

Multimodal Text Style Transfer for Outdoor VLN

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

[1]University of California, Santa Barbara

[2]University of California, San Diego, [3]Google

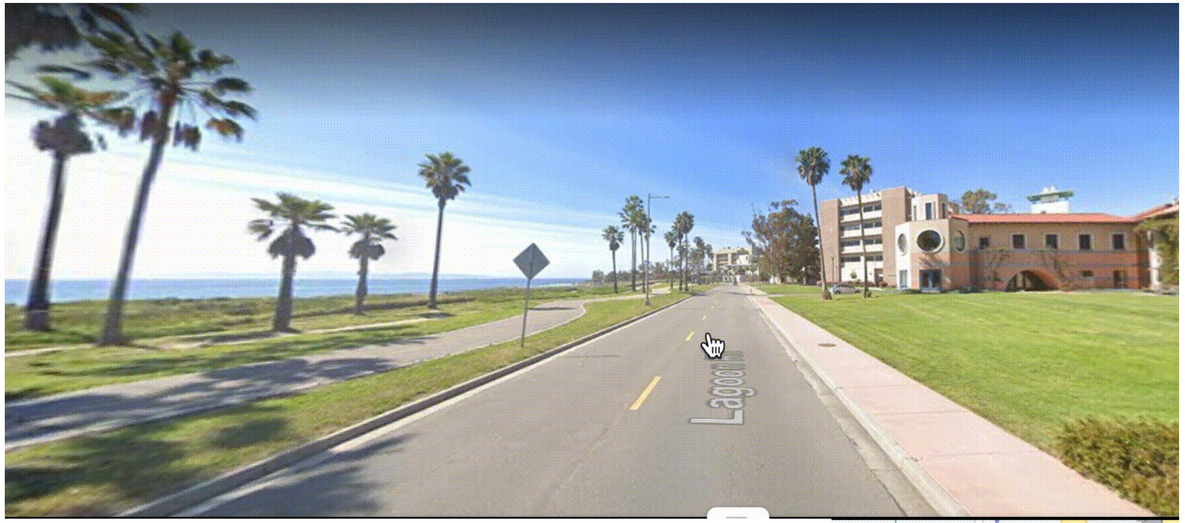
Outdoor Vision-and-Language Navigation

- **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 Views
- 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

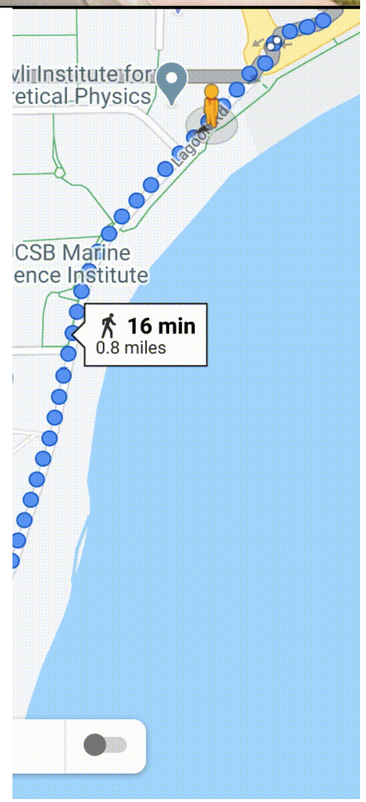
At the traffic circle, take the 2nd exit onto Lagoon Rd

0.7 mi

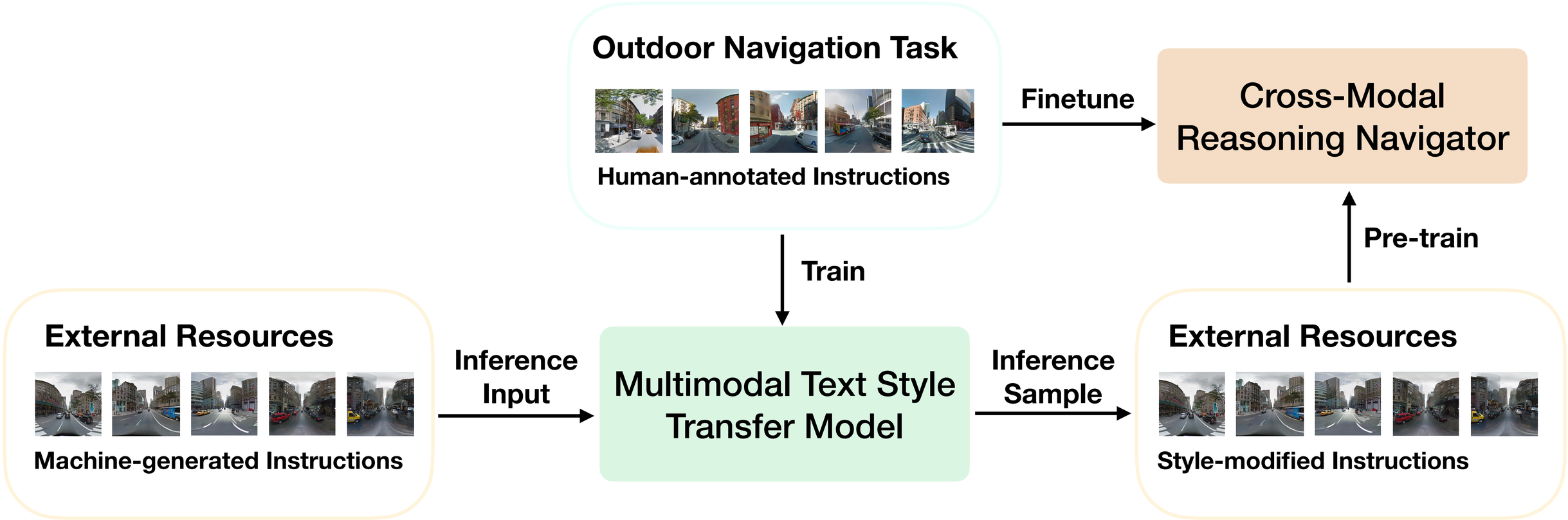
Turn left

69 ft

Goleta Point
California 93106



Multimodal Text Style Transfer Framework Overview



Multimodal Text Style Transfer: Training

Go straight. There will be a red wall to your right. Take a right. Stop at the intersection.

Masking

[MASK]. There will be a [MASK] to [MASK] right. Take a right. Stop at the [MASK].

Recovering

Multimodal Text Style Transfer Model

Go straight. There will be a red wall to your right. Take a right. Stop at the intersection.

Multimodal Text Style Transfer: Inference

Head southwest on 5th Ave toward E 49th St.
Turn right onto W 47th St.

Masking

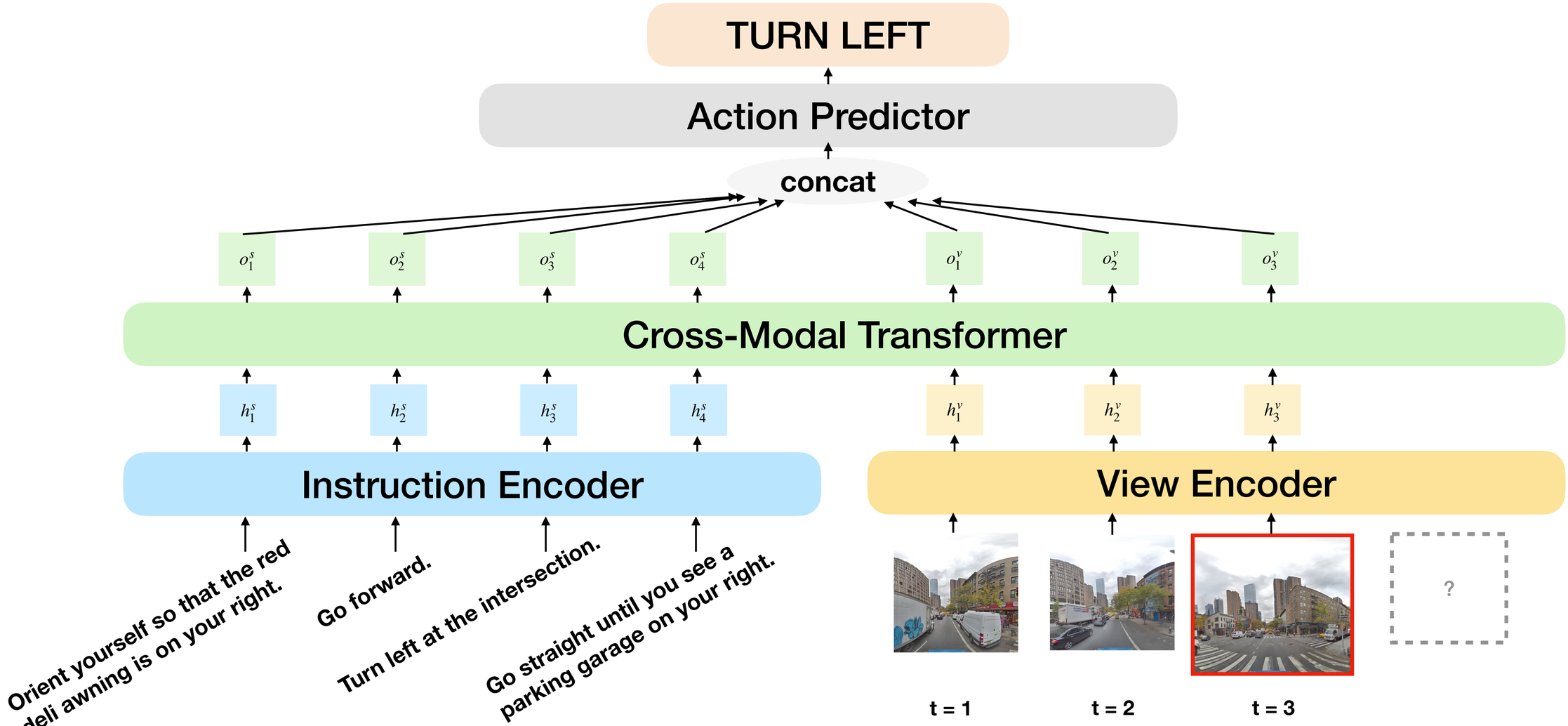
[MASK] on [MASK] toward [MASK].
[MASK] right onto [MASK].

Transferring Text Style

Multimodal Text Style Transfer Model

Head down the street with traffic on your right. Turn right onto the street.

Cross-Modal Reasoning Navigator



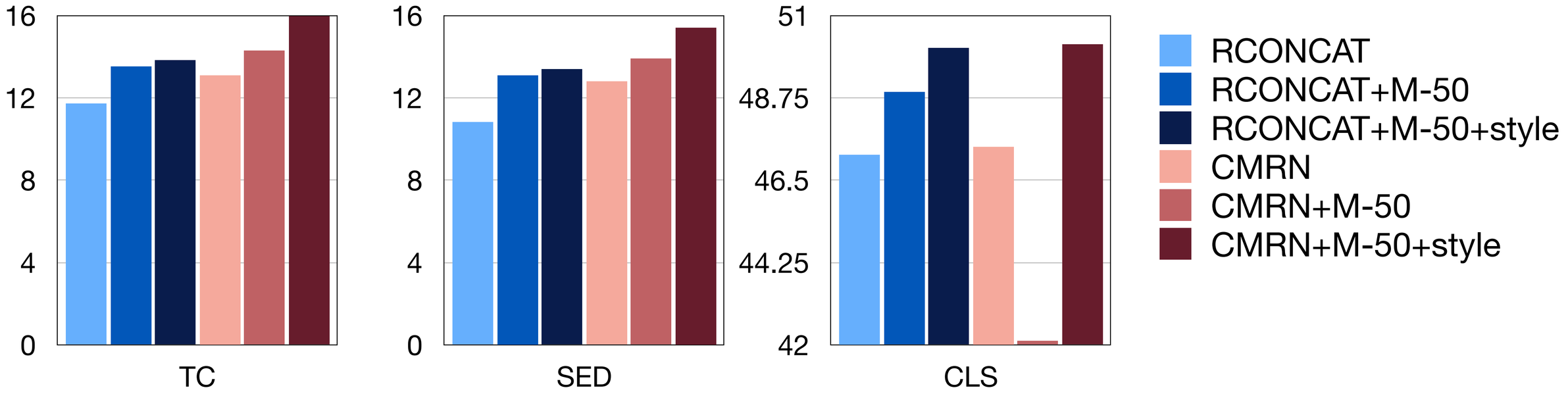
Task & Datasets

- **Task:** Touchdown dataset
- **External Resource:** StreetLearn dataset

Dataset	Trajectory Source	Instruction Source
Touchdown	Google Street Views	Human Annotator
StreetLearn	Google Street Views	Google Map API

Experimental Results

- **Baseline model:** RCONCAT
- **+M-50:** pre-train on a StreetLearn subset with machine-generated instructions
- **+M-50 +style:** pre-train on a StreetLearn subset with style-modified instructions
- **TC:** task completion rate
- **SED:** success weighted by edit distance
- **CLS:** a measurement of the fidelity of the agent's path with respect to reference path



Case Study

- **Baseline model:** Speaker
- **Red tokens:** contradictions with ground truth
- **Blue tokens:** alignments with ground truth

Ground Truth	**Head northwest** on W 35th St toward Hudson Blvd E. **Turn right** at the 1st cross street onto Hudson Blvd E.
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.

Google

Thanks!