ART: Adaptive Relation Tuning for Generalized Relation Prediction

Supplementary Material

Gopika Sudhakaran[1,2] Hikaru Shindo[1] Patrick Schramowski[1,3] Simone Schaub-Meyer[1,2]
Kristian Kersting[1,2,3] Stefan Roth[1,2]

[1]Department of Computer Science, TU Darmstadt, Germany
[2]Hessian Center for AI (hessian.AI) [3]German Research Center for AI (DFKI)

A. Overview

We provide supplementary experimental details, followed by an in-depth explanation of the balanced sampling algorithm used for initial sampling. We then include a dedicated section — *Understanding ART through examples* — which offers an intuitive walkthrough of ART's core sampling strategy using illustrative cases. Next, we analyze the relation predictions based on their diversity and whether they are unseen (*i.e.*, not contained in the training annotations). We then discuss the computational cost of ART and its baselines, analyze the trade-off between data usage and performance, and clarify the behavior of ART on certain recall metrics. Finally, we conclude with a qualitative comparison between ART and its baselines.

B. Additional experimental details

Training details. In addition to the hyperparameters outlined in Sec. 4 of the main paper, we set the initial z-score threshold to 1.96, which corresponds to 95% of the data. This threshold was chosen because a z-score of 1.96 is more sensitive to potential outliers, making it useful for capturing more subtle deviations from the norm. For training, we use an initial learning rate of $1e-3$ and a linear warmup for 3000 steps. We optimize with Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$) and apply a weight decay of 0.05. Additionally, we chose 12% of training data for instruction tuning as mR@k saturates near this point as analyzed in Fig. 5. We use the LAVIS library [19] for implementation, training, and evaluation. The models are trained using four Nvidia A100 (40Gb) GPUs within two days.

Dataset details. We adopt the VG150 split for Visual Genome (VG) [14], which includes 150 object classes and 50 predicates, aligned with established baselines [34, 36, 44, 45, 48]. In comparison, GQA [12] (GQA200 split) includes 100 predicates and 200 object classes. VG is a subset of GQA with overlapping categories. Testing on the expanded set of GQA allows us to assess the model's generalization to new predicates and object categories, a more rigorous test of robustness than the reverse (training on GQA and testing on VG). Additionally, we test on Open Images (OI) [16], where OI-v4 includes 9 predicates and 57 object classes, and OI-v6 expands to 31 predicates and 601 object classes. Since OI's data distribution is entirely distinct from VG and GQA, it serves as a fully out-of-distribution benchmark, presenting increased complexity and enabling us to comprehensively evaluate model adaptability and robustness to unseen categories and relationships.

On semantic similarity for evaluation. We threshold the semantic similarity $\mathbf{S}$ at 95% to ensure that only highly semantically similar predictions are counted. For example, in Fig. 3, FPs such as Ⓐ that are semantically similar to the ground truth are counted as TPs. The similarity is computed over subject–predicate–object triplets with only the predicate varying. Even small differences (*e.g.*, "bag on table" *vs.* "bag under table") yield noticeable drops in similarity. The threshold 0.95 was selected based on qualitative analysis, which confirmed that high-similarity matches preserved meaningful semantics and did not introduce false positives. Notably, ART frequently predicts semantically rich alternatives (*e.g.*, "girl petting dog" *vs.* GT: "girl interacts with dog"), which may be underrecognized by current metrics — pointing to potential improvements in future evaluation design.

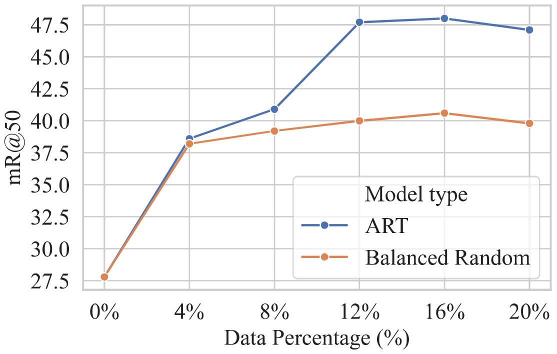

Figure 5. **Training data subsampling analysis.** We plot mR@50 for ART and Naive-RT (balanced random) as a function of the training data percentage used (y-axis) on Vicuna model variants.

Algorithm 2 Balanced Sampling Module

```
Input:
    $\mathcal{P} = \{p_i\}_{i=1}^N$    ▷ Set of $N$ predicate categories
    $\mathcal{S}_{\text{train}} \leftarrow \emptyset$    ▷ Initial training set
    $\mathcal{S}_{\text{pool}}$    ▷ Remaining dataset (excluding $\mathcal{S}_{\text{train}}$)
    $N_p \leftarrow$ Available samples for each predicate $p_i$
    $B \leftarrow$ Total sampling budget per iteration
Output: Updated training set $\mathcal{S}_{\text{train}}$ and pool set $\mathcal{S}_{\text{pool}}$
Initialization:
    $B'_p \leftarrow 0, \quad \forall p \in \mathcal{P}$    ▷ Initialize per-predicate budget
    $i \leftarrow 1$
while $B > 0$ do
    if $N_{p_i} > 0$ then
        $B'_{p_i} \leftarrow B'_{p_i} + 1$    ▷ Allocate one sample to predicate $p_i$
        $N_{p_i} \leftarrow N_{p_i} - 1$    ▷ Decrement available samples
        $B \leftarrow B - 1$    ▷ Reduce remaining budget
    end if
    $i \leftarrow (i + 1) \mod N$    ▷ Move to the next predicate
end while
Assign samples to training set:
for each predicate $p_i \in \mathcal{P}$ do
    $\mathcal{S}_{\text{train}} \leftarrow \mathcal{S}_{\text{train}} \cup \{$Randomly selected $B'_{p_i}$ samples from $\mathcal{S}_{\text{pool}}\}$
    $\mathcal{S}_{\text{pool}} \leftarrow \mathcal{S}_{\text{pool}} \setminus \mathcal{S}_{\text{train}}$
end for
```

C. Balanced sampling

As described in Sec. 3.2 of the main paper, the ART pipeline begins with a balanced sampling algorithm, described in Algorithm 2, to provide an unbiased and balanced understanding of the relations during the initial loop. This step ensures that the subsequent adaptive sampling loop is better guided to select informative samples rather than being influenced by the biases of the underlying data distribution. The balanced sampling distributes a fixed sampling budget across multiple predicates fairly. At first, the predicates are sorted in descending order of frequency, and their allocated budgets are set to zero. The algorithm then allocates a single sampling slot to a predicate with a non-zero frequency, decreases its frequency (*i.e.*, availability), and reduces the remaining budget. If a predicate's availability is exhausted, the algorithm skips it and continues assigning slots to the remaining predicates in a round-robin manner. This ensures that sampling focuses on predicates whose availability has not been exhausted while maintaining a balanced distribution as much as possible.

D. Understanding ART through examples

To help understand the inner workings of Adaptive Relation Tuning (ART), we illustrate the core sampling choices that guide learning. ART's goal is to adapt vision-language models for robust and generalizable visual relation detection. It does this by selecting training instances that are not only informative but also help the model learn from its weaknesses.

We categorize predictions into three groups — True Positives (TP), False Negatives (FN), and False Positives (FP) — and strategically sample from each using a combination of entropy (model uncertainty) and semantic similarity (to ground truth). Below, we explain the reasoning behind each sampling choice with concrete examples:

D.1. High-Entropy True Positives (TPs): Improve uncertain correct predictions

These are predictions where the model gets the relation right, but shows uncertainty (high entropy) in doing so. Including them in training reinforces correct behavior and improves model confidence.

Example: The model correctly predicts "boy riding bike" but assigns nearly equal probability to "boy on bike". This shows uncertainty despite being correct. Sampling this TP helps the model reinforce the right prediction with more certainty.

D.2. Low- and High-Entropy False Negatives (FNs): Correct missed relations

False Negatives occur when a relation exists in the ground truth, but the model says that no prominent relation exists.

Example: If the ground truth is "man holding umbrella", the model may either hesitate (high entropy) or confidently predict "no prominent relation exists" (low entropy). Both cases are important — uncertain misses highlight confusion, while confident misses expose overfitting or bias. Sampling both types improves robustness.

D.3. Low-Similarity False Positives (FPs): Penalize semantically incorrect predictions

False positives are predicted relations that do not appear in the ground truth. However, not all FPs are equally harmful. Some are semantically close — or even more descriptive — and may still reflect a correct understanding of the scene. Others are misleading and indicate poor generalization.

Example: Given the ground truth "man in canoe", predicting "man under canoe" is a low-similarity FP — it is spatially incorrect and misleading. On the other hand, predicting "man paddling canoe" is a high-similarity FP that, while not an exact match, is semantically rich and even more informative than the original label. ART distinguishes between such cases and focuses on refining the misleading ones.

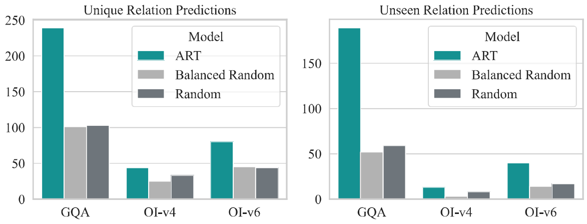

Figure 6. **Comparison of unique relation predictions** *(left)* **and unseen relation predictions** *(right)* for the ART (FlanT5) model across different datasets.

These sampling decisions are made adaptively per predicate using dynamically computed thresholds (based on per-predicate entropy and similarity distributions). This ensures flexible and targeted learning.

E. Analysis of predicted relations

To further evaluate the effectiveness of ART in predicting informative, diverse, and unseen relations, we compared its predictions against random and balanced random baseline methods for both Vicuna [49] (see Fig. 4) and FlanT5 [4] model variants. As discussed in Sec. 4, ART's superiority in predicting diverse and unseen relations extends from Vicuna to FlanT5. Fig. 6 (left) illustrates the total number of unique relations predicted by ART and its baselines. As shown, ART consistently predicts a greater variety of relations across all datasets. A similar pattern can be observed in Fig. 6 (right), where ART predicts more relations unseen during training on VG.

Notably, GQA has the most test samples, followed by OI-v6 and OI-v4, leading to variations in total predictions. The larger GQA test set allows inference across broader scenarios, increasing the likelihood of predicting more diverse and unseen relations.

F. Computational cost and predictive performance

In this section, we analyze both the computational characteristics and the predictive performance behavior of ART. We provide a breakdown of training and inference time, examine the trade-off between data usage and predictive performance, and explain the observed drop in R@k metrics due to biased relation distributions in evaluation datasets.

F.1. Computational cost

As depicted in Tab. 6, while ART incurs higher training costs due to adaptive sampling, it does not increase inference time, making it practical for real-world deployment. The added training complexity is offset by ART's superior generalization, ensuring improved relation prediction without sacrificing efficiency during inference. This trade-off is crucial, as ART enhances mean Recall (mR) by prioritizing informative samples, ultimately leading to a more robust VRD model that generalizes well to unseen data.

Table 6. **Comparison of training and inference time** on a single A100 GPU.

Method	**Train (hrs)**	**Inference (sec/Itr)**
SGGs (Motifs, VTransE, VETO)	18–22	0.07–0.075
VLM (Random/Balanced)	32	0.45
VLM (Adaptive)	96	0.45

F.2. Computational cost *vs.* performance trade-off

As shown in Fig. 7, using just 12% of the training data provides an excellent trade-off between computational cost and predictive performance. This setting achieves near-peak accuracy while requiring only 1.5 days of training on four DGX-A100 GPUs. Beyond this point, additional data yields diminishing returns.

Notably, the 0% baseline incurs negligible computational cost but delivers limited predictive performance, whereas the 12% configuration offers substantial gains at a reasonable expense.

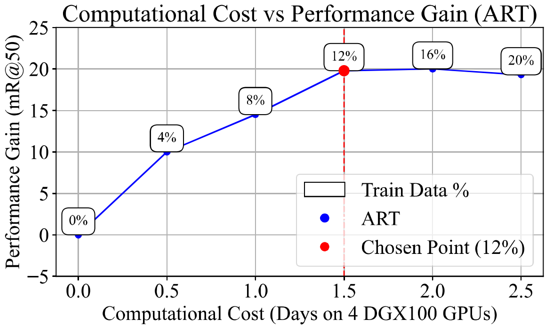

Figure 7. **Trade-off between computational cost and predictive performance** as a function of training data usage.

F.3. On R@k and gR@k performance trade-offs

While ART achieves strong generalization and diversity, it may show lower R@k and gR@k compared to random baselines in cases where ground-truth annotations are skewed toward frequent but semantically shallow relations. Random sampling tends to exploit this bias by favoring head-predicate predictions, leading to inflated recall scores without improving meaningful understanding. In contrast, ART explicitly counteracts this bias through balanced and adaptive sampling, resulting in more informative and diverse predictions. This is evidenced by the higher number of unique and unseen relations predicted by ART across datasets (Figs. 4 and 6) and illustrated qualitatively in the relation prediction examples (Sec. H).

Table 7. **Adaptive *vs.* fixed thresholding.** t_{FP}: low-similarity FP threshold, t_{FN}: low-entropy FN threshold, h_{FN}: high-entropy FN threshold, h_{TP}: high-entropy TP threshold. From the mid point threshold, we increase (higher-h) or decrease (lower-t) the respective thresholds to analyse the effect of fixed thresholds.

Fixed Thresholding	t_{FP}	t_{FN}	h_{FN}	h_{TP}	R@20/50	mR@20/50
Lower-t	0.9	0.25	0.5	0.5	**42.1/42.7**	44.8/45.9
Mid point	0.95	0.5	0.5	0.5	37.2/37.4	43.2/44.3
Higher-h	0.95	0.5	0.75	0.75	34.5/34.8	44.7/45.9
Lower-t Higher-h	0.9	0.25	0.75	0.75	34.5/34.9	44.7/45.9
Adaptive Thresholding					41.1/41.4	**46.4/47.7**

G. Additional analysis: Adaptive *vs.* fixed thresholding

As depicted in Tab. 7, we begin with fixed midpoints (2$^{\text{nd}}$ row) for threshold values: 0.5 for entropy scores (h_{FN}, t_{FN}, t_{FP}) and 0.95 for similarity scores (t_{FP}). The higher similarity threshold accounts for the fact that similarity is computed on predicate phrases, not standalone predicates, resulting in generally higher values (*e.g.*, Fig. 3, instance Ⓑ). We then lower t and increase h from their midpoint values to explore fixed thresholding. However, all fixed-threshold variations yield lower mR than adaptive thresholding, highlighting the difficulty of selecting an optimal fixed threshold. In contrast, adaptive thresholding dynamically adjusts per predicate, ensuring optimal tuning of the VLM.

H. Qualitative results

Next, we present some qualitative results of ART on downstream segmentation, followed by a comparative analysis of relation predictions between ART and its baselines.

H.1. ART-enhanced segmentation reasoning

As shown in Fig. 8, ART-enhanced scene graphs enable DeiSAM [32] to produce higher-quality segmentations. While ground-truth scene graphs fail to capture the relations in the segmentation prompt, ART's unseen relation prediction allows DeiSAM to accurately segment the referenced object in the deictic prompt.

H.2. Comparative analysis of ART and its baselines

We compare the relationship predictions from the ART Vicuna model against its strongest baseline, Naive-RT (Naive Relation Tuning), which includes Naive-RT (balanced random) and Naive-RT (random), as well as the ground truth. Predictions are evaluated on the GQA, OI-v4, and OI-v6 test sets. Examples are shown in Figs. 9 to 14.

Overall, we observe that ART not only identifies new relationships but also produces predictions that are more meaningful than existing ground-truth annotations. Predictions that are either similar to or more meaningful than the ground-truth annotation are highlighted in green, while those that are both informative and unseen are additionally highlighted in yellow. Incorrect predictions are marked in red.

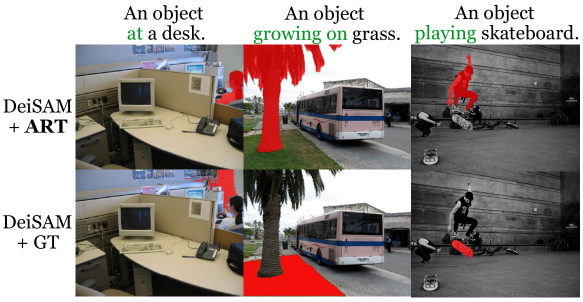

Figure 8. **ART can be used to label missing annotations and predict new unseen predicates.** Segmentation results with textual prompts *(top)* using DeiSAM [32], which segments objects via reasoning on scene graphs. ART successfully detects new relations and improves the segmentation quality, while ground-truth scene graphs fail to capture relations in the prompt.

ART consistently outperforms its baselines across the GQA, OI-v4, and OI-v6 datasets by providing more meaningful and informative relationship predictions. For example, on the GQA dataset (see Fig. 9), ART predicts the sensible spatial relation *under* between water and sky and more detailed interactions such as *water reflecting sky* and *boat sailing under sky*, while Naive-RT (random) and (balanced random) perform poorly. Fig. 10 highlights that ART predicts the more descriptive interaction *swimming in* between the animal and water, whereas the Naive-RT baselines, as well as the ground truth, only identify the spatial relation *in*. On the OI-v4 dataset (see Fig. 11), ART clarifies ambiguous ground-truth relations like *interacts with*, which raises the question "What kind of interaction?" by providing clarity that the interaction is *petting* and also predicts the spatial relation *near*, whereas Naive-RT baselines fail to provide clarity. Similarly, in Fig. 12, the ground-truth relation *holds* between man and beer raises the question, "What does he intend to do with the beer?" This ambiguity is resolved by ART's prediction of the more specific relation *drinking*. On the OI-v6 dataset (see Fig. 13), ART identifies the action *paddling* between man and canoe, along with the spatial relation *in*, outperforming the baselines, which lack specificity in describing the interaction. Another example from the OI-v6 dataset, shown in Fig. 14, once again shows that the ground-truth relation *contains* between mug and beer is less detailed compared to ART's prediction of *filled with*, which conveys that the mug is full or nearly full of beer. The reasonable predictions *with* and *holding* made by Naive-RT (random) and (balanced random) are also less descriptive.

Overall, the qualitative examples support the substantial quantitative gains reported in Tab. 1 of the main paper.

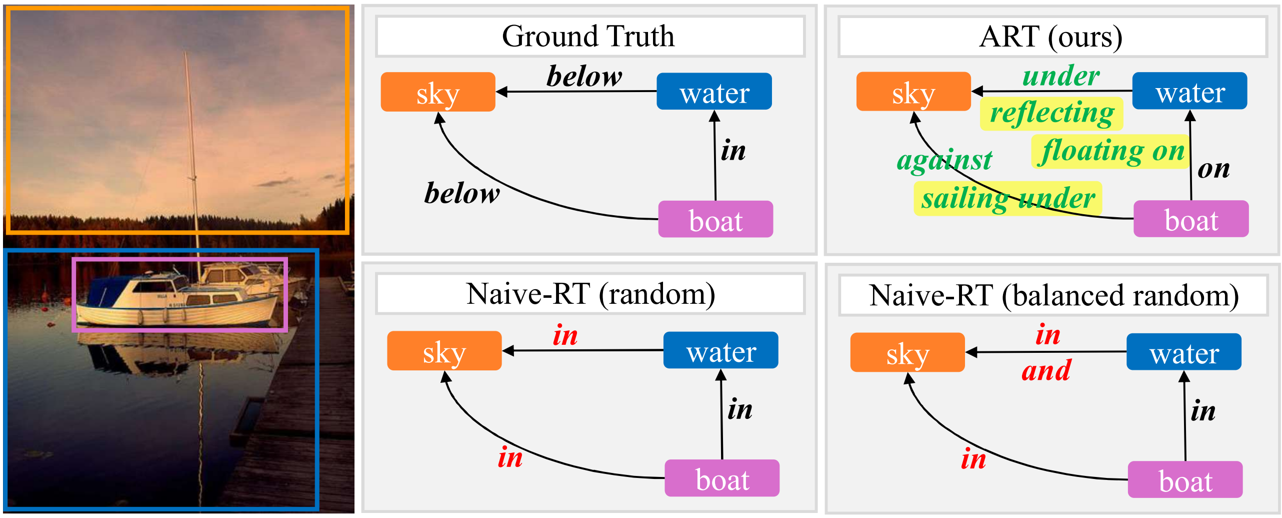

Figure 9. **Comparison of ART and its baselines on the GQA dataset.** ART predicts sensible spatial relations similar to the ground-truth annotation such as, *water under sky*, while also identifying more informative relations than the ground truth, such as *water reflecting sky*, *boat floating on water*, and *boat sailing under sky*. In contrast, both Naive-RT (random) and (balanced random) perform poorly. Informative relation predictions are highlighted in green, while those that are both informative and unseen are additionally highlighted in yellow. Incorrect predictions are marked in red.

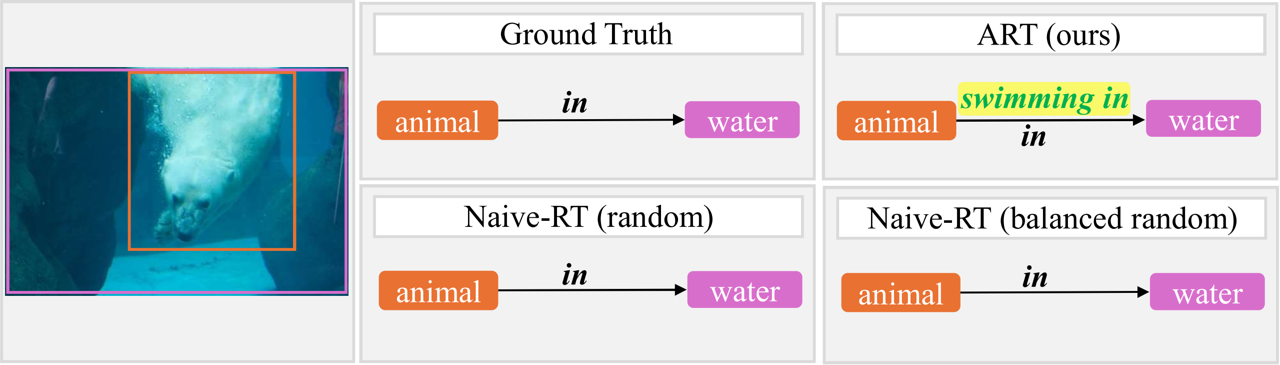

Figure 10. **Comparison of ART and its baselines on the GQA dataset.** The ground truth only provides a spatial relation *in* between animal and water, while ART predicts the descriptive interaction *swimming in*. Informative relation predictions are highlighted in green, while those that are both informative and unseen are additionally highlighted in yellow.

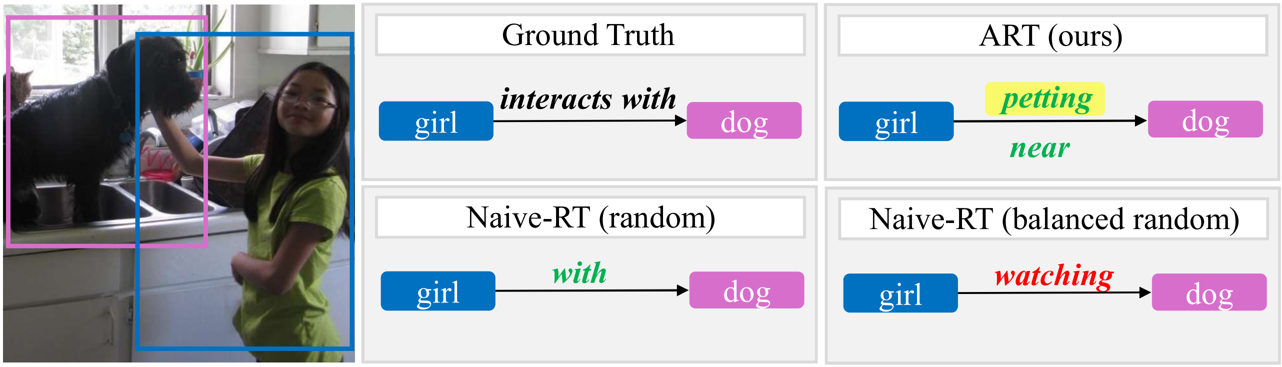

Figure 11. **Comparison of ART and its baselines on the OI-v4 dataset.** In contrast to the provided less informative ground-truth relation *interacts with* in *girl interacts with dog*, which raises the question "What kind of interaction?", ART provides a much clearer interpretation that the interaction is *petting*, *i.e. girl petting dog*, while also identifying the sensible spatial relation *near*. In contrast, while Naive-RT (random) suggests the less meaningful relation *with*, Naive-RT (balanced random) produces an entirely incorrect prediction. Informative relation predictions are highlighted in green, while those that are both informative and unseen are additionally highlighted in yellow. Incorrect predictions are marked in red.

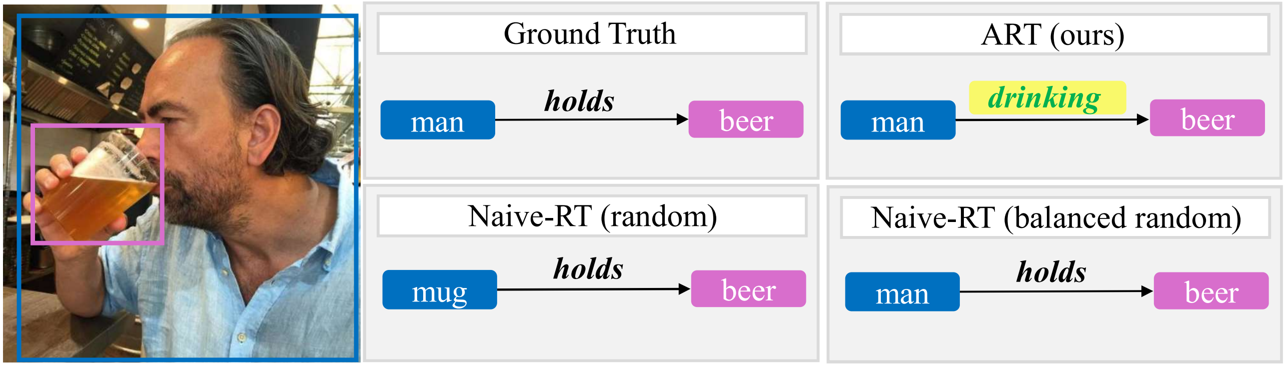

Figure 12. **Comparison of ART and its baselines on the OI-v4 dataset.** The ground truth relation *holds* between *man* and *beer* leaves an open question "What he intends to do with the beer?", while the prediction *drinking* made by ART gives more context and the ongoing action. The Naive-RT baselines also predict the less descriptive relation *holds*. Informative relation predictions are highlighted in green, while those that are both informative and unseen are additionally highlighted in yellow.

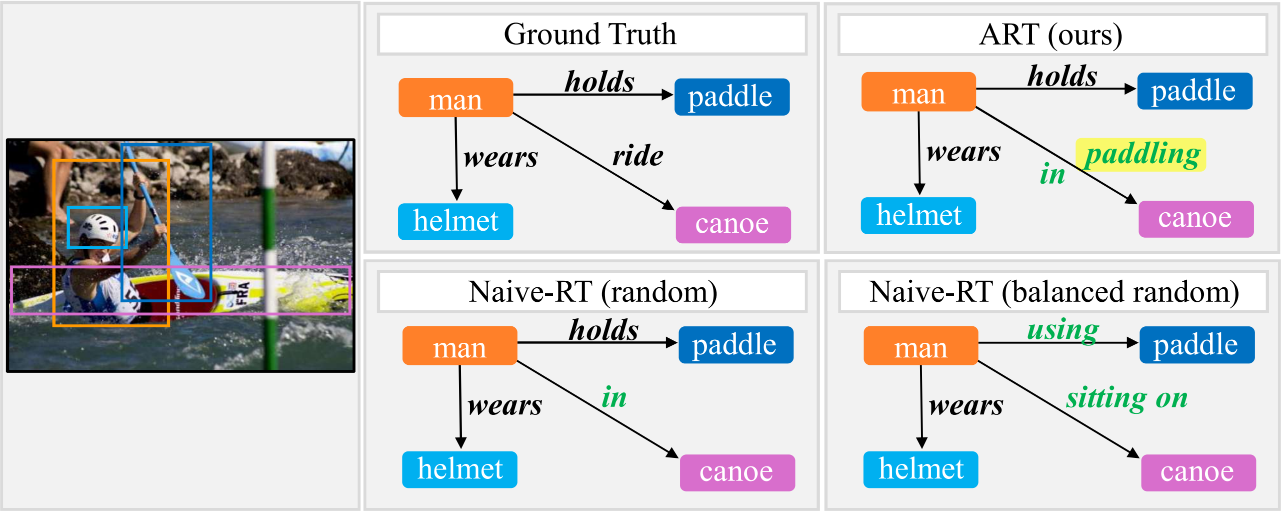

Figure 13. **Comparison of ART and its baselines on the OI-v6 dataset.** ART predicts the more informative relation *paddling* between the man and canoe while also identifying sensible spatial relation *in*. In contrast, although both Naive-RT (random) and (balanced random) make reasonable spatial predictions, they fail to clarify the action taking place between the man and the canoe. Informative relation predictions are highlighted in green, while those that are both informative and unseen are additionally highlighted in yellow.

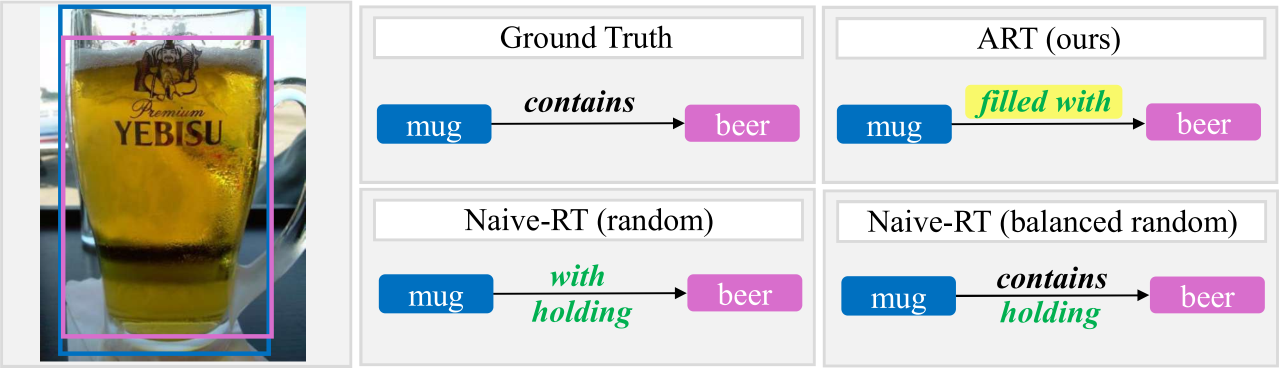

Figure 14. **Comparison of ART and its baselines on the OI-v6 dataset.** While the ground-truth relation *contains* between the mug and beer merely indicates the presence of beer in the mug, the relation *filled with*, predicted by ART, provides more detail by suggesting that the mug is full or nearly full of beer. The predictions *with* and *holding*, made by Naive-RT (random) and (balanced random) respectively, are reasonable but lack the level of descriptiveness conveyed by *filled with*. Informative relation predictions are highlighted in green, while those that are both informative and unseen are additionally highlighted in yellow.