Multi-View Images Suffice 3D Reasoning Through Chain-of-Thought Selection and Question-Guided Fusion

Boqiang Xu, Jinlin Wu, Wei Zhang, Chenyang Su, Jian Liang, *Member, IEEE*, Zhenan Sun, *Senior Member, IEEE*, and Zhen Lei, *Fellow, IEEE*

***Abstract*—3D reasoning is crucial in areas like robotics and autonomous driving. Due to the high cost of 3D data acquisition, some recent methods attempt to enable LLMs to perform 3D reasoning through multi-view images, thereby transferring the powerful 2D reasoning capabilities of LLMs to 3D environments. However, these methods face challenges: either they use redundant views that contain many perspectives irrelevant to the question, or they rely on globally aggregated multi-view representations, losing the fine-grained vision-language correlations. To tackle these challenges, we propose 3DMulti-LLM, which mainly consists of three components: a COT selector, a question-guided fusion block, and pre-trained LLMs. Specifically, first, the COT selector leverages the powerful chain-of-thought reasoning capabilities of LLMs to identify question-related multi-view images. In this way, 3DMulti-LLM can eliminate a substantial amount of interference from unnecessary viewpoints. Then, we propose a question-guided fusion block for integrating multi-view features via question-guided interaction among various viewpoints. Finally, the pre-trained LLMs are utilized to reason in 3D scenes directly through multi-view features. Notably, our approach understands the 3D scene solely through multi-view images, without requiring the input of point cloud information or additional 3D feature extraction. Through our experiments, 3DMulti-LLM achieves impressive performance and surpasses existing 3D-input-free methods by + 12.2% and + 7.1% on ScanQA and 3DMV-VQA datasets, respectively.**

***Index Terms*—Large language models, image analysis, commonsense reasoning.**

Received 28 June 2025; revised 13 October 2025, 11 December 2025, 16 January 2026, and 5 March 2026; accepted 31 March 2026. Date of publication 17 April 2026; date of current version 22 April 2026. This work was supported by the InnoHK Program and the National Natural Science Foundation of China under Grant 62276256, Grant 62306313, Grant U2441251, and Grant U1836217. The associate editor coordinating the review of this article and approving it for publication was Prof. Yong Man Ro. *(Corresponding authors: Wei Zhang; Zhen Lei.)*

Boqiang Xu and Jinlin Wu are with the Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing 100190, China, and also with the Centre for Artificial Intelligence and Robotics, Hong Kong Institute of Science & Innovation, Chinese Academy of Sciences, Hong Kong, China (e-mail: boqiang.xu@cripac.ia.ac.cn; jinlin.wu@nlpr.ia.ac.cn).

Wei Zhang is with Hunan University, Changsha 410082, China (e-mail: wzhang0716@outlook.com).

Chenyang Su is with the Centre for Artificial Intelligence and Robotics, Hong Kong Institute of Science & Innovation, Chinese Academy of Sciences, Hong Kong, China (e-mail: chenyansu5-c@my.cityu.edu.hk).

Jian Liang and Zhenan Sun are with the New Laboratory of Pattern Recognition and the State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS), Beijing 100049, China (e-mail: liangjian92@gmail.com; znsun@nlpr.ia.ac.cn).

Zhen Lei is with the State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing 100190, China, also with the School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS), Beijing 100049, China, also with the Centre for Artificial Intelligence and Robotics, Hong Kong Institute of Science & Innovation, Chinese Academy of Sciences, Hong Kong, China, and also with the School of Computer Science and Engineering and the Faculty of Innovation Engineering, Macau University of Science and Technology, Macau, China (e-mail: zhen.lei@ia.ac.cn).

Digital Object Identifier 10.1109/TIP.2026.3683282

I. Introduction

3D REASONING plays a pivotal role in fields such as robotics [5], [22] and autonomous driving [42], [44], where embodied agents are expected to comprehend the 3D layout and perform planning and navigation tasks based on human instructions. Reasoning within 3D environments encompasses rich concepts such as spatial relationships, interactions, affordances, physics, etc [17]. This enables intelligent agents to better emulate human-like reasoning, enhancing their utility in complex, real-world applications. Recently, Large Language Models (LLMs) have been shown to possess strong understanding and reasoning capabilities for 2D scenes [2], [24], sparking a growing interest in extending these reasoning abilities to 3D environments.

Some approaches [6], [18] attempt to extract 3D features from point clouds and feed them into LLMs for reasoning. However, the acquisition of point cloud data is costly, and embedding 3D features into 2D pre-trained LLMs inevitably causes feature space misalignment, resulting in the loss of crucial 3D information. To tackle these challenges, several efforts [17], [53] have been made to enable LLMs to understand 3D scenes through multi-view images, thereby transferring their 2D reasoning capabilities to 3D contexts. As illustrated in Fig. 1, when provided with appropriate views, the information they offer allows for a complete understanding of the 3D scene and the accurate inference of the correct answer. Building upon this, Zheng et al. [53] extract features from multi-view images, perform multi-view fusion using a transformer encoder, and then input the result into an LLM along with task-specific instructions to generate responses. References [17] and [18] construct scene representations from multi-view images and align them with the image encoder of 2D pre-trained LLMs using a proposed alignment loss. However, the aforementioned methods have two main issues: 1) **Multi-view redundancy**, a typical 3D scene often includes hundreds of multi-view

Question Type	Question	Answer
Concept	Is there a piano?	Yes.
Counting	How many cushions are there on the sofa?	One.
Relation	In what part of the room is the piano located?	In the corner of the room.
Comparison	Which is closer to the long sofa, the piano or the wooden chair?	The wooden chair.

Fig. 1. An example scene of 3D reasoning from multi-view images. We have demonstrated four typical types of questions in the 3D-MVQA dataset. Multi-view images from appropriate perspectives provide sufficient information for a complete understanding of the 3D scene and the accurate inference of the correct answer.

images, many of which are irrelevant to the posed question and contain noise, potentially hindering model performance. 2) **Multi-view fusion**, previous globally aggregated multi-view representations fail to preserve fine-grained vision-language correlations, leading to inefficient utilization of the pretrained 2D LLM.

To tackle the two issues above, we propose a novel model called 3DMulti-LLM. For the multi-view redundancy, we introduce a COT selector to filter noise through hundreds of multi-view images. The COT selector harnesses the powerful chain-of-thought reasoning abilities of LLMs to quantitatively evaluate the relevance of different images to the query. This selector identifies the images most relevant to the input question, filtering out noise and enhancing the model's performance. For the multi-view fusion, we introduce a learnable question-guided fusion block for integrating multi-view features in a question-guided manner. Specifically, we extract the compact question-aware global feature of the 3D scene via a Q-Former [24] structure. Based on the global representation, the optimized multi-view features are generated via a linear layer. In this way, each view is equipped with the question-aware overall understanding and also combines complementary information from different perspectives. Finally, the question and multi-view features are input to the pre-trained LLMs for precise 3D reasoning.

We conduct experiments on ScanQA [4] dataset and observe that directly inputting multi-view images into BLIP2 [24], followed by fine-tuning, results in an accuracy rate of only 29.7%. However, with the help of our proposed COT selector and question-guided fusion, which address the aforementioned multi-view redundancy and multi-view fusion issues when inputting multi-view images into LLMs, the accuracy rate is enhanced to 41.9%. This performance is comparable to other state-of-the-art methods that require 3D point cloud inputs, demonstrating that our approach can effectively understand and reason about 3D scenes using only multi-view images, without the need for additional 3D feature extraction.

The contributions of our paper are summarized as follows:

- We propose 3DMulti-LLM to empower pre-trained LLMs with 3D reasoning capabilities through multi-view images and without the need for 3D feature extraction or point cloud data inputs.
- We design a COT selector to identify question-relevant multi-view images, effectively filtering out unnecessary viewpoints and minimizing interference, thereby enhancing the model's performance.
- We introduce a question-guided fusion block for integrating multi-view features via question-guided interaction among various viewpoints and improving the performance of 3D reasoning.
- Comprehensive experiments are conducted on ScanQA [4] and 3DMV-VQA [17], which indicate 3DMulti-LLM's potential for 3D reasoning.

II. Related Work

A. Multi-Modal Large-Language Pre-Trained Models

Our work is related to multi-modal large-language pre-trained models that integrate images with natural language [10], [13], [21], [24], [36]. Several studies [21], [36] have focused on training models from scratch using extensive image-language datasets, subsequently applying these models to downstream tasks such as visual question answering [14], [48], captioning [7], and referring expression comprehension [47] through fine-tuning. Alternatively, some researchers have combined pre-trained vision models with pre-trained large language models (LLMs) using adaptable neural modules like Perceiver [3] and QFormers [24], taking advantage of the

perceptual strengths found in pre-trained vision models and the reasoning and generalization abilities inherent in LLMs. Inspired by these previous works, we attempt to apply the powerful reasoning capabilities of pre-trained multi-modal LLMs to the 3D world. This is not trivial and we have to overcome challenges such as the scarcity of 3D text-image pairs and the issue of adapting 2D pre-trained LLMs to reason within the 3D world.

B. 3D Reasoning With Large Language Models

Recently, there have been several works [17], [18], [31], [50] attempting to integrate LLMs into the 3D world, leveraging the powerful reasoning capabilities of LLMs for 3D reasoning. Some of them [17], [18] leverage 3D features to facilitate LLMs' understanding of 3D scenes. 3D-LLM [18] takes 3D point clouds and their features as input, and then uses 2D LLMs as backbones to train the model. With the integration of a 3D localization mechanism, 3D-LLM enhances its ability to grasp 3D spatial information. 3D-CLR [17] firstly learns a compact 3D scene representation from multi-view images. Then, the model obtains per-pixel 2D features and aligns these features with the 3D representation via a 3D-2D alignment loss. Finally, the reasoning process is executed through a collection of neural reasoning operators. Despite the utility of 3D features in these methods, the acquisition of point cloud data is expensive, and incorporating 3D features into 2D pre-trained LLMs inevitably results in feature space misalignment, leading to the loss of critical 3D information.

To tackle these challenges, some approaches [31], [50] have explored enabling LLMs to understand 3D scenes through multi-view images. Agent3D-Zero [50] leverages an LLM to determine the extrinsic parameters of the camera for the most informative viewpoint using the Bird's Eye View (BEV) image of the scene and subsequently renders the corresponding multi-view images as input. However, directly inferring the camera's extrinsic parameters using an LLM is challenging and often yields suboptimal results. BridgeQA [31] employs the pre-trained BLIP [25] model for image-text retrieval to identify the most relevant multi-view image corresponding to a given question. However, image-text retrieval approach treats viewpoint selection as a simple semantic similarity matching task, lacking the logical reasoning capabilities. As a result, the selected viewpoints often provide superficial matches, containing relevant objects but insufficient comprehensive 3D scene information for accurate reasoning. Different from these methods, our COT selector leverages the chain-of-thought reasoning capabilities of LLMs to select question-related multi-views. As a result, the selection is more accurate, which better facilitates LLMs in understanding 3D scenes through multi-view images.

III. Methodology

A. Overview

In this section, we introduce the structure of our 3DMulti-LLM. To fully leverage the capabilities of the pre-trained 2D LLMs, we use multi-view images of the scene as input and leverage a COT selector to choose relevant views for addressing specific questions. Then, we propose the question-guided fusion block to optimize the extracted multi-view features. Finally, the pre-trained LLMs are utilized for 3D reasoning directly through multi-view features. The structure of the 3DMulti-LLM is illustrated in Fig. 2.

B. Selecting Question-Related Viewpoints

We introduce the COT selector, which leverages the chain-of-thought reasoning capabilities of LLMs to select question-related viewpoints.

1) Prompt Design: The purpose of the COT selector is to utilize LLMs to choose question-related perspectives from numerous multi-view images, thereby reducing interference from irrelevant perspectives. We harness the powerful chain-of-thought reasoning capabilities of LLMs to assess whether a specific multi-view image is relevant to addressing the question posed. To achieve this, our approach requires the COT selector to assign each image a relevance score, and finally, we rank and filter the most relevant multi-view images based on the scores.

Specifically, we select the question-related films by constructing a prompt template as:

Question: Imagine yourself in a real room.

First, think about the steps necessary to answer the question {question}. (### Do not display steps)

Then, considering those steps and the provided image, how likely is the image to be related to answering the question on a scale from 0 to 100? (### Display score)

Answer: [Score]

where {*question*} is replaced by the input question. For example, given the question of *"What color is the chair near the table?"*, we complete the command as:

Question: Imagine yourself in a real room.

First, think about the steps necessary to answer the question *"What color is the chair near the table?"*. (### Do not display steps)

Then, considering those steps and the provided image, how likely is the image to be related to answering the question on a scale from 0 to 100? (### Display score)

Answer: [Score]

2) Multi-View Selection: By this command, we obtain LLMs' response as *"I would rate this picture as **90** in terms of its relevance to answering the question"*. Finally, we will rank all the scores and select the top N films with scores greater than 0 as the images most relevant to the question. Importantly, through our experiments, we found that when the COT selector is required to display intermediate reasoning steps while computing the relevance score, it tends to favor assigning excessively high scores and attempts to justify them, even though many of the justifications are tenuous and unreasonable. This leads to low reliability of the resulting relevance scores. In contrast, when the intermediate reasoning steps are hidden, the scores are more objective and accurate. Therefore, we ultimately chose to hide the intermediate reasoning steps in the COT selector and directly output the relevance score. We present several examples of multi-view image selection in Fig. 3. As shown in Fig. 3 (a), the cushions on the sofa are clearly visible, so the COT Selector assigns it a high relevance

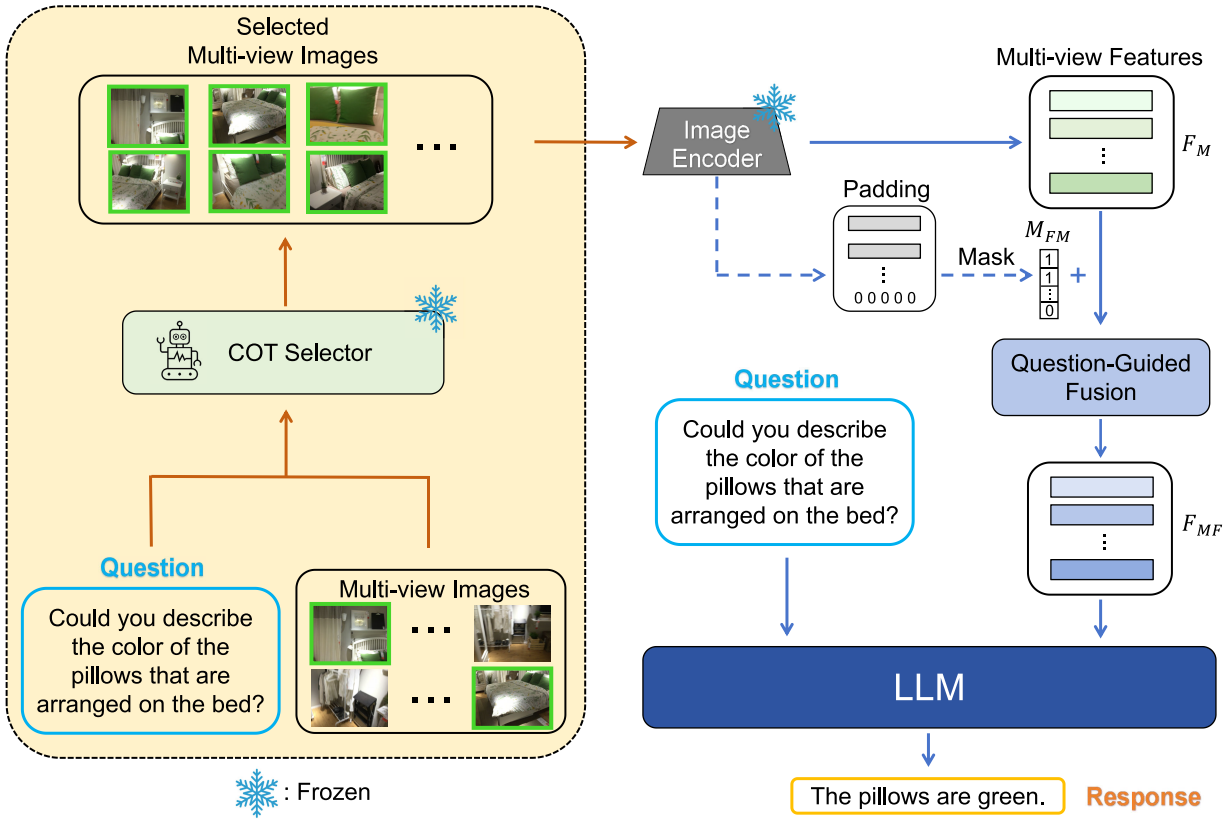

Fig. 2. **Overall pipeline of 3DMulti-LLM.** We use multi-view images of the scene as input. Firstly, a COT selector is employed to choose question-related multi-view images in an offline preprocessing step. After that, we extract multi-view features, and when the number of selected multi-view images is insufficient, we pad the multi-view features and compute the attention mask M_{FM} during this process. We also propose a question-guided fusion block to boost inter-view communication. Finally, the features and questions are input to the large language model to generate responses.

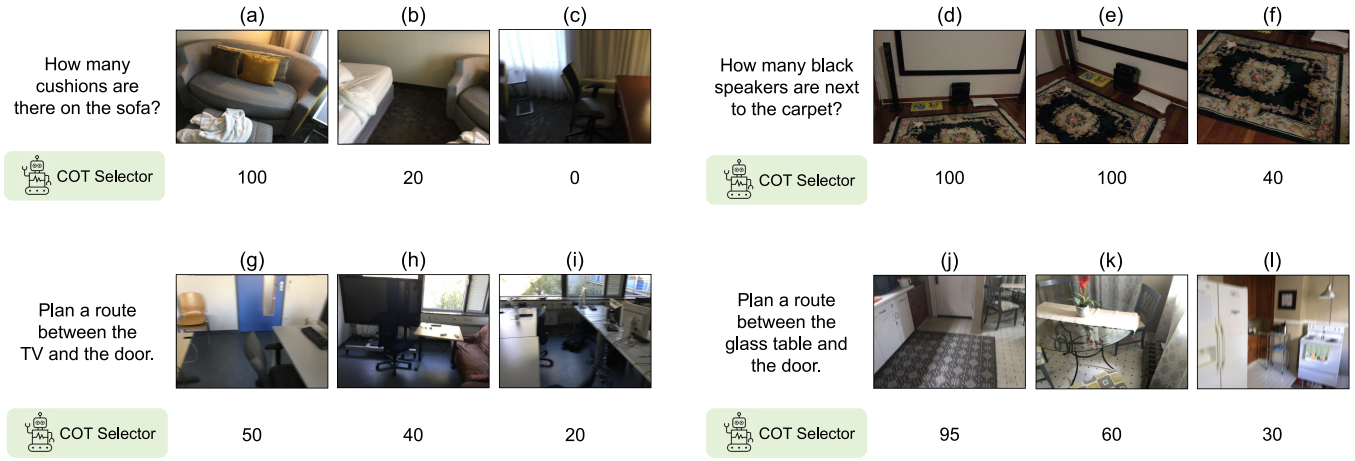

Fig. 3. **Examples of the COT selector.** The number below each image represents the relevance score given by the COT selector, with a maximum score of 100. A higher score indicates a stronger relevance between the image and the question. The results demonstrate that the COT selector can accurately filter question-related images by leveraging the powerful chain-of-thought reasoning capabilities of LLMs.

score of 100. In contrast, in Fig. 3 (b), only a small portion of the cushions are visible, and thus the COT Selector assigns it a lower score of 20. Similarly, in Fig. 3 (j), both the 'glass table' and the 'door' mentioned in the question are visible, leading the COT Selector to assign a high score of 95. In contrast, Fig. 3 (k) shows only the 'glass table', offering less support for route-planning compared to Fig. 3 (j), and thus the COT Selector assigns a lower score of 60 to it. The results have demonstrated the COT selector's strong capacity for chain-of-thought reasoning, as well as its ability to accurately select multi-view images relevant to the question.

C. Obtaining Multi-View Features

We use frozen and pre-trained CLIP [36] as our image encoder to extract multi-view features F_M. If the number of multi-view images selected by the COT selector, denoted by M, is less than N, we will pad $F_M \in \mathbb{R}^{B\times M\times C\times HW}$ with 0 to

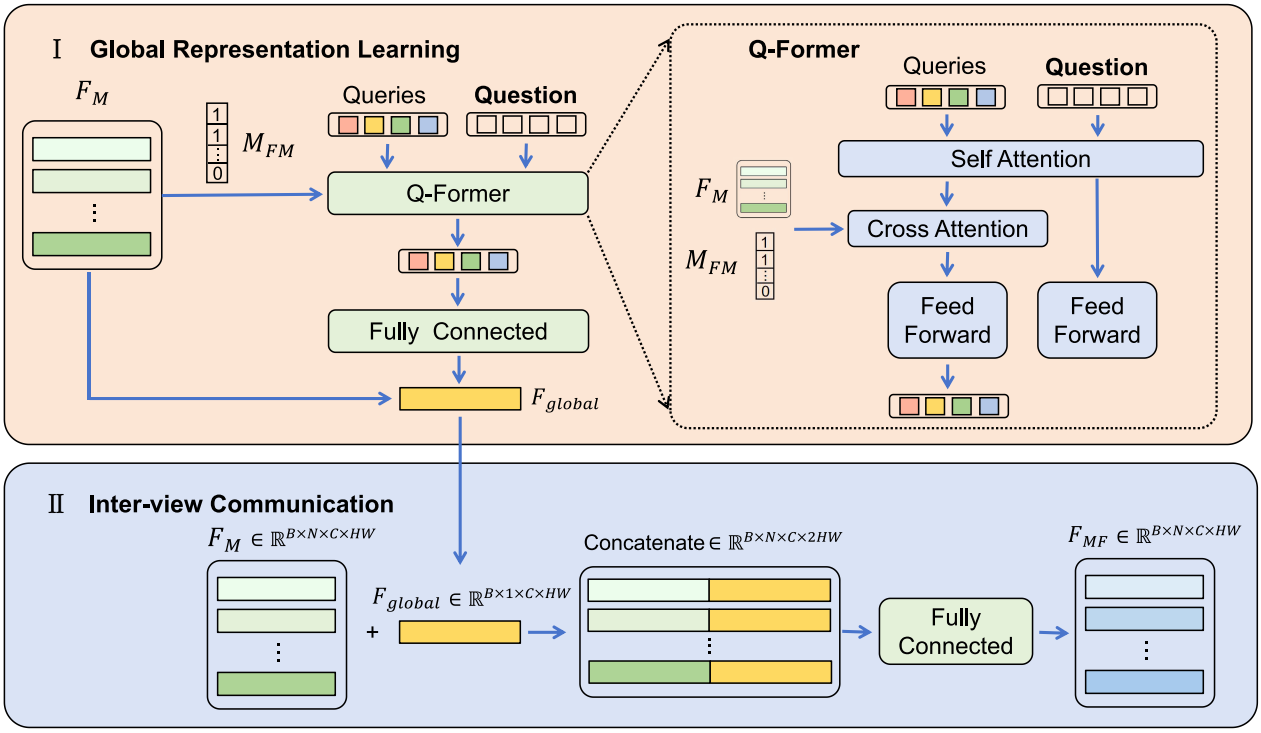

Fig. 4. **Detailed structure of the proposed question-guided fusion block.** The question-guided fusion block consists of two stages. Given multi-view features F_M and the corresponding attention mask M_{FM}, we first extract the global representation through a Q-former [24] structure. Next, the global representation is fused into the multi-view features and the inter-view communication is enhanced via a fully connected layer.

$F_M \in \mathbb{R}^{B\times N\times C\times HW}$, where B, N, C, H, W respectively indicate the batch size, number of views, channels, the height, and the width of features. During this process, we also calculate the attention masks $M_{FM} \in \{0,1\}^{B\times N}$ of F_M based on the position of the padding, where 0 indicates that the corresponding view is padded.

D. Question-Guided Multi-View Fusion

To integrate multi-view features F_M for better 3D comprehension, we introduce a question-guided fusion block, which strengthens inter-view interactions and facilitates more comprehensive feature integration. The question-guided fusion block consists of two stages. In the first stage, we employ a Q-Former [24] structure to extract a question-aware global feature. In the second stage, we propose a global-guided fusion strategy to enhance inter-view communications, facilitating the fusion of features from different perspectives. The structure of the question-guided fusion block is illustrated in Fig. 4.

1) Global Representation Learning: To obtain the global representation, we employ a Q-Former [24] architecture, which adapts to the variable number of valid views in F_M. Our goal is to derive a question-aware global feature. To achieve this, we leverage the InstructBLIP [10] approach in the design of the Q-Former.

To be specific, we generate a set of learnable query embeddings, which engage with the multi-view features F_M via cross-attention layers utilizing the attention mask M_{FM}. Moreover, the query embeddings can also interact with the question T_q through the self-attention layers for extracting question-aware global feature. F_{global} is obtained via the Q-Former followed by a fully connected layer, formulated as

$$F_{global} = (\text{Q-Former}(F_M, M_{FM}, T_q))W_1^T, \tag{1}$$

where T_q is the posed question, W_1^T is the weight of the fully connected layer, $F_M \in \mathbb{R}^{B\times N\times C\times HW}$, $M_{FM} \in \{0,1\}^{B\times N}$ and $F_{global} \in \mathbb{R}^{B\times 1\times C\times HW}$. By this aggregation, features from various perspectives blend into a summative representation in a manner sensitive to the posed question.

2) Inter-View Communication: To boost the inter-view communication and blend question-aware overall understanding into multi-view features, we concatenate the global feature and multi-view features, then input them into a two-layer fully connected layer for fusion. Firstly, we broadcast $F_{global} \in \mathbb{R}^{B\times 1\times C\times HW}$ to $F'_{global} \in \mathbb{R}^{B\times N\times C\times HW}$. Next we obtain optimized multi-view features F_{MF} as

$$F_{MF} = \text{ReLU}(\text{Concate}(F_M, F'_{global})W_2^T)W_3^T, \tag{2}$$

where Concate$(\cdot)$ denotes the concatenation of F_M and F'_{global} along the HW dimention, $F_{MF} \in \mathbb{R}^{B\times N\times C\times HW}$ and $W_2 \in \mathbb{R}^{c\times 2HW}$, $W_3 \in \mathbb{R}^{HW\times c}$ are the weights of a two-layer fully connected network. On the one hand, F_{MF} integrates global-guided question-aware features into F_M to enhance the overall understanding of the 3D scene, leading to improved view-wise predictions. On the other hand, through this global-guided fusion method, inter-view communication is strengthened, enabling complementary information from different perspectives.

TABLE I
EXPERIMENTAL RESULTS ON SCANQA [4] DATASET. CHECKMARKS INDICATE THAT THE MODEL OPERATES WITHOUT 3D POINT CLOUD INPUT. THE BEST RESULT IS MARKED IN **BOLD**

Methods	3D-input-free	Backbone	BLEU-1	BLEU-2	BLEU-3	METEOR	EM
ScanQA[4]		PointNet++[35]+LSTM[15]	30.2	20.4	15.1	13.1	21.0
SIG3D[29]		OpenScene-LSeg[33]+SBERT-MPNet[38]	39.5	-	-	13.4	-
LL3DA[6]		3DETR[30]+OPT-1.3B[51]	-	-	-	15.9	-
3D-LLM[18]		gradslam[20]/NeRF[39]+BLIP2 Vit-g FlanT5-XL[24]	39.3	25.2	18.4	14.5	20.5
BridgeQA[31]		VoteNet[34]+BLIP-L[25]+GPT4[1]	29.2	18.4	16.5	14.1	26.9
LEO[19]		PointNet++[35]+Vicuna-7B[32]	-	-	-	20.0	24.5
3D-LLaVA [11]		3D U-Net[8]+LLaVA-1.5-7B[27]	-	-	-	18.4	-
LLaVA-3D [54]		LLaVA-Video-7B[52]	-	-	-	**20.8**	**30.6**
3UR-LLM [43]		3DETR[30]+Flan-T5-XL[9]	**43.7**	**30.1**	**22.1**	18.4	21.5
LLaVA[28](Zero-Shot)	✓	LLaVA-13B[28]	7.1	2.6	0.9	10.5	0.0
BLIP2-flant5[24]	✓	BLIP2 Vit-g FlanT5-XL[24]	29.7	16.2	9.8	11.3	13.6
Flamingo[2]	✓	Flamingo-9B[2]	25.6	15.2	9.2	11.3	18.8
Agent3D-Zero[50]	✓	GPT4-V[1]	28.6	-	-	16.0	17.5
NaviLLM[53]	✓	EVA-CLIP-02-Large[40]+Vicuna-7B[32]	-	-	-	15.4	23.0
3DMulti-LLM (Ours)	✓	BLIP2 Vit-g FlanT5-XXL[24]+BLIP2 Vit-g FlanT5-XL[24]	**41.9**	**27.4**	**19.3**	**16.7**	**24.3**

IV. EXPERIMENTS

A. Datasets

1) ScanQA: The ScanQA [4] is a specialized dataset designed for 3D reasoning. It's built upon 3D scans of indoor environments and each entry in the ScanQA [4] pairs a 3D representation with one or more natural language questions about the environment. The questions are designed to probe the model's understanding of spatial relationships, object properties, and the ability to integrate visual and textual information in 3D contexts. The ScanQA [4] consists of 25,563 questions and 562 3D scenes in the training set. We test on the val set of ScanQA, which contains 4,675 questions and 71 3D scenes.

2) 3DMV-VQA: The 3DMV-VQA [12] dataset is designed for reasoning in 3D environments, leveraging a combination of 3D mesh data and multi-view images to facilitate the understanding of complex spatial relationships and object characteristics within 3D scenes. 3DMV-VQA [12] dataset includes 5k 3D scenes from the Habitat-Matterport 3D Dataset (HM3D) [37] dataset, and around 600k images rendered from these 3D environments. 3DMV-VQA [12] dataset encompasses four types of questions: concept, counting, relation, and comparison.

3) CLEVER3D-REAL: The CLEVER3D-REAL [45] dataset is derived from 3RScan [41] dataset for 3D VQA task. The training and test sets include 38,806 and 5,765 questions from 1,176 and 157 scenes, respectively. The questions are grouped into six types.

B. Experimental Settings

1) Evaluation Metrics: We report BLEU-1, BLEU-2, BLEU-3, METEOR, and EM scores on the ScanQA [4] dataset, and present the visual question answering accuracy on the 3DMV-VQA [12] dataset for four types of questions separately.

2) Implementation Details: In our experiments, we used four A100 GPUs with 80 GB memory each. We train the network for 50 epochs with a batch size of 36. Our COT selector employs BLIP2 Vit-g FlanT5-XXL to select views relevant to the question and was conducted using deterministic greedy decoding ($do\_sample = False, temperature = 1.0$). The number of multi-view features N is set to be 80. We use BLIP2 Vit-g FlanT5-XL as the backbone, initializing the model from 3D-LLM (BLIP2-flant5) checkpoints released in [18], and finetune the parameters for the QFormer. Our loss function is the same as that of BLIP2 [24], with no additional loss functions added. Training on ScanQA, 3DMV-VQA, and CLEVR3D took approximately 90, 160, and 150 GPU-hours, respectively, while inference consumed about 0.4, 0.9, and 0.5 GPU-hours. For ScanQA, 3DMV-VQA, and CLEVR3D, the multi-view selector took on average 55 seconds, 120 seconds, and 101 seconds per scene during multi-view images selection, and each question required scoring an average of 18, 142, and 62 multi-view images, respectively. We perform batched scoring in our implementation, where multiple candidate images per question are scored in parallel. This allows us to process on average 8 images per second.

C. Quantitative Analysis

1) Performance on ScanQA: We report our results on ScanQA [4] dataset in Table I. We compare with various representative baseline models. For a fair comparison, we demonstrate the backbone of each method. Particularly,

TABLE II

EXPERIMENTAL RESULTS ON 3DMV-VQA [12] DATASET. CHECKMARKS INDICATE THAT THE MODEL OPERATES WITHOUT 3D POINT CLOUD INPUT. QUESTION-ANSWERING ACCURACY ON DIFFERENT QUESTION TYPES IS REPORTED. THE BEST RESULT IS MARKED IN **BOLD**

Methods	3D-input-free	Backbone	Concept	Counting	Relation	Comparison	Overall
3D-Feature+LSTM[15]		3D-CNN+LSTM[15]	61.2	22.4	49.9	61.3	48.2
3D-LLM[18]		gradslam[20]/NeRF[39]+BLIP2 Vit-g FlanT5-XL[24]	**68.9**	**32.4**	**61.6**	**68.3**	**58.6**
3D-CLR[17]	✓	NeRF[39]+CLIP-LSeg[23]	66.1	**41.3**	57.6	72.3	57.7
NS-VQA[46]	✓	ResNet-34[16]+LSTM[15]	59.8	21.5	33.4	61.6	38.0
Flamingo[2]	✓	Flamingo-9B[2]	60.0	18.3	40.2	61.4	41.6
BLIP2-flant5[24]	✓	BLIP2 Vit-g FlanT5-XL[24]	61.9	21.1	48.0	62.3	47.1
3DMulti-LLM (Ours)	✓	BLIP2 Vit-g FlanT5-XXL[24]+BLIP2 Vit-g FlanT5-XL[24]	**75.5**	30.5	**64.9**	**84.2**	**64.8**

ScanQA [4] and **SIG3D** [29] do not use LLMs. We conduct a zero-shot evaluation of **LLaVA** [28] by leveraging its pre-trained model and providing it with a single randomly selected image. **3D-LLM** [18] employs a 3D feature extractor to obtain 3D features from rendered multi-view images, subsequently utilizing a 2D VLM as the backbone for model training. **BridgeQA** [31] utilizes the pre-trained BLIP [25] model for image-text retrieval to select the most relevant multi-view image for the given question. It also designs a Twin-Transformer architecture to separately extract 2D and 3D features. **Pretrained LLMs** utilize multi-view images as the input of the pre-trained LLMs, and then finetune on ScanQA [4] dataset. **Agent3D-Zero** [50] utilizes an LLM to identify the most informative viewpoint based on the BEV (Bird's Eye View) image of the scene and renders the corresponding multi-view images as input. **NaviLLM** [53] is a model for embodied navigation. It utilizes the agent's historical observations of the scene as input, where these historical observations are based on multi-view 2D images. **LEO** [19] is an embodied, multimodal generalist agent that operates in the 3D world. **3D-LLaVA** [11] proposes an Omni Superpoint Transformer (OST) to perform visual token selection, visual prompt encoding, and mask decoding. **LLaVA-3D** [54] employs joint 2D and 3D vision–language representations for instruction tuning. **3UR-LLM** [43] directly takes 3D point clouds as input and projects 3D features fused with textual instructions into a compact set of tokens. We report BLEU, METEOR and EM scores for robust answer matching.

Specifically, the results show that compared to other 3D-input-free methods, 3DMulti-LLM achieves the best performance across all evaluation metrics. 3DMulti-LLM significantly outperforms other 3D-input-free methods by at least 12.2%, 11.2%, 9.5%, 0.7% and 1.3% in BLEU-1, BLEU-2, BLEU-3, METEOR and EM respectively. 3DMulti-LLM also achieves comparable performance to methods that require 3D point cloud input. The result shows that our model could perform state-of-the-art 3D reasoning with simply multi-view 2D images as input, without the need for point clouds or 3D coordinates.

2) Performance on 3DMV-VQA: We report our results on 3DMV-VQA [12] dataset in Table II. We compare with various representative baseline models. To ensure a fair comparison, we present the backbone architecture for each method. Particularly, 3D-Feature+LSTM [15], 3D-CLR [17] and NS-VQA [46] do not use LLMs. **3D-CLR** [17] leverages NeRF [39] and CLIP-Lseg [23] to learn 3D and 2D features, respectively, from multi-view images, subsequently using a 3D-2D alignment loss to assign features to the 3D compact representation. **NS-VQA** [46] is a 2D version of 3D-CLR model. **3D-Feature+LSTM** utilizes 3D features obtained from 3D-2D alignment, with voxel grids downsampled using a 3D-CNN as input. These are concatenated with language features from the LSTM and then fed into an MLP.

Specifically, the results indicate that 3DMulti-LLM outperforms all other methods in accuracy across all question types except for 'Counting'. 3DMulti-LLM significantly surpasses other methods by at least 6.6%, 3.3%, and 15.9% in "Concept", "Relation" and "Comparison" types of questions, and overall exceeds other methods by 6.2%. The results indicate that our model achieves state-of-the-art 3D reasoning performance by utilizing only multi-view 2D images as input, without the need for point clouds or 3D coordinates.

3) Performance on CLEVER3D-REAL: We report our results on CLEVER3D-REAL [45] dataset in Table III and we report results for each type separately. As shown in Table III, 3DMulti-LLM outperforms prior methods by at least 1.8% in overall accuracy. This result further demonstrates the generalizability of our model and shows that 3DMulti-LLM achieves state-of-the-art 3D reasoning using only multi-view 2D images as input, with no need for point clouds or 3D coordinates.

D. Qualitative Analysis

In Fig. 5, we present some qualitative examples. To verify whether our method can truly understand the 3D world through multi-view images, in Fig. 5 (a), we specifically removed images where both the chair and bedside lamp appear simultaneously from the selected images. The experimental results show that, despite the chair and bedside lamp not appearing together in any selected images, 3DMulti-LLM is still able to correctly determine their positional relationship. These examples demonstrate that our model can precisely

TABLE III

EXPERIMENTAL RESULTS ON CLEVER3D-REAL DATASET. CHECKMARKS INDICATE THAT THE MODEL OPERATES WITHOUT 3D POINT CLOUD INPUT. THE BEST RESULT IS MARKED IN **BOLD**

Method	3D-input-free	Backbone	Existence	Counting	Compare Integer	Query Attr.	Query Object	Compare Attr.	Overall
Number			*474*	*1,386*	*602*	*2,339*	*355*	*609*	*5,765*
ReferIt3D		PointNet+++RNN	85.4	26.6	57.5	38.3	11.3	**58.5**	41.8
ScanQA		PointNet+++LSTM	84.8	26.3	55.6	**41.6**	13.2	54.7	42.6
TransVQA3D		PointNet+++BERT	**88.2**	**30.7**	**63.1**	37.3	**21.7**	57.1	**43.7**
BEV+MCAN	✓	MnasNet	85.7	27.3	60.0	39.5	14.4	53.7	42.5
3DMulti-LLM (Ours)	✓	BLIP2 Vit-g FlanT5-XXL+BLIP2 Vit-g FlanT5-XL	**89.1**	**30.9**	**61.5**	**42.2**	**22.8**	**55.3**	**45.5**

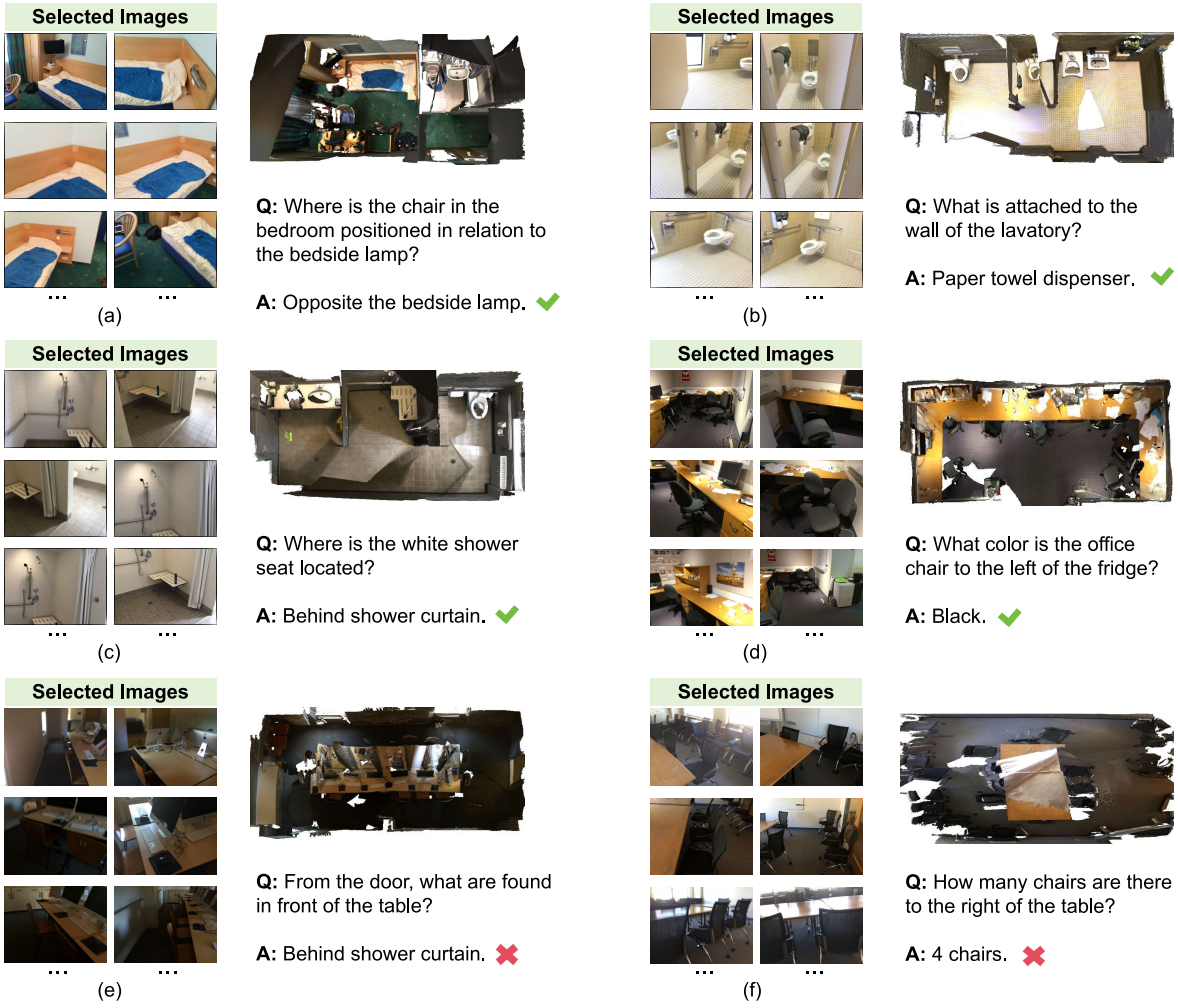

Fig. 5. Qualitative examples of our 3DMulti-LLM. We can see that 3DMulti-LLM can effectively comprehend 3D scenes and answer most questions correctly. However, it still fails sometimes, mainly because it is difficult to count accurately when the counting problem involves images from multiple perspectives.

select multi-view images related to the question and accurately understand and infer 3D scenes from multi-view images. However, 3DMulti-LLM also fails on some questions. It cannot accurately count the instances when the counting problem involves images from multiple perspectives sometimes. This difficulty primarily arises from the challenge of cross-view instance association: when the same object appears in different views, the model must correctly recognize and associate them as one instance. However, our current architecture lacks explicit geometric or pose information to guide cross-view alignment. As a result, the model may either double-count the same object across views or fail to aggregate instances consistently. This architectural limitation highlights a key area for improvement, suggesting that incorporating

TABLE IV

ABLATION STUDY CONDUCTED ON THE SCANQA DATASET. 'FC LAYERS', 'Q-FORMER' AND 'SELF-ATTENTION' INDICATE WE REPLACE QUESTION-GUIDED FUSION BLOCK WITH FC LAYERS, Q-FORMER AND SELF-ATTENTION MODULE FOR MULTI-VIEW INTERACTION. 'IMAGE-TEXT RETRIEVAL SELECTION' REFER TO REPLACING THE COT SELECTOR WITH VIEWPOINT SELECTION METHODS BASED ON IMAGE-TEXT RETRIEVAL

Index	Method	ScanQA						
		BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	EM
I	w/o COT selector	33.5	20.2	13.5	8.6	12.5	30.2	11.5
	w/o question-guided fusion	38.3	23.2	15.4	9.8	14.5	35.1	16.6
II	w/o Instruction Tunning	38.9	23.5	15.2	10.7	14.6	35.4	16.6
III	FC Layers	37.5	23.3	15.8	10.4	16.4	34.3	14.4
	Q-Former	36.5	23.3	15.3	10.5	15.8	33.8	14.3
	Self-Attention	36.1	23.5	14.7	10.1	14.3	32.9	13.8
IV	Image-Text Retrieval Selection	35.3	22.5	14.9	10.1	14.3	32.1	13.5
	VinVL	33.5	20.8	13.1	9.2	13.1	31.2	11.8
	ALBEF	34.2	21.8	14.2	9.8	13.6	31.3	12.2
V	llama3-llava-next-8b	41.5	27.2	18.8	12.8	16.5	38.5	18.2
	Qwen2.5-VL-7B	41.2	26.9	19.1	12.5	16.2	38.9	18.5
	3DMulti-LLM	**41.9**	**27.4**	**19.3**	**12.9**	**16.7**	**39.1**	**18.8**

explicit geometric cues or object-level cross-view matching could mitigate such failures in future work.

E. Ablation Studies

To explore the effectiveness of each design in 3DMulti-LLM, we conduct extension ablation studies on the ScanQA [4] and 3DMV-VQA [12] datasets.

1) Effectiveness of Each Component: We conduct ablation studies to evaluate the effectiveness of each component on the ScanQA [4] and 3DMV-VQA [12] datasets, as shown in Index I of Table IV and Table V, respectively. 'w/o COT selector' indicates we replace COT selector with random selection. The results indicate that removing the COT selector and the question-guided fusion block leads to an 8.4% and 3.6% decrease in BLEU-1 on the ScanQA dataset, and a 11.3% and 5.3% decrease in overall accuracy on the 3DMV-VQA dataset, respectively. This demonstrates the effectiveness of both the COT selector and the question-guided fusion block.

2) Effectiveness of Instruction Tuning: We conduct ablation studies to evaluate the impact of Instruction Tuning in the question-guided fusion block on the ScanQA [4] and 3DMV-VQA [12] dataset, as shown in Index II of Table IV and Table V, respectively. We train the question-guided fusion block without instruction tuning, *i.e.*, only queries are inputted without inputting the question to the Q-former in Fig. 4, for comparison. The results indicate that omitting instruction tuning leads to a 3.0% decrease in BLEU-1 on the ScanQA dataset and a 2.2% drop in overall accuracy on the 3DMV-VQA dataset. The experimental results demonstrate the importance of extracting question-related global features for enhancing the model's 3D understanding capabilities.

3) Effectiveness of the Question-Guided Fusion Block: To validate the impact of the question-guided fusion block, we conduct comparisons by replacing it with FC layers, Q-Former, and the self-attention module, respectively. The results are shown in Index III of Table IV and Table V. Among them, 'Q-Former' performs feature fusion by applying cross-attention between learnable queries and multi-view features, whereas 'Self-Attention' enables view-to-view interaction by allowing the multi-view features to attend to each other. From the results, we can find that replacing the question-guided fusion block with FC layers, Q-Former, and Self-Attention leads to a 4.4%, 5.4%, and 5.8% decrease in BLEU-1 on the ScanQA dataset, and a 2.0%, 3.3%, and 4.5% decrease in overall accuracy on the 3DMV-VQA dataset, respectively. The results demonstrate the effectiveness of the question-guided fusion block, showing that incorporating question-aware overall understanding into multi-view features is crucial for the QA task. Moreover, question-guided interaction across different views enables a more accurate and question-specific understanding of 3D scenes from 2D images compared to simple fully connected, Q-Former, or self-attention fusion.

4) Effectiveness of the COT Selector: We conducted ablation experiments on the COT selector. Some approaches [31] employ pre-trained LLMs for image-text retrieval to select the most relevant multi-view image for a given question. To validate the effectiveness of our COT selector, we conduct comparative experiments by replacing it with image-text retrieval selection method and two cross-encoder methods (*i.e.*, VinVL [49], ALBEF [26]) for viewpoint selection. The results are shown in Index IV of Table IV and Table V. From the results, we can find that the overall accuracy drops 6.6%, 8.4%, and 7.7% in BLEU-1 on the ScanQA dataset and 3.6%, 4.7%, and 4.0% in overall accuracy on the 3DMV-VQA dataset, for Image-Text Retrieval Selection, VinVL, and ALBEF, respectively. The results demonstrate the

TABLE V

ABLATION STUDY CONDUCTED ON THE 3DMV-VQA DATASET. 'FC LAYERS', 'Q-FORMER' AND 'SELF-ATTENTION' INDICATE WE REPLACE QUESTION-GUIDED FUSION BLOCK WITH FC LAYERS, Q-FORMER AND SELF-ATTENTION MODULE FOR MULTI-VIEW INTERACTION. 'IMAGE-TEXT RETRIEVAL SELECTION' REFER TO REPLACING THE COT SELECTOR WITH VIEWPOINT SELECTION METHODS BASED ON IMAGE-TEXT RETRIEVAL

Index	Method	3DMV-VQA				
		Concept	Counting	Relation	Comparison	Overall
I	w/o COT selector	66.9	24.3	56.8	68.2	53.5
	w/o question-guided fusion	72.5	28.5	59.8	72.8	59.5
II	w/o Instruction Tuning	73.1	27.2	63.8	72.6	62.6
III	FC Layers	72.2	29.6	63.7	74.7	62.8
	Q-Former	70.5	28.3	64.2	73.3	61.5
	Self-Attention	69.8	27.6	63.1	72.1	60.3
IV	Image-Text Retrieval Selection	71.1	29.5	61.2	78.9	61.2
	VinVL	70.2	29.3	59.5	76.7	60.1
	ALBEF	70.8	30.2	60.8	77.5	60.8
V	llama3-llava-next-8b	75.2	30.4	64.5	84.0	64.5
	Qwen2.5-VL-7B	75.0	30.1	64.6	83.7	64.3
VI	w/o Counting-Type Questions	75.2	9.2	61.2	76.1	59.4
	3DMulti-LLM	**75.5**	**30.5**	**64.9**	**84.2**	**64.8**

effectiveness of the COT selector, showing that leveraging the chain-of-thought reasoning capabilities of LLMs enables more accurate selection of question-related viewpoints. This, in turn, enhances the model's capacity for 3D reasoning from multi-view images.

Furthermore, to verify the generalizability of the COT selector, we replaced the VLLM in the COT selector with llama3-llava-next-8b and Qwen2.5-VL-7B while keeping the backbone fixed for comparative experiments. The results are shown in Index V of Table IV and Table V. From the results, we can find that the COT selector with BLIP2 Vit-g FlanT5-XXL achieves the best performance. Moreover, the COT selector combined with other VLLMs also demonstrates comparable performance, which indicates the strong generalizability of the COT selector. The underlying reason is that we select N = 80 multi-view images for each question, which provides considerable tolerance for the VLLMs. As long as the selected 80 multi-view images contain sufficient valid information, the 3DMulti-LLM is able to generate accurate responses.

5) Effectiveness of the Number of Selected Multi-View Images: We perform a hyperparameter study on the number of selected multi-view images N on the ScanQA and 3DMV-VQA datasets, and the results are shown in Fig. 6. From the results, we can find that the impact of N on model performance exhibits a similar trend on both the ScanQA and 3DMV-VQA datasets. Initially, as the number of input multi-view images increases, more useful information is introduced, leading to a gradual improvement in performance. The model achieves its best performance when $N = 80$. However, when N exceeds 80, the inclusion of additional multi-view images introduces more noise, causing performance to decline. Additionally, the influence of N diminishes as the dataset size increases. The 3DMV-VQA dataset is approximately ten times larger than

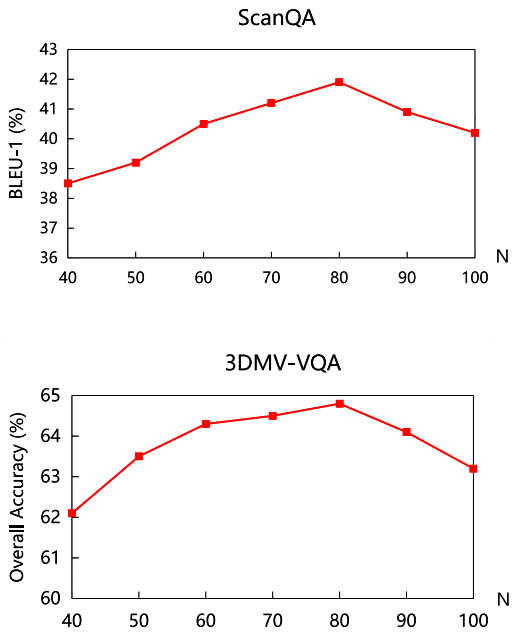

Fig. 6. Effect of the number of selected multi-view images N on the ScanQA and 3DMV-VQA datasets.

ScanQA. When N ranges from 60 to 90, the variation in BLEU-1 remains within roughly 2% on ScanQA, whereas the fluctuation is within about 1% on 3DMV-VQA. Moreover, both datasets exhibit optimal performance at $N = 80$. Therefore, for plug-and-play scenarios, we recommend setting $N = 80$ as a robust and generally effective choice.

6) Computational Efficiency of the COT Selector: We conduct ablation studies to investigate the computational efficiency of the COT selector on the 3DMV-VQA dataset. We con-

duct experiments using four NVIDIA A100 (80G) cards. To begin with, we filter out similar images within the same scene based on image features, and then employ the COT selector to choose the multi-view images relevant to the given question. Importantly, this preprocessing is carried out entirely offline and therefore does not affect the actual training or testing speed. The preprocessing takes approximately 110 hours. Additionally, unlike 3D-LLM [18], which relies on Direct Voxel Grid Optimization (DVGO) [39] to build 3D representations from multi-view images, our method significantly reduces the time required for constructing 3D features. DVGO takes approximately 10 minutes to process a single scene. Given that the 3DMV-VQA dataset includes 3,150 scenes, our approach saves approximately 415 hours compared to 3D-LLM [18].

7) Human–VLM Agreement and Inter-View Consistency of the COT Selector: We sampled 50 scenes from the test split to conduct a human study quantifying Human-VLM agreement. Five annotators assigned image–question relevance scores. We then averaged the human ratings and computed Spearman correlation ρ and Kendall-Tau τ between the averaged human scores and the COT Selector's scores. The results show that on the ScanQA dataset, $\rho = 0.56$ and $\tau = 0.48$; on the 3DMVVQA dataset, $\rho = 0.53$ and $\tau = 0.45$. The results demonstrate a positive Human–VLM agreement, suggesting that the COT Selector's scoring aligns reasonably well with human judgments.

Additionally, we quantify inter-view consistency by sampling 50 scenes, each comprising five near-duplicate views. For each scene, we obtain the COT Selector's score and compute the within-set standard deviation SD, the average pairwise differences APD, and the consistency ratio CR_{10} under a predefined tolerance of $\delta = 10$ on a 0 to 100 scale. The results show that on the ScanQA dataset, $\overline{\text{SD}} = 5.5$, $\overline{\text{APD}} = 4.5$ and $\overline{\text{CR}_{10}} = 86.2\%$; on the 3DMV-VQA dataset, $\overline{\text{SD}} = 7.2$, $\overline{\text{APD}} = 6.8$ and $\overline{\text{CR}_{10}} = 81.3\%$. The results indicate that the COT Selector assigns similar scores to near-duplicate views.

8) The Impact of Counting-Type Questions: In Table II, we observe that 3DMulti-LLM performs relatively poorly on counting-type questions. Therefore, we removed all counting-type questions from the 3DMV-VQA training set to examine their impact on the model's performance on the remaining question types. The results are shown in Index VI of Table V. From the results, we can find that the accuracy drops 0.3%, 3.7%, and 8.1% for the "Concept", "Relation", and "Comparison" question types, respectively. This indicates that although 3DMulti-LLM performs relatively poorly on counting-type questions, training on such questions still enhances the model's overall capability for 3D scene comprehension.

V. Conclusion

In this paper, we propose **3DMulti-LLM**, which utilizes pre-trained LLMs as the backbone and directly conducts 3D reasoning through multi-view images without the need for 3D feature extraction. Via COT selector, 3DMulti-LLM accurately selects question-related multi-view images. To integrate multi-view information for better 3D comprehension, we propose a question-guided fusion block that enhances inter-view communication in a question-guided manner. Finally, the pre-trained LLMs are utilized to reason in 3D scenes directly through multi-view features. Extensive experiments are conducted to demonstrate the superiority of 3DMulti-LLM on 3D reasoning.

References

[1] J. Achiam et al., "GPT-4 technical report," 2023, *arXiv:2303.08774*.

[2] J.-B. Alayrac et al., "Flamingo: A visual language model for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst. 35*, 2022, pp. 23716–23736.

[3] A. Awadalla et al., "OpenFlamingo: An open-source framework for training large autoregressive vision-language models," 2023, *arXiv:2308.01390*.

[4] D. Azuma, T. Miyanishi, S. Kurita, and M. Kawanabe, "ScanQA: 3D question answering for spatial scene understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 19107–19117.

[5] M. Beetz, "AI reasoning methods for robotics," in *Springer Handbook of Robotics*. Cham, Switzerland: Springer, 2016, pp. 329–356.

[6] S. Chen et al., "LL3DA: Visual interactive instruction tuning for omni-3D understanding reasoning and planning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 26428–26438.

[7] X. Chen et al., "Microsoft COCO captions: Data collection and evaluation server," 2015, *arXiv:1504.00325*.

[8] C. Choy, J. Gwak, and S. Savarese, "4D spatio-temporal ConvNets: Minkowski convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3070–3079.

[9] H. W. Chung et al., "Scaling instruction-finetuned language models," *J. Mach. Learn. Res.*, vol. 25, no. 70, pp. 1–53.

[10] W. Dai et al., "InstructBLIP: Towards general-purpose vision-language models with instruction tuning," 2023, *arXiv:2305.06500*.

[11] J. Deng, T. He, L. Jiang, T. Wang, F. Dayoub, and I. Reid, "3D-LLaVA: Towards generalist 3D LMMs with omni superpoint transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 3772–3782.

[12] Y. Etesam, L. Kochiev, and A. X. Chang, "3DVQA: Visual question answering for 3D environments," in *Proc. 19th Conf. Robots Vis. (CRV)*, May 2022, pp. 233–240.

[13] T. Gong et al., "MultiModal-GPT: A vision and language model for dialogue with humans," 2023, *arXiv:2305.04790*.

[14] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the V in VQA matter: Elevating the role of image understanding in visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6325–6334.

[15] A. Graves, "Long short-term memory," in *Supervised Sequence Labelling With Recurrent Neural Networks*, 2012, pp. 37–45.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[17] Y. Hong, C. Lin, Y. Du, Z. Chen, J. B. Tenenbaum, and C. Gan, "3D concept learning and reasoning from multi-view images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 9202–9212.

[18] P. Chen et al., "3D-LLM: Injecting the 3D world into large language models," in *Proc. Adv. Neural Inf. Process. Syst. 36*, 2023, pp. 20482–20494.

[19] J. Huang et al., "An embodied generalist agent in 3D world," in *Proc. 41st Int. Conf. Mach. Learn.*, 2024, pp. 20413–20451.

[20] K. M. Jatavallabhula, G. Iyer, and L. Paull, "∇SLAM: Dense SLAM meets automatic differentiation," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 2130–2137.

[21] C. Jia et al., "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 4904–4916.

[22] C. Landsiedel, V. Rieser, M. Walter, and D. Wollherr, "A review of spatial reasoning and interaction for real-world robotics," *Adv. Robot.*, vol. 31, no. 5, pp. 222–242, Mar. 2017.

[23] B. Li, K. Q. Weinberger, S. Belongie, V. Koltun, and R. Ranftl, "Language-driven semantic segmentation," 2022, *arXiv:2201.03546*.

[24] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," 2023, *arXiv:2301.12597*.

[25] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 12888–12900.
[26] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 9694–9705.
[27] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 26296–26306.
[28] H. Liu, C. Li, Q. Wu, and Y. Jae Lee, "Visual instruction tuning," 2023, *arXiv:2304.08485*.
[29] Y. Man, L. Y. Gui, and Y. X. Wang, "Situational awareness matters in 3D vision language reasoning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2024, pp. 13678–13688.
[30] I. Misra, R. Girdhar, and A. Joulin, "An end-to-end transformer model for 3D object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 2886–2897.
[31] W. Mo and Y. Liu, "Bridging the gap between 2D and 3D visual question answering: A fusion approach for 3D VQA," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 5, 2024, pp. 4261–4268.
[32] B. Peng, C. Li, P. He, M. Galley, and J. Gao, "Instruction tuning with GPT-4," 2023, *arXiv:2304.03277*.
[33] S. Peng, K. Genova, C. Jiang, A. Tagliasacchi, M. Pollefeys, and T. Funkhouser, "OpenScene: 3D scene understanding with open vocabularies," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 815–824.
[34] C. R. Qi, O. Litany, K. He, and L. Guibas, "Deep Hough voting for 3D object detection in point clouds," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9276–9285.
[35] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5105–5114.
[36] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.
[37] S. K. Ramakrishnan et al., "Habitat-matterport 3D dataset (HM3D): 1000 large-scale 3D environments for embodied AI," 2021, *arXiv:2109.08238*.
[38] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," 2019, *arXiv:1908.10084*.
[39] C. Sun, M. Sun, and H.-T. Chen, "Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5449–5459.
[40] Q. Sun, Y. Fang, L. Wu, X. Wang, and Y. Cao, "EVA-CLIP: Improved training techniques for CLIP at scale," 2023, *arXiv:2303.15389*.
[41] J. Wald, A. Avetisyan, N. Navab, F. Tombari, and M. Niessner, "RIO: 3D object instance re-localization in changing indoor environments," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7657–7666.
[42] S. Wang et al., "OmniDrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning," 2024, *arXiv:2405.01533*.
[43] H. Xiong, Y. Zhuge, J. Zhu, L. Zhang, and H. Lu, "3UR-LLM: An end-to-end multimodal large language model for 3D scene understanding," *IEEE Trans. Multimedia*, vol. 27, pp. 2899–2911, 2025.
[44] J.-R. Xue, J.-W. Fang, and P. Zhang, "A survey of scene understanding by event reasoning in autonomous driving," *Int. J. Autom. Comput.*, vol. 15, no. 3, pp. 249–266, Jun. 2018.
[45] X. Yan et al., "Comprehensive visual question answering on point clouds through compositional scene manipulation," *IEEE Trans. Vis. Comput. Graph.*, vol. 30, no. 12, pp. 7473–7485, Dec. 2023.
[46] K. Yi, J. Wu, C. Gan, A. Torralba, P. Kohli, and J. Tenenbaum, "Neural-symbolic VQA: Disentangling reasoning from vision and language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1039–1050.
[47] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, "Modeling context in referring expressions," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 69–85.
[48] P. Zhang, Y. Goyal, D. Summers-Stay, D. Batra, and D. Parikh, "Yin and Yang: Balancing and answering binary visual questions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5014–5022.
[49] P. Zhang et al., "VinVL: Revisiting visual representations in vision-language models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5575–5584.
[50] S. Zhang et al., "Agent3D-zero: An agent for zero-shot 3D understanding," 2024, *arXiv:2403.11835*.
[51] S. Zhang et al., "OPT: Open pre-trained transformer language models," 2022, *arXiv:2205.01068*.
[52] Y. Zhang et al., "LLaVA-video: Video instruction tuning with synthetic data," 2024, *arXiv:2410.02713*.
[53] D. Zheng, S. Huang, L. Zhao, Y. Zhong, and L. Wang, "Towards learning a generalist model for embodied navigation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 13624–13634.
[54] C. Zhu et al., "LLaVA-3D: A simple yet effective pathway to empowering LMMs with 3D capabilities," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2025, pp. 4295–4305.