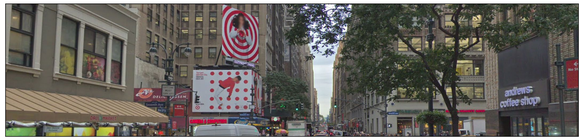
Touchdown: Natural Language Navigation and Spatial Reasoning in Visual Street Environment

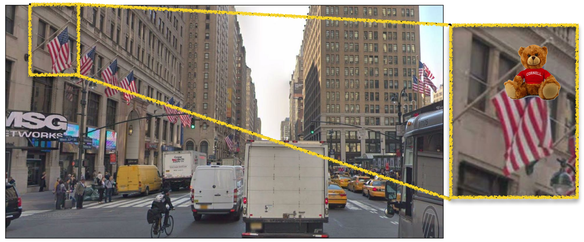
Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi

http://touchdown.ai

Tasks and Environment

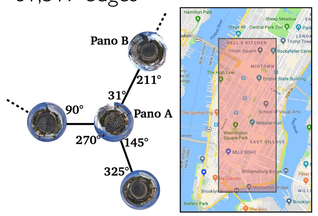
Turn and go with the flow of traffic. At the first traffic light turn left. Go past the next two traffic light. As you come to the third traffic light you will see a white building on your left with many American flags on it. Touchdown is sitting in the stars of the first flag.





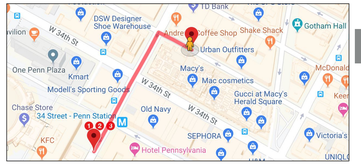
We create a graph of Manhattan using Google Street View

- 29,641 nodes, each 360° panorama
- 61,319 edges



Navigation

- **Goal**: follow instructions to get to the goal position
- **Actions**: forward, left, right, and stop
- **Evaluation**:
  1) Task competition
  2) Shortest path distance
  3) Success wighted by trajectory edit distance (SED)



Spatial Description Resolution (SDR)

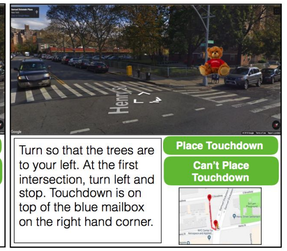
- **Goal**: locate Touchdown based on the description
- **Evaluation**:
  1) Accuracy
  2) Consistency
  3) Mean distance error

Data Collection

Task I: Writing

(a) the worker starts at the beginning of the route facing north

(b) the worker turns to face the correct direction and follows the path while writing instructions



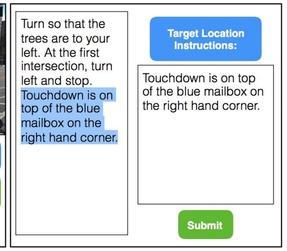
(c) the worker reaches the goal position, places Touchdown and completes the description

Task II: Pano Propagation

The worker annotates the location of Touchdown in the neighboring panoramas

Task III: Route Validation

The worker follows the instructions and guesses where Touchdown is when reaching the goal



Task IV: Text Segmentation

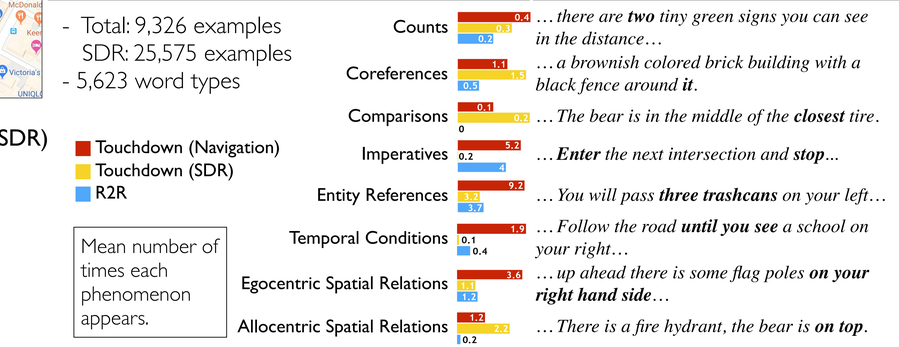
The worker highlights segments corresponding to the navigation and target location subtasks

Dataset Analysis

- Total: 9,326 examples
  SDR: 25,575 examples
- 5,623 word types

Touchdown (Navigation)
Touchdown (SDR)
R2R

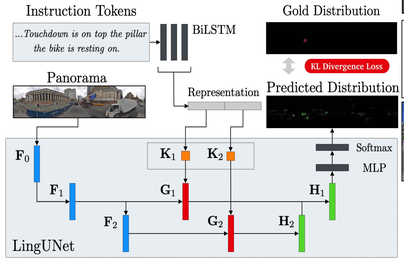
Mean number of times each phenomenon appears.



Experiments: Spatial Description Resolution

Systems

- **Text2Conv**: text-conditioned convolution filter generation
- **LingUNet**: multi-level text-conditioned image-feature reconstruction



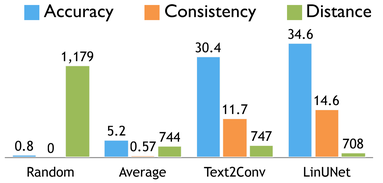
a black doorway with red brick to the right of it, and green brick to the left of it. it has a light just above the doorway, and on that light is where you will find touchdown.

the dumpster has a blue tarp draped over the end closest to you. touchdown is on the top of the blue tarp on the dumpster.

- LingUNet performs best
- Hard cases: object appears many times in a scene

Metrics

- **Accuracy**: correct pixel prediction
- **Consistency**: proportion of unique SDRs, for which predictions are correct for all paired panorama



Experiments: Navigation

Systems

- **GA:** gated attention
- **RConcat:** image-text concatenation

Metrics

- **Task Completion (TC)**: reaching the goal
- **Success Weighted by Edit Distance (SED)**:

$$\frac{1}{N}\sum_{i=1}^{N} S_i\left(1 - \frac{\text{lev}(\text{PredPath}, \text{TargetPath})}{\max(|\text{PredPath}|, |\text{TargetPath}|)}\right)$$

- **Shortest Path Distance (SPD)**: mean distance away from the goal

- RConcat performs best
- A lot of room for improvement