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%

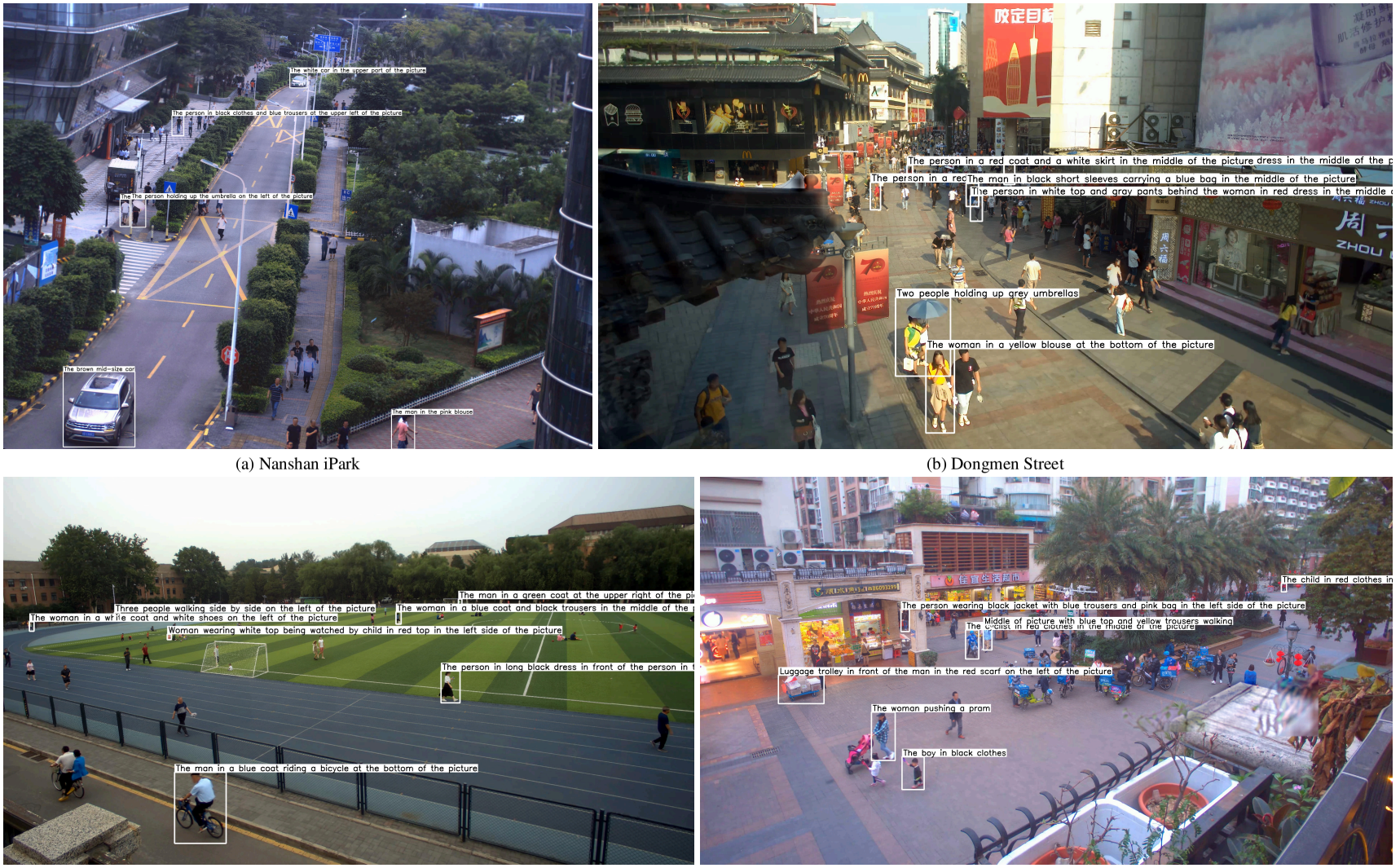
Table 7. Performance stratification by bounding box scale for all methods. Here, $Q1$, $Q2$, and $Q3$ denote the first, second, and third quartiles of bounding box scale, respectively.

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

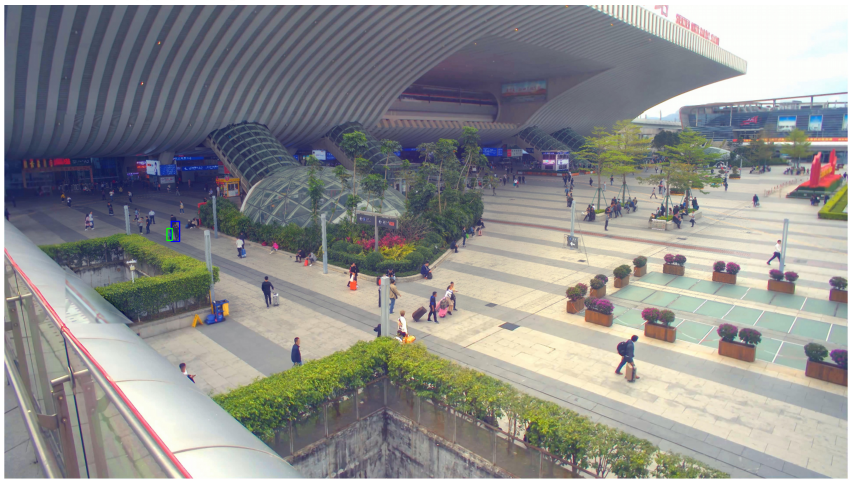
(d) Xili Street

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.