

MINGLE: VLMs for Semantically Complex Region Detection in Urban Scenes

Liu Liu^{1*}, Alexandra Schild^{2*}, Marco Cipriano², Fatimeh Al Ghannam¹, Freya Tan¹,
Gerard de Melo², Andres Sevtsuk¹

¹Massachusetts Institute of Technology

²Hasso Plattner Institute

lyons66@mit.edu, aleksandra.kudaeva@hpi.de

Abstract

Understanding group-level social interactions in public spaces is crucial for urban planning, informing the design of socially vibrant and inclusive environments. Detecting such interactions from images involves interpreting subtle visual cues such as relations, proximity, and co-movement—semantically complex signals that go beyond traditional object detection. To address this challenge, we introduce a social group region detection task, which requires inferring and spatially grounding visual regions defined by abstract interpersonal relations. We propose MINGLE (Modeling INterpersonal Group-Level Engagement), a modular three-stage pipeline that integrates: (1) off-the-shelf human detection and depth estimation, (2) VLM-based reasoning to classify pairwise social affiliation, and (3) a lightweight spatial aggregation algorithm to localize socially connected groups. To support this task and encourage future research, we present a new dataset of 100K urban street-view images annotated with bounding boxes and labels for both individuals and socially interacting groups. The annotations combine human-created labels and outputs from the MINGLE pipeline, ensuring semantic richness and broad coverage of real-world scenarios.

Code and Dataset — <https://github.com/lyons66/MINGLE>

Introduction

Object detection is a foundational task in computer vision, and state-of-the-art models, such as YOLO and Grounding DINO, excel at localizing discrete, well-defined objects, typically described by simple nouns or phrases (Liu et al. 2024; Ren et al. 2016; Jocher, Qiu, and Chaurasia 2023; Cheng et al. 2024). However, an increasing number of real-world applications require the detection of regions defined by semantically complex or abstract criteria, e.g., interactions of objects, social behaviors, or environmental context. Such tasks require a deeper understanding of context, relationships, and semantics than conventional object detection models are equipped to handle.

Vision-language models (VLMs), trained on massive image–text datasets, possess strong contextual reasoning and

semantic abstraction capabilities. Yet, they are generally incapable of accurately grounding rich visual descriptions to specific regions in an image. This limitation arises from two main factors: (1) sequential training objectives tailored to text, and (2) the lack of datasets and data creation methods for semantically complex region grounding during training.

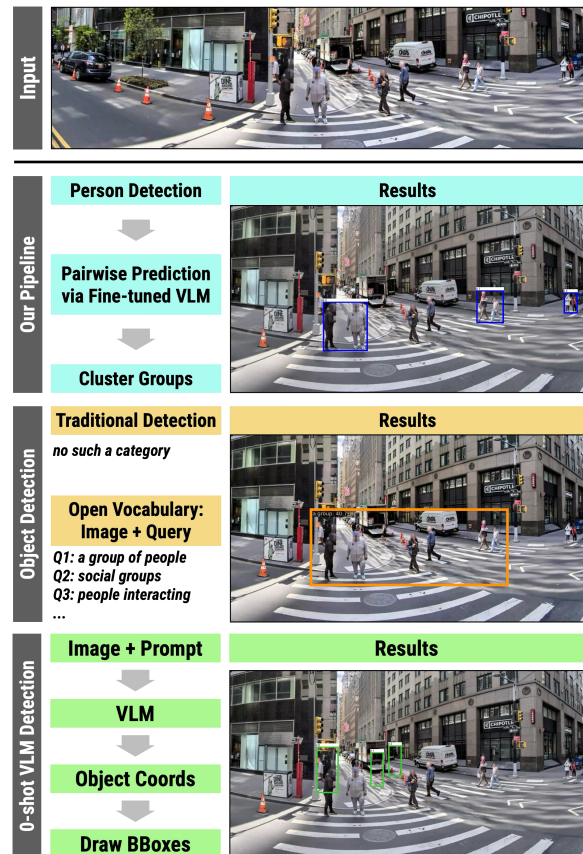


Figure 1: Comparison of our method with zero-shot object detection and standard VLM-based methods. Our pipeline enables localized detection of socially interacting human groups in complex urban scenes.

In this paper, we propose the task of social group region detection in urban scenes—detecting and localizing groups

*These authors contributed equally.

of individuals who are socially interacting. Unlike traditional object detection, which targets physical entities, this task requires identifying abstract, emergent visual patterns involving multiple people. These include diverse visual cues such as co-movement, proximity, body posture, and spatial formation—signals that define social groups rather than isolated entities.

From an urban design perspective, detecting social groups in public spaces has long been a key focus in pedestrian observational studies. Seminal works by Whyte (Whyte et al. 1980), Gehl (Gehl 1987), and Jacobs (Jacobs 1961, 1985) emphasize that observing how people gather, linger, and interact in cities is crucial to understanding urban vitality. These studies go far beyond simply counting individuals: Inferring social groups is central to evaluating how built environments support or inhibit human interaction. Such knowledge informs the creation of healthier, more inclusive, and socially vibrant urban environments.

Still, research in this area remains constrained by the lack of scalable, semantically rich annotations. Existing datasets typically either lack sufficient scale, fail to capture complex interaction cues, or require labor-intensive manual labeling that does not scale to a city- or country-level analysis.

To address this gap, we propose MINGLE, a modular three-stage pipeline for detecting socially interacting human groups in street-view imagery. MINGLE leverages both existing state-of-the-art object detection and depth estimation models as well as the semantic comprehension and reasoning capabilities of VLMs. First, we apply the best available off-the-shelf object detection and depth estimation models to detect all humans in the scene and create an image depth map. Second, using a novel human-annotated dataset, we fine-tune a vision-language model to classify whether a pair of individuals belong to the same social group. Finally, we apply an efficient pair-aggregation algorithm to identify social group regions.

Using MINGLE, we construct a new dataset of 100K images with annotations for both individual people and social groups, making it the first large-scale dataset focused on semantically complex social interaction regions in urban outdoor scenes, not limited to a particular interaction type. Our labels combine high-quality manual annotations and carefully validated pseudo-labels propagated through our pipeline.

This dataset serves as a valuable benchmark and resource for research in relational vision, social perception, and urban computing. It can be used to fine-tune VLMs for grounding relational concepts, to evaluate multi-object interaction models, or to explore high-level visual understanding.

Our contributions are threefold:

- We introduce a new vision-language task: social group region detection, which involves grounding relational and semantic concepts in visual space;
- We propose a scalable pipeline that combines object detection, depth estimation, and VLM reasoning to localize social interaction regions;
- We release a large-scale dataset of urban scenes with bounding boxes for both individual people and socially

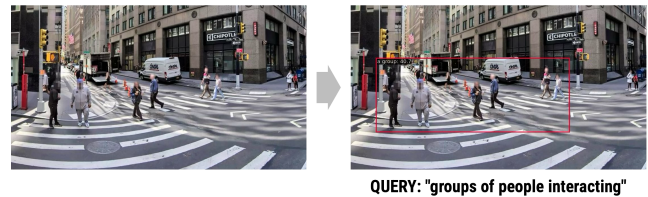


Figure 2: The result of OVD for social group detection.

interacting groups, enabling future work in this domain.

Beyond urban planning, social group region detection has broad applications in social dynamics analysis, autonomous navigation, AR/VR, robotics, as well as potentially harmful applications such as surveillance. We hope that this work encourages the vision community to explore more nuanced and relational forms of scene understanding—moving beyond object detection toward semantically complex scene comprehension and region detection.

Related Work

Social Interaction Research

Detecting social groups has long been a key focus in pedestrian observational studies. Starting from classic works by Whyte, Gehl, and Jacobs, more recent studies in urban planning revisit criteria of public space liveliness and emphasize that social interactions play a central role (Whyte et al. 1980; Jacobs 1961, 1985; Gehl 1987; Mehta and Bosson 2021; Qi, Mazumdar, and Vasconcelos 2024).

However, conducting large-scale analyses of group-level social behavior has been limited by the labor- and cost-intensive nature of manual fieldwork. Advances in AI and computer vision aim to address this challenge, offering tools to automate the detection and analysis of social dynamics at scale. Despite significant progress in human–human interaction detection, most computer vision approaches focus on relatively constrained scenarios. With existing street-view benchmarks being limited in scale (Hosseini et al. 2024), large-scale datasets such as JRDB (Ehsanpour et al. 2022; Jahangard et al. 2024) and AVA (Gu et al. 2018) primarily support action or gesture recognition in surveillance or indoor contexts. Others target pairwise or small-group social signal understanding, often framed as classification problems based on a closed set of body pose, gaze, or proximity cues (Wang et al. 2023; Wei et al. 2024). Although these datasets and models offer valuable insights, they do not generalize to open-world, urban street environments, where social formations are more ambiguous and embedded in complex spatial and visual contexts.

Open-Vocabulary Object Detection

Language-based object detection expands the fixed categories of traditional closed-set detection to free-form descriptions. Current commonly used open-vocabulary object detection and segmentation models, such as YOLO-World, Grounding-DINO, and Grounded SAM, have demonstrated reliable performance on standard object detection or segmentation benchmarks (Cheng et al. 2024; Liu et al. 2024;

Ren et al. 2024a). However, reliance on BERT-based language models limits their ability to comprehend semantically complex language structures. As a result, OVD models struggle to interpret abstract natural language instructions for precise visual grounding, particularly when handling complex instructions and visual scenarios. In particular, for social group region detection, they often fail to interpret and reason between complex inter-group relationships and the surrounding urban context (Figure 2).

VLMs for Localization, Object Detection, and Segmentation

Large language models (LLMs) and vision-language models (VLMs) excel at understanding semantically complex commands and visual contexts. However, most general-purpose VLMs lack built-in support for precise spatial tasks such as segmentation and object detection, mostly due to the lack of such training data and specialized training procedures. Recent studies confirm that a Transformer architecture trained with a sufficiently large amount of localization data can achieve superior performance on object detection and segmentation tasks (Xiao et al. 2023).

The literature highlights several key limitations of current VLMs in region-level localization tasks. Most generalist VLMs output text-based responses without explicit or reliable spatial grounding, limiting their direct applicability to detection or segmentation. Specialized adaptations have emerged, including learnable coordinate tokens and decoupled localization pipelines, e.g., ROD-MLLM (Yin et al. 2025), designed to better align textual instructions with precise visual regions.

Moreover, models such as LISA (Lai et al. 2024) and u-LLaVA (Xu et al. 2024) introduce segmentation tokens as prompts within VLMs to improve mask prediction, yet these often still focus on single-object instances, lacking robust mechanisms for multi-object or group-level reasoning. Subsequent approaches such as GLaMM (Rasheed et al. 2024), PerceptionGPT (Pi et al. 2023), and PixelLM (Ren et al. 2024b) extend this with condition tokens for multi-object grounding, but their effectiveness degrades with increasing instruction complexity and relational reasoning demands.

Semantically Complex Object Detection and Segmentation

A recent review of VLM usage for object detection and segmentation underscores that despite progress achieved through multi-modal pretraining, semantically complex region detection—especially for abstract or relational concepts—remains an open challenge (Feng et al. 2025). Their analysis reveals persistent difficulties in grounding abstract queries, rejecting false positives, and accurately localizing regions defined by high-level semantics rather than explicit category labels.

To address this, recent work such as Ground-V (Zong et al. 2025) has taken a step forward by explicitly training VLMs to align natural language descriptions—including complex and compositional instructions—with pixel-level regions. Ground-V demonstrates improved grounding for in-

structions involving actions and scene context, but it primarily targets synthetic or simple compositional examples and lacks evaluation in socially grounded, real-world scenarios.

As existing methods fail when regions are described using semantically rich concepts, there is a clear need for methods that extend VLMs toward reasoning-driven localization in complex real-world scenes. Our work fills the gap by focusing on social interaction regions in urban scenes—an example of semantically complex region detection with real-world societal impact. By combining pairwise reasoning with spatial aggregation and introducing the MINGLE dataset, we contribute toward enabling robust detection of abstract, relational visual concepts that current models struggle to handle, as well as to urban planning analysis, enabling the design of healthier, more inclusive cities.

Method

In this work, we introduce a new task—social group region detection, which requires generating a bounding box given a query text involving implicit reasoning capabilities. The main difference to the conventional object detection task is that the query text is not limited to a straightforward object reference (e.g., “the person”), but is a more complex description requiring complex reasoning and world knowledge, for example:

Based on visual cues, determine whether the individuals are actively interacting (e.g., engaged in conversation or displaying clear interactive behavior) or if they are merely near each other without meaningful interaction.

This task is characterized by three key challenges: 1) the complexity of the text query and visual context; 2) multiple regions that satisfy criteria simultaneously, and 3) the potential for hallucinated references of regions not present in the scene. In the following section, we describe a three-stage workflow that enables overcoming these challenges to efficiently detect socially affiliated groups of people in street-view images (see Figure 3).

Social Group Region Detection Task

We formulate the problem as follows. Let

$$\mathcal{I} = \{I_1, I_2, \dots, I_M \mid I_m \in \mathbb{R}^{H \times W \times 3}\}$$

be a set of RGB streetview images. For each image $I_m \in \mathcal{I}$, let

$$\mathcal{P}^m = \{p_1^{(m)}, p_2^{(m)}, \dots, p_K^{(m)} \mid K \geq 0\},$$

$$\mathcal{G}^m = \{g_1^{(m)}, g_2^{(m)}, \dots, g_L^{(m)} \mid L \geq 0\}$$

denote, respectively, the set of detected single person bounding boxes and socially affiliated group bounding boxes in image I_m , where each g_i^m for image I_m contains at least two individual detections from \mathcal{G}^m such that $|\{p_j^m \in \mathcal{P}^m \mid p_j^m \subseteq g_i^m\}| \geq 2$.

The goal is to detect all such regions containing social groups that consist of two or more socially connected individuals in each image I_m .

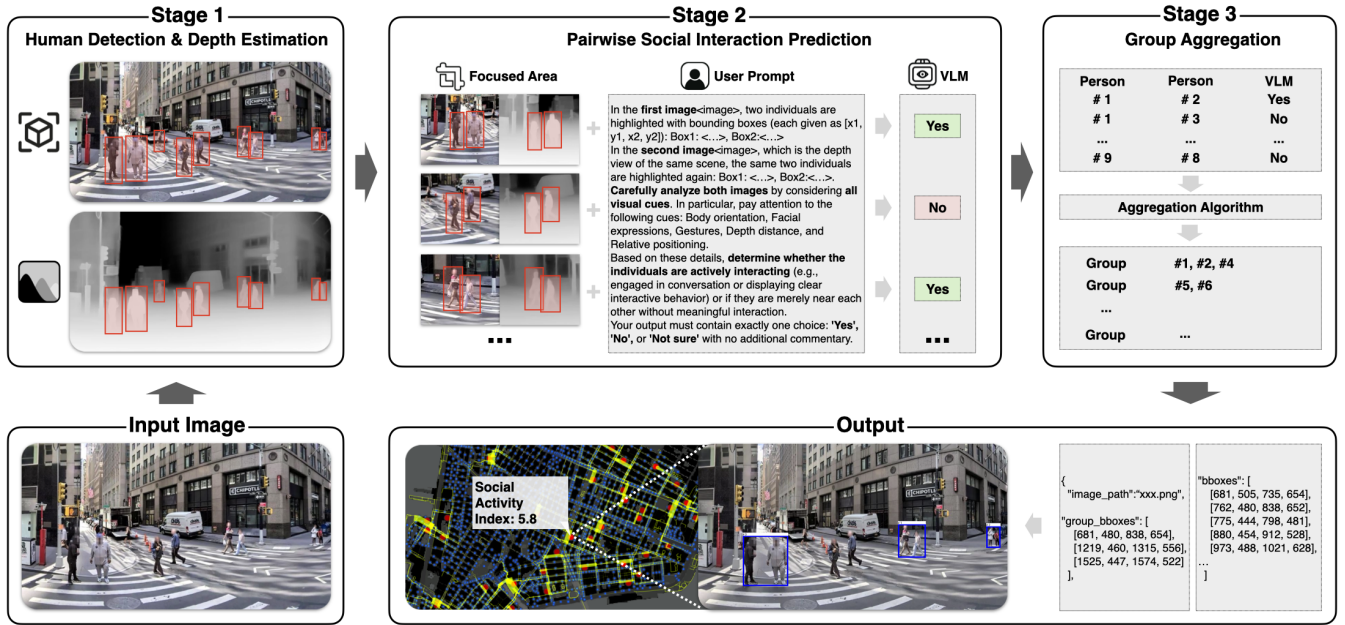


Figure 3: Illustration of the three-stage pipeline for detecting semantically complex social interaction regions.

Let $\mathcal{R} : \mathbb{R}^{H \times W \times 3} \rightarrow \mathcal{G}$ be the full pipeline that maps an image I_m to a set of socially affiliated group region bounding boxes \mathcal{G} .

In order to ensure that pipeline results do not suffer from hallucinated references or struggle with multi-region detections, as is common for VLM-based object detection methods (Zong et al. 2025), we propose a three-stage pipeline leveraging existing object-detection models’ capabilities.

Three-Stage Pipeline

The proposed pipeline combines off-the-shelf human detection, VLM-based pairwise reasoning, and a lightweight spatial aggregation algorithm to detect socially affiliated groups in street-view images (see Figure 3).

Stage 1: Single Person Detection

In order to address two common critical issues faced by VLMs applied to object detection—hallucinated references and multiple detections, we design our pipeline based on a robust, high-precision, and recall human detector.

Given an image $I_m \in \mathcal{I}$, we first apply a pre-trained human detector

$$\mathcal{D} : \mathbb{R}^{H \times W \times 3} \rightarrow \mathcal{P}$$

to predict a set of bounding boxes for the person category, where each p_i is an axis-aligned bounding box b_i with an associated confidence score $s_i \in [0, 1]$. We retain only high-certainty predictions by discarding detections with $s_i^{(m)} < \tau_{\text{det}}$, where τ_{det} can be empirically chosen based on the human-predictor model and the desired precision–recall trade-off.

We instantiate \mathcal{D} with the ATSS-Swin-L-DyHead model (Zhang et al. 2020). However, Stage 1 is detector-agnostic: \mathcal{D} may be replaced by any modern human detector, e.g., a YOLOv8 variant (Jocher, Qiu, and Chaurasia 2023),

for further improvements without requiring any modifications to the downstream pipeline.

Stage 2: Pairwise Social-Interaction Prediction

For every unordered pair of bounding boxes $(p_i^{(m)}, p_j^{(m)})$, we fine-tune Qwen2-VL-7B and Qwen2.5-VL-7B to perform classification (“Yes”, “No”, or “Not sure”), indicating whether the pair belongs to the same social group using a human-annotated dataset (see Dataset Construction section).

Visual inputs. We first form a padded union box

$$U_{ik}^{(m)} = \text{Pad}(\text{BBoxUnion}(b_i^{(m)}, b_k^{(m)})),$$

then crop the identical region from (i) the RGB image I_m and (ii) the predicted depth map D_m obtained with EVP (Lavreniuk et al. 2024). These two crops,

$$F_{ik,\text{rgb}}^{(m)} = \text{Crop}(I_m, U_{ik}^{(m)}),$$

$$F_{ik,\text{depth}}^{(m)} = \text{Crop}(D_m, U_{ik}^{(m)})$$

are both overlaid with 1-pixel bounding boxes for the two individuals, so the model receives precisely aligned RGB–depth views focused on the candidate pair.

Depth-aware textual cues. Within each original bounding box, we take all depth pixels in D_m and compute their medians, $d_i^{(m)}$ and $d_k^{(m)} \in [0, 255]$. These two scalars, together with their absolute difference $|d_i^{(m)} - d_k^{(m)}|$, are inserted verbatim into the prompt text (see Figure 3).

Thus, the model simultaneously accounts for (i) fine-grained RGB appearance, (ii) relative geometry in the depth crop, and (iii) explicit numeric depth values, enabling robust reasoning even when visually adjacent persons are separated in 3-D space.

In order to eliminate highly unlikely pairs and make the inference more scalable, we propose an algorithm that allows reducing the number of VLM calls needed for reliable group detection (see Algorithm 1).

Algorithm 1: Distance and Depth Filtering Algorithm

Input: List of bounding boxes $\mathcal{P} = \{p_1, p_2, \dots, p_N\}$ with depth values z_i

Parameters: Distance threshold τ_d , depth threshold τ_z

Output: Pairwise relation matrix $\mathcal{R} \in \{\text{Yes, No, Not Sure}\}^{N \times N}$

```

1: Initialize  $\mathcal{R}_{i,j} \leftarrow \text{Not Sure}$  for all  $i, j$ 
2: for each  $p_i \in \mathcal{P}$  do
3:   for each  $p_j \in \mathcal{P}$  do
4:     if  $i = j$  then
5:       continue
6:     end if
7:     if  $d(p_i, p_j) > \tau_d$  then
8:        $\mathcal{R}_{i,j} \leftarrow \text{No}$ 
9:       continue
10:    end if
11:    if  $|z_i - z_j| > \tau_z$  then
12:       $\mathcal{R}_{i,j} \leftarrow \text{No}$ 
13:      continue
14:    end if
15:    result  $\leftarrow$  VLM( $p_i, p_j$ ) {Returns Yes/No/Not Sure}
16:     $\mathcal{R}_{i,j} \leftarrow$  result
17:  end for
18: end for
19: return  $\mathcal{R}$ 

```

We choose thresholds τ_z and τ_d empirically based on the trade-off between computational speed vs. group detection accuracy (see Results section).

Stage 3: Group Aggregation and Region Generation

Given the answers from Stage 2 for all unordered person pairs $(p_i^{(m)}, p_k^{(m)})$, we construct the social groups by applying a greedy clustering algorithm.

Let $J^{(m)} : \mathcal{P}_{I_m} \times \mathcal{P}_{I_m} \rightarrow \{\text{Yes, No, Not sure}\}$ be the set of pairwise judgments indicating whether two individuals belong to the same group.

We initialize each person as a singleton cluster, then iteratively merge the pair of clusters (C_a, C_b) that maximally increases an agreement score. The score is defined based on whether merging the clusters aligns with the pairwise annotations in $J^{(m)}$. If $(p_i^{(m)}, p_k^{(m)}) = \text{“Yes”}$, then placing i and k in the same cluster increases the score, while if the answer is “No” or “Not sure”, it does not.

This process continues greedily until no further improvement can be made. We define valid social groups as those clusters with cardinality at least 2, i.e., $|g_j^{(m)}| \geq 2$.

For each group $g_j^{(m)}$, we compute a group bounding box $B_j^{(m)}$ by taking the smallest and largest coordinates from all individual person boxes in the group:

$$B_j^{(m)} = \left(\min x_p^{(1)}, \min y_p^{(1)}, \max x_p^{(2)}, \max y_p^{(2)} \right), p \in g_j^{(m)}$$

where each person’s box is $b_p = (x_p^{(1)}, y_p^{(1)}, x_p^{(2)}, y_p^{(2)})$. These group boxes provide a compact spatial summary of

social groups for visualization and evaluation.

Experiments

Fine-Tuning Dataset Construction

To construct the fine-tuning dataset, we collected 79,265 human-labeled annotations for pairwise social interactions between single people in the images. In each training image, exactly two individuals are highlighted with bounding boxes. Human annotators were instructed to identify whether the indicated individuals belong to the same social group (“yes”), are independent from each other (“no”), or it is not possible to determine (“not sure”). To ensure that the resulting dataset is balanced with regard to “yes” and “no” answers, the annotation process consisted of two iterations: (1) Human annotators received 46K randomly selected street-view images. (2) Based on the analysis of the first iteration (skewed towards “no” answers), we pre-selected pairs with individuals potentially belonging to the same group based on the first-stage fine-tuned Qwen-2 (7B) model answers. The resulting dataset distribution is described in Table 1.

Split	Pairs	“Yes”	“No”	“Not Sure”
Overall	79,265	29,716	45,029	4,520
Train	76,265	28,341	43,654	4,270
Test	2,000	925	925	150
Validation	1,000	450	450	100

Table 1: Pairwise Relationships Dataset Overview.

For the model training, the pairwise-relationships dataset is additionally split into different subsets. Variants differ by dataset sampling (balanced vs. unbalanced) and pre-processing (deletion or replacement of low-confidence pairs, i.e., those labeled as “not sure”, where “YN” indicates that all “not sure” answers were omitted).

Model Selection

To identify a strong yet efficient model for pairwise social affiliation classification, we benchmarked several open-source and commercial VLMs on a curated set of 2,000 manually validated image pairs. As shown in Table 2, larger models like Qwen2-VL (72B) outperform others in overall F_1 . However, smaller variants—especially Qwen2-VL (7B)—obtain just marginally lower effectiveness, while offering greater robustness to overfitting, lower memory consumption, and significantly faster inference. The latter is crucial for urban research,¹ and therefore Qwen2-VL (7B) and Qwen2.5-VL (7B) were selected for further fine-tuning.

Prompt Generation

To create suitable text prompts for guiding the model towards social group understanding, we evaluated a variety of formulations augmented with additional visual and textual

¹Urban analysis requires working with vast amounts of data; e.g., for New York City alone, the number of street views exceeds 8M.

Model	Model Size	Fine-tuning Method	Fine-tuning Data	F_1		Precision		Recall	
				Yes	No	Yes	No	Yes	No
Evaluation result of first round fine-tuning									
Qwen2-VL	7B	LoRA (rank 64)	29k-YN	0.64	0.57	0.56	0.69	0.75	0.49
Qwen2.5-VL	7B	LoRA (rank 64)	29k-YN	0.65	0.53	0.55	0.72	0.82	0.42
Qwen2.5-VL	7B	LoRA (rank 128)	29k-YN	0.66	0.39	0.52	0.78	0.91	0.26
Evaluation result of second round fine-tuning									
Qwen2-VL	7B	LoRA (rank 64)	61k-Balanced	0.68	0.62	0.59	0.75	0.80	0.53
Qwen2-VL	7B	LoRA (rank 64)	76k-Unbalanced	0.70	0.72	0.67	0.75	0.73	0.69
Qwen2-VL	7B	LoRA (rank 64)	61k-Balanced-YN	0.72	0.70	0.65	0.78	0.79	0.63
Qwen2-VL	7B	LoRA (rank 64)	76k-Unbalanced-YN	0.68	0.65	0.61	0.75	0.78	0.57
Baseline results (no fine-tuning)									
Phi 3.5	4.2B	—	—	0.57	0.24	0.45	0.47	0.79	0.16
Qwen2-VL	7B	—	—	0.53	0.54	0.50	0.58	0.58	0.50
Qwen2.5-VL	7B	—	—	0.10	0.69	0.57	0.54	0.05	0.97
Qwen2-VL	72B	—	—	0.62	0.58	0.56	0.67	0.71	0.51
Qwen2.5-VL	72B	—	—	0.59	0.56	0.54	0.64	0.68	0.49
Pixtral	12B	—	—	0.52	0.51	0.48	0.56	0.46	0.57
Gemma 3	24B	—	—	0.27	0.67	0.49	0.54	0.18	0.84
LLama3.2-Vision	11B	—	—	0.46	0.30	0.40	0.38	0.55	0.25
GPT 4o	—	—	—	0.00	0.70	0.00	0.54	0.00	1.00
Claude-Sonnet	—	—	—	0.24	0.66	0.47	0.54	0.16	0.85

Table 2: Model Performance on pairwise relationship classification task. Fine-tuning results in the upper part of the table, zero-shot VLM performance in the lower part.

cues (e.g., bounding box coordinates for individual people, numeric depth and distance values, etc.). We determined that the best results are achieved when the model receives inputs and prompts as illustrated in Figure 3. We release the nine best-performing prompt versions in the project repository.

Evaluation

To assess the effectiveness of our proposed approach—from pairwise interaction reasoning to full-region social group detection—we conducted two complementary evaluations: one at the relationship level and one at the group region level.

Pairwise-relationships We first evaluated the performance of the fine-tuned VLM on the task of predicting whether two individuals in a street-view image belong to the same social group. For this purpose, we used the test split of the fine-tuning dataset, which was additionally revised to ensure a high quality of labels. For each of the answer options (‘Yes’ and ‘No’), we compute common binary classification metrics: F_1 , precision, and recall. We compare our results to the zero-shot performance of available open-source and commercial VLMs as a baseline.

Social Group Region Detection To evaluate the quality of group region detections produced by our pipeline, we adopt the ELSA benchmark (Hosseini et al. 2024), as it provides the most relevant content—urban street-view imagery annotated with social group bounding boxes across a range of activities and group sizes.

We evaluate the predicted group region bounding boxes using the mean Intersection over Union (mIoU) to measure spatial alignment with annotated group regions. In addition, we report precision, recall, and F_1 score, using an IoU threshold of 0.5, consistent with common object detection evaluation practices.

Implementation Details

We tested several fine-tuning configurations, including LoRA adapters with rank 64 and 128, and full fine-tuning. For fine-tuning, we used 8 NVIDIA A100 GPUs and repeated all experiments twice. Each model was trained for 5 epochs and the best-performing one is selected based on the validation set performance.

Results

Fine-tuning

To determine the best fine-tuning setup, we conducted a set of experiments on different subsets of the dataset and different preprocessing options (see selected experiments in Table 2). The results illustrate that fine-tuning confers substantial gains over the zero-shot performance. While increasing the dataset size from 29K to 66K or 76K allowed for improving the model’s effectiveness and reducing overfitting, data quality and dataset balance are crucial for achieving superior and more robust results.

Our experiments with full parameter fine-tuning showed promising results. However, given the relatively small train-

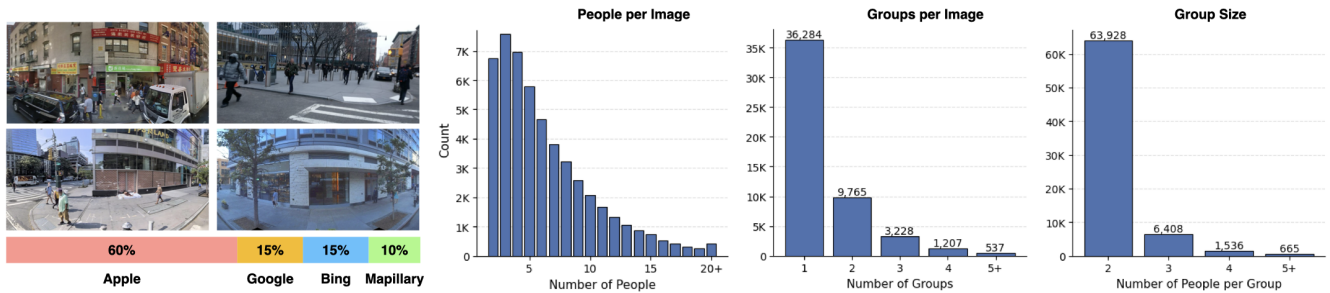


Figure 4: Descriptive statistics pertaining to our Social Group Region dataset.

ing dataset size and resource limitations, they are currently excluded from the scope of this study.

Social Group Region Detection

To measure the real-world performance of our full pipeline, we evaluate detected social group regions and compare our results to the zero-shot performance of state-of-the-art open source and commercial VLMs (see Table 3). Our approach substantially outperforms the baselines and illustrates the current inability of all examined models to accurately interpret and localize interpersonal relationships, showing that they remain inadequate for this challenging task.

Model	mIoU	F_1	Precision	Recall
MINGLE	0.64	0.60	0.75	0.61
GPT-4o	0.01	0.01	0.19	0.01
Claude-Sonnet	0.02	0.02	0.02	0.02
Claude-Opus	0.01	0.02	0.20	0.02
LLama3.2-Vision	0.00	0.00	0.91	0.00
Qwen2-VL (7B)	0.00	0.00	0.88	0.00

Table 3: Currently available VLMs are unable to reliably detect semantically complex regions.

Distance and Depth Filtering Effects

Using thresholds for depth and distance allows us to exclude pairs of individuals that are highly unlikely to belong to the same social group, thus decreasing the computational cost and improving the overall effectiveness by reducing false positives.

Filtering	τ_d	τ_z	mIoU	F_1	Pairs
No Filtering	255	1	0.635	0.604	1071
Optimal Depth	80	1	0.634	0.603	1070
Optimal Distance	255	0.4	0.636	0.608	1068
Combined	80	0.4	0.634	0.607	1068

Table 4: Summary of depth and distance filtering effects on the social group region detection.

We ran a sweep over all (1) distance values [0,1] with step size 0.1, and (2) depth values [0,255] with step size 20.

The results demonstrate that the optimal effectiveness is attained at distance values 0.3–0.5 and depth values 80–100.

However, the limited size of the benchmark dataset may limit the generalization of this finding, and we recommend re-running a similar analysis on a relevant dataset of interest.

Social Group Region Dataset

We introduce a large-scale street scene dataset comprising 100K annotated images for social group detection research. The dataset combines manual annotations with pipeline-propagated labels to ensure both quality and scale. Images were systematically collected from four major mapping platforms to capture diverse urban environments and imaging conditions: Apple Lookaround (60%), Google Street View (15%), Bing Streetside (15%), and Mapillary (10%).

To optimize social interaction visibility, we extract lateral viewpoints (left and right) from panoramic imagery, as these perspectives provide superior coverage of pedestrian interactions compared to frontal or rear views. Each image contains a minimum of two human detections, resulting in 588,430 total human instances across the dataset. Ground truth annotations include 72,537 social group regions, with detailed statistics on people-per-image, groups-per-image, and group composition distributions presented in Figure 4.

This dataset addresses the critical gap in training data for semantically complex region detection, providing the first large-scale benchmark specifically designed for social interaction analysis in urban environments from street-view imagery. The combination of diverse data sources, viewpoint selection, and comprehensive annotation coverage makes this dataset suitable for both model training and systematic evaluation of social group detection algorithms.

Conclusion

In this work, we address the previously unsolved urban analysis challenge: social group region detection. Our extensive zero-shot evaluation confirms that existing models fail to reliably localize social groups, highlighting the semantic complexity of the task. To tackle this, we have presented MINGLE: a modular three-stage VLM-based pipeline, which reframes the problem from direct region-level detection to pairwise social affiliation classification.

To encourage future work, we release a large-scale street-view dataset annotated with granular information about social groups created through a combination of manual and pipeline-based labeling.

Acknowledgments

The project was supported by the Designing for Sustainability research program, a joint initiative of the Morningside Academy for Design at MIT and the Hasso Plattner Institute, generously funded by the Hasso Plattner Foundation.

This work used the Jetstream2 GPU and NCSA Delta GPU resources through allocation CIS250088 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, supported by the U.S. National Science Foundation (NSF) under grants #2138259, #2138286, #2138307, #2137603, and #2138296.

References

- Cheng, T.; Song, L.; Ge, Y.; Liu, W.; Wang, X.; and Shan, Y. 2024. YOLO-World: Real-Time Open-Vocabulary Object Detection. In *Proceedings of the IEEE/CVF CVPR*, 16901–16911.
- Ehsanpour, M.; Saleh, F.; Savarese, S.; Reid, I.; and Rezatofighi, H. 2022. JRDB-Act: A Large-Scale Dataset for Spatio-Temporal Action, Social Group and Activity Detection. In *Proceedings of the IEEE/CVF CVPR*, 20983–20992.
- Feng, Y.; Liu, Y.; Yang, S.; Cai, W.; Zhang, J.; Zhan, Q.; Huang, Z.; Yan, H.; Wan, Q.; Liu, C.; Wang, J.; Lv, J.; Liu, Z.; Shi, T.; Liu, Q.; and Wang, Y. 2025. Vision-Language Model for Object Detection and Segmentation: A Review and Evaluation. arXiv:2504.09480.
- Gehl, J. 1987. *Life Between Buildings*. Danish Architectural Press.
- Gu, C.; Sun, C.; Ross, D. A.; Vondrick, C.; Pantofaru, C.; Li, Y.; Vijayanarasimhan, S.; Toderici, G.; Ricco, S.; Sukthankar, R.; Schmid, C.; and Malik, J. 2018. AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Actions. arXiv:1705.08421.
- Hosseini, M.; Cipriano, M.; Eslami, S.; Hodczak, D.; Liu, L.; Sevtuk, A.; and de Melo, G. 2024. ELSA: Evaluating Localization of Social Activities in Urban Streets using Open-Vocabulary Detection. arXiv:2406.01551.
- Jacobs, A. B. 1985. *Looking at Cities*. Harvard University Press.
- Jacobs, J. 1961. *The Death and Life of Great American Cities*. Random House.
- Jahangard, S.; Cai, Z.; Wen, S.; and Rezatofighi, H. 2024. JRDB-Social: A Multifaceted Robotic Dataset for Understanding of Context and Dynamics of Human Interactions within Social Groups. In *Proceedings of the IEEE/CVF CVPR*, 22087–22097.
- Jocher, G.; Qiu, J.; and Chaurasia, A. 2023. Ultralytics YOLO.
- Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; and Jia, J. 2024. LISA: Reasoning Segmentation via Large Language Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9579–9589.
- Lavreniuk, M.; Bhat, S. F.; Müller, M.; and Wonka, P. 2024. EVP: Enhanced Visual Perception using Inverse Multi-attentive Feature Refinement and Regularized Image-Text Alignment. In *European Conference on Computer Vision*, 206–225. Springer.
- Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. 2024. Grounding DINO: Marrying DINO with Grounded Pre-training for Open-Set Object Detection. In *European Conference on Computer Vision*, 38–55. Springer.
- Mehta, V.; and Bosson, J. K. 2021. Revisiting Lively Streets: Social Interactions in Public Space. *Journal of Planning Education and Research*, 41(2): 160–172.
- Pi, R.; Yao, L.; Gao, J.; Zhang, J.; and Zhang, T. 2023. PerceptionGPT: Effectively Fusing Visual Perception into LLM. arXiv:2311.06612.
- Qi, J.; Mazumdar, S.; and Vasconcelos, A. C. 2024. Understanding the Relationship between Urban Public Space and Social Cohesion: A Systematic Review. *International Journal of Community Well-Being*, 7(2): 155–212.
- Rasheed, H.; Maaz, M.; Mullappilly, S. S.; Shaker, A.; Khan, S.; Cholakkal, H.; Anwer, R. M.; Xing, E.; Yang, M.-H.; and Khan, F. S. 2024. GLaMM: Pixel Grounding Large Multimodal Model. arXiv:2311.03356.
- Ren, S.; He, K.; Girshick, R.; and Sun, J. 2016. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6): 1137–1149.
- Ren, T.; Liu, S.; Zeng, A.; Lin, J.; Li, K.; Cao, H.; Chen, J.; Huang, X.; Chen, Y.; Yan, F.; et al. 2024a. Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks. arXiv preprint arXiv:2401.14159.
- Ren, Z.; Huang, Z.; Wei, Y.; Zhao, Y.; Fu, D.; Feng, J.; and Jin, X. 2024b. PixelLM: Pixel Reasoning with Large Multimodal Model. arXiv:2312.02228.
- Wang, Z.; Ying, K.; Meng, J.; and Ning, J. 2023. Human-to-Human Interaction Detection. arXiv:2307.00464.
- Wei, J.; Zhou, T.; Yang, Y.; and Wang, W. 2024. Nonverbal Interaction Detection. In *European Conference on Computer Vision*, 277–295. Springer.
- Whyte, W. H.; et al. 1980. *The Social Life of Small Urban Spaces*, volume 116. Conservation Foundation Washington, DC.
- Xiao, B.; Wu, H.; Xu, W.; Dai, X.; Hu, H.; Lu, Y.; Zeng, M.; Liu, C.; and Yuan, L. 2023. Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks. arXiv:2311.06242.
- Xu, J.; Xu, L.; Yang, Y.; Li, X.; Wang, F.; Xie, Y.; Huang, Y.-J.; and Li, Y. 2024. u-LLaVA: Unifying Multi-Modal Tasks via Large Language Model. arXiv:2311.05348.
- Yin, H.; Ren, Y.; Yan, K.; Ding, S.; and Hao, Y. 2025. ROD-MLLM: Towards More Reliable Object Detection in Multimodal Large Language Models. In *Proceedings of CVPR*, 14358–14368.
- Zhang, S.; Chi, C.; Yao, Y.; Lei, Z.; and Li, S. Z. 2020. Bridging the Gap Between Anchor-based and Anchor-free Detection via Adaptive Training Sample Selection. In *Proceedings of the IEEE/CVF CVPR*, 9759–9768.
- Zong, Y.; Zhang, Q.; An, D.; Li, Z.; Xu, X.; Xu, L.; Tu, Z.; Xing, Y.; and Dabeer, O. 2025. Ground-V: Teaching VLMs to Ground Complex Instructions in Pixels. arXiv:2505.13788.