

Peer Review

Review of: "Grounding-IQA: Multimodal Language Grounding Model for Image Quality Assessment"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

The paper proposes a new task for multimodal models in image quality assessment. Instead of a score, or a general paragraph about blur and noise, the model should say *where* the quality problem is. The authors call this grounding-IQA and split it into two subtasks: GIQA-DES, a quality description with bounding boxes attached to the objects that affect quality, and GIQA-VQA, question answering about local regions, where either the question carries a box (referring) or the answer must return one (grounding). To train models for this, they build GIQA-160K with an automated pipeline on top of the human-written descriptions in Q-Pathway and DQ-495K (Sec. 3.3). Llama3 extracts object tags, Grounding DINO finds the boxes, Q-Instruct filters boxes of the wrong quality, a merge step removes redundant boxes, and the coordinates are written into the text as discretised grid indices (Sec. 3.2, Alg. 1). Table 1 gives 42,960 images and 167,657 samples: 66,689 DES, 50,484 Yes/No VQA, and 50,484 open-ended VQA.

For evaluation, they release GIQA-Bench, 100 held-out images with 100 DES and 150 VQA samples, annotated by at least three experts. Four open MLLMs (LLaVA-v1.5-7B, LLaVA-v1.5-13B, LLaVA-v1.6-7B, mPLUG-Owl2-7B) are fine-tuned on GIQA-160K and compared with general, grounding, and IQA-specialised models (Table 5). The central claim is that fine-tuning on the new data gives models both quality perception and localisation, which no existing model group has at once.

I think the task itself is the real contribution. It is a sensible and practical idea: a restoration or editing tool can act on "the hand on the left is blurred, here is its box" far more easily than on a score of 3.2. The pipeline is described stage by stage, the IQA-Filter step addresses a concrete failure (same-class objects

with different quality, Fig. 3a), and the ablations in Tables 2 to 4 test the design choices one by one. The code link is given in the abstract.

2. Whether the conclusion holds on the data

Much of it holds. I checked the internal arithmetic first, and it is clean. Table 1 adds up: $66,689 + 50,484 + 50,484 = 167,657$, and the bench row gives $100 + 90 + 60 = 250$, matching the breakdown in Sec. 3.4 ($35 + 55 = 90$ Yes/No; $30 + 18 + 12 = 60$ open-ended). The Acc (Total) column in Table 5 is the question-weighted mean of the two parts; for the row Grounding-IQA (mPLUG-Owl2-7B), $(90 \times 0.8444 + 60 \times 0.5875) / 150 = 111.246 / 150 = 0.7416$, reported as 0.7417. The fine-tuned rows of Table 4 agree exactly with the "Ours" rows of Table 5, and the Ref-Box row of Table 2a (Tag-Recall 0.5497, LLM-Score 61.75) matches Disc-Coord in Table 2b and Only-DES in Table 3, so the three ablations clearly share one run. That consistency is reassuring.

The strongest result is on VQA. In Table 5, with the same base model, Grounding-IQA (mPLUG-Owl2-7B) reaches Acc (Y) 0.8444 against 0.6111 for Q-Instruct (mPLUG-Owl2-7B), which is 76 versus 55 correct answers out of 90. That gap is far outside sampling noise. The binomial standard error at 0.8444 with $n = 90$ is $\sqrt{(0.8444 \times 0.1556 / 90)} \approx 0.038$. The grounding columns tell the same story for GIQA-VQA: Tag-Recall 0.7372 for the fine-tuned model, where the IQA group cannot ground at all.

Where I would ask the authors to soften the text is the summary sentence in Sec. 4.3: "our models outperform MLLMs specialized in grounding or IQA tasks in both aspects." Table 5 does not fully support "in both aspects".

- On GIQA-DES grounding, Ferret-7B has Tag-Recall 0.6778, higher than every Grounding-IQA row (the best is 0.5981, LLaVA-v1.6-7B). On mIoU, Ferret-7B (0.6458) is beaten only by Grounding-IQA (LLaVA-v1.6-7B) at 0.6583.
- On description quality, Q-Instruct (LLaVA-v1.5-7B) has BLEU@4 22.69, while Grounding-IQA on the same base has 19.02 (only the mPLUG-Owl2-7B row, 22.87, is above it). On LLM-Score, the mPLUG-Owl2 pair differs by one point (63.00 versus 62.00). On the 0 to 4 scale summed over 100 samples, that is $(63.00 - 62.00) / 25 \times 100 = 4$ raw points in total.
- On Acc (W), DepictQA-Wild-7B scores 0.5458, identical to Grounding-IQA (LLaVA-v1.5-7B).

So the fair reading is: similar description quality to the best IQA model, much better VQA, and grounding that is competitive with dedicated grounding models but not uniformly better on DES. That is still a good

result. It just needs to be stated that way.

Two ablation sentences have the same problem. Sec. 4.2 says Disc-Coord improves "grounding accuracy (Tag-Recall)" over Norm-Coord. In Table 2b the Tag-Recall values are 0.5497 and 0.5490, a difference of 0.0007, while mIoU goes the other way (0.5851 for Disc-Coord, 0.6046 for Norm-Coord). I would describe this as a trade: fewer tokens and slightly better text metrics, at some cost in box precision, which is exactly what the discretisation to a 20×20 grid would predict. Similarly, "multi-task training (GIQA-160K) enhances performance on both GIQA-DES/VQA" (Sec. 4.2) is true for three of the four columns of Table 3, but GIQA-DES Tag-Recall is 0.5474 for GIQA-160K against 0.5497 for Only-DES.

Several evaluation choices deserve a sentence or a control.

Class balance in the bench. Training is balanced (25,242 "Yes" and 25,242 "No", Sec. 3.3), but GIQA-Bench has 35 "Yes" and 55 "No" (Sec. 3.4). A model that always answers "No" would score $55 / 90 = 0.6111$ on Acc (Y). That is exactly the value of Q-Instruct (mPLUG-Owl2-7B) in Table 5. It may be a coincidence, but I would add the majority-class baseline to Table 5 and report a balanced accuracy or per-class accuracy for Yes/No.

The same LLM builds the data and judges it. Llama3 extracts tags (Stage-1), writes the VQA pairs (Sec. 3.2), and then scores LLM-Score (Sec. 3.4). Acc (W) is also scored by "LLM", which is not named, so I assume it is the same Llama3. A model fine-tuned on Llama3-shaped text may be favoured by a Llama3 judge. A small human check of the judge (for example, agreement on a subset of the 100 DES outputs) or a second judge model would settle this.

Source overlap with Q-Instruct. The bench DES descriptions "are from Q-Pathway and adjusted" (Sec. 3.4), and Q-Instruct was trained on Q-Pathway. The bench images are excluded from GIQA-160K, but I could not tell from the text whether they were also absent from Q-Instruct's training data. If they were not, the IQA-group numbers in Table 5 may be inflated, which would work against the authors, but it should still be stated.

No variance. Tables 2 to 5 give one value per cell, with no seeds, repeated runs, or spread reported, and the bench is small. With 60 open-ended questions and 100 descriptions, differences of a point or two in LLM-Score or 0.01 in Tag-Recall are within what I would expect from reruns. Confidence intervals, even bootstrap ones over bench samples, would tell the reader which gaps are real.

Quality of the automatic labels. Fig. 6 shows that refinement moves the box-area distribution closer to the bench. That is useful, but it is indirect. I could not find a manual audit of GIQA-160K itself: for

instance, on a few hundred random samples, how often is the box on the right object, and how often is the quality attribute correct? The pipeline starts from 80K image-text pairs (53K + 27K, Sec. 3.3) and ends with 66,689 DES samples; the fraction dropped at each stage would also show how much the filter does.

3. Reproducibility

The paper is better than average here. The training section (Sec. 4.1) gives the loss, AdamW with $\beta_2 = 0.999$, a cosine schedule with an initial learning rate of 2×10^{-5} and a warmup ratio of 0.03, batch size 64, 2 epochs, full fine-tuning, and four A100-80G GPUs. The base models are named with sizes. The code repository is linked. The thresholds of Box-Merge are given (area 0.256, overlap 95%, Sec. 3.2).

What I could not find in the Qeios text:

- Which Llama3 model is used (size, instruct version) for tag extraction, VQA generation and judging, and with what prompts and decoding settings. The paper refers to the supplement for pipeline details and for the testing prompts; the supplement is not part of the Qeios version I reviewed.
- The Grounding DINO checkpoint and its box and text thresholds, and which Q-Instruct variant serves as the IQA-Filter.
- Inference decoding for all models in Table 5 (greedy or sampling, temperature, maximum length). This matters for BLEU@4 in particular.
- Seeds, and whether the dataset and GIQA-Bench (with boxes and gold answers) are released, not only the code.
- How predicted and ground-truth boxes are matched for mIoU when counts differ, and what "object name similarity" means in Tag-Recall (string match, embedding cosine, which model).
- Inter-annotator agreement for the bench. "At least three experts" in "multiple rounds" is described, but no agreement statistic is given.

The discretisation in Sec. 3.2 is stated clearly ($n = m = 20$, at most 9 tokens per box), but Eq. (1) and Eq. (2) as written use plain products and "/" where a floor or integer division is needed to get a grid index and a row number. Algorithm 1 appears in the Qeios text only as its title, "IQA-Filter & Box-Merge", with no body.

4. What specifically to fix

1. **Sec. 4.3, "Quantitative Results", last sentence.** Rephrase the claim of outperforming grounding and IQA models "in both aspects". Name the cells where this does not hold: Ferret-7B on GIQA-DES Tag-Recall (0.6778) and mIoU against three of four fine-tuned models, Q-Instruct (LLaVA-v1.5-7B) on BLEU@4 (22.69 versus 19.02), and the tie on Acc (W) with DepictQA-Wild-7B (0.5458).
2. **Sec. 4.2, "Box Optimization", second paragraph:** Tag-Recall in Table 2b is 0.5497 versus 0.5490, and mIoU favours Norm-Coord (0.6046 versus 0.5851). Describe it as a trade-off.
3. **Sec. 4.2, "Multi-Task Training".** Note that GIQA-DES Tag-Recall drops slightly with joint training (Table 3: 0.5474 versus 0.5497), so "enhances performance on both" should be limited to the columns where it is true.
4. Add a majority-class row to Table 5 for Acc (Y). With 55 "No" out of 90, it is 0.6111. I would also report balanced accuracy.
5. **Tables 2 to 5.** Report bootstrap confidence intervals over bench samples, and if compute allows, at least two seeds for the mPLUG-Owl2-7B run used in the ablations of Tables 2 and 3.
6. **Sec. 3.4, LLM-Score and Acc (W):** validate the LLM judge (Llama3 for LLM-Score; the Acc (W) judge is not named) against human ratings on a subset, or repeat the scoring with a second judge from a different model family. Give the judge prompt in the main text or an appendix that travels with the preprint.
7. **Sec. 3.3 or a new subsection.** A manual audit of a random sample of GIQA-160K (box correctness, attribute correctness), plus the number of samples removed at each pipeline stage between the 80K source pairs and the final 66,689 DES answers.
8. State whether the 100 GIQA-Bench images appear in Q-Pathway training data used by Q-Instruct.
9. **Table 5, Acc (Y) column.** The values 0.3888 (LLaVA-v1.5-13B, General) and 0.7777 (Grounding-IQA LLaVA-v1.5-7B) look truncated rather than rounded: $35 / 90 = 0.38889$ and $70 / 90 = 0.77778$, and the Kosmos-2 row prints the same $35 / 90$ as 0.3889. Please round consistently. The table caption refers to red and blue colouring for best and second best; in the Qeios text, I read that the colours do not come through, so bold and underline would survive better.
10. Question-type naming is inconsistent. Sec. 3.2 says "What/Why/How"; Sec. 3.3 and 3.4 say "What/Which/How", yet the bench breakdown lists What, Why, and How only. Pick one.
11. **Sec. 3.2, discretisation.** Add the floor (or integer division) to Eq. (1) and Eq. (2), and restore the body of Algorithm 1, which is missing from the Qeios text.

12. Small text issues I noticed. Sec. 2.1: "Q-Instruct constructs a multimodal dataset to enhance." ends mid-thought. Sec. 3.2: "Is the image quality is <...>?" has a doubled verb, and "from top-left to bottom-right" is repeated. Sec. 3.1 cites Fig. 5b for both referring and grounding. In Sec. 2.1, reference [21] is cited twice in one bracket, and references [16] and [52] are the same paper ("Visual instruction tuning").
13. Add the reproducibility items listed in Section 3 above (Llama3 variant, Grounding DINO checkpoint and thresholds, inference decoding, Tag-Recall similarity measure, box matching for mIoU, inter-annotator agreement), and say whether GIQA-160K and GIQA-Bench are released.

5. Recommendation

Minor revisions. The task is new and useful, the dataset and bench are clearly constructed, the tables are internally consistent, and the VQA gain over the strongest IQA baseline is large enough to be real. What needs work is how the results are described (Sec. 4.2 and 4.3 claim more than Tables 2b, 3, and 5 show), plus a few controls that are cheap to add: a majority-class baseline, a check of the Llama3 judge, confidence intervals, and a small audit of the automatic labels. None of this changes the main contribution.

Declarations

Potential competing interests: No potential competing interests to declare.