

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

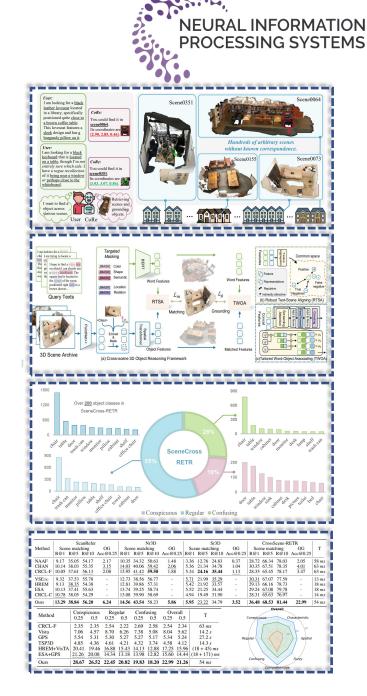






Overview of Our Work

- ❖ 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



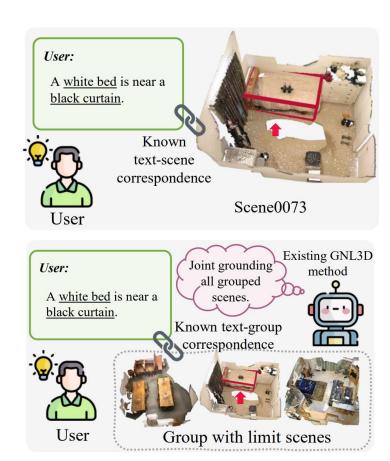




Motivation

* 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 Pratical 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.



Retrieving scenes and grounding objects.





Hundreds of arbitrary scenes without known correspondence.













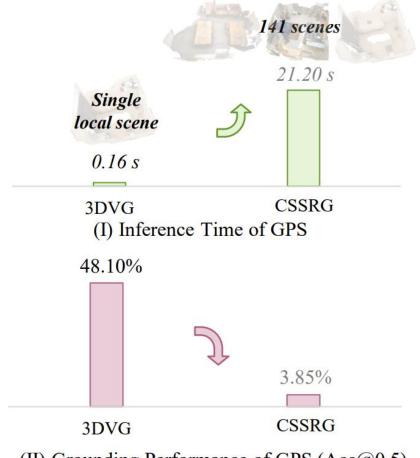


We propose a new task: CSSRG, which empowers textto-3D with cross-scene alignment and reasoning capabilities.

Challenges

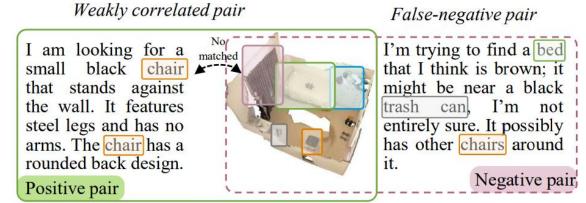


Challenge 1: Prohibitive computational costs.

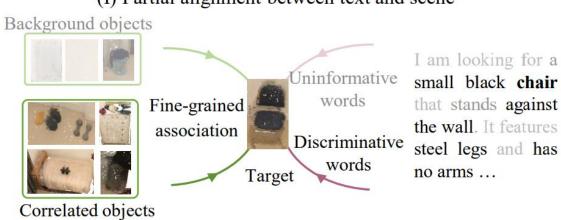


(II) Grounding Performance of GPS (Acc@0.5)

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)

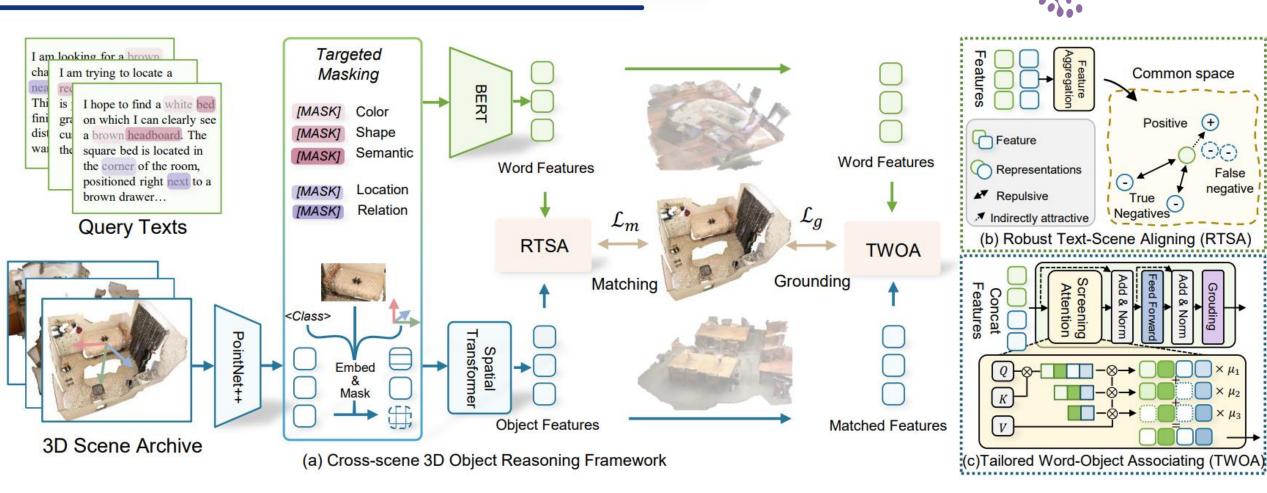






CoRe





Text-Scene Aligning: RTSA module \rightarrow Word-Object Associating: TWOA module

Matching-then-Grounding Pipeline



Dataset: CrossScene-RETR





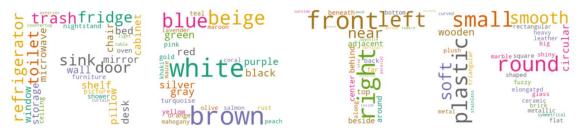


CrossScene-RETR

Construct pipeline:



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



Phase 2 Phase 3 Phase 1 Phase 4 **Corpus Collection** Scene Analysis **Text Generation** Verification Object: Confusing Validation tag 3D info. to text Other: Prompt Text-generation Chairs around requirement template table; chair next to Manual inspection above table. Intra-scene Analysis Additional corpus PROMPTI ScanRefer: Sentence coherence there is a beige Specific prompt wooden working table. placed on the side of the Text topics and grammar room. GPT 4o Nr3d: select the chair that Cross-scene discrimination is close to the end Inter-scene Analysis Constructed corpus Existing corpus

CrossScene-RETR

Characteristic-focused description:

I am looking for a small black chair that stands against the wall. It features steel legs and has no arms, giving it a sleek and modern profile. The chair has a rounded back design. which adds a touch of comfort despite its minimalist appearance.

Spatial-information-focused description:

I need assistance finding a chair. is positioned next to the door, directly adjacent to a radiator, and it can be observed if you stand facing the bed. The chair is on the left side as you look at the front of the bed, making it easy to spot among the other furnishings in the area.

Comprehensive description:

Looking for the chair means noting its critical location; it lies in the center of the space between the door, the curtain, and the radiator. That chair is not just tucked away but prominently set between these elements, thus making it an integral part of the room's layout. It's vital to focus on its position relative to surrounding items, which should help narrow down where I might find it. On the right side of the dresser, there are a white door. Near the foot of the bed, there is a dresser. The blue trash bin is located beside the radiator. The back wall of the bed is adorned with a curtain. The trash bin itself is blue in color.

I'm not entirely sure where I last saw the chair, but I think it was somewhere near the door, maybe around a radiator? It was a small black chair, I believe.

Characteristic-focused description:

I hope to find a specific white bed on which I can clearly see a brown headboard. The bed is located in the corner of the room, positioned right next to a brown drawer. In the center of the bed, there is a free weight with identical disks at each end, giving it a unique appearance among other furniture in the room.

Spatial-information-focused

I'm looking for the bed that is nestled in the corner, bordered by dark brown curtains that hang behind it. This small corner bed is strategically placed such that it aligns closely with a dresser at the foot, creating a compact and functional setting.

Characteristic-focused description:

Comprehensive description.

The bed I seek has a clean white color and a sturdy brown headboard. It's in a cozy room, where the curtains provide a warm ambiance. I remember it being in front of the curtains, offering a nice view while integrating smoothly with the surrounding furniture, including the drawer nearby. The radiator is situated behind the trash can. To the right of the curtain is a bed. The chair has four legs. On the left side of the table is the radiator. The dresser and closet are both white in color. In the middle of the bed, there is a free weight, The curtains behind the bed are dark brown.

Fuzzy description:

I believe the bed is somewhat in the corner, perhaps close to some clothes. I remember something about the curtains being dark or brown, but I'm not entirely sure. There's definitely a dresser near it, and I want to say the bed might be in the middle of everything.

Challenge-level: conspicuous

Challenge-level: confusing

Characteristic-focused description:

I am looking for a brown chair that is positioned nearest to me, facing left. This chair has a smooth finish and is distinguishable by its warm brown color. The chair is located on the right side of the room and is directly adjacent to a middle black table, providing a nice contrast against the darker furniture in the area. Its shape is quite standard for a dining chair.

Spatial-information-focused description:

I need help finding a chair that is placed at the right of a black table in the mid-back of the room. It is closely positioned next to another chair, which is to its right. The brown chair I'm searching for faces inward, suggesting it's intended for conversation or interaction with others seated nearby

My desired chair is a brown one, located near a black table, in the mid-back area of a room. While I remember that it has a relatively simple design and sits next to two other chairs. I can't recall the exact layout surrounding it. I think it's positioned such that it complements the furniture around it, especially the table, making it feel inviting for anyone who sits there.

Fuzzy description:

I'm trying to find a chair that I think is brown it might be near a black table, I'm not entirely sure. It possibly has other chairs around it, and I vaguely remember it facing a certain direction. The details are a bit unclear, but it should be somewhere in the right side

Comprehensive description:

I would like you to help me find a specific door that is light brown in color and has a rectangular shape. This brown door is situated in the northwest-most part of the room, specifically at the far-right corner. It has swung open against the wall, allowing for easy access, and I should mention that there is a chair positioned to the left of it, which enhances its visibility.

Spatial-information-focused description:

The door I'm looking for is located in the back right corner of the room. It's important to note that this door is not just any door; it is the only one in that particular corner, resting between a chair and a board that is mounted on the wall to the left. The layout in this area makes it quite distinct, and it stands out in its spot when you

Comprehensive description:

I'm hoping my description helps you find this light brown door, which is a rectangular shape, located in the far right corner of the room. To give you additional context, it's nestled there with a chair adjacent to it, and there's a wall board that could serve as a landmark as well. The way it's positioned really catches your eye if you happen

Fuzzy description

I think the door I'm referring to is somewhat back right area of the room, but I can't exactly recall all the details.

to be in the corner.

light brown, located in a corner, but I'm not completely sure if there was a chair next to it or a board nearby. It feels like it's at the



(a) Characteristic-focused

white feature locate front creating and desktop istand to be a stand of the stand o



(b) Spatial-information-focused



Over 200 object classes in 1200 600 SceneCross-RETR 300 600 **SceneCross** RETR 600 ■ Conspicuous ■ Regular ■ Confusing

NEURAL INFORMATION PROCESSING SYSTEMS

	Statistical indicators	ScanRefer	Nr3D	Sr3D	RETR
	Average length	17.9	11.4	9.7	77.7
	Number of samples	46,173	41,503	83,572	39,526
Overall	Vocabulary size	6,919	6,951	196	22,485
	Free-form	$$	\checkmark	×	
	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
Richness of	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
		1			ı



Experiment







Comparative Experiments



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

			ScanRefer			Nr3D			Sr3D			CrossScene-RETR			1		
Method	Scen	ne mate	hing	OG	Scei	ne mato	hing	OG	Sce	ne mate	ching	OG	Sce	ne mate	ching	OG	T
	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	3.15	14.03	40.06	59.42	2.06	5.36	21.34	34.78	1.04	30.35	67.51	78.35	4.01	63 ms
CRCL-F	10.05	37.61	56.13	2.08	13.95	41.42	59.50	1.88	5.34	24.16	35.44	1.13	28.35	65.45	78.17	3.47	63 ms
VSE∞	9.32	37.53	55.78	-	12.73	38.56	56.77	-			35.29	-	30.31	67.07	77.59	-	13 ms
HREM	9.13	38.35	54.38	1-1	12.81	39.86	57.31	-	5.42	21.92	33.57	-	29.13	66.16	78.73	-	18 ms
ESA	10.13	37.41	55.63	(<u>-</u>)	13.74	39.25	58.74	12	5.52	21.25	34.44	-	29.24	67.08	79.78	2	18 ms
CRCL-C	10.78	38.05	54.29	-	13.08	39.90	58.69	Ħ	4.94	19.49	31.90	-	28.31	65.63	76.97	- 100 Tes	14 ms
Ours	13.29	38.84	56.20	6.24	14.56	43.54	58.23	5.86	5.95	23.22	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		Т
Method	0.25	0.5	0.25	0.5	0.25	0.5	0.25	0.5	1
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	21.26	20.08	14.34	13.18	13.98	12.82	15.60	14.44	(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.

5 5	[Scan	Refer		Nr3D	Sr3D		
Method	Unique		Multiple		Overall		Overall	Overall	T
	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	5.34	5.23	5.36	0.63	0.25	12.1 s
GPS	5.04	4.11	3.84	3.80	4.04	3.85	0.20	0.35	21.2 s
TSP3D	3.17	2.82	1.53	1.32	1.78	1.55	1.14	2	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	6.60	5.86	5.21	4.91	5.43	5.06	4.78	2.97	(18 + 164) ms
Ours	7.39	6.88	5.98	5.54	6.24	5.79	5.86	3.52	51 ms

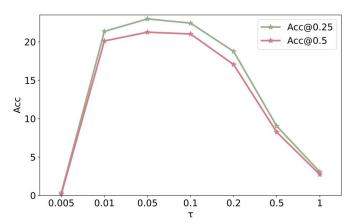
Ablation Study

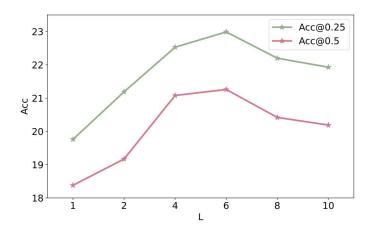
Ablation stuty

	Statistical indicators	ScanRefer	Nr3D	Sr3D	RETR
	Average length	17.9	11.4	9.7	77.7
Overall	Number of samples	46,173	41,503	83,572	39,526
	Vocabulary size	6,919	6,951	196	22,485
	Free-form	$\overline{}$		×	\checkmark
	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
Richness of	Number of info. points per text	4.5	4.4	2.4	23.2
description	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



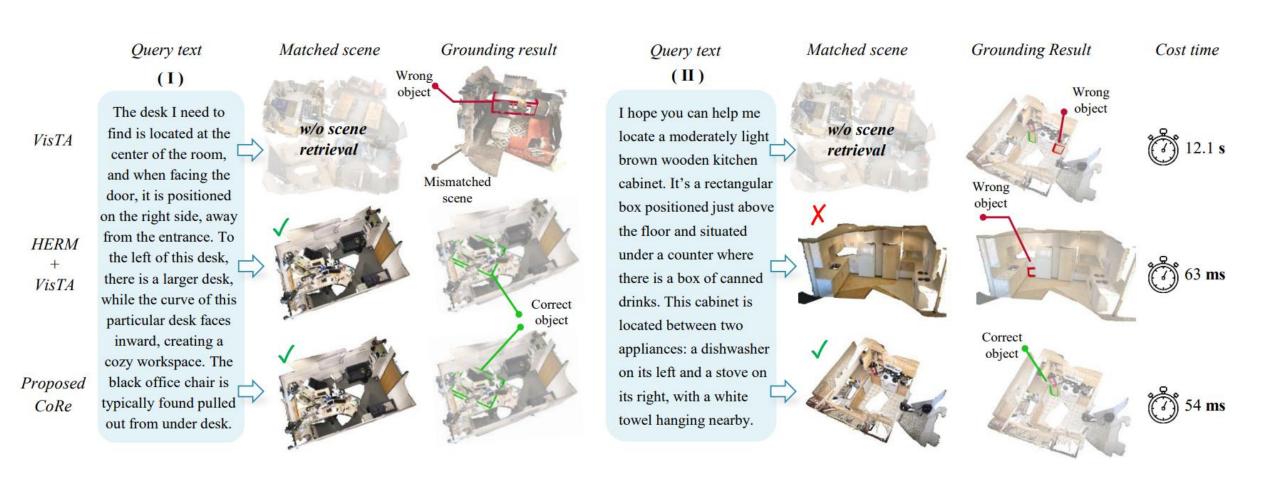
Parameter analysis





Visualization Experiments

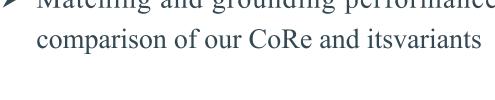


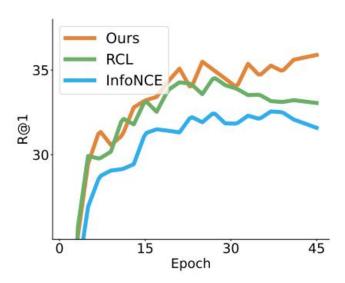


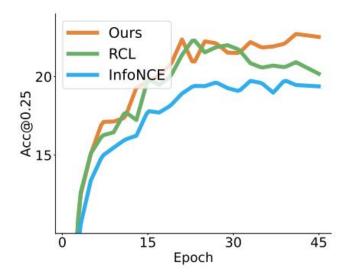
Visualization Experiments



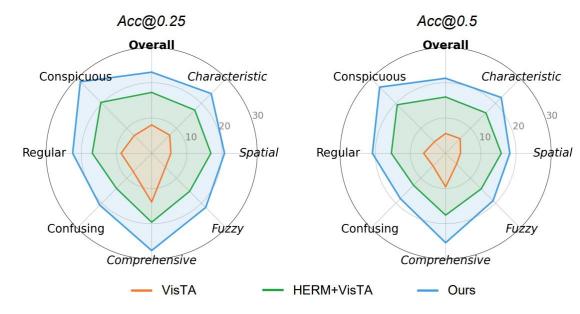
> Matching and grounding performance comparison of our CoRe and its variants







> Performance comparison on different subsets of CrossScene-RETR dataset



Thanks for watching!

