3D-MoRe: Unified Modal-Contextual Reasoning for Embodied Question Answering and Dynamic Dense Captioning

Rongtao Xu, Han Gao, Mingming Yu, Dong An, Shunpeng Chen, Changwei Wang,
Li Guo, Xiaodan Liang†, Shibiao Xu

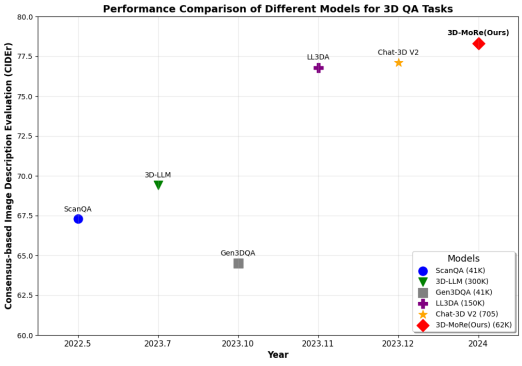

Fig. 1. The agent's success rate in 3D Question Answering improves with increased data size. Our method generates 62K QA pairs, substantially enhancing performance and bringing it closer to human-level results.

***Abstract*— With the growing need for diverse and scalable data in indoor scene tasks, such as question answering and dense captioning, we propose 3D-MoRe, a novel paradigm designed to generate large-scale 3D-language datasets by leveraging the strengths of foundational models. The framework integrates key components, including multi-modal embedding, cross-modal interaction, and a language model decoder, to process natural language instructions and 3D scene data. This approach facilitates enhanced reasoning and response generation in complex 3D environments. Using the ScanNet 3D scene dataset, along with text annotations from ScanQA and ScanRefer, 3D-MoRe generates 62,000 question-answer (QA) pairs and 73,000 object descriptions across 1,513 scenes. We also employ various data augmentation techniques and implement semantic filtering to ensure high-quality data. Experiments on ScanQA demonstrate that 3D-MoRe significantly outperforms state-of-the-art baselines, with the CIDEr score improving by 2.15%. Similarly, on ScanRefer, our approach achieves a notable increase in CIDEr@0.5 by 1.84%, highlighting its effectiveness in both tasks. Our code and generated datasets will be publicly released to benefit the community, and both can be accessed on the project page.**

I. INTRODUCTION

In 3D question answering (3DQA) and dense captioning tasks, models must tackle complex multimodal reasoning within 3D environments. The 3DQA task demands deep scene understanding and spatial reasoning to answer text-based questions, while dense captioning requires detailed descriptions of objects and their relationships in 3D space. Existing methods often rely on multimodal fusion techniques such as semantic-level data augmentation, spatial attention, and cross-modal encoding to enhance object localization and scene comprehension. Unlike Gen3DQA [1], which focuses on small datasets, our 3D-MoRe method leverages diverse data augmentation strategies to broaden dataset variety. In contrast to Vote2Cap-DETR++ [2], which depends heavily on spatial features, our approach integrates both spatial and linguistic information to achieve robust performance across various environments. Advanced semantic filtering ensures high-quality data, significantly improving contextual accuracy.

Generating large-scale datasets poses challenges, including prompt construction [3][4], accurate annotation extraction [5][6], and data quality filtering [7]. To address these, we propose the Adaptive Multimodal Fusion Paradigm, which incorporates three generation methods: QA Generation (expanding ScanQA), Captioning Generation (transforming ScanRefer captions into QA pairs), and Scene Generation (using vision-language models to generate QA pairs from 3D scene data). Additionally, we introduce two data filtering techniques—semantic similarity and search—to ensure data quality. This approach generates 62,000 triplets for the 3DQA task and 73,000 for captioning, as shown in Figure 1. Using these triplets, we train a 3D-LLM model that encodes the triplets across three branches, aligns the modalities through interaction, and decodes responses with an LLM, achieving a significant performance improvement, approaching human-level proficiency.

In summary, our key contributions lie in:

- We introduce 3D-MoRe, an innovative framework that leverages foundational models to generate large-scale 3D-language datasets, integrating multimodal embedding, cross-modal interaction, and a language model decoder to enhance reasoning in complex 3D environments.
- 3D-MoRe synthesizes 62,000 question-answer pairs and 73,000 object descriptions from the ScanNet dataset, significantly increasing data diversity and improving performance on 3D question answering and 3D dense captioning tasks.
- Through advanced data augmentation techniques, including synonym substitution, sentence reordering, and semantic filtering, 3D-MoRe achieves a 2.15% improve-

†Xiaodan Liang is the corresponding author (xdliang328@gmail.com). Rongtao Xu and Han Gao contributed equally. Rongtao Xu, Mingming Yu, Dong An, and Changwei Wang are with the Institute of Automation, Chinese Academy of Sciences, China. Han Gao, Shunpeng Chen, Li Guo and Shibiao Xu are with the Beijing University of Posts and Telecommunications, China. Xiaodan Liang is with Sun Yat-Sen University, China.

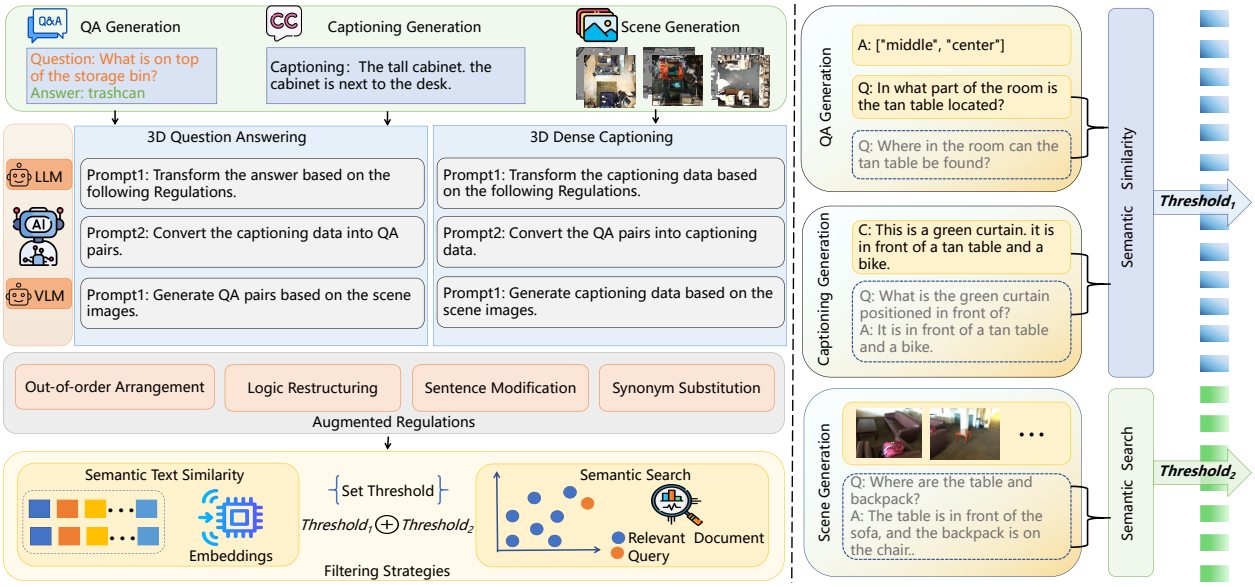

Fig. 2. **Generation paradigm pipeline.** To expand the datasets, we combine ScanNet scene data with textual annotations from ScanQA and ScanRefer. We apply semantic search and similarity filtering to rigorously select generated data and obtain high-quality text embeddings. The figure on the right illustrates the filtering strategies for the 3D Question Answering task, with parentheses indicating the compared text data.

ment in CIDEr on ScanQA and a 1.84% increase in CIDEr@0.5 on ScanRefer, effectively enhancing model accuracy.

II. Related Work

3D Question Answering and 3D Dense Captioning. 3D Question Answering and Dense Captioning involve interpreting 3D scenes by utilizing depth and point cloud data to enhance spatial understanding. 3DQA models align visual and linguistic data to improve response accuracy and reduce uncertainty [8][9]. Datasets like ScanQA [10], CLEVR3D [11], and FE-3DGQA [12] provide essential benchmarks. Dense Captioning generates detailed descriptions of objects and their spatial relationships, leveraging depth information for better object geometry capture [13][14]. Transformer architectures combined with point cloud networks enhance the alignment of 3D visual features with language representations [15]. Datasets like ScanRefer and ScanNet support model training with rich 3D annotations [16][17].

Large Language Models. Recent advancements in large language models (LLMs) have enabled complex reasoning and conversational understanding, fueled by internet-scale data. Recent work extends LLM capabilities to visual reasoning tasks, advancing multimodal processing [18][19]. Vision-language models like LLaVA [20] use LLMs to generate question-answer pairs from image descriptions, improving performance in 2D tasks. However, research on 3D scene instruction-tuning remains limited. Our work enhances model capabilities by generating triplets of 3D point clouds, visual prompts, and text instructions to improve 3D understanding and reasoning.

III. Methodology

A. Problem Formulation

We aim to train a generalist agent capable of handling various 3D-language tasks using samples from our proposed scaling data paradigm. The agent processes a 3D scene context represented as point clouds, visual prompts such as 3D bounding boxes and instance prompts, and natural language instructions. It needs to understand both the textual instructions and the 3D scene, interpreting spatial and contextual information to generate an appropriate natural language response.

B. Adaptive Multimodal Fusion Paradigm

Our framework implements quality-controlled data augmentation through metric-guided transformations. By integrating multi-source inputs $\mathcal{D} = \{D_{\text{scene}}, D_{\text{QA}}, D_{\text{cap}}\}$ from ScanNet [17], ScanQA [10], and ScanRefer [16], the augmentation pipeline applies task-specific transformations Φ_k followed by metric-based filtering Ψ_k, yielding the final dataset $\mathcal{D}_{\text{final}} = \bigcup_{k=1}^{3} \Phi_k \circ \Psi_k(D_k)$.

1) Semantic Quality Control: To ensure high-quality data generation, we rely on two key metrics. First, semantic similarity is measured as $S_Q(Q_{\text{orig}}, Q_{\text{gen}}) = \cos\big(f_{\text{BERT}}(Q_{\text{orig}}), f_{\text{BERT}}(Q_{\text{gen}})\big)$, where BERT embeddings [21] quantify the alignment between the original and generated questions. Second, we assess semantic consistency via a semantic search approach, defined as $S_{\text{cap}}(C_{\text{orig}}, C_{\text{gen}}) = \frac{1}{n}\sum_{i=1}^{n} \text{NLI}(C_{\text{orig}}, C_{\text{gen}}^{(i)})$, which leverages RoBERTa inference [22] to evaluate the correctness of caption-derived QA pairs. Task-specific thresholds, determined from human-annotated statistics as $\tau_k = \mu_k + 1.96\sigma_k$, are set to $\tau_{\text{QA}} = 0.82$ for QA tasks and $\tau_{\text{cap}} = 0.77$ for captioning.

2) Augmentation Architectures: For QA generation, we apply transformations such as synonym replacement, logical reversal, and order shuffling, with relevance scoring computed as $\text{rel}(A, Q) = \sigma(\mathbf{W}_r[f_{\text{BERT}}(A); f_{\text{BERT}}(Q)])$, where σ denotes the sigmoid function. Caption-to-QA conversion employs a T5 model [23], where the generated question $Q_{\text{gen}} = \text{T5}(C \oplus \mathcal{P}_{\text{template}})$ is produced using 32 handcrafted

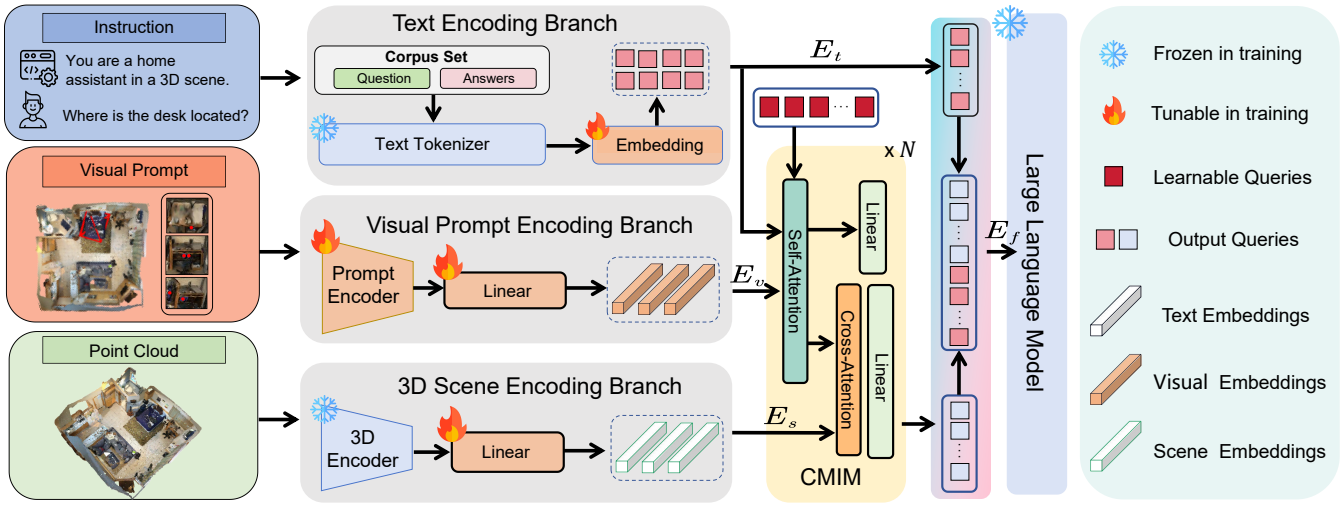

Fig. 3. The architecture processes multi-modal inputs, combining natural language and 3D scene data for reasoning in 3D environments. It features three main components: Multi-Modal Embedding, Cross-Modal Interaction, and LLM Decoder. **CMIM**: Cross-Modal Interaction Module.

templates $\mathcal{P}_{\text{template}}$, and answers are projected via $\mathbf{W}_{\text{ans}} \in \mathbb{R}^{d\times|\mathcal{V}|}$.

3) Multimodal Integration: For scene-to-QA generation, we adopt CogVLM [24], where QA likelihood is computed through 3D-text cross-attention as $p(Q, A \mid S) = \text{softmax}(\text{CrossAtt}(\mathbf{E}_{\text{3D}}, \mathbf{E}_{\text{text}}))$. The final dataset $\mathcal{D}_{\text{final}}$ integrates over 62,000 QA pairs and 73,000 caption annotations, surpassing the original ScanQA and ScanRefer datasets.

C. Model Architecture

The proposed architecture establishes a hierarchical fusion framework for embodied reasoning, comprising three core components: 1) Multi-Modal Embedding Module, 2) Cross-Attention Fusion Module, and 3) LLM Decoder. As formalized in Figure. 3, the processing pipeline $\mathcal{R} = \text{LLM}_{\text{dec}}(\mathcal{F}_{\text{fusion}}(\mathcal{E}_t(T) \parallel \mathcal{E}_v(V) \parallel \mathcal{E}_s(S)))$ integrates text T, visual prompts V, and 3D scene S through tensor concatenation.

1) Multi-Modal Embedding Module:

- **Text Encoding**: For input tokens $T = \{w_i\}_{i=1}^{L}$, standard Transformer encoding generates embeddings $\mathbf{E}_t = \text{Transformer}_{\text{enc}}(\text{Embed}(T) + \mathbf{P}_t) \in \mathbb{R}^{L\times d}$ with positional encoding $\mathbf{P}_t$.
- **Visual Prompt Encoding**: Spatial guidance signals $V = \{v_j\}_{j=1}^{N_v}$ (3D bounding boxes $v_j = (\mathbf{c}_j, \mathbf{d}_j) \in \mathbb{R}^6$) are obtained through:
  1) User annotations
  2) Mask3D detector outputs

  Encoded via $\mathbf{E}_v = \text{MLP}(\text{Flatten}(V)) \in \mathbb{R}^{N_v\times d}$ to establish spatial priors.
- **3D Scene Encoding**: Point cloud $S \in \mathbb{R}^{N_p\times 3}$ is processed by Vote2Cap-DETR++:

$$\mathbf{E}_s = \text{Vote2Cap-DETR++}(S) \in \mathbb{R}^{N_s\times d} \tag{1}$$

2) Cross-Attention Fusion Module: The fusion module implements three-stage feature integration:

1. **Cross-Modal Alignment**:

$$\mathbf{E}_{f1} = \text{softmax}\left(\frac{\mathbf{E}_t\mathbf{W}_q(\text{Concat}(\mathbf{E}_v, \mathbf{E}_s)\mathbf{W}_k^\top}{\sqrt{d}}\right)\cdot(\mathbf{E}_v \parallel \mathbf{E}_s)\mathbf{W}_v \tag{2}$$

2. **Context Preservation**: Self-attention processes $[\mathbf{E}_{f1}; \mathbf{E}_t]$ through Transformer layers to maintain linguistic coherence while enabling adaptive fusion control [25].
3. **Residual Fusion**:

$$\mathbf{E}_f = \text{LayerNorm}(\text{TransformerLayer}(\mathbf{E}_{f1} \parallel \mathbf{E}_t) + \mathbf{E}_t) \tag{3}$$

preserving instructional details via residual connections [26].

3) LLM Decoder: The decoder implements visual-grounded generation through dynamic prefix projection:

$$\mathbf{h}_t = \text{LLM}(\mathbf{p}_{1:t}; \text{LinearProj}(\mathbf{E}_f)) \tag{4}$$

where visual-spatial context flows via projected prefix embeddings to condition token probabilities $p(w_{t+1}) = \text{softmax}(\mathbf{h}_t\mathbf{W}_{\text{vocab}})$.

IV. Experiments

A. Implementation Details

We trained our model on the ScanQA [10] dataset and object descriptions from ScanRefer [16], extending the data with an additional 36,437 QA pairs and 36,635 captioning instances. Generated textual responses were evaluated using BLEU [27], ROUGE [28], METEOR [29], and CIDEr [30] metrics. The LL3DA framework [31] was employed, with 40,000 points randomly sampled from each 3D scene as input. The model used the frozen OPT-1.3B [32] as the LLM decoder and Vote2Cap-DETR [33] as the object detector in the visual prompt encoder. Training was conducted over 10,000 iterations using the AdamW optimizer [34], with a cosine annealing learning rate schedule, on two Nvidia RTX 3090 GPUs with a batch size of 24.

TABLE I

COMPARISON OF RESULTS ON SCANQA FOR RELATED WORK. **MULTIMODAL COMBINATION:** THE GENERATED DATA COMES FROM SCANQA, SCANREFER TEXT ANNOTATION DATA, AND SCANNET SCENE DATA. **PARAMS**: MODEL PARAMETER SIZE.

Method	Version	Params	Validation			
			CIDEr↑	BLEU-4↑	METEOR↓	ROUGE-L↓
3D-LLM[35]	flamingo	3B	59.20	7.20	12.20	32.30
	BLIP2-opt	3B	63.80	9.40	13.80	34.00
	BLIP2-flant5	3B	69.40	12.00	14.50	35.70
LL3DA[31]	scratch	1.3B	74.80	13.68	15.40	36.25
	fine-tuned	1.3B	76.79	13.53	15.88	37.31
GPT-4[36]	GPT Blind	1800B	53.59	3.81	13.54	30.92
	Vocab-agnostic	1800B	34.22	0.98	8.75	20.03
	Vocab-grounded	1800B	58.32	1.63	14.23	33.43
Gen3DQA[1]	single object	-	64.91	10.52	13.62	33.39
	multiple objects	-	64.51	10.21	13.68	32.84
3DMIT[37]	Vicuna-7b	-	44.38	6.44	10.40	24.64
	LLaVA1.5+IMG	-	46.42	5.98	10.64	24.46
	Vicuna-7b	-	48.03	5.24	10.70	26.22
NaviLLM[38]	Vicuna-7b	7B	75.9	12.5	15.40	38.40
Chat-3D[39]	Chat-3D	7B	53.20	6.40	11.90	28.50
	Chat-3D V2	7B	77.10	7.30	**16.10**	**40.10**
3D-MoRe(Ours)	multimodal combination	1.3B	**78.94**	**14.17**	16.07	37.89

TABLE II

COMPARISON OF RESULTS ON SCANREFER FOR RELATED WORK. †:THE DATASET USED COMES FROM MULTIMODAL COMBINATION.

Method	Params	ScanRefer			
		CIDEr@0.5↑	BLEU-4@0.5↑	METEOR@0.5↓	ROUGE-L@0.5↓
Vote2Cap-DETR[2]	60M	61.19	34.46	**26.22**	54.40
3D-VisTA[40]	0.12B	61.60	34.10	26.80	**55.00**
X-Trans2Cap[15]	0.12B	43.87	25.05	22.46	45.28
UniT3D[14]	-	46.69	27.22	21.91	45.98
3D-VLP[41]	-	54.94	32.31	24.83	51.51
LL3DA[31]	1.3B	62.76	35.00	25.68	54.23
MORE[42]	-	40.94	22.93	21.66	44.42
3D-MoRe†(ours)	1.3B	**64.08**	**35.52**	25.90	54.49

B. Comparison With Leading Methods

Table.I and Table.II present the performance of our approach with existing methods. Here, "Versions" refer to different model architectures or experimental techniques. Our model consistently outperformed previous methods, especially on the validation set, using the CIDEr metric as the primary indicator. Notably, our approach surpassed the generation-based Chat-3D V2 model [39], improving the CIDEr@0.5 score by 1.84%. Additionally, in a comparison on the ScanRefer dataset, our model outperformed LL3DA and other prior approaches, showing a 2.15% improvement in CIDEr performance on low-parameter LLMs.

C. Ablation Study

Data Augmentation and Filtering Strategies

In this study, we adopted the Adaptive Multimodal Fusion Paradigm (Sec. III-B) to expand our datasets for 3D Question Answering and 3D Dense Captioning tasks. To ensure a fair comparison among different data generation methods, we uniformly sampled 3,000 instances from each of the three types—QA Generation (derived from question-answer pairs), Captioning Generation (from dense captions), and Scene Generation (based on image scene data)—implemented via the Sentence Transformers framework. For 3D Question Answering, additional question-answer pairs were generated and subsequently filtered using semantic search to remove low-quality data based on their alignment with the original pairs, whereas for 3D Dense Captioning, semantic similarity measures were employed to select high-quality captions. As demonstrated in Tables III , our combined data augmentation and filtering approach achieved the highest CIDEr scores of 78.43% and 63.21% for 3D Question Answering and Dense Captioning, respectively, while consistently improving performance across all augmented data. Our code and dataset have been made openly available.

Effectiveness of the Visual Prompt. Table IV shows that incorporating object detection as a visual prompt significantly improves performance. Without it, key metrics in 3D Question Answering (CIDEr, BLEU-4, METEOR, ROUGE-L) and 3D Dense Captioning (CIDEr@0.5, ME-

TABLE III

COMPARISON OF DATASET TYPES USING 28K SAMPLES FOR 3DQA AND 39K FOR 3D DENSE CAPTIONING (CONTROL THE COMPARISON QUANTITY,FULL RESULTS IN TABLE I). **QUALITY CONTROL**: 3DQA USES SEMANTIC SEARCH; 3D DENSE CAPTIONING USES SEMANTIC SIMILARITY. **COMBINATION GEN**: A MIX OF QA, CAPTIONING, AND SCENE GENERATION METHODS.

Dataset	Quality Control	Core Metrics				Data Size
		CIDEr(QA) CIDEr@0.5(DC)	BLEU-4(QA) BLEU-4@0.5(DC)	METEOR(QA) METEOR@0.5 (DC)	ROUGE-L(QA) ROUGE-L@0.5(DC)	
3D Question Answering (28K Data)						
ScanQA	-	76.79	13.53	**15.88**	37.31	25K
ScanQA + QA Gen	✗	77.81	13.69	15.58	37.67	28K
	✓	78.18	14.52	15.65	37.82	28K
ScanQA + Captioning Gen	✗	77.97	14.21	15.55	37.63	28K
	✓	78.24	14.50	15.60	37.66	28K
ScanQA + Scene Gen	✗	77.74	14.61	15.60	37.59	28K
	✓	78.32	14.59	15.73	37.78	28K
ScanQA + Combination Gen	✓	**78.43**	**14.62**	15.75	**37.84**	28K
3D Dense Captioning (39K Data)						
ScanRefer	-	62.76	35.00	25.68	54.23	36K
ScanRefer + QA Gen	✗	62.71	34.89	25.41	54.12	39K
	✓	62.77	35.16	25.70	54.28	39K
ScanRefer + Captioning Gen	✗	63.08	35.11	25.29	54.01	39K
	✓	63.15	35.23	25.69	54.33	39K
ScanRefer + Scene Gen	✗	62.75	35.14	25.53	54.17	39K
	✓	62.79	35.21	25.74	54.30	39K
ScanRefer + Combination Gen	✓	**63.21**	**35.33**	**25.77**	**54.38**	39K

TABLE IV

THE IMPACT OF OBJECT DETECTION ON EXPERIMENTAL OUTCOMES.

Task	Object Detection	Evaluation Criteria			
		C*	B-4*	M*	R*
3D Question Answering	✗	77.58	14.11	15.85	37.70
	✓	78.69	14.07	16.04	38.04
3D Dense Captioning	✗	63.26	36.44	25.99	54.81
	✓	63.74	36.16	26.16	54.97

TEOR@0.5, ROUGE-L@0.5) decline noticeably. These results underscore that object detection is vital for capturing spatial context and generating accurate, contextually relevant descriptions, thereby enhancing multimodal reasoning in 3D environments.

D. Qualitative Results

We evaluated our model's performance in 3D Dense Captioning and 3D Question Answering (see Figure 4) by analyzing successful and challenging cases. This examination clarified the model's spatial reasoning strengths and limitations, particularly in crowded scenes. We further improved consistency using adaptive learning rates, task-specific fine-tuning, and controlled sampling (max_token, top-k, top-p).

V. CONCLUSIONS

We introduce a novel data generation paradigm to overcome data scarcity and limited diversity. By synthesizing 62,000 QA pairs and 73,000 object descriptions from ScanNet, ScanQA, and ScanRefer, our approach enriches training data and enhances performance in vision-and-language tasks. Furthermore, our model effectively encodes 3D point clouds with attention mechanisms and object detection cues, reducing ambiguities and establishing a robust framework for future 3D vision and language integration.

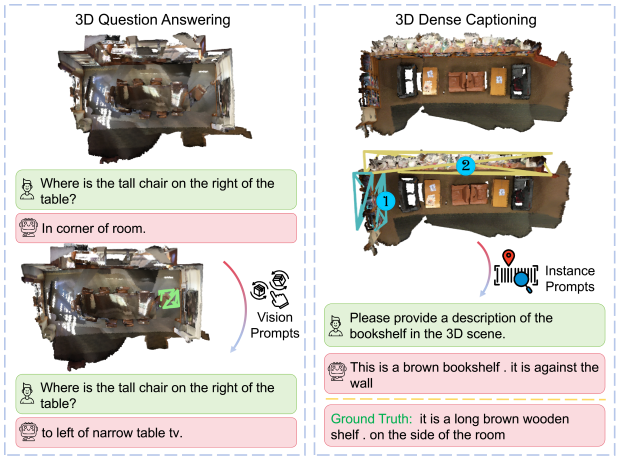

Fig. 4. Vision and instance prompts enhance object localization and differentiation, improving model accuracy in 3D captioning and question answering.

REFERENCES

[1] M. M. Dwedari, M. Niessner, and D. Z. Chen, "Generating context-aware natural answers for questions in 3d scenes," *arXiv preprint arXiv:2310.19516*, 2023.

[2] S. Chen, H. Zhu, M. Li, X. Chen, P. Guo, Y. Lei, Y. Gang, T. Li, and T. Chen, "Vote2cap-detr++: Decoupling localization and describing for end-to-end 3d dense captioning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[3] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim, "Visual prompt tuning," in *European Conference on Computer Vision*. Springer, 2022, pp. 709–727.

[4] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, and D. C. Schmidt, "A prompt pattern catalog to enhance prompt engineering with chatgpt," *arXiv preprint arXiv:2302.11382*, 2023.

[5] A. Falcon, G. Serra, and O. Lanz, "A feature-space multimodal data augmentation technique for text-video retrieval," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 4385–4394.

[6] W. Hu, Y. Xu, Y. Li, W. Li, Z. Chen, and Z. Tu, "Bliva: A simple multimodal llm for better handling of text-rich visual questions," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 3, 2024, pp. 2256–2264.

[7] P. Vadlapati, "Autopuredata: Automated filtering of web data for llm fine-tuning," *arXiv preprint arXiv:2406.19271*, 2024.

[8] J. Ma, P. Wang, D. Kong, Z. Wang, J. Liu, H. Pei, and J. Zhao, "Robust visual question answering: Datasets, methods, and future challenges," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[9] J. Luo, J. Fu, X. Kong, C. Gao, H. Ren, H. Shen, H. Xia, and S. Liu, "3d-sps: Single-stage 3d visual grounding via referred point progressive selection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 454–16 463.

[10] D. Azuma, T. Miyanishi, S. Kurita, and M. Kawanabe, "Scanqa: 3d question answering for spatial scene understanding," in *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 19 129–19 139.

[11] X. Yan, Z. Yuan, Y. Du, Y. Liao, Y. Guo, S. Cui, and Z. Li, "Comprehensive visual question answering on point clouds through compositional scene manipulation," *IEEE Transactions on Visualization & Computer Graphics*, no. 01, pp. 1–13, 2023.

[12] L. Zhao, D. Cai, J. Zhang, L. Sheng, D. Xu, R. Zheng, Y. Zhao, L. Wang, and X. Fan, "Toward explainable 3d grounded visual question answering: A new benchmark and strong baseline," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 6, pp. 2935–2949, 2022.

[13] T. Yu, X. Lin, S. Wang, W. Sheng, Q. Huang, and J. Yu, "A comprehensive survey of 3d dense captioning: Localizing and describing objects in 3d scenes," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[14] Z. Chen, R. Hu, X. Chen, M. Nießner, and A. X. Chang, "Unit3d: A unified transformer for 3d dense captioning and visual grounding," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 18 109–18 119.

[15] Z. Yuan, X. Yan, Y. Liao, Y. Guo, G. Li, S. Cui, and Z. Li, "X-trans2cap: Cross-modal knowledge transfer using transformer for 3d dense captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8563–8573.

[16] D. Z. Chen, A. X. Chang, and M. Nießner, "Scanrefer: 3d object localization in rgb-d scans using natural language," in *European conference on computer vision*. Springer, 2020, pp. 202–221.

[17] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5828–5839.

[18] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma *et al.*, "Scaling instruction-finetuned language models," *Journal of Machine Learning Research*, vol. 25, no. 70, pp. 1–53, 2024.

[19] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin *et al.*, "Opt: Open pre-trained transformer language models, 2022," *URL https://arxiv.org/abs/2205.01068*, vol. 3, pp. 19–0, 2023.

[20] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, 2024.

[21] J. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[22] L. Yinhan, O. Myle, G. Naman, D. Jingfei, J. Mandar, C. Danqi, L. Omer, and L. Mike, "Roberta: A robustly optimized bert pretraining approach (2019)," *arXiv preprint arXiv:1907.11692*, pp. 1–13, 2019.

[23] M. T. Pilehvar and J. Camacho-Collados, "Wic: the word-in-context dataset for evaluating context-sensitive meaning representations," *arXiv preprint arXiv:1808.09121*, 2018.

[24] W. Hong, W. Wang, Q. Lv, J. Xu, W. Yu, J. Ji, Y. Wang, Z. Wang, Y. Dong, M. Ding *et al.*, "Cogagent: A visual language model for gui agents," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 281–14 290.

[25] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[27] K. Papinesi, "Bleu: A method for automatic evaluation of machine translation," in *Proc. 40th Actual Meeting of the Association for Computational Linguistics (ACL), 2002*, 2002, pp. 311–318.

[28] L. Chin-Yew, "Rouge: A package for automatic evaluation of summaries," in *Proceedings of the Workshop on Text Summarization Branches Out, 2004*, 2004.

[29] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.

[30] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.

[31] S. Chen, X. Chen, C. Zhang, M. Li, G. Yu, H. Fei, H. Zhu, J. Fan, and T. Chen, "Ll3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 428–26 438.

[32] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin *et al.*, "Opt: Open pre-trained transformer language models," *arXiv preprint arXiv:2205.01068*, 2022.

[33] S. Chen, H. Zhu, X. Chen, Y. Lei, G. Yu, and T. Chen, "End-to-end 3d dense captioning with vote2cap-detr," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 11 124–11 133.

[34] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[35] Y. Hong, H. Zhen, P. Chen, S. Zheng, Y. Du, Z. Chen, and C. Gan, "3d-llm: Injecting the 3d world into large language models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 20 482–20 494, 2023.

[36] S. Singh, G. Pavlakos, and D. Stamoulis, "Evaluating zero-shot gpt-4v performance on 3d visual question answering benchmarks," *arXiv preprint arXiv:2405.18831*, 2024.

[37] Z. Li, C. Zhang, X. Wang, R. Ren, Y. Xu, R. Ma, and X. Liu, "3dmit: 3d multi-modal instruction tuning for scene understanding," *arXiv preprint arXiv:2401.03201*, 2024.

[38] D. Zheng, S. Huang, L. Zhao, Y. Zhong, and L. Wang, "Towards learning a generalist model for embodied navigation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 624–13 634.

[39] H. Huang, Z. Wang, R. Huang, L. Liu, X. Cheng, Y. Zhao, T. Jin, and Z. Zhao, "Chat-3d v2: Bridging 3d scene and large language models with object identifiers," *arXiv preprint arXiv:2312.08168*, 2023.

[40] Z. Zhu, X. Ma, Y. Chen, Z. Deng, S. Huang, and Q. Li, "3d-vista: Pre-trained transformer for 3d vision and text alignment," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2911–2921.

[41] Z. Jin, M. Hayat, Y. Yang, Y. Guo, and Y. Lei, "Context-aware alignment and mutual masking for 3d-language pre-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 984–10 994.

[42] Y. Jiao, S. Chen, Z. Jie, J. Chen, L. Ma, and Y.-G. Jiang, "More: Multi-order relation mining for dense captioning in 3d scenes," in *European Conference on Computer Vision*. Springer, 2022, pp. 528–545.