# Multimodal Text Style Transfer for Outdoor Vision-and-Language Navigation

Wanrong Zhu[1], Xin Wang[2], Tsu-Jui Fu[1], An Yan[3], Pradyumna Narayana[4], Kazoo Sone[4], Sugato Basu[4], William Yang Wang[1]

[1]*UC Santa Barbara ,*[2]*UC Santa Cruz,* [3]*UC San Diego,* [4]*Google*

UC SANTA BARBARA UC SANTA CRUZ UC San Diego Google

# Outdoor Vision-and-Language Navigation

- Challenges:
  - Complicated visual input
  - Lack of annotated instructions


*Orient yourself so that the umbrellas are to the right. Go straight and take a right at the first intersection. At the next intersection there should be an old-fashioned store to the left. There is also a dinosaur mural to the right. Touchdown is on the back of the dinosaur.*

# External Resource

- Google Street View
  - Street view images
  - Machine-generated instructions

**University of California, Santa Barbara**
Santa Barbara, CA 93106

Head southwest on CA-217 W/Ward Memorial Blvd
95 ft

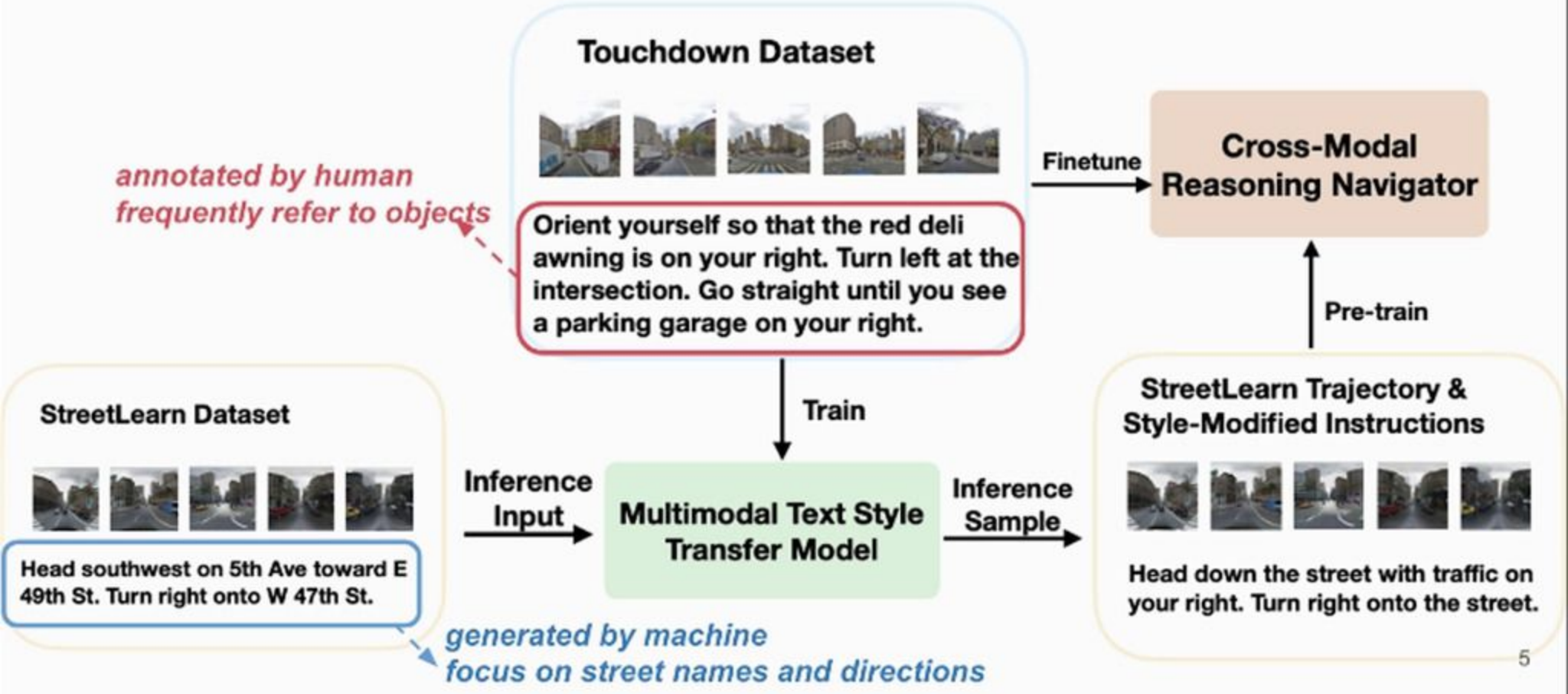
Slight left onto Ward Memorial Blvd
207 ft

At the traffic circle, take the 2nd exit onto Lagoon Rd
0.7 mi

Turn left
69 ft

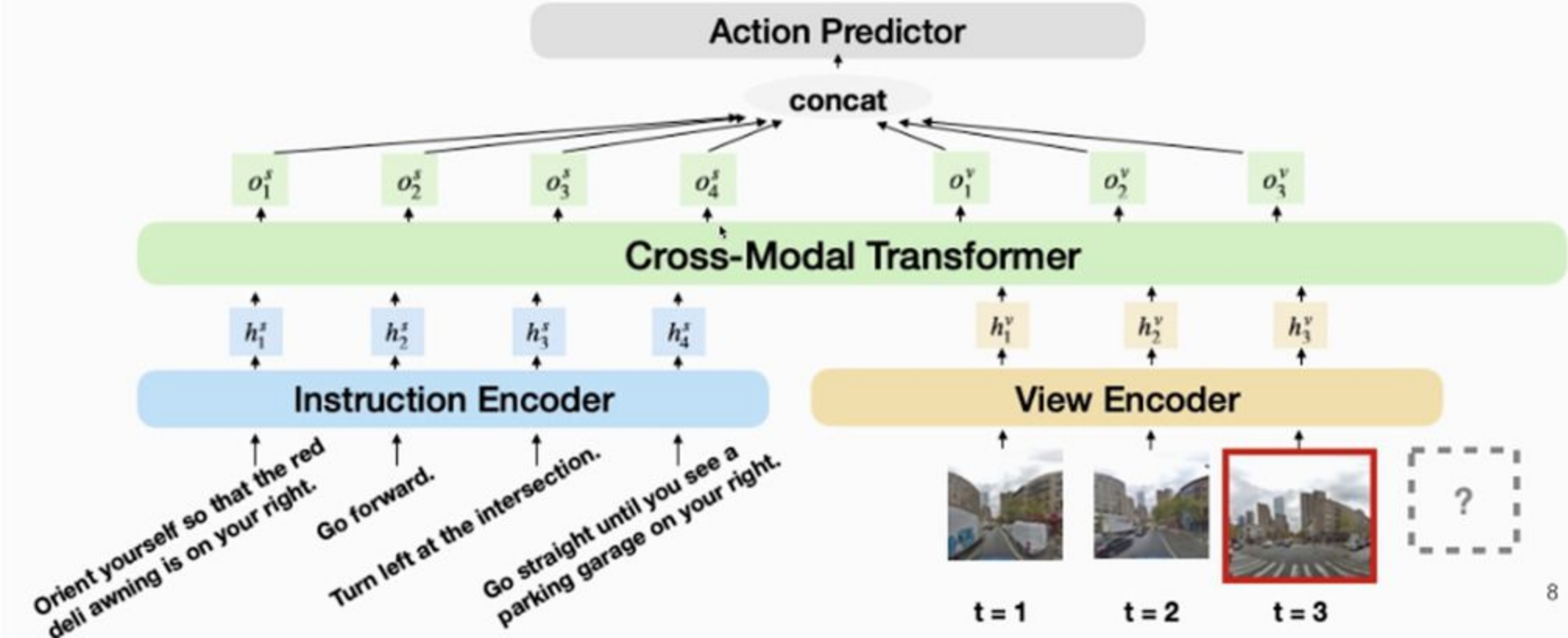
**Goleta Point**
California 93106

# Multimodal Text Style Transfer Framework Overlook



# Multimodal Text Style Transfer: Inference

# Cross Modal Reasoning Navigator



# Tasks & Datasets

- Task: Touchdown dataset [1]
- External resource: StreetLearn dataset [2]

- Datasets Comparison

| Dataset | Trajectory source | Instruction source |
|---|---|---|
| Touchdown | Google Street Views | Human-written |
| StreetLearn | Google Street Views | Google Map API |

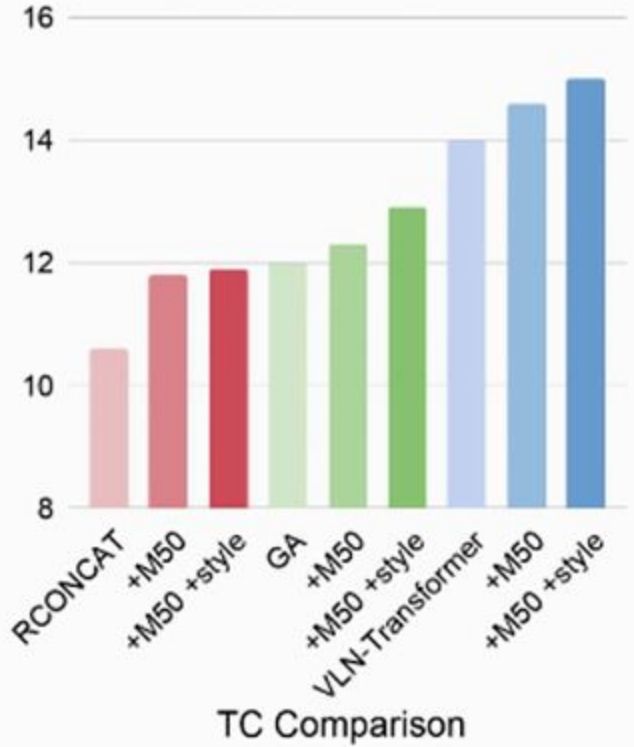
[1] Howard Chen et al., *TOUCHDOWN: Natural Language Navigation and Spatial Reasoning in Visual Street Environments*, CVPR 2019
[2] Piotr Mirowski et al., *Learning to Navigate in Cities Without A Map*, NeurIPS 2018

# Experiment Settings

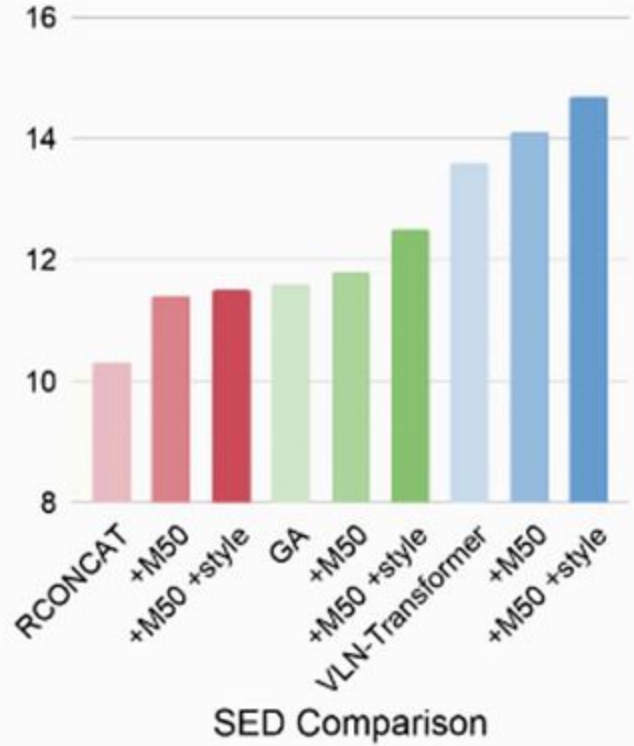
- Models:
  - RCONCAT
  - GA
  - VLN-Transformer (ours)
- Metrics
  - TC: task completion rate
  - SED: success weighted by edit distance
  - CLS: coverage weighted by length score

# Experiment Results

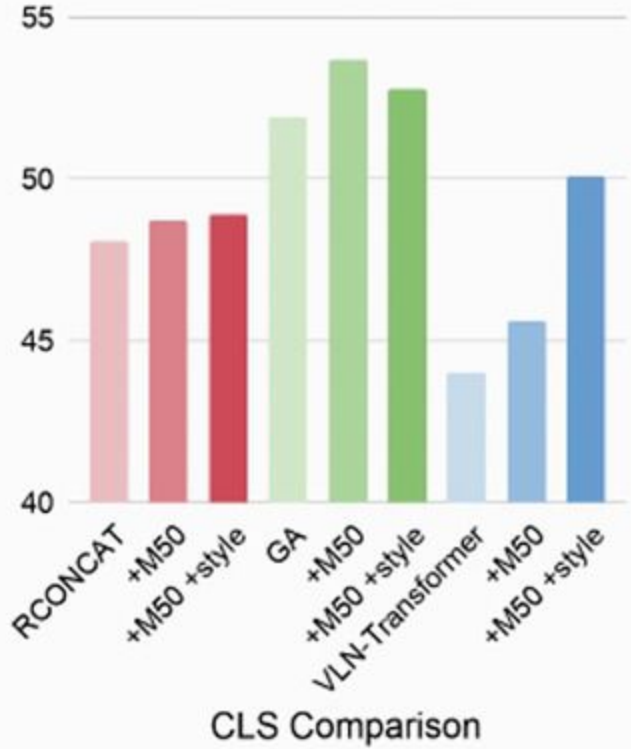
- ***+M50***: pre-train on a StreetLearn subset with machine-generated instructions
- ***+M50 +style***: pre-train on a StreetLearn subset with style-modified instructions



TC Comparison



SED Comparison



CLS Comparison

# Case Study

- Red tokens: contradictions with ground truth.
- Blue tokens: alignment with ground truth.

| | |
|---|---|
| StreetLearn | **Head northwest** on W 35th St toward Hudson Blvd E. **Turn right** at the 1st cross street onto Hudson Blvd E. |
| Original Speaker | Turn so the red construction is on your left and the red brick building is on your right. Go forward to the intersection and turn right. You'll have a red brick building with a red awning on your right. |
| Multimodal Text Style Transfer | Move forward with traffic on the right turn right at the light. Continue straight. |

# Thanks!

UC SANTA BARBARA UC SANTA CRUZ UC San Diego Google