ELSA: EVALUATING LOCALIZATION OF SOCIAL ACTIVITIES IN URBAN STREETS USING OPEN-VOCABULARY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce ELSA: Evaluating Localization of Social Activities, to our knowledge the first multi-label benchmark for detecting individual and group activities in still street-level images. ELSA draws on theoretical frameworks in urban sociology and design, and covers in-the-wild scenes in which group sizes and activity types vary significantly. The dataset includes 1,500 manually annotated images with more than 6,000 bounding boxes. Existing Open-Vocabulary Detection (OVD) models face a number of challenges. They often struggle with semantic consistency across diverse inputs and are sensitive to slight variations in input phrasing, leading to inconsistent performance. The calibration of their predictive confidence, especially in complex multi-label scenarios, remains suboptimal, frequently resulting in overconfident predictions that do not accurately reflect their context understanding. We expose these limitations and show that modern OVD models perform poorly on ELSA. We further propose Normalized Log-Sum-Exp (N-LSE), a context-aware confidence score for these models. Unlike the commonly used Max-Logit, which relies on a single token, N-LSE accounts for every label in the query, yielding a less inflated ranking and better detection results. Finally, we present Dynamic Box Aggregation (DBA), an evaluation algorithm that penalizes contradictory predictions for the same person or group and recovers correct boxes that NMS discards, providing a more reliable AP. We release our dataset and code[1].

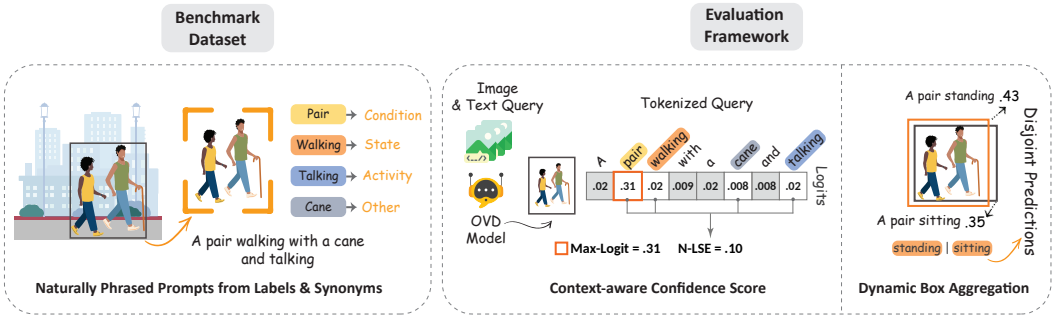


Figure 1: We present ELSA: Evaluating Localization of Social Activities—a novel *benchmark dataset* and *evaluation framework* for assessing open-vocabulary detection (OVD) models in recognizing and localizing social interactions on urban streets from still images. ELSA includes a multi-label annotation scheme spanning four categories: Condition, State, Activity, and Other. It features natural language prompts derived from these labels, along with synonymous variations to rigorously test models' semantic comprehension. Our *N-LSE* context-aware confidence score surpasses Max-Logit scoring by yielding scores less dominated by a single token and effectively reducing false positives. Our *DBA* algorithm dynamically groups overlapping predictions, ensuring semantic coherence and recovering correct predictions that otherwise would be missed by class-agnostic NMS.

[1] https://anonymous.4open.science/r/elsa-release/

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.

1 INTRODUCTION

Recently, increased focus on the human scale of the cities has drawn more attention to public spaces and pedestrian facilities. For decades, urban scholars have been fascinated by the complex interplay between public spaces and the social interactions they support (Mehta & Bosson, 2021; Whyte, 1980; Jacobs, 1961). However, scientific inquiry into the distribution of social activities has been hampered by data collection costs and time requirements.

Advances in computer vision and publicly available street-level imagery have opened new avenues for observational studies at reduced cost and increased scale. Activity recognition techniques are mostly designed for videos (Lan et al., 2012; Sun et al., 2023; Jahangard et al., 2024; Kim et al., 2024b), since human activity inherently involves motion and sequences of actions. Yet, continuous video footage across an entire city is difficult to procure and incurs high storage and processing costs. Object detection on still images offers a low-cost and efficient alternative that can identify and localize complex social interactions in diverse settings (Choi et al., 2014; Pratt et al., 2020; Lei et al., 2024), where the environmental context shapes the range of possible interactions and each image can contain a large number of people engaged in diverse activities (Wang et al., 2020).

While conventional object detectors are limited to a closed set of predefined classes, open-vocabulary detection (OVD; Zhu & Chen, 2024; Wu et al., 2024) utilize the abundance of language data to facilitate the detection of uncommon classes in standard benchmark training data. A robust OVD model is expected to handle a wide range of input terms and phrases that were not explicitly part of its training set. This is crucial for models deployed in real-world settings, such as urban streets, where unpredictable and varied interactions are common. The absence of benchmark data for OVD of social and individual actions in still images 'in the wild' hinders the development of models that generalize well across diverse and spontaneous urban scenarios, where the context and variability of human activities are far greater than those typically encountered in controlled environments. Furthermore, OVDs pose new challenges in both localization and semantic understanding of unseen new categories. They often struggle with semantic consistency across diverse inputs, are sensitive to slight variations in input phrasing, and have suboptimal predictive confidence calibration in out-of-distribution scenarios, resulting in overconfident predictions that do not accurately reflect their actual accuracy (Schulter et al., 2023; Dave et al., 2021).

In response to these challenges, we propose ELSA, a new benchmark dataset and evaluation framework to evaluate the performance of OVD models in recognizing and localizing human activity in urban streets from still images. Our dataset employs a multi-label annotation scheme encompassing 26 distinct labels that can be concurrently assigned to each bounding box. This results in over 6,000 bounding boxes annotated with 265 unique combinations of human activities across 1,500 street view images. We used a template-based generator to produce precise, naturally phrased sentences for each label combination and their near synonyms, totaling 1,361 unique prompts. These prompts were applied to each image during evaluation, providing a comprehensive and nuanced assessment of the models' ability to handle varied and contextually rich descriptions of human activities.

Recognizing the intertwined nature of OVD models with language features—and the evaluation complexities this presents—we further introduce Dynamic Box Aggregation (DBA) to address overlapping detections and disjoint predictions. Unlike Non-Maximum Suppression (NMS), which underlies NMS-AP (Yao et al., 2024b), DBA retains predictions whose confidence is within a threshold of the maximum, while penalizing disjoint predictions for the same box. DBA integrates into metrics such as Average Precision (AP) and F1-score, capturing both localization and semantic understanding. We also propose N-LSE, a confidence score for phrase grounding and referring expressions. N-LSE prioritizes predictions based on the most salient sub-phrases and tokens within the query, aligning token-level language features with visual features in the predicted regions.

Our contributions are threefold: **1)** ELSA, to our knowledge the first multi-label benchmark for detecting individual and group activities in still street-level images; **2)** N-LSE, a context-aware confidence score that accounts for every label in the query; and **3)** DBA, an evaluation algorithm that penalizes disjoint predictions for the same box and recovers correct boxes that NMS discards.

These innovations address key challenges in OVD model evaluation and offer valuable tools for advancing future research in this domain.

2 Related Work

Social Interactions in Public Spaces. Vibrant streets rich in interpersonal exchange have fascinated urban scholars because of their social qualities as well as fundamental indicators of sustainable urban environments (Mehta & Bosson, 2021). Whyte (1980) along with Jacobs (1961) highlight the intrinsic value of public spaces in fostering vibrant social life. Gehl (1968) describes activities in the public spaces as a spectrum between optional activities, such as talking with friends, and necessary activities, like walking to work. The public space observational method (Gehl & Svarre, 2013) distinguishes active social group activities, such as dining or talking together, from passive activities, such as strangers sitting on a bench checking their cell-phones. Inspired by this research, we define the target set of social activities in ELSA.

Open-Vocabulary Object Detection. OVD, first introduced by Zareian et al. (2021), primarily tackles the limitation of traditional object detection models that rely on a pre-defined closed set of objects. OVD models (Bravo et al., 2023; Minderer et al., 2023; Kim et al., 2024a) are tested on various OVD benchmark datasets (Schulter et al., 2023; Yao et al., 2024b). At their core, these models often use a vision–language contrastive loss to align semantics and concepts in the two modalities (Kamath et al., 2021; Liu et al., 2024; Cheng et al., 2024; Minderer et al., 2022; Li et al., 2022), with additional soft token prediction in MDETR (Kamath et al., 2021). Using a dual-encoder-single-decoder architecture, Grounding DINO (Liu et al., 2024) extends DINO (Zhang et al., 2023) such that, given a text prompt, a language-guided query selection picks the image features most relevant to the text as queries for the cross-modal decoder. A contrastive loss between the decoder outputs and the text tokens, together with L1 and generalized intersection over union (GIoU) box losses, is optimized end-to-end. Detic (Zhou et al., 2022b) trains the classifier of an image detector on classification data, expanding the detector's vocabulary to tens of thousands of concepts. In OWL-ViT (Minderer et al., 2022), the fixed classification weights of a pre-trained Vision Transformer are replaced with text embeddings generated by the pre-trained text encoder. More recent models are trained with LLM supervision (LLMDet; Fu et al., 2025), generate object descriptions (Yao et al., 2024a), or unify detection and grounding in a single sequence-to-sequence model (Xiao et al., 2024). Alternatively, a two-stage pipeline scores the boxes of a detector with an image–text encoder such as SigLIP 2 (Tschannen et al., 2025).

OVD Evaluation. The standard evaluation metric for object detection is the mean of the per-class average precision (mAP; Everingham et al., 2010). As shown by Dave et al. (2021), standard AP is sensitive to changes in cross-category ranking. Furthermore, Yao et al. (2024b) highlight the problem of inflated AP and propose to suppress it using the NMS-AP metric, which unifies multiple predictions of the same box and assigns the most confident label to that box. Relying on the maximum-logit confidence, this method is also prone to misrepresenting the correct ranking of relevant boxes and can inaccurately represent the robustness and stability of the model in predicting the correct class, as it merely considers the maximum-logit token from the query. In contrast, our approach ranks the predicted boxes with respect to all tokens in the query, which is crucial for multi-label scenarios.

Activity Localization Datasets. Activity localization involves analyzing the activities in a sequence of images (Bagautdinov et al., 2017; Barekatain et al., 2017; Ehsanpour et al., 2022; Zhou et al., 2022a). A seminal study by Choi et al. (2009) classifies the collective activities of pedestrians in real-world videos. Recent work recognizes interactions from skeleton sequences (Zhou et al., 2022a) or jointly detects people, their actions, and interacting groups in videos with a transformer (Wang et al., 2024). HAKE (Li et al., 2023b) maps pixels to body-part-level activity primitives and infers human–object interactions from them with logic rules. Most existing models are tested on video datasets, including Volleyball (Ibrahim et al., 2016), AVA-Interaction (Wang et al., 2024), NTU RGB+D (Shahroudy et al., 2016; Liu et al., 2019), SBU-Kinect-Interaction (Yun et al., 2012), and MatchNMingle (Cabrera-Quiros et al., 2018), while image-based benchmarks such as HICO-DET (Chao et al., 2018), V-COCO (Gupta & Malik, 2015), and SWIG-HOI (Wang et al., 2021) focus on human–object interactions. Ehsanpour et al. (2022) introduced JRDB-Act, a video dataset for group-based social activities in university campus scenes, and recent group benchmarks are likewise video-based (Jahangard et al., 2024; Kim et al., 2024b). In contrast, ELSA focuses on the more challenging task of localizing social activities of individuals and groups in images, where models must infer activities from a snapshot without the temporal cues available in videos.

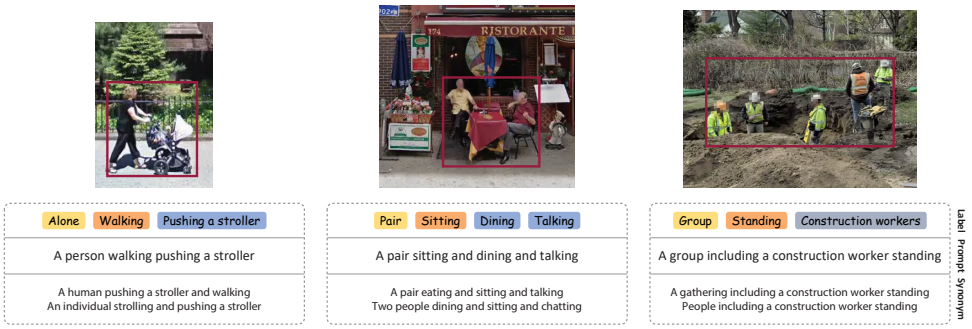


Figure 2: Examples of individual annotations extracted from larger images in ELSA. Each bounding box has a set of labels, a base natural language prompt, and synonyms, two of which are shown.

3 Evaluating Open-Vocabulary Localization

3.1 ELSA Construction Process

Image Resources. We selected New York City as our site of interest due to its vibrant streets and public spaces. We compiled street-level images from three different sources: Microsoft Bing Streetside (Kopf et al., 2010), Google Street View (Google Maps Platform, 2026; Anguelov et al., 2010), and Apple Look Around (Apple Inc., 2019). The Bing imagery provides time-stamps, making it possible to choose days and times with a higher probability of encountering pedestrians on the streets.

Target Labels. We draw on the literature on active design and urban vibrancy (see Section 2) to select our primary individual labels. ELSA exhibits non-disjoint label spaces, where multiple concurrent labels can be applied to the same object in a multi-labeling scheme that encompasses the nuances of human behavior and context. Labels are grouped into four main categories: 1) Condition: defines the social configuration of the subjects as *alone*, *pair*, or *group*. These labels are disjoint and denote mutually exclusive social settings, establishing the primary context for potential interactions, such as solo activities, limited interactions, or group dynamics; 2) State: captures the physical disposition or activity mode of the subjects, such as *walking* or *sitting*. While disjoint for individuals, these labels can co-occur in pair or group scenarios, indicating stationary engagement (*standing*, *sitting*) or transient interactions (*walking*, *biking*); 3) Activity: reflects specific behaviors or activities, such as *dining* or *talking*; 4) Other: any other information that can be of interest in urban activity analysis such as the usage of mobility aids or the presence of construction workers. We report additional information about the label categories in Appendix A.1.

Annotation Process. We customized the open source Label Studio tool (Tkachenko et al., 2020–2022) for annotation and integrated YOLOv8 (Jocher et al., 2023) for pre-detecting the initial objects. A team of seven trained annotators manually corrected the initial boxes and annotated the label combinations according to predefined guidelines. An urban planning expert then reviewed all annotations to ensure accuracy and consistency. Examples of ELSA's annotations are depicted in Figure 2. Additional examples are included in Appendix A.2.

Annotation Cleaning. Following the initial annotations, we applied sanity checks to ensure label consistency. For example, a single person cannot be labeled as both *sitting* and *walking*. The complete list of sanity rules is provided in Appendix A.4. Annotations failing these checks were re-evaluated and corrected, and this process was repeated until all boxes met the defined criteria.

Dataset Statistics. ELSA consists of 1,500 images with 6,384 annotated bounding boxes for social and individual activities, encompassing 26 distinct labels. Each box averages 2.44 labels, with 'alone' and 'walking' being the most frequent. This results in 265 unique combinations of human activities in the dataset. Figure 3 displays the distribution of individual labels and their combinations in ELSA.

Prompt Formation. Unlike physical objects, activities and human–human or human–object interactions frequently cannot be characterized by a single word or label. To investigate this, we conducted

a series of tests on various models, examining their responses to prompts with verbs like "walking," "talking," or "standing," and phrases like "walking alone" or "talking in groups." As expected, the results were often inaccurate or non-existent. The models require more detailed natural language descriptions to detect these activities correctly, such as "an individual sitting on a bench."

To address this need, we augmented ELSA with a template-based generator that produces precise, naturally phrased sentences for each label combination and their near-synonyms. This ensures that the models receive comprehensive descriptions, significantly improving their detection accuracy.

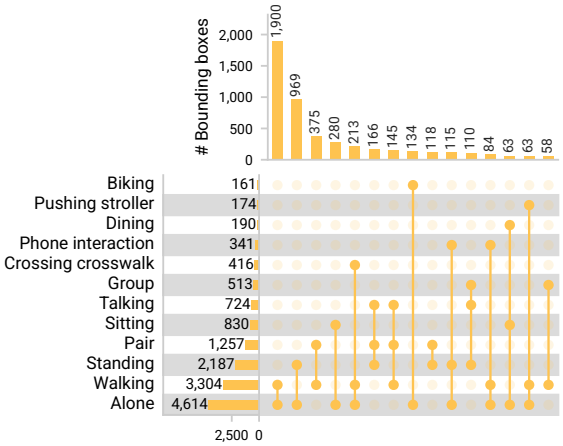


Figure 3: Overview of label distribution and combinations in ELSA, showing the 15 most frequent label combinations. Connected dots represent a combination, with its frequency plotted above. For example, "walking alone" appears 1,900 times.

3.2 Evaluation Framework

Evaluating open-vocabulary detection (OVD) models on complex human activities and interactions presents unique challenges. Unlike traditional object detection tasks, OVD models must interpret rich natural language queries and accurately localize corresponding instances in images. In this section, we introduce a robust evaluation methodology designed to assess both the semantic understanding and localization accuracy of OVD models in multi-label scenarios. An effective OVD model should excel in these two key aspects: 1) **Semantic Understanding**: Accurately interpret detailed query phrases to detect the correct targets, 2) **Localization Accuracy**: Precisely localize target objects or interactions in images by effectively bridging natural language and visual features.

3.2.1 Context-aware Confidence Score

Unlike single-object detection, multi-label human activity and interaction detection presents additional challenges for identifying multiple overlapping targets, activities, and interactions within the same scene. In particular, boxes must reflect not only the presence of the targets but also their attributes such as state and condition with higher confidence. In open-vocabulary detection, specifically, phrase grounding models, each predicted box is typically associated with a confidence score and an array of logits. These logits quantify the model's confidence in the relationship between the visual features within the box and specific tokens. Often, the confidence score of a box is determined by the highest logit value, i.e., Max-Logit, among all tokens (Liu et al., 2024). While effective in some contexts, the Max-Logit approach can bias the confidence towards prevalent object classes (e.g., "person"), potentially overlooking nuanced attributes critical in multi-label scenarios. This is consistent with the observation that language-based detectors rely mostly on the object name and largely ignore the rest of the description (Li et al., 2023a).

To address the limitations inherent in these methods—such as biases towards prevalent classes or insufficient emphasis on nuanced attributes—we introduce a *Context-Aware Confidence Score* that aggregates the logits of selected relevant tokens to compute a more representative confidence score. Specifically, we propose the **Normalized Log-Sum-Exp (N-LSE)** score. Let a prompt consist of M labels and let s_l be the mean confidence of the tokens of label l (a label such as "pushing a stroller" spans several tokens). The score is defined as

$$\text{N-LSE}_\beta(\mathbf{s}) = \frac{1}{\beta} \log \left(\frac{1}{M} \sum_{l=1}^{M} e^{\beta s_l} \right), \tag{1}$$

a soft maximum over the labels. Since we use prompts of different lengths, averaging within a label keeps labels with more tokens from outweighing those with fewer tokens. For $\beta \to 0$, it yields the mean of label confidences, for $\beta \to \infty$, the largest one, similar to Max-Logit over the label confidences. We use $\beta = 1$ throughout; Appendix A.9 analyzes the temperature.

Our scoring method involves mapping the tokens to boxes and comparing the native scoring of each model like Max-Logit in Grounding DINO and LLMDet with N-LSE methods. By capturing the combined evidence from pertinent tokens, our approach mitigates the bias towards any single prevalent token (Liu et al., 2024) or the limitations of a "no object" probability (Kamath et al., 2021). This not only decreases false positives but also retrieves under-ranked boxes.

Following COCO (Lin et al., 2014), we evaluate at a fixed maximum number of detections rather than pruning using a threshold, retaining the same number of boxes for every model and score. As in Dave et al. (2021), this budget covers the whole dataset rather than each image: we keep the $K = 2{,}041{,}500$ highest-scoring boxes over all images and prompts, i.e., one box per image–prompt pair on average. OWL-ViT is the only model that returns fewer boxes and retains all of them. Fixing the count makes scores of different scale directly comparable. Halving or doubling the budget changes at most one of the twelve comparisons reported below (Appendix A.8).

To manage overlapping predictions, for each ground truth (GT) bounding box, we identify the predicted bounding box with the highest Intersection over Union (IoU) relative to that GT box—designated as the *anchor*. Other predicted boxes with an IoU greater than 0.85 relative to the anchor are grouped together, forming a cluster of candidates representing the same object or interaction. This grouping allows us to consider multiple high-confidence predictions that may correspond to the same GT instance. A predicted box matches a GT box if their IoU is at least `iou_thr` and every label of its prompt is among the GT box's labels (Algorithm 1). As in COCO, every AP we report is the mean over several IoU thresholds, here $\texttt{iou_thr} \in \{0.75, 0.80, 0.85, 0.90\}$. Since every unique label combination is queried with multiple synonymous prompts, a GT box may be matched once per prompt: correct boxes of different synonymous prompts on the same person each count as true positives. The number of positives of a label combination is thus its number of GT boxes times its number of prompts, so that AP rewards consistency across phrasings instead of a single lucky phrasing.

3.2.2 Dynamic Box Aggregation (DBA)

A significant challenge in evaluating OVD models is handling multiple predictions for the same object arising from different prompts or overlapping boxes. Yao et al. (2024b) proposed combining "class-ignored non-maximum suppression" (C-NMS) with average precision computation, which, for any set of overlapping prediction boxes, selects the box with the highest confidence, suppresses the rest, and computes AP over the subset of the prediction boxes instead.

While more effective than the traditional AP, this approach has notable drawbacks: 1) It does not expose the model's susceptibility to making mutually exclusive predictions with close confidence levels, and 2) It may incorrectly suppress true positives with confidence levels close to the highest prediction as false positives, as also raised by the original paper (Yao et al., 2024b).

To address the limitations of existing methods like NMS-AP, we introduce the *Dynamic Box Aggregation* (DBA; Algorithm 1), which considers both confidence and semantic coherence by grouping the disjoint predictions for the same object among overlapping boxes. Unlike methods that rely solely on C-NMS, DBA retains boxes with confidence scores within a specified threshold (`score_thr`) of the maximum, suppressing the rest. We fix the threshold for confidence retention in our DBA at score_thr $= 0.2$ for all models; an extensive hyperparameter sweep across a range from 0.05 to 0.50, selecting the value that maximizes the precision–recall balance (DBA-F_1), shows no interior optimum. The comparison with NMS-AP is the same at every margin. We therefore keep 0.2.

An important component of DBA is its ability to account for mutually exclusive predictions with similar confidence scores. It exposes the model's vulnerability in understanding the target by penalizing cases where multiple inconsistent predictions are made for the same bounding box: if the retained boxes predict more than one condition for a box, more than one state for a single person, or more than two states for a pair, all of them are counted as false positives. For example, DBA would identify and penalize predictions of mutually exclusive classes such as a person being detected as "sitting," "walking," and "running" with close confidence, as in Figure 6 (Appendix A.7). This approach enhances evaluation accuracy by highlighting such failures in semantic understanding, which are otherwise overlooked by traditional C-NMS methods, ensuring that only predictions with sufficiently high confidence are considered, reducing false positives while maintaining high recall.

Algorithm 1: Dynamic Box Aggregation (DBA)

Input: Grouped bboxes O, ground truth set GT, IoU threshold iou_thr, score threshold score_thr
Output: TP, FP
Initialize empty dictionary data structures for TP, FP
for each group $o \in O$ **do**
    $T = \max(\text{Scores}(o)) - \text{score_thr}$
    $R = \{b_i \in o \mid \text{Score}(b_i) \geq T\}$
    if predicted labels in R are mutually exclusive (in condition or state) **then**
        Add all $b_i \in R$ to the FP
    else
        for each bounding box b_i in R **do**
            if $\text{IoU}(b_i, g) \geq$ iou_thr and $\text{label}(b_i) \subseteq \text{label}(g)$ for any $g \in GT$ **then**
                Add b_i to the TP
            else
                Add b_i to the FP
return TP, FP

Algorithm 2: DBA Integrated Evaluation

Input: Non-overlapping bounding boxes N, GT, TP and FP from DBA, IoU threshold iou_thr
Output: AP
Initialize empty dictionary data structure for FN
Initialize empty dictionary data structure for matched GT per prompt $matchedGT$
for each $n \in N$ **do**
    if $\text{IoU}(n, g) \geq$ iou_thr and $\text{label}(n) \subseteq \text{label}(g)$ for any $g \in GT$ not yet in $matchedGT[\text{prompt}(n)]$ **then**
        Add n to TP, add g to $matchedGT[\text{prompt}(n)]$
    else
        Add n to FP
for each prompt p and each $g \in GT$ not in $matchedGT[p]$ **do**
    Add (g, p) to FN
Compute precision and recall at each threshold
Compute AP
return AP

Table 1: Baseline mAP (without C-NMS or DBA), precision and recall at the detection budget under our N-LSE score and the maximum confidence among all tokens (Max), at four levels: global, CS (Condition + State), CSA (CS + Activity), and CSO (CS + Other); Swin-T without (1) and with (2) finetuning on COCO. Chance and Ceil.: global mAP of the same retained boxes in random order and with all correct boxes first. All scores in %; best mAP per model and level in yellow.

Model	Variation	Ranking	Global	Chance	Ceil.	CS	CSA	CSO	Prec.	Recall
Gdino	Swin-B	N-LSE	0.240	0.131	14.65	1.20	0.128	0.143	0.49	31.8
		Max	0.230	0.135	13.83	1.01	0.142	0.134	0.31	20.1
	Swin-T (1)	N-LSE	0.103	0.076	6.42	0.629	0.037	0.086	0.31	20.4
		Max	0.096	0.057	3.84	0.453	0.058	0.039	0.13	8.4
	Swin-T (2)	N-LSE	0.104	0.068	5.35	0.628	0.042	0.054	0.31	19.9
		Max	0.052	0.039	2.15	0.258	0.031	0.012	0.07	4.7
LLMDet	Swin-B	N-LSE	0.209	0.114	16.16	0.833	0.141	0.127	0.46	30.0
		Max	0.161	0.108	14.35	0.571	0.118	0.115	0.26	16.8

DBA handles overlapping predictions, while non-overlapping predictions are processed separately and later combined with DBA results for final metric computation, as shown in Algorithm 2.

In summary, ELSA provides a challenging and comprehensive benchmark for evaluating the detection of social interactions and individual activities in complex urban scenes. By addressing both visual data and prompt-level challenges, and introducing novel evaluation methods such as context-aware re-ranking and DBA, our benchmark facilitates a deeper understanding of model performance in open-vocabulary detection tasks. In the following section, we present experimental results demonstrating the effectiveness of our approach.

4 RESULTS

In this section, we first compare our N-LSE re-ranking with the Max-Logit scoring of prior work (Liu et al., 2024; Cheng et al., 2024; Li et al., 2022), showing how N-LSE better maps logits to box

confidences. We then show DBA's benefits across all benchmark levels on eight backbone variants of five detectors, and finally discuss each model's zero-shot performance on ELSA.

4.1 Context-aware Confidence Score

We investigate the benefit of our N-LSE re-ranking on Grounding DINO and LLMDet, since, unlike the other OVD models, they expose a confidence score for each prompt token. For each image in ELSA, we run all 1,361 prompts and keep all 900 boxes that each model returns per prompt, yielding roughly 1.8B boxes per model. Our detection budget finally retains roughly 0.11% of them.

Table 1 compares the mAP, precision, and recall obtained when box confidences are computed with our N-LSE versus the commonly used Max-Logit. N-LSE ranks the boxes better, outperforming Max-Logit in 14 of the 16 mAP cases, including the global level of every variant and all four levels of LLMDet. Within the same budget, it retains 1.6, 2.4, 4.2, and 1.8 times as many true positives as Max-Logit for Swin-B, Swin-T (1), Swin-T (2), and LLMDet, respectively, raising both precision and recall. By construction, a box scored with Max-Logit can rank high because of a single token, typically the condition noun (e.g., "person" or "pair"), regardless of the state and activity words, whereas N-LSE fully accounts for the evidence of all labels, so a single matching token no longer decides the score.

Table 1 also reports two references on the same retained boxes: their mAP in random order (*chance*) and with all correct boxes first (*ceiling*). Both stay low; even the ceiling is at most 16.2%, as most targets have no matching retained box, which shows how hard ELSA is. Yet both scores remain above chance for every variant, and N-LSE reaches a higher mAP and ceiling in all four.

These results show that N-LSE yields a more faithful ranking than Max-Logit, whose scores are often inflated and fail to reflect the model's true confidence in multi-label scenarios.

4.2 Analysis of DBA

Table 2 compares NMS-AP and DBA-AP for the eight variants of the five OVD models, and, in its last row, for a two-stage baseline discussed below. For consistency with the other OVD models, we rank the boxes of Grounding DINO and LLMDet by their native Max-Logit score, also making the analysis of DBA independent of our scoring; Appendix A.11 reports similar results with N-LSE. DBA-AP is generally the stricter score: NMS evaluates only the most confident box in each group of overlapping predictions, whereas DBA retains every box scored within `score_thr` of that maximum, and, if these boxes assign contradictory conditions or states to the same target, counts all of them as false positives. Accordingly, DBA-AP is lower in most cases, including every level for Swin-B, the strongest variant. The exceptions are cases in which NMS-AP is zero or close to it, e.g., the CSO level for Swin-T (1) and the CSA level for OWL-ViT. These stem from a weakness of NMS-AP already noted by its authors (Yao et al., 2024b): for weaker models and harder cases, the most confident box is often not the correct one, and NMS discards the correct prediction even when it is scored just below. DBA keeps such predictions as long as they do not contradict each other, recovering 1.3 to 40 times as many true positives as NMS for every variant except OWL-ViT (e.g., 679 against 316 for Swin-B and 160 against 4 for MDETR). At the level of individual label combinations, DBA-AP is non-zero for 181 combinations on which NMS-AP is zero, whereas the reverse occurs for only 15. At the CS level, NMS-AP puts Swin-B seven times above LLMDet, but DBA-AP puts LLMDet ahead.

The right half of Table 2 breaks the results down by the condition in which people appear ("alone", "pair", or "group"), showing a similar pattern: DBA-AP is generally lower and exceeds NMS-AP only where the latter nearly vanishes. This is most visible for the "pair" condition, where NMS-AP is zero for both MDETR variants and OWL-ViT, whereas DBA-AP is not. The two-stage baseline rescores Grounding DINO Swin-B's person, pair, and group boxes with SigLIP 2. This lifts both NMS-AP and DBA-AP well above those of Swin-B: Grounding DINO finds many people but fails to match their descriptions. Unlike for the eight variants, its DBA-AP falls below NMS-AP only for the "pair" condition and at the CSO level. Yet even this baseline stays below 1%.

Taken together, these results highlight the reliability of DBA-AP, unlike NMS-AP, which overstates performance when a model gives contradictory labels to the same target with similar confidence, and understates it, down to zero, when the correct box is not the most confident one.

Table 2: Comparison of NMS-AP and DBA-AP at the global, CS (Condition + State), CSA (CS + Activity), and CSO (CS + Other) levels (left) and on the Alone, Pair, and Group conditions (right), using each model's native score (Max-Logit for Grounding DINO and LLMDet). All scores in %. DBA-AP is highlighted in red where lower than NMS-AP and in green where higher.

			Level				Condition		
Model	**Variation**	**Scoring**	**Global**	**CS**	**CSA**	**CSO**	**Alone**	**Pair**	**Group**
Gdino	Swin-B	DBA-AP	0.0023	0.0068	0.0015	0.0054	0.0032	0.0015	0.0024
		NMS-AP	0.0435	0.177	0.0291	0.0286	0.0449	0.0721	0.0097
	Swin-T (1)	DBA-AP	0.0007	0.0010	0.0007	0.0008	0.0015	0.0007	0
		NMS-AP	0.0049	0.0225	0.0032	0	0.0089	0.0055	0
	Swin-T (2)	DBA-AP	0	0.0002	0	0	0.0001	0	0
		NMS-AP	0	0	0	0	0	0	0
MDETR	EfficientNet	DBA-AP	0.0005	0.0014	0.0004	0.0001	0.0012	0.0001	0.0001
		NMS-AP	0.0096	0.0005	0.0123	0.0041	0.0288	0	0
	ResNet 101	DBA-AP	0.0009	0.0054	0.0004	0	0.0022	0.0005	0
		NMS-AP	0.0041	0.0001	0.0001	0.0481	0.0122	0	0.0001
OWL	ViT-B/32	DBA-AP	0.0003	0.0021	0.0001	0	0.0004	0.0005	0
		NMS-AP	0.0017	0.0155	0	0	0.0051	0	0
Detic	ResNet 50	DBA-AP	0.0001	0.0001	0.0001	0	0.0003	0	0
		NMS-AP	0.0011	0.0048	0.0008	0	0.0034	0	0
LLMDet	Swin-B	DBA-AP	0.0018	0.0105	0.0008	0.0006	0.0040	0.0003	0.0011
		NMS-AP	0.0037	0.0244	0.0006	0.0074	0.0098	0.0011	0.0002
Gdino + SigLIP 2	Swin-B	DBA-AP	0.448	0.865	0.443	0.014	0.80	0.02	0.55
		NMS-AP	0.448	0.586	0.401	0.953	0.74	0.35	0.24

4.3 Performance on ELSA

Finally, we discuss the general performance of models on our benchmark task, for both NMS-AP and DBA-AP. Tables 1 and 2 show that, overall, Grounding DINO with the Swin-B backbone is the strongest model, whereas Detic is the weakest. The effectiveness of most models tends to degrade with an increase in complexity, with CSA and CSO being the hardest levels to predict for most models, showing AP values close to zero.

The overall performance of all models on ELSA appears significantly lower compared to other common benchmarks where those models excel. This highlights the substantial gap in detection capabilities when it comes to recognizing social activities in complex urban scenarios, consistent with findings on other out-of-distribution domains (Robicheaux et al., 2025).

Table 2 further provides an in-depth analysis on conditions. We observe that all models struggle across all conditions, with consistently low performance. The presence of disjoint prompts in the evaluation adds further challenges, as models must process and detect individuals in complex scenarios while seeking to predict non-overlapping categories in the prompts. These results reinforce the challenges our benchmark poses for these models and highlight the need for improved detection methods.

5 Conclusion

This paper introduces ELSA, a novel dataset specifically curated for the detection of social activities from still images within urban environments. Employing a multi-labeling scheme, ELSA comprises 1,500 annotated images and more than 6,000 bounding boxes, annotated with 265 unique combinations of social activities. ELSA comes with a new re-ranking approach, namely N-LSE, specifically designed for multi-label scenarios and OVD models, for which the effect of each token in a query is accounted for in the final logit score calculations of the predicted bounding boxes. We show that N-LSE, in contrast to the Max-Logit approach in prior work, reaches a higher AP in 14 out of 16 cases with 1.6 to 4.2 times more true positives. Together with DBA-AP, it exposes the models' tendencies to predict contradictory labels with high confidence. With this work, we enable a more profound evaluation of OVD models in challenging real-world settings and encourage further research on applying such models in applications such as assessing human activity in still images.

AI USE STATEMENT

In this work, we used generative AI tools (Claude) to give feedback on our evaluation methodology and experiments, and to help in part with implementing the methods and baselines, including finding bugs and refactoring the code. We did not use generative AI tools to generate, annotate or clean the dataset: we chose the wording of the prompts in ELSA and their synonyms manually and assembled the prompts with our own template-based code. We also did not use them to devise the label taxonomy or the core ideas of N-LSE and DBA, or to formulate hypotheses or mathematical claims. Proofs, translation, and qualitative or thematic data analysis are not applicable to this work.

Additionally, we used generative AI tools to draft and revise passages of the paper, mainly in the description of related work and of the evaluation protocol, and to improve its readability.

We have reviewed all AI-assisted work. We revised all AI-drafted or AI-edited text and checked it against our results and the cited papers. We verified every methodological comment before acting on it, and inspected every AI-suggested code change before adopting it. All design decisions are our own. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

All images in ELSA come from publicly available street-level imagery services: Microsoft Bing Streetside, Google Street View and Apple Look Around. The providers blur faces and license plates before publishing the imagery, to prevent people from being identified. Our annotations add no identity information: each bounding box only carries the labels described in Section 3. ELSA is intended solely for research on social interactions in public spaces and should not be used for military or surveillance applications, or to identify or track individuals.

Due to the providers' terms of use, we do not redistribute the images. For each image, ELSA provides the source service, the panorama ID, the geographic coordinates (latitude, longitude) and the heading, together with the labeled bounding boxes. Users who have the necessary permissions from the respective provider can download the panoramas with the third-party library `streetlevel`[2]; our scripts then reconstruct the exact images on which the boxes were annotated. Since the panoramas are retrieved directly from the providers, any blurring or removal that the providers apply later, e.g., at the request of a person depicted, also applies to ELSA.

REPRODUCIBILITY STATEMENT

The dataset statistics, prompt formation and evaluation protocol are described in Section 3; the full label list, the annotation rules and the implementation details, including the model checkpoints and the runtime of prediction and evaluation, are given in Appendix A.3, A.4 and A.12. The annotations, metadata and scripts are publicly available at the link given on the first page. The full set of images from all datasets used in this work is publicly available online and has been verified for completeness.

[2] `https://github.com/sk-zk/streetlevel`

REFERENCES

Dragomir Anguelov, Carole Dulong, Daniel Filip, Christian Frueh, Stéphane Lafon, Richard Lyon, Abhijit Ogale, Luc Vincent, and Josh Weaver. Google street view: Capturing the world at street level. *Computer*, 43(6):32–38, 2010.

Apple Inc. Apple previews iOS 13. Apple Newsroom, June 2019. URL `https://www.apple.com/newsroom/2019/06/apple-previews-ios-13/`. Introduces the Look Around street-level imagery of Apple Maps.

Timur Bagautdinov, Alexandre Alahi, Francois Fleuret, Pascal Fua, and Silvio Savarese. Social scene understanding: End-to-end multi-person action localization and collective activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Mohammadamin Barekatain, Miquel Martí, Hsueh-Fu Shih, Samuel Murray, Kotaro Nakayama, Yutaka Matsuo, and Helmut Prendinger. Okutama-action: An aerial view video dataset for concurrent human action detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 28–35, 2017.

Maria A Bravo, Sudhanshu Mittal, Simon Ging, and Thomas Brox. Open-vocabulary attribute detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7041–7050, 2023.

Laura Cabrera-Quiros, Andrew Demetriou, Ekin Gedik, Leander van der Meij, and Hayley Hung. The matchnmingle dataset: a novel multi-sensor resource for the analysis of social interactions and group dynamics in-the-wild during free-standing conversations and speed dates. *IEEE Transactions on Affective Computing*, 12(1):113–130, 2018.

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)*, pp. 381–389. IEEE, 2018.

Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. YOLO-World: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16901–16911, 2024.

Wongun Choi, Khuram Shahid, and Silvio Savarese. What are they doing?: Collective activity classification using spatio-temporal relationship among people. In *2009 IEEE 12th international conference on computer vision workshops, ICCV Workshops*, pp. 1282–1289. IEEE, 2009.

Wongun Choi, Yu-Wei Chao, Caroline Pantofaru, and Silvio Savarese. Discovering groups of people in images. In *European Conference on Computer Vision (ECCV)*, volume 8692 of *Lecture Notes in Computer Science*, pp. 417–433. Springer, 2014.

Achal Dave, Piotr Dollár, Deva Ramanan, Alexander Kirillov, and Ross Girshick. Evaluating large-vocabulary object detectors: The devil is in the details. *arXiv preprint arXiv:2102.01066*, 2021.

Mahsa Ehsanpour, Fatemeh Saleh, Silvio Savarese, Ian Reid, and Hamid Rezatofighi. Jrdb-act: A large-scale dataset for spatio-temporal action, social group and activity detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20983–20992, 2022.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010.

Shenghao Fu, Qize Yang, Qijie Mo, Junkai Yan, Xihan Wei, Jingke Meng, Xiaohua Xie, and Wei-Shi Zheng. LLMDet: Learning strong open-vocabulary object detectors under the supervision of large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 14987–14997, 2025.

Jan Gehl. People on foot. *Architecture*, 20:429–446, 1968.

Jan Gehl and Birgitte Svarre. *How to study public life*, volume 2. Springer, 2013.

Google Maps Platform. Street View Static API, 2026. URL https://developers.google.com/maps/documentation/streetview/overview. Accessed: 2026-09-26.

Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015.

Mostafa S Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A hierarchical deep temporal model for group activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1971–1980, 2016.

Jane Jacobs. *The Death and Life of Great American Cities*. Random House, New York, 1961.

Simindokht Jahangard, Zhixi Cai, Shiki Wen, and Hamid Rezatofighi. JRDB-Social: A multifaceted robotic dataset for understanding of context and dynamics of human interactions within social groups. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22087–22097, 2024.

Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLOv8, 2023. URL https://github.com/ultralytics/ultralytics. Version 8.0.0.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1780–1790, 2021.

Dahun Kim, Anelia Angelova, and Weicheng Kuo. Region-centric image-language pretraining for open-vocabulary detection. In *European Conference on Computer Vision (ECCV)*, volume 15121 of *Lecture Notes in Computer Science*, pp. 162–179. Springer, 2024a.

Dongkeun Kim, Youngkil Song, Minsu Cho, and Suha Kwak. Towards more practical group activity detection: A new benchmark and model. In *European Conference on Computer Vision (ECCV)*, 2024b.

Johannes Kopf, Billy Chen, Richard Szeliski, and Michael Cohen. Street slide: browsing street level imagery. *ACM Transactions on Graphics (TOG)*, 29(4):1–8, 2010.

Tian Lan, Leonid Sigal, and Greg Mori. Social roles in hierarchical models for human activity recognition. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1354–1361. IEEE, 2012.

Ting Lei, Shaofeng Yin, and Yang Liu. Exploring the potential of large foundation models for open-vocabulary HOI detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16657–16667, 2024.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10965–10975, 2022.

Liunian Harold Li, Zi-Yi Dou, Nanyun Peng, and Kai-Wei Chang. DesCo: Learning object recognition with rich language descriptions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 37511–37526, 2023a.

Yong-Lu Li, Xinpeng Liu, Xiaoqian Wu, Yizhuo Li, Zuoyu Qiu, Liang Xu, Yue Xu, Hao-Shu Fang, and Cewu Lu. HAKE: A knowledge engine foundation for human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2684–2701, 2019.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision (ECCV)*, 2024.

Vikas Mehta and Jennifer K Bosson. Revisiting lively streets: Social interactions in public space. *Journal of Planning Education and Research*, 41(2):160–172, 2021.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection. In *European Conference on Computer Vision (ECCV)*. Springer, 2022.

Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. Grounded situation recognition. In *European Conference on Computer Vision (ECCV)*, volume 12349 of *Lecture Notes in Computer Science*, pp. 314–332. Springer, 2020.

Peter Robicheaux, Matvei Popov, Anish Madan, Isaac Robinson, Joseph Nelson, Deva Ramanan, and Neehar Peri. Roboflow100-VL: A multi-domain object detection benchmark for vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2025.

Samuel Schulter, Yumin Suh, Konstantinos M Dafnis, Zhixing Zhang, Shiyu Zhao, Dimitris Metaxas, et al. Omnilabel: A challenging benchmark for language-based object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11953–11962, 2023.

Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1010–1019, 2016.

Zehua Sun, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. Human action recognition from various data modalities: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3), 2023. doi: 10.1109/TPAMI.2022.3183112.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. Label Studio: Data labeling software, 2020–2022. URL `https://github.com/heartexlabs/label-studio`. Open source software available from https://github.com/heartexlabs/label-studio.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Suchen Wang, Kim-Hui Yap, Henghui Ding, Jiyan Wu, Junsong Yuan, and Yap-Peng Tan. Discovering human interactions with large-vocabulary objects via query and multi-scale detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13475–13484, 2021.

Xueyang Wang, Xiya Zhang, Yinheng Zhu, Yuchen Guo, Xiaoyun Yuan, Liuyu Xiang, Zerun Wang, Guiguang Ding, David Brady, Qionghai Dai, and Lu Fang. PANDA: A gigapixel-level human-centric video dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3268–3278, 2020.

Zhenhua Wang, Kaining Ying, Jiajun Meng, and Jifeng Ning. Human-to-human interaction detection. In *Neural Information Processing (ICONIP 2023)*, volume 14450 of *Lecture Notes in Computer Science*, pp. 120–132. Springer, 2024. doi: 10.1007/978-981-99-8070-3_10.

William H. Whyte. *The Social Life of Small Urban Spaces*. Conservation Foundation, Washington, D.C., 1980.

Jianzong Wu, Xiangtai Li, Shilin Xu, Haobo Yuan, Henghui Ding, Yibo Yang, Xia Li, Jiangning Zhang, Yunhai Tong, Xudong Jiang, Bernard Ghanem, and Dacheng Tao. Towards open vocabulary learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4818–4829, 2024.

Lewei Yao, Renjie Pi, Jianhua Han, Xiaodan Liang, Hang Xu, Wei Zhang, Zhenguo Li, and Dan Xu. DetCLIPv3: Towards versatile generative open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27391–27401, 2024a.

Yiyang Yao, Peng Liu, Tiancheng Zhao, Qianqian Zhang, Jiajia Liao, Chunxin Fang, Kyusong Lee, and Qing Wang. How to evaluate the generalization of detection? A benchmark for comprehensive open-vocabulary detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 6630–6638, 2024b. doi: 10.1609/aaai.v38i7.28485.

Kiwon Yun, Jean Honorio, Debaleena Chattopadhyay, Tamara L Berg, and Dimitris Samaras. Two-person interaction detection using body-pose features and multiple instance learning. In *2012 IEEE computer society conference on computer vision and pattern recognition workshops*, pp. 28–35. IEEE, 2012.

Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14393–14402, 2021.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In *International Conference on Learning Representations (ICLR)*, 2023.

Jin Zhou, Zhenhua Wang, Jiajun Meng, Sheng Liu, Jianhua Zhang, and Shengyong Chen. Human interaction recognition with skeletal attention and shift graph convolution. In *International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2022a. doi: 10.1109/IJCNN55064.2022.9892292.

Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, pp. 350–368. Springer, 2022b.

Chaoyang Zhu and Long Chen. A survey on open-vocabulary detection and segmentation: Past, present, and future. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 8954–8975, 2024. doi: 10.1109/TPAMI.2024.3413013.

A Appendix

In this appendix, we provide supplementary information about our work. We first give further details on the dataset: the label categories (Section A.1), our annotation strategies (Section A.2), the full list of labels (Section A.3), and the sanity rules used to clean the annotations (Section A.4). We then expand on the evaluation framework: how we select the tokens used by N-LSE, with an example comparing N-LSE and Max-Logit (Section A.5), and the key features of DBA (Section A.6). Section A.7 presents qualitative results. Section A.8 studies the effect of the detection budget. Section A.9 shows results with different values of temperature β of N-LSE. Section A.10 reports

the chance and ceiling of every model, and Section A.11 repeats the DBA analysis with the boxes of Grounding DINO and LLMDet ranked by N-LSE instead of Max-Logit. Finally, Section A.12 provides the implementation details.

A.1 Label categories

In the realm of social interaction recognition, the labels under the "Activity" category are instrumental in identifying engagement patterns and interaction types, distinguishing, for example, between conversational engagement and co-active behavior.

Activity labels are non-disjoint, capturing the complexity of human behavior, where multiple actions can co-occur, like *talking* while *pushing a stroller*.

We also have another category of labels, namely, "Other", which represents characteristics of the scene that do not fall under the previous categories and are still important for understanding the features of the urban area. For example, the label *construction workers* can indicate an area under active development.

A.2 Annotation Strategies

As shown in Figure 4, for activities that are described with another non-stationary object, e.g., *pushing a wheelchair* or *biking*, the annotated ground truth bounding box includes the object as well as the person performing the action (see Figure 4-a), whereas for actions without an object that is actively a part of the action, the annotated bounding box merely captures the person (see Figure 4-b *sitting*).

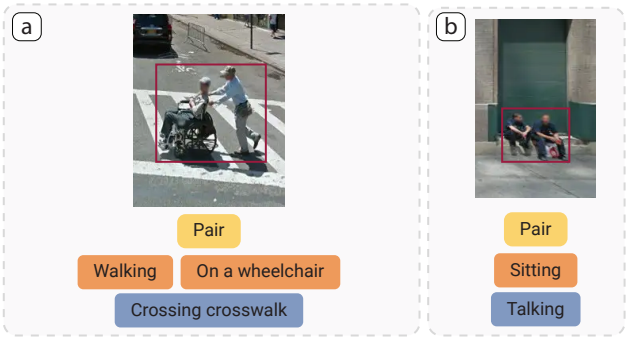


Figure 4: Example of rules of capture in annotation. a) Two people crossing a crosswalk and one sitting in a wheelchair. The wheelchair is captured in the annotation. b) Two people sitting and the stairs are not captured as an annotation.

A.3 Full list of labels

Table 3 reports the full list of labels used during the annotation process in ELSA. We omit some additional meta-labels, which supported the annotation process and the collection of statistics, such as "no people" and "model hint".

A.4 Sanity Rules for Annotation Cleaning

To make sure that all the annotated labels for bounding boxes are correct, we performed a sanity-check using a predefined set of sanity rules. In the following, we summarize the full set of rules we considered at this stage:

1. Each bounding box must have a condition label;
2. Each bounding box must have at least one state label;
3. Each bounding box can only have one condition label associated, e.g., "alone" and "group" cannot appear together;

Condition	State	Activity	Others
Alone	Sitting	Dining	Street vendors
Pair	Standing	Talking	Construction workers
Group	Walking	Shopping	Mobility aids
	Running	Taking photo	
	Biking	Talking on phone	
	On wheelchair	Phone interaction	
		Pushing stroller	
		Sport activities	
		Crossing crosswalk	
		Pushing shopping cart	
		Waiting in bus station	
		Public service/cleaning	
		Load/unload packages from car/truck	
		Pet interactions	

Table 3: Full list of labels in ELSA divided by category

4. If a bounding box is associated with the "alone" condition, then it can only have one state label associated, e.g., "alone walking running" is not allowed;
5. If a bounding box is associated with a "pair" condition, then it can only have at most two state labels associated, e.g., "couple walking sitting running" is not allowed;
6. If a bounding box is associated with the "shopping" activity, then state should include either one of "sitting" or "standing" labels;
7. If a bounding box is associated with "street vendors", then state should include either one of "sitting" or "standing" labels;
8. If a bounding box is associated with the "load/unload packages" activity, then state should include either one of "sitting" or "standing" labels;
9. If a bounding box is associated with the "waiting in bus station" activity, then state should include either one of "sitting" or "standing" labels.

A.5 Additional Results

Selecting Relevant Logits. Grounding DINO and LLMDet use the BERT model for tokenization. We keep the mapping between logits and tokens and their category of condition, state, activity. Using this mapping, we only keep the relevant tokens in our metric calculation. Figure 5 shows our metric being applied to relevant tokens, as well as the Max-Logit. In both prompts, one target (the red box) was predicted with the highest confidence. The ground truth for that target comprises the following labels: *C: Alone + S: Standing + A: Phone interaction*. In this example, we showcase how the same target is assigned two disjoint conditions, with high confidence. The same individual is returned as the highest-confidence prediction for the first prompt, "a group eating and sitting on a chair", with 49% confidence for "group" and 11% for "eating". For the second prompt, "a person with a dog walking alone", only the condition matches the ground truth, and the model assigns 50% confidence to "person". The model is thus almost equally confident in two mutually exclusive conditions for the same target, which reveals a poor understanding of the condition in this image.

A.6 DBA

Key features of DBA include:

- **Confidence Range Consideration**: Instead of selecting only the box with the highest confidence, DBA retains all boxes whose confidence scores are within a specified threshold (`score_thr`) of the maximum confidence in the group. This approach prevents the unnecessary suppression of true positives that have slightly lower confidence scores.



Figure 5: An example of the top five predictions of the Grounding DINO model for two distinct prompts on the same image is provided. Top tables present the model's confidence in the presence of the tokens within each box. The selected tokens used to compute the N-LSE metric are highlighted with boxes annotated by C: condition, S: state, A: activity. Bottom tables display the overall score for each color-coded box, comparing N-LSE on selected tokens (ours), and the maximum logit of all tokens.

- **Semantic Consistency Check**: DBA examines the predicted labels for disjointness in critical categories (e.g., condition or state). If overlapping boxes predict contradictory labels (e.g., *sitting* vs. *walking*), all of them are flagged as false positives.

A.7 Qualitative Results

As a prompt increases in level from CS to CSA or CSO, the likelihood that the prompt contains labels that the model has a low confidence on from its training increases, lowering the N-LSE score for the box. The outcome is that the most basic-level prompts (CS) are overrepresented among the boxes retained within the detection budget, and higher-level prompts are underrepresented. Conversely, when more conventionally determining the score by the maximum logit for the box, higher-level prompts have more logits and therefore tend to be better represented among the retained boxes.

When a prompt includes an object that is among the pre-trained vocabulary, the model can more easily detect and localize it. This is a case where contextual cueing leads to better predictions. For instance, when we query for "a group of people sitting", the model less frequently finds the correct target, but the prompt "a group of people sitting on a chair" can lead to a better prediction.

The most challenging part for the models was recognizing *state*. The confidence of the model in associating the area inside each box with the labels in the *state* category is very low across all images and queries.

To further analyze the model's understanding of people's states ("sitting," "standing," "walking," etc.), we prompt it using its native Max-Logit scoring.

Here, we used variations of our original prompt "a group of people sitting on a bench": "a group of people standing on a bench" and "a group of people running on a bench". These prompts do not match the state of the people in the image and are not in our prompt set.

In all three cases, one target was in common and had the highest confidence, as shown in Figure 6. When prompted with *people sitting on a bench*, the model returned one result with 44% confidence; however, the model assigned higher confidence to the same target with *people standing on a bench* with 52.98% confidence and 53.08% confidence in the box showing *people running on a bench*. The Max-Logit method results in false positive predictions with high confidence and undermines the actual context of the query by allowing the logit with the maximum confidence to represent the whole query.

Our findings also highlight the need for the incorporation of uncertainty estimation techniques during model fine-tuning and training to mitigate the risk of overconfident false predictions.



Figure 6: Highest-scoring box of Grounding DINO (Swin-T (1), Max-Logit) for three prompts that differ only in the states. The same group scores higher for the two wrong states (0.53) than for the correct one (0.44).

A.8 Detection Budget

Table 1 uses a budget of one box per image-prompt pair. At half and twice this budget (1,020,750 and 4,083,000 boxes respectively), N-LSE remains above Max-Logit in 10 and 9 of the 12 cells; the cells it loses are the CSA level of Swin-B and Swin-T (1) and, at twice the budget, the global mAP of Swin-B (0.263 against 0.268). The conclusions of Section 4 therefore do not depend on the budget.

A.9 N-LSE Temperature

Table 4: Global mAP of Table 1 for the per-label N-LSE at different temperatures β, at the same detection budget; $\beta \to \infty$ is Max-Logit over the label confidences. The last column counts the cells of Table 1 (all four levels, three models) in which N-LSE is above Max-Logit. All scores are in percentage. The best mAP of each model is highlighted in yellow.

β	**Swin-B**	**Swin-T (1)**	**Swin-T (2)**	**Cells above Max**
1	0.240	0.103	0.104	10
2	0.245	0.107	0.104	10
3	0.247	0.105	0.101	10
5	0.252	0.104	0.092	11
10	0.245	0.100	0.076	10
20	0.237	0.103	0.068	10
50	0.231	0.103	0.063	11
Max-Logit	0.230	0.096	0.052	–

Table 4 reports the global mAP of Table 1 for $\beta \in \{1, 2, 3, 5, 10, 20, 50\}$ at the same budget; $\beta \to \infty$ is Max-Logit. Raising β lifts the CSA level and lowers CS for every model, and N-LSE stays above Max-Logit in 10 of the 12 cells for every β (11 at $\beta = 5$ and 50). The mean over the four levels does not improve beyond $\beta = 2$ for any model, and a value chosen on one half of the images and scored on the other lies between 1 and 7 depending on the model; we therefore keep $\beta = 1$.

A.10 Comparison with Chance and Ceiling

Table 5 extends the chance and ceiling of Table 1 to every model, at the same budget and with the same matching. For each label combination, chance is the AP of its retained boxes in random order (mean over 10 shuffles), and ceiling is their AP with all correct boxes first, the best that any order of these boxes can reach. Every model stays above chance, and every ceiling is at most 16.2%, the highest being that of LLMDet under N-LSE. After Grounding DINO and LLMDet, Detic has the highest mAP, although its NMS-AP and DBA-AP are among the lowest.

Table 5: Global mAP (without C-NMS or DBA) of every model at the detection budget of Table 1, with the chance and ceiling of its retained boxes: their mAP in random order (mean over 10 shuffles) and with all correct boxes first. MDETR, OWL and Detic use their native score. All scores in %.

Model	Variation	Ranking	Global	Chance	Ceiling
Gdino	Swin-B	N-LSE	0.240	0.131	14.65
		Max	0.230	0.135	13.83
	Swin-T (1)	N-LSE	0.103	0.076	6.42
		Max	0.096	0.057	3.84
	Swin-T (2)	N-LSE	0.104	0.068	5.35
		Max	0.052	0.039	2.15
MDETR	EfficientNet	Native	0.049	0.025	3.11
	ResNet 101	Native	0.051	0.027	2.49
OWL	ViT-B/32	Native	0.038	0.022	1.74
Detic	ResNet 50	Native	0.074	0.029	3.58
LLMDet	Swin-B	N-LSE	0.209	0.114	16.16
		Max	0.161	0.108	14.35

A.11 DBA UNDER N-LSE

Table 2 ranks the bounding boxes of every model using the respective native score. Table 6 repeats the Grounding DINO and LLMDet rows with N-LSE. N-LSE recovers label combinations at the CSA level that NMS suppresses under Max-Logit (Swin-B: 47 combinations recovered against 4 lost, Swin-T (1): 44 against 2, Swin-T (2): 14 against 1; under Max-Logit 47/4, 33/3 and 1/0), and keeps Swin-T (2) evaluable, whose Max-Logit ranking places almost no correct box at the budget. At the same time it retains more competing boxes per person, so DBA exposes more contradictions at the CS level, where DBA-AP falls below NMS-AP for every variant.

Table 6: NMS-AP versus DBA-AP for the three Grounding DINO variants and LLMDet of Table 2 when the boxes are ranked with N-LSE instead of Max-Logit; the other models are unchanged. All scores are in percentage. DBA-AP is highlighted in red where it is lower than NMS-AP and in green where it is higher.

			Level				Condition		
Model	Variation	Scoring	Global	CS	CSA	CSO	Alone	Pair	Group
Gdino	Swin-B	DBA-AP	0.0040	0.0077	0.0029	0.0121	0.0084	0.0014	0.0024
		NMS-AP	0.0191	0.149	0.0002	0.0308	0.0298	0.0177	0.0093
	Swin-T (1)	DBA-AP	0.0045	0.0026	0.0053	0.0024	0.0085	0.0042	0.0004
		NMS-AP	0.0025	0.0229	0	0	0.0056	0.0018	0.0001
	Swin-T (2)	DBA-AP	0.0002	0.0015	0	0.0001	0.0005	0.0001	0
		NMS-AP	0.0049	0.0447	0	0	0.0139	0.0007	0
LLMDet	Swin-B	DBA-AP	0.0080	0.0254	0.0066	0.0034	0.0191	0.0014	0.0038
		NMS-AP	0.0102	0.0920	0.0001	0	0.0277	0.0008	0.0021

A.12 IMPLEMENTATION DETAILS

The generation of all the predictions with Grounding DINO, MDETR, Detic, and OWL takes up to twenty hours per Grounding DINO backbone, five hours per MDETR backbone, two hours for Detic, and eight hours for OWL on one B200 or RTX PRO 6000 GPU. The evaluation of one model takes between a few minutes and one hour on an Intel Xeon Gold 5220S with 256 GB of memory, depending on the number of retained boxes.

We used the Open Grounding DINO implementation[3]. Our inference was conducted using the configuration from the official repository with Swin-T backbone, pre-trained on O365, GoldG, and

[3] https://github.com/longzw1997/Open-GroundingDino

Cap4M datasets. For MDETR[4] and Detic[5] we used the official repository and checkpoints, whereas for OWL we employed the ViT Patch 32 version from the Hugging Face hub[6]. For LLMDet we used the Swin-B checkpoint from the Hugging Face hub[7] through the `transformers` library. The two-stage baseline scores Grounding DINO Swin-B proposals with SigLIP 2[8].

[4] https://github.com/ashkamath/mdetr
[5] https://github.com/facebookresearch/Detic
[6] https://huggingface.co/docs/transformers/model_doc/owlvit
[7] https://huggingface.co/iSEE-Laboratory/llmdet_base
[8] https://huggingface.co/google/siglip2-so400m-patch14-384