A Appendix

In this appendix we provide supplementary information about our work. The first Section addresses a small typo from the main submission, A.1 offers further details on the labels used in the study, examples of annotations are provided in A.2, while a comprehensive table listing all labels in the dataset can be found in A.3, additional results are presented in A.5, complemented by qualitative results in A.6. In A.7 we include some additional discussion and limitations of ELSA. Finally, in A.8 we provide information on the implementation details.

Erratum

Table 3 reports an updated version of the semantic stability scores. We wish to point a small typo from table in 4.3. The reported values are in percentage.

Method	*CS*	*CSA*	*All*
Grounding DINO (N-LSE)	64%	65%	64%
Grounding DINO (Max-Logit)	57%	56%	56%

Table 3: Fixed typo in Semantic Stability scores.

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 *kid* can indicate a family-friendly area.

A.2 Annotation examples

As shown in Figure 5, 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 5-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 5-b *sitting*).

A.3 Full list of labels

Table 4 reports the full list of labels used during the annotation process in ELSA. We omit some additional meta-label which supported the annotation process and the statistic collection such as "no people" and "model hint".

A.4 Sanity Rules for Annotation Cleaning

In order 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, unless it is a “pet”;
2. Each bounding box must have at least one state label, unless it is a “pet”;
3. Each bounding box can only have one condition label associated, e.g., “alone” and “group” cannot appear together;

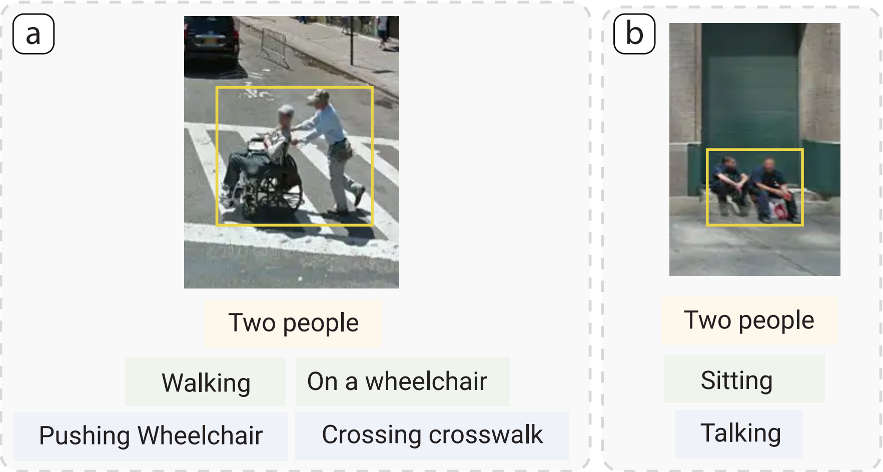


Figure 5: Example of rules of capture in annotation. a) two people sitting and the stairs are not captures as an annotation. b) two people crossing a crosswalk and one pushing a wheelchair. The wheelchair is captured in the annotation.

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 the “couple/two person” condition, then it can only have 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 the “street vendors” activity, 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 uses 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 6 shows our metric being applied to relevant tokens (selected.loglse), to all tokens (whole.loglse) as well as the Max-logit (whole.argmax). In all three prompts, one target (the red box) was predicted with highest confidence. The ground truth for that target comprises of the following labels: *C: Alone* + *S: Standing* + *A: Phone interaction*. In this example, we showcase how the same target, is assigned three disjoint conditions, with high confidence. The same individual is returned as the highest confidence prediction for first prompt: “a group eating and sitting on a chair", with 49% confidence in representing a “group", and 11% eating. While in the second prompt, “two people including a child walking", the model showed a high confidence in the red box showing two people (two:45% & people: 62%). The third prompt, has a matching condition only, “alone", which was returned by the model with 50% confidence. All predictions have pretty close confidence in the target representing disjoint conditions, highlighting the low understanding of the model in interpreting the condition in this image.

None of the people in this image match any of our queries. However, using the max log score, for the first prompt (Figure 6-top), all five boxes would pass the 0.3 threshold and be counted as likely

Condition	State	Activity	Others
Alone	Sitting	Dining	Pet
Couple	Standing	Snacking	Kid
Group	Walking	Talking	Police
	Running	Playing	Infant
	Biking	Shopping	Elderly
	On wheelchair	Hugging	Teenager
	Mobility aids	Taking photo	With bike
	Riding carriage	Talking on phone	
	Riding motorcycle	Taking Taxi	
		Pet interactions	
		Street vendors	
		Phone interaction	
		Waving to camera	
		Pushing stroller	
		Sport activities	
		Crossing crosswalk	
		Pushing wheelchair	
		Working with laptop	
		Construction workers	
		Pushing shopping cart	
		Waiting in bus station	
		At petrol/gas station	
		Public service/cleaning	
		Load/unload packages from car/truck	

Table 4: Full list of labels in ELSA divided by category

candidates. However, using our score (N-LSE), none of the boxes would be selected. Same goes for the other two prompts. There is a notable difference between the two scores, highlighting the important role of the taking relevant query terms into account.

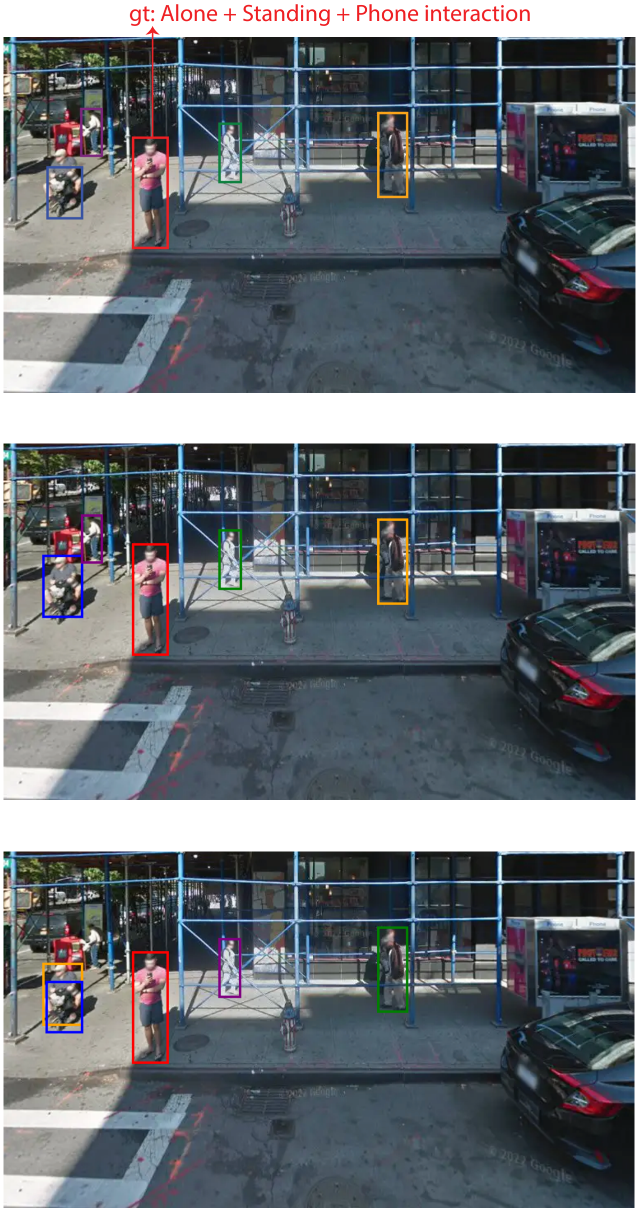
A.6 Qualitative results

As a prompt increases in level from *condition* to *condition, state, activity, and others,* the likelihood that the prompt contains labels which the model has low confidence trained on increases, lowering the computed score for the box. The outcome is that the most basic-level prompts are overrepresented among the predictions that pass score-based filters, and high-level prompts are extremely uncommon. *Condition* prompts accounted for less than 2% of the total prompts generated, but were 20% of the bounding boxes that passed initial thresholding on score. Conversely, when more conventionally determining the score by the maximum logit for the box, higher-level prompts have more logits and therefore always result in higher representation in the predictions that pass the threshold.

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 "group of people sitting" the model less frequently finds the correct target, but the prompt "groups 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 *state* group is very low across all images and all set of 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 and the 0.3 threshold. 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 have semantically valid *state*



selected.loglse	whole.loglse	whole.argmax
0.18	0.15	0.49
0.17	0.15	0.47
0.17	0.15	0.46
0.14	0.14	0.33
0.14	0.12	0.38

a	group	eating	and	sitting	on	a	chair
0.43	0.49	0.11	0.01	0.01	0.01	0.01	0.01
0.42	0.47	0.12	0.01	0.01	0.01	0.01	0.01
0.42	0.46	0.12	0.01	0.01	0.01	0.01	0.02
0.14	0.13	0.06	0.04	0.03	0.03	0.28	0.33
0.34	0.38	0.10	0.01	0.01	0.01	0.01	0.02
	C	A		S			

selected.loglse	whole.loglse	whole.argmax
0.27	0.24	0.62
0.25	0.22	0.55
0.25	0.23	0.52
0.20	0.18	0.40
0.17	0.15	0.37

two	people	including	a	child	walking
0.45	0.62	0.01	0.07	0.07	0.02
0.48	0.55	0.01	0.04	0.05	0.02
0.43	0.52	0.01	0.12	0.14	0.03
0.36	0.40	0.01	0.09	0.11	0.03
0.35	0.37	0.01	0.04	0.05	0.02
C				O	S

selected.loglse	whole.loglse	whole.argmax
0.16	0.17	0.50
0.14	0.14	0.40
0.12	0.12	0.35
0.10	0.10	0.22
0.10	0.11	0.31

a	person	with	a	dog	walking	alone
0.44	0.50	0.01	0.02	0.03	0.01	0.01
0.35	0.40	0.01	0.05	0.08	0.02	0.01
0.31	0.35	0.01	0.04	0.05	0.01	0.01
0.10	0.11	0.03	0.18	0.22	0.04	0.03
0.28	0.31	0.01	0.03	0.04	0.01	0.01
	C			A	S	

Figure 6: An example of the top five predictions of the model for three distinct prompts on the same image is provided. Each image is accompanied by two tables. The first table displays the overall score for each color-coded box, including three different metrics: N-LSE on selected tokens (ours), N-LSE on all tokens, and the maximum logit of all tokens. The second table presents 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 and O:others.

verbs and are not among our set of prompt list. In all three cases, one target was in common and had the highest confidence, as shown in Figure 7. When prompted *people sitting on a bench*, the model returned one result 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

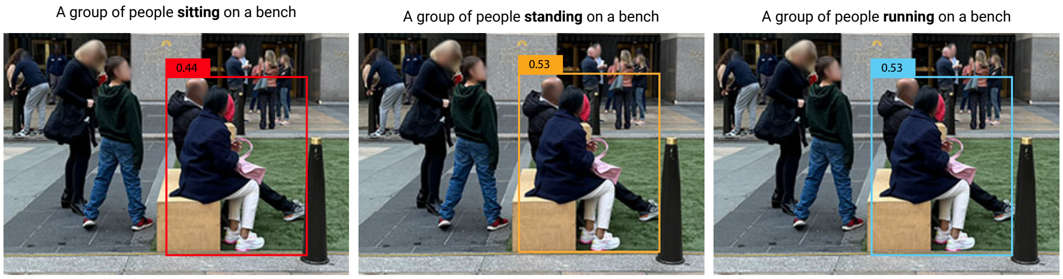

A group of people **sitting** on a bench
0.44
A group of people **standing** on a bench
0.53
A group of people **running** on a bench
0.53

Figure 7: Using the native Grounding DINO model with Swin-T backbone and Max-logit scoring to run variations of the same prompt with different states.

with very high confidence and undermine the actual context of the query by allowing the logit with the maximum confidence to represent the whole query.

A.7 Discussion

Existing OVDs exhibit a number of challenges. They often struggle with semantic consistency across diverse inputs, showing limited adaptability to novel or unseen categories, and can suffer from high computational costs during inference. Additionally, these models may demonstrate sensitivity to slight variations in input phrasing, leading to inconsistent performance. The calibration of their predictive confidence, especially in out-of-distribution scenarios, remains suboptimal, frequently resulting in overconfident predictions that do not accurately reflect their actual accuracy.

Following Desai et al. [9], we categorize target interactions into spatial relationship (people sit “on" something), spatial co-occurrence (pedestrians usually co-occur, a stroller should co-occur with a human), and mutual exclusion or disjoint (an individual cannot be sitting and walking at the same time). We incorporated the main non-stationary objects like bike, wheelchair, stroller, luggage, or shopping card in our annotation boxes.

Aside from the challenging nature of human activity and interaction detection, the lower quality of large-scale publicly available street-level images impact the detection results. On top of that, the anonymization process to blur faces creates artifacts that can impact the other people in the scene, making them difficult to be detected.

Although the metrics and evaluation protocols presented herein are applicable to any OVD model, this study was confined to a single model. Future work will encompass the inclusion of additional OVD models in our benchmark, enabling a comprehensive comparison of their understanding, stability, and localization accuracy in detecting social activities.

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.

A.8 Resource requirements implementation details?

The generation of all the predictions with Grounding DINO takes around eight hours on three H100 with 80GB of memory. The generation of the results on an Intel(R) Xeon(R) Platinum 8480CL takes around ten minutes.

We used the Open Grounding DINO implementation, which is also featured on the official repository of the paper [1]. Our inference was done using the configuration from the official repository with Swin-T backbone, pre-trained on O365, GoldG, and Cap4M dataset.

[1] https://github.com/urban-submissions/elsa