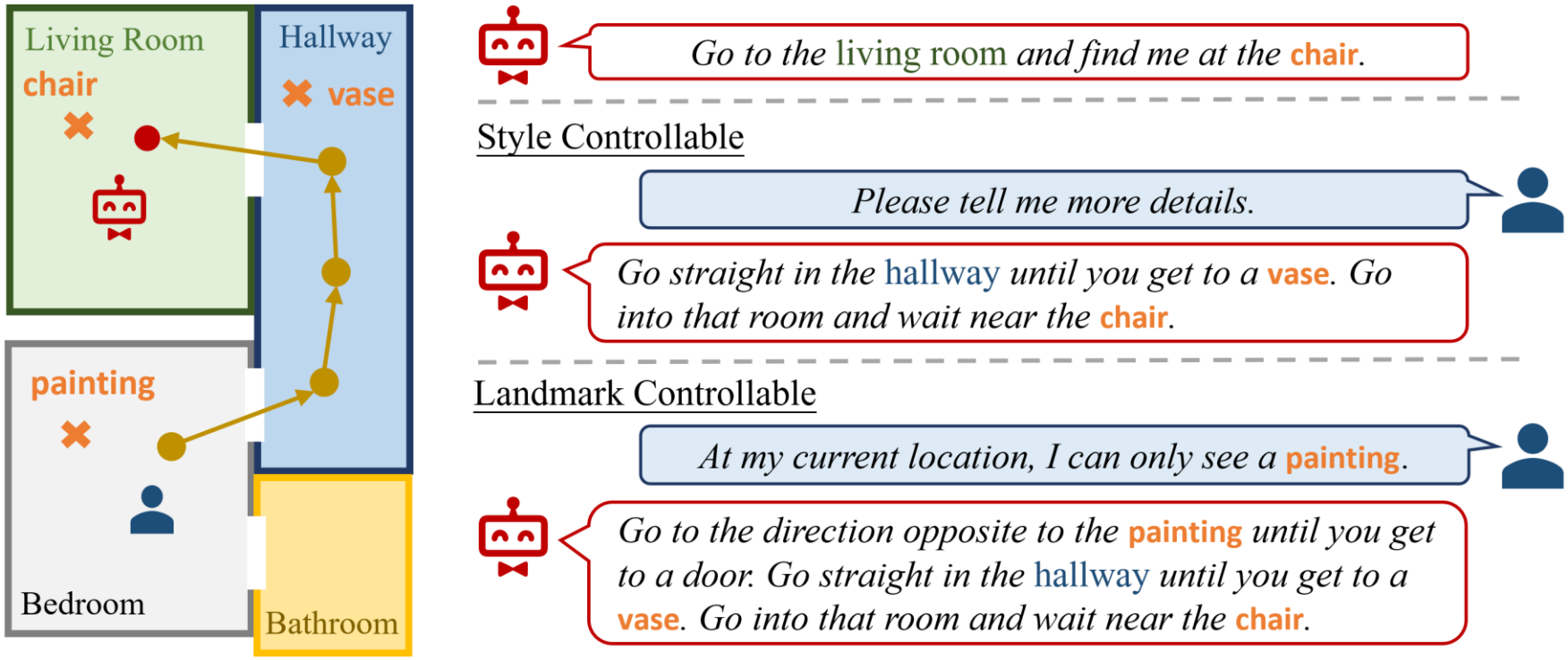


Controllable Navigation Instruction Generation with Chain of Thought Prompting

Xianghao Kong[1]*, Jinyu Chen[1]*, Wenguan Wang[2]†, Hang Su[3], Xiaolin Hu[3], Yi Yang[2], Si Liu[1]†

[1] School of Artificial Intelligence, Beihang University

[2] Zhejiang University, [3] Tsinghua University

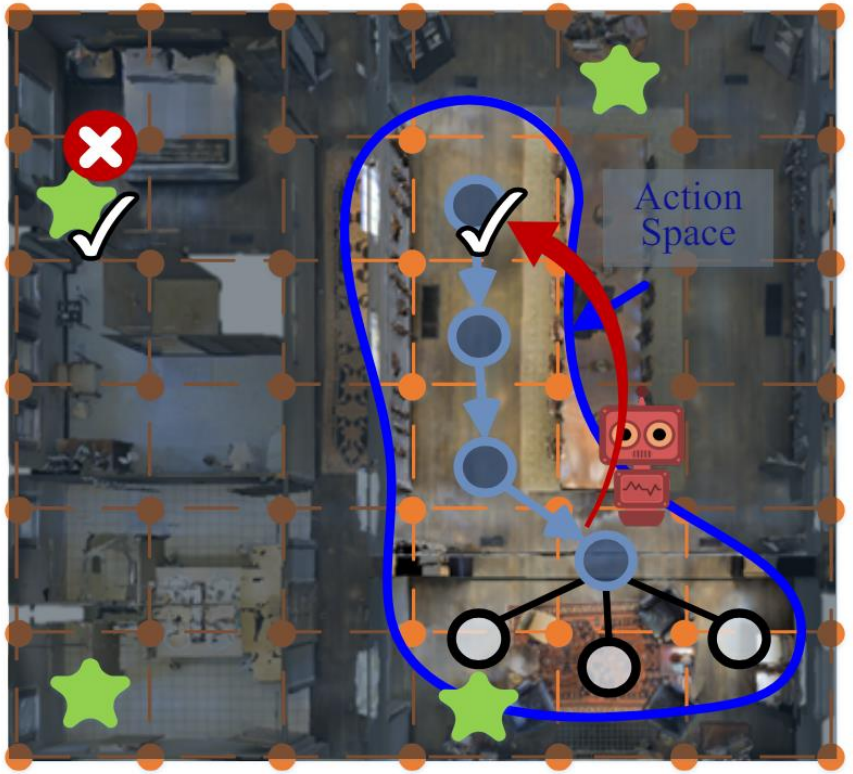
Problem Setting: Vision-Language Navigation

Follower

- **Input**
  - Current panoramic view and position
  - Instruction
  - Navigable views
- **Output**
  - Select an action view from navigable views

Speaker (Our Focus)

- **Input**
  - Panoramic views, positions, and action views of each step along a trajectory
- **Output**
  - Navigation instruction



Instruction: Get out of the dining room and walk into the bedroom. Wait at the window beside the bed.

Motivation

Enhancing Executability and Controllability of Instruction Generation Models

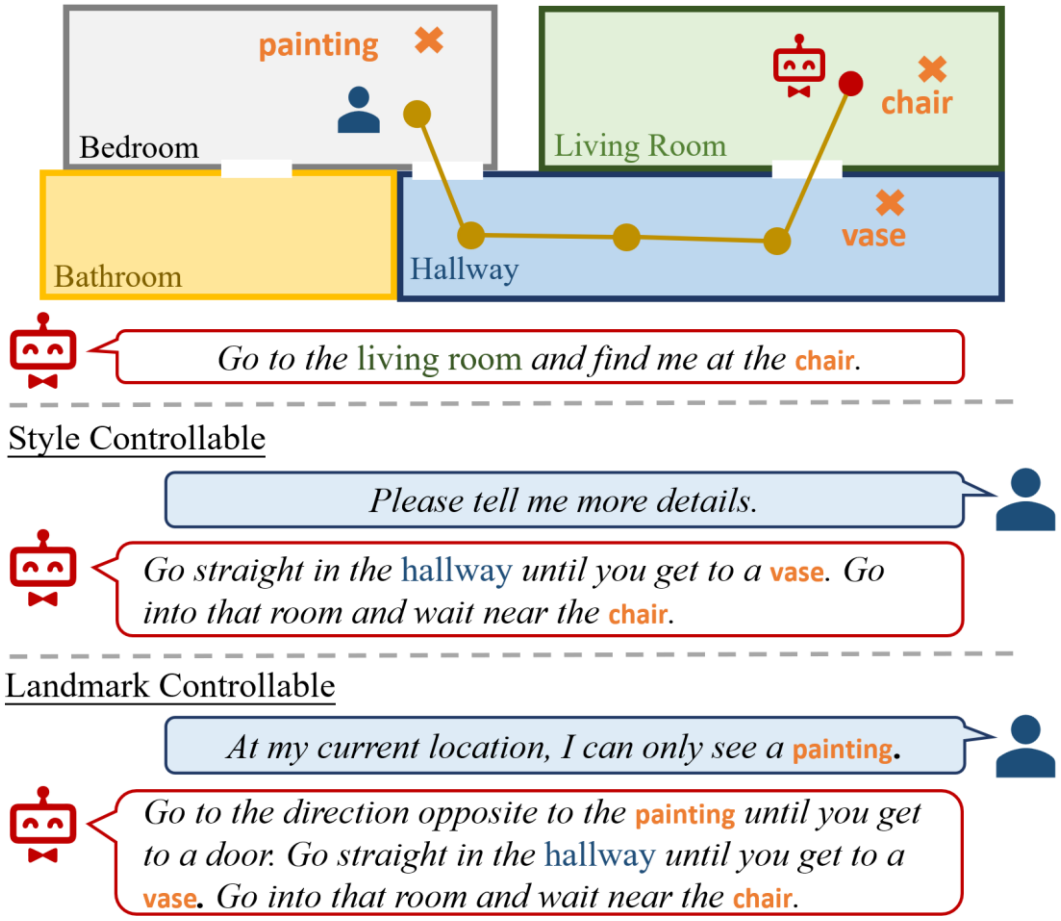
- **Executability**
  - High linguistic quality
  - Clear guidance at navigational junctions
- **Controllability**
  - **Style Control**: Whether the instruction recipient is acquainted with the environment → level of abstraction
  - **Content Control**: What type of landmarks the instruction follower focuses on

- **Existing Methods**: No Controllability
  - One model instance, one single style
  - Cannot specify landmarks

Solution

Controllable Navigation Instructor (C-Instructor)

- **Adapter Structure**
  - Incorporate path information into LLM
- **Chain-of-Thought with Landmarks Mechanism**
  - Better Executability: Identifying crucial landmarks before generating full instructions
  - Provide Content Controllability: Modifying landmarks
- **Spatial Topology Modeling Task**
  - Better Executability: Incorporating spatial connectivity prediction to help understand environment layout
- **Style-Mixed Training Policy**
  - Generate instructions in various styles within a single model instance
  - Provide Style Controllability: Prompts as differentiation



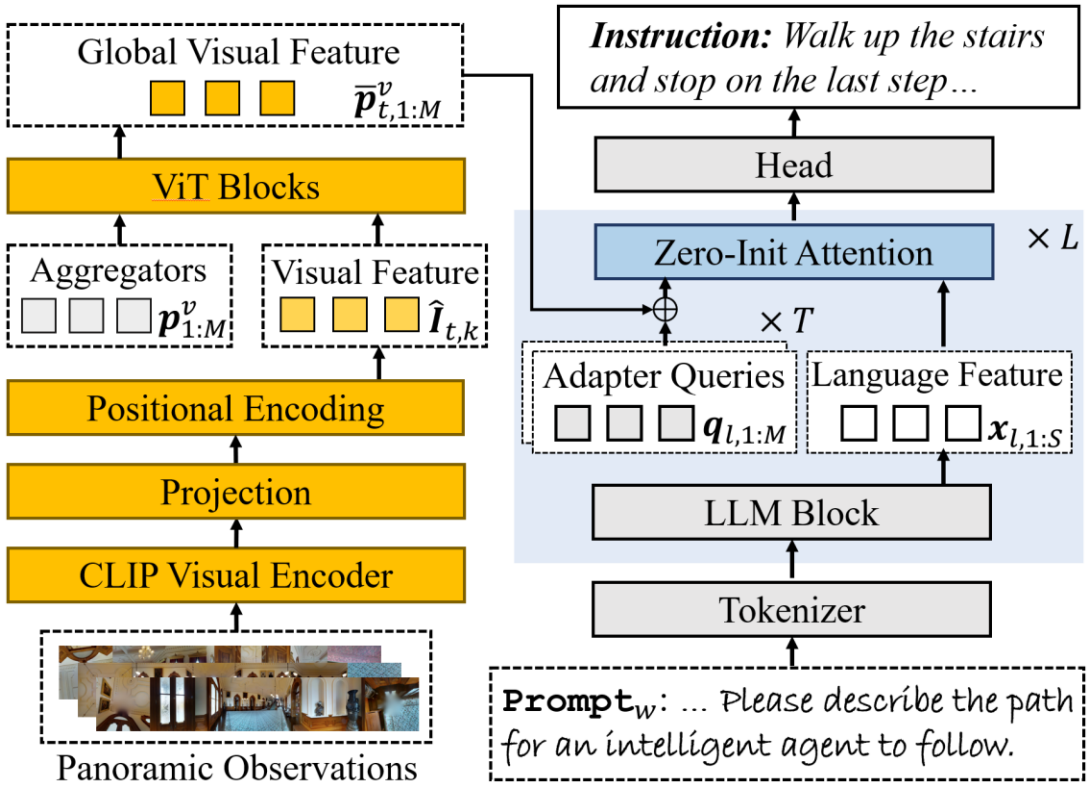
Overall Framework

Trajectory Encoder (Left)

- **Encode** each sub-view with CLIP visual encoder
- Add spatial and temporal **PE**, add action PE
- **Aggregate** global visual feature of each panoramic observations using ViT blocks

LLM Adapter (Right)

- Add **adapter queries** of each layer with global visual feature, then do linear projection
- Conduct **zero-initialized attention** with language feature to inject trajectory information



Chain of Thought with Landmarks

Landmark Selection

- **Linguistic Landmarks**: nouns in annotated instructions
- **Visual Landmarks**: selected according to spatial and temporal importance

Spatial Selection: Objects distinct in action view

Temporal Selection: Action leads to new scene

CoT Training and Inference

- **Step 1**: Identify landmarks
- **Step 2**: Generate full instruction

CoT Inference

prompt$_\lambda$: You are given a sequence of views of a path. Please extract critical landmarks in the path.
Landmarks **:** ***< stair, living room, coffee table >***

⬇

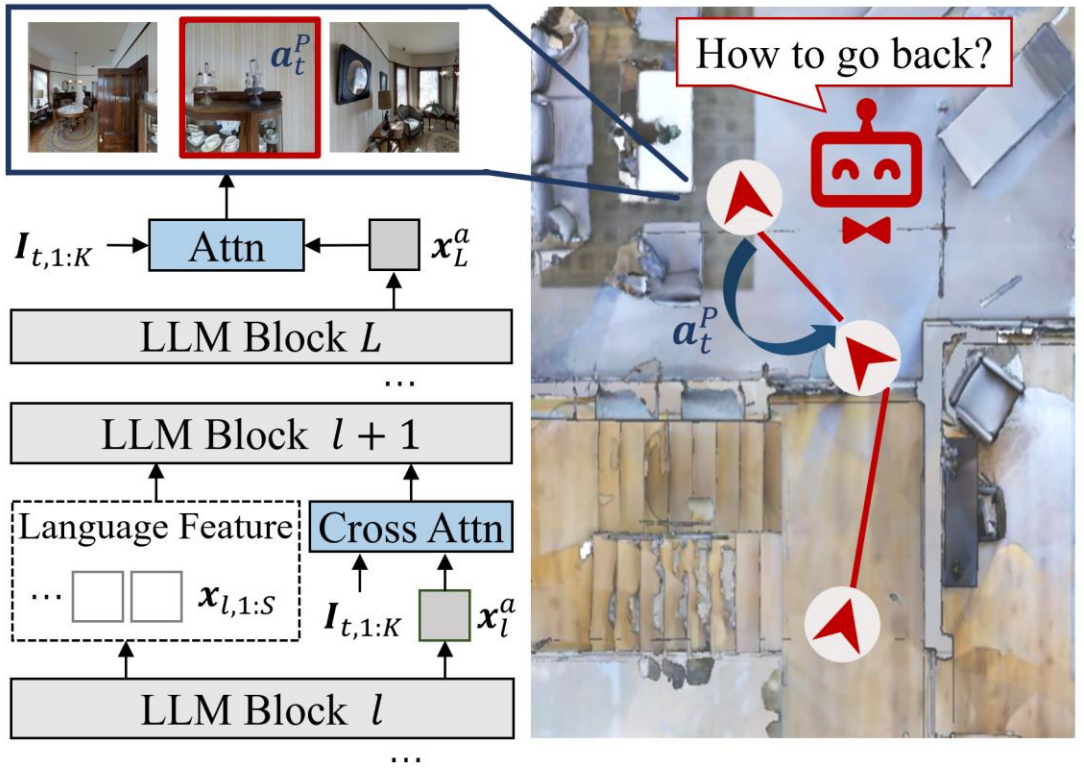
prompt$_w$: You are given a sequence of views of a path in an indoor environment. Please describe the path according to the given landmarks in details for an intelligent agent to follow. ***<Landmarks>***
Instruction*:** Go down the ***stairs and take a right. Go into the ***living room***. Stop next to the ***coffee table***.

Spatial Topology Modeling Task

Enhancing Spatial Perception Capability

- LLMs and visual encoders are typically trained on Internet data with **few embodied-type data**
- Understanding **spatial relationships** is essential for guiding agents

- Predict how to **return** to the previous viewpoint
  - Can be done on a random trajectory without instruction annotation
- Introduce a special token for action prediction
- Inject view information through cross attention



Comparison in Text Similarity Metrics

Assessing the Similarity between Generated Instructions and Human Annotations

- **4 datasets with different linguistic styles**
  - 3 indoor datasets, 1 outdoor dataset
- **SOTA performance**

Methods	R2R val seen						R2R val unseen					
	SPICE↑	BLEU-1↑	BLEU-4↑	CIDEr↑	Meteor↑	Rouge↑	SPICE↑	BLEU-1↑	BLEU-4↑	CIDEr↑	Meteor↑	Rouge↑
BT-speaker [13] [NeurIPS2018]	0.173	0.670	0.236	0.373	0.209	0.443	0.113	0.600	0.149	0.113	0.167	0.376
EDrop-speaker [41] [NAACL2019]	0.168	0.660	0.228	0.362	0.208	0.447	0.117	0.590	0.157	0.160	0.174	0.389
CCC-speaker [47] [CVPR2022]	0.194	0.698	0.265	0.449	0.218	0.467	0.108	0.591	0.139	0.120	0.164	0.375
Lana [49] [CVPR2023]	0.170	0.657	0.215	0.265	0.205	0.433	0.174	0.667	0.236	0.295	0.213	0.448
C-Instructor *w/o* SMT	0.230	**0.732**	0.270	0.511	0.237	0.475	**0.217**	**0.715**	0.263	**0.453**	0.234	0.470
C-Instructor (Ours)	**0.233**	0.726	**0.276**	**0.529**	**0.247**	**0.480**	0.212	0.713	**0.266**	0.447	**0.239**	**0.473**

Methods	RxR val seen					RxR val unseen				
	BLEU-1↑	BLEU-4↑	CIDEr↑	Meteor↑	Rouge↑	BLEU-1↑	BLEU-4↑	CIDEr↑	Meteor↑	Rouge↑
BT-speaker [13] [NeurIPS2018]	0.514	0.188	0.026	0.204	0.365	0.566	0.211	0.024	0.208	0.372
EDrop-speaker [41] [NAACL2019]	0.595	0.197	0.047	0.213	0.378	0.568	0.184	0.038	0.205	0.370
CCC-speaker [47] [CVPR2022]	0.526	0.194	0.024	0.185	0.355	0.518	0.187	0.026	0.184	0.353
Lana [49] [CVPR2023]	0.342	0.123	0.040	0.128	0.275	0.319	0.115	0.043	0.124	0.273
C-Instructor *w/o* SMT	0.683	0.233	0.081	**0.243**	0.381	0.667	0.224	**0.080**	0.236	0.379
C-Instructor (Ours)	**0.685**	**0.234**	**0.082**	0.238	**0.382**	**0.678**	**0.233**	0.077	**0.239**	**0.382**

Methods	REVERIE val seen						REVERIE val unseen					
	SPICE↑	BLEU-1↑	BLEU-4↑	CIDEr↑	Meteor↑	Rouge↑	SPICE↑	BLEU-1↑	BLEU-4↑	CIDEr↑	Meteor↑	Rouge↑
BT-speaker [13] [NeurIPS2018]	0.121	0.693	0.347	0.269	0.223	0.602	0.103	0.664	0.302	0.190	0.200	0.569
EDrop-speaker [41] [NAACL2019]	0.138	0.641	0.360	0.523	0.277	0.597	0.114	0.648	0.319	0.333	0.233	0.546
CCC-speaker [47] [CVPR2022]	0.144	0.727	0.408	0.502	0.272	0.589	0.115	0.681	0.357	0.334	0.232	0.548
Lana [49] [CVPR2023]	0.137	0.707	0.404	0.627	0.282	0.619	0.107	0.696	0.345	0.327	0.239	0.582
C-Instructor *w/o* SMT	**0.184**	**0.785**	**0.480**	**0.844**	**0.319**	**0.649**	0.139	0.739	0.369	0.464	0.259	0.577
C-Instructor (Ours)	0.182	0.775	0.459	0.805	0.311	0.647	**0.141**	**0.754**	**0.419**	**0.545**	**0.267**	**0.591**

Methods		UrbanWalk				
		SPICE ↑	BLEU-1 ↑	BLEU-4 ↑	Meteor ↑	Rouge ↑
BT-speaker [13]	[NeurIPS2018]	0.524	0.649	0.408	0.350	0.620
EDrop-speaker [41]	[NAACL2019]	0.531	0.689	0.435	0.358	0.634
ASSISTER [19]	[ECCV2022]	0.451	0.576	0.164	0.319	0.557
Kefa-speaker [52]	[Arxiv2023]	0.566	0.711	0.450	0.378	0.655
C-Instructor	(Ours)	**0.645**	**0.771**	**0.534**	**0.461**	**0.781**

Diagnostic Experiments

Comparing the Full Model with Several Ablative Designs

- **Vanilla LLM**
  - Caption (BLIP) then generate (LLaMA), with fine-tuning
- **Baseline with our basic structure to inject path information**
- **Baseline + proposed modules**

Methods	REVERIE val unseen					R2R val unseen				
	BLEU-1 ↑	BLEU-4 ↑	CIDEr ↑	Meteor ↑	Rouge ↑	BLEU-1 ↑	BLEU-4 ↑	CIDEr ↑	Meteor ↑	Rouge ↑
Vanilla LLM	0.399	0.131	0.432	0.156	0.400	0.307	0.059	0.292	0.139	0.303
Baseline	0.648	0.308	0.347	0.248	0.547	0.676	0.232	0.356	0.225	0.449
Baseline + SMT	0.679	0.344	0.397	0.254	0.562	0.685	0.254	0.407	0.233	0.466
Baseline + SMT + STMT	0.737	0.402	0.490	0.258	0.590	0.689	0.262	0.445	0.228	**0.479**
Baseline + SMT + STMT + CoTL	**0.754**	**0.419**	**0.545**	**0.267**	**0.591**	**0.713**	**0.266**	**0.447**	**0.239**	0.473

Instruction Quality Analysis

Assessing the Semantic Alignment between Instructions and Trajectories

- **Path Guiding Proficiency**
  - Regenerate instructions on validation splits to guide followers
- **Data Augmentation**
  - Generate instructions on new trajectories sampled from training scenes as additional training data

Data Source	REVERIE val unseen			
	SR↑	SPL↑	RGS↑	RGSPL↑
Original [38]	32.95	30.20	18.92	17.28
+BT-speaker [13]	31.84	28.37	17.35	15.14
+EDrop-speaker [41]	30.45	27.18	18.60	16.24
+CCC-speaker [47]	29.65	26.20	16.33	14.58
+Lana [49]	33.05	29.76	19.14	17.20
+**C-Instructor** (Ours)	**34.25**	**31.25**	**19.99**	**18.08**

Instruction Generator	Follower			
	HAMT [7]		DUET [8]	
	SR↑	SPL↑	SR↑	SPL↑
Human annotation [38]	32.95	30.20	46.98	33.73
BT-speaker [13]	24.85	21.74	30.47	21.46
EDrop-speaker [41]	26.19	23.55	27.89	17.00
CCC-speaker [47]	23.29	20.69	29.74	19.55
Lana [49]	26.84	24.38	31.39	20.44
C-Instructor (Ours)	**31.35**	**29.27**	**43.34**	**30.13**

- **User Study**
  - Ask volunteers to rate the semantic alignment between instructions and trajectories

Qualitative Results

Demonstrating the Controllability of C-Instructor


R2R Style

Annotation: Walk past the sitting area and wait in the kitchen by the island.

Generation: < ***couch, chairs, kitchen*** > Walk straight past the ***couch*** and ***chairs***. Stop in front of the ***kitchen***.

LM. Control: < ***hallway, sofa, ~~couch~~, ~~chairs~~, ~~kitchen~~, island*** > Walk down the ***hallway*** past the sectional ***sofa*** and stop by the *dining room* ***island***.

REVERIE Style

Annotation: Go into the kitchen and clean the table nearest the couch.

Generation: < ***kitchen, chair, stool*** > Go to the ***kitchen*** and pull out the ***chair*** closest to the ***stool***.

LM. Control: < ***kitchen, counter, ~~chair~~, stool*** > Go to the ***kitchen*** on level 1 and clean the ***counter*** with the ***stools***.

Summary

Controllable Navigation Instructor (C-Instructor)

- **Controllability**
  - Style
  - Content (Landmark)
- **High Linguistic Quality**
- **Key Contributions**
  - Adapter Structure
  - Chain-of-Thought with Landmarks Mechanism
  - Spatial Topology Modeling Task
  - Style-Mixed Training Policy
- **Experiments**
  - Text Similarity Metrics
  - Follower Experiments
  - User Study