ART: Adaptive Relation Tuning for Generalized Relation Prediction

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)

https://github.com/visinf/ART

Abstract

*Visual relation detection (VRD) is the task of identifying the relationships between objects in a scene. VRD models trained solely on relation detection data struggle to generalize beyond the relations on which they are trained. While prompt tuning has been used to adapt vision-language models (VLMs) for VRD, it uses handcrafted prompts and struggles with novel or complex relations. We argue that instruction tuning offers a more effective solution by fine-tuning VLMs on diverse instructional data. We thus introduce ART, an **A**daptive **R**elation **T**uning framework that adapts VLMs for VRD through instruction tuning and strategic instance selection. By converting VRD datasets into an instruction-tuning format and employing an adaptive sampling algorithm, ART directs the VLM to focus on informative relations while maintaining generalizability. Specifically, we focus on the relation classification, where subject-object boxes are given and the model predicts the predicate between them. We tune on a held-in set and evaluate across multiple held-out datasets of varying complexity. Our approach strongly improves over its baselines and can infer unseen relation concepts, a capability absent in mainstream VRD methods. We demonstrate ART's practical value by using the predicted relations for segmenting complex scenes.*

1. Introduction

Visual relation detection (VRD) plays a key role in visual scene understanding by enabling the recognition of relationships between entities as triplets of ⟨subject, predicate, object⟩ [10, 22, 35, 44]. This structured understanding of visual content is essential for downstream tasks like visual question answering (VQA) [2], image captioning [11], and referring object detection and segmentation [32]. While mainstream VRD models have advanced through relation-centric training, they predominantly rely only on fixed datasets such as Visual Genome [14] and GQA [12]. This dependence introduces three key limitations: *(i)* Overfitting to frequent in-distribution relations, weakening performance on rare relations [3]; *(ii)* Inability to infer novel, unseen relations; *(iii)* Coarse annotations that fail to capture fine-grained relational semantics. For example, given the scene in Fig. 1, GQA annotations describe generic relations like *"boat in water"* or *"flag on boat"*. A model trained solely on these overlooks finer distinctions, such as whether the boat is *floating on water* or the flag is positioned *on the back of the boat*. Humans naturally infer such nuances using broad context, yet, manually extending VRD datasets with fine-grained annotations is costly and impractical, making it crucial to develop models that generalize beyond predefined relations without exhaustive retraining.

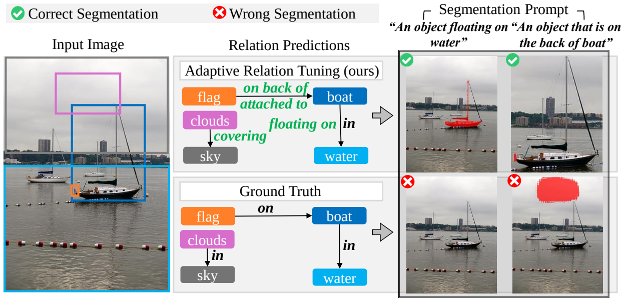

Figure 1. **ART predicts detailed, context-rich relationships, enhancing downstream reasoning, *e.g.*, for segmentation.** From the input image, ART (trained on VG) predicts unseen and informative relations like *floating on* and *on back of* (highlighted in green). This richer relational context, facilitated by careful tuning of a VLM, *e.g.*, allows the DeiSAM [32] segmentation model, which relies on scene graphs for spatial reasoning, to accurately identify even small objects like the flag on the boat's back. When prompted to find an object floating on water, the ART scene graph also enables correct segmentation, while the less specific ground-truth scene graphs [12] may miss such segmentation.

A promising alternative is to leverage vision-language models (VLMs), which excel in generalization across multimodal tasks by learning from large-scale image-text corpora [1, 5, 24, 29]. Thus, VLMs could help overcome VRD

To appear in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Honolulu, Hawai'i, USA, 2025.

generalization challenges. Although recent studies have attempted to extract knowledge from VLMs to enhance VRD [10, 22], they rely on off-the-shelf VLMs through prompt engineering, crafting prompts tailored to the cues of an in-distribution or otherwise limited set of relations.

To address these challenges, we reframe the VRD task as an instruction tuning (IT) problem, leveraging the proven ability to enhance VLMs [5, 39]. We first generate instruction-tuning instances from established VRD benchmarks, followed by fine-tuning with these tailored instances. We propose a novel framework for adaptively selecting the most informative instances for relation tuning. Our approach, named Adaptive Relation Tuning (ART), enhances the relation classification capabilities of VLMs by focusing on the most informative aspects of the data. This is crucial for safety-critical scenarios like autonomous driving, where distinguishing subtle relational differences, *e.g.*, "pedestrian waiting at crosswalk" versus "pedestrian stepping onto crosswalk," can be life-saving. By enabling the detection of fine-grained relations while preserving generalization, ART moves beyond simplistic relational inference, enabling more robust and adaptable visual reasoning.

Our key contributions are: *(i)* We convert benchmark relation detection datasets into an effective format for instruction tuning. *(ii)* We present ART, an innovative framework for relation tuning of VLMs, which preserves the model's generalization by adaptively selecting informative samples. *(iii)* We train ART on a held-in dataset and evaluate it on held-out datasets of varying complexity. Our quantitative analysis highlights ART's strong generalization capabilities. *(iv)* We deploy ART on a downstream segmentation task, where both quantitative and qualitative results highlight its real-world effectiveness.

2. Related work

Visual relation detection (VRD). The prediction of relationships between subject-object pairs has been widely studied in the domain of scene graph generation (SGG) [23, 26, 36, 44] and human object interaction (HOI) [9, 13, 38, 41]. Previous work largely focused on learning the object and predicate categories in the training data distribution and testing on the same distribution. However, such models suffer from noisy annotations [21] and biased predictions resulting from the long-tailed predicate distribution [34, 36, 48]. Thus, current mainstream models fall short in their ability to generalize beyond the specific relations they were trained on, highlighting a critical gap in the field. This underscores the need for innovative approaches that can enhance model robustness and generalization capabilities, enabling effective inference over completely unseen object classes and relationships. Recently, some efforts were made in this direction by adopting prompt tuning [25] of VLMs [10, 22]. However, these approaches rely on off-the-shelf VLMs with prompts tailored to limited relational cues. Conversely, instruction tuning [5] offers a more comprehensive solution by fine-tuning models on diverse instructional data, significantly enhancing their ability to generalize and infer unseen relationships. Hence, we here adapt VLMs for VRD with adaptive instruction tuning.

Uncertainty estimation is essential in setups that require informative sample selection. At a high level, uncertainty can be classified into aleatoric (data) and epistemic (model) uncertainty. Aleatoric uncertainty reflects intrinsic variability in the data, while epistemic uncertainty arises from the model's limited understanding of the underlying task. Some of the commonly used uncertainty estimators are Monte-Carlo dropout [8], deep ensembles [7], and the entropy of the softmax predictions [27]. These uncertainty measures are used predominately to compute model uncertainty. In our work, we use entropy as an uncertainty measure due to its cost-effective use in a VLM setup compared to dropout and ensembles. We also use a cosine similarity-based approach [18] as an additional uncertainty measure.

Active learning (AL) iteratively selects and annotates the most informative instances from an unlabeled pool to improve training [28, 33, 47]. Kung *et al.* [15] proposed selecting informative instructions using an active learning-inspired setup. Our work is related to these techniques but differs in a key aspect: we assume the pool data is already labeled. Our goal is to identify the most informative samples within the provided distribution to achieve optimal performance in out-of-distribution scenarios, while AL aims for optimal in-distribution performance with limited data.

3. Adaptive relation tuning framework

VRD models must generalize well to be effective in real-world scenarios. Yet, models trained solely on VRD datasets either overfit the underlying distribution or learn spurious correlations. Though VLMs excel in generalization due to their large training corpora, they often struggle with fine-grained relations, favoring frequent patterns. To address this, ART follows a two-step process: *(1)* Relation-tuning data creation, where the training data is structured with high-level relation categories (spatial, semantic, possessive) and a counter-negative mining strategy to refine negative samples; and *(2)* Adaptive relation tuning of VLMs, which optimizes the model to focus on rare, diverse relations, improving generalization to unseen scenarios.

Typically, VRD involves object detection followed by relation classification. While generalization in object detection—such as zero-shot or cross-domain detection—has been widely studied [31, 42, 43], we focus on relation classification. Given object categories and bounding boxes, we predict the relation between subject-object pairs, denoted as s (subject), o (object), and p (predicate/relation).

Question: Is there a prominent $\langle r\rangle$ relation between $\langle s\rangle$ (subject) and $\langle o\rangle$ (object) in the image?
Positive response: Yes, $\langle s, r, o\rangle$.
Negative response: No, there is no prominent $\langle nr\rangle$ relation between $\langle s\rangle$ and $\langle o\rangle$.

Figure 2. **Relation tuning format template.**

3.1. Relation-tuning data creation

A key aspect of our approach is relation-tuning data construction, which involve carefully crafting questions to capture fine-grained object relations. A simple transformation of the ground-truth relation triplets $\langle s, p, o\rangle$ into a question format such as "Is there a relation between $\langle s\rangle$ and $\langle o\rangle$ in the image?"—with the answers "Yes, $\langle s, p, o\rangle$" or "No" for valid or invalid relations, respectively, might seem sufficient. However, this naive approach overlooks important nuances in relation classification, as shown below (Tab. 2 in Sec. 4.2). To address this, we introduce a more structured approach to relation-tuning data creation, where we design a question format that incorporates two key components: *(1)* high-level relation categories and *(2)* counter-negative mining. The final relation tuning format template, shown in Fig. 2, is derived from this structured formulation, which is empirically supported by findings in Tab. 2.

Prevalent VRD relations can be classified into high-level relation types [44] such as spatial (*above, behind, under, . . .* , denoting spatial arrangements of objects), semantic (*carrying, eating, using, . . .* , corresponding to activities), and possessive (*wearing, part of, has, . . .* , depicting a sense of ownership). These high-level relation types can help the model learn a richer set of relations. Furthermore, because multiple relations can co-exist between subject-object pairs, we carefully construct questions to capture these complexities. *E.g.*, the phrase "person walking on street" has both semantic (action) and spatial (location) relations, and a broad question format might fail to capture possible predictions.

Another challenge arises from the incompleteness of VRD benchmark annotations, where many potential object relations remain unlabelled. Treating all unlabelled pairs as negatives can mislead the model and degrade performance. To address this, we introduce a counter-negative mining approach that selectively assigns negatives based on the relation type. We assume that *semantic* vs. *possessive* and *spatial* vs. *possessive* relations are mutually exclusive—semantic and spatial relations serve as negatives for possessive ones, and vice versa. *E.g.*, in *person wearing hat* (possessive), spatial orientation is implicit, while action verbs are less relevant, as possession describes a static state rather than an active interaction. In contrast, semantic relations often depend on spatial positioning, which can vary and impact their meaning. *E.g.*, in *person watching TV*, the spatial position of the subject relative to the object is crucial—whether they are near, far, or beside the TV changes the nature of the interaction. Thus, *semantic* vs. *spatial* is not treated as mutually exclusive. Specifically, we define the following exclusivity for high-level relation categories, grounded in the above assumptions and empirically validated through our instruction set analysis in Tab. 2:

$$\text{positive: } r \in \{\text{spatial}, \text{possessive}, \text{semantic}\}, \tag{1}$$

$$\text{negative: } nr = \begin{cases} \{\text{possessive}\}, & \text{if } r = \text{spatial}, \\ \{\text{spatial}, \text{semantic}\}, & \text{if } r = \text{possessive}, \\ \{\text{possessive}\}, & \text{if } r = \text{semantic}, \end{cases}$$

where r and nr denote positive resp. negative categories.

3.2. Adaptive relation tuning (ART)

Given that we have converted a benchmark VRD dataset into instruction-tuning format (*cf*. Sec. 3.1), a straightforward approach would be to tune the VLM on the full instruction set. However, this leads to overfitting and loss of generalization (*cf*. Tab. 3 below). To prevent this, we propose adaptive relation tuning (ART), which selectively refines structurally relevant relational patterns while maintaining the VLM's broad generalization for VRD.

Inspired by active learning, ART aims to select the most informative samples through multiple learning iterations, as illustrated in Fig. 3. While both prioritize uncertain instances, ART differs by operating on a fully labeled dataset rather than an unlabeled pool. We aim to achieve optimal out-of-distribution performance without overfitting to in-distribution data. As depicted in Fig. 3, ART comprises three key components (highlighted in orange): a balanced sampling module, an adaptive sampling module, and an uncertainty and dissimilarity computation module.

Balanced sampling module. ART operates through multiple learning iterations, selecting a fixed budget B of training samples per loop to refine the model. However, VRD datasets, such as Visual Genome (VG) [14], are highly imbalanced, dominated by a few frequent classes. If the initial training samples were randomly selected, the model would be biased toward these head classes, limiting its ability to learn rare relations. To counteract this, our balanced sampling module (see Algorithm 2 in Appendix C) serves as the initialization step, ensuring an even distribution of relations at the start of training. Let $\mathcal{P} = \{p_i\}_{i=1}^{N}$ represent the set of predicates, where N is the number of predicate categories, and let B'_p and N_p represent the sampling budget and available data samples per predicate. B'_p is evenly assigned across predicates, ensuring all are included. If N_p is

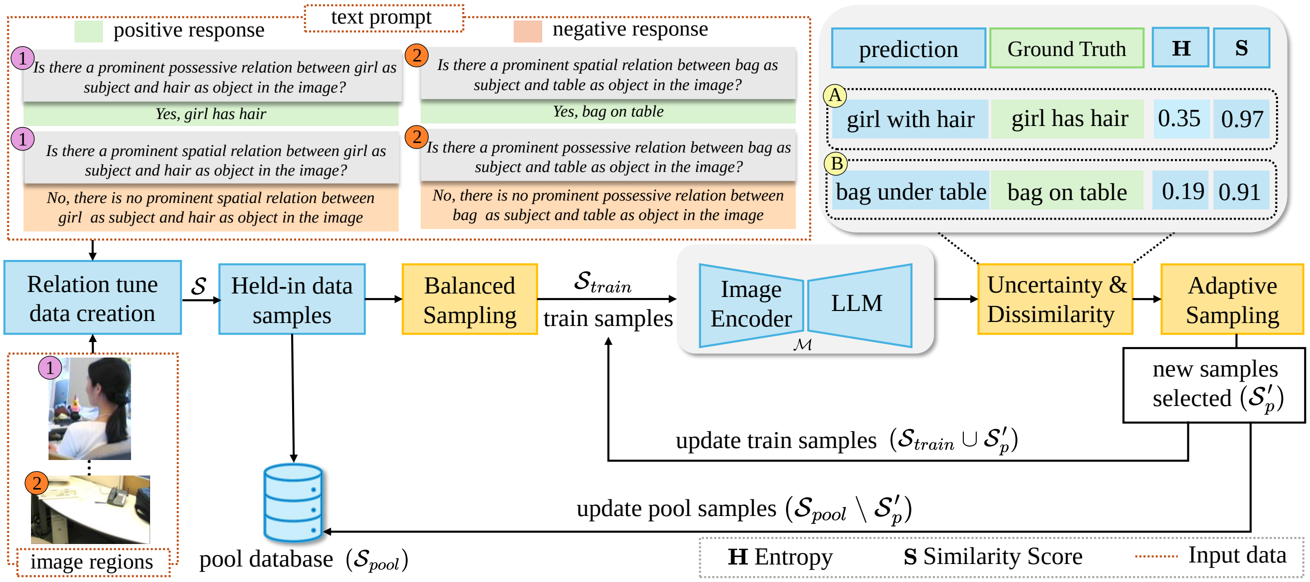
text prompt
positive response
negative response
1
Is there a prominent possessive relation between girl as subject and hair as object in the image?
Yes, girl has hair
1
Is there a prominent spatial relation between girl as subject and hair as object in the image?
No, there is no prominent spatial relation between girl as subject and hair as object in the image
2
Is there a prominent spatial relation between bag as subject and table as object in the image?
Yes, bag on table
2
Is there a prominent possessive relation between bag as subject and table as object in the image?
No, there is no prominent possessive relation between bag as subject and table as object in the image
prediction
Ground Truth
H
S
A
girl with hair
girl has hair
0.35
0.97
B
bag under table
bag on table
0.19
0.91
Relation tune data creation
$\mathcal{S}$
Held-in data samples
Balanced Sampling
$\mathcal{S}_{train}$
train samples
Image Encoder
LLM
$\mathcal{M}$
Uncertainty & Dissimilarity
Adaptive Sampling
new samples selected ($\mathcal{S}'_p$)
update train samples ($\mathcal{S}_{train} \cup \mathcal{S}'_p$)
update pool samples ($\mathcal{S}_{pool} \setminus \mathcal{S}'_p$)
1
2
image regions
pool database ($\mathcal{S}_{pool}$)
H Entropy
S Similarity Score
Input data

Figure 3. **ART Framework.** To construct the relation-tuning dataset $\mathcal{S}$, we create question-response prompts for each image region (*e.g.*, ① and ②), including both positive and negative examples, which we call instruction sets. Balanced sampling ensures that instruction sets span all predicate categories in $\mathcal{S}$, forming the training data $\mathcal{S}_{\text{train}}$. This training data is used to fine-tune the vision-language model (VLM), $\mathcal{M}$, while unused data is stored in the pool $\mathcal{S}_{\text{pool}}$. After training, $\mathcal{M}$ is evaluated on $\mathcal{S}_{\text{pool}}$, where uncertainty and dissimilarity are estimated using entropy $\mathbf{H}$ and similarity scores $\mathbf{S}$. In example Ⓐ, the model assigns higher entropy to a similar but uncertain prediction compared to an incorrect prediction Ⓑ, but similarity scores help prioritize the incorrect example for further training. These estimates guide adaptive sampling, creating a refined set of samples $\mathcal{S}'_p$ for each predicate p to expand $\mathcal{S}_{\text{train}}$, enabling further rounds of training for $\mathcal{M}$.

exhausted for a predicate p, sampling continues from the remaining predicates until B is met. This prevents early overfitting to frequent relations, allowing adaptive sampling to refine predicate selections in later iterations.

Adaptive sampling module. Only using balanced sampling can still miss out on informative samples, as the model quickly learns some relations while struggling with others. To address this, we propose *adaptive sampling*, which prioritizes samples where the VRD model lacks confidence. This approach selects the necessary samples from each category based on how well it performs and how uncertain the model is about the samples in each category.

Uncertainty and dissimilarity computation module. Inspired by uncertainty-based active learning [27], we leverage entropy to measure the uncertainty of the VRD model's output. To achieve this, the output from the LLM component of the VLM is further processed. Since our VLMs employ beam search, we assess uncertainty across multiple possible sequences. Let M be the beam width, a hyperparameter controlling the number of candidate sequences. To comprehensively understand the model's uncertainty, we compute the entropy over all possible sequences. Let $\mathbf{E} \in \mathbb{R}^{M\times L\times V}$ represent the output logits from the LLM, where L is the sequence length and V is the vocabulary size. We calculate the overall entropy $\mathbf{H}$ of the predicted sentence as $\mathbf{H} = -\frac{1}{M\cdot L}\sum_{m,l,v}\mathbf{P}_{mlv}\log(\mathbf{P}_{mlv})$, where $\mathbf{P}_{mlv} = \frac{\exp(\mathbf{E}_{mlv})}{\sum_{v'}\exp(\mathbf{E}_{mlv'})}$ is the softmax probability for each vocabulary element v. This formula averages the entropy across all beams and sequence positions, providing a single entropy value that summarizes the uncertainty regarding the words in the predicted sentence.

As illustrated in Fig. 3 (*cf*. Ⓐ and Ⓑ), while entropy can indicate model-level uncertainty, it may also produce low entropy values for overly confident predictions. Relying solely on entropy could result in neglecting informative samples. Since our objective is to develop a generalizable VRD model capable of handling out-of-distribution data, retaining predictions that semantically align with the ground truth is preferable. For instance, the phrases *girl with hair* and *girl has hair* are semantically similar; hence, such False Positives can be considered valid and sampled less often. Conversely, the prediction *bag under table* for the ground truth *bag on table* is evidently invalid, but the model's overconfidence may assign it low entropy, overlooking valuable samples. To overcome this issue, we use a similarity metric $\mathbf{S}$, measuring the cosine similarity between predicted and ground-truth embeddings, ensuring better selection of informative samples.

Adaptive sampling algorithm. We summarize the adaptive sampling algorithm in Algorithm 1. Given a fully labeled pool $\mathcal{S}_{\text{pool}}$, we begin by categorizing the samples into True Positives (TP), False Positives (FP), and False Negatives (FN) by inference on $\mathcal{S}_{\text{pool}}$. A TP occurs when the predicted positive response matches the ground truth. An FP arises when the predicted positive response does not align with the ground truth. An FN happens when the model predicts a negative response for a relation that actually exists in the ground truth. This categorization is essential for identifying areas where the model can be improved. Subsequently, our approach adaptively adjusts sampling thresholds based on entropy and similarity scores of $\mathcal{S}_{\text{pool}}$, ensuring appropriate sample selection at each iteration. The following steps outline the sampling algorithm in detail.

Step 1: Sampling budget allocation. To prioritize predicates requiring greater supervision, we compute a per-predicate sampling probability as $P_p = \frac{1-R_p}{\sum_j (1-R_j)}$, where R_p is the recall of predicate p on the validation set. Each predicate is allocated budget $B'_p = \min(B \cdot P_p, N_p)$, ensuring that predicates with lower recall receive a higher sampling budget while preventing over-sampling beyond available instances (Lines 7–10).

Step 2: Sampling process. We iteratively select informative samples by leveraging predicate-specific uncertainty and dissimilarity. This process consists of three key components: *(i) Adaptive thresholding*, where thresholds are adapted based on entropy $\mathbf{H}$ and similarity $\mathbf{S}$ distributions; *(ii) Sampling criteria*, which define selection rules for TPs, FNs, and FPs based on their statistics; and *(iii) Threshold refinement*, ensuring sufficient sample selection when the allocated budget is not met.

Step 2.1: Adaptive thresholding. Fixed thresholds may not generalize well across predicates due to varying entropy and similarity distributions. To address this, we estimate per-predicate sampling thresholds by fitting normal distributions to entropy $\mathbf{H}$ and similarity $\mathbf{S}$, computing means and standard deviations for TPs, FNs, and FPs (Lines 16–23). These statistics guide adaptive thresholding.

Step 2.2: Sampling criteria. Sample selection ensures that the most informative samples are chosen for fine-tuning. *(i) True Positives (TPs):* We use entropy $\mathbf{H}$ to measure uncertainty. High-entropy TPs ($\mathbf{H}(\mathbf{s}) > h_{\text{TP}}$) indicate cases where the model is uncertain despite correct predictions, suggesting further refinement (Line 26). Therefore, we sample these instances to enhance model confidence and robustness. Conversely, low-entropy TPs represent confidently correct predictions and do not require further sampling. *(ii) False Negatives (FNs):* Both high-entropy ($\mathbf{H}(\mathbf{s}) > h_{\text{FN}}$) and low-entropy ($\mathbf{H}(\mathbf{s}) < t_{\text{FN}}$) FNs are valuable – high-entropy FNs signal model confusion, meaning the model is uncertain about its incorrect predictions, necessitating further training. In contrast, low-entropy FNs expose overconfident misclassifications, where the model is incorrectly certain about a wrong prediction. Addressing both extremes enhances the model's ability to distinguish correct and incorrect relationships (Lines 27–28). *(iii) False Positives (FPs):* Instead of entropy, similarity $\mathbf{S}$ is used, as FPs vary in their semantic closeness to the ground truth. High-similarity FPs ($\mathbf{S}(\mathbf{s}) > t_{\text{FP}}$, *e.g.*, Fig. 3, instance Ⓐ, where GT: "girl with hair" and Pred: "girl has hair") retain meaningful semantics, making them less critical for correction. In contrast, low-similarity FPs ($\mathbf{S}(\mathbf{s}) < t_{\text{FP}}$, *e.g.*, Fig. 3, instance Ⓑ, where GT: "bag on table" and Pred: "bag under table") indicate weak generalization and require

Algorithm 1 Adaptive Sampling Algorithm

```
 1: Input:
 2:     $\mathcal{S}_{\text{pool}} \leftarrow \mathcal{S} \setminus \mathcal{S}_{\text{train}}$
 3:     $B \leftarrow$ Total sampling budget per iteration
 4:     $\mathcal{R} \leftarrow$ Recall per predicate on validation set $\mathcal{S}_{\text{val}}$
 5:     $\text{TP}_{\text{pool}}, \text{FN}_{\text{pool}}, \text{FP}_{\text{pool}}$    ▷ True Positive, False Negative, and False Positive samples in $\mathcal{S}_{\text{pool}}$
 6: Output: Updated training set $\mathcal{S}_{\text{train}}$ and pool set $\mathcal{S}_{\text{pool}}$
 7: Step 1: Sampling Budget Allocation
 8: for each predicate $p$ do
 9:     $P_p \leftarrow \frac{1-R_p}{\sum_j (1-R_j)}$    ▷ Higher weight for lower recall
10:     $B'_p \leftarrow \min(B \cdot P_p, N_p)$    ▷ $N_p$: available data samples for $p$
11: end for
12: Step 2: Sampling Process
13: for each predicate $p$ with allocated budget $B'_p$ do
14:     Initialize $z \leftarrow z_{\text{init}}$, $\mathcal{S}'_p \leftarrow \emptyset$
15:     while $|\mathcal{S}'_p| < B'_p$ do ▷ Continue until budget is met
16:         Step 2.1: Adaptive Thresholding
17:         Compute mean ($\mu$) and std. dev. ($\sigma$) of entropy ($\mathbf{H}$) resp. similarity ($\mathbf{S}$) of the sample pools:
18:             $\mu_{\text{TP}}, \sigma_{\text{TP}} \leftarrow$ Mean, std. dev. of $\mathbf{H}$ in $\text{TP}_{\text{pool}}$
19:             $\mu_{\text{FN}}, \sigma_{\text{FN}} \leftarrow$ Mean, std. dev. of $\mathbf{H}$ in $\text{FN}_{\text{pool}}$
20:             $\mu_{\text{FP}}, \sigma_{\text{FP}} \leftarrow$ Mean, std. dev. of $\mathbf{S}$ in $\text{FP}_{\text{pool}}$
21:         Compute sampling thresholds:
22:         $h_{\text{TP}} \leftarrow \mu_{\text{TP}} + z\sigma_{\text{TP}}$,    $h_{\text{FN}} \leftarrow \mu_{\text{FN}} + z\sigma_{\text{FN}}$
23:         $t_{\text{FN}} \leftarrow \mu_{\text{FN}} - z\sigma_{\text{FN}}$,    $t_{\text{FP}} \leftarrow \mu_{\text{FP}} - z\sigma_{\text{FP}}$
24:         Step 2.2: Sampling Criteria
25:         Select samples ($\mathbf{s}$) with high-entropy TP, high-entropy FN, low-entropy FN, and low-similarity FP
26:         $\mathcal{S}'_p \leftarrow \mathcal{S}'_p \cup \{\mathbf{s} \in \text{TP}_{\text{pool}} \mid \mathbf{H}(\mathbf{s}) > h_{\text{TP}}\}$
27:         $\mathcal{S}'_p \leftarrow \mathcal{S}'_p \cup \{\mathbf{s} \in \text{FN}_{\text{pool}} \mid \mathbf{H}(\mathbf{s}) > h_{\text{FN}}\}$
28:         $\mathcal{S}'_p \leftarrow \mathcal{S}'_p \cup \{\mathbf{s} \in \text{FN}_{\text{pool}} \mid \mathbf{H}(\mathbf{s}) < t_{\text{FN}}\}$
29:         $\mathcal{S}'_p \leftarrow \mathcal{S}'_p \cup \{\mathbf{s} \in \text{FP}_{\text{pool}} \mid \mathbf{S}(\mathbf{s}) < t_{\text{FP}}\}$
30:         Step 2.3: Iterative Threshold Refinement
31:         if $|\mathcal{S}'_p| < B'_p$ then
32:             $z \leftarrow z - 0.1$    ▷ adjust threshold if insufficient samples
33:         end if
34:     end while
35:     Step 3: Updating the Training Set
36:     $\mathcal{S}_{\text{train}} \leftarrow \mathcal{S}_{\text{train}} \cup \mathcal{S}'_p$
37:     $\mathcal{S}_{\text{pool}} \leftarrow \mathcal{S}_{\text{pool}} \setminus \mathcal{S}'_p$
38: end for
```

refinement. Since similarity is undefined for FNs (due to missing predictions) and trivially 1 for TPs (as predictions match the ground truth), it is applied exclusively to FPs. We sample FPs below the similarity threshold to ensure that the model improves on the most misleading cases (Line 29).

Step 2.3: Iterative threshold refinement. If the selected samples fall short of the budget B'_p, the threshold z is iteratively reduced to ensure sufficient yet controlled, per-predicate sampling (Lines 30–33).

Step 3: Updating the training set. The selected samples $\mathcal{S}'_p$ are added to $\mathcal{S}_{\text{train}}$ and removed from $\mathcal{S}_{\text{pool}}$ to ensure non-redundant sampling in future iterations (Lines 35–37).

Refer to Appendix D for an intuitive overview of ART.

4. Experiments

We aim to answer the following questions: **(Q1)** Does ART outperform its baselines on in-distribution samples with unseen predicates? **(Q2)** How well does it handle out-of-distribution data? **(Q3)** What happens when the complexity of out-of-distribution data increases? **(Q4)** How effective is the adaptive relation tuning framework? **(Q5)** How robust and beneficial is ART in real-world settings involving downstream tasks and noisy object detections?

4.1. Experimental setup

Datasets. We train on Visual Genome (VG) [14] and test on GQA [12] and Open Images (OI) v4, v6 [16] to evaluate model generalization across datasets with increasing complexity and out-of-distribution scenarios. Since VG is a subset of GQA with overlapping categories, it is easier than OI-v4 and v6, which are entirely distinct from VG and GQA. See Appendix B for data splits and further details.

Evaluation protocol and metrics. As our aim is to improve predicate prediction in a generalizable manner, we evaluate model performance on the predicate classification task using ground-truth/predicted bounding boxes and objects, with a focus on generalization across data distributions, object, and predicate categories. Following previous works [34, 44], we report Recall@k (R@k) and mean Recall (mR@k). *Mean recall (mR@k) is particularly important* as it reflects a model's ability to perform across the full predicate distribution. To further assess generalization, we propose generalized Recall (gR@k) and mean generalized Recall (mgR@k), which treat false positives with high semantic similarity (measured by **S** from Sec. 3.2) as true positives. See Appendix B for details.

Implementation details. We leverage BLIP-2 [20] models fine-tuned for the captioning task with strong generalization capabilities, specifically two InstructBLIP [5] variants. Both share the ViT/14 [6] image encoder but differ in language models: One integrates a Vicuna-7B [49] LLM, instruction-tuned from LLaMA [37], while the other uses FlanT5 [4], based on Transformer T5 [30]. We fine-tune both variants with our adaptive instruction tuning. The models reached the best results with about 12% (see Appendix B) of the training data samples, with a sampling budget of 2% per adaptive learning loop.

4.2. Experimental results

Using the above experimental setup, we are able to address the questions **(Q1–5)**. We establish fair comparisons by using several baselines, including naive relation-tuned VLMs (Naive-RT) with random and balanced-random sampling setups. Naive-RT (random) selects training samples through basic random sampling, while Naive-RT (balanced random) ensures even sampling across predicate categories, with random selection within each category. Additionally, we evaluate mainstream VRD models (Motifs [44], VTransE [45]) and recent models with balanced mR (PENET [48], VETO [34]). Higher mR reflects diverse relational learning, avoiding overfitting to dominant predicates. To assess generalization, all models are trained on VG and evaluated on GQA and OI.

(Q1) In-distribution samples with unseen predicates. We analyze the methods in a simpler generalization setting using GQA, which shares images with VG (in-distribution) but has more object classes and predicates. Like VG, GQA is dominated by a few head predicates. Consequently, models that generalize well should achieve higher mR and gmR, indicating diverse relational learning rather than overfitting to head categories, which results in high R but low mR. As seen in Tab. 1, Vicuna-7B+ARToutperforms the strongest baseline, Vicuna-7B+Naive-RT (balanced random), with more than a 40% and 20% relative improvement in mR and gmR, respectively. Similar gains are observed for FlanT5. In contrast, mainstream models and Naive-RT (random) achieve high R but lower mR, revealing their bias toward frequent predicates. Additional evaluations (Fig. 4) confirm ART's superior relational diversity, with more unique and unseen predicates than baselines.

(Q2) Out-of-distribution samples with unseen predicates. In this setting, we assess a more challenging generalization scenario using the OI-v4 dataset. Since OI contains distinct image samples from VG, it is considered out-of-distribution data and includes unseen object and predicate categories. As shown in Tab. 1, ART consistently outperforms its baselines across all metrics for both the Vicuna and FlanT5 model variants. Unique and unseen predictions on OI-v4 (Fig. 4) further confirm ART's ability to predict diverse relations in out-of-distribution settings.

(Q3) Out-of-distribution samples with increased complexity. In this scenario, we further evaluate the out-of-distribution generalization of ART on OI-v6. As shown in Tab. 1, mainstream models drop significantly across all

Table 1. **Evaluation of VG-trained models on GQA, Open Images v4 (OI-v4), and Open Images v6 (OI-v6),** showing Recall (R), mean Recall (mR), and their generalized forms (gR, mgR, see text) — all in %, higher values indicate better performance. The superscripts '†' and '*' denote methods that use TDE and re-weighting strategies, respectively. Double citations refer to the original model and the used TDE-debiased version. The superscript '‡' denotes models using VLMs with the same image encoder (ViT-g/14) but different LLMs: FlanT5 [4] and Vicuna [49]. The best results for each VLM variant are highlighted in **bold**.

Model	Dataset	R@k: 20/50	mR@k: 20/50	gR@k: 20/50	gmR@k: 20/50
Motifs† [36, 44]	GQA	41.7 / 41.7	12.7 / 12.8	55.7 / 55.8	23.5 / 23.6
VTransE† [36, 45]		37.9 / 38.0	9.8 / 9.9	51.3 / 51.3	18.7 / 18.8
PENET* [48]		45.8 / 45.8	10.2 / 10.3	62.7 / 62.7	21.7 / 21.8
VETO* [34]		47.7 / 47.8	10.9 / 10.9	63.4 / 63.5	23.2 / 23.3
FlanT5-XL‡ + Naive-RT (random)		**51.4 / 53.4**	8.9 / 9.5	**65.8 / 69.5**	20.0 / 22.6
FlanT5-XL‡ + Naive-RT (balanced random)		36.1 / 37.7	13.6 / 14.7	53.7 / 56.8	25.6 / 29.2
FlanT5-XL‡ + **ART (ours)**		30.3 / 31.5	**15.5 / 16.1**	61.1 / 62.0	**30.8 / 32.2**
Vicuna-7B‡ + Naive-RT (random)		**59.0 / 61.3**	8.3 / 9.1	**70.0 / 73.9**	18.8 / 21.7
Vicuna-7B‡ + Naive-RT (balanced random)		31.9 / 33.0	13.2 / 14.1	55.1 / 58.3	26.0 / 29.7
Vicuna-7B‡ + **ART (ours)**		40.1 / 40.4	**18.9 / 19.4**	61.5 / 62.2	**33.2 / 34.7**
Motifs† [36, 44]	OI-v4	30.0 / 30.1	13.4 / 13.4	56.9 / 57.0	31.0 / 31.1
VTransE† [36, 45]		29.7 / 29.8	13.3 / 13.4	55.7 / 56.0	29.3 / 29.6
PENET* [48]		2.1 / 2.2	0.6 / 0.7	18.1 / 18.2	0.6 / 0.7
VETO* [34]		12.1 / 12.1	6.7 / 6.7	46.3 / 46.4	28.3 / 28.4
FlanT5-XL‡ + Naive-RT (random)		30.6 / 31.8	14.5 / 14.9	70.0 / 72.5	40.5 / 40.9
FlanT5-XL‡ + Naive-RT (balanced random)		40.1 / 42.2	15.0 / 15.6	69.7 / 72.2	40.0 / 40.9
FlanT5-XL‡ + **ART (ours)**		**54.0 / 54.7**	**17.1 / 17.3**	**81.0 / 82.2**	**44.9 / 45.4**
Vicuna-7B‡ + Naive-RT (random)		43.5 / 44.6	12.7 / 13.1	73.9 / 76.0	37.8 / 38.6
Vicuna-7B‡ + Naive-RT (balanced random)		42.5 / 44.7	11.8 / 12.4	72.4 / 74.9	35.5 / 36.5
Vicuna-7B‡ + **ART (ours)**		**46.6 / 47.5**	**26.0 / 26.2**	**79.0 / 80.4**	**54.3 / 54.5**
Motifs† [36, 44]	OI-v6	6.3 / 6.3	4.0 / 4.1	54.8 / 54.9	13.9 / 13.9
VTransE† [36, 45]		4.5 / 4.6	2.9 / 3.0	27.4 / 27.6	9.2 / 9.2
PENET* [48]		1.3 / 1.4	0.1 / 0.1	19.9 / 19.9	5.9 / 6.0
VETO* [34]		3.4 / 3.4	2.5 / 2.5	50.0 / 50.0	12.2 / 12.2
FlanT5-XL‡ + Naive-RT (random)		7.2 / 7.4	5.5 / 5.6	54.5 / 55.2	21.7 / 21.9
FlanT5-XL‡ + Naive-RT (balanced random)		19.6 / 19.9	7.0 / 7.1	56.0 / 56.8	22.0 / 22.3
FlanT5-XL‡ + **ART (ours)**		**21.0 / 21.2**	**10.4 / 10.5**	**57.4 / 57.7**	**25.5 / 25.7**
Vicuna-7B‡ + Naive-RT (random)		9.7 / 9.8	3.8 / 3.9	56.3 / 57.5	15.4 / 16.7
Vicuna-7B‡ + Naive-RT (balanced random)		23.4 / 25.0	8.5 / 8.7	55.0 / 55.8	22.1 / 22.5
Vicuna-7B‡ + **ART (ours)**		**27.3 / 27.4**	**9.5 / 9.6**	**63.2 / 63.4**	**25.6 / 25.8**

Table 2. **Instruction set analysis of ART (Vicuna-7B) on VG.** *Rel* indicates relations, and *Neg* denotes negative samples. Removing components lowers Recall (R@k) and mean Recall (mR@k). High-level relations slightly boost both metrics, while counter-negatives notably improve mR@k. Best results are in **bold**. The proposed final instruction set is highlighted in cyan.

High-level	Random	Counter	Per-sample	R@k	mR@k
Rel	*Neg*	*Neg*	*Neg*	(20/50)	(20/50)
✗	✗	✗	–	37.5/38.0	37.5/38.0
✓	✗	✗	–	38.6/39.1	39.4/39.9
✓	✓	✗	1	**41.2/41.8**	40.9/41.3
✓	✓	✗	2	27.9/28.0	41.3/41.7
✓	✗	✓	1	41.1/41.4	**46.4/47.7**
✓	✗	✓	2	37.0/37.3	40.3/41.4

metrics, while ART remains robust, consistently improving mR and gmR over its baselines. Notably, FlanT5+ART achieves a 50% mR improvement over the second-best Naive-RT (balanced random), reinforcing ART's effectiveness in complex out-of-distribution settings.

(Q4) Effectiveness of adaptive relation tuning (ART) framework. We conduct three analyses to evaluate the effectiveness of the ART framework.

Instruction set analysis. We examine the impact of our proposed instruction set from Sec. 3.1. As shown in Tab. 2, incorporating high-level relation types (2nd row) in the prompt improves both R and mR. Adding a single random negative response per positive response (3rd row) further enhances these metrics. However, increasing the number of random negatives leads to a notable drop in R. Replacing random negatives with a single counter-negative response yields the highest mR; adding more counter-negatives reduces both R and mR. Thus, we use one counter-negative per sample in our final instruction set.

Sampling component analysis. We compare our full ART model with all sampling components in Tab. 3, (*i.e.* h_{FN}, t_{FP}, t_{FN}, h_{TP}, last row) against its variants. When comparing a model trained on the full VG train instructions (1st row) to a model trained on a reduced random 12% sampling budget (2nd row), the model trained on a random subset outperforms the fully trained one. This indicates that

Table 3. **Ablation study of ART on VG.** Adaptive relation tuned model variants. *entropy*: only entropy-based uncertainty is used, h_{TP}: entropy-based head samples from TP, h_{FN}: entropy-based head samples from FN, t_{FN}: entropy-based tail samples from FN, t_{FP}: similarity score-based tail samples from FP. TP: True Positive, FN: False Negative, FP: False Positive. Except the 1st row that uses 100% train data, all other models use 12% of train data. The proposed final ART model is highlighted in cyan.

Sampling strategy	R@k: 20/50	mR@k: 20/50
full train data	59.9 / 60.2	14.1 / 14.7
random	**63.2 / 63.4**	19.6 / 19.8
balanced random	36.4 / 36.6	42.4 / 43.5
ART: *entropy*	34.7 / 34.9	44.6 / 44.8
ART: h_{FN}, t_{FP}	37.5 / 37.6	45.5 / 45.6
ART: h_{FN}, t_{FP}, t_{FN}	38.3 / 38.5	45.2 / 45.4
ART: h_{FN}, t_{FP}, t_{FN}, h_{TP}	41.1 / 41.4	**46.4 / 47.7**

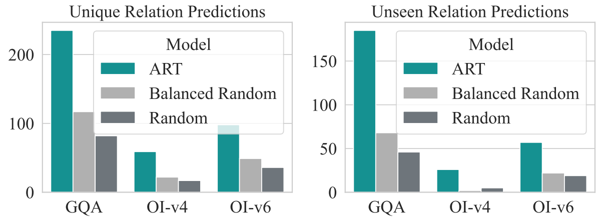

Figure 4. **Comparison of unique relation predictions** *(left)* **and unseen relation predictions** *(right)* for the ART (Vicuna-7B) model across different datasets.

tuning a VLM on more data can lead to suboptimal results if the data distribution is biased. Comparing random to balanced random sampling reveals that the balanced random model better learns various relation concepts, resulting in improved mR. Introducing our ART strategy results in further improved mR. The ART: *entropy* model that picks the highest entropy samples from the entire sampling pool is less efficient than the variants that diligently pick based on the uncertainty or dissimilarity (**H** or **S**) and the head or tail region to sample from the prediction types (TP, FP, or FN).

Additionally, a detailed analysis of adaptive *vs.* fixed thresholding is provided in Appendix G.

(Q5) Robustness across tasks and inputs. We evaluate ART's effectiveness in a downstream application, specifically a *deictic*[1] segmentation task [32] using ART. Given visual input and a complex textual prompt, the task is to produce a segmentation mask for the object specified by the prompt. We evaluate ART on a deictic segmentation task [32] using 10k image–prompt pairs with expressive relations absent from training data (*e.g.*, *floating on*, *attached to*). We compare: (1) LISA [17], a state-of-the-art VLM-based reasoning model; (2) DeiSAM, a neuro-symbolic segmentation model using ground-truth scene graphs (DeiSAM + GT); and using ART-generated graphs (DeiSAM + ART). Results with and without ground-truth boxes are reported. As shown in Tab. 4, ART outperforms all baselines, demonstrating strong generalization to novel relations, especially valuable in dynamic environments.

Table 4. **ART enhances segmentation qualities.** Mean Average Precision (mAP) comparison of DeiSAM+ART, DeiSAM+GT, and LISA on VG and GQA datasets.

mAP (in %, ↑)	VG	GQA
LISA [17]	7.87	18.92
DeiSAM + GT (w/o bbox)	4.77	11.84
DeiSAM + ART (w/o bbox)	**25.07**	**26.62**
DeiSAM + GT (with bbox)	35.04	43.84
DeiSAM + ART (with bbox)	**99.98**	**97.96**

To further assess robustness, we evaluate ART using noisy object detections from real-world detectors. Specifically, we test with boxes from Detectron2 [40] and the zero-shot RAM model [46]. As shown in Tab. 5, ART maintains strong performance across both detectors. Remarkably, ART significantly outperforms its baselines even when using RAM's open-world detections, confirming its adaptability and resilience to detection noise.

Table 5. **Performance on detected bounding boxes** from Detectron2 [40] and RAM [46]. R: Random sampling, B: Balanced Random sampling, and ART (ours).

Model	Dataset	R@20	mR@20	gR@20	gmR@20
R	OI-v4	42.6	13.0	64.0	24.5
BR		35.3	11.1	51.1	22.6
ART		**53.7**	**16.3**	**69.5**	**26.8**
R	OI-v4	19.1	6.7	55.0	22.3
BR		29.2	7.3	59.4	26.5
ART		**40.6**	**24.0**	**69.1**	**42.6**

5. Conclusion

In this work, we presented ART, a novel framework for the relation classification task of VRD that leverages vision-language models through instruction tuning by converting benchmark VRD datasets into an instruction-tuning format. We ensured that our model focuses on the most relevant instances by incorporating adaptive sampling and fine-tuning techniques, thus enhancing its robustness and generalization capabilities. ART consistently outperformed baseline models, improving the handling of both diverse and unseen object classes and relations. This underscores the importance of innovative VRD approaches, especially for real-world applications with diverse data distributions. Future research can enhance ART by refining instruction tuning strategies and extending it to other vision-language tasks, potentially unlocking new applications in visual understanding.

[1] A deictic representation refers to an object depending on the overall context, *e.g.*, "An object on the table and next to a cup."

Acknowledgments. This work was funded by the Hessian Ministry of Science and the Arts (HMWK) through the projects "The Third Wave of Artificial Intelligence – 3AI" and hessian.AI. The work was further supported by the Deutsche Forschungsgemeinschaft (German Research Foundation, DFG) under Germany's Excellence Strategy (EXC 3057/1 "Reasonable Artificial Intelligence", Project No. 533677015). SR acknowledges support by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 866008).

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikołaj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: A visual language model for few-shot learning. In *NeurIPS*, 2022. 1

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *ICCV*, pages 2425–2433, 2015. 1

[3] Xiaojun Chang, Pengzhen Ren, Pengfei Xu, Zhihui Li, Xiaojiang Chen, and Alex Hauptmann. A comprehensive survey of scene graphs: Generation and application. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(1):1–26, 2021. 1

[4] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *J. Mach. Learn. Res.*, 25, 2024. 6, 7, iii

[5] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N. Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. *NeurIPS*, 2024. 1, 2, 6

[6] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. EVA: Exploring the limits of masked visual representation learning at scale. In *CVPR*, 2023. 6

[7] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep Ensembles: A loss landscape perspective. *arXiv:1912.02757 [stat.ML]*, 2019. 2

[8] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, pages 1050–1059, 2016. 2

[9] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ICAN: Instance-centric attention network for human-object interaction detection. In *BMVC*, 2018. 2

[10] Tao He, Lianli Gao, Jingkuan Song, and Yuan-Fang Li. Towards open-vocabulary scene graph generation with prompt-based finetuning. In *ECCV*, 2022. 1, 2

[11] Md. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys*, 51(6): 1–36, 2019. 1

[12] Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, pages 6700–6709, 2019. 1, 6, i

[13] A. S. M. Iftekhar, Satish Kumar, R. Austin McEver, Suya You, and B. S. Manjunath. GTNet: Guided transformer network for detecting human-object interactions. In *Pattern Recognition and Tracking XXXIV*, pages 192–205. SPIE, 2023. 2

[14] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, et al. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73, 2017. 1, 3, 6, i

[15] Po-Nien Kung, Fan Yin, Di Wu, Kai-Wei Chang, and Nanyun Peng. Active instruction tuning: Improving cross-task generalization by training on prompt sensitive tasks. *arXiv:2311.00288 [cs.CL]*, 2023. 2

[16] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *Int. J. Comput. Vis.*, 128:1956–1981, 2020. 6, i

[17] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: Reasoning segmentation via large language model. In *CVPR*, 2024. 8

[18] Baoli Li and Liping Han. Distance weighted cosine similarity measure for text classification. In *Intelligent Data Engineering and Automated Learning*, 2013. 2

[19] Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven C. H. Hoi. LAVIS: A library for language-vision intelligence. *arXiv:2209.09019 [cs.CV]*, 2022. i

[20] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 6

[21] Lin Li, Long Chen, Yifeng Huang, Zhimeng Zhang, Songyang Zhang, and Jun Xiao. The devil is in the labels: Noisy label correction for robust scene graph generation. In *CVPR*, pages 18869–18878, 2022. 2

[22] Lin Li, Jun Xiao, Guikun Chen, Jian Shao, Yueting Zhuang, and Long Chen. Zero-shot visual relation detection via composite visual cues from large language models. *NeurIPS*, 2024. 1, 2

[23] Wentong Liao, Bodo Rosenhahn, Ling Shuai, and Michael Ying Yang. Natural language guided visual relationship detection. In *CVPRW*, 2019. 2

[24] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1

[25] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 2023. 2

[26] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *ECCV*, pages 852–869, 2016. 2

[27] Vu-Linh Nguyen, Mohammad Hossein Shaker, and Eyke Hüllermeier. How to measure uncertainty in uncertainty sampling for active learning. *Machine Learning*, 111(1):89–122, 2022. 2, 4

[28] Fredrik Olsson. A literature survey of active machine learning in the context of natural language processing. *SICS Technical Report*, 2009. 2

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1

[30] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21, 2020. 6

[31] Shafin Rahman, Salman Khan, and Fatih Porikli. Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts. In *ACCV*, pages 547–563, 2018. 2

[32] Hikaru Shindo, Manuel Brack, Gopika Sudhakaran, Devendra Singh Dhami, Patrick Schramowski, and Kristian Kersting. DeiSAM: Segment anything with deictic prompting. In *NeurIPS*, 2024. 1, 8, iv

[33] Aditya Siddhant and Zachary C. Lipton. Deep Bayesian active learning for natural language processing: Results of a large-scale empirical study. *arXiv:1808.05697 [cs.CL]*, 2018. 2

[34] Gopika Sudhakaran, Devendra Singh Dhami, Kristian Kersting, and Stefan Roth. Vision relation transformer for unbiased scene graph generation. In *ICCV*, 2023. 2, 6, 7, i

[35] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *CVPR*, pages 6619–6628, 2019. 1

[36] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *CVPR*, pages 3716–3725, 2020. 2, 7, i

[37] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and efficient foundation language models. *arXiv:2302.13971 [cs.CL]*, 2023. 6

[38] Oytun Ulutan, A. S. M. Iftekhar, and Bangalore S. Manjunath. VSGNet: Spatial attention network for detecting human object interactions using graph convolutions. In *CVPR*, pages 13617–13626, 2020. 2

[39] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. *arXiv:2109.01652 [cs.CL]*, 2021. 2

[40] Yuxin Wu et al. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019. 8

[41] Bingjie Xu, Yongkang Wong, Junnan Li, Qi Zhao, and Mohan S. Kankanhalli. Learning to detect human-object interactions with knowledge. In *CVPR*, 2019. 2

[42] Caixia Yan, Qinghua Zheng, Xiaojun Chang, Minnan Luo, Chung-Hsing Yeh, and Alexander G. Hauptman. Semantics-preserving graph propagation for zero-shot object detection. *IEEE Trans. Image Process.*, 29:8163–8176, 2020. 2

[43] Caixia Yan, Xiaojun Chang, Minnan Luo, Huan Liu, Xiaoqin Zhang, and Qinghua Zheng. Semantics-guided contrastive network for zero-shot object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(3):1530–1544, 2024. 2

[44] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural Motifs: Scene graph parsing with global context. In *CVPR*, pages 5831–5840, 2018. 1, 2, 3, 6, 7, i

[45] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. Visual translation embedding network for visual relation detection. In *CVPR*, pages 5532–5540, 2017. 6, 7, i

[46] Youcai et al. Zhang. Recognize anything: A strong image tagging model. In *CVPR*, pages 1724–1732, 2024. 8

[47] Zhisong Zhang, Emma Strubell, and Eduard Hovy. A survey of active learning for natural language processing. *arXiv:2210.10109 [cs.CL]*, 2022. 2

[48] Chaofan Zheng, Xinyu Lyu, Lianli Gao, Bo Dai, and Jingkuan Song. Prototype-based embedding network for scene graph generation. In *CVPR*, 2023. 2, 6, 7, i

[49] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *arXiv:2306.05685 [cs.CL]*, 2023. 6, 7, iii

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.

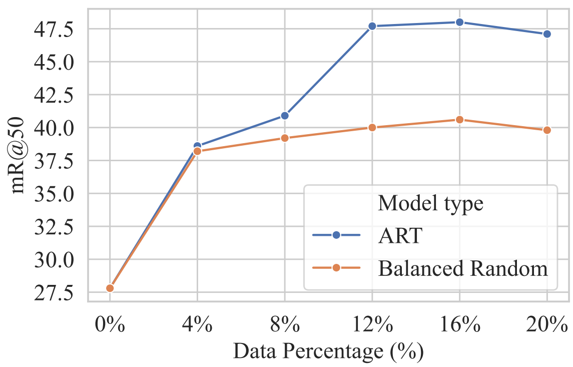

Figure 5. **Training data subsampling analysis.** We plot mR@50 for ART and Naive-RT (balanced random) as a function of the training data percentage used (y-axis) on Vicuna model variants.

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.

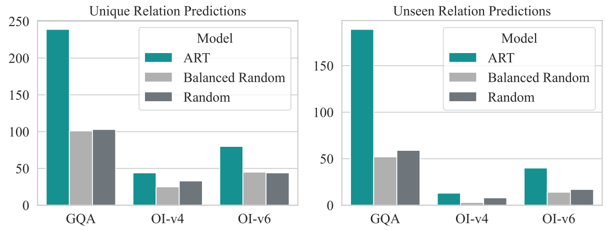

Figure 6. **Comparison of unique relation predictions** ***(left)*** **and unseen relation predictions** *(right)* for the ART (FlanT5) model across different datasets.

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.

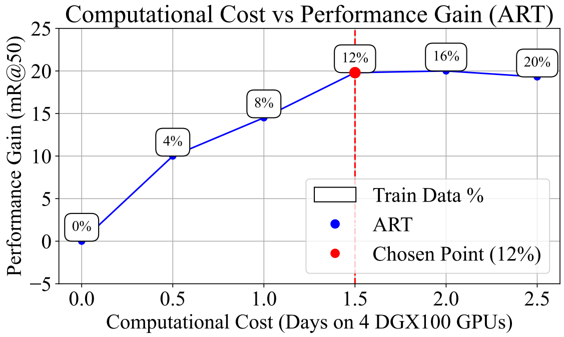

Figure 7. **Trade-off between computational cost and predictive performance** as a function of training data usage.

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.

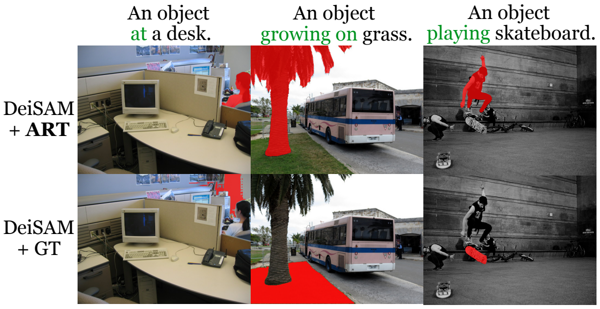

Figure 8. **ART can be used to label missing annotations and predict new unseen predicates.** Segmentation results with textual prompts *(top)* using DeiSAM [32], which segments objects via reasoning on scene graphs. ART successfully detects new relations and improves the segmentation quality, while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

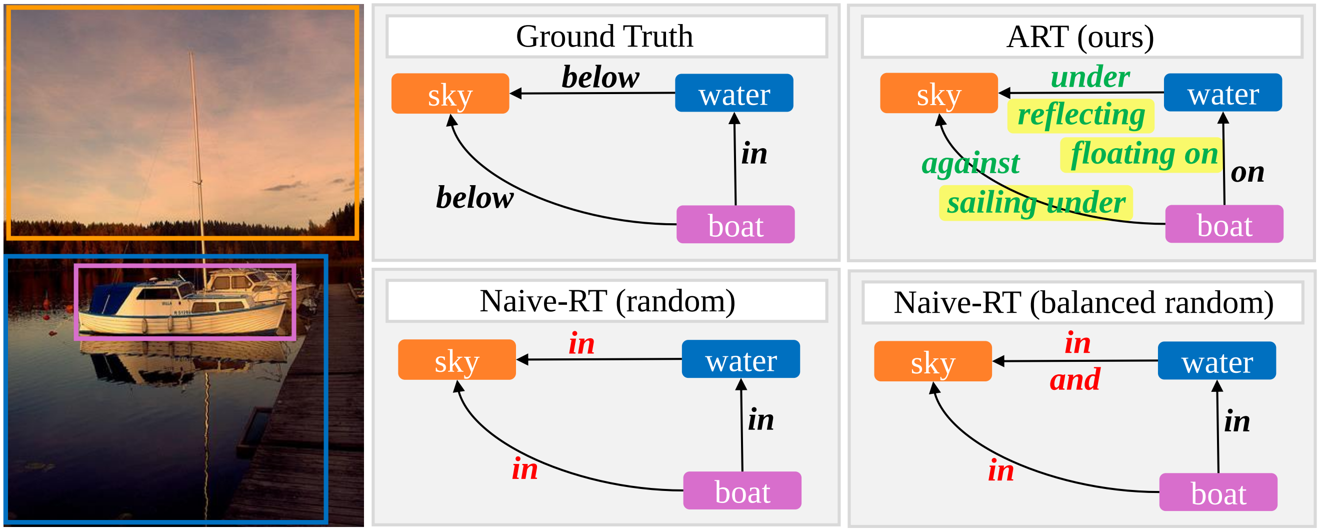

Figure 9. **Comparison of ART and its baselines on the GQA dataset.** ART predicts sensible spatial relations similar to the ground-truth annotation such as, *water under sky*, while also identifying more informative relations than the ground truth, such as *water reflecting sky*, *boat floating on water*, 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.

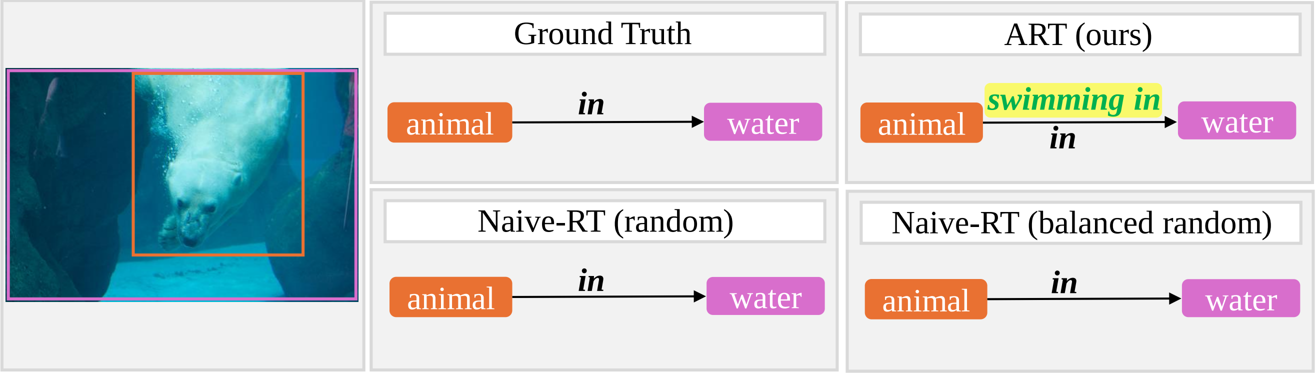

Figure 10. **Comparison of ART and its baselines on the GQA dataset.** The ground truth only provides a spatial relation *in* between animal and water, while ART predicts the descriptive interaction *swimming in*. Informative relation predictions are highlighted in green, while those that are both informative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

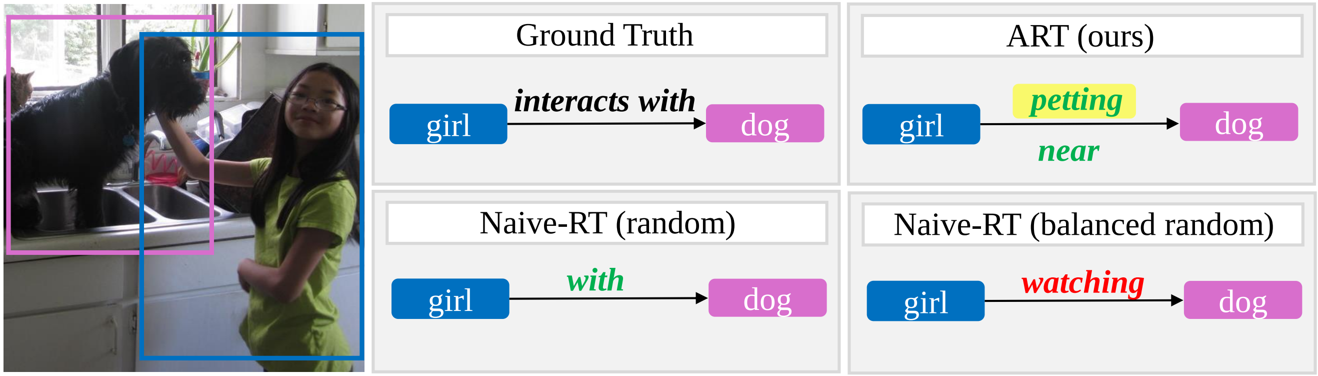

Figure 11. **Comparison of ART and its baselines on the OI-v4 dataset.** In contrast to the provided less informative ground-truth relation *interacts with* in *girl interacts with dog*, which raises the question "What kind of interaction?", ART provides a much clearer interpretation that the interaction 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.

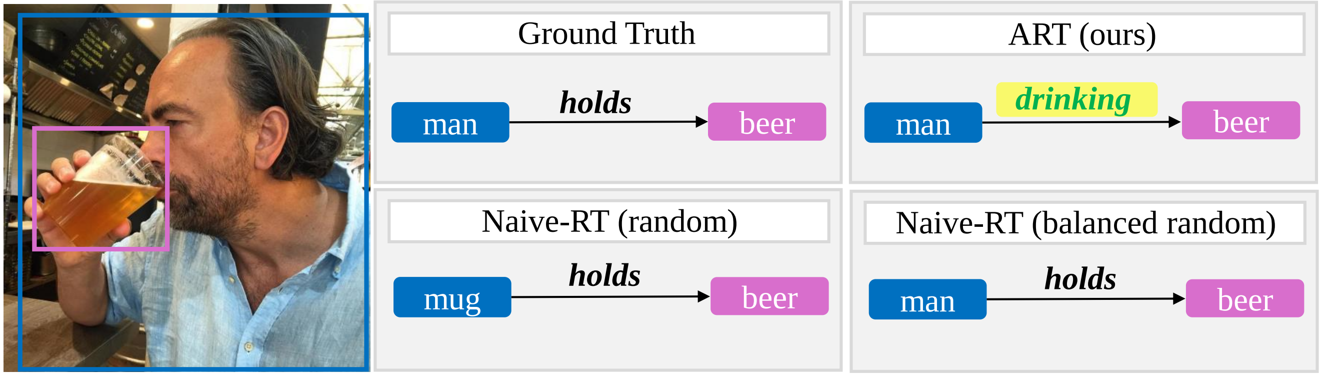

Figure 12. **Comparison of ART and its baselines on the OI-v4 dataset.** The ground truth relation *holds* between *man* and *beer* leaves an open question "What he intends to do with the beer?", while the prediction *drinking* made by ART gives more context and the ongoing action. The Naive-RT baselines 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.

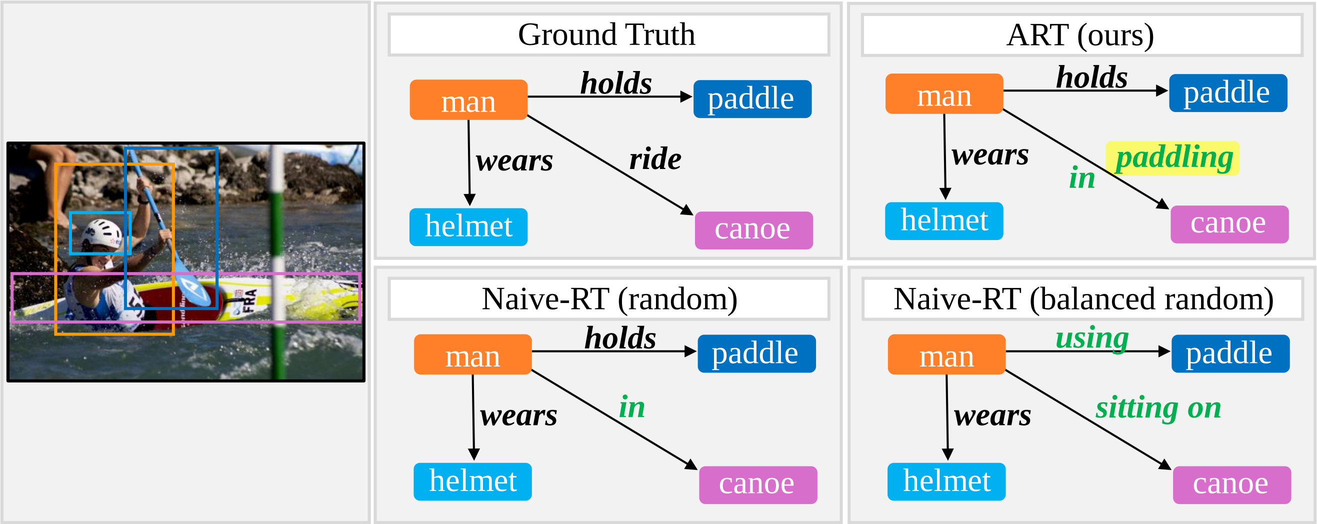

Figure 13. **Comparison of ART and its baselines on the OI-v6 dataset.** ART predicts the more informative relation *paddling* between the man and canoe while also identifying sensible spatial relation *in*. In contrast, although both Naive-RT (random) and (balanced random) make reasonable spatial predictions, 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.

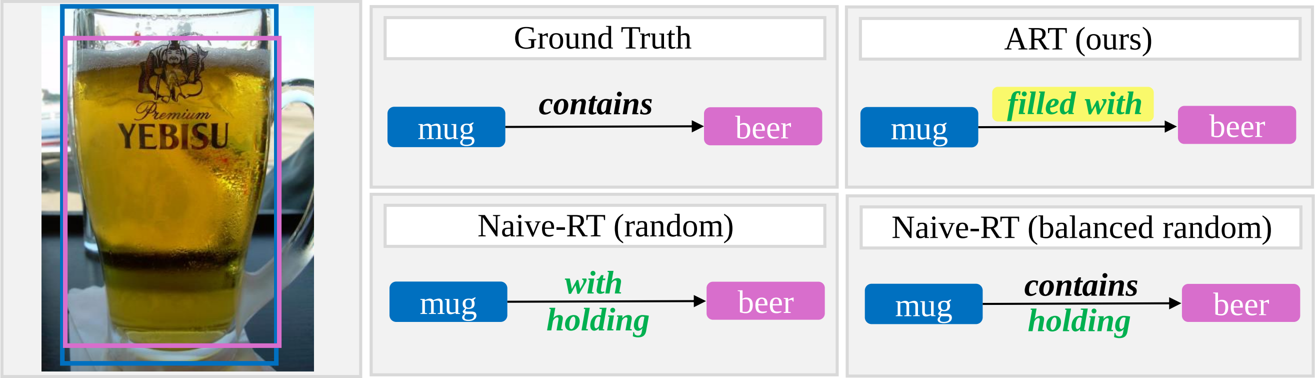

Figure 14. **Comparison of ART and its baselines on the OI-v6 dataset.** While the ground-truth relation *contains* between the mug and beer merely indicates the presence of beer in the mug, the relation *filled with*, predicted by ART, provides more detail by suggesting that the mug is full or nearly 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.