

Robust Cross-modal Alignment Learning for Cross-Scene Spatial Reasoning and Grounding

Yanglin Feng Hongyuan Zhu Dezhong Peng Xi Peng Xiaomin Song Peng Hu*



Overview of Our Work



- ❖ Overview of Our Work
- ❖ Task: Cross-Scene Spatial Reasoning and Grounding
- ❖ Baseline: Cross-Scene 3D Object Reasoning Framework
 - Robust Text-Scene Aligning module (RTSA)
 - Tailored Word-Object Associating module (TWOA)
- ❖ Dataset: CrossScene-RETR
- ❖ Experiments



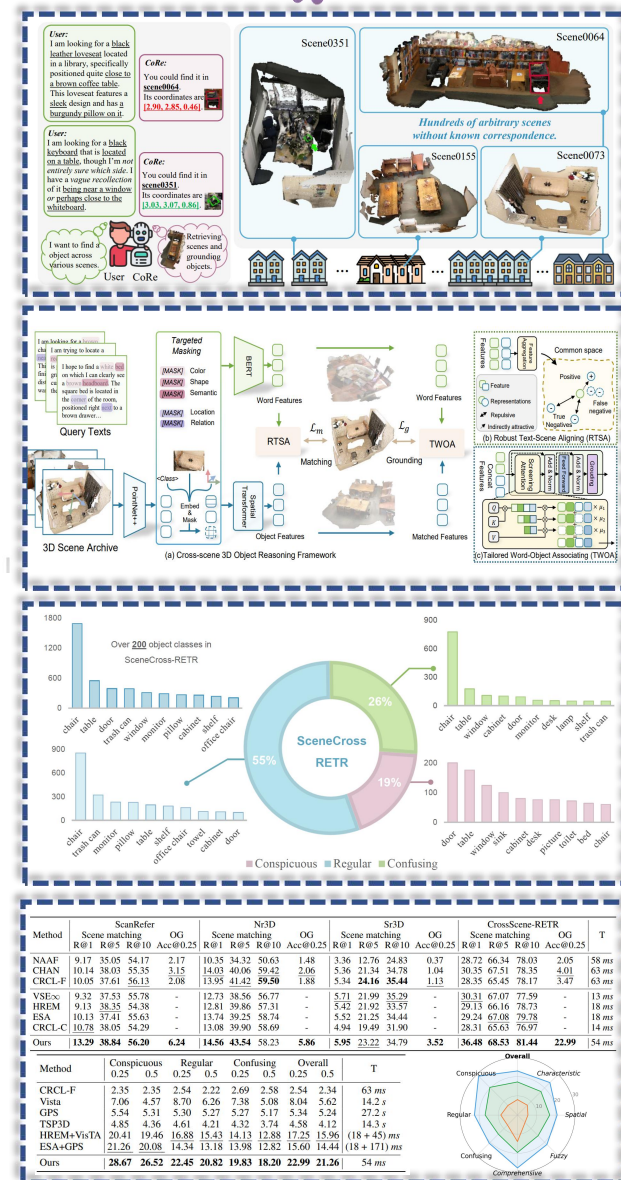


Overview of Our Work



Overview of Our Work

- ❖ We extend the 3D visual grounding task to the more general Cross-Scene Spatial Reasoning and Grounding (CSSRG) task, which aims to ground a described object anywhere across an entire collection of 3D scenes instead of predetermined scenes.
- ❖ We propose the novel two-stage Cross-Scene 3D Object Reasoning Framework (CoRe), following a matching-then-grounding paradigm to effectively mitigate computational costs. CoRe includes a Robust Text-Scene Aligning module (RTSA) for robust scene matching and a Tailored Word-Object Associating module (TWOA) for object grounding.
- ❖ We present the CrossScene-RETR dataset to facilitate complex cross-modal spatial alignment in the data aspect, offering a comprehensive evaluation for CSSRG.
- ❖ Extensive experiments on four multimodal datasets demonstrate the superiority and effectiveness of our CoRe in CSSRG, remarkably outperforming state-of-the-art baselines.



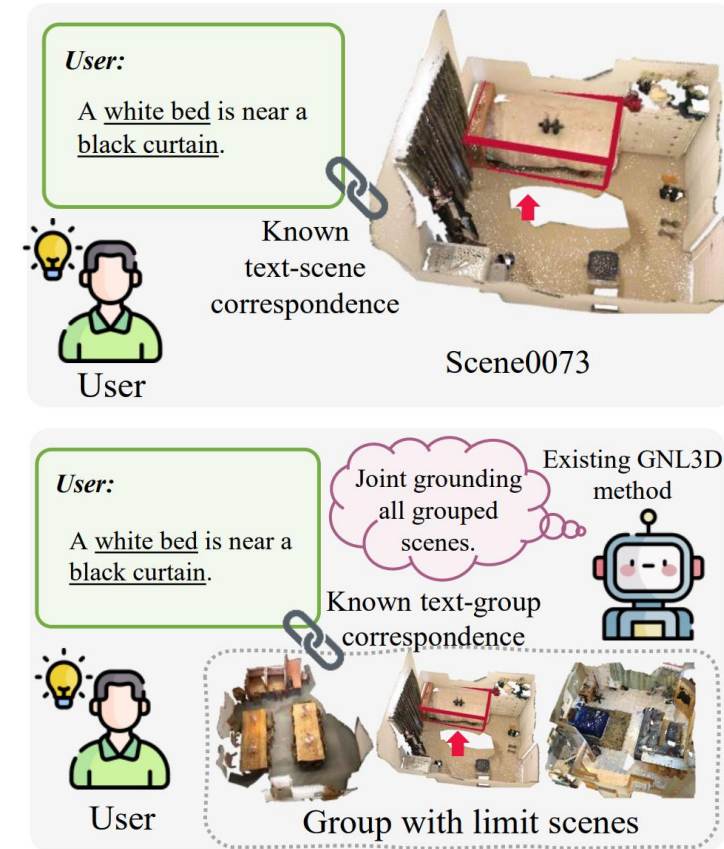
Task:

Cross-Scene Spatial Reasoning and Grounding



Motivation

- ❖ Existing 3DVG methods often assume the availability of predefined corresponding scenes.
- ❖ Even the latest multi-scene approaches still assume that the text-group correspondence is known.



If cross-scene object reasoning can be achieved, it would serve as a technical foundation for building-scale indoor navigation and task planning, thereby enabling broader applications in smart homes and robotics.

A Novel Task!

New Practical Task:

Cross-Scene Spatial
Reasoning and Grounding

New Challenging Dataset:

CrossScene-RETR

Novel Baselines:

CoRe

User:

I am looking for a black leather loveseat located in a library, specifically positioned quite close to a brown coffee table. This loveseat features a sleek design and has a burgundy pillow on it.

CoRe:

You could find it in scene0064.

Its coordinates are [2.90, 2.85, 0.46].



User:

I am looking for a black keyboard that is located on a table, though I'm *not entirely sure which side*. I have a *vague recollection* of it being near a window or perhaps close to the whiteboard.

CoRe:

You could find it in scene0351.

Its coordinates are [3.03, 3.07, 0.86].



I want to find a
object across
various scenes.



User CoRe



Retrieving
scenes and
grounding
objects.

Scene0351



Scene0064



*Hundreds of arbitrary scenes
without known correspondence.*

Scene0155



Scene0073



...



...



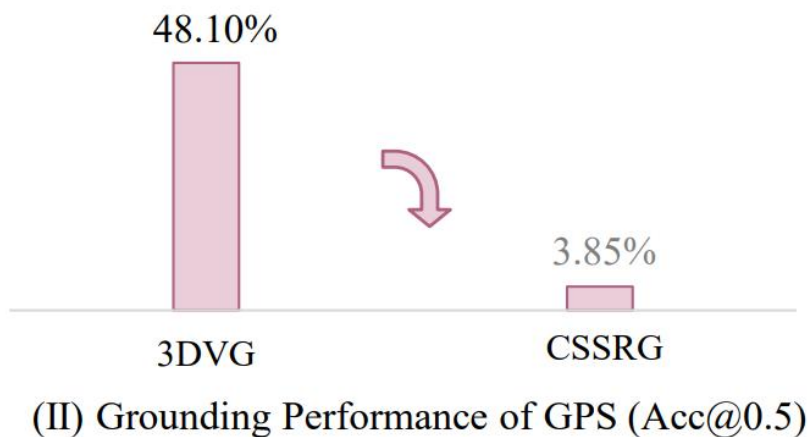
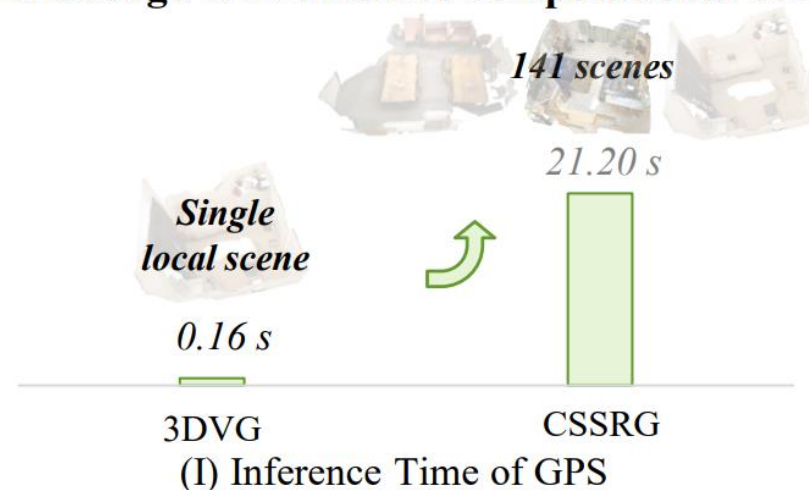
...



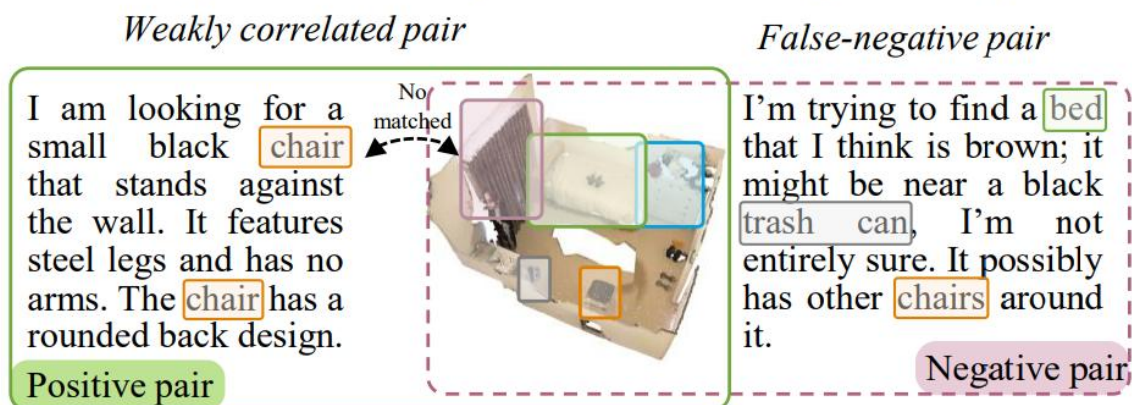
We propose a
new task:
CSSRG, which
empowers text-
to-3D with
cross-scene
alignment and
reasoning
capabilities.

Challenges

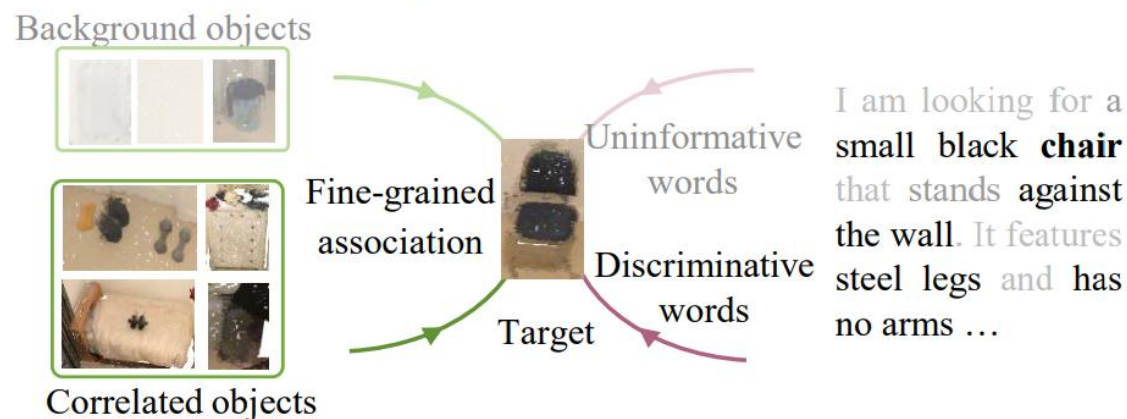
Challenge 1: Prohibitive computational costs.



Challenge 2: More complex cross-modal spatial alignment.



(I) Partial alignment between text and scene

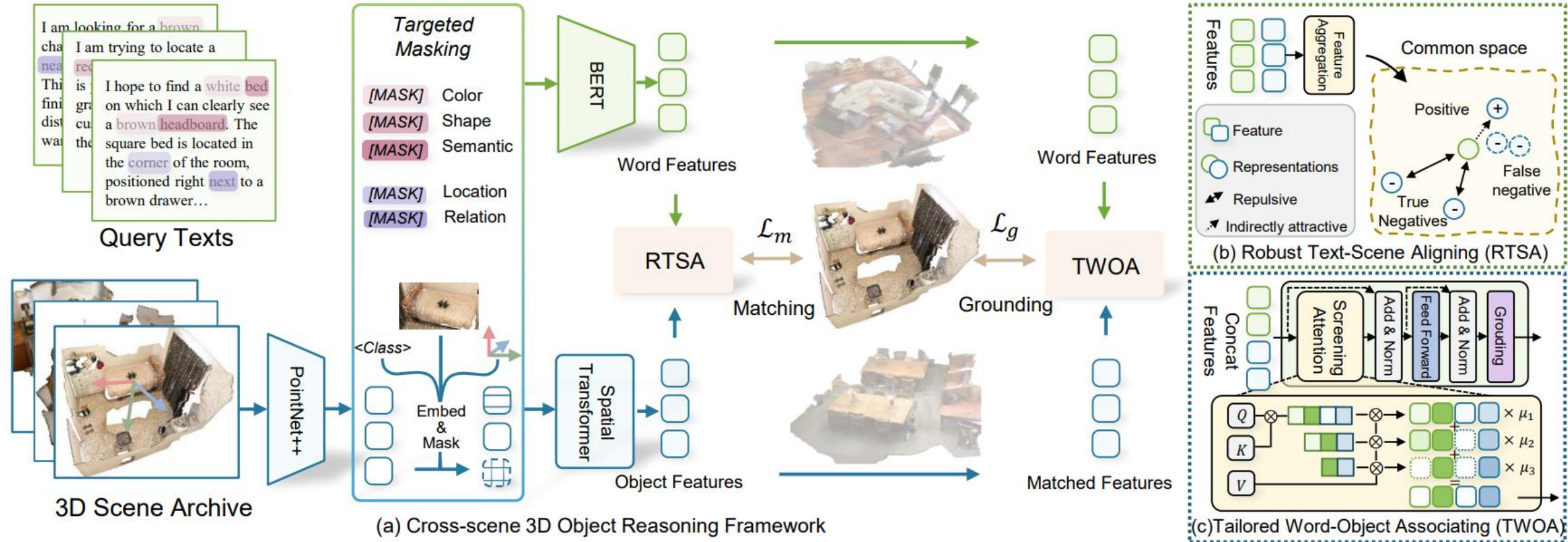


(II) Alignment between text and target object

Baseline:

**Cross-Scene 3D Object Reasoning Framework
(CoRe)**





Text-Scene Aligning: RTSA module → Word-Object Associating: TWOA module

Matching-then-Grounding Pipeline



Dataset:

CrossScene-RETR

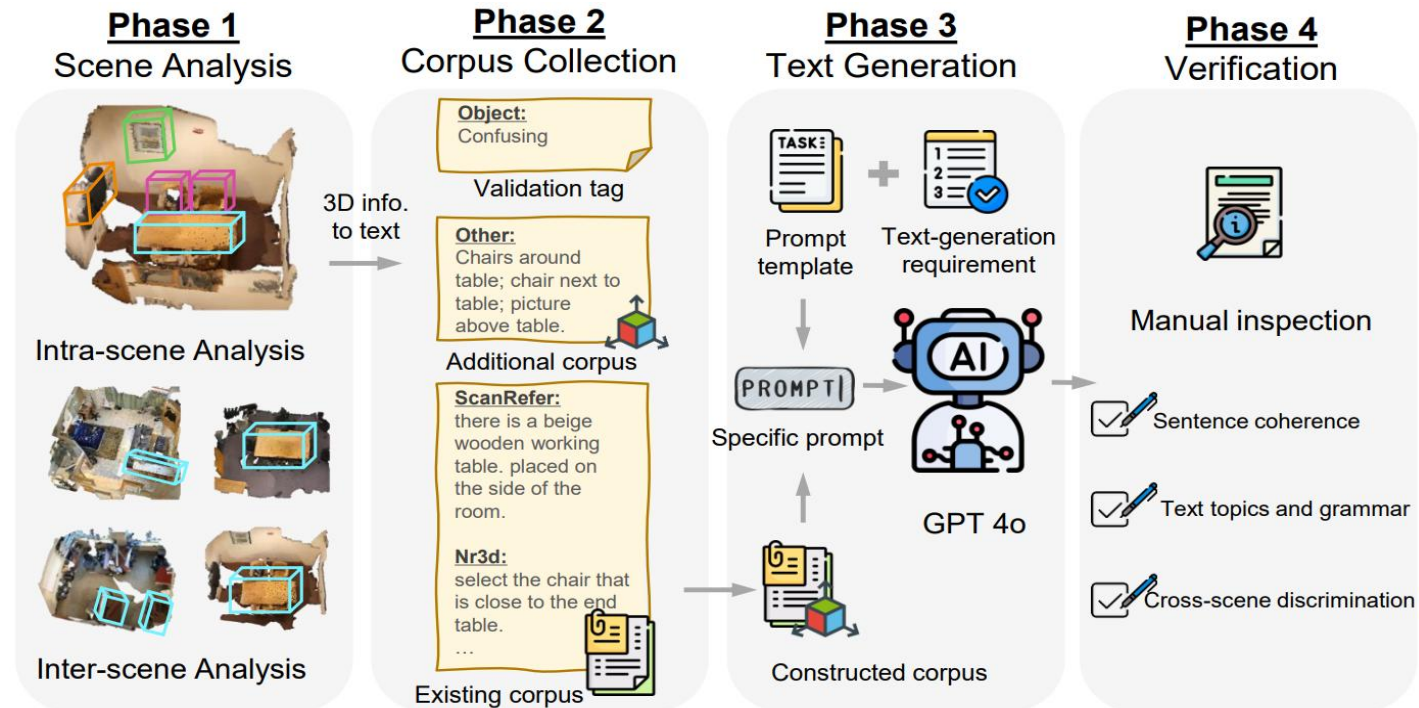


CrossScene-RETR

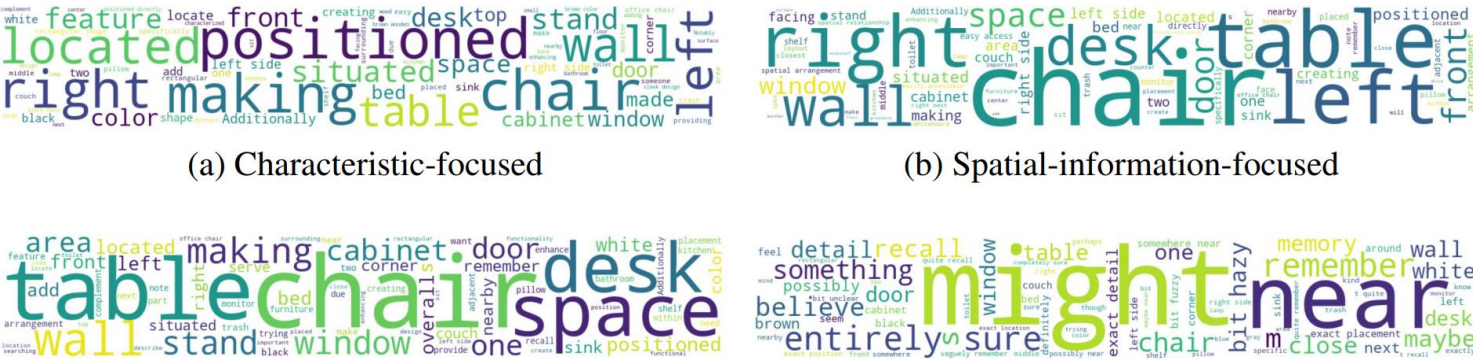
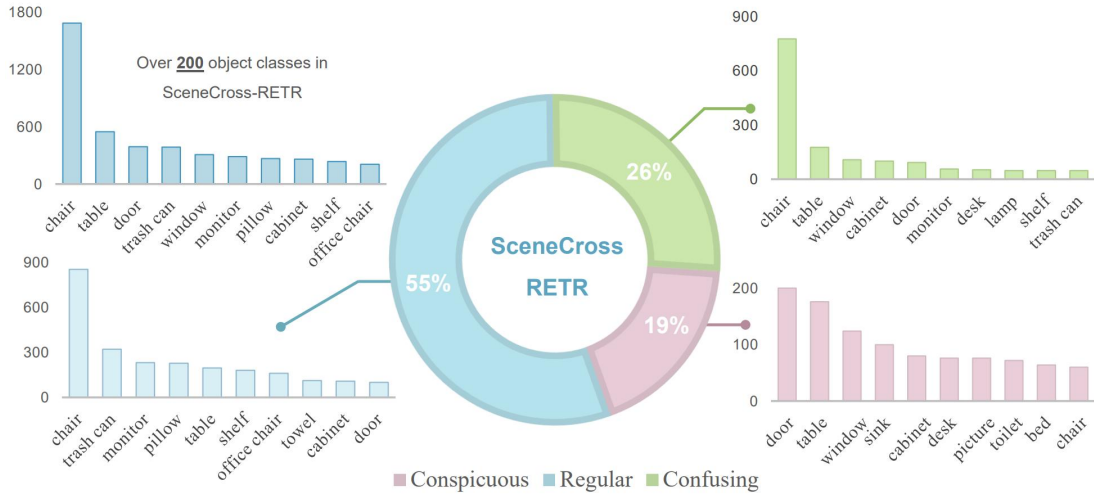
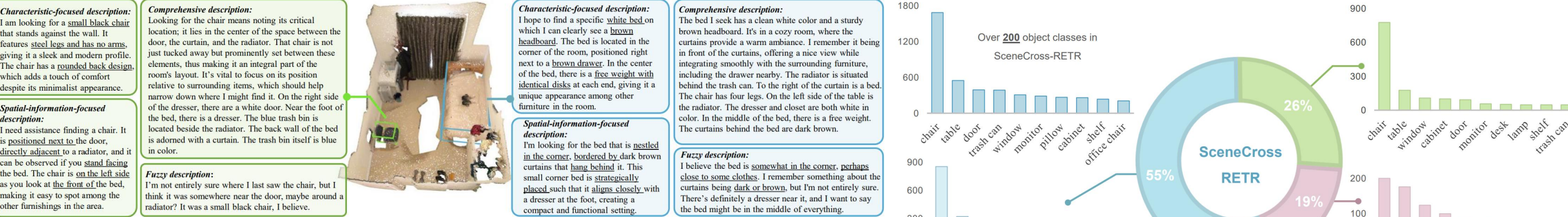
The descriptions in existing 3DVG datasets undoubtedly intensify the challenges and complexity of the cross-modal spatial alignment.



Construct pipeline:



CrossScene-RETR



Statistical indicators		ScanRefer	Nr3D	Sr3D	RETR
Overall	Average length	17.9	11.4	9.7	77.7
	Number of samples	46,173	41,503	83,572	39,526
	Vocabulary size	6,919	6,951	196	22,485
	Free-form	✓	✓	×	✓
Richness of description	Number of objects per text	1.8	1.7	1.8	11.4
	Number of characteristics per text	1.5	1.6	0.0	5.9
	Number of Spatial info. per text	1.2	1.1	0.6	6.3
	Number of info. points per text	4.5	4.4	2.4	23.2
	Text with Spatial info. (%)	69.8	47.5	55.6	97.5
	Text with color description (%)	58.2	29.7	0.0	81.2
	Text with shape description (%)	20.3	6.5	0.0	45.0
	Text with material description (%)	13.0	2.1	0.7	38.5



Experiment



Comparative Experiments

Performance comparison with existing cross-modal matching baselines on the ScanRefer, Nr3D, Sr3D, and CrossScene-RETR datasets.

Method	ScanRefer				Nr3D				Sr3D				CrossScene-RETR				T
	Scene matching			OG	Scene matching			OG	Scene matching			OG	Scene matching			OG	
	R@1	R@5	R@10	Acc@0.25	R@1	R@5	R@10	Acc@0.25	R@1	R@5	R@10	Acc@0.25	R@1	R@5	R@10	Acc@0.25	
NAAF	9.17	35.05	54.17	2.17	10.35	34.32	50.63	1.48	3.36	12.76	24.83	0.37	28.72	66.34	78.03	2.05	58 ms
CHAN	10.14	38.03	55.35	<u>3.15</u>	<u>14.03</u>	40.06	59.42	<u>2.06</u>	5.36	21.34	34.78	1.04	30.35	67.51	78.35	4.01	63 ms
CRCL-F	10.05	37.61	<u>56.13</u>	<u>2.08</u>	<u>13.95</u>	<u>41.42</u>	59.50	1.88	5.34	24.16	35.44	<u>1.13</u>	28.35	65.45	78.17	<u>3.47</u>	63 ms
VSE ∞	9.32	37.53	55.78	-	12.73	38.56	56.77	-	<u>5.71</u>	21.99	<u>35.29</u>	-	<u>30.31</u>	67.07	77.59	-	13 ms
HREM	9.13	<u>38.35</u>	54.38	-	12.81	39.86	57.31	-	<u>5.42</u>	21.92	<u>33.57</u>	-	<u>29.13</u>	66.16	78.73	-	18 ms
ESA	10.13	<u>37.41</u>	55.63	-	13.74	39.25	58.74	-	5.52	21.25	34.44	-	29.24	<u>67.08</u>	<u>79.78</u>	-	18 ms
CRCL-C	<u>10.78</u>	38.05	54.29	-	13.08	39.90	58.69	-	4.94	19.49	31.90	-	28.31	<u>65.63</u>	<u>76.97</u>	-	14 ms
Ours	13.29	38.84	56.20	6.24	14.56	43.54	58.23	5.86	5.95	<u>23.22</u>	34.79	3.52	36.48	68.53	81.44	22.99	54 ms

Performance comparison with existing baselines on the CrossScene-RETR datasets.

Method	Conspicuous		Regular		Confusing		Overall		T
	0.25	0.5	0.25	0.5	0.25	0.5	0.25	0.5	
CRCL-F	2.35	2.35	2.54	2.22	2.69	2.58	2.54	2.34	63 ms
Vista	7.06	4.57	8.70	6.26	7.38	5.08	8.04	5.62	14.2 s
GPS	5.54	5.31	5.30	5.27	5.27	5.17	5.34	5.24	27.2 s
TSP3D	4.85	4.36	4.61	4.21	4.32	3.74	4.58	4.12	14.3 s
HREM+VisTA	20.41	19.46	16.88	15.43	14.13	12.88	17.25	15.96	(18 + 45) ms
ESA+GPS	<u>21.26</u>	<u>20.08</u>	<u>14.34</u>	<u>13.18</u>	<u>13.98</u>	<u>12.82</u>	<u>15.60</u>	<u>14.44</u>	(18 + 171) ms
Ours	28.67	26.52	22.45	20.82	19.83	18.20	22.99	21.26	54 ms

Performance comparison with existing 3DVG baselines on the ScanRefer, Nr3D, Sr3D datasets.

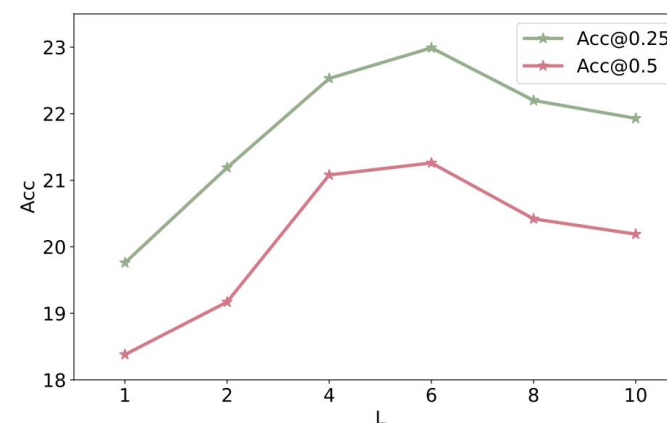
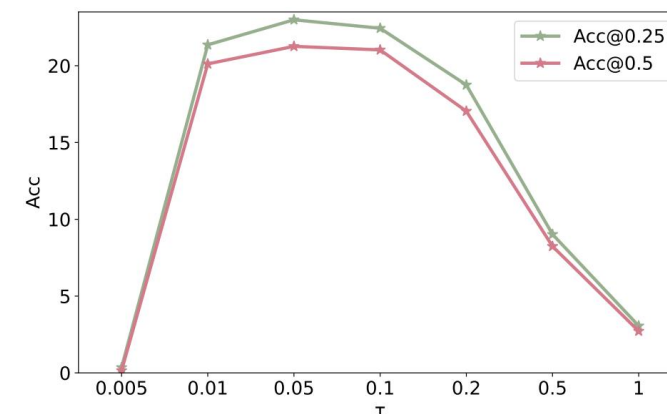
Method	ScanRefer						Nr3D	Sr3D	T
	Unique		Multiple		Overall		Overall	Overall	
	0.25	0.5	0.25	0.5	0.25	0.5	0.25	0.25	
3D-BUTD	6.14	2.90	2.17	2.03	2.90	2.19	0.00	0.20	13.2 s
EDA	3.31	0.00	0.40	0.00	0.70	0.00	2.50	0.66	23.1 s
VisTA	5.63	5.44	5.16	<u>5.34</u>	5.23	<u>5.36</u>	0.63	0.25	12.1 s
GPS	5.04	4.11	3.84	<u>3.80</u>	4.04	<u>3.85</u>	0.20	0.35	21.2 s
TSP3D	3.17	2.82	1.53	1.32	1.78	1.55	1.14	-	14.3 s
HREM+VisTA	6.14	5.34	4.74	4.14	5.00	4.36	4.57	2.14	(18 + 42) ms
ESA+GPS	<u>6.60</u>	<u>5.86</u>	<u>5.21</u>	4.91	<u>5.43</u>	5.06	<u>4.78</u>	<u>2.97</u>	(18 + 164) ms
Ours	7.39	6.88	5.98	5.54	6.24	5.79	5.86	3.52	51 ms

Ablation Study


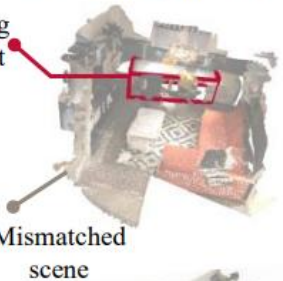

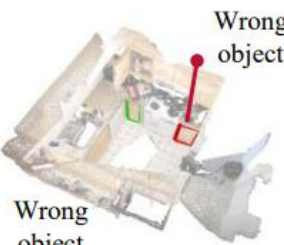


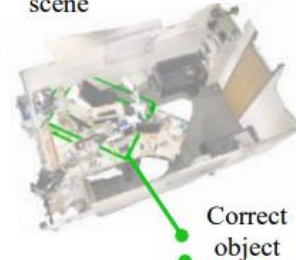






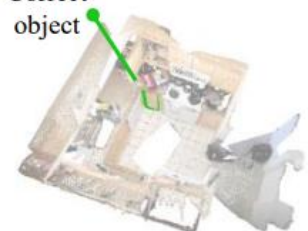

❖ Ablation study

Statistical indicators		ScanRefer	Nr3D	Sr3D	RETR
Overall	Average length	17.9	11.4	9.7	77.7
	Number of samples	46,173	41,503	83,572	39,526
	Vocabulary size	6,919	6,951	196	22,485
	Free-form	✓	✓	×	✓
Richness of description	Number of objects per text	1.8	1.7	1.8	11.4
	Number of characteristics per text	1.5	1.6	0.0	5.9
	Number of Spatial info. per text	1.2	1.1	0.6	6.3
	Number of info. points per text	4.5	4.4	2.4	23.2
	Text with Spatial info. (%)	69.8	47.5	55.6	97.5
	Text with color description (%)	58.2	29.7	0.0	81.2
	Text with shape description (%)	20.3	6.5	0.0	45.0
	Text with material description (%)	13.0	2.1	0.7	38.5

❖ Parameter analysis

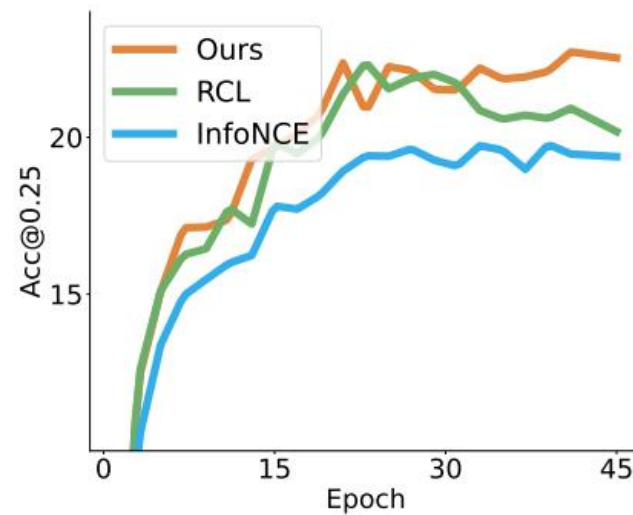
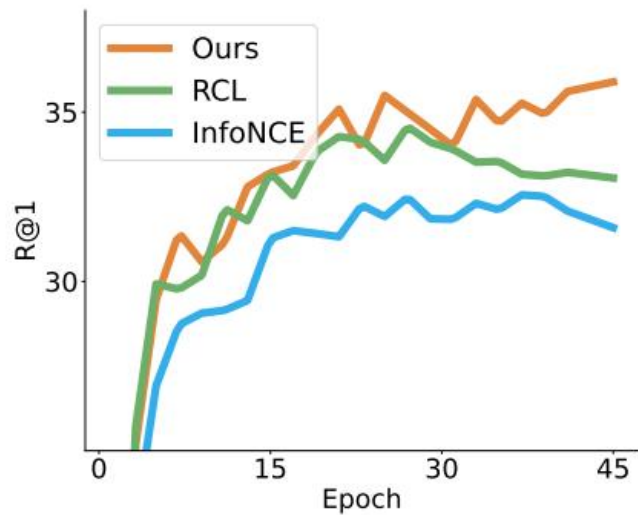


Visualization Experiments

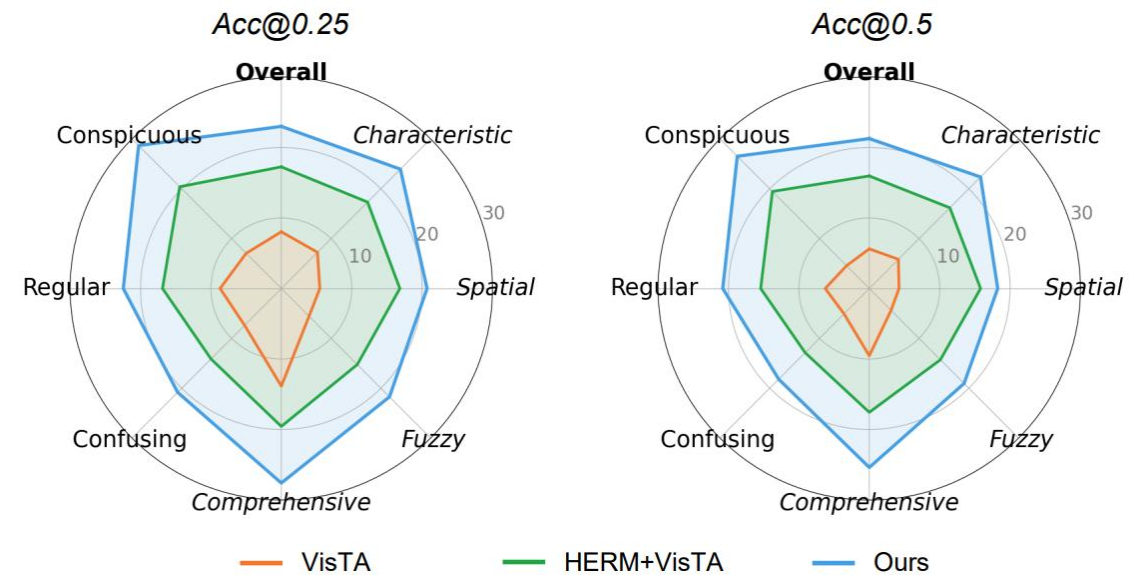
	Query text (I)	Matched scene	Grounding result	Query text (II)	Matched scene	Grounding Result	Cost time
VisTA	<p>The desk I need to find is located at the center of the room, and when facing the door, it is positioned on the right side, away from the entrance. To the left of this desk, there is a larger desk, while the curve of this particular desk faces inward, creating a cozy workspace. The black office chair is typically found pulled out from under desk.</p>	 <p><i>w/o scene retrieval</i></p>	 <p>Wrong object</p> <p>Mismatched scene</p>	<p>I hope you can help me locate a moderately light brown wooden kitchen cabinet. It's a rectangular box positioned just above the floor and situated under a counter where there is a box of canned drinks. This cabinet is located between two appliances: a dishwasher on its left and a stove on its right, with a white towel hanging nearby.</p>	 <p><i>w/o scene retrieval</i></p>	 <p>Wrong object</p>	 12.1 s
HERM + VisTA		 <p>Correct object</p>	 <p>Correct object</p>			 <p>Wrong object</p>	 63 ms
Proposed CoRe		 <p>Correct object</p>	 <p>Correct object</p>			 <p>Correct object</p>	 54 ms

Visualization Experiments

- Matching and grounding performance comparison of our CoRe and its variants



- Performance comparison on different subsets of CrossScene-RETR dataset



Thanks for watching!

