Interactive Cross-modal Learning for Text-3D Scene Retrieval

Yanglin Feng [1] Yongxiang Li [1] Yuan Sun [2] Yang Qin [1] Dezhong Peng [1,3] Peng Hu* [1]

[1] College of Computer Science, Sichuan University, Chengdu, China.

[2] National Key Laboratory of Fundamental Algorithms and Models for Engineering Numerical Simulation, Sichuan University, Chengdu, China.

[3] Tianfu Jincheng Laboratory, Chengdu, China.

NEURAL INFORMATION PROCESSING SYSTEMS

Contents

- Overview of Our Work
- Background
- Method: Interactive Text-3D Scene Retrieval Method
  - Interactive Retrieval Refinement framework (IRR)
  - Interaction Adaptation Tuning strategy (IAT)
- Experiments

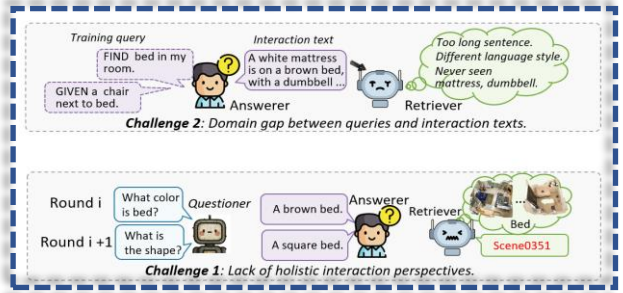
NEURAL INFORMATION PROCESSING SYSTEMS

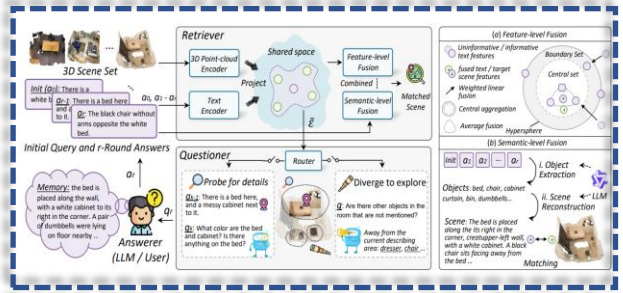
Overview of Our Work

Overview of Our Work

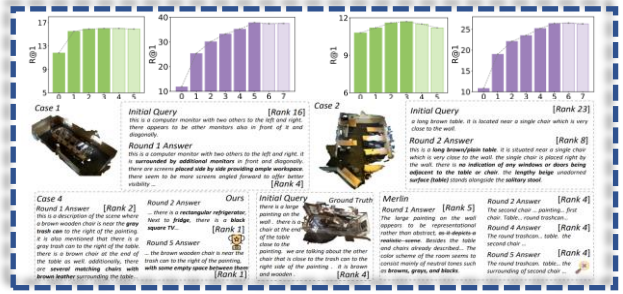
NEURAL INFORMATION PROCESSING SYSTEMS

- We propose a novel Interactive Text-3D Scene Retrieval Method (IDeal), which actively enhances alignment between text queries and 3D scenes through ongoing interactions.
- An Interactive Retrieval Refinement framework (IRR) is presented to enable a deep interaction for comprehensive scene exploration, leading to progressively improved retrieval.
- An Interaction Adaptation Tuning strategy (IAT) is proposed, which facilitates the transfer of the retriever to the interaction text domain, promoting improved interaction.
- We conduct extensive comparison experiments on text-3D scene datasets. Our IDeal remarkably outperforms the existing methods, demonstrating its superiority.





Methods	ScanRefer R@1	R@5	R@10	Rsum	Nr3D R@1	R@5	R@10	Rsum	Sr3D R@1	R@5	R@10	Rsum
w/ coarse-grained descriptions:												
VSE∞ (CVPR'21)	9.7	33.1	50.2	93.0	5.8	21.5	32.5	59.8	5.5	18.9	27.4	51.8
CHAN (CVPR'23)	9.4	32.3	52.1	93.8	7.5	15.0	32.1	54.6	5.4	21.3	34.8	61.5
HREM (CVPR'23)	10.2	34.0	51.4	95.6	7.3	18.1	31.9	57.3	5.4	21.9	33.6	60.9
CRCL (NeurIPS'23)	10.3	32.4	49.8	92.5	8.1	22.5	33.2	63.8	4.9	19.5	31.9	56.3
RoMa (TMM'25)	11.4	34.8	54.4	100.6	6.5	24.8	37.6	68.9	7.8	27.3	39.3	74.4
IDeal	**16.0**	**42.7**	**59.8**	**118.5**	**11.7**	**34.8**	**50.4**	**96.9**	**10.3**	**30.2**	**48.5**	**89.0**
w/ fine-grained descriptions:												
ChatIR† (NeurIPS'23)	21.8	55.4	73.1	150.3	15.6	40.3	58.1	114.0	12.1	35.9	50.3	98.3
Rewrite† (ICMR'24)	17.4	47.0	63.7	128.1	17.1	31.4	45.8	94.3	12.4	28.1	40.4	80.9
MERLIN† (EMNLP'24)	31.1	68.8	83.8	183.7	21.8	55.0	71.8	148.6	14.2	42.0	60.3	116.5
BASELINE: IR†	29.9	68.0	83.3	181.2	18.0	51.6	60.2	129.8	14.6	39.2	55.3	109.1
BASELINE SUM†	34.4	69.5	85.1	189.0	22.5	55.2	67.5	145.2	16.4	41.5	62.1	120.0
IDeal†	**37.8**	**71.8**	**86.4**	**196.0**	**26.4**	**62.7**	**78.7**	**167.8**	**20.2**	**43.1**	**63.1**	**126.4**



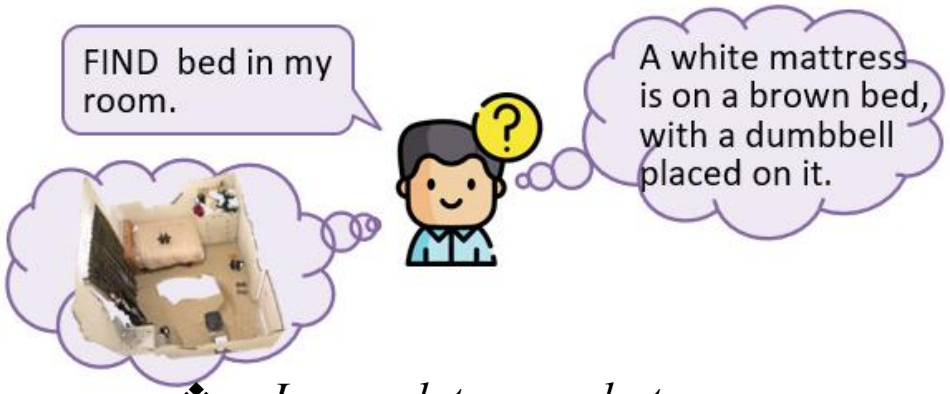

NEURAL INFORMATION
PROCESSING SYSTEMS

Background

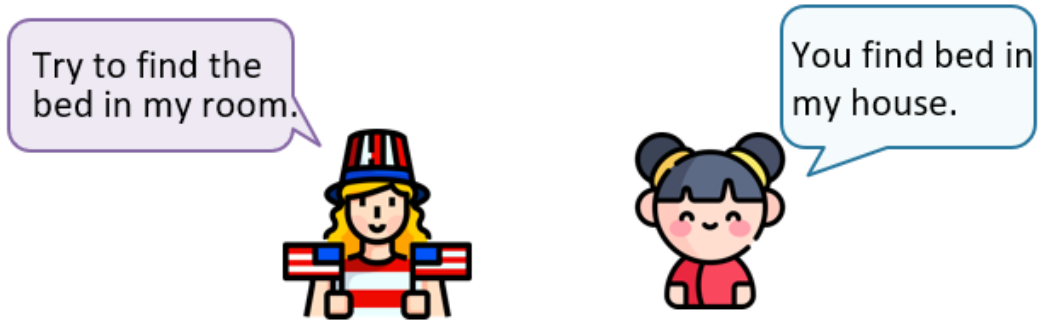

四川大学
SICHUAN UNIVERSITY
1896


TIANFU JINCHENG LABORATORY
天府锦城实验室

Background

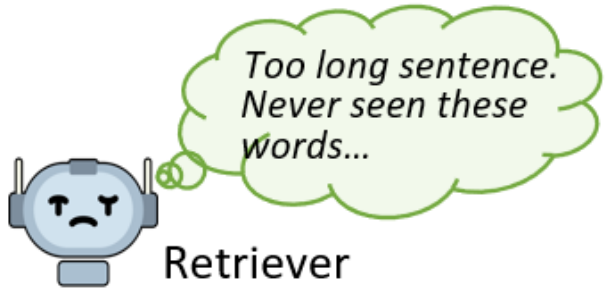


❖ *Incomplete one-shot descriptions of user intent*



❖ *Domain shifts*

❖ *Ambiguous descriptions*



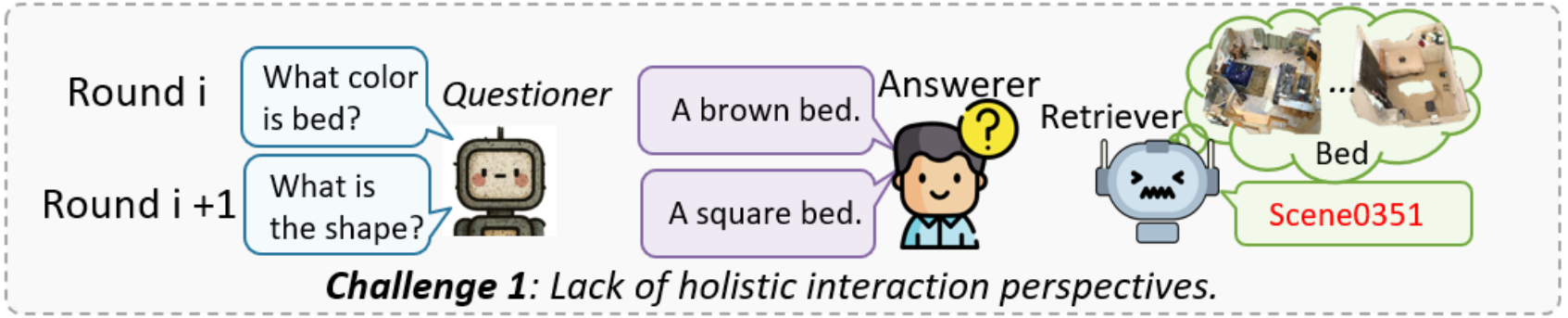
❖ *Limited generalization of the models*

The effectiveness of single-round static retrieval methods is often challenged by complex real-world conditions.

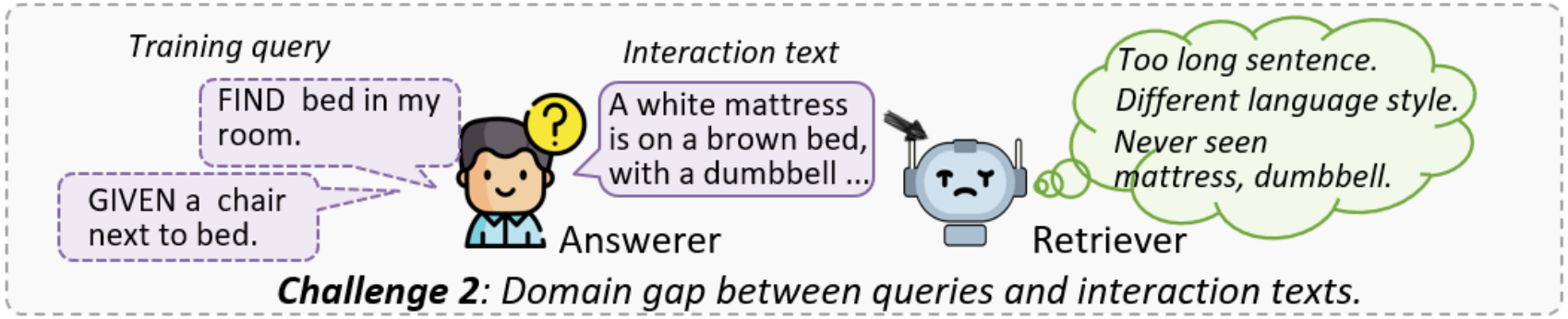
NEURAL INFORMATION PROCESSING SYSTEMS

Challenges

- Challenges in applying existing interactive methods



***Challenge 1**: Lack of holistic interaction perspectives.*



***Challenge 2**: Domain gap between queries and interaction texts.*

- Challenges in making existing static methods interactive

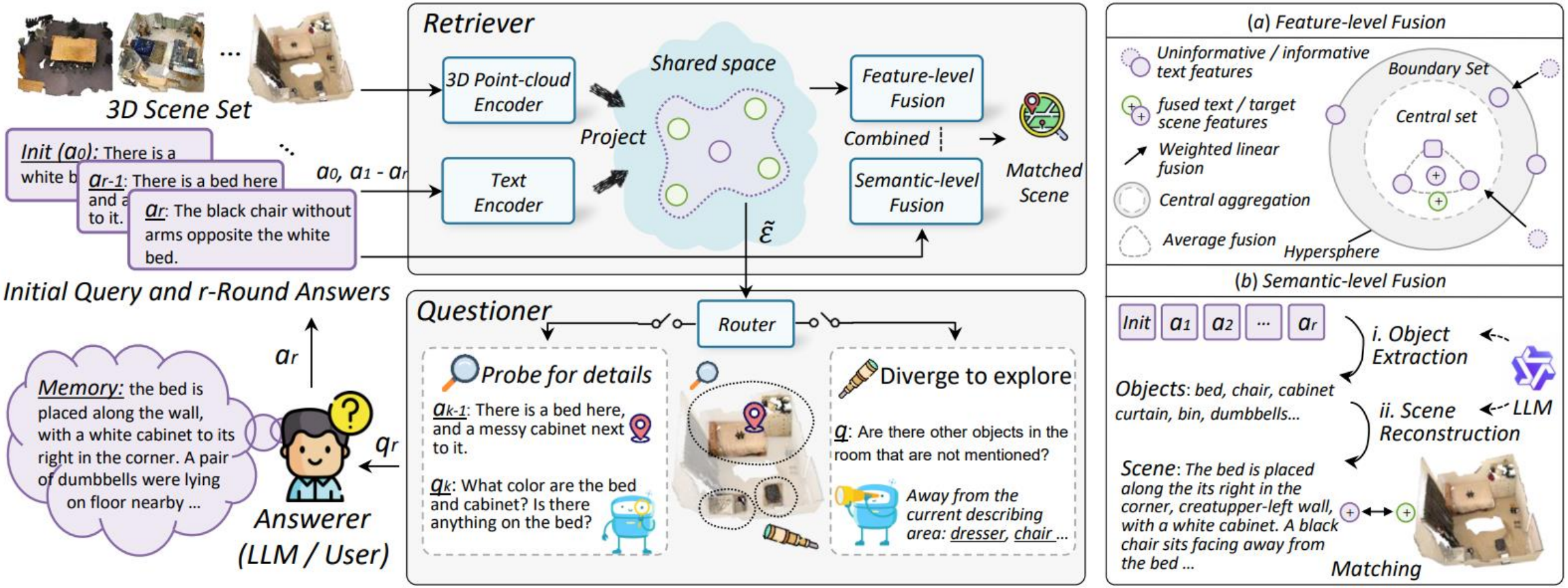
NEURAL INFORMATION PROCESSING SYSTEMS

Method:
Interactive Text-3D Scene Retrieval Method (IDeal)

Ideal - Interactive Retrieval Refinement framework (IRR)

NEURAL INFORMATION PROCESSING SYSTEMS

❖ Framework



❖ Pipeline

Round 1

Questioner —*Ask*→ Answerer —*Answer*→ Retriever —*Matching*→ .....

Round N

→ Questioner —*Ask*→ Answerer —*Answer*→ Retriever

Ideal - Interaction Adaptation Tuning strategy (IAT)

NEURAL INFORMATION PROCESSING SYSTEMS

- Interaction Adaptation Tuning
  - **Interaction-like text generation**

    It begins by integrating scene information and descriptions from the training data to construct simulated memory. Then, we continuously interact with it using to obtain augmented texts that closely approximate real interaction scenarios.

  - **Trained model domain and interaction text domain alignment**

    Contrastive domain adaptation criterion:

$$\mathcal{R}(\boldsymbol{\theta}) = \mathcal{R}_{dis}(\boldsymbol{\theta}) + \mathcal{R}_{div}(\boldsymbol{\theta}) = \mathbb{E}_{\tilde{\mathcal{U}}}\left[\left(-\mathbb{E}_{\tilde{\mathcal{U}}^+}\left\{\mathcal{S}(\tilde{\boldsymbol{u}}_i^+, \tilde{\boldsymbol{u}}_i)\right\}\right) + \left(\mathbb{E}_{\tilde{\mathcal{U}}^-}\left\{\mathcal{S}(\tilde{\boldsymbol{u}}_i^-, \tilde{\boldsymbol{u}}_i)\right\}\right)\right]$$

    Optimizable loss function for contrastive domain adaptation:

$$\mathcal{L} = \lambda \mathcal{L}_{dis} + (1-\lambda)\mathcal{L}_{div}$$

    (a) *Positive aligning*: log-based proxy loss term of discriminability risk $\mathcal{R}_{dis}(\boldsymbol{\theta})$:

$$\mathcal{L}_{dis} = -\sum_{i=1}^{b}\sum_{j=1}^{n_c} y_{ij} \log \mathcal{S}(\tilde{\boldsymbol{u}}_i, \boldsymbol{v}_j)$$

    (b) *Negative diverging*: surrogate objective for the lower bound of divergence risk $\mathcal{R}_{div}(\boldsymbol{\theta})$:

$$\mathcal{L}_{div} = \sum_{i=1}^{b}\sum_{j\neq i}^{b} \underbrace{\exp\left(-\max\left(0, \mathcal{S}(\tilde{\boldsymbol{u}}_i, \tilde{\boldsymbol{u}}_j) - \gamma\right)\right)}_{\text{Weighting term}} \underbrace{\log\left(1 - \mathcal{S}(\tilde{\boldsymbol{u}}_i, \tilde{\boldsymbol{u}}_j)\right)}_{\text{Complementary contrastive term}}$$


NEURAL INFORMATION
PROCESSING SYSTEMS

Experiment


四川大学
SICHUAN UNIVERSITY
1896


TIANFU JINCHENG LABORATORY
天府锦城实验室

Comparative Experiments

NEURAL INFORMATION PROCESSING SYSTEMS

Tab. 1 Performance comparison on ScanRefer, Nr3D, and Sr3D in terms of R@1, R@5, R@10, and their sum (Rsum). † denotes the use of coarse-grained descriptions as memory

Methods	ScanRefer				Nr3D				Sr3D			
	R@1	R@5	R@10	Rsum	R@1	R@5	R@10	Rsum	R@1	R@5	R@10	Rsum
w/ coarse-grained descriptions:												
VSE∞ (CVPR'21)	9.7	33.1	50.2	93.0	5.8	21.5	32.5	59.8	5.5	18.9	27.4	51.8
CHAN (CVPR'23)	9.4	32.3	52.1	93.8	7.5	15.0	32.1	54.6	5.4	21.3	34.8	61.5
HREM (CVPR'23)	10.2	34.0	51.4	95.6	7.3	18.1	31.9	57.3	5.4	21.9	33.6	60.9
CRCL (NeurIPS'23)	10.3	32.4	49.8	92.5	8.1	22.5	33.2	63.8	4.9	19.5	31.9	56.3
RoMa (TMM'25)	11.4	34.8	54.4	100.6	6.5	24.8	37.6	68.9	7.8	27.3	39.3	74.4
IDeal	**16.0**	**42.7**	**59.8**	**118.5**	**11.7**	**34.8**	**50.4**	**96.9**	**10.3**	**30.2**	**48.5**	**89.0**
w/ fine-grained descriptions:												
ChatIR† (NeurIPS'23)	21.8	55.4	73.1	150.3	15.6	40.3	58.1	114.0	12.1	35.9	50.3	98.3
Rewrite† (ICMR'24)	17.4	47.0	63.7	128.1	17.1	31.4	45.8	94.3	12.4	28.1	40.4	80.9
MERLIN† (EMNLP'24)	31.1	68.8	83.8	183.7	21.8	55.0	71.8	148.6	14.2	42.0	60.3	116.5
BASELINE: IR†	29.9	68.0	83.3	181.2	18.0	51.6	60.2	129.8	14.6	39.2	55.3	109.1
BASELINE SUM†	34.4	69.5	85.1	189.0	22.5	55.2	67.5	145.2	16.4	41.5	62.1	120.0
IDeal†	**37.8**	**71.8**	**86.4**	**196.0**	**26.4**	**62.7**	**78.7**	**167.8**	**20.2**	**43.1**	**63.1**	**126.4**

The advancement of the interactive methods~

Plug-and-play boost~

Table 2: Performance comparison on ScanRefer and Nr3D in terms of R@1, R@5, R@10, and their sum. +Ideal indicates plugging the model into our IDeal. † denotes the use of fine-grained descriptions as memory.

Methods	ScanRefer				Nr3D			
	R@1	R@5	R@10	Rsum	R@1	R@5	R@10	Rsum
VSE∞	9.7	33.1	50.2	93.0	5.8	21.5	32.5	59.8
+IDeal	13.3	38.9	57.6	109.8	8.7	27.5	42.1	78.3
VSE∞†	14.9	42.3	61.5	118.7	16.4	47.5	55.2	119.1
+IDeal†	35.8	70.6	85.0	191.4	21.2	52.1	68.4	141.7
CRCL	10.3	32.4	49.8	92.5	8.1	22.5	33.2	63.8
+IDeal	13.4	35.5	56.1	105.0	7.4	25.4	38.3	71.1
CRCL†	17.5	45.1	58.3	120.9	13.4	44.5	51.5	109.4
+IDeal†	31.7	66.9	83.5	182.1	15.8	50.4	64.4	130.6
RoMa	9.7	33.1	50.2	93.0	8.3	27.9	37.2	73.4
+IDeal	16.0	42.7	59.8	118.5	11.7	34.8	50.4	96.9
RoMa†	16.7	44.8	61.6	123.1	17.4	48.5	57.5	123.4
+IDeal†	37.8	71.8	86.4	196.0	25.4	60.7	75.7	161.8

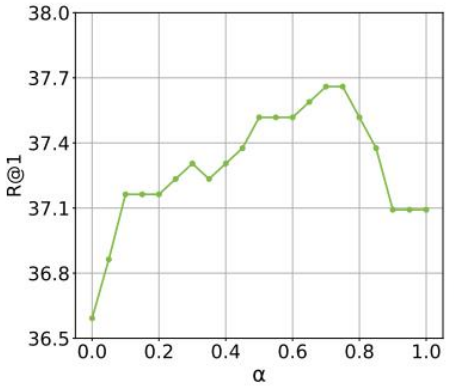
Ablation Study

NEURAL INFORMATION PROCESSING SYSTEMS

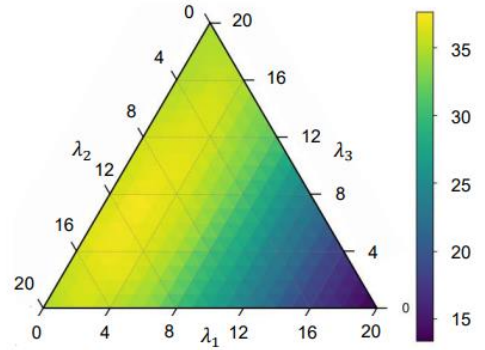
❖ Ablation stuty

Configurations		ScanRefer R@1	R@5	R@10	Rsum
Questioner	*w/o* $\mathcal{Q}_1$	36.0	71.2	86.7	193.9
	w/o $\mathcal{Q}_2$	26.3	61.5	77.3	165.1
Retriever	*w/o* $\hat{\boldsymbol{p}}_1(\boldsymbol{u}_i)$	35.2	70.1	86.5	191.8
	w/o $\hat{\boldsymbol{p}}_2(\bar{\boldsymbol{u}}_i)$	28.1	63.3	80.6	172.0
	w/o $\hat{\boldsymbol{p}}_3(\boldsymbol{s}_i)$	31.8	67.8	84.2	183.8
	w/o CoT	35.7	71.5	85.9	193.1
Adaptation	*w/o* IAT	16.6	48.4	64.4	129.4
	w/o $\mathcal{L}_{dis}$	34.9	69.5	84.4	188.8
	w/o $\mathcal{L}_{div}$	35.1	69.4	84.1	191.7
Full	IDeal	37.8	71.8	86.4	196.0

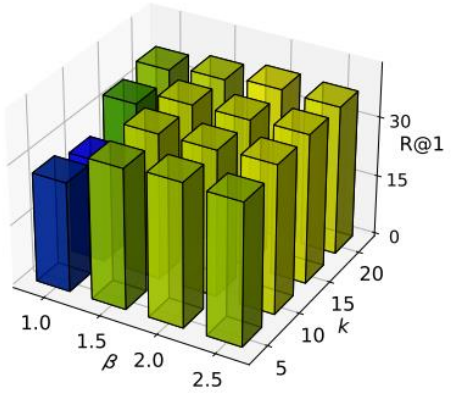
❖ Parameter analysis



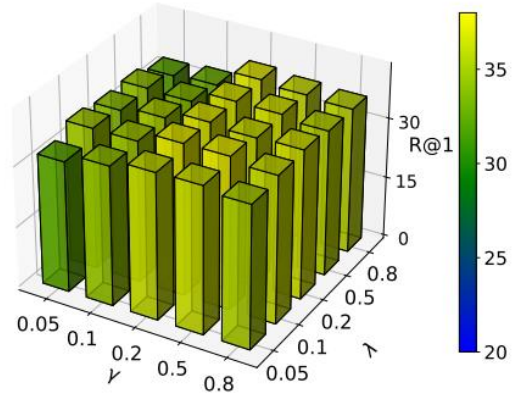
(a) α in *retriever*.



(b) λ_1, λ_2, λ_3 in *retriever*.



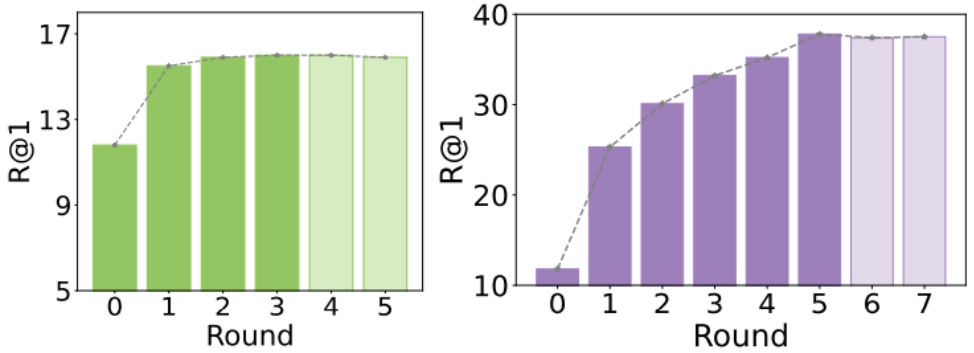
(c) k, β in *questioner*.



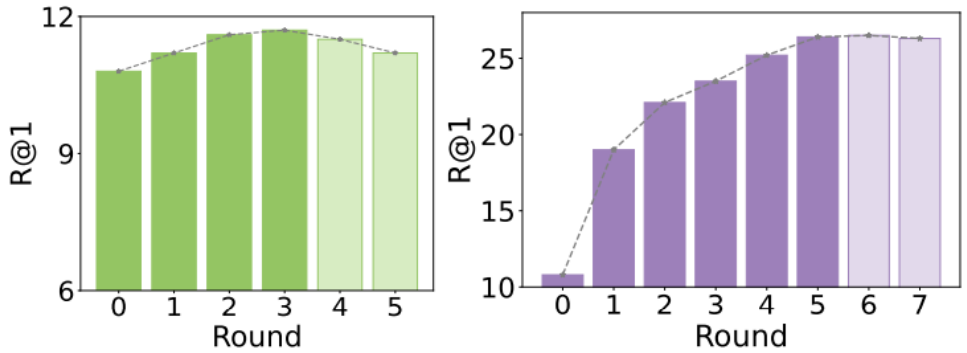
(d) λ, γ in IAT.

Visualization Experiments

❖ *Interaction Analysis*



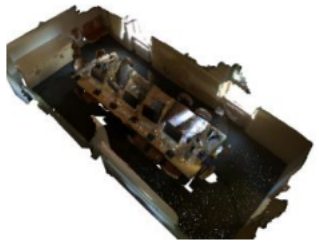
(a) ScanRefer



(b) Nr3D

❖ *Case Analysis*

Case 1



Initial Query [Rank 16]

this is a computer monitor with two others to the left and right. there appears to be other monitors also in front of it and diagonally.

Round 1 Answer

this is a computer monitor with two others to the left and right. it is ***surrounded by additional monitors*** *in front and diagonally. there are screens* ***placed side by side providing ample workspace****. there seem to be more screens angled forward to offer better visibility ...* [Rank 4]

Case 2

Initial Query [Rank 23]

a long brown table. it is located near a single chair which is very close to the wall.

Round 2 Answer [Rank 8]

this is a ***long brown/plain table****. it is situated near a single chair which is very close to the wall. the single chair is placed right by the wall. there is* ***no indication of any windows or doors being adjacent to the table or chair****. the* ***lengthy beige*** *unadorned* ***surface (table)*** *stands alongside the* ***solitary stool****.*

Case 4

Round 1 Answer [Rank 2]

this is a description of the scene where a brown wooden chair is near the ***gray trash can*** *to the right of the painting. it is also mentioned that there is a gray trash can to the right of the table. there is a brown chair at the end of the table as well. additionally, there are* ***several matching chairs with brown leather*** *surrounding the table...*

Ours

Round 2 Answer

... there is a ***rectangular refrigerator****, Next to* ***fridge****, there is a* ***black square TV****...* [Rank 1]

Round 5 Answer

... the brown wooden chair is near the trash can to the right of the painting, ***with some empty space between them*** [Rank 1]

Initial Query

there is a large painting on the wall . there is a chair at the end of the table close to the painting. we are talking about the other chair that is close to the trash can to the right side of the painting . it is brown and wooden . [Rank 4]

Ground Truth

Merlin

Round 1 Answer [Rank 5]

The large painting on the wall appears to be representational rather than abstract, ~~*as it depicts a realistic scene*~~*. Besides the table and chairs already described... The color scheme of the room seems to consist mainly of neutral tones such as* ***browns, grays, and blacks****.*

Round 2 Answer [Rank 4]

The second chair ... painting... first chair. Table... round trashcan...

Round 4 Answer [Rank 4]

The round trashcan... table. the second chair ...

Round 5 Answer [Rank 4]

The round trashcan. table... the surrounding of second chair ...

Thanks for watching!