that a robust referring model should have two essential capabilities: **i) robust perception ability**, where the model can reliably detect all individuals in an image, and **ii) strong language comprehension**, where the model can accurately interpret complex natural language descriptions of people.

For the first capability, modern object detection models [26, 41, 60, 61] are highly effective at identifying people within images. However, these models often lack the necessary language comprehension abilities to process intricate and nuanced referring expressions. On the other hand, while MLLMs are proficient in understanding natural language, they often struggle with fine-grained object detection tasks. Inspired by ChatRex [25], we propose a hybrid framework, RexSeek, which integrates the strengths of both object detection models and LLMs. RexSeek combines a high-performance detection model with a multimodal LLM to achieve both accurate detection and effective language understanding.

4.2. Architecture

Following ChatRex, we formulate the referring task as a retrieval-based process [25, 49]. As illustrated in Figure 6, RexSeek consists of three main components: vision encoders, a person detector, and a large language model. Given an input image, we first pass it through a dual vision encoder module used in ChatRex. This module consists of a CLIP [58] to extract low-resolution image features $\mathcal{F}_{\text{low}}$ and a ConvNeXt [43] to extract high-resolution image features $\mathcal{F}_{\text{high}}$. We adjust the input resolutions for both vision encoders to ensure they generate the same number of tokens at the last scale. The final vision tokens $\mathcal{F}$ is obtained by concatenating these features at the channel dimension:

$$\mathcal{F} = \text{Concat}(\mathcal{F}_{\text{low}}, \mathcal{F}_{\text{high}})$$

Next, we prompt DINO-X [60] to get the bounding boxes of persons $\{B_i\}_{i=1}^{K}$ in the image. For each bounding box, we extract its RoI features $\mathcal{C}_i$ and add their positional embeddings to generate object tokens $\mathcal{O}_i$, which capture both the content and spatial context of each detected person:

$$\mathcal{O}_i = \mathcal{C}_i + \text{PE}(B_i)$$

Specifically, the RoI feature is extracted from the high-resolution vision features using a multi-scale RoI Align operation [24]. The positional embedding is computed by encoding the bounding box coordinates (x, y, w, h) using a sinusoidal encoding function and concatenating the encoded values along the channel dimension.

Finally, the vision tokens $\mathcal{F}$, object tokens $\mathcal{O}$, and text tokens $\mathcal{T}$ are projected using different MLPs and then fed into the LLM. By default, we use Qwen2.5 [71] as the LLM. The LLM decodes the input to produce the corresponding box indices $\mathcal{I}$:

$$\mathcal{I} = \text{LLM}(\mathcal{F}, \mathcal{O}, \mathcal{T})$$

The output $\mathcal{I}$ consists of object indices that correspond to the bounding boxes of the target persons corresponding to the referring. This sequence is structured as follows:

```
<g>referring</g><o><objm>...<objn></o>
```

Here, <objm> and <objn> refer to specific object index tokens that correspond to the detected persons. The special tokens <g>, </g>, <o>, and </o> are used to format the output, linking the referring expression with the relevant object indices.

4.3. Four Stage Training

Similar to other VLMs, we adopt a pretraining followed by supervised fine-tuning approach [39]. Our training process consists of four stages. In the first stage, we align the visual and textual modalities using image-captioning data. In the second stage, we focus on perception training with detection-oriented data, enabling the model to retrieve relevant objects from input bounding boxes. In the third stage, we incorporate multimodal data to enhance the model's general understanding abilities. Finally, in the fourth stage, we fine-tune the model using the HumanRef dataset, resulting in the final RexSeek model. The data, task, and trainable modules for each stage are shown in Table 5.

5. Experiments

In this section, we first introduce the evaluation metrics used in our study and assess the performance of multimodal models on HumanRef. We perform a comprehensive analysis to

Method	Attribute			Position			Interaction			Reasoning			Celebrity			Average			Rejection
	R	P	DF1	R	P	DF1	R	P	DF1	R	P	DF1	R	P	DF1	R	P	DF1	Score
Baseline†	100.0	37.2	24.2	100.0	28.5	15.9	100.0	32.5	19.4	100.0	42.6	30.3	100.0	14.4	4.9	100.0	31.0	18.9	0.0
DINOX [60]	59.5	28.8	20.9	78.8	28.1	17.6	67.3	28.5	18.9	76.2	32.1	22.2	94.1	48.0	37.0	75.2	33.1	23.3	36.0
InternVL-2.5-8B [14]	23.5	39.0	27.1	23.0	28.0	24.3	27.8	40.1	31.3	17.5	22.8	18.9	57.4	59.3	58.0	29.8	37.8	31.9	54.9
Ferret-7B [74]	27.9	44.4	30.4	30.2	36.2	29.8	30.8	41.8	31.2	19.7	33.7	22.8	63.2	60.0	57.5	34.4	43.2	34.3	2.0
Groma-7B [49]	67.5	47.8	38.6	63.2	43.1	37.2	66.6	48.1	40.6	59.1	41.4	34.8	73.2	63.3	59.1	65.9	48.7	42.1	0.0
ChatRex-7B [25]	44.3	78.0	51.8	48.0	66.7	52.5	49.6	74.8	56.5	36.6	65.1	42.8	73.7	76.5	74.2	50.4	72.2	55.6	0.0
Qwen2.5-VL-7B [3]	49.1	71.3	54.4	50.2	61.7	52.8	48.2	66.3	53.2	34.6	61.2	40.3	80.3	81.9	80.1	52.5	68.5	56.2	7.1
DeepSeek-VL2-small [70]	52.3	78.0	57.7	56.4	66.1	58.1	55.4	75.7	60.7	46.6	61.7	50.1	85.9	74.3	70.7	59.3	71.2	59.5	3.1
Molmo-7B-D* [20]	82.7	86.4	76.3	78.0	80.6	72.4	69.9	77.7	66.1	72.1	80.4	65.5	85.9	87.5	82.9	77.7	82.5	72.6	68.6
RexSeek-7B	87.2	86.8	81.5	86.1	86.3	83.8	84.8	84.6	80.7	87.8	84.7	81.5	83.4	86.5	84.2	85.9	85.8	82.3	54.1

Table 4. Benchmarking multimodal models on HumanRef Benchmark. R, P, and DF1 represent Recall, Precision, and DensityF1, respectively. † A simple baseline that uses the bounding boxes of all persons in the image as results, simulating a person detection model that does not follow the referring description. *Molmo-7B-D predicts point coordinates as output and use point-in-mask evaluation criteria.

Stage	Trainable Modules	Task	# Samples	Datasets
Stage1	MLPs	Image Captioning	976K	ALLAVA-4V-Caption [6]
Stage2	MLPs + LLM + Vision Encoders	Grounding & Region Understanding	2.07M	COCO [37], LVIS [22], O365 [62], Rexverse-2M [25]
Stage3	MLPs + LLM + Vision Encoders	General Knowledge & Grounding & Region Understanding	2.15M	LLAVA-665K [38] Rexverse-2M [25]
Stage4	MLPs + LLM + Vision Encoders	Referring	103K	HumanRef

Table 5. Data, task, and trainable modules for each stage.

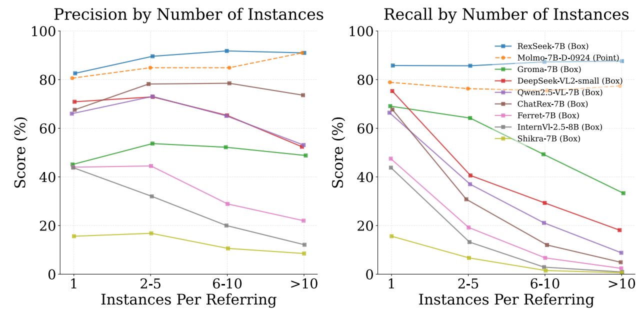


Figure 7. Visualizing the trend of recall and precision variations across different models as the number of instances corresponding to each referring expression increases.

explore the challenges faced by existing models in handling the referring task. Additionally, we perform ablation experiments on RexSeek for model design choices.

5.1. Metrics

We evaluate the referring task using Precision, Recall, and DensityF1 Score. Given a referring expression, the model predicts one or more bounding boxes, and a prediction is considered correct if its IoU with any ground truth box exceeds a predefined threshold. Following the evaluation protocol in COCO [37], we report the average performance across IoU thresholds from 0.5 to 0.95 in increments of 0.05. For models that only output points, such as Molmo [20], a prediction is considered correct if the predicted point falls within the mask of the corresponding instance. However, this evaluation is less strict than the IoU-based metric, as point-in-mask criteria impose looser spatial constraints, making direct comparisons less fair. For the rejection subset, we calculate the number of referring expressions that the model does not predict any boxes and divide it by the number of total expressions.

To penalize models that indiscriminately detect all persons in an image to achieve a high F1 score through high recall, we introduce the DensityF1 Score, which modifies the standard F1 Score with a density-aware penalty:

$$\text{DensityF1} = \frac{1}{N}\sum_{i=1}^{N} 2 \times \frac{\text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \times D_i \quad (1)$$

where D_i is the density penalty factor, defined as:

$$D_i = \min(1.0, \frac{\text{GT Count}_i}{\text{Predicted Count}_i}) \quad (2)$$

Here, GT Count is the total number of persons in an image, and Predicted Count is the number of predicted boxes for a given referring expression. This penalty discourages over-detection by reducing the score when the predicted box count significantly exceeds the ground truth count.

5.2. Benchmarking on HumanRef

In Table 4, we evaluate the performance of various multimodal models on the HumanRef benchmark. While these models perform well on the widely used RefCOCO, RefCOCO+, and RefCOCOg benchmarks, their performance significantly degrades on HumanRef. Our analysis reveals two common issues among these models:

Low Recall for Multi Instance: We observe a common issue among most models: when a referring expression corresponds to multiple instances, recall drops significantly, as shown in Figure 7. This suggests that when multiple objects need to be detected, most models tend to predict only a few bounding boxes, limiting their applicability in real-world scenarios. A key factor contributing to this behavior is the nature of the training data. Most multimodal models are trained on RefCOCO, RefCOCO+, and RefCOCOg,

Model	With Rejection Data	Rejction Score
RexSeek-7B	No	0
RexSeek-7B	Yes	541

Table 6. Rejection score comparison under different model scales with and without rejection data during training.

Loading Stage	HumanRef Average		
	R	P	DF1
stage1	73.9	73.5	68.2
stage2	77.0	77.3	72.2
stage3	77.9	78.0	73.0

Table 7. Ablation experiments on multi-stage training by loading models from different training stages and fine-tuning them on the HumanRef dataset. We Qwen2.5-3B as the base LLM.

Method	RefCOCOg	
	val	test
Shikra-7B [9]	82.3	82.2
InternVL2-8B [14]	82.7	82.7
Grounding DINO-L [42]	86.1	87.0
Qwen2.5-VL-7B [3]	87.2	87.2
MM1.5-7B [82]	-	87.1
ChatRex-7B [25]	88.8	88.6
RexSeek-7B	84.0	84.4

Table 8. Zero-shot evaluation of RexSeek on RefCOCOg. We use the open-set detector DINOX to detect the subject object in the image and use the detected bounding box as input to RexSeek.

where referring expressions rarely correspond to multiple instances. As a result, these models become biased toward single-instance predictions. In contrast, RexSeek has been trained on datasets that explicitly include multi-instance referring expressions, demonstrate a significantly improved ability to handle these real-world cases.

Hallucination Issue: On the rejection subset, we observe that most models perform poorly with low rejection score. This indicates that regardless of whether the referred object is actually present in the image, these models tend to predict a bounding box, exhibiting a severe hallucination issue. In real-world referring applications, such as referring in video streams, it is crucial for models to accurately determine whether the specified object exists in the image. Additionally, we find that the rejection capability can be significantly improved by incorporating appropriate training data. As shown in Table 6, when trained without the rejection data in HumanRef, RexSeek also demonstrates strong hallucination tendencies. This highlights the critical role of dataset design in the referring task, as inadequate dataset construction can lead to overconfident predictions.

5.3. Ablations on RexSeek

Ablation of Multi-stage Training: We analyzed the impact of the four-stage training approach used in RexSeek. As shown in Table 7, we conducted supervised fine-tuning on the HumanRef dataset after each training stage. The results demonstrate that the model achieves its best performance after undergoing SFT with general multimodal data (LLaVA-665K [38]). We attribute this improvement to the model acquiring richer general knowledge from multimodal data, which enhances its ability to accurately refer to persons in complex scenarios.

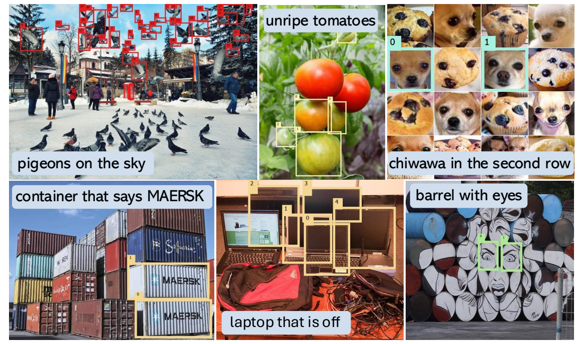


Figure 8. RexSeek can refer to arbitrary objects beyond person.

Generalization to Any Object Referring: Although RexSeek is trained exclusively on human-related referring data, we find that it also demonstrates the ability to refer to arbitrary objects. We first evaluate the performance of RexSeek on RefCOCOg. Given a referring expressions, we apply DINO-X to detect the main object in the image, using the detected bounding box as input to RexSeek. As shown in Table 8, RexSeek achieves competitive performance on RefCOCO/+/g, despite not being explicitly trained on general object referring. Additionally, Figure 8 presents visualizations illustrating that RexSeek can also detect multiple instances even for non-human objects. We attribute this generalization ability to our multi-stage training approach, where perception and multimodal understanding training develop object comprehension, and fine-tuning on HumanRef effectively extends it to arbitrary objects.

6. Conclusion

In this work, we identify the fundamental limitations of existing referring datasets and models, demonstrating that they fail to meet real-world application demands, particularly in multi-instance referring. To address this, we introduce HumanRef, a large-scale benchmark reflecting real-world complexity, and propose RexSeek, a retrieval-based detection MLLM integrating person detection with a language model. Our multi-stage training approach equips RexSeek with strong generalization capabilities, allowing it to excel in human-centric referring while extending effectively to arbitrary object referring. Extensive evaluations highlight the struggles of state-of-the-art models with multi-instance detection and hallucination, underscoring the importance of dataset design and training strategies for more reliable and generalizable referring expression models.

References

[1] Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Devendra Chaplot, Jessica Chudnovsky, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, et al. Pixtral 12b. *arXiv preprint arXiv:2410.07073*, 2024. 3

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *NeurIPS*, 35: 23716–23736, 2022.

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 3, 4, 7, 8

[4] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 561–578. Springer, 2016. 2

[5] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. `https://github.com/kakaobrain/coyo-dataset`, 2022. 4

[6] Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv preprint arXiv:2402.11684*, 2024. 7

[7] Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12538–12547, 2019. 3

[8] Jierun Chen, Fangyun Wei, Jinjing Zhao, Sizhe Song, Bohuai Wu, Zhuoxuan Peng, S-H Gary Chan, and Hongyang Zhang. Revisiting referring expression comprehension evaluation in the era of large multimodal models. *arXiv preprint arXiv:2406.16866*, 2024. 3

[9] 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. 2, 3, 8

[10] Ling-Hao Chen, Jiawei Zhang, Yewen Li, Yiren Pang, Xiaobo Xia, and Tongliang Liu. Humanmac: Masked motion completion for human motion prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9544–9555, 2023. 2

[11] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 403–412, 2017.

[12] Weihua Chen, Xianzhe Xu, Jian Jia, Hao Luo, Yaohua Wang, Fan Wang, Rong Jin, and Xiuyu Sun. Beyond appearance: a semantic controllable self-supervised learning framework for human-centric visual tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15050–15061, 2023. 2

[13] Zhenfang Chen, Peng Wang, Lin Ma, Kwan-Yee K Wong, and Qi Wu. Cops-ref: A new dataset and task on compositional referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10086–10095, 2020. 3

[14] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 2, 3, 7, 8

[15] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16901–16911, 2024. 3

[16] Yuanzheng Ci, Yizhou Wang, Meilin Chen, Shixiang Tang, Lei Bai, Feng Zhu, Rui Zhao, Fengwei Yu, Donglian Qi, and Wanli Ouyang. Unihcp: A unified model for human-centric perceptions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17840–17852, 2023. 2

[17] Volkan Cirik, Taylor Berg-Kirkpatrick, and Louis-Philippe Morency. Refer360: A referring expression recognition dataset in 360 images. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7189–7202, 2020. 3

[18] Samyak Datta, Karan Sikka, Anirban Roy, Karuna Ahuja, Devi Parikh, and Ajay Divakaran. Align2ground: Weakly supervised phrase grounding guided by image-caption alignment. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2601–2610, 2019. 3

[19] Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5503–5512, 2017. 3

[20] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024. 3, 5, 7

[21] Riza Alp Guler and Iasonas Kokkinos. Holopose: Holistic 3d human reconstruction in-the-wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10884–10894, 2019. 2

[22] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 7

[23] Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive learning for

weakly supervised phrase grounding. In *European Conference on Computer Vision*, pages 752–768. Springer, 2020. 3

[24] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 6

[25] Qing Jiang, Yuqin Yang, Yuda Xiong, Yihao Chen, Zhaoyang Zeng, Tianhe Ren, Lei Zhang, et al. Chatrex: Taming multimodal llm for joint perception and understanding. *arXiv preprint arXiv:2411.18363*, 2024. 2, 3, 6, 7, 8

[26] Qing Jiang, Feng Li, Zhaoyang Zeng, Tianhe Ren, Shilong Liu, and Lei Zhang. T-rex2: Towards generic object detection via text-visual prompt synergy. In *European Conference on Computer Vision*, pages 38–57. Springer, 2025. 3, 6

[27] Xuan Ju, Ailing Zeng, Jianan Wang, Qiang Xu, and Lei Zhang. Human-art: A versatile human-centric dataset bridging natural and artificial scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 618–629, 2023. 2

[28] Xuan Ju, Ailing Zeng, Chenchen Zhao, Jianan Wang, Lei Zhang, and Qiang Xu. Humansd: A native skeleton-guided diffusion model for human image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15988–15998, 2023. 2

[29] 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. 2, 3

[30] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. In *European Conference on Computer Vision*, pages 206–228. Springer, 2024. 2

[31] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123:32–73, 2017. 3

[32] Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, Zhiqi Shen, Bowen Qu, Xinyao Niu, Guoyin Wang, Bei Chen, and Junnan Li. Aria: An open multimodal native mixture-of-experts model. *arXiv preprint arXiv:2410.05993*, 2024. 3

[33] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024. 3

[34] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 3

[35] Zhiqi Li, Guo Chen, Shilong Liu, Shihao Wang, Vibashan VS, Yishen Ji, Shiyi Lan, Hao Zhang, Yilin Zhao, Subhashree Radhakrishnan, et al. Eagle 2: Building post-training data strategies from scratch for frontier vision-language models. *arXiv preprint arXiv:2501.14818*, 2025. 3

[36] Yue Liao, Si Liu, Guanbin Li, Fei Wang, Yanjie Chen, Chen Qian, and Bo Li. A real-time cross-modality correlation filtering method for referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10880–10889, 2020. 3

[37] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, pages 740–755, 2014. 3, 7

[38] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv: 2310.03744*, 2023. 7, 8

[39] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 6

[40] Runtao Liu, Chenxi Liu, Yutong Bai, and Alan L Yuille. Clevr-ref+: Diagnosing visual reasoning with referring expressions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4185–4194, 2019. 3

[41] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 6

[42] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024. 3, 8

[43] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, pages 11976–11986, 2022. 6

[44] Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, et al. Nvila: Efficient frontier visual language models. *arXiv preprint arXiv:2412.04468*, 2024. 3

[45] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 2

[46] Shunlin Lu, Ling-Hao Chen, Ailing Zeng, Jing Lin, Ruimao Zhang, Lei Zhang, and Heung-Yeung Shum. Humantomato: Text-aligned whole-body motion generation. *arXiv preprint arXiv:2310.12978*, 2023. 2

[47] Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797*, 2024. 3

[48] 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. 3

[49] Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. Groma: Localized visual tokenization for grounding multimodal large language models. *arXiv preprint arXiv:2404.13013*, 2024. 2, 3, 6, 7

[50] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, pages 11–20, 2016. 2, 3, 5

[51] Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, Anton Belyi, Haotian Zhang, Karanjeet Singh, Doug Kang, Ankur Jain, Hongyu Hè, Max Schwarzer, Tom Gunter, Xiang Kong, Aonan Zhang, Jianyu Wang, Chong Wang, Nan Du, Tao Lei, Sam Wiseman, Guoli Yin, Mark Lee, Zirui Wang, Ruoming Pang, Peter Grasch, Alexander Toshev, and Yinfei Yang. MM1: methods, analysis & insights from multimodal LLM pretraining. *arXiv: 2403.09611*, 2024. 3

[52] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *European conference on computer vision*, pages 728–755. Springer, 2022. 3

[53] OpenAI. Gpt-4v(ision) system card. `https://cdn.openai.com/papers/GPTV_System_Card.pdf`, 2023. 3

[54] 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. 3

[55] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9982–9991, 2020. 3

[56] Yanyuan Qiao, Chaorui Deng, and Qi Wu. Referring expression comprehension: A survey of methods and datasets. *IEEE Transactions on Multimedia*, 23:4426–4440, 2020. 3

[57] Heqian Qiu, Hongliang Li, Taijin Zhao, Lanxiao Wang, Qingbo Wu, and Fanman Meng. Refcrowd: Grounding the target in crowd with referring expressions. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4435–4444, 2022. 3

[58] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 6

[59] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 5

[60] Tianhe Ren, Yihao Chen, Qing Jiang, Zhaoyang Zeng, Yuda Xiong, Wenlong Liu, Zhengyu Ma, Junyi Shen, Yuan Gao, Xiaoke Jiang, et al. Dino-x: A unified vision model for open-world object detection and understanding. *arXiv preprint arXiv:2411.14347*, 2024. 2, 3, 4, 6, 7

[61] Tianhe Ren, Qing Jiang, Shilong Liu, Zhaoyang Zeng, Wenlong Liu, Han Gao, Hongjie Huang, Zhengyu Ma, Xiaoke Jiang, Yihao Chen, et al. Grounding dino 1.5: Advance the" edge" of open-set object detection. *arXiv preprint arXiv:2405.10300*, 2024. 3, 6

[62] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019. 7

[63] Shixiang Tang, Cheng Chen, Qingsong Xie, Meilin Chen, Yizhou Wang, Yuanzheng Ci, Lei Bai, Feng Zhu, Haiyang Yang, Li Yi, et al. Humanbench: Towards general human-centric perception with projector assisted pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21970–21982, 2023. 2

[64] Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024. 3

[65] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1653–1660, 2014. 2

[66] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 3

[67] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023. 3

[68] Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. Phrasecut: Language-based image segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10216–10225, 2020. 3

[69] Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. Aligning bag of regions for open-vocabulary object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15254–15264, 2023. 3

[70] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024. 2, 3, 7

[71] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang,

Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. 2, 4, 5, 6

[72] Sibei 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. 3

[73] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. *Advances in Neural Information Processing Systems*, 35:9125–9138, 2022. 3

[74] 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. 2, 3, 7

[75] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In *ECCV*, pages 69–85, 2016. 2, 3, 5

[76] 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. 3

[77] Zhihan Yu and Ruifan Li. Revisiting counterfactual problems in referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13438–13448, 2024. 3

[78] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14393–14402, 2021. 3

[79] Yufei Zhan, Yousong Zhu, Hongyin Zhao, Fan Yang, Ming Tang, and Jinqiao Wang. Griffon v2: Advancing multimodal perception with high-resolution scaling and visual-language co-referring. *arXiv preprint arXiv:2403.09333*, 2024. 3

[80] 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. 3

[81] Chao Zhang, Weiming Li, Wanli Ouyang, Qiang Wang, Woo-Shik Kim, and Sunghoon Hong. Referring expression comprehension with semantic visual relationship and word mapping. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1258–1266, 2019. 3

[82] Haotian Zhang, Mingfei Gao, Zhe Gan, Philipp Dufter, Nina Wenzel, Forrest Huang, Dhruti Shah, Xianzhi Du, Bowen Zhang, Yanghao Li, et al. Mm1. 5: Methods, analysis & insights from multimodal llm fine-tuning. *arXiv preprint arXiv:2409.20566*, 2024. 3, 8

[83] Haotian Zhang, Haoxuan You, Philipp Dufter, Bowen Zhang, Chen Chen, Hong-You Chen, Tsu-Jui Fu, William Yang Wang, Shih-Fu Chang, Zhe Gan, et al. Ferret-v2: An improved baseline for referring and grounding with large language models. *arXiv preprint arXiv:2404.07973*, 2024. 3

[84] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1367–1376, 2017. 2

[85] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7739–7749, 2019. 2