TRAINING-FREE TEST-TIME RETRIEVAL AUGMENTATION FOR VISUAL RELATIONSHIP DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual relationship detectors learn relational knowledge from annotated examples, yet these examples are typically discarded as explicit evidence after training. We revisit their value at inference time and introduce Training-Free Test-Time Retrieval Augmentation, *i.e.*, TF-TRA, a recipe that complements frozen detectors with non-parametric relation memory constructed solely from existing training data. TF-TRA organizes training instances into a relation memory that associates semantic and visual representations with relation annotations while maintaining broad configuration coverage. For each query, it retrieves exemplars through joint semantic and visual matching, with diversity-aware reranking encouraging complementary relation evidence. The retrieved annotations are aggregated into query-specific support and calibrated with the memory prior to refine the frozen detector's predictions. Extensive experiments on three distinct VRD tasks, *i.e.*, human-object interaction (HOI) detection, scene graph generation (SGG), and nonverbal interaction (NVI) detection validate the effectiveness of TF-TRA. It improves overall mAP under both fully supervised and four zero-shot settings on HICO-DET, increases mean Recall and F-score on Visual Genome for SGG, and yields substantial gains in NVI detection. These results show that training examples remain valuable beyond parameter learning, providing explicit relational evidence that strengthens existing detectors at test time. Our source code will be released.

1 INTRODUCTION

Relational reasoning has long been recognized as a fundamental component of human intelligence, playing a central role in how we understand the visual world (Biederman et al., 1982; Kaiser et al., 2019). In recent years, the pursuit of richer scene understanding has brought this relational perspective to the forefront of computer vision research, spanning human-object interaction (HOI) detection (Chao et al., 2018), scene graph generation (SGG) (Xu et al., 2017), and nonverbal interaction (NVI) detection (Wei et al., 2024). Although existing models (Zhao et al., 2023; Zheng et al., 2023; Wei et al., 2024; Kim et al., 2025) have made substantial progress, relation recognition remains constrained by the limited coverage of available datasets relative to the complexity of the visual world. Scaling training data and model capacity (Yuan et al., 2023; Yang et al., 2024; Park et al., 2025) can alleviate this limitation but incurs considerable computational and annotation costs. This line of thinking rests on the premise that the acquired knowledge is encoded solely in network parameters. Such knowledge constitutes a form of *parametric memory*: rich in learned patterns, yet biased toward frequent relations and brittle on underrepresented ones.

Faced with this dilemma, we seek to answer a question: how can we extend a model's relational capacity beyond what it learns during training? Our starting point is simple: *knowledge need not reside entirely in model parameters*. This perspective echoes complementary learning systems theory (Kumaran et al., 2016) in human cognition, which highlights the complementary roles of learning general patterns and retaining specific experiences. Guided by this intuition, we turn to test-time retrieval augmentation for VRD, seeking a training-free recipe to integrate non-parametric exemplar memory into parametric relational reasoning. In general, a retrieval-augmented pipeline entails three sequential stages: memory construction, evidence retrieval, and augmented prediction (Long et al., 2022; Zhang et al., 2022). Yet, adapting this paradigm to VRD poses three non-trivial challenges: **First**, in memory construction, given the vast scale and long-tail distribution in the training data, the primary challenge is to retain broad configuration coverage while preventing head categories from

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

overcrowding the memory. **Second**, during evidence retrieval, since top-ranked neighbors are easily dominated by repetitive categories, the key challenge is to retrieve instances relevant to the target relation while ensuring pattern diversity. **Third**, for prediction augmentation, as retrieved exemplars often vary in relevance and carry pervasive prior bias from the memory bank, the remaining challenge involves aggregating these exemplars into informative context to facilitate prediction.

To tackle these challenges, TF-TRA introduces three targeted solutions. **i)** Group-capped memory construction. TF-TRA encodes each exemplar as a tuple containing semantic and visual representations paired with relation labels. It groups exemplars by relational configuration and caps each group's quota. This preserves broad coverage while curbing head-group dominance. **ii)** Diversity-aware exemplar retrieval. TF-TRA ranks all memory entries using combined relational and visual similarity. It then applies diversity-aware reranking to select complementary exemplars and reduce redundancy. **iii)** Prior-calibrated logit rectification. TF-TRA aggregates retrieved evidence through similarity weighting and calibrates it against the memory's label prior. It integrates positive log-odds residuals into detector predictions. This emphasizes instance-level support over global label prevalence. Together, these designs establish TF-TRA as a training-free, plug-and-play enhancement for pre-trained VRD detectors, constructed solely from the training corpus without parameter updates.

Experiments across HOI detection, SGG, and NVI detection show that existing detectors retain substantial room for improvement through retrieval alone, without further parameter updates. On HICO-DET (Chao et al., 2018), TF-TRA improves overall mAP by an average of 1.32% across all four zero-shot settings, demonstrating the value of reusing annotated relational examples even when evaluation extends beyond the training categories. Under fully supervised evaluation, the method shows substantial improvements on rare categories, gaining 3.26% and 2.57% under the Default and Known Object protocols, respectively, while also lifting overall performance (Full mAP by 1.19% and 1.20%). For SGG on Visual Genome (Xu et al., 2017), TF-TRA promotes much more balanced predicate predictions across head and tail categories, improving mean Recall and F-score across all evaluated prediction budgets, with gains of 7.2% in mR@100 and 6.7% in F@100. On NVI detection, TF-TRA improves all reported metrics on both validation and test sets, increasing AR by 1.83% and 3.91% and mR@25 by 4.87% and 3.94%, respectively. These results establish explicit relation memory as an effective complement to parametric knowledge: training examples remain useful beyond model fitting and can directly support stronger relational predictions at inference time.

2 Related Work

Visual Relationship Detection (VRD). VRD extends object recognition by identifying the semantic relationships among entities in an image (Lu et al., 2016). The exploration of various relationship types has led to several focused research directions in this field, such as human-object interaction (HOI) detection (Chao et al., 2018), scene graph generation (SGG) (Xu et al., 2017), and nonverbal interaction (NVI) detection (Wei et al., 2024). SGG provides a general framework for relational understanding by detecting entity pairs and predicting their ⟨subject, predicate, object⟩ relationships. Focusing on human-centric activities, HOI detection grounds interactions between humans and objects, while NVI detection extends relational understanding to nonverbal behaviors and social interactions among individuals and groups. Despite their varied objectives, these paradigms collectively grapple with severe data scarcity, long-tail distributions, and generalization hurdles. Existing solutions predominantly resort to leveraging large-scale pretrained models (Kim et al., 2024; Hu et al., 2025) and expanding training datasets (Yang et al., 2024; Park et al., 2025). We argue that such paradigms fundamentally address the problem through the lens of parametric memory, which is an effective yet prohibitively expensive route. In contrast, we investigate a complementary approach based on non-parametric memory, retrieving relevant evidence from existing training examples to support visual relationship detection without additional parameter updates.

Retrieval-Augmented Prediction. Since not all knowledge can be encoded in model parameters, retrieval-augmented prediction offers a complementary mechanism for leveraging external knowledge. By retrieving relevant examples or representations, these methods provide additional evidence to inform model predictions. Initially pioneered in natural language processing (Sukhbaatar et al., 2015), such memory-augmented approaches have since been widely embraced across computer vision tasks, such as classification (Karunaratne et al., 2021; Zhang et al., 2024b), segmentation (Xie et al., 2021; Ding et al., 2025), and detection (Chen et al., 2020; Cai et al., 2022). Existing approaches

can be broadly categorized into training-based and training-free methods. Training-based methods incorporate retrieved evidence through additional optimization and/or learnable components, whereas training-free methods use retrieved information to support inference without further parameter updates. In this work, we focus on the latter setting, which enables the reuse of existing knowledge without the computational overhead of additional training.

To realize such training-free adaptation, a prominent mechanism is to retain training examples or their representations in non-parametric memory, making these details accessible through retrieval. For instance, Tip-Adapter (Zhang et al., 2022) constructs a key-value cache from downstream few-shot examples to augment CLIP's classification scores, while APE (Zhu et al., 2023) refines cached representations by selecting discriminative feature channels for training-free prediction. This principle naturally connects to visual relationship detection (VRD) by reusing relational experiences to alleviate the data scarcity inherent in the combinatorial triplet space (Lu et al., 2016). Inspired by this insight, we propose TF-TRA, which equips the relation detector with an explicit non-parametric memory of training annotation to ground and regularize predicate prediction.

3 METHOD

3.1 PROBLEM STATEMENT AND OVERVIEW

Problem Statement. VRD (Lu et al., 2016) aims to identify entities and characterize their relationships in an image. Given an image x, each relationship is represented as a triplet (s, c, o), where s and o denote the subject and object, respectively, each associated with a spatial localization and a semantic category, and c denotes the relationship category. This formulation encompasses HOI detection, SGG, and NVI detection, each characterized by distinct entity configurations and relation categories. For each candidate entity pair i, a frozen base model produces relation logits $\boldsymbol{z}_i^{\text{base}} \in \mathbb{R}^C$, where C is the number of relation categories. Our objective is to refine these predictions by retrieving query-specific evidence from a relation memory $\mathcal{M}$ constructed from annotated training instances. Notably, our method is training-free and involves no parameter updates. The memory is constructed before evaluation without using validation or test annotations and remains fixed throughout inference.

Overview. As illustrated in Fig. 1, TF-TRA comprises three stages: **i)** *group-capped memory construction* organizes annotated training instances into memory entries by identical entity metadata and complete relation label sets, retaining a fixed maximum number of exemplars per group (§3.2); **ii)** *diversity-aware exemplar retrieval* ranks all memory entries using semantic and visual similarities, then reranks them to balance query relevance and relation-label diversity (§3.3); **iii)** *prior-calibrated logit rectification* aggregates the retrieved evidence, calibrates it against memory class priors, and integrates it with the frozen base model's predictions (§3.4). Together, these stages enable training-free prediction refinement by incorporating relation information retrieved from training instances.

3.2 GROUP-CAPPED MEMORY CONSTRUCTION

Relation-Instance Encoding. An image may contain multiple relationships, and different relationships in the same image may be supported by different visual regions. Accordingly, the source bank is organized at the level of relation instances rather than entire images. For each relation instance j, we retain its entity metadata η_j, relation annotation $\boldsymbol{y}_j$, relation representation $\boldsymbol{s}_j$, and visual region representation $\boldsymbol{v}_j$. The metadata η_j records the ordered subject and object categories, while $\boldsymbol{y}_j \in \{0,1\}^C$ encodes the positive relation labels over C categories. The source bank is defined as:

$$\mathcal{M}_{\text{src}} = \{(\boldsymbol{s}_j, \boldsymbol{v}_j, \eta_j, \boldsymbol{y}_j)\}_{j=1}^{M_{\text{src}}}. \tag{1}$$

The representations $\boldsymbol{s}_j$ and $\boldsymbol{v}_j$ are extracted using the frozen base model and CLIP visual encoder.

Group-Capped Selection. Retrieval over the entire source bank is computationally expensive, and its imbalanced relation distribution may bias retrieval toward frequent categories. We thus construct a compact relation memory $\mathcal{M}$ by imposing a budget of at most B instances per group, where each group contains entries with identical entity metadata and the same complete relation-label set, *i.e.*, $g = (\eta_j, \boldsymbol{y}_j)$. The number of instances in group g is given by:

$$n_g = \sum\nolimits_{j=1}^{M_{\text{src}}} \mathbb{I}[\eta_j = \eta_g \wedge \boldsymbol{y}_j = \boldsymbol{y}_g], \tag{2}$$

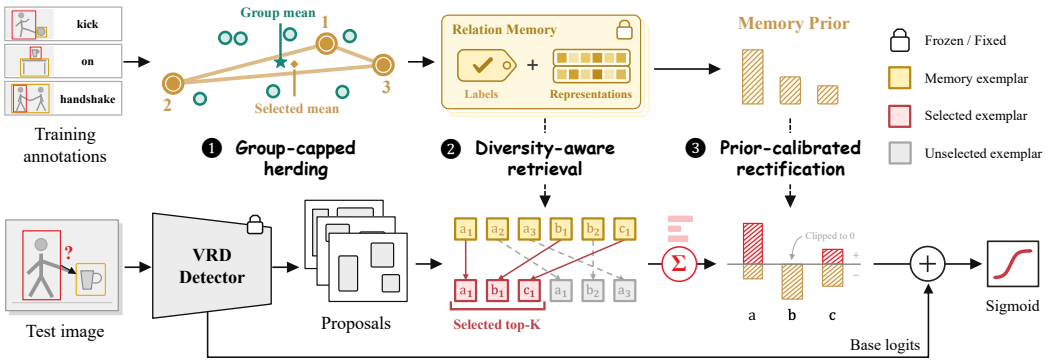


Figure 1: **Overview of TF-TRA.** Before inference, ❶ TF-TRA selects representative relation instances from training annotations via group-capped herding to construct a compact relation memory. During inference, ❷ candidate proposals are first extracted via a VRD detector and used to query the relation memory through diversity-aware retrieval, which accounts for both relevance and diversity to retrieve informative exemplars. ❸ The retrieved exemplars' labels are then aggregated using relevance-based weights to form relation support, from which a residual is computed by calibrating against the memory prior and retaining only positive values. Finally, the calibrated residual is added to the detector's base logits, followed by a sigmoid function to obtain the relation probabilities.

with its retention quota defined as $t_g = \min(n_g, B)$. This retains all exemplars from groups with at most B instances and caps larger groups, preserving coverage of observed relational configurations while limiting the dominance of frequent groups. For each group g, we select t_g instances using a without-replacement variant of the herding procedure adopted in Rebuffi et al. (2017). This procedure greedily selects instances whose mean relation representation approximates the empirical mean of the full group. Let $\bar{s}_j$ denote the ℓ_2-normalized relation representation and $\boldsymbol{\mu}_g$ its empirical mean over group g. Starting from an empty selected set $\mathcal{S}_g$, at step $k = 1, \ldots, t_g$, we select:

$$j^* = \arg\min_{j \in \mathcal{I}_g \setminus \mathcal{S}_g} \|\boldsymbol{\mu}(\mathcal{S}_g \cup \{j\}) - \boldsymbol{\mu}(\mathcal{I}_g)\|_2^2, \tag{3}$$

where $\mathcal{I}_g$ indexes group g and $\boldsymbol{\mu}(\cdot)$ averages its indexed ℓ_2-normalized relation representations. Each selected instance is added to $\mathcal{S}_g$ until $|\mathcal{S}_g| = t_g$. The relation memory $\mathcal{M}$ is constructed as the union of these group-wise subsets and remains fixed during test-time inference.

3.3 Diversity-Aware Exemplar Retrieval

For each candidate entity pair i, we obtain the bounding boxes and category labels of both entities from the detector outputs. Then, we extract its relation representation $\boldsymbol{s}_i$ and visual region representation $\boldsymbol{v}_i$ using the same frozen encoders as in memory construction. To retrieve relevant exemplars from the relation memory $\mathcal{M}$, we first rank all memory entries by their fused similarity to the query and then apply diversity-aware reranking to select the final neighborhood.

Relation Memory Scoring. First, we rank memory entries based on two signals: semantic similarity (S) and visual similarity (V). For a query proposal i and a memory entry j, we compute scores as:

$$s_{ij}^S = \cos(\boldsymbol{s}_i, \boldsymbol{s}_j), \qquad s_{ij}^V = \cos(\boldsymbol{v}_i, \boldsymbol{v}_j), \tag{4}$$

where $\boldsymbol{s}$ is the relation representation extracted by the frozen base model, and $\boldsymbol{v}$ is the visual region representation extracted by a frozen CLIP visual encoder. To resolve scale discrepancies, the visual scores s_{ij}^V are normalized to match the mean and standard deviation of the semantic scores s_{ij}^S across the relation memory, yielding the aligned scores $\widetilde{s}_{ij}^V$. The final similarity score is then computed as the average of the semantic and aligned visual scores, *i.e.*, $s_{ij}^{SV} = \frac{1}{2}(s_{ij}^S + \widetilde{s}_{ij}^V)$.

Diversity-Aware Reranking. Highly ranked exemplars may concentrate on a few relation categories, limiting the coverage of other plausible relations. We therefore apply a soft label-diversity penalty that balances retrieval relevance with relation coverage. For each candidate entry j, we define its redundancy o_{ij} by counting how often its own labels have already appeared in preceding exemplars:

$$o_{ij} = 1 + \frac{\sum_{\ell \prec_{ij}} \boldsymbol{y}_j^\top \boldsymbol{y}_\ell}{\|\boldsymbol{y}_j\|_1}, \tag{5}$$

where $\boldsymbol{y}_j$ is the multi-hot label vector of entry j, and $\ell \prec_i j$ means that candidate ℓ ranks ahead of j in the relation memory $\mathcal{M}$ according to the initial descending order of the fused similarity scores. The fraction is the average number of times each label of j appears among these preceding candidates. Thus, o_{ij} increases as these labels recur across higher-ranked candidates, allowing us to downweight entries with greater label redundancy.

We use this redundancy factor to apply a soft label-diversity penalty and select the final neighborhood:

$$\mathcal{N}_i = \mathrm{TopK}_{j\in\mathcal{M}}[s_{ij}^{SV} - \lambda_d\, \sigma_i^{SV} \log(o_{ij})], \tag{6}$$

where σ_i^{SV} is the standard deviation of s_{ij}^{SV} for query i over $\mathcal{M}$, and $\lambda_d = 0.5$ balances retrieval relevance against label diversity. The logarithmic penalty $\log(o_{ij})$ grows sublinearly with label repetition and vanishes when none of an entry's labels occurs among preceding candidates.

By coupling relation memory scoring with diversity-aware reranking, our method builds a neighborhood that harmonizes retrieval relevance and relation diversity for subsequent prediction refinement.

3.4 Prior-Calibrated Logit Rectification

Given the retrieved neighborhood $\mathcal{N}_i$, we aggregate exemplar labels into a query-specific relation support and use it to refine the predictions of the frozen base model, without updating any parameters.

Neighborhood Aggregation. To estimate relation support from the retrieved neighborhood, we aggregate exemplar annotations while accounting for their relevance to the query. Specifically, each retrieved exemplar is assigned a normalized weight based on its similarity score:

$$w_{ij} = \frac{\exp(s_{ij}^{SV}/T)}{\sum_{k\in\mathcal{N}_i}\exp(s_{ik}^{SV}/T)}, \qquad j \in \mathcal{N}_i, \tag{7}$$

where T denotes a temperature that controls the concentration of the weights. The retrieval-based relation support for category c is then computed as the weighted sum of exemplar annotations:

$$p_{i,c}^{\mathrm{ret}} = \sum_{j\in\mathcal{N}_i} w_{ij} y_{j,c}. \tag{8}$$

This formulation accommodates both one-hot and multi-hot annotations. For multi-label relations, each category receives its own support estimate; no normalization across categories is imposed.

Prior Calibration. Although group-capped construction curbs configuration-level dominance and diversity-aware retrieval mitigates label repetition, neither guarantees a balanced memory distribution, as intrinsic factors like relation co-occurrence cannot be fully resolved by selection strategies alone. Consequently, the relation support reflects both query-specific relevance and category prevalence in the relation memory. To account for the latter, we compute a memory prior directly from the annotations of the stored exemplars:

$$\pi_c^{\mathrm{mem}} = \frac{1}{|\mathcal{M}|}\sum_{j\in\mathcal{M}} y_{j,c}. \tag{9}$$

This prior measures the fraction of memory entries supporting class c. It is computed separately for each evaluation protocol from the corresponding relation memory and remains fixed during testing.

We convert the relation support into a non-negative log-odds residual relative to this memory prior:

$$r_{i,c} = [\mathrm{logit}(p_{i,c}^{\mathrm{ret}}) - \mathrm{logit}(\pi_c^{\mathrm{mem}})]_+, \tag{10}$$

where $[x]_+ = \max(x, 0)$ and $\mathrm{logit}(p) = \log\frac{p}{1-p}$. For numerical stability, both $p_{i,c}^{\mathrm{ret}}$ and π_c^{mem} are clipped to $[\epsilon, 1-\epsilon]$ before the logit transformation, with $\epsilon = 10^{-4}$. Thus, retrieval provides a positive correction only when the query-specific support exceeds its memory prior.

Logit Rectification. Let $z_{i,c}^{\mathrm{base}}$ denote the frozen base model's logit for relation category c. The base logit is refined by adding the prior-calibrated residual:

$$z_{i,c}^{\mathrm{aug}} = z_{i,c}^{\mathrm{base}} + \gamma r_{i,c}, \tag{11}$$

where we empirically set $\gamma = 0.5$. The refined relation probability is obtained via:

$$q_{i,c} = \sigma(z_{i,c}^{\mathrm{aug}}), \tag{12}$$

where $\sigma(\cdot)$ denotes the sigmoid function. Final prediction scores and candidate selection strictly adhere to the default output protocol of each task (see Appendix for further details).

Table 1: Fully supervised HOI detection results on the HICO-DET (Chao et al., 2018) TEST set. R50 denotes ResNet-50; ViT-B and ViT-L denote CLIP visual backbones. **Bold** indicates the best results.

Method	Backbone	Default			Known Object		
		Full	Non-Rare	Rare	Full	Non-Rare	Rare
DOQ	R50+ViT-B/16	33.28	34.50	29.19	–	–	–
GEN-VLKT	R50+ViT-B/16	33.75	35.10	29.25	37.80	38.71	34.76
ADA-CM	R50+ViT-B/16	33.80	34.42	31.72	–	–	–
HOICLIP	R50+ViT-B/32	34.69	35.74	31.12	37.61	38.54	34.47
VIL	R50+ViT-B/16	34.21	35.30	30.58	37.67	38.50	34.88
CQL	R50+ViT-B/16	35.36	36.07	32.97	–	–	–
LOGICHOI	R50+ViT-B/16	35.47	36.22	32.03	38.21	39.03	35.29
CMMP	R50+ViT-B/16	33.24	33.53	32.26	–	–	–
HOIGen	R50+ViT-B/16	34.84	34.94	34.52	–	–	–
CEFA	R50+ViT-B/16	35.00	35.81	32.30	38.23	39.02	35.62
GroupHOI	R50+ViT-B/16	36.70	37.26	34.86	39.42	39.91	37.78
ADA-CM	R50+ViT-L/14	38.40	38.66	37.52	–	–	–
UniVRD	R50+ViT-L/14	37.41	**39.95**	28.90	–	–	–
CMMP	R50+ViT-L/14	38.14	38.25	37.75	–	–	–
EZ-HOI	R50+ViT-L/14	38.56	38.89	37.45	41.86	41.99	41.39
+TF-TRA	R50+ViT-L/14	**39.75**	39.46	**40.71**	**43.06**	**42.79**	**43.96**

4 EXPERIMENTS

4.1 HUMAN-OBJECT INTERACTION DETECTION

Dataset. HICO-DET (Chao et al., 2018) serves as the evaluation benchmark, which contains 38,118 training images and 9,658 test images, covering 600 HOI categories across 117 verbs and 80 object classes. Evaluation includes both zero-shot and fully supervised settings, with performance measured by mean average precision (mAP). Zero-shot evaluation covers rare-first unseen composition (RF-UC), non-rare-first unseen composition (NF-UC), unseen object (UO), and unseen verb (UV), with mAP reported over Full, Seen, and Unseen categories. RF and NF assess generalization to unseen HOI combinations, whereas UO and UV assess generalization to unseen objects and verbs, respectively. Fully supervised evaluation follows the Default and Known Object protocols, reporting mAP over Full (600 categories), Rare (138 categories with fewer than 10 training instances), and Non-Rare (462 categories). Default evaluates each HOI category over all test images, whereas Known Object considers only images containing the corresponding object category.

Setup. LAIN (Kim et al., 2025) serves as the zero-shot baseline, while EZ-HOI (Lei et al., 2024a) serves as the fully supervised baseline. The offline relation memory is constructed from HICO-DET (Chao et al., 2018) training annotations. All 600 HOI categories are used under the fully supervised setting. For each zero-shot split, held-out annotations are removed before constructing multi-hot verb labels, retaining only entity pairs with at least one remaining valid interaction. For both zero-shot and fully supervised HOI settings, we use a per-group memory budget of $B = 20$, an aggregation temperature of $T = 0.05$, and retrieve $K = 128$ exemplars per query.

Fully-Supervised Results. As shown in Table 1, TF-TRA improves Full mAP over EZ-HOI (Lei et al., 2024a) by **1.19**% and **1.20**% under the Default and Known Object protocols, reaching 39.75% and 43.06%, respectively. These gains are driven by substantial improvements on rare categories: Rare mAP increases by **3.26**% and **2.57**%, respectively. The pronounced gains on rare interactions suggest that retrieved exemplars provide useful evidence for categories with limited training coverage. These results demonstrate the effectiveness of non-parametric relation memory in alleviating the long-tail distribution by complementing the detector's parametric knowledge with relational evidence.

Zero-Shot Results. As shown in Table 2, TF-TRA improves the overall performance of frozen LAIN (Kim et al., 2025) across all four zero-shot settings without additional training. Compared with the LAIN baseline, Full mAP increases by **1.14**%, **1.17**%, **1.12**%, and **1.85**% under RF-UC, NF-UC, UO, and UV, respectively, yielding an average improvement of **1.32**%. These overall gains are primarily driven by improvements on seen categories, with Seen mAP increasing by **1.23**%, **1.75**%,

Table 2: Zero-shot HOI detection results on HICO-DET (Chao et al., 2018) TEST set. R50 denotes ResNet-50; ViT-B and ViT-L denote CLIP visual backbones. **Bold** indicates the best results.

Method	Backbone	RF-UC			NF-UC			UO			UV		
		Full	Seen	Unseen	Full	Seen	Unseen	Full	Seen	Unseen	Full	Seen	Unseen
FCL	R50	22.01	24.23	13.16	19.37	19.55	18.66	19.87	20.74	15.54	–	–	–
ATL	R50	21.57	24.67	9.18	18.67	18.78	18.25	20.47	21.54	15.11	–	–	–
RLIP	R50	30.52	33.35	19.19	26.19	27.67	20.27	–	–	–	–	–	–
GEN-VLKT	R50+ViT-B	30.56	32.91	21.36	23.71	23.38	25.05	25.63	28.92	10.51	28.74	30.23	20.96
LOGICHOI	R50+ViT-B	33.17	34.93	25.97	27.95	27.86	26.84	28.23	30.42	15.67	30.77	31.88	24.57
ADA-CM	R50+ViT-B	33.01	34.35	27.63	31.39	31.13	32.41	–	–	–	–	–	–
HOICLIP	R50+ViT-B	32.99	34.85	25.53	27.75	28.10	26.39	28.53	30.99	16.20	31.09	32.19	24.30
CLIP	ViT-B	23.36	22.00	28.79	23.36	22.06	28.52	23.36	22.29	28.66	23.36	22.90	26.16
CLIP4HOI	R50+ViT-B	34.08	35.48	28.47	28.90	28.26	31.44	32.58	32.73	31.79	30.42	31.14	26.02
CMMP	R50+ViT-B	32.18	32.87	29.45	30.18	29.71	32.09	31.59	31.15	33.76	31.84	32.75	26.23
EZ-HOI	R50+ViT-B	33.13	34.15	29.02	31.17	30.55	33.66	32.27	32.06	33.28	32.32	33.49	25.10
HOLa	R50+ViT-B	34.19	35.08	30.61	32.36	31.64	35.25	33.59	33.02	36.45	34.09	35.09	27.91
VDRP	R50+ViT-B	33.78	34.41	31.29	32.57	31.60	**36.45**	33.39	32.84	36.13	32.73	33.72	26.69
LAIN	R50+ViT-B	34.41	35.06	31.83	33.23	32.44	36.41	34.27	33.55	37.88	33.12	33.80	**28.96**
+TF-TRA	R50+ViT-B	**35.55**	**36.29**	**32.61**	**34.40**	**34.19**	35.23	**35.39**	**34.88**	**37.92**	**34.97**	**35.95**	**28.96**

1.33%, and **2.15**% under RF-UC, NF-UC, UO, and UV, respectively. Meanwhile, performance on unseen categories remains largely stable, with changes in Unseen mAP ranging from -1.18% to $+0.78\%$. These results demonstrate that TF-TRA enhances recognition of seen interactions while largely preserving generalization to unseen interactions.

4.2 Scene Graph Generation

Dataset. Evaluation is conducted on Visual Genome (Xu et al., 2017), following the standard SGG split used by EGTR (Im et al., 2024). This split retains the 150 most frequent object categories and 50 predicate categories, with approximately 57K training, 5K validation, and 26K test images. The scene graph detection task requires localizing and classifying objects and predicting directed predicates between them. Performance is evaluated on the test split using Recall (R), mean Recall (mR), and their harmonic mean (F-score) at $K \in \{20, 50, 100\}$.

Setup. EGTR (Im et al., 2024) serves as the baseline SGG detector. The relation memory is constructed offline from the Visual Genome training split, storing ordered object-pair representations and their associated predicate annotations. The per-group memory budget is set to $B = 80$, the aggregation temperature to $T = 0.05$, and the number of retrieved exemplars per query to $K = 16$.

Results. As shown in Table 3, incorporating TF-TRA into EGTR consistently boosts mean Recall and F-score across all prediction budgets in a training-free manner, without updating model parameters. Specifically, mR@20, mR@50, and mR@100 improve by **3.3**%, **5.5**%, and **7.2**% (e.g., mR@100 rises from 10.1% to **17.3**%). Crucially, the substantial gains on rare categories far outweigh the marginal loss on frequent ones, driving F@20, F@50, and F@100 up to **12.5**%, **18.1**%, and **22.3**% (from 8.9%, 12.5%, and 15.6%). Since standard Recall is heavily dominated by frequent head predicates, a compromise in it is expected; however, the resulting increase in their harmonic mean indicates a better balance between instance-weighted and category-averaged recall, supporting the effectiveness of our TF-TRA. Overall, these results show that non-parametric relation memory serves as an effective complement to parametric representations, enabling more balanced predicate prediction.

4.3 Nonverbal Interaction Detection

Dataset. The NVI dataset (Wei et al., 2024) serves as the evaluation benchmark, containing 13,711 images divided into 9,634 training, 1,418 validation, and 2,659 test images. It annotates human individuals, social groups, and 22 nonverbal interactions, including 16 individual-wise and 6 group-wise behaviors across five interaction types. Following the NVI-DET protocol, the task is formulated as detecting $\langle \text{individual}, \text{group}, \text{interaction} \rangle$ triplets. Performance is evaluated using mean Recall at

Table 3: SGG results on the Visual Genome (Xu et al., 2017) TEST set. **Bold** indicates the best results.

Model	F-score			Mean Recall			Recall		
	F@20	F@50	F@100	mR@20	mR@50	mR@100	R@20	R@50	R@100
FCSGG	4.6	6.2	7.2	2.7	3.6	4.2	16.1	21.3	25.1
IMP (EBM)	4.8	7.2	9.2	2.8	4.2	5.4	18.1	25.9	31.2
Motifs	7.0	9.4	11.5	4.1	5.5	6.8	25.1	32.1	**36.9**
VCTree	8.2	10.9	12.7	4.9	6.6	7.7	24.8	31.8	36.1
VTransE	8.4	11.2	13.1	5.1	6.8	8.0	24.5	31.3	35.5
Iterative SGG	–	12.6	13.8	–	8.0	8.8	–	29.7	32.1
GPS-Net	–	11.0	13.9	–	6.7	8.6	–	31.1	35.9
VCTree (EBM)	9.2	12.4	14.5	5.7	7.7	9.1	24.2	31.4	35.9
VCTree (TDE)	9.2	12.6	15.0	6.9	9.3	11.1	14.0	19.4	23.2
SSR-CNN	9.9	13.4	15.7	6.1	8.4	10.0	**25.8**	**32.7**	**36.9**
Relationformer	7.6	14.0	15.9	4.6	9.3	10.7	22.2	28.4	31.3
BGNN	11.3	15.9	18.6	7.5	10.7	12.6	23.3	31.0	35.8
SGTR	–	16.1	19.8	–	12.0	15.2	–	24.6	28.4
RelTR	10.3	15.5	–	6.8	10.8	–	21.2	27.5	–
EGTR	8.9	12.5	15.6	5.5	7.9	10.1	23.5	30.2	34.3
+TF-TRA	**12.5**	**18.1**	**22.3**	**8.8**	**13.4**	**17.3**	21.8	27.7	31.4

Table 4: NVI-DET results on the NVI (Wei et al., 2024) VAL and TEST sets (%). AR denotes the average of mR@25, mR@50, and mR@100. **Bold** indicates the best results.

Method	Val				Test			
	mR@25	mR@50	mR@100	AR	mR@25	mR@50	mR@100	AR
m-QPIC	56.89	69.52	78.36	68.26	59.44	71.46	80.07	70.32
m-CDN	55.57	71.06	78.81	68.48	59.01	72.94	82.61	71.52
m-GEN-VLKT	50.59	70.87	80.08	67.18	56.68	74.32	84.18	71.73
NVI-DEHR	54.85	73.42	85.33	71.20	59.46	76.01	88.52	74.66
GroupHOI	55.67	76.73	87.16	73.19	62.67	76.57	86.42	75.22
+TF-TRA	**60.54**	**76.95**	**87.56**	**75.02**	**66.61**	**80.19**	**90.60**	**79.13**

K (mR@K), where $K \in \{25, 50, 100\}$, together with their average recall (AR). We report results for all interaction categories on both the validation and test sets.

Setup. GroupHOI (Hong et al., 2025) serves as the baseline NVI detector. The relation memory is built offline exclusively from the NVI training set, and stores representations of individual-/group-level relation instances with their associated labels. The per-group memory budget is set to $B = 180$, the aggregation temperature to $T = 0.1$, and the number of retrieved exemplars per query to $K = 1024$.

Results. As shown in Table 4, integrating TF-TRA into GroupHOI (Hong et al., 2025) improves all reported metrics on both the validation and test sets without updating model parameters, achieving the best performance among the compared methods. On the validation set, TF-TRA increases AR from 73.19% to **75.02**% and mR@25 from 55.67% to **60.54**%, yielding gains of **1.83**% and **4.87**%, respectively. On the test set, AR improves from 75.22% to **79.13**%, with gains of **3.94**%, **3.62**%, and **4.18**% in mR@25, mR@50, and mR@100, respectively. These consistent gains on both splits support the effectiveness of non-parametric relation memory in enhancing nonverbal interaction detection.

4.4 Diagnostic Study

Memory Budget. We investigate the effect of memory capacity by varying the number of retained instances per group. As shown in Table 5a, even $B = 1$ improves average Full mAP from 33.76% to 34.45% over the baseline without retrieval ($B = 0$). Increasing the budget improves performance up to $B = 20$, which achieves the highest average Full mAP of 35.08%. Further increasing B to 50 or 100 yields slightly lower averages of 35.02% and 35.00%, respectively, suggesting that a compact memory can retain sufficient reference instances for effective prediction refinement.

Table 5: Diagnostic study on HICO-DET TEST set with LAIN as the base detector. Full mAP (%) is reported under four zero-shot settings; S: semantic similarity; V: visual similarity; D: diversity-aware reranking. For $B=0$, the retrieval residual is disabled. **Bold** indicates the best results.

B	RF	NF	UO	UV	Avg.
0	34.41	33.23	34.27	33.12	33.76
1	35.11	33.75	34.86	34.07	34.45
5	35.39	34.27	35.29	34.89	34.96
10	35.44	34.32	**35.39**	34.92	35.02
20	35.55	**34.40**	**35.39**	**34.97**	**35.08**
50	35.55	34.31	35.33	34.88	35.02
100	**35.58**	34.29	35.29	34.84	35.00

(a) Memory budget.

S	V	D	RF	NF	UO	UV	Avg.
✓	×	×	34.66	33.86	34.79	33.34	34.16
×	✓	×	34.83	33.87	34.83	34.21	34.43
✓	×	✓	34.88	33.90	34.76	33.84	34.34
×	✓	✓	34.86	33.84	34.82	34.23	34.44
✓	✓	×	35.37	34.32	35.31	34.77	34.94
✓	✓	✓	**35.55**	**34.40**	**35.39**	**34.97**	**35.08**

(b) Retrieval components.

Aggregation	RF	NF	UO	UV	Avg.
Uniform	35.01	33.76	34.70	33.68	34.29
Weighted	**35.55**	**34.40**	**35.39**	**34.97**	**35.08**

(c) Neighborhood aggregation.

Residual	RF	NF	UO	UV	Avg.
TF-TRA	**35.55**	**34.40**	**35.39**	**34.97**	**35.08**
w/o prior	34.66	33.48	34.48	33.39	34.00
w/o clipping	35.52	33.73	**35.39**	34.90	34.88

(d) Prior-calibrated residual extraction.

Retrieval Components. Table 5b examines the contributions of semantic similarity (S), visual similarity (V), and diversity-aware reranking (D). Combining the two similarities consistently outperforms either alone, improving average Full mAP by 0.73% and 0.64% over the semantic-only and visual-only variants, respectively. Furthermore, incorporating diversity-aware reranking into the combined similarity yields an additional gain of 0.14% in average Full mAP across all four settings. These results support combining semantic and visual relevance, with label-diversity reranking providing a further improvement over similarity-only retrieval.

Neighborhood Aggregation. Table 5c compares uniform aggregation with the similarity-weighted aggregation in Eq. 8, using the same retrieved neighbors and prediction settings. Similarity weighting improves Full mAP across all four settings, increasing the average from 34.29% to 35.08%, with the largest gain on UV (1.29%). These results show that distinguishing neighbor contributions by query relevance is more effective than treating all retrieved exemplars equally.

Prior-Calibrated Residual. Table 5d examines the contributions of memory-prior correction and positive residual clipping in Eq. 10. Removing prior correction reduces average Full mAP by 1.08%, from 35.08% to 34.00%, highlighting the benefit of correcting for category prevalence in the relation memory. Similarly, removing positive clipping, thereby allowing negative residual updates, degrades performance on three of the four evaluation splits, with the largest drop of 0.67% observed on NF. Together, these results support calibrating relation support against the memory prior and retaining only positive residual signals for effective prediction refinement.

5 Conclusion

This paper introduces TF-TRA, a training-free recipe that complements the parametric memory of visual relationship detectors with non-parametric relation memory constructed from training annotations. TF-TRA first constructs the relation memory by grouping exemplars and capping per-group capacity, mitigating the dominance of frequent relation configurations. During inference, it retrieves relevant exemplars from this memory by combining semantic and visual similarity with diversity-aware reranking to reduce redundancy. The retrieved annotations are then aggregated and calibrated against memory priors to refine the detector's predictions. Extensive experiments on HOI detection, SGG, and NVI detection, including both fully supervised and zero-shot HOI settings, show consistent performance gains over existing detectors without parameter updates. These findings reveal that training annotations remain underutilized when confined to offline supervision. Overall, TF-TRA charts an alternative roadmap for visual relationship detection, highlighting the untapped potential of test-time memory retrieval alongside purely parametric modeling.

AI USE STATEMENT

We use generative AI tools to assist with translation and language editing of the manuscript, including improving grammar, clarity, and readability. Their use is limited to these writing tasks. The authors review the AI-assisted text for accuracy and consistency with the intended scientific meaning. We take full responsibility for the final content of this paper, including all AI-assisted text.

REPRODUCIBILITY STATEMENT

Section 3 describes the group-capped relation memory construction, diversity-aware exemplar retrieval, and prior-calibrated logit rectification procedures. Section 4 specifies the datasets, evaluation protocols, base detectors, and task-specific hyperparameter settings. Appendix A provides additional details on annotation processing, instance matching, feature extraction, and task-specific prediction protocols. All relation memories are constructed exclusively from the training splits; for zero-shot HOI evaluation, held-out annotations are removed before constructing the memory labels. The base models and relation memories remain fixed during inference. Our source code will be released.

REFERENCES

Irving Biederman, Robert J Mezzanotte, and Jan C Rabinowitz. Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive psychology*, 1982.

Jiarui Cai, Mingze Xu, Wei Li, Yuanjun Xiong, Wei Xia, Zhuowen Tu, and Stefano Soatto. Memot: Multi-object tracking with memory. In *CVPR*, 2022.

Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018.

Yihong Chen, Yue Cao, Han Hu, and Liwei Wang. Memory enhanced global-local aggregation for video object detection. In *CVPR*, 2020.

Yuren Cong, Michael Ying Yang, and Bodo Rosenhahn. Reltr: Relation transformer for scene graph generation. *IEEE TPAMI*, 2023.

Shuangrui Ding, Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Yuwei Guo, Dahua Lin, and Jiaqi Wang. Sam2long: Enhancing sam 2 for long video segmentation with a training-free memory tree. In *ICCV*, 2025.

Shuman Fang, Shuai Liu, Jie Li, Guannan Jiang, Xianming Lin, and Rongrong Ji. Improving human-object interaction detection via virtual image learning. In *ACM MM*, 2023.

Yixin Guo, Yu Liu, Jianghao Li, Weimin Wang, and Qi Jia. Unseen no more: Unlocking the potential of clip for generative zero-shot hoi detection. In *ACM MM*, 2024.

Jiajun Hong, Jianan Wei, and Wenguan Wang. Learning human-object interaction as groups. *NeurIPS*, 2025.

Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Affordance transfer learning for human-object interaction detection. In *CVPR*, 2021a.

Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Detecting human-object interaction via fabricated compositional learning. In *CVPR*, 2021b.

Yupeng Hu, Changxing Ding, Chang Sun, Shaoli Huang, and Xiangmin Xu. Bilateral collaboration with large vision-language models for open vocabulary human-object interaction detection. In *ICCV*, 2025.

Jinbae Im, JeongYeon Nam, Nokyung Park, Hyungmin Lee, and Seunghyun Park. Egtr: Extracting graph from transformer for scene graph generation. In *CVPR*, 2024.

Daniel Kaiser, Genevieve L Quek, Radoslaw M Cichy, and Marius V Peelen. Object vision in a structured world. *Trends in cognitive sciences*, 2019.

Geethan Karunaratne, Manuel Schmuck, Manuel Le Gallo, Giovanni Cherubini, Luca Benini, Abu Sebastian, and Abbas Rahimi. Robust high-dimensional memory-augmented neural networks. *Nature communications*, 2021.

Siddhesh Khandelwal and Leonid Sigal. Iterative scene graph generation. *NeurIPS*, 2022.

Kibum Kim, Kanghoon Yoon, Jaehyeong Jeon, Yeonjun In, Jinyoung Moon, Donghyun Kim, and Chanyoung Park. Llm4sgg: Large language models for weakly supervised scene graph generation. In *CVPR*, 2024.

Sanghyun Kim, Deunsol Jung, and Minsu Cho. Locality-aware zero-shot human-object interaction detection. In *CVPR*, 2025.

Dharshan Kumaran, Demis Hassabis, and James L McClelland. What learning systems do intelligent agents need? complementary learning systems theory updated. *Trends in cognitive sciences*, 2016.

Qinqian Lei, Bo Wang, and Robby T Tan. Ez-hoi: Vlm adaptation via guided prompt learning for zero-shot hoi detection. *NeurIPS*, 2024a.

Qinqian Lei, Bo Wang, and Robby T Tan. Hola: Zero-shot hoi detection with low-rank decomposed vlm feature adaptation. In *ICCV*, 2025.

Ting Lei, Fabian Caba, Qingchao Chen, Hailin Jin, Yuxin Peng, and Yang Liu. Efficient adaptive human-object interaction detection with concept-guided memory. In *ICCV*, 2023.

Ting Lei, Shaofeng Yin, Yuxin Peng, and Yang Liu. Exploring conditional multi-modal prompts for zero-shot hoi detection. In *ECCV*, 2024b.

Liulei Li, Jianan Wei, Wenguan Wang, and Yi Yang. Neural-logic human-object interaction detection. *NeurIPS*, 2023.

Rongjie Li, Songyang Zhang, Bo Wan, and Xuming He. Bipartite graph network with adaptive message passing for unbiased scene graph generation. In *CVPR*, 2021.

Rongjie Li, Songyang Zhang, and Xuming He. Sgtr: End-to-end scene graph generation with transformer. In *CVPR*, 2022.

Yue Liao, Aixi Zhang, Miao Lu, Yongliang Wang, Xiaobo Li, and Si Liu. Gen-vlkt: Simplify association and enhance interaction understanding for hoi detection. In *CVPR*, 2022.

Xin Lin, Changxing Ding, Jinquan Zeng, and Dacheng Tao. Gps-net: Graph property sensing network for scene graph generation. In *CVPR*, 2020.

Hengyue Liu, Ning Yan, Masood Mortazavi, and Bir Bhanu. Fully convolutional scene graph generation. In *CVPR*, 2021.

Alexander Long, Wei Yin, Thalaiyasingam Ajanthan, Vu Nguyen, Pulak Purkait, Ravi Garg, Alan Blair, Chunhua Shen, and Anton Van den Hengel. Retrieval augmented classification for long-tail visual recognition. In *CVPR*, 2022.

Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *ECCV*, 2016.

Yunyao Mao, Jiajun Deng, Wengang Zhou, Li Li, Yao Fang, and Houqiang Li. Clip4hoi: Towards adapting clip for practical zero-shot hoi detection. *NeurIPS*, 2023.

Shan Ning, Longtian Qiu, Yongfei Liu, and Xuming He. Hoiclip: Efficient knowledge transfer for hoi detection with vision-language models. In *CVPR*, 2023.

Jae Sung Park, Zixian Ma, Linjie Li, Chenhao Zheng, Cheng-Yu Hsieh, Ximing Lu, Khyathi Chandu, Quan Kong, Norimasa Kobori, Ali Farhadi, et al. Synthetic visual genome. In *CVPR*, 2025.

Xian Qu, Changxing Ding, Xingao Li, Xubin Zhong, and Dacheng Tao. Distillation using oracle queries for transformer-based human-object interaction detection. In *CVPR*, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *CVPR*, 2017.

Suprosanna Shit, Rajat Koner, Bastian Wittmann, Johannes Paetzold, Ivan Ezhov, Hongwei Li, Jiazhen Pan, Sahand Sharifzadeh, Georgios Kaissis, Volker Tresp, et al. Relationformer: A unified framework for image-to-graph generation. In *ECCV*, 2022.

Mohammed Suhail, Abhay Mittal, Behjat Siddiquie, Chris Broaddus, Jayan Eledath, Gerard Medioni, and Leonid Sigal. Energy-based learning for scene graph generation. In *CVPR*, 2021.

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. *NeurIPS*, 2015.

Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In *CVPR*, 2021.

Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *CVPR*, 2019.

Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *CVPR*, 2020.

Yao Teng and Limin Wang. Structured sparse r-cnn for direct scene graph generation. In *CVPR*, 2022.

Jianan Wei, Tianfei Zhou, Yi Yang, and Wenguan Wang. Nonverbal interaction detection. In *ECCV*, 2024.

Chi Xie, Fangao Zeng, Yue Hu, Shuang Liang, and Yichen Wei. Category query learning for human-object interaction classification. In *CVPR*, 2023.

Guo-Sen Xie, Huan Xiong, Jie Liu, Yazhou Yao, and Ling Shao. Few-shot semantic segmentation with cyclic memory network. In *ICCV*, 2021.

Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *CVPR*, 2017.

Chanhyeong Yang, Jihwan Park, Hyunwoo J Kim, et al. Visual diversity and region-aware prompt learning for zero-shot hoi detection. *NeurIPS*, 2025.

Jie Yang, Bingliang Li, Ailing Zeng, Lei Zhang, and Ruimao Zhang. Open-world human-object interaction detection via multi-modal prompts. In *CVPR*, 2024.

Hangjie Yuan, Jianwen Jiang, Samuel Albanie, Tao Feng, Ziyuan Huang, Dong Ni, and Mingqian Tang. Rlip: Relational language-image pre-training for human-object interaction detection. *NeurIPS*, 2022.

Hangjie Yuan, Shiwei Zhang, Xiang Wang, Samuel Albanie, Yining Pan, Tao Feng, Jianwen Jiang, Dong Ni, Yingya Zhang, and Deli Zhao. Rlipv2: Fast scaling of relational language-image pre-training. In *ICCV*, 2023.

Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *CVPR*, 2018.

Aixi Zhang, Yue Liao, Si Liu, Miao Lu, Yongliang Wang, Chen Gao, and Xiaobo Li. Mining the benefits of two-stage and one-stage hoi detection. *NeurIPS*, 2021.

Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. Visual translation embedding network for visual relation detection. In *CVPR*, 2017.

Lijun Zhang, Wei Suo, Peng Wang, and Yanning Zhang. A plug-and-play method for rare human-object interactions detection by bridging domain gap. In *ACM MM*, 2024a.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *ECCV*, 2022.

Yabin Zhang, Wenjie Zhu, Hui Tang, Zhiyuan Ma, Kaiyang Zhou, and Lei Zhang. Dual memory networks: A versatile adaptation approach for vision-language models. In *CVPR*, 2024b.

Long Zhao, Liangzhe Yuan, Boqing Gong, Yin Cui, Florian Schroff, Ming-Hsuan Yang, Hartwig Adam, and Ting Liu. Unified visual relationship detection with vision and language models. In *ICCV*, 2023.

Chaofan Zheng, Xinyu Lyu, Lianli Gao, Bo Dai, and Jingkuan Song. Prototype-based embedding network for scene graph generation. In *CVPR*, 2023.

Xiangyang Zhu, Renrui Zhang, Bowei He, Aojun Zhou, Dong Wang, Bin Zhao, and Peng Gao. Not all features matter: Enhancing few-shot clip with adaptive prior refinement. In *ICCV*, 2023.

SUMMARY OF THE APPENDIX

This appendix contains additional details for the ICLR 2027 submission, titled *"Training-Free Test-Time Retrieval Augmentation for Visual Relationship Detection"*. The appendix is organized as follows:

- §A introduces more implementation details of TF-TRA.
- §B provides references for all compared methods in Tables 1, 2, 3, and 4.
- §C provides additional experimental results of TF-TRA.
- §D discusses our limitations, broader impact, and directions of our future work.

A More Implementation Details

A.1 Relation Memory Construction

We describe how the four components of each source entry $(\boldsymbol{s}_j, \boldsymbol{v}_j, \eta_j, \boldsymbol{y}_j)$ are constructed for HOI, SGG, and NVI. Source instances are obtained exclusively from the training split of the corresponding dataset. Entity metadata and relation labels are derived from ground-truth annotations, while relation and visual representations are extracted using frozen models in evaluation mode.

Entity Metadata η_j. The metadata η_j encodes the entity categories used for group-capped memory construction. For HOI and SGG, it records the ordered subject and object categories. The subject category is fixed to *person* in HOI, whereas both categories vary in SGG. For NVI, both individual- and group-level instances involve only persons, so we use a constant category code corresponding to *person* for all entries. Consequently, grouping by $(\eta_j, \boldsymbol{y}_j)$ reduces to grouping by the complete positive-label set in NVI.

Relation Annotation $\boldsymbol{y}_j$. The relation annotation is formulated as a multi-hot binary vector $\boldsymbol{y}_j \in \{0,1\}^C$, where C denotes the task-specific relation vocabulary size. Annotations associated with the same instance are merged, allowing an entry to contain multiple positive labels. For HOI, the vector spans 117 verb categories, and annotations associated with the same human–object pair are merged. Under each zero-shot protocol, annotations belonging to held-out HOI categories are removed before constructing the multi-hot verb labels; pairs with no remaining valid annotations are discarded. For SGG, the vector encompasses 50 predicate classes, with multiple predicates between the same ordered entity pair merged into one label vector. For NVI, the vector contains 22 categories: 16 individual-level and six group-level categories. Individual annotations are merged for the same person, while group annotations are merged when they share the same subject and group membership.

Relation Representation $\boldsymbol{s}_j$. Each ground-truth instance is paired with a candidate prediction generated by the corresponding frozen base model, from which the relation representation is extracted. For HOI and SGG, candidate matching enforces endpoint category agreement and maximizes $\min(\text{IoU}_{\text{sub}}, \text{IoU}_{\text{obj}})$. Both subject and object IoUs must exceed a threshold of 0.1; unassigned instances are discarded. For NVI, individual instances are matched strictly via subject-box IoU, whereas group instances maximize $\min(\text{IoU}_{\text{sub}}, \text{IoU}_{\text{grp}})$ without an explicit IoU threshold. Matching is conducted independently across all ground-truth instances. For HOI, we use the matched candidate's interaction feature; for SGG, the ordered concatenation of the projected subject and object features; and for NVI, the matched candidate's query feature. In our implementation, relation representations are 512-dimensional for all three tasks and are ℓ_2-normalized before similarity computation.

Visual Region Representation $\boldsymbol{v}_j$. Visual features are extracted from annotated image regions using a frozen CLIP ViT-L/14 image encoder. For HOI and SGG, we extract visual features from the union of the ground-truth subject and object boxes. For NVI, we use the annotated person region for individual-level predictions and the region enclosing all annotated group members for group-level predictions. These regions are cropped with surrounding context and processed by a frozen CLIP ViT-L/14 visual encoder, yielding ℓ_2-normalized, 768-dimensional representations $\boldsymbol{v}_j$.

A.2 VRD Prediction Protocols

Let $q_{i,c}$ denote the refined probability of relation category c for candidate i. Retrieval updates relation predictions while preserving the base model's predicted boxes and entity categories. The refined probabilities are converted into final prediction scores using the task-specific procedures below.

Human-Object Interaction Detection. For each detected human–object pair i with human and object detection confidences d_i^h and d_i^o, the final score for verb c is:

$$S_{i,c}^{\mathrm{HOI}} = q_{i,c} \cdot d_i^h \cdot d_i^o. \tag{13}$$

Each predicted verb is combined with the predicted object category through the dataset's verb–object correspondence to produce an HOI category. Only valid combinations in the 600-category HICO-DET vocabulary are emitted. Multiple verbs may be predicted for the same entity pair; no per-pair argmax over verbs is applied. The final detections retain the base detector's bounding boxes and are ranked by $S_{i,c}^{\mathrm{HOI}}$ within each HOI category for mean average precision (mAP) evaluation.

Scene Graph Generation. For an ordered subject–object pair i, let d_i^s and d_i^o denote the predicted entity class confidences, and let a_i represent the base model's pairwise connectivity score. The final predicate prediction score is given by:

$$S_{i,c}^{\mathrm{SGG}} = q_{i,c} \cdot a_i \cdot d_i^s \cdot d_i^o. \tag{14}$$

Entity categories and their confidences are obtained from the foreground entity-class probabilities. Both the entity predictions and connectivity scores remain fixed during retrieval augmentation. Following the standard graph-constrained protocol, each ordered pair contributes only its highest-scoring predicate, and self-connections are disallowed. The triplets are ranked within each image, and Recall@k and mean Recall@k are computed from the top-k predictions for $k \in 20, 50, 100$.

Nonverbal Interaction Detection. Each candidate instance specifies a person box, a group box, and classification scores over 22 interaction categories, comprising 16 individual-level and 6 group-level categories. Individual- and group-level refinements use subject- and group-region visual representations, respectively, and are concatenated into a unified 22-dimensional refined probability vector $\boldsymbol{q}_i$. Unlike HOI and SGG, NVI does not apply additional entity-confidence multipliers:

$$S_{i,c}^{\mathrm{NVI}} = q_{i,c}. \tag{15}$$

All candidate-category pairs are ranked jointly within each image, so that individual- and group-level predictions compete under a shared output budget without branch-specific quotas or heuristics. Evaluation uses the top $k \in \{25, 50, 100\}$ predictions at IoU thresholds $\{0.25, 0.50, 0.75\}$. An individual-level prediction is matched according to its person box, whereas a group-level prediction requires both its person and group boxes to satisfy the specified IoU threshold. For each output budget k, recall is computed separately for each interaction category and IoU threshold and then averaged over all categories and the three IoU thresholds to obtain mean Recall@k (mR@k). Average recall (AR) is defined as the arithmetic mean of mR@25, mR@50, and mR@100.

B Compared Methods

This section provides references for the methods included in Tables 1, 2, 3, and 4.

Human-Object Interaction Detection. The methods in Tables 1 and 2 include FCL (Hou et al., 2021b), ATL (Hou et al., 2021a), RLIP (Yuan et al., 2022), DOQ (Qu et al., 2022), GEN-VLKT (Liao et al., 2022), LOGICHOI (Li et al., 2023), ADA-CM (Lei et al., 2023), HOICLIP (Ning et al., 2023), CLIP (Radford et al., 2021), CLIP4HOI (Mao et al., 2023), VIL (Fang et al., 2023), CQL (Xie et al., 2023), UniVRD (Zhao et al., 2023), HOIGen (Guo et al., 2024), CEFA (Zhang et al., 2024a), CMMP (Lei et al., 2024b), EZ-HOI (Lei et al., 2024a), HOLa (Lei et al., 2025), VDRP (Yang et al., 2025), LAIN (Kim et al., 2025), and GroupHOI (Hong et al., 2025).

Scene Graph Generation. The comparison in Table 3 includes IMP (Xu et al., 2017), VTransE (Zhang et al., 2017), Motifs (Zellers et al., 2018), VCTree (Tang et al., 2019), GPS-Net (Lin et al., 2020), BGNN (Li et al., 2021), FCSGG (Liu et al., 2021), RelTR (Cong et al., 2023), SGTR (Li et al., 2022), Relationformer (Shit et al., 2022), Iterative SGG (Khandelwal & Sigal, 2022),

SSR-CNN (Teng & Wang, 2022), and EGTR (Im et al., 2024). The IMP (EBM) and VCTree (EBM) entries use the energy-based learning framework of Suhail et al. (2021). VCTree (TDE) additionally refers to the total direct effect formulation of Tang et al. (2020).

Nonverbal Interaction Detection. The methods in Table 4 include m-QPIC, m-CDN, m-GEN-VLKT, NVI-DEHR (Wei et al., 2024), and GroupHOI (Hong et al., 2025). The prefix m denotes the NVI adaptations introduced by Wei et al. (2024), based on QPIC (Tamura et al., 2021), CDN (Zhang et al., 2021), and GEN-VLKT (Liao et al., 2022), respectively.

C Additional Experimental Results

C.1 Memory Selection

Table 6 compares herding with uniform random sampling without replacement. All other settings remain unchanged. Herding improves all three validation recall metrics, increasing AR from 74.69% to 75.02%. On the test set, herding achieves higher mR@25, while random sampling obtains slightly higher mR@50 and mR@100; their overall AR is nearly identical (79.13% vs. 79.09%). These results support retaining herding as our default configuration.

Table 6: Memory selection on NVI. AR is the average of mR@25, mR@50, and mR@100. **Bold** indicates the best result in each column.

Memory selection	Val				Test			
	mR@25	mR@50	mR@100	AR	mR@25	mR@50	mR@100	AR
Random	60.31	76.92	86.83	74.69	66.42	**80.20**	**90.64**	79.09
Herding	**60.54**	**76.95**	**87.56**	**75.02**	**66.61**	80.19	90.60	**79.13**

D Discussion

Limitations. TF-TRA is designed to complement the parametric knowledge learned by the base model with non-parametric memory. Its gains depend on the coverage and quality of the annotated relation memory: If a relation category has no positive annotations in the memory, its retrieval support will be zero. Moreover, TF-TRA refines relation predictions while preserving the base detector's entity predictions, and therefore cannot recover relationships whose constituent entities are missed or incorrectly localized. Finally, although TF-TRA is training-free, memory retrieval incurs inference overhead, alongside storage costs for the non-parametric memory.

Social Impact. TF-TRA enables existing visual relationship detectors to incorporate explicit references in a training-free manner, potentially reducing the resources required to improve their predictions. However, the relation memory may inherit annotation biases and imbalanced coverage from its source data, which can influence predictions despite diversity-aware retrieval and prior calibration. In applications involving people, especially nonverbal interaction analysis, incorrect predictions may lead to unwarranted interpretations of human behavior. Responsible use therefore requires attention to data consent and privacy, evaluation across relevant populations and contexts, and human oversight in consequential applications.

Future Work. Looking ahead, we plan to extend relation memory to be more compact, structured, and adaptive. First, exploring coreset-based memory selection is a promising avenue to retain diverse interaction patterns and long-tail rare relations under strict storage budgets. Second, moving beyond simple pairwise similarity matching, we aim to investigate structured memory designs, such as hierarchical organization and multi-hop retrieval. Third, we plan to extend the framework to few-shot scenarios, leveraging sparse annotated exemplars to enrich relation memory and enable rapid adaptation to novel relation categories without updating the base model.