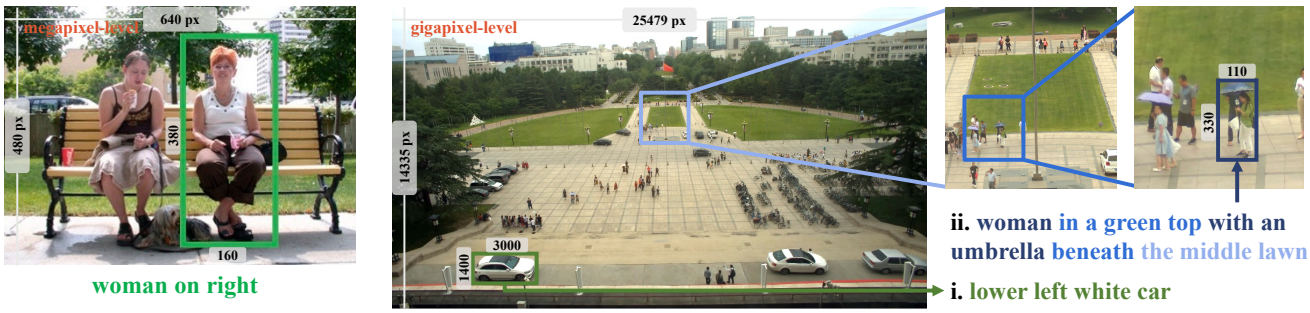
When Visual Grounding Meets Gigapixel-level Large-scale Scenes: Benchmark and Approach

Tao Ma[1*], Bing Bai[1*], Haozhe Lin[1*], Heyuan Wang[2], Yu Wang[1], Lin Luo[2], Lu Fang[1†]
[1]Tsinghua University [2]Peking University
fanglu@tsinghua.edu.cn



(a) RefCOCO (b) GigaGrounding

Figure 1. Comparison of the proposed GigaGrounding and the RefCOCO benchmark. GigaGrounding presents distinctive challenges, including large-scale scene understanding, high-resolution with significant scale variation, and multi-hop expressions.

Abstract

Visual grounding refers to the process of associating natural language expressions with corresponding regions within an image. Existing benchmarks for visual grounding primarily operate within small-scale scenes with a few objects. Nevertheless, recent advances in imaging technology have enabled the acquisition of gigapixel-level images, providing high-resolution details in large-scale scenes containing numerous objects. To bridge this gap between imaging and computer vision benchmarks and make grounding more practically valuable, we introduce a novel dataset, named GigaGrounding, designed to challenge visual grounding models in gigapixel-level large-scale scenes. We extensively analyze and compare the dataset with existing benchmarks, demonstrating that GigaGrounding presents unique challenges such as large-scale scene understanding, gigapixel-level resolution, significant variations in object scales, and the "multi-hop expressions". Furthermore, we introduced a simple yet effective grounding approach, which employs a "glance-to-zoom-in" paradigm and exhibits enhanced capabilities for addressing the GigaGrounding task. The dataset is available at www.gigavision.ai.

1. Introduction

Visual grounding (VG), also known as referring expression comprehension [1, 2], phrase localization [3, 4], and natural language object retrieval [5, 6], aims to identify the region within an image that corresponds to a given natural language expression. This field offers considerable potential for enhancing our capacity to connect the human language and visual elements of the tangible world [7]. The application of VG in scenarios such as rescue and robotics can prove to be exceedingly valuable, for example, in searching for a missing child in crowds.

While the VG community has made significant progress in recent years, the benchmarks employed thus far have been predominantly limited to small-scale scenes [1–4]. As illustrated in Figure 1a, the number of objects is limited, and the objects cover the dominant portion of the image. However, recent advances in imaging technology, such as array cameras [8] and even consumer-grade smartphones [9], have enabled the capture of gigapixel-level photos, providing high-resolution detail information while covering large-scale scenes up to km^2-level [10], as shown in Figure 1b. The characteristics of conventional VG benchmarks have restricted current models' application to relatively simple and vanilla settings, hindering the development of new technologies and limiting the applicability of grounding in broader contexts [11].

*These authors contributed equally to this work.
†Lu Fang is the corresponding author (www.luvision.net).

To address this limitation, we introduce a novel benchmark dataset, named GigaGrounding, for visual grounding in large-scale scenes. The dataset is constructed based on PANDA [11], a gigapixel-level human-centric video dataset. An illustration of the GigaGrounding dataset is displayed in Figure 1b. In contrast to conventional visual grounding benchmarks such as RefCOCO [2], GigaGrounding poses unique and non-trivial challenges:

- **Large-scale scene understanding**. Large-scale scenes require models to deal with a much larger spatial and semantic context and a significantly larger number of objects in the image. Hundreds of objects can be distributed over a vast area with possible occlusions, clutter, and other distractions, making it a difficult task to recognize and understand large-scale scenes.
- **High resolution with significant scale variation**. Processing high-throughput data up to gigapixel-level images poses new challenges to models in terms of efficiency [10]. Moreover, objects may be presented at drastically different scales. For instance, as shown in Figure 1b, the ground-truth bounding box (B-box) for *Case i* spans approximately 3000×1400 pixels, whereas, for *Case ii*, it spans only about 110×330 pixels. This significant variation necessitates models to be adaptable in scaling to accommodate diverse object sizes.
- **Multi-hop expressions**. In addition to *direct expressions*, which refer to descriptions of independent objects as illustrated in *Case i*, GigaGrounding also includes a considerable number of *multi-hop expressions*. These expressions challenge grounding models to locate target objects by first identifying some other *reference objects*. For example, in *Case ii*, the model needs first to identify "the middle lawn" and then zoom in to locate the correct person based on the relationship with the lawn.

We benchmark the performance of cutting-edge VG models with the GigaGrounding dataset. Empirical findings reveal that while existing VG models have demonstrated satisfactory performance in small-scale scenes, they face distinct and intricate challenges when being applied to GigaGrounding. To address these challenges, we propose GlaZing (GLAnce-to-Zoom-IN Grounding), a simple yet effective model inspired by human cognitive processes. GlaZing employs a "glance-to-zoom-in" cascading strategy to ground target bounding boxes in high-resolution images. It begins by taking a global view of the thumbnail image, to identify the region of interest based on the given expression. Subsequently, it zooms in on the local region with adaptive scaling for the final grounding process. Experiments demonstrated that GlaZing outperformed existing models in the GigaGrounding dataset, effectively addressing major challenges such as high resolution, significant scale variation, and multi-hop expressions.

In conclusion, this paper presents several contributions:

- We highlight the significance of deploying visual grounding models in large-scale scenes, and introduce a novel benchmark dataset, GigaGrounding, characterized by unique properties that set it apart from existing datasets.
- We empirically assess current VG models using the proposed GigaGrounding dataset, highlighting the limitations of state-of-the-art models in this context.
- We introduce GlaZing, a tailored model for GigaGrounding, which leverages a "glance-to-zoom-in" approach. Our experimental results demonstrate the efficacy of this approach in addressing the current challenges, while also indicating room for further improvements.

2. Related Work

In this section, we review two relevant research domains: visual grounding and high-resolution deep learning.

2.1. Visual Grounding

Visual grounding is a field focused on predicting the location within an image that corresponds to a natural language expression. Benchmark datasets for VG have predominantly centered on small-scale scenes. For example, the Flickr30K Entities dataset [4] extends the original Flickr30K dataset [12] by incorporating annotations that establish correspondence between short region phrases and images. ReferItGame [3] comprises 20,000 images collected from the SAIAPR-12 dataset [13]. RefCOCO [2], RefCOCO+ [2], and RefCOCOg [1] are all built upon the MS COCO dataset [14]. Notably, the image resolutions in these datasets do not exceed 640×480. In addition, there are related benchmarks in fields such as remote sensing [15, 16] and biomedical science [17, 18], with their resolutions also limited to 1024×1024 or lower.

For VG models, existing work can be broadly categorized into two-stage and one-stage methods [7, 19]. Two-stage methods [20–24] initiate the process by generating a set of region proposals using object detectors. Subsequently, they leverage natural language expressions to rank these proposals and identify the best matching proposal. In contrast, one-stage methods [7, 19, 25–29] densely integrate both visual and textual information and directly output bounding boxes.

2.2. High-resolution Deep Learning

In recent years, significant progress has been made in the field of imaging, leading to gigapixel-level images with both a wide FoV and high resolution [8, 30, 31]. Employing high-resolution images and videos as direct inputs for deep learning models introduces challenges in both training and inference stages [11, 32–34]. To address these

Dataset	Typical res.	# Images	# B-boxes	# Expressions
Flickr30K Entities [4]	500×375	31,783	276k	427k
ReferItGame [3]	480×360	19,894	96,654	130,525
RefCOCO [2]	640×480	19,994	50,000	142,210
RefCOCO+ [2]	640×480	19,992	49,856	141,564
RefCOCOg [1]	640×480	25,799	49,822	95,010
GigaGrounding	>25k×14k	3,775	61,353	61,353

Table 1. Statistics of existing benchmarks and GigaGrounding.

challenges, several approaches have been explored. Rudimentary strategies like "uniform downsampling (UD)" and "cutting into patches (CIP)" may be employed, but they may result in the loss of detailed information and relational cues, respectively [10]. Alternatively, more sophisticated techniques, including non-uniform downsampling [35–37], selective zooming and skipping [38–40], and the use of lightweight scanner networks [41, 42], have been proposed. Bakhtiarnia *et al*. [10] have provided a comprehensive review of these methods, their relevant applications, and datasets pertaining to high-resolution deep learning.

3. GigaGrounding Benchmark

This section offers a comprehensive delineation of the GigaGrounding benchmark, including an overview, an elucidation of the data collection strategy, and analyses of its unique properties.

3.1. Overview of GigaGrounding

Dataset overview GigaGrounding is built upon images pre-extracted from PANDA-Video dataset by the dataset providers [11]. A summary of key statistics comparing conventional VG benchmarks with GigaGrounding is provided in Table 1. The analysis reveals a significant limitation in conventional VG datasets, where image resolutions are constrained to the megapixel level. In contrast, GigaGrounding ambitiously extends its scope to gigapixel-level images, encompassing intricate details and large-scale scenes. Consequently, the total number of images naturally decreases. In terms of bounding boxes, GigaGrounding offers a quantity comparable to other benchmark datasets. Additionally, in traditional benchmarks, multiple expressions may correspond to a single bounding box, while in GigaGrounding, each bounding box is associated with only one expression.

We illustrate representative samples from conventional VG datasets in Figure 2, and also provide examples of GigaGrounding annotations in Figure 3. Specifically, GigaGrounding includes two types of annotated expressions:

- **Direct expressions**: These expressions directly describe each target object without introducing dependencies on other objects. Position modifiers, such as "in the top left of the image", may be used when describing small and challenging-to-locate objects, but only when necessary. In total, there are 41,040 direct expressions.
- **Multi-hop expressions**: These expressions require grounding models to respect additional restrictive constraints dependent on their relationship with other *reference objects*. In other words, the model needs to locate objects by first identifying specific reference objects and then finding the target objects based on constraints defined by relative positions (*e.g*., "to the left") or relationships (*e.g*., "watching", "holding") with those reference objects. There are 20,313 multi-hop expressions in total. It is noteworthy that we did not identify a design intentionally analogous to multi-hop expressions in traditional VG benchmarks. Notably, RefCOCO+ was explicitly designed to "focus on purely appearance-based description" [2].

For more details regarding the PANDA-Video dataset and additional examples of GigaGrounding, please refer to the Supplementary Material.

Privacy GigaGrounding is built based on the publicly available PANDA [11] dataset. PANDA is collected in public areas where photography is officially approved and is published under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 License [43]. GigaGrounding has been created by further processing the data in PANDA with additional annotations. To protect individual privacy, PANDA has already anonymized the faces in the images. However, it is necessary to acknowledge that certain individuals may still be identifiable based on other less invasive characteristics, such as clothing or body shape [44].

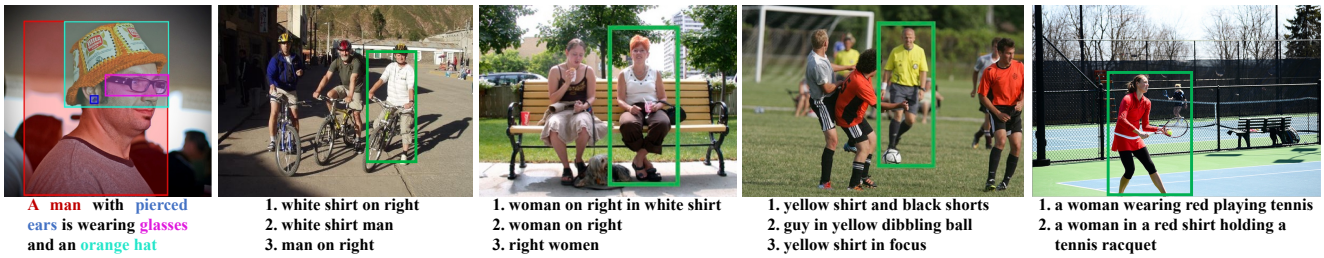
3.2. Data Collection Strategy

We have devised the following annotation procedure and quality control strategies to diligently uphold non-ambiguity and maintain high annotation quality.

Annotation protocol The annotation protocol, as depicted in Figure 4, comprises the following steps:

1. Given the gigapixel images, *Group A* of workers was asked to write natural language expressions and draw bounding boxes for the corresponding objects.
2. Given the gigapixel images and natural language expressions from Step 1, *Group B* of workers was then asked to draw corresponding bounding boxes within 40 seconds independently.
3. If the Intersection over Union (IoU) rate of bounding boxes from Steps 1 and 2 was greater than 0.5, the sample were deemed valid. Then the valid expressions were manually translated into English with the highest possible fidelity.

This procedure tries to ensure non-ambiguity, by excluding data instances with object selection discrepancies between Steps 1 and 2.



(a) Flickr30K Entities (b) ReferItGame (c) RefCOCO (d) RefCOCO+ (e) RefCOCOg

Figure 2. Example data from conventional visual grounding datasets.

Figure 3. Example annotation results of GigaGrounding. Blue boxes indicate direct expressions and orange ones are multi-hop expressions.

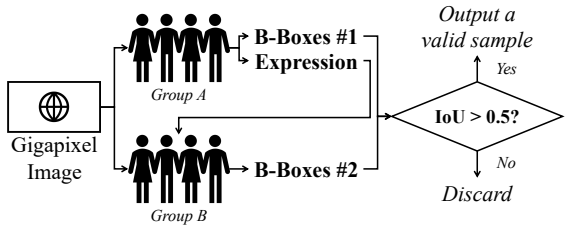


Figure 4. The overview of our annotation procedure.

Quality control strategies We employ the following strategies for quality control.

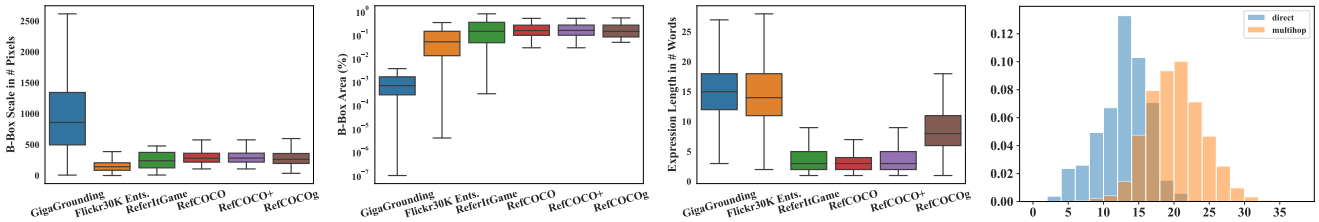
- **Terminology standardization**. We introduced a standardized lexicon to encourage uniform terminology usage among annotators, reducing the prevalence of ambiguous or domain-specific terms. For example, the use of specific color names like "gamboge" and the adoption of specialized terminology such as "coupe" should be avoided.
- **Sampling inspection and repair**. We performed a sampling inspection for each annotation batch to assess its quality. Specifically, for expression and bounding box annotations, we established a sampling pass rate threshold of 96%, whereas, for expression translation, the threshold was set to 99%. The sampling inspection process included no fewer than 100 samples per round, and the batch of data was accepted once it met the pass rate threshold or underwent repair.
- **Box ensemble and de-duplication**. In Step 3 of the annotation procedure, we obtained the ground truth by averaging the valid bounding boxes annotated by both groups of annotators. Any expressions that did not result in a valid bounding box were discarded. Furthermore, as data annotation is carried out in parallel among many annotation workers, we performed a deduplication procedure to eliminate objects with overlapping bounding boxes and identical expressions across multiple images to enhance the data variance, with an IoU threshold empirically established at 0.3.

Despite potential undetected ambiguities, the implemented methodologies are believed to substantially mitigate their effects, thereby enhancing the dataset's overall quality.

3.3. Analyses of GigaGrounding

This section analyzes the salient characteristics differentiating GigaGrounding from existing VG benchmarks. Our analysis encompasses the following viewpoints.

Quantities of objects per image Among conventional datasets, Flickr30K Entities in Figure 2a requires the model



(a) Pixel-count distribution of bounding box scales. (b) Relative area of bounding boxes *w.r.t.* the images. (c) Word-count distribution of expression lengths. (d) Word-count distribution of direct and multi-hop expressions.

Figure 5. A comparative analysis of GigaGrounding against established benchmarks RefCOCO, RefCOCO+, and RefCOCOg. (a) and (b) elucidate the distribution patterns of bounding box scale (in terms of the max value of height and width) and relative area with respect to the image, (c) presents the distribution of expression lengths, and (d) shows the lengths' distribution of both direct and multi-hop expressions within the GigaGrounding dataset. GigaGrounding exhibits a notably elevated level of complexity, primarily stemming from its pronounced scale variation and intricate expression structures.

to identify and locate multiple objects corresponding to short phrases within a single sentence. However, we observe that the model may not need to fully understand the global semantics before grounding the corresponding object, particularly when there exists only one object of that type. On the other side, RefCOCO and RefCOCO+ include an average of 3.9 same-type objects per image, while RefCOCOg displays a relatively lower average of 1.63 same-type objects per image [2].

Scales of bounding boxes In Figures 5a and 5b, we respectively illustrate the absolute scale of bounding boxes and their relative area with respect to the images. It is evident that GigaGrounding exhibits greater scale variance in comparison to other datasets. Furthermore, bounding box scale in GigaGrounding adheres to a long-tail distribution, while other datasets generally show a more concentrated distribution. On average, the RefCOCO dataset features bounding boxes with dimensions of 212×268 pixels, accounting for approximately 18.49% of the total image area. In contrast, the GigaGrounding dataset exhibits an average bounding box size of 478×978 pixels, representing just 0.12% of the image area. This statistic reveals the inherent challenges associated with grounding models tasked with large-scale scenes. We also checked the distribution of direct and multi-hop bounding boxes and noticed that they share a similar distribution pattern, while the bounding boxes corresponding to multi-hop expressions are slightly smaller, with an average resolution of 425×938, compared to 503×997 of direct expression bounding boxes.

Lengths of expressions We demonstrate the word count distribution of expression lengths in Figure 5c. We note that the average expression length in RefCOCOg is 8.46 words, while in RefCOCO and RefCOCO+, it is only 3.50 and 3.53, respectively. In contrast, in order to locate objects in large-scale scenes without ambiguity, GigaGrounding provides significantly longer expressions, and the average expression length is 14.78 words. Flickr30K Entities presents comparatively lengthy expressions, while it is tasked with ground objects referred by short region phrases. Figure 5d further illustrates the length distribution of direct expressions and multi-hop expressions, and the average lengths are 12.33 and 18.74, respectively. Longer expressions may accommodate more details about the objects' location, appearance, and properties, which can increase the semantic space that the model needs to consider, leading to challenges in both textual and imagery comprehension.

Other statistics We notice that GigaGrounding is certainly not offering the most number of images, bounding boxes, and annotated expressions among the compared datasets (Table 1). However, we believe that GigaGrounding's unique characteristics have opened the door for new modeling paradigms with practically useful and broader real-life applications.

4. Approach

In this section, we present an initial step towards effective visual grounding in the complex context of gigapixel-level large-scale scenes. We introduce a simple yet effective model, referred to as GlaZing.

4.1. Architecture Overview

In the context of GigaGrounding, human operators commonly begin with an initial coarse search across the image to identify regions of interest. Following this initial search, they then allocate their attention to the selected regions, subjecting them to thorough scrutiny. This methodical narrowing of focus significantly increases search efficacy. Our GlaZing model is conceived by analogizing this human cognitive heuristic.

An overview of GlaZing is presented in Figure 6. Given an input image, it initiates by inputting the downsampled thumbnail and the corresponding expression into the *Glance Grounding Module* (GGM) for preliminary localization and grounding. Subsequently, the *Adaptive Cropping Module* (ACM) dynamically extracts the high-resolution patch

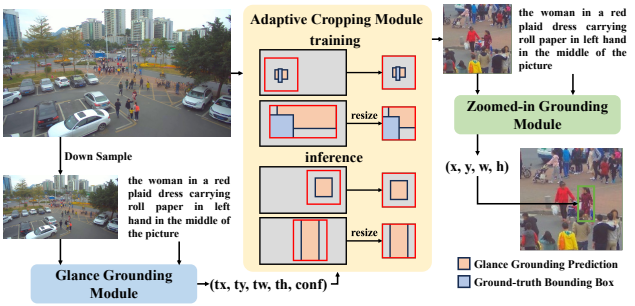


Figure 6. Overview of the proposed GlaZing model.

of interest. This extracted patch, along with the expression, is then forwarded to the *Zoomed-in Grounding Module* (ZGM) to generate the final prediction.

4.2. Detailed Implementations

The architectural framework expounded in Section 4.1 allows for a multitude of implementations. Notably, GGM and ZGM can be seamlessly integrated with a wide array of conventional grounding models. In this section, we present our specific implementation, which yielded favorable results in our experiments.

Glance grounding module We design GGM based on ReSC [28], which adopts an anchor-based grounding paradigm that divides the image into multiple grids and determines the grid to which the target box belongs. This design effectively locates the target-relevant region in the thumbnail image.

Adaptive cropping module This module is devised for the adaptive extraction of a region of interest from the original image, guided by the GGM's output. More precisely, ACM initiates this process by first computing the coordinates for a patch of specific size located at the center of the bounding box predicted by the GGM. It subsequently proceeds to crop the minimum rectangular region that encloses both this designated patch and the bounding box predicted by the GGM. During model training, this process also encapsulates the ground-truth bounding box. Finally, the extracted region is resized to a specified target size. Through the utilization of ACM, our model can adaptively focus on regions of varying scales and locations.

Zoomed-in grounding module After the first glance and zooming-in, the target bounding box should account for a significant proportion of the input patch of ZGM, so we employ a coordinate regression-based visual grounding method, *i.e.*, TransVG [7] to accomplish this task.

Training objectives The overall training objective can be defined as $\mathcal{L} = \mathcal{L}_{\text{GGM}} + \mathcal{L}_{\text{ZGM}}$. We follow the training objectives employed in ReSC and TransVG for $\mathcal{L}_{\text{GGM}}$ and $\mathcal{L}_{\text{ZGM}}$ respectively. Specifically, $\mathcal{L}_{\text{GGM}}$ comprises the YOLOv3 loss [45] designed to optimize anchor association confidence and prediction offset, and a diversity loss [28] intended to guide sub-queries in each iteration to concentrate on distinct elements within the query. Concerning $\mathcal{L}_{\text{ZGM}}$, given that ZGM directly generates a 4-dimensional vector representing the relative coordinates of the grounded target box with respect to the input patch, we employ both the smooth $\ell 1$ loss and the generalized IoU loss [46] to minimize the disparity between the predicted and target values.

4.3. Discussion

As an initial stride for GigaGrounding, GlaZing tackles the challenges posed by high-resolution input and scale variation through a "glance-to-zoom-in" cascading strategy coupled with adaptive cropping. Empirical findings in Section 5.3 attest to its effectiveness. Nevertheless, substantial opportunities for further refinement remain, which we leave as future work, including: (a) more tailored training objectives for GGM and ZGM, (b) end-to-end pipelines for seamless glancing and zooming-in, and (c) the feasibility of multiple iterations of glance and zoom adjustments.

5. Experiments

This section delineates the evaluation protocols employed, followed by results' presentation and analysis.

5.1. Baselines

We conducted an evaluation of several publicly available baselines. As discussed in Section 2, two-stage methods rely on object detectors to generate region proposals. Given that object detection in large-scale scenes remains a challenge in its own right [40], we conducted brief experiments involving a representative open-source two-stage method, namely, **MAttNet** [47]. On the other hand, one-stage methods directly provide bounding box predictions without relying on separate object detection. In this regard, GlaZing is more aligned with one-stage methods. We evaluated state-of-the-art open-source one-stage methods, including **ReSC** [28], **TransVG** [7], **RefTR** [48], **SeqTR** [29], **QRNet** [49], and **SimREC** [50]. A concise introduction to these baselines is provided in the Supplementary Material.

In addition to visual grounding models, we also conducted a **human evaluation** to assess the ability of human subjects to complete the task when presented with downsampled images as used by deep learning models.

5.2. Evaluation Protocol

Dataset split We randomly selected 70% distinct expressions for training, 10% for validation, and 20% for testing. This process resulted in 42,641 samples for the training set, 6,254 for the validation set, and 12,458 for the testing set.

Metric Following the standard protocol, we assessed the performance using Precision@0.5, where the prediction is deemed correct if its IoU with the ground-truth box is larger than 0.5. We also provide information on inference time and GFLOPs to demonstrate the computational efficiency of each model.

Image preprocessing In our empirical analysis, for the two-stage baseline MAttNet, features were extracted with 8-times downsampled images. We followed Yu *et al*. [47], who used ground-truth object bounding boxes to eliminate region proposal bottlenecks from the object detector.* On the other side, one-stage baseline models were evaluated using two distinct image resolutions: 640×640 and 1536×1536. The first resolution was in accordance with established benchmarks, while the latter was the feasible upper limit for processing by most baseline models given our computational constraints. At 640×640, the mean bounding box dimensions were roughly 12×25 pixels. This size expanded to 30×60 pixels at 1536×1536 resolution. To ensure fair comparison, for GlaZing, we employed a 640×640 thumbnail image for the GGM, and the ACM cropped the region of interest from 1536×1536 images. The cropped region was subsequently resized to 640×640 for the ZGM to identify the referred region.

Implementation details For all baseline methods, we conducted experiments based on their official code. However, for the object detection component of MAttNet, due to the lack of maintenance, we substituted it with the corresponding implementation of Faster R-CNN [51] from Detectron2 [52]. The substitution did not hurt the performance on RefCOCO based on our experiments. For one-stage methods, we tried to optimized their performance within a certain hyper-parameter tuning budget. For instance, for models with official weights on RefCOCO provided, we experimented with whether or not to use these weights for initialization and reported the better results. At the resolution of 1536×1536, we also tried initializing with weights from GigaGrounding at 640×640, and reported the better outcomes. We also adjusted some important hyperparameters including learning rate, *etc*. As for GlaZing, we trained the model with similar settings as ReSC. For all experiments, we set the maximum query length to 30, and the computations were performed on a server equipped with 8 NVIDIA RTX 3090 GPUs. Please refer to the Supplementary Material for more details.

5.3. Overall Performance

We present the evaluation results of our approach and the baselines in Table 2. Several key observations and analyses are as follows:

- **Baseline models' limited efficacy in GigaGrounding:** None of the baseline models achieved Precision@0.5 scores surpassing 50%. In contrast, human testers consistently achieved scores exceeding 60%, even on images of size 640×640. It is also noteworthy that the performance on GigaGrounding dramatically diverged from that on other benchmark datasets, such as RefCOCO. For instance, the classical ReSC model outperformed other baselines on GigaGrounding, even surpassing models initialized with pretrained weights.
- **Substantial advantages of GlaZing:** GlaZing demonstrated a significant performance advantage over all baseline models. When compared to the highest-performing deep learning baseline at a resolution of 1536×1536 (*i.e*., ReSC), GlaZing achieveed an impressive relative improvement of 28.6%. This substantial gain can be attributed to its glance-to-zoom-in design.
- **Incompatibility of existing two-stage methods:** We noticed that MAttNet exhibited marginal improvements over random guessing (6.8% vs. 6.2%), despite that it used the ground-truth object bounding boxes as region proposals. Our hypothesis is that the bottleneck lies in the feature extractors. Two-stage methods typically use frozen feature extractors pretrained on MS COCO or ImageNet [53], which may not be effective enough for large-scale scene understanding.

Additionally, our study reveals the following insights:

- **Influence of resolution on performance:** Our findings highlight the critical role of high-resolution input in enhancing the performance of both deep learning models and human evaluators.
- **Performance variation with expression complexity:** Notably, the GlaZing model demonstrated consistent performance across both direct and multi-hop expressions. In contrast, baseline models exhibited a marked performance decline when handling multi-hop expressions. This disparity underscores the need for models to adapt to expressions with differing levels of complexity.

5.4. In-depth Analysis

Effective high-resolution processing of GlaZing. GlaZing's adaptive cropping mechanism allows for higher-resolution image processing without significant storage and computation overhead. Experiments with 8-times downsampled images in ACM and ZGM showed promising results: 66.5% on the test set, with 66.2% for direct and 67.5% for multi-hop expressions. These findings suggest that processing at higher resolutions may further enhance the model's performance.

Benefits of RefCOCO initialization. We report the results of TransVG, RefTR, and SeqTR with and without RefCOCO weight initialization in Table 3. The adoption of

*Consequently, it is inappropriate to directly compare the results of one-stage and two-stage methods.

Category	Method	Backbone	Val			Test			Infer Time	GFLOPs	Trainable Params.
			overall	direct	multi-hop	overall	direct	multi-hop			
Two-stage	MAttNet	ResNet-101	11.4%	11.6%	10.7%	6.8%	6.4%	7.6%	268.6ms	1592.6	13.0M
One-stage (640×640)	ReSC	DenseNet-53	30.8%	35.5%	21.2%	29.0%	33.3%	20.3%	53.4ms	101.3	158.2M
	QRNet	Swin-S	16.2%	19.5%	9.4%	13.8%	16.8%	7.9%	54.8ms	79.2	247.1M
	SimREC	CSPDarkNet-53	23.3%	27.6%	14.4%	20.8%	24.7%	12.7%	16.5ms	48.9	40.2M
	TransVG‡	ResNet-101	13.1%	15.9%	7.1%	11.6%	13.7%	7.3%	25.9ms	41.4	122.6M
	RefTR‡	ResNet-50	25.4%	29.0%	18.1%	21.8%	25.0%	15.3%	45.0ms	23.5	123.4M
	SeqTR‡	Darknet-53	17.6%	20.1%	12.4%	15.9%	18.1%	11.6%	39.8ms	48.9	100.9M
	Human		– –	– –	– –	62%§	– –	– –	– –	– –	– –
One-stage (1536×1536)	ReSC	DenseNet-53	52.0%	57.5%	40.8%	49.7%	56.0%	37.0%	96.6ms	566.3	158.2M
	QRNet	Swin-S	28.8%	33.3%	19.5%	26.5%	31.0%	17.3%	141.6ms	435.0	247.1M
	SimREC	CSPDarkNet-53	35.4%	41.7%	22.5%	33.3%	39.5%	20.8%	48.4ms	280.9	40.2M
	TransVG‡	ResNet-101	26.3%	31.2%	16.5%	24.6%	29.1%	15.4%	72.6ms	225.3	122.6M
	RefTR‡	ResNet-50	42.4%	48.4%	30.0%	41.0%	46.7%	29.5%	53.0ms	119.7	123.4M
	SeqTR‡	Darknet-53	35.6%	41.4%	23.5%	33.2%	38.3%	22.8%	51.2ms	263.7	100.9M
	Human		– –	– –	– –	84%§	– –	– –	– –	– –	– –
GlaZing (Ours)			**65.0%**	**65.3%**	**64.3%**	**63.9%**	**64.1%**	**63.7%**	110.2ms	172.1	280.8M

‡We used the released weights pretrained on RefCOCO for initialization. §Estimated with 100 randomly selected samples from the testing set.

Table 2. Performance Evaluation on the GigaGrounding Dataset. Both accuracy-related and efficiency-related metrics are reported.

	TransVG	RefTR	SeqTR
w/o pretraining	5.1%	0.0%	5.3%
w/ pretraining	11.6%	21.8%	15.9%
Absolute Gain	+6.5%	+21.8%	+10.6%

Table 3. Performance improvements on 640×640 input images achieved through pretraining. We initiated each model with publicly available weights trained on the RefCOCO dataset.

Expr. length & proportion	ReSC	TransVG	GlaZing
1-10 (16.99%)	65.9%	42.4%	66.8%
11-20 (62.77%)	47.8%	22.6%	63.5%
21+ (20.24%)	39.7%	13.8%	62.6%

Table 4. Performance stratification by expression length.

B-box scale	ReSC	TransVG	GlaZing
$s < Q1$	29.4%	4.5%	48.1%
$Q1 \leq s < Q2$	50.2%	19.3%	63.6%
$Q2 \leq s < Q3$	58.2%	31.1%	72.8%
$s \geq Q3$	60.7%	43.6%	70.9%

Table 5. Performance stratification by bounding box scale. Here, $Q1$, $Q2$, and $Q3$ denote the first, second, and third quartiles of bounding box scale, respectively.

pretrained weights from RefCOCO led to performance enhancements for three methods, despite the substantial differences between RefCOCO and GigaGrounding.

Performance about expression length and B-box scale. We conducted an in-depth analysis of the performance variability about expression length and B-box scale at 1536×1536, specifically focusing on the ReSC, TransVG, and GlaZing. The results are presented in Tables 4 and 5. Our findings indicate that GlaZing demonstrated a notable robustness across varying difficulty levels. Conversely, TransVG exhibited a more pronounced performance degradation, particularly in scenarios involving longer expressions and smaller B-box sizes. Complete results of all baselines can be found in the Supplementary Material.

Analysis of the glance-to-zoom-in strategy in GlaZing. Our investigation reveals that GlaZing's ZGM effectively rectifies prediction errors observed in the GGM, underscoring the effectiveness of the glance-to-zoom-in approach for GigaGrounding. Specifically, we identified that approximately 54% of errors fall under "*target feature mismatch*", with ZGM correcting 48% of these. Additionally, another 20% of errors are categorized as "*over-fuzziness due to downsampling*", of which ZGM amends about 50%. Due to space constraints, detailed explanations and analyses are provided in the Supplementary Material.

6. Conclusion

This paper presents GigaGrounding, a novel and challenging benchmark for visual grounding in gigapixel-level large-scale scenes, accompanied by an efficient solution strategy. The distinctive challenges posed by GigaGrounding establish it as a pioneering benchmark in this domain. Our goal is to stimulate further research in visual grounding, aiming to propel the development of models that are both more accurate and efficient. These advancements are particularly crucial for real-world applications such as rescue operations.

Acknowledgements This work is supported in part by Natural Science Foundation of China (NSFC) under contract No. 62125106 and 62088102, in part by Ministry of Science and Technology of China under contract No. 2021ZD0109901, in part by Tsinghua-Zhijiang joint research center, in part by Young Elite Scientists Sponsorship Program by CAST under contract No. 2022QNRC001.

References

[1] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016.

[2] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, 2016.

[3] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014.

[4] 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 *ICCV*, 2015.

[5] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object retrieval. In *CVPR*, 2016.

[6] Jianan Li, Yunchao Wei, Xiaodan Liang, Fang Zhao, Jianshu Li, Tingfa Xu, and Jiashi Feng. Deep attribute-preserving metric learning for natural language object retrieval. In *ACM MM*, 2017.

[7] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. TransVG: End-to-end visual grounding with transformers. In *ICCV*, 2021.

[8] Xiaoyun Yuan, Mengqi Ji, Jiamin Wu, David J Brady, Qionghai Dai, and Lu Fang. A modular hierarchical array camera. *Light: Sci. Appl.*, 2021.

[9] Byford Sam. Samsung announces 200-megapixel phone camera sensor. https://www.theverge.com/2021/9/2/22653558/samsung-isocell-hp1-gn5-200-megapixel-camera-sensor-announced, 2021.

[10] Arian Bakhtiarnia, Qi Zhang, and Alexandros Iosifidis. Efficient high-resolution deep learning: A survey. *arXiv preprint arXiv:2207.13050*, 2022.

[11] Xueyang Wang, Xiya Zhang, Yinheng Zhu, Yuchen Guo, Xiaoyun Yuan, Liuyu Xiang, Zerun Wang, Guiguang Ding, David Brady, Qionghai Dai, and Lu Fang. PANDA: A gigapixel-level human-centric video dataset. In *CVPR*, 2020.

[12] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2014.

[13] Hugo Jair Escalante, Carlos A Hernández, Jesus A Gonzalez, Aurelio López-López, Manuel Montes, Eduardo F Morales, L Enrique Sucar, Luis Villasenor, and Michael Grubinger. The segmented and annotated IAPR TC-12 benchmark. *CVIU*, 2010.

[14] 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 *ECCV*, 2014.

[15] Yuxi Sun, Shanshan Feng, Xutao Li, Yunming Ye, Jian Kang, and Xu Huang. Visual grounding in remote sensing images. In *ACM MM*, 2022.

[16] Yang Zhan, Zhitong Xiong, and Yuan Yuan. RSVG: Exploring data and models for visual grounding on remote sensing data. *TGRS*, 61:1–13, 2023.

[17] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *ECCV*, 2022.

[18] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *CVPR*, 2017.

[19] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *ICCV*, 2019.

[20] Richang Hong, Daqing Liu, Xiaoyu Mo, Xiangnan He, and Hanwang Zhang. Learning to compose and reason with language tree structures for visual grounding. *TPAMI*, 2019.

[21] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. Modeling relationships in referential expressions with compositional modular networks. In *CVPR*, 2017.

[22] Sibei Yang, Guanbin Li, and Yizhou Yu. Dynamic graph attention for referring expression comprehension. In *ICCV*, 2019.

[23] Bohan Zhuang, Qi Wu, Chunhua Shen, Ian Reid, and Anton Van Den Hengel. Parallel attention: A unified framework for visual object discovery through dialogs and queries. In *CVPR*, 2018.

[24] Peng Wang, Qi Wu, Jiewei Cao, Chunhua Shen, Lianli Gao, and Anton van den Hengel. Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks. In *CVPR*, 2019.

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

[26] 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 *CVPR*, 2020.

[27] Arka Sadhu, Kan Chen, and Ram Nevatia. Zero-shot grounding of objects from natural language queries. In *ICCV*, 2019.

[28] Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. Improving one-stage visual grounding by recursive sub-query construction. In *ECCV*, 2020.

[29] 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 *ECCV*, 2022.

[30] Jianing Zhang, Tianyi Zhu, Anke Zhang, Xiaoyun Yuan, Zihan Wang, Sebastian Beetschen, Lan Xu, Xing Lin, Qionghai Dai, and Lu Fang. Multiscale-VR: Multiscale gigapixel 3d panoramic videography for virtual reality. In *ICCP*, 2020.

[31] Xiaoyun Yuan, Lu Fang, Qionghai Dai, David J Brady, and Yebin Liu. Multiscale gigapixel video: A cross resolution image matching and warping approach. In *ICCP*, 2017.

[32] Jianing Zhang, Jinzhi Zhang, Shi Mao, Mengqi Ji, Guangyu Wang, Zequn Chen, Tian Zhang, Xiaoyun Yuan, Qionghai Dai, and Lu Fang. GigaMVS: a benchmark for ultra-large-scale gigapixel-level 3d reconstruction. *TPAMI*, 2021.

[33] Guangyu Wang, Jinzhi Zhang, Kai Zhang, Ruqi Huang, and Lu Fang. GiganticNVS: Gigapixel large-scale neural rendering with implicit meta-deformed manifold. *TPAMI*, 2023.

[34] Xueyang Wang, Xuecheng Chen, Puhua Jiang, Haozhe Lin, Xiaoyun Yuan, Mengqi Ji, Yuchen Guo, Ruqi Huang, and Lu Fang. The group interaction field for learning and explaining pedestrian anticipation. *Engineering*, 2023.

[35] Adria Recasens, Petr Kellnhofer, Simon Stent, Wojciech Matusik, and Antonio Torralba. Learning to zoom: a saliency-based sampling layer for neural networks. In *ECCV*, 2018.

[36] Dmitrii Marin, Zijian He, Peter Vajda, Priyam Chatterjee, Sam Tsai, Fei Yang, and Yuri Boykov. Efficient segmentation: Learning downsampling near semantic boundaries. In *ICCV*, 2019.

[37] Chen Jin, Ryutaro Tanno, Thomy Mertzanidou, Eleftheria Panagiotaki, and Daniel C Alexander. Learning to downsample for segmentation of ultra-high resolution images. In *ICLR*, 2022.

[38] Mingfei Gao, Ruichi Yu, Ang Li, Vlad I Morariu, and Larry S Davis. Dynamic zoom-in network for fast object detection in large images. In *CVPR*, 2018.

[39] Burak Uzkent and Stefano Ermon. Learning when and where to zoom with deep reinforcement learning. In *CVPR*, 2020.

[40] Kai Chen, Zerun Wang, Xueyang Wang, Dahan Gong, Longlong Yu, Yuchen Guo, and Guiguang Ding. Towards real-time object detection in gigapixel-level video. *Neurocomputing*, 2022.

[41] Maria Tzelepi and Anastasios Tefas. Improving the performance of lightweight CNNs for binary classification using quadratic mutual information regularization. *Pattern Recogn.*, 2020.

[42] Huangjing Lin, Hao Chen, Simon Graham, Qi Dou, Nasir Rajpoot, and Pheng-Ann Heng. Fast ScanNet: Fast and dense analysis of multi-gigapixel whole-slide images for cancer metastasis detection. *TMI*, 2019.

[43] Using Creative Commons Public Licenses. Attribution-noncommercial-sharealike 4.0 international.

[44] Baowei Jiang, Bing Bai, Haozhe Lin, Yu Wang, Yuchen Guo, and Lu Fang. DartBlur: Privacy preservation with detection artifact suppression. In *CVPR*, 2023.

[45] Joseph Redmon and Ali Farhadi. YOLOv3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[46] Hamid Rezatofighi, 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 *CVPR*, 2019.

[47] 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 *CVPR*, 2018.

[48] Muchen Li and Leonid Sigal. Referring transformer: A one-step approach to multi-task visual grounding. In *NeurIPS*, 2021.

[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 *CVPR*, 2022.

[50] Gen Luo, Yiyi Zhou, Jiamu Sun, Shubin Huang, Xiaoshuai Sun, Qixiang Ye, Yongjian Wu, and Rongrong Ji. What goes beyond multi-modal fusion in one-stage referring expression comprehension: An empirical study. *arXiv preprint arXiv:2204.07913*, 2022.

[51] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015.

[52] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

[53] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015.

When Visual Grounding Meets Gigapixel-level Large-scale Scenes: Benchmark and Approach

Supplementary Material

A. More Information about GigaGrounding

For the construction of GigaGrounding, we used images pre-extracted from PANDA-Video by the providers. Both the training and testing split are included, and there are 20 videos from varied scenes with different camera positions. 120-238 frames are extracted from each video, and each image exhibited real-world scenes with a broad field-of-view, high-resolution details, and numerous objects, thereby offering diverse semantics. Readers may refer to [11] for more detailed information about PANDA.

More example annotation results of GigaGrounding is presented in Figure 7.

B. Brief Introduction to Baselines

The considered two-stage method is

- **MAttNet** [47]. This method decomposes expressions into three modular components related to subject appearance, location, and relationship to other objects, and then performs visual grounding.

The one-stage baselines comprised

- **ReSC** [28]. This model uses a recursive sub-query construction framework, which reasons between image and query for multiple rounds and reduces the referring ambiguity step by step. It employs an *anchor-based* grounding module.
- **TransVG** [7]. This model employs a visual transformer, a linguistic transformer, and a visual-linguistic transformer to encode and fuse information from the two modalities and predict the corresponding spatial coordinates through a *coordinate regression* process.
- **RefTR** [48].This model leverages a transformer architecture. Features from two modalities are fused in the visual-lingual encoder, and then the model learns to generate contextualized lingual queries in the decoder, which are decoded to directly regress the bounding box coordinates.
- **SeqTR** [29]. This model provides a universal network for visual grounding by sequentially predicting *coordinate tokens*.
- **QRNet** [49].This method has a transformer architecture similar to TransVG, and proposes a query-modulated refinement network for adjusting intermediate features in the visual backbone.
- **SimREC** [50].This paper builds a simple REC network and explores multiple design variations, then properly combines these findings to improve the grounding performance.

C. Human evaluation strategy

We conducted a human evaluation to assess the performance of our proposed method, which involved 10 participants. Prior to the evaluation, participants received brief training to acquaint themselves with the task. During the evaluation, they were presented with randomly selected expressions from the test dataset and were instructed to identify objects within downsampled images to the best of their abilities. Correct identifications were rewarded with monetary compensation.

To mitigate potential cognitive fatigue, several measures were implemented. Each participant was limited to evaluating 20 samples, and intermittent breaks were scheduled during the testing process. Instead of annotating bounding boxes, participants were solely tasked with identifying the target objects, omitting the need for precise localization, which is a trivial task for human. To maintain evaluation consistency and accuracy, guidance and supervision were provided throughout the duration of the experiment.

This experimental setup was devised to ensure the robustness and reliability of the human evaluation, given the intricate cognitive demands associated with the task.

D. More Implementation Details

For the baselines, we adhered to the official settings and conducted appropriate hyperparameter tuning to optimize the performance on GigaGrounding:

- **MAttNet**: For MAttNet, we adhered to all the official settings with the exception of replacing the detector with a Faster R-CNN produced by Detectron2 to extract visual features. The attribute labels in the training set were produced using the same template parser as MAttNet to identify color and generic attribute words. We also maintained the training settings, establishing a batch size of 1 and a learning rate of 4×10^{-4}.
- **ReSC**: At the resolution of 640×640, we set the batch size to 8 and the learning rate to 10^{-4}. At the resolution of 1536×1536, we used the weight obtained from GigaGrounding at 640×640 for model initialization, and we set the batch size to 2 and the learning rate to 10^{-4}.
- **TransVG**: We eliminated the random crop augmentation used in official work. At the resolution of 640×640, we

used the weight obtained from RefCOCO for model initialization, and we set the batch size to 8. The learning rate was set to 10^{-5} for visual CNN and BERT, and 10^{-4} for other parameters. At the resolution of 1536×1536, we used the weight obtained from GigaGrounding at 640×640 for model initialization, and we set the batch size to 4. The learning rate was set to 5×10^{-5}.

- **RefTR**: We used the ResNet-50 as the visual backbone because the official source only provided the training results of RefCOCO on ResNet-50. We used the weight obtained from RefCOCO to initialize the 640×640 GigaGrounding model. We set the batch size to 32, and the learning rate was set to 10^{-4}, while the learning rate of the image backbone and context encoder was set to 10^{-5}. At the resolution of 1536×1536, we used the weight obtained from RefCOCO, set the batch size to 8, and applied the same learning rate to 10^{-4}.
- **SeqTR**: For both 640×640 and 1536×1536 resolutions, we initialized the model using the weight obtained from RefCOCO. For the 640×640 resolution, we set the batch size to 64 and the learning rate to 2.5×10^{-4}. As for the 1536×1536 resolution, we set the batch size to 8 and the learning rate to 2.5×10^{-4}.
- **QRNet**: For both 640×640 and 1536×1536 resolutions, the learning rate was set to 10^{-5} for pre-trained parameters and 10^{-4} for other parameters. For the 640 × 640 resolution, we set the batch size to 16, for the 1536 × 1536 resolution, we set the batch size to 8.
- **SimREC**: We implemented the multi-scale training strategy mentioned in the paper, which significantly improved the model's performance on GigaGrounding. For the 640× 640 resolution, we used a scale range from 480×480 to 640×640. For the 1536×1536 resolution, we used a scale range from 640×640 to 1536×1536. The learning rate was set to 5×10^{-5} and the batch size was set to 8 for both resolutions.
- **GlaZing**: We trained the model with similar settings as ReSC. We set the batch size to 4 and the learning rate to 10^{-4}.

E. Complete Performance Stratification Results

The complete performance stratification results are delineated in Tables 6 and 7. We undertook a comprehensive analysis of the performance variability concerning expression length and B-box scale at a resolution of 1536×1536. The two-stage method, MAttNet, exhibited poor performance across all categories of breakdown. Two-stage methods typically employ frozen feature extractors pre-trained on datasets such as MS COCO, which may not be suffice for large-scale scene comprehension. All one-stage methods exhibited a substantial performance decline as the expression length increased. For example the efficacy of ReSC with expressions exceeding 21 words was only 60.2% of its efficacy with expressions between 1 to 10 words. Regarding the B-box scale, all one-stage methods performed suboptimally when processing boxes smaller than the first quartile in scale, with QRNet, TransVG, and SeqTR exhibiting particularly poor results. We hypothesize that certain design choices, such as strategies based on coordinate token prediction for bounding box delineation and direct coordinate regression for decoding, may struggle with accommodating variations in the bounding box scale. Conversely, GlaZing showcased remarkable resilience, maintaining consistent performance across a diverse array of difficulty levels.

F. Failure Cases Analyses of ReSC and GlaZing

Figure 8 illustrates example failure cases by ReSC at 1536×1536, each attributable to distinct causes. Approximately 40% of these failures originate from a "target object feature mismatch", suggesting that the model selected an object with properties that only partially match the ground truth. 34% of these failures are due to "inaccurate position prediction". Approximately 14% can be attributed to "over-fuzziness due to downsampling", where the target object becomes too small to be clearly visible after downsampling. An additional 12% of failures can be linked to a "reference object feature mismatch", indicating that the model incorrectly matched the reference object in multi-hop expressions.

Adhering to the same protocol, we conducted an analysis of GlaZing, which revealed that 50% of the failures could be attributable to target object feature mismatch, 40% of the failures are due to inaccurate position prediction. 4% can be ascribed to over-fuzziness due to downsampling, and 6% are a consequence of reference object feature mismatch.

GlaZing's performance was significantly enhanced by the glance-to-zoom-in strategy. In Figure 9, we demonstrate the effectiveness of this strategy. In Figure 9a, given the expression: "the woman in a red plaid dress carrying roll paper in left hand in the middle of the picture", the glance grounding module (GGM) identified a woman in red in the center of the image, which could be easily confused with the described target due to the similarity in location and attributes. Upon locking the region, in the zoomed-in grounding phase, by grounding on a high-resolution patch, the described target was successfully identified. Figure 9b presents a case involving a multi-hop expression. The GGM located the reference object mentioned in the expression, and with further refinement by the Zoomed-in Grounding Module (ZGM), the correct target was located.

Expr. length & proportion	MAttNet	ReSC	QRNet	SimREC	TransVG	RefTR	SeqTR	GlaZing
1-10 (16.99%)	6.5%	65.9%	44.4%	50.0%	42.4%	63.5%	52.9%	66.8%
11-20 (62.77%)	6.6%	47.8%	24.0%	31.6%	22.6%	37.3%	30.7%	63.5%
21+ (20.24%)	7.8%	39.7%	19.3%	21.8%	13.8%	34.9%	21.7%	62.6%

Table 6. Performance stratification by expression length for all methods.

B-box scale	MAttNet	ReSC	QRNet	SimREC	TransVG	RefTR	SeqTR	GlaZing
$s < Q1$	7.0%	29.4%	4.0%	17.6%	4.5%	16.9%	9.5%	48.1%
$Q1 \leq s < Q2$	6.2%	50.2%	20.2%	32.7%	19.3%	36.0%	27.7%	63.6%
$Q2 \leq s < Q3$	7.0%	58.2%	34.9%	36.7%	31.1%	50.8%	41.0%	72.8%
$s \geq Q3$	6.9%	60.7%	46.4%	46%	43.6%	60.8%	54.2%	70.9%

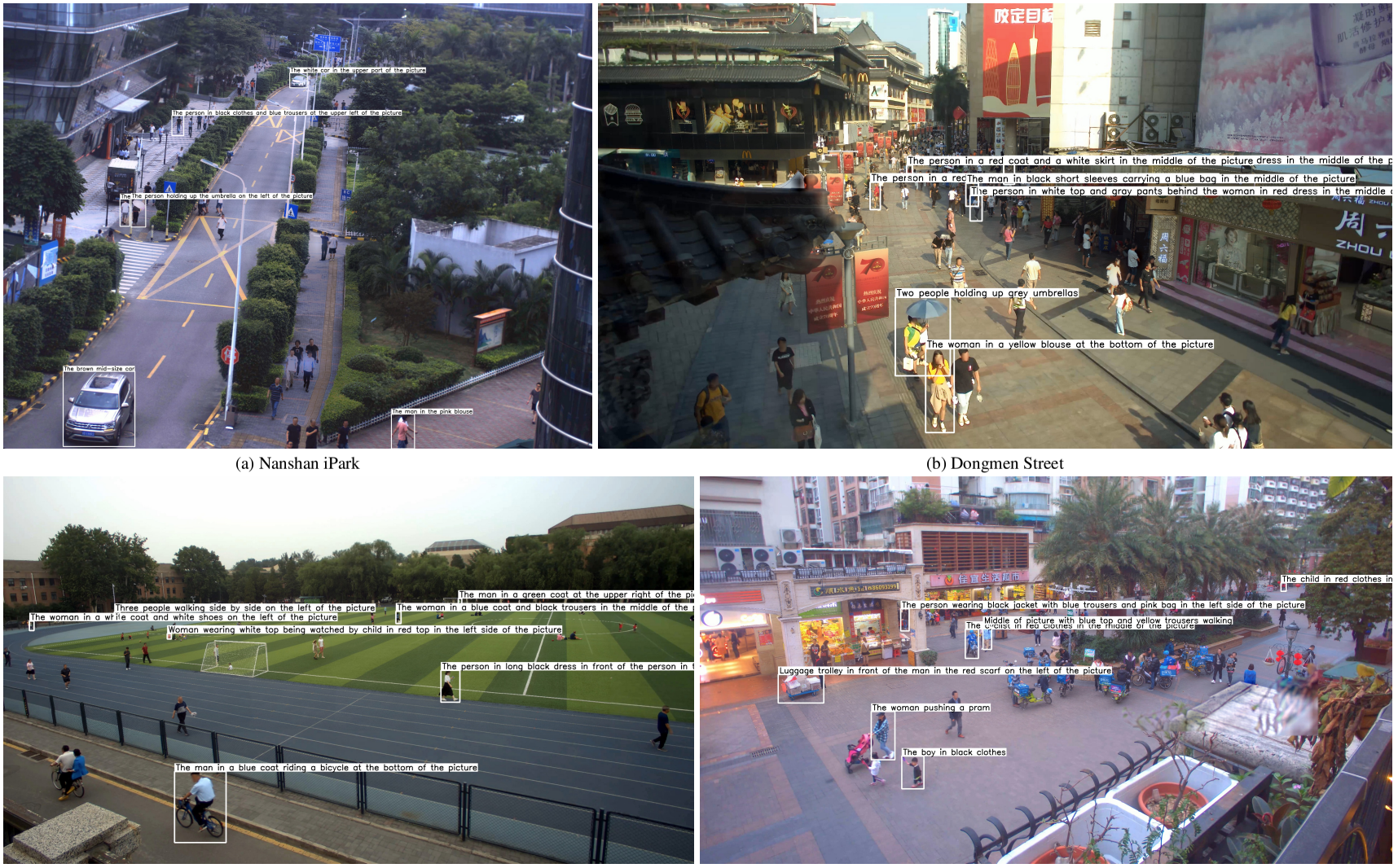
Table 7. Performance stratification by bounding box scale for all methods. Here, $Q1$, $Q2$, and $Q3$ denote the first, second, and third quartiles of bounding box scale, respectively.

G. Performance on Existing Visual Grounding Benchmarks

Method	RefCOCO		RefCOCO+		RefCOCOg
	testA	testB	testA	testB	val-g
ReSC	80.5%	72.3%	68.7%	56.8%	63.1%
TransVG	82.7%	78.4%	70.7%	56.9%	67.0%
SeqTR	86.5%	81.2%	76.3%	64.9%	71.5%
GlaZing	86.4%	82.3%	69.3%	60.6%	72.5%

Table 8. Performance evaluations on previous VG benchmarks.

We also present the benchmark results of GlaZing on previous VG datasets in Table 8. To conduct the evaluation, we employed the GGM algorithm to process 640×640 images, while ACM accurately extracted a region of interest measuring 320×320 from the original image. The results depicted in the table affirm the commendable performance of GlaZing across these established benchmarks.



(a) Nanshan iPark

(b) Dongmen Street

(c) University Playground

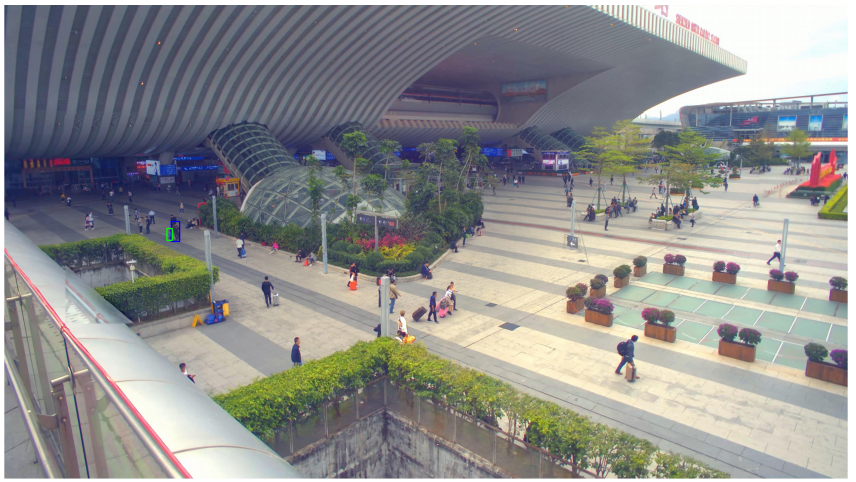
(d) Xili Street

Figure 7. Example annotation results from 4 scenes. Please zoom in and read on computers.

(a) Example failure case of target object feature mismatch. Expression: the woman in a white blouse and a grey skirt at the bottom of the picture.

(b) Example failure case of inaccurate location. Expression: the woman in grey clothes and blue trousers on the right of the picture.

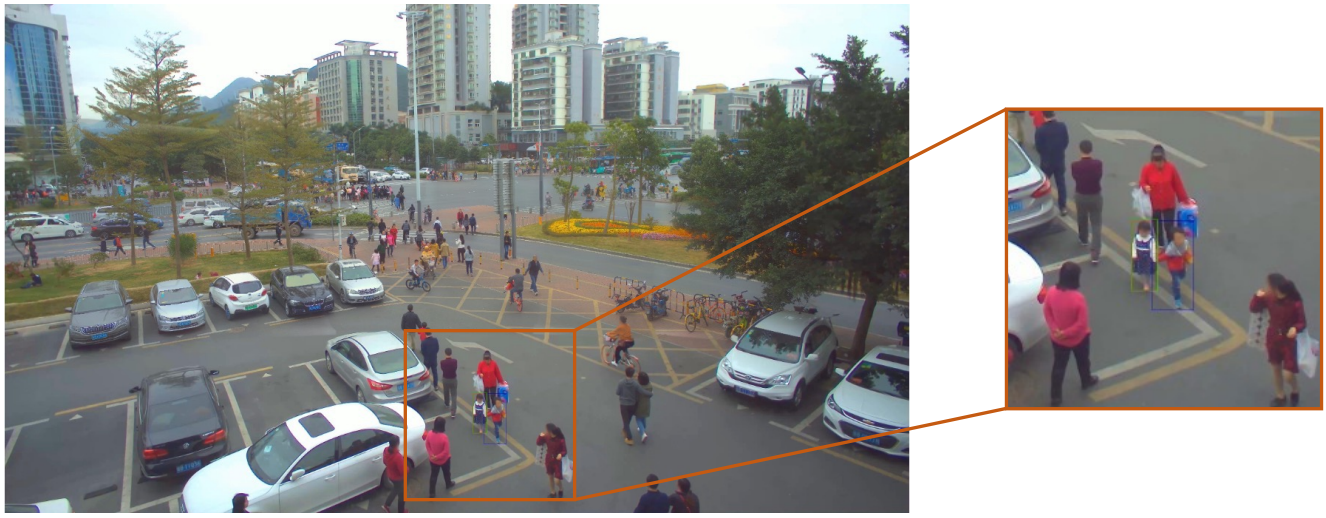
(c) Example failure case of over-fuzziness due to downsampling. Expression: the woman with short hair wearing a blue top and black trousers on the left of the picture.



(d) Example failure case of reference object feature mismatch. Expression: child in blue top in the left side of the picture holding hands with the man wearing black jacket.

Figure 8. Example failure cases of different reasons. Blue boxes indicate the prediction results by ReSC, and green boxes indicate the grouth-truth boxes in GigaGrounding.

(a) Expression: the woman in a red plaid dress carrying roll paper in left hand in the middle of the picture.



(b) Expression: the girl holding the boy in red to her left at the bottom of the picture.

Figure 9. Prediction results benefited from the glance-to-zoom-in strategy. Green boxes represent the ground-truth boxes in GigaGrounding, blue boxes denote the prediction results from the Glance Grounding Module, and red boxes signify the prediction results from the Zoomed-in Grounding Module.