Appendix I

Human Annotation for Grouping Detection

Labeling Instruction

We prepared labeling instructions for human annotators. In this instruction, we asked the annotators to answer whether the two highlighted persons within the bounding boxes are part of a group, which is defined as people who are interacting or closely associated in a way that suggests they are together intentionally. We also encouraged the annotator to infer from their appearance, proximity, and behavior.

For each question, the annotator could choose one out of the four options:

1) **Yes:**
   - The two individuals appear to be interacting or engaging with each other.
2) **No:**
   - The two individuals are not interacting and do not appear to be together.
3) **Not sure:**
   - The image is unclear or obstructed, or the boxes are not correct.
   - It is ambiguous whether the two individuals are interacting.
4) **Wrong Photos:**
   - No person in the bounding box or the person in the bounding box isn't real.

We also gave the annotator examples for each answer (see Figure I-1, Figure I-2, and Figure I-3).

Figure I-1 "Yes" Examples: The two people could be walking side by side, facing towards each other, or they could be a part of a larger chatting group.

Figure I-2 "No" Examples: One person is walking one way, and the other is walking in the opposite direction, looking in a different direction.

Figure I-3 "Not sure" Examples (left two): It is ambiguous whether the two individuals are interacting, and "Wrong Photos" Examples (right three): There is no person in either of the boxes, or the people in the boxes are unreal (banner or something else)

Data Preparation

There are two rounds of human annotation in this study.

The first round was conducted in December 2024, in which the annotation company labeled 40k images. 40k images were collected randomly from Google streetview, Bing Streetside, and Mapillary. After a person detection with the model "ATSS SWIN Large" pretrained on the COCO datasets, we got all the person's bounding boxes within all these images. Then, we randomly picked 40k pairs of people from each image to make this dataset. So there are some duplicate images in this dataset, but with different pairs of bounding boxes. From this dataset, we get 39999 annotations, including 5172 "Yes", 31231 "No", 2606 "Not sure", and 990 "Wrong Photos."

The second round was conducted in March 2025. This time, we created a dataset including 38755, but to increase the positive annotations, we adopted one of our fine-tuned models to predict and only keep those images with a prediction of "Yes" for this dataset. The annotation for this round has 18215 "Yes," 16277 "No," 2325 "Not sure," and 1938 "Wrong Photos."

For the specific Annotation Instruction, please see the "labeling_Instruction.pdf" under the "Supplementary Materials" folder.

Appendix II

Prompt Engineering for Pairwise Relationship Judgment

So far, I have tried 7 versions of prompts for model inference and finetuning, since different model prompts can yield different performance.

Below are prompts 1 to 3. Unfortunately, these prompts are not used since they could not constrain the model to only answer "Yes / No / Not sure" when fed with an image.

Prompt 1: Can you tell me whether the two persons in the bounding boxes belong to the same social group or not?

Prompt 2: Based on visual evidence of interaction, coordination, proximity and body language of two people in red...same group, or are they independently walking near each other? Answer with a simple yes or no, based on what is more probable.

Prompt 3: Based on visual evidence of interaction, coordination, proximity, and body language of the two individuals in the red bounding boxes in this image, do they appear to be interacting with each other and part of the same group, or are they independently near each other? Answer with a simple 'yes,' 'no,' or 'not sure,' based on what is most likely.

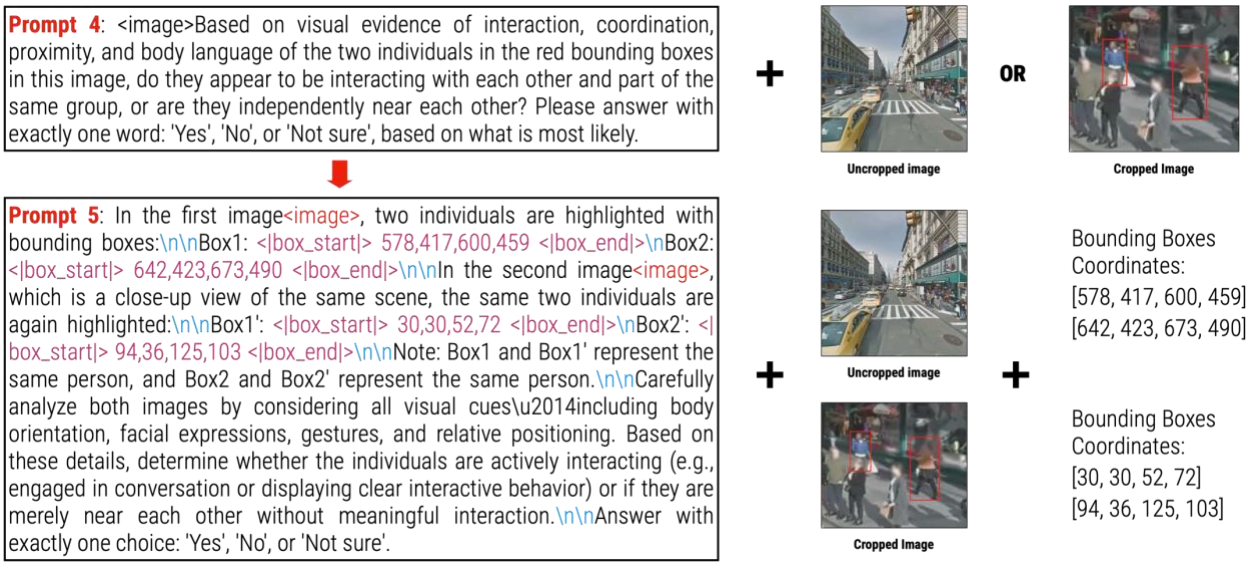
Starting from Prompt 4, the fine-tuning and inference tests began to work. In this prompt, either an uncropped image or a cropped image with 30px buffer can be given to a model. And the evaluation of zero-shot models (GPT, Qwen72b, Phi3.5, etc) is also based on this prompt plus the uncropped version og images.

Prompt 4: <image>Based on visual evidence of interaction, coordination, proximity, and body language of the two individuals in the red bounding boxes in this image, do they appear to be interacting with each other and part of the same group, or are they independently near each other? Please answer with exactly one word: 'Yes', 'No', or 'Not sure', based on what is most likely.

Prompt 5 are exploring the potentiality of using focused area in supporting model reasoning.

Prompt 5: In the first image<image>, two individuals are highlighted with bounding boxes:\n\nBox1: <|box_start|> 578,417,600,459 <|box_end|>\nBox2: <|box_start|> 642,423,673,490 <|box_end|>\n\nIn the second image<image>, which is a close-up view of the same scene, the same two individuals are again

highlighted:\n\nBox1': <|box_start|> 30,30,52,72 <|box_end|>\nBox2': <|box_start|> 94,36,125,103 <|box_end|>\n\nNote: Box1 and Box1' represent the same person, and Box2 and Box2' represent the same person.\n\nCarefully analyze both images by considering all visual cues\u2014including body orientation, facial expressions, gestures, and relative positioning. Based on these details, 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.\n\nAnswer with exactly one choice: 'Yes', 'No', or 'Not sure'.



Prompt 6 are exploring the potentiality of using depth map in supporting model reasoning.

Prompt 6: In the first image<image>, two individuals are highlighted with bounding boxes: Box1: <|box_start|> 426, 117, 485, 263 <|box_end|>\nBox2: <|box_start|> 577, 103, 619, 220 <|box_end|>\n\nIn the second image<image>, which is a close-up view of the same scene, the same two individuals are again highlighted:\n\nBox1': <|box_start|> 30, 44, 89, 190 <|box_end|>\nBox2': <|box_start|> 181, 30, 223, 147 <|box_end|>\n\nNote: Box1 and Box1' represent the same person, and Box2 and Box2' represent the same person.\n\nDepth Values (from 0-255, 0 means far and 255 means close): Box1 and Box1' = 249, Box2 and Box2' = 248\n\nCarefully analyze both images by considering all visual cues—including body orientation, facial expressions, gestures, depth distance, and relative positioning. Based on these details, 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.\n\nAnswer with exactly one choice: 'Yes', 'No', or 'Not sure'.

Prompt 7 is the latest version of our model since it combines the depth map with a focused area, which can give the model most important information in a concised way.

Prompt 7: In the first image<image>, two individuals are highlighted with bounding boxes (each given as [x1, y1, x2, y2]):\n\nBox1: <|box_start|> 135,117,194,263 <|box_end|>\nBox2: <|box_start|> 286,103,328,220 <|box_end|>\n\nIn the second image<image>, which is the depth view of the same scene, the same two individuals are highlighted again:\n\nBox1': <|box_start|> 135,117,194,263 <|box_end|>\nBox2': <|box_start|> 286,103,328,220 <|box_end|>\n\nNote: Box1 and Box1' represent the same person, and Box2 and Box2' represent the same person.\n\nDepth Values (from 0-255, where 0 means far and 255 means close; these values are critical for determining proximity): Box1 and Box1' = 179, Box2 and Box2' = 142, so the Depth Difference is 37\n\nCarefully analyze both images by considering all visual cues. In particular, pay attention to the following cues:\n- Body orientation\n- Facial expressions\n- Gestures\n- Depth distance\n- Relative positioning\n\nBased on these details, 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.\n\nYour output must contain exactly one choice: 'Yes', 'No', or 'Not sure' with no additional commentary.

Prompt 6: In the first image<image>, two individuals are highlighted with bounding boxes: Box1: <|box_start|> 426, 117, 485, 263 <|box_end|>\nBox2: <|box_start|> 577, 103, 619, 220 <|box_end|>\n\nIn the second image<image>, which is a close-up view of the same scene, the same two individuals are again highlighted:\n\nBox1': <|box_start|> 30, 44, 89, 190 <|box_end|>\nBox2': <|box_start|> 181, 30, 223, 147 <|box_end|>\n\nNote: Box1 and Box1' represent the same person, and Box2 and Box2' represent the same person.\n\nDepth Values (from 0-255, 0 means far and 255 means close): Box1 and Box1' = 249, Box2 and Box2' = 248\n\nCarefully analyze both images by considering all visual cues—including body orientation, facial expressions, gestures, depth distance, and relative positioning. Based on these details, 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.\n\nAnswer with exactly one choice: 'Yes', 'No', or 'Not sure'.

Uncropped Image | Cropped Image

Focused Image | Focused Image

Appendix III

The Street-Level Imagery Collection

Google Street View collects up to five directional images (front, back, left, right, and top) of each street segment, often covering nearly 20 years. Monthly and yearly time stamps enable a longitudinal analysis of urban change.

Bing Streetside stores similar directional imagery with precise date-and-time metadata, facilitating temporal comparisons across different periods of the day.

Mapillary Data is crowd-sourced imagery that extends coverage to pedestrian-only areas such as parks, plazas, and sidewalks that vehicular street views do not capture routinely.

Apple Look Around Data offers immersive, high-resolution, street-level imagery with interactive navigation. One significant advantage is that every panorama is timestamped.

As a result, over 8.2 million streetview images and geo-tagged photos have been aggregated for NYC (see Figure III-1), forming the backbone of computer vision analysis.

The Composition of the 100k Datasets

As the entire database is quite extensive, we selectively chose the following two peak hours based on the intensity of human activities:

1) Weekday and weekend midday periods (12:00 – 2:00 PM), when public spaces tend to attract lunchtime activity.

2) Evening peak hours (5:00 – 7:00 PM), when people are commuting, socializing, or returning home.

These sources enable a detailed exploration of the physical context, including architectural features, street design elements, and indications of human activity. In addition, Figures III-2, III-3, and III-4 present the temporal coverage of the three datasets—Bing, Mapillary, and Apple. We observed that Bing's imagery was predominantly captured on weekdays. Apple's images, by contrast, were strictly confined to the time window between noon and 10 p.m. Mapillary, likely due to its crowdsourced nature, exhibited a more balanced temporal distribution, including some coverage during nighttime hours.

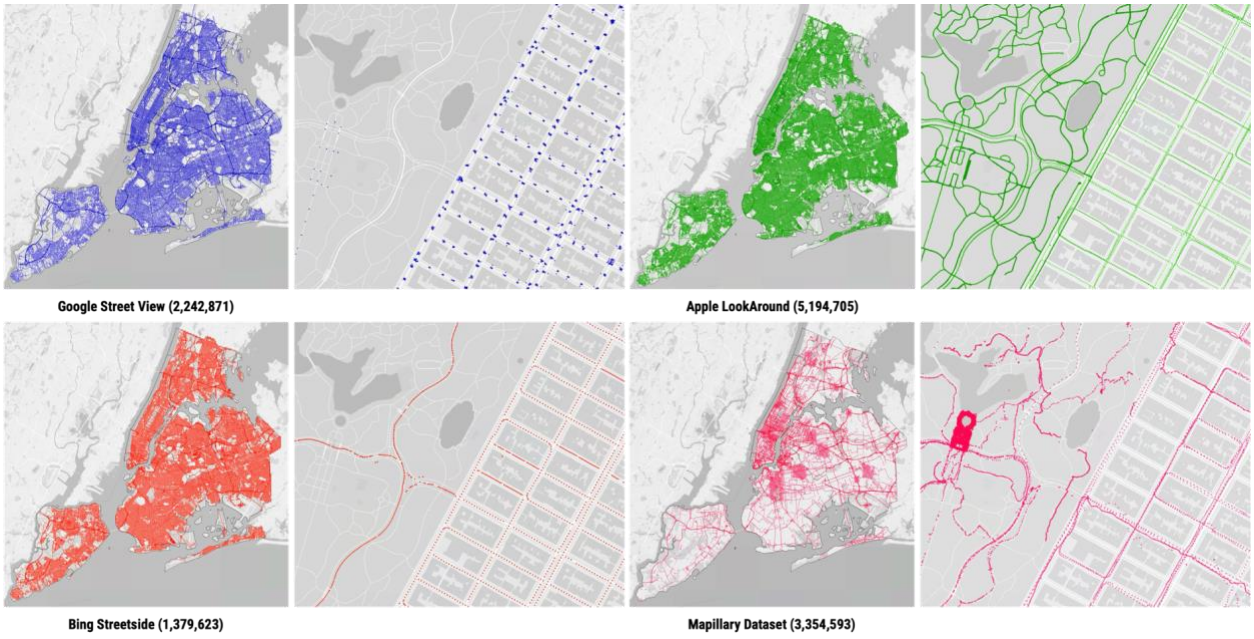


Figure III-1 The distribution of all kinds of streetview images

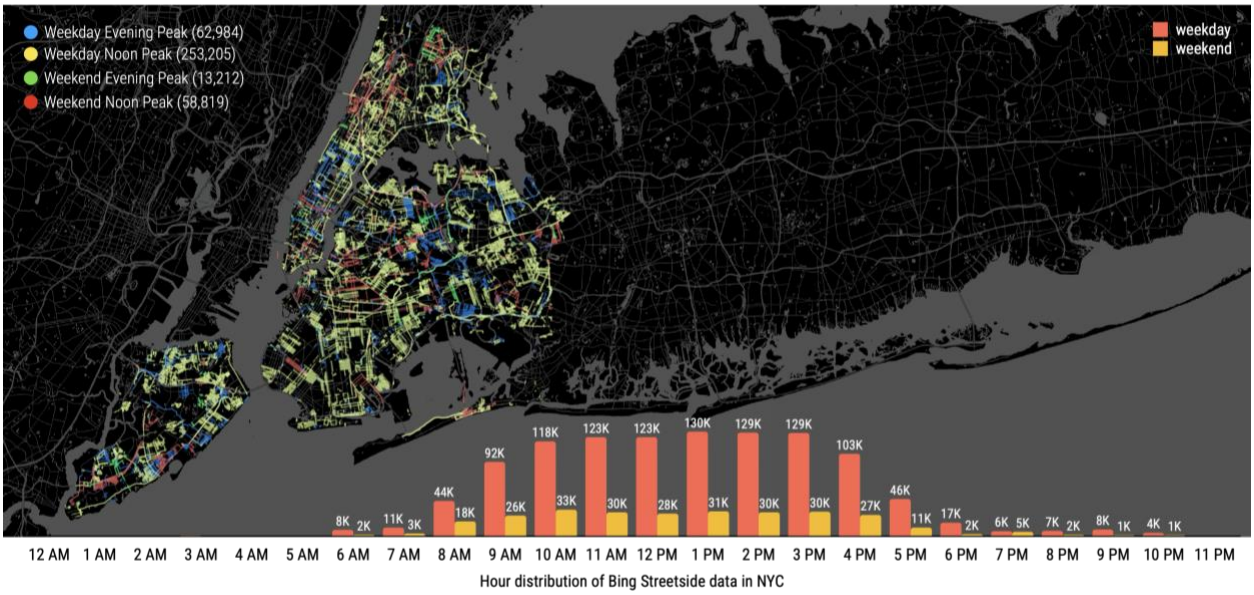


Figure III-2 Timestamps coverage from Bing Streetside and the Geospatial distribution of the Peak Hours

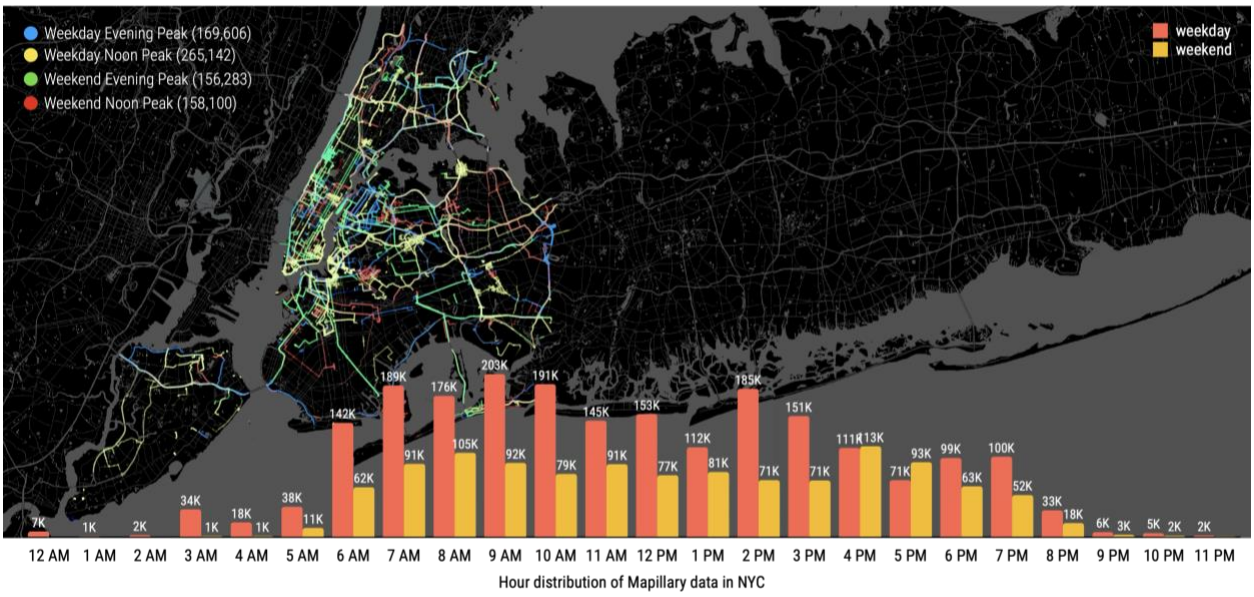


Figure III-3 Timestamps coverage from Mapillary and the Geospatial distribution of the Peak Hours

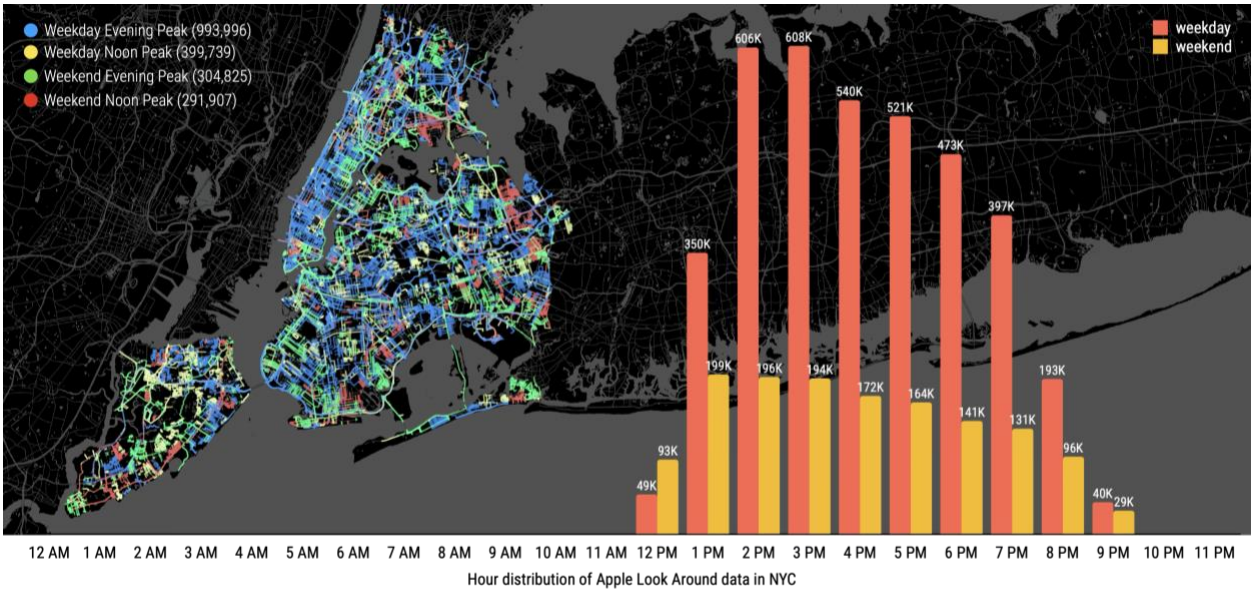


Figure III-4 Timestamps coverage from Apple Look Around and the Geospatial distribution of the Peak Hours

The Composition of the 100k Collection

Figure III-5 shows an overview of how the dataset is composed from all the different sources.

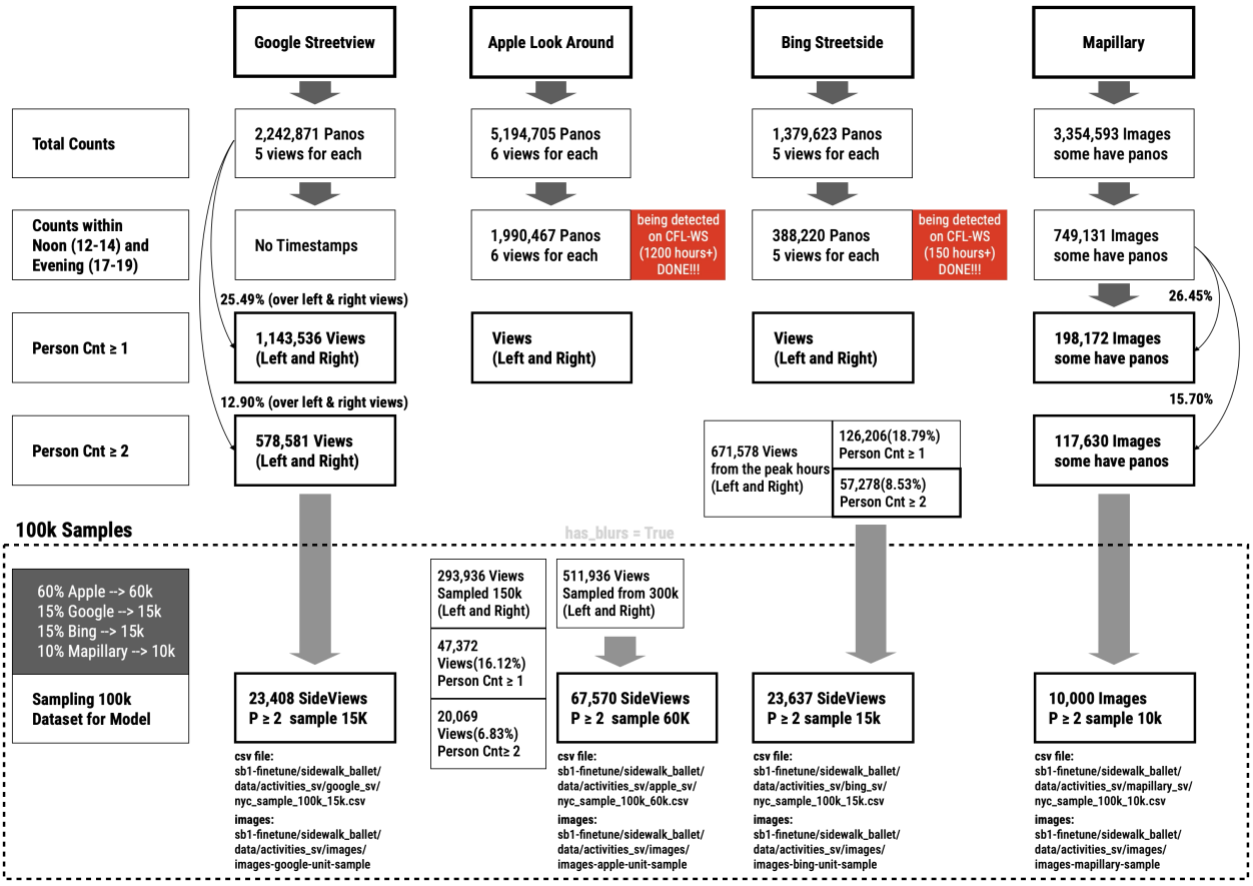


Figure III-5 The Pipeline of Composing the 100k Samples

The Stitching of Streetview into a wider view

Taking into account the differences in field of view among various street view images, we focused on a strategy that merges street view images from Apple, Google, and Bing. This approach enables us to perform sampling within specific neighborhoods and at defined spatial intervals. A specific example of how these wider-view images are composed is demonstrated in the GitHub Repo.

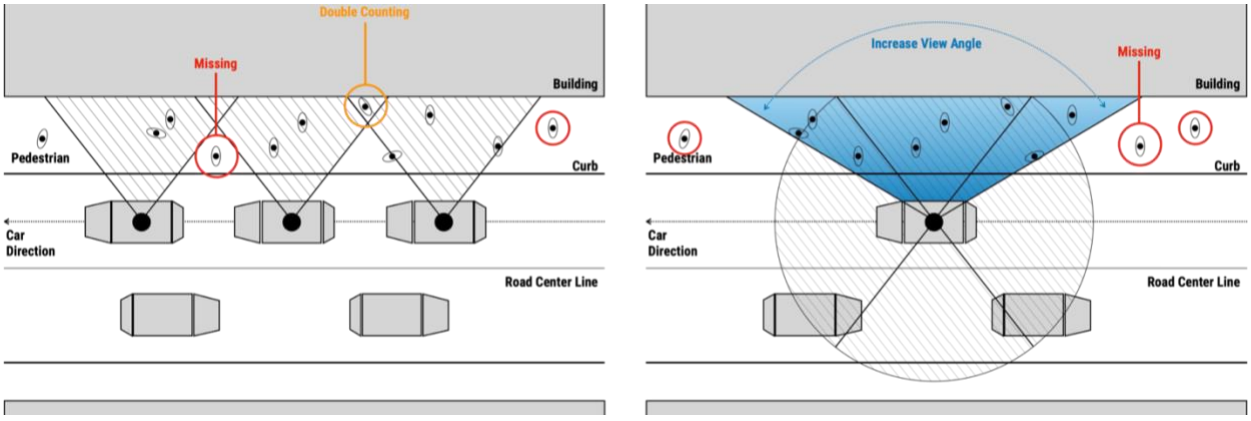


Figure III-6 The Comparison between the original Street View Extraction and the Stitched Wider View

Figure III-7 Get a wider view of the sidewalk and apply that as a Spatial Unit of Analysis

2) Rough Estimation of "in-the-view" Sidewalk Coverage

After empirical testing, a 150-degree horizontal field of view was selected as the optimal setting for image capture. This configuration balances a relatively broad viewing angle—sufficient to include both sides of the streetscape—with an acceptable level of visual distortion. In this study, each sidewalk segment is represented by a panoramic image captured at its midpoint using this 150-degree perspective. Given the typical urban geometry in New York City, including the average segment length (approximately 120 meters) and the constraints imposed by street width and visual occlusions, each image is estimated to capture approximately 20~25% of the total segment length. This estimation provides a rough measure of the proportion of the sidewalk that falls within the effective field of view at each observation point.

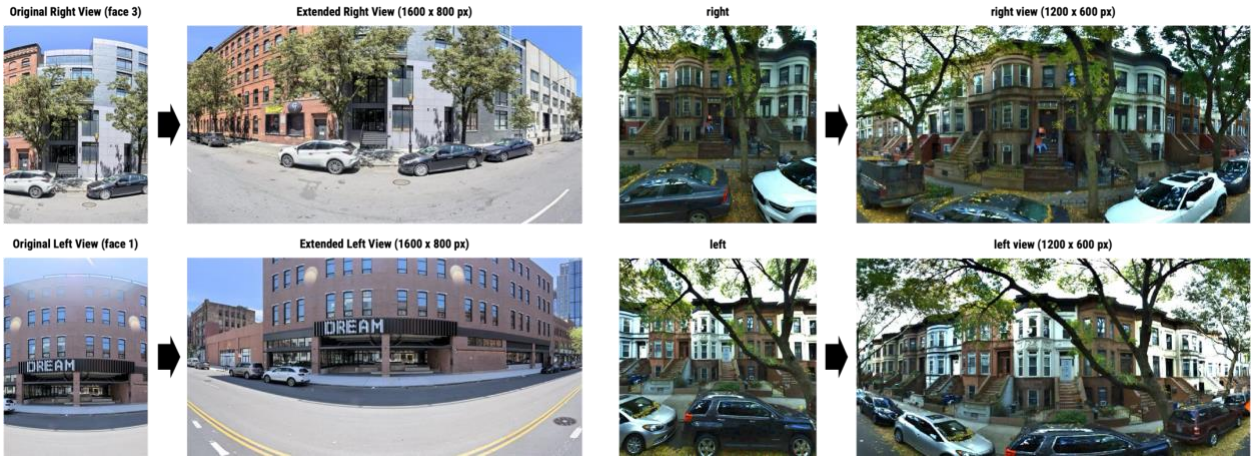


Figure III-8 The original side views and the 150-degree extended side views (left Apple and right Bing)