Robust 3D Visual Question Answering via Bias Learning

Aihua Mao, Shuyi Wen, Feng Chen, Ran Yi, and Yong-Jin Liu, *Senior Member, IEEE*

***Abstract*—Visual question answering (VQA) tasks have witnessed significant advancements in recent years. So far, enhancing the robustness of models on diverse datasets and improving their performance in *3D environments* remains a challenging research direction. In this paper, we propose a high-performance framework called Bias3D-VQA for 3D-VQA based on generative adversarial networks via bias learning, addressing the inherent biases that arise from the model's dependency on dataset-specific patterns or tendencies during training. Such biases often lead the model to focus on more frequently occurring but incorrect answers. Our framework comprises a target model, a bias model, and a generative adversarial component. In each training iteration, we employ an alternating training approach for the target and bias models. When training the bias model, fake point cloud data is generated from random noise, and then we accumulate biases present in language modality and various modules through adversarial training. When training the target model, both the question and the 3D point cloud are inputted into the bias model simultaneously, and the output of the bias model is utilized to correct the loss of the target model. Our approach (Bias3D-VQA) is the first to focus on enhancing model robustness by addressing diverse biases in the 3D-VQA domain. Our target model demonstrates superior performance compared to state-of-the-art models, showing significant improvements in classification accuracy and text generation quality. Notably, in the metrics such as EM@1 and CIDEr, our model even surpasses some pre-trained models with large additional datasets. The source code is available at https://github.com/coderr727/bias_3DQA**

***Index Terms*—3D-VQA, robustness, generative adversarial networks, bias learning.**

I. Introduction

In RECENT years, research on vision-language (V-L) tasks has developed rapidly, driven by the increasing demand for AI systems capable of understanding and generating natural language descriptions of visual content. This multi-modal field encompasses various downstream tasks, including visual grounding [1] [2], dense captioning [3], [4] [5], and visual question answering (VQA) [6] [7], [8]. In particular, VQA poses a significant challenge as it requires machines to automatically answer questions about given visual contents. This demands a joint understanding of both visual content and textual queries, enabling machines to predict accurate answers through complex visual reasoning processes.

Traditionally, much of the research in VQA has been centered around processing 2D images, where algorithms are developed for analyzing static images to generate responses to questions. In such 2D-VQA, models typically rely on visual features extracted from images using convolutional neural networks (CNNs) [9] and textual features derived from questions using recurrent neural networks (RNNs) [10]. Furthermore, advancements in attention mechanisms [10] and transformer-based approaches [11] enables more effective fusion of image and text features and enhances the overall quality of question answering.

The shift from 2D-VQA to 3D-VQA introduces extraordinary complexities due to the spatial nature of 3D scenes, which require specialized techniques for feature extraction, often involving point cloud processing methods such as PointNet++ [12]. In contrast, 2D-VQA typically involves answering questions based on entire 2D images without explicitly performing object-level classification or localization. The models rely on attention mechanisms to highlight relevant image regions, guiding the fusion of image and text features to directly predict answers. Recent advances in 3D spatial reasoning [13] mark a paradigm shift toward cognitive-driven intelligence, where systems dynamically integrate exploration, perception, and reasoning in unstructured environments. 3D-VQA fundamentally requires the model to reason over rich spatial structures and complex object interactions in 3D space, often through multi-task learning. This involves not only generating answers but also accurately identifying and localizing specific objects within the 3D environment, which are essential for comprehensively understanding the scene and producing correct answers. While some approaches [14], [15] [16] have shown promising results by employing pre-trained models or large language models(LLMs), these methods rely heavily on extensive external datasets, requiring substantial computational and human resources. In the context of 3D-VQA, which is actually a multi-class vision-language task, achieving both high answer accuracy and high-quality answer generation is critical. The emerging emphasis on process-oriented evaluation [17] further highlights the need to bridge geometric reasoning with causal logic validation. Therefore, it is essential to

Received 10 November 2024; revised 17 February 2025, 24 April 2025, and 18 June 2025; accepted 10 July 2025. Date of publication 18 July 2025; date of current version 8 December 2025. This work was supported in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2024A1515012791, in part by Guangzhou Municipal Science and Technology Program Key Projects under Grant 2023B01J1001, in part by the Natural Science Foundation of China under Grant 62332019, Grant 62461160309, and Grant 62302297; and in part by the Young Elite Scientists Sponsorship Program by China Association for Science and Technology (CAST) under Grant 2022QNRC001. This article was recommended by Associate Editor M. Teutsch. *(Corresponding authors: Aihua Mao; Ran Yi.)*

Aihua Mao, Shuyi Wen, and Feng Chen are with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China (e-mail: ahmao@scut.edu.cn; 202321044412@mail.scut.edu.cn; chenfeng771@163.com).

Ran Yi is with the School of Computer Science, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: ranyi@sjtu.edu.cn).

Yong-Jin Liu is with BNRist, Department of Computer Science and Technology, MOE-Key Laboratory of Pervasive Computing, Tsinghua University, Beijing 100190, China (e-mail: liuyongjin@tsinghua.edu.cn).

Digital Object Identifier 10.1109/TCSVT.2025.3590456

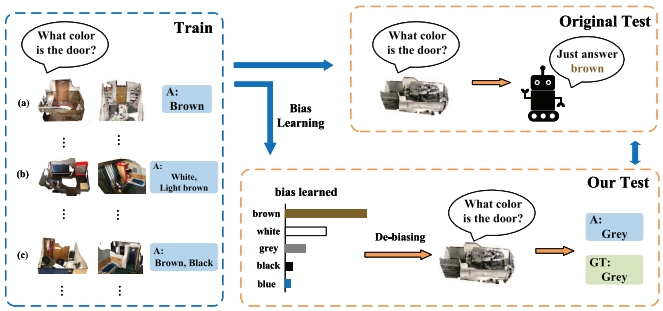

Fig. 1. An example of erroneous predictions caused by bias in 3D indoor scene datasets and the improvement through bias learning. (a), (b), and (c) respectively illustrate the impact of the three types of bias ((a) language modality bias, (b) classification and localization bias from multi-task joint training, and (c) manually categorized answer distribution bias) on the final prediction results.

develop strategies that maintain strong performance across both objectives.

Ensuring the model's robustness is crucial for achieving precise answers and effective object retrieval. Prior studies in the 2D-VQA domain [18], [19] [20] have shown that models often exploit superficial language patterns to predict answers, rather than performing genuine visual reasoning. In 3D-VQA tasks that require different modeling techniques, unique biases that do not exist or are less pronounced in 2D settings can be introduced. Therefore, bias learning in the context of 3D-VQA is not only necessary but also opens new research directions that cannot be sufficiently addressed by simply adapting existing 2D debiasing frameworks. As illustrated in Figure 1, it explicitly presents the individual impact of different types of bias through subfigures (a), (b), and (c). These three types of bias can influence the question-answering outcome in combination. The model tends to believe that most of the doors are brown because of existed bias, leading to incorrect predictions for the test sample where doors are actually gray. We note that such issue can be resolved through bias learning. Unlike 2D-VQA methods that tackles biases by training debiased models to capture solely language biases, 3D-VQA introduces new challenges that make bias mitigation more complex. Firstly, the most commonly used ScanQA dataset [8] does not explicitly assign categories to each question, making it hard to identify biases associated with specific question types. Secondly, it is insufficient to fully reflect the biases learned by the model if we only focus on a single language modality in models with multiple auxiliary task losses to incorporate additional information such as object detection and localization.

By addressing the above challenges, in this paper, we introduce a novel approach Bias3D-VQA to address the difficulties of 3D-VQA. It is based on generative adversarial networks (GANs) [21] and operates on the ScanQA dataset [8] undergoing preprocessing with pre-classification and SQA3D dataset [22]. The goal of our framework is to achieve high classification accuracy and good answer generation quality simultaneously. Specifically, inspired by GenB [23], which introduces a GAN-based model to capture language modality bias in 2D-VQA, our proposed method significantly expands the scope and functionality of bias modeling in the context of 3D-VQA. Unlike GenB, which focuses solely on static linguistic biases, we not only consider language modality bias but also focus on object classification and localization biases, as well as statistically derived answer distribution biases. In our framework, the bias model, the target model, and the generative adversarial component jointly form the architecture. In detail, the bias model is designed to learn biases, the target model performs the core 3D-VQA task, and the generative adversarial component ensures that the biases learned by the bias model closely resemble the real-world biases. In each training iteration, the bias model generates biased outputs through adversarial training, and the learned biases are then utilized in the subsequent alternating training phase of the target model to mitigate the impact of biases on performance.

The architecture of the bias model includes a generator that creates false point clouds from random noise to simulate the distribution of real input clouds, allowing the bias model to capture various biases. In order to distinguish between the real and generated answers, we utilize an adversarial discriminator to guide the bias model towards convergence with the target model. After achieving a balance between the two models, the bias model has effectively learned the inherent biases present in the task. We then use a debiased loss function to train the target model, optimizing through iterative backward propagation. In each training iteration, the bias model and the target model undergo alternating dual-stage training, leading to continuous iterative optimization. Upon completion of the training phase, inference is performed using only the target model for 3D-VQA. Notably, compared to existing single-branch model architectures, our proposed dual-branch structure, where the bias model interacts with the target model, achieves superior performance. By integrating bias mitigation techniques into our framework, we demonstrate the effectiveness of our approach by addressing the complexities and challenges in 3D-VQA. Furthermore, our method outperforms many pre-trained and even LLM-based methods in terms of answer accuracy, while significantly saving computational resources.

Our contributions can be summarized as follows:

- To the best of our knowledge, the proposed Bias3D-VQA in this study represents the pioneering effort in enhancing the performance of 3D-VQA through bias mitigation. We introduce a GAN-based bias model to capture various biases inherent in 3D-VQA. Moreover, our debiasing module is general and can be readily adapted to other 3D-VQA models.
- During the model training, we pay attention to the construction of debiased loss function that effectively mitigates the impact of language modality bias, multi-task joint bias, and manually-stated dataset distribution biases. To support this, we manually categorize the questions in the ScanQA dataset and analyze answer distributions to identify statistical bias patterns.
- We train our model on the ScanQA and SQA3D datasets and achieve state-of-the-art performance on almost all metrics. Moreover, compared to methods using large amounts of additional data to train pre-trained models

for VQA, our model even surpasses them while saving computational resources.

II. Related Work

A. Visual Question Answering (VQA)

VQA has received widespread attention in recent years, which involves answering questions about images, requiring machines to understand both visual content and textual queries. Many datasets and approaches have been proposed in this research field. The most common datasets include manually annotated VQA v1.0 [24] and VQA v2.0 [25], as well as synthetic question-answer datasets like GQA [26] and CLEVR [27], which obtain questions and answers from real-world or generated images, and can provide a comprehensive understanding of visual scenes. Early VQA implementations [28] focused on simplistic approaches, directly fusing image and text features for classification, but these methods had limited efficiency. Subsequently, approaches [9] inspired by human reasoning processes were proposed, but they struggled to adapt to real-world scenes containing complex information.

Inspired by the widely used attention mechanism, many attention-based VQA methods [29] [30] emerged as a promising solution to address the shortcomings of earlier approaches. These models aims to identify relevant visual elements and contextual cues within images, enhancing the model's ability to comprehend intricate visual scenes. Furthermore, the adoption of Transformer architectures [31], initially introduced by Google, revolutionized the VQA field. Models like MCAN [7] and MCAN-PA [32], utilized such structure to capture complex correlations between visual and textual features, were proposed. Our model builds upon the strengths of attention mechanisms and Transformer networks, leveraging their capabilities to integrate point cloud and text features in the 3D-VQA domain.

B. Bias Learning for Robust VQA

The robustness of VQA has been a subject of ongoing research, particularly in addressing biases and superficial semantic correlations that may affect model performance. While VQA v2.0 aimed to reduce some of these problems, models still had the potential to excel on the test set by learning prior answer distributions from the training set. To address this issue, the VQA-CP v2.0 dataset [18] was introduced, which is specially designed to enhance model robustness by reducing superficial semantic correlations and encourage models to rely on genuine understanding rather than memorization.

Additionally, researchers have improved model robustness by modifying VQA model structures. Ensemble-based methods [33], [34] utilize extra branches to capture biases between questions and answers. LMH [35] is a typical loss function for reducing language modality biases. It adjusts the model output using learnable neural network layers to align it with the language patterns generated by pre-trained language models. Recently, Bi et al. [19] introduced a network called FAN-VQA, which simultaneously estimates language and visual biases through dedicated branches. Grounding-based methods [36], [37] leverage or generate synthetic data to enhance model robustness, although some studies, such as the introduction of greedy gradient ensemble(GGE) [38] loss function, argue that the performance improvement of this method is from regularization rather than learning multi-modal information between images and language. GGE assumes that the objective is to maximize the match between questions and images and uses mathematical gradient descent to minimize the difference between model predictions and ground truth. Building upon the above two methods, a combined counterfactual-based method [39] was proposed, which trains models using ensemble methods on artificially generated datasets. More recently, Liu et al. [20] employed adaptive entropy minimization based on question types to identify unreliable samples and used an answer entropy change rate to flag biased samples. However, these studies on the robustness of VQA are primarily focused on 2D images. Our model adopts a bias-based ensemble approach, marking the first study on the model robustness in the field of 3D-VQA.

C. Generative Adversarial Networks (GANs)

GANs have gained significant attention in the field of machine learning since their introduction by Goodfellow et al. [21]. GANs consist of two neural networks, namely the generator and the discriminator, which are trained simultaneously in a minimax framework. Through adversarial training, GANs learn to generate high-quality data samples that are indistinguishable from real data. Early developments in GANs focused on enhancing training stability and improving the quality of generated images. For example, techniques like PGGAN [40] have contributed to it and have better convergence properties. Moreover, some studies [41], [42] demonstrate that attention mechanisms can be integrated into GAN architecture to improve image synthesis quality.

In the context of VQA, GANs have been used to bridge the gap between visual and textual modalities, thereby enhancing the diversity and accuracy of generated answers. Specifically, Conditional GANs [43] have been employed to generate image features conditioned on questions, enabling VQA models to reason about visual content effectively. Subsequently, the adversarial attention network like VQA-AA [44] was proposed, which produces more diverse visual attention maps to better generalize the maps to questions.

More recently, there has been a growing interest in leveraging the adversarial training strategy to address the biases in VQA systems. Typically, Cho et al. [23] introduced a bias model that can accumulate biases from language modality, thereby improving the robustness and generalization capabilities of VQA models. Despite remarkable success in 2D-VQA, GANs have yet to find significant applications in the domain of 3D-VQA due to the complexity and unique challenges posed by 3D data. In our study, through the training of the bias model, we utilize a generative adversarial component to enable input to simulate the distribution of real point clouds. Obviously, this component is crucial to the accumulation of various biases and will have a great influence on the final results.

D. 3D VQA

While significant progress has been made in the field of 2D-VQA, there remains considerable room for development in addressing 3D scenes. The introduction of the VQA task for 3D scenes by Azuma et al. [8] marked a crucial moment in this domain, accompanied by the release of the ScanQA dataset. However, ScanQA focuses on basic question-answer pairs with limited object annotations. Zhao et al. [45] improved this issue by designing a framework that uses enhanced self-attention to capture complex object relationships in 3D environments. Subsequently, the 3DQA-TR model [46] was proposed, which improved BERT [47] to jointly process information of textual and visual inputs. Utilizing pre-trained V-L models in VQA has emerged as an effective strategy, with significant improvements observed in the pre-training phase. For instance, inspired by the MotionCLIP [48], Parelli et al. [14] recognized the benefits of incorporating 2D information for better 3D scene understanding. Consequently, a 3D scene encoder was pre-trained to align scene embeddings with text and 2D image features extracted by the CLIP model. Jin et al. [15] employed a context-aware spatial semantic alignment strategy to find interactive relations between visual and language masks, facilitating cross-modal information exchange in 3D-VQA tasks. Additionally, the 3D-VisTa model [16] introduced a self-supervised pre-training approach, comprising masked language/object modeling losses and scene-text matching losses to align point clouds and text effectively.

Recently, BridgeQA [49] was introduced as a unified structure for 2D-3D question answering, aiming to capture richer object information by pre-training on 2D images. Importantly, 3D-VQA systems often incorporate multi-task learning modules, which enable simultaneous training for tasks such as answering questions, object classification, and object localization, further enhancing the model's understanding of the scene. However, existing 3D-VQA methods primarily focus on improving overall performance through network architecture adjustments or enhancements, and have not adequately addressed biases present in the dataset and during the training process. Learning these various biases can be useful to reduce their impact on answer generation. To improve model robustness, we introduce a bias model inspired by GANs, providing reverse supervision to the target model, enabling effective bias correction and enhancing the overall performance of 3D-VQA systems.

III. Method

In this section, we propose our method Bias3D-VQA using alternating dual-branch model training scheme (Figure 2), which builds the bias and target models as two interactive modules using the same backbone network to better capture various biases from multiple sources in the 3D-VQA task.

First, to facilitate the identification of answer distributions based on question types, in Sec. III-A, we manually count the question categories on the ScanQA dataset. Then in Sec. III-B, we introduce the overall framework and training strategies of our method. In Sec. III-C, the basic architecture of the backbone network is presented. In Sec. III-D, different biases including answer distribution bias are presented. In Sec. III-E, we describe the training of the bias model based on the generative adversarial process. In Sec. III-F, we present the training of the target model under the debiased loss.

TABLE I

Adaptability of ScanQA Dataset Problems to 65 Question Types in the VQA-CP v2.0 Dataset

Dataset type	Number of samples	Number of matched samples	Matched rate
Train	25563	22055	86.28%
Val	4675	4013	85.84%
Test	4976	4347	87.36%

A. Statistics of Dataset Question Categories

Unlike the manual categorization of questions in the VQA-CP v2.0 dataset, the ScanQA dataset lacks explicit question categorization, making it challenging to identify answer distributions based on question types. We apply the 65 question categories from the VQA-CP v2.0 dataset to the 3D ScanQA dataset and test the adaptability, yielding results shown in Table I. It is worth noting that the matching rates are consistently above 85% across the training, validation, and test sets. For unmatched questions, we combined the first two words of the question as the question type. The visualization of the distribution of various question types is summarized in Figure 3. Specifically, for better visual clarity, we selectively display a subset of question categories as axis labels. In the legend, we annotate the entire bar chart by grouping question categories based on their count as a clustering method.

B. Overall Architecture

In this study, we aim to improve answer accuracy in the 3D-VQA task by mitigating biases. As illustrated in Figure 2, the framework Bias3D-VQA consists of a bias model, a target model, and the adversarial training component.

1) Bias Model Design and Motivation: The bias model is introduced to explicitly learn and capture biases that affect 3D-VQA performance. Unlike typical training pipelines that rely on real input data, the bias model starts from random Gaussian noise z and uses a generator to produce a fake point cloud r_b. This design ensures that the bias model focuses on learning the biases not directly dependent on the actual scene structure, such as inherent biases in the dataset and biases in the model's backbone. The generated point cloud is passed through the shared backbone to produce biased outputs y_b. A discriminator is used to distinguish between biased outputs y_b and real outputs y from the target model. This adversarial setup guides the bias model to better simulate biases observed in the real task.

2) Target Model and Debiasing Mechanism: The target model is responsible for producing the final answer predictions. It receives real input point clouds r along with the question. During training, it is guided by the outputs of the bias model and optimized using a debiased loss function. The loss function is designed to incorporate the knowledge learned by the bias model, adjusting the target model's outputs toward less biased and more accurate predictions.

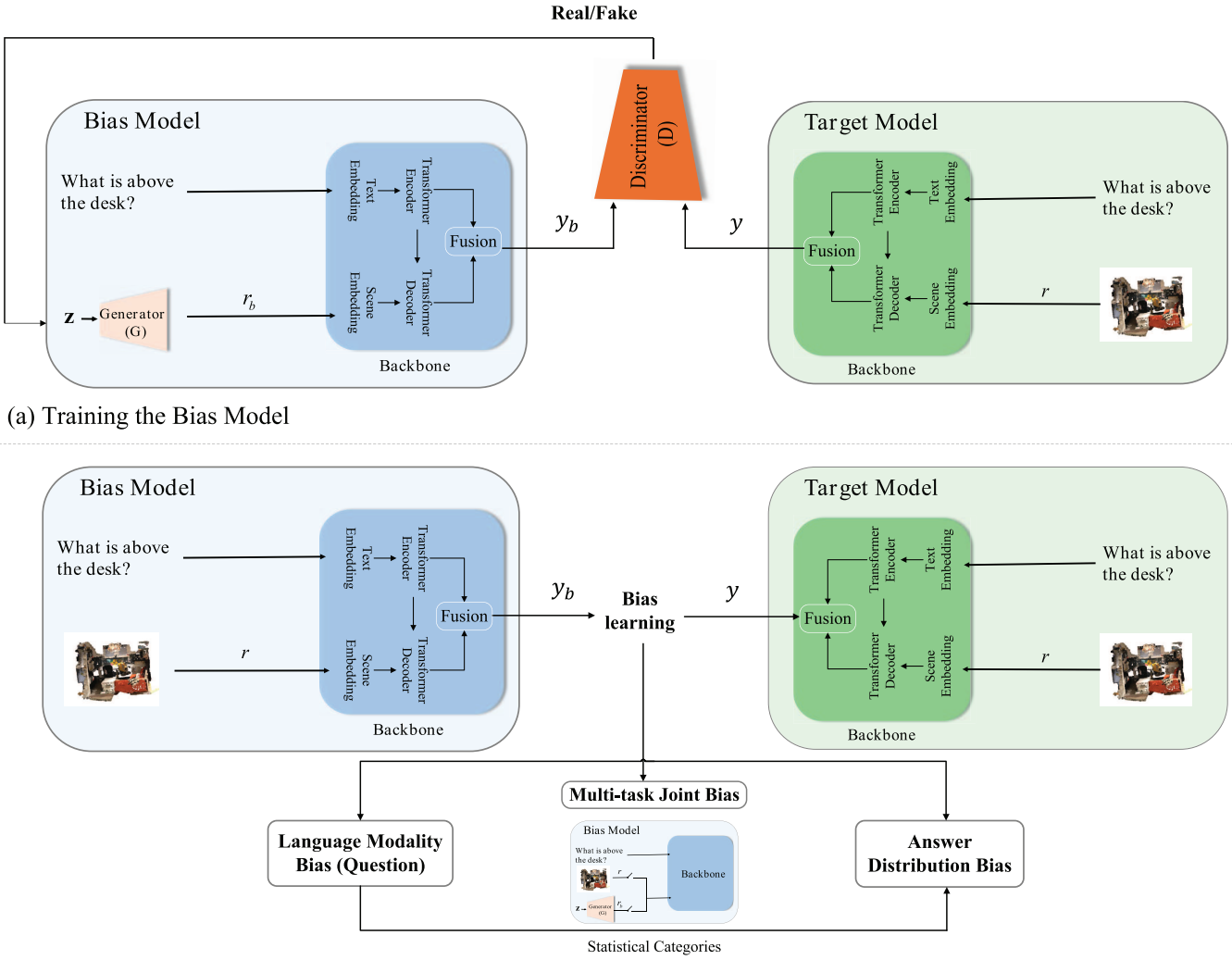

Fig. 2. Overview of our framework Bias3D-VQA. (a) and (b) demonstrate the alternating training of the two stages within each iteration. When training the bias model, adversarial component makes fake point cloud as close to the real input, and various biases are accumulated through this process. After the stage, a debiased loss function is used to train the target model to reduce the influence of biases. y_b and y correspond to the outputs of the bias model and the target model, respectively. Both of them include essential components such as answer prediction, object classification, and object localization.

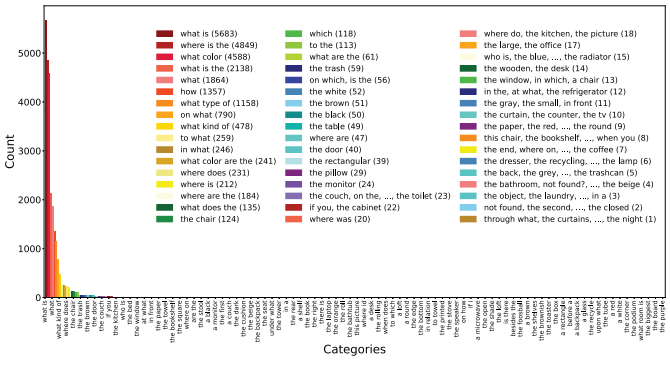

Fig. 3. Histogram of the proportion of question types in the ScanQA dataset.

3) Two-Stage Alternating Training Process: The two-stage alternating training process for one batch is illustrated in Algorithm 1. In the first stage, the bias model and the generative adversarial component are trained. The generator produces fake point clouds from noise, which are passed through the backbone to produce biased outputs. The discriminator is trained to differentiate between these and the real outputs, improving the bias model's ability to simulate realistic bias patterns. In the second stage, the target model is trained using real point clouds. It is optimized using a debiased loss informed by the bias model's predictions, aiming to reduce the impact of learned biases on answer generation. Each iteration corresponds to one mini-batch and consists of both stages, with separate optimization for the bias and target models.

4) Backbone Sharing: Both the bias model and target model share the same backbone network architecture, which is trained from scratch. This ensures a consistent feature representation space and enables the bias model to effectively influence the target model.

Algorithm 1 Alternating Training Framework in One Iteration

Ensure: X :=training data contains question q and point cloud r

for batch in training data do
  Feed X to target model(forward propagation)
  //Stage1:
  Use G to generate fake point cloud r_b
  Feed q and r_b to bias model
  Discriminate the authenticity of outputs from the target model and the bias model
  Calculate $\mathcal{L}_{GAN}$ based on Equation 3
  Calculate $\mathcal{L}_b$ based on Equation 4
  Backpropagate and optimize the parameters of the bias model, G and D
  //Stage2:
  Disable G
  Feed X to bias model without accumulating gradients
  Calculate $\mathcal{L}$ based on Equation 5
  Backpropagate and optimize the parameters of target model
end for

C. Model Backbone for Visual Question Answering

In our work, we leverage ScanQA [8], one of the state-of-the-art architectures, as the backbone network in our framework. This network first utilizes VoteNet [50] with PointNet++ [12] to extract point cloud features r, which describe the input point cloud's 3D coordinates along with auxiliary information, to generate proposals for object detection. Simultaneously, the representation q of the question is obtained by GloVE [51] for word representation, followed by LSTM [52] to capture contextual relationships, ultimately obtaining sentence features. Subsequently, through two layers of Transformer [31] for representing the relationship between object proposals and text, guided by attention mechanisms, we fuse the outputs through layer normalization. Finally, the network's performance is optimized through multi-module losses, with the backbone network's output y including answer results s^{ans}, object classification results s^{obj}, and object localization results s^{loc}. Since our research is based on indoor scene datasets, which involves selecting the highest-probability answer from the answer set as the final answer, making it a multi-class classification task. Each module focuses on predicting specific aspects relevant to the question: s^{loc} identifies the most relevant object bounding boxes, s^{obj} predicts associated object classes, and s^{ans} selects the correct answer among multiple candidates.

We note that as long as the above processing can be performed on point clouds and text, various networks used for 3D-VQA tasks can serve as the backbone. In this work, our research focuses on performance enhancement via bias learning, and the backbone network is only employed as a carrier for extracting bias. Therefore, considering the enhancement of multi-module losses, we opt for ScanQA [8] as our backbone network, as it has been demonstrated with robust performance in capturing the intricate relationships between textual queries and 3D scenes.

D. Sources and Analysis of Bias

1) Language Modality Bias: Many studies [18], [53] have indicated the reliance of VQA on language biases present in the dataset to predict answers accurately. Ensemble-based methods utilize multiple models to learn different aspects of language modality and generate ensemble predictions, thereby enhancing generalization to linguistic variations. The objective of ensemble-based methods is to create a more robust model capable of handling language modality biases. Our emphasis on reducing language modality bias is evident in our choice to use pure textual questions as input. In the process of capturing biases with the bias model, the generator creates random noise to simulate point cloud distributions. At this stage, the only meaningful input is the question text, which inherently possesses language modality biases. Experimental results from static bias analysis in Table IX provide further insights into the influence of language modality bias on 3D-VQA results.

2) Multi-Task Joint Bias: As outlined in Sec. III-C, the 3D-VQA task encompasses three auxiliary tasks: answer generation, target object classification, and localization. These tasks are integral components of the computation and optimization of losses within our framework. Considering that the bias model generates multi-task results during the training process, it also involves multiple loss functions as optimization objectives. In the training pipeline, during the training of the bias model, the architecture is designed to mirror that of the target model via a generative adversarial framework. By optimizing the multi-task loss within the adversarial training setup, the bias model learns superficial correlations (biases) not only in answers, but crucially also in object classification and localization predictions. When training the target model, the bias model's outputs (containing biased predictions for object classification and localization) are leveraged. Through the debiased loss, the target model is explicitly encouraged to learn representations that are adversarial to these biased classification and localization predictions, directly counteracting the identified classification and localization biases within the target model's outputs for these sub-tasks. Consequently, the bias model not only captures static biases inherent in language modality, but also obtains dynamic biases arising from multi-task joint learning. This holistic approach ensures that our model is robust across various aspects of the 3D-VQA task.

3) Manual Statistics of Answer Distribution Bias: In addition to addressing biases inherent in language modality and multi-task learning, we also consider manual statistics of answer distribution bias. Since the fact that the original ScanQA dataset did not categorize question types, we manually analyzed and categorized the questions into distinct types in Sec. III-A. Based on the manual categorization results, we conducted statistical analysis on the number of answers per question category to gain insights into additional answer distribution bias. When calculating losses, we incorporate the distribution of answers from the training set, enhancing our model's ability to capture and mitigate bias effectively. Although bias models capture language biases inherent in the original dataset, our additional consideration of answer distribution bias based on question-answer correspondences significantly enhances overall bias collection and leads to a

more comprehensive understanding and management of biases in the overall task.

E. Adversarial Training of the Bias Model

Our framework includes a bias model $F_b(\cdot,\cdot)$ (which produces fake output y_b) and a target model $F(\cdot,\cdot)$ (which generates real output y); see also the pseudo-code in Algorithm 1. Initially, since the model has not yet learned effective representations for point cloud and text inputs, both y_b and y are meaningless random outputs. Since each iteration involves optimizing the parameters of both the bias model and the target model, multiple iterations of training transform these outputs into meaningful answers for the 3D-VQA task. Existing work on 3D-VQA has not fully addressed biases present in the dataset itself or during the training process. To ensure that the bias model learns biases similar to those of the target model, we use ScanQA model as the basic structure for both $F_b(\cdot,\cdot)$ and $F(\cdot,\cdot)$.

The adversarial training process is listed from line 2 to line 8 in Algorithm 1. We introduce Gaussian random noise $z \sim N(0,1)$, the generator G generates fake point cloud features r_b of the same dimension as the original point cloud features r based on this noise, where the bias model's output y_b can be represented as:

$$y_b = F_b(q, r_b) = F_b(q, G(z)) \tag{1}$$

The target model then accepts text and real point cloud features r as input, with the model's real output y represented as:

$$y = F(q, r) \tag{2}$$

To capture biases in the target model, we propose an adversarial training method to train our bias model. In order that the point clouds generated by the generator G are more realistic, we employ a discriminator D to differentiate between the authenticity of y and y_b, denoted as $D(y)$ and $D(y_b)$ respectively. The objective of the adversarial network, based on the generator and discriminator, can be expressed as:

$$\begin{aligned}&\min_{F_{b,G}} \max_{D} \mathcal{L}_{GAN}(F_{b,G}, D)\\ &= E_y[log(D(y))] + E_{y_b}[log(1 - D(y_b))]\end{aligned} \tag{3}$$

Ideally, we expect the discriminator to accurately differentiate between real answer y and fake answer y_b, where $D(y)$ tends toward 1 and $D(y_b)$ tends toward 0. Thus, maximizing $\mathcal{L}_{GAN}$, i.e., maximizing the discriminator's classification accuracy, is desired. For the generator, maximizing $E_{y_b}[log(D(y_b))]$ is equivalent to minimizing $E_{y_b}[log(1-D(y_b))]$. Through iterative training, such min-max problem reaches a balance, where the quality of data generated by the generator is good enough to make it difficult for the discriminator to distinguish between real and fake. Since the generator does not receive real point clouds as input and shares the same backbone network as the target model, it can be considered that the bias model learns complex biases which occur even in the absence of real point cloud inputs, including not only biases inherent in language modality but also biases accumulated from other multi-task learning stages. As an iterative optimization objective, the multi-task joint loss can be used to address biases introduced by multi-task joint learning.

The total loss for the training of bias model can be expressed as:

$$\mathcal{L}_b = \mathcal{L}_b^{vote} + \mathcal{L}_b^{obj} + \mathcal{L}_b^{loc} + \mathcal{L}_b^{ans} + \mathcal{L}_{GAN} \tag{4}$$

where $\mathcal{L}_b^{vote}$ corresponds to the VoteNet object detection loss, while $\mathcal{L}_b^{obj}$, $\mathcal{L}_b^{loc}$, and $\mathcal{L}_b^{ans}$, basing on cross-entropy (CE) loss, respectively accord with the three tasks (object detection, object localization, answer generation) in the output section of the ScanQA backbone network, and $\mathcal{L}_{GAN}$ corresponds to the adversarial training loss as described above. Through multiple iterations of training, each iteration enables learning more accurate biases to support the alternating training of the target model.

F. Mitigating Bias in the Target Model

As shown from line 9 to line 12 in Algorithm 1, during the alternating training of the target model in one iteration, the parameters of the bias model are frozen and only participate in forward operations, without back propagation and gradient accumulation. The data generated by the bias model is used solely for the bias model's optimization and does not require backpropagation to the bias model again. When the bias model receives real complete point clouds r as input, the output y_b is considered to contain a large amount of bias, including language bias and accumulated biases of our modules. The loss formula for training the target model is represented as:

$$\mathcal{L} = \mathcal{L}^{vote} + \mathcal{L}^{obj} + \mathcal{L}^{loc} + \mathcal{L}^{ans} \tag{5}$$

where the calculation of the VoteNet detection loss $\mathcal{L}^{vote}$, object detection loss $\mathcal{L}^{obj}$, and object localization loss $\mathcal{L}^{loc}$ is the same as that in the bias model with ScanQA as the backbone, while the answer classification loss $\mathcal{L}^{ans}$ is calculated using the debiased loss function, represented as:

$$\begin{aligned}\mathcal{L}^{ans} &= Debiased - loss(y^{ans}, s^{ans}, s_b^{ans})\\ &= LMH(y^{ans}, s^{ans}, s_b^{ans})\\ &= BCE(softmax(log(y^{ans}) + s^{ans} log(s_b^{ans}), s^{ans}) + R\end{aligned} \tag{6}$$

where we pay attention to the loss, which is based on the existing bias, between correct answer y^{ans} and the predicted answer s^{ans}. Given that the bias model is currently able to capture various inherent biases, such bias can be represented as s_b^{ans}. The R term in the formula represents an entropy penalty introduced to prevent the model from ignoring the effects of the bias. R can be expressed as:

$$R = \omega H(softmax(s^{ans} \log(s_b^{ans})) \tag{7}$$

where the entropy is calculated as $H(z) = \sum_r z_r \log(z_r)$, and ω is a hyperparameter used to adjust the influence of the penalty term. The debiased loss function is designed to mitigate the biases captured by the bias model during training. The core principle behind this loss function is to penalize predictions that align too closely with the biases identified by the bias model. By incorporating this debiased loss into the training process of the target model, the model is encouraged to

focus on the underlying features of the data rather than the superficial correlations that may exist due to dataset or model biases, leading to more accurate and fairer predictions. In this paper, we choose LMH as the debiased loss function due to experimental findings that the utilization of LMH loss leads to the optimal performance of the target model, which is demonstrated in the ablation study in Sec. IV-E.3. Besides, LMH effectively reduces biases by dynamically adjusting the model's predictions through learned mixing coefficients, which are designed to counteract the biases identified, making the model more focused on genuine multi-modal information.

Based on our research experience in VQA robustness, static answer distribution biases often have a significant impact on classification results. For a specific question category i, this distribution can be represented as:

$$B_i = \left[\frac{a_{i,1}}{A_i}, \frac{a_{i,2}}{A_i}, \ldots, \frac{a_{i,j}}{A_i}, \ldots, \frac{a_{i,C}}{A_i}\right] \tag{8}$$

where $a_{i,j}$ represents the number of candidate answers for the j-th candidate answer in the i-th question category, C represents the total number of candidate answers across all question categories, and A_i represents the total occurrence times of all candidate answers in this category. Under the influence of static answer distribution biases, the answer classification loss can be optimized as:

$$\mathcal{L}^{ans} = Debiased - loss(y^{ans}, s^{ans}, s_b^{ans} + B_i) \tag{9}$$

For category label i, we select B_i from the i-th question category and combine it with the answer distribution s_b^{ans} across all answer categories. This results in a bias representation that incorporates manually calculated dataset distribution biases. In this situation, the bias model has already captured various biases including those in the language modality, making the additional statistical answer distribution biases appear redundant. However, this redundancy help prevent network degradation or overfitting.

IV. Experiment

A. Experiment Setup

1) Datasets and Metrics: Our study focuses on the 3D-VQA task and primarily leverages the ScanQA dataset [8], with additional comparisons conducted on the SQA3D dataset [22]. The two datasets, designed to enable models to answer text-based questions about 3D indoor scenes, are all derived from the ScanNet dataset [54] and automatically generated many question-answer pairs. After multiple rounds of filtering, editing, and manual annotation of answers and objects, the data quality is enhanced. The ScanQA dataset includes 41,363 questions and 58,191 answers, covering 800 indoor scenes. Each scene originates from colorful RGB-D indoor scans, containing various types of questions related to object attributes, spatial relationships, and numerical statistics, ensuring dataset diversity and richness. Differently, the SQA3D dataset, intended for situated question answering, provides around 20,400 descriptions of 6,800 various situations which are based on 650 scenes in ScanNet. The dataset can be applied to embodied AI due to it's questioner location annotations and descriptive situation texts.

Given the deeper understanding of 3D space required for the 3D-VQA task, evaluation metrics include EM@1 and EM@10 for assessing classification accuracy, along with BLEU [55], ROUGE-L [56], METEOR [57], and CIDEr [58] for evaluating text generation quality. Notably, since the ScanQA dataset tends to generate long-text and free-form answers, it offers a wider array of metrics for assessing text quality compared to the SQA3D dataset.

2) Implementation Details: Our framework consists of alternating bias model training and target model training in one iteration, with the backbone network settings consistent with ScanQA scheme. Input point cloud features are represented as $(b, c, 10)$, where b denotes batch size, c denotes the number of features per sample, and 10 channels encompass point coordinates, heights, color, and normals. Multi-view features are excluded due to a significant increase in training parameters and time, with only marginal performance improvement observed.

The generator G comprises three layers, each consisting of a combination of linear layers, batch normalization layers, and LeakyReLU activation layers (with a negative slope set to 2). The noise z is represented as $(b, c, 32)$, which is then mapped to the same dimension as the point cloud through linear layers after three layers of expansion to obtain fake point clouds. The discriminator D maps the data to 1024-dimensional features, followed by dimension reduction in hidden layers and compression in the output layer, ultimately using Sigmoid to map a 1-dimensional feature to between 0 and 1 as the basis for judging authenticity. During the construction of the LMH debiased loss function, the hyperparameter ω in the entropy penalty term R is set to 0.36.

The discriminator's weights are initialized using the Kaiming method. The bias model employs the Adamax optimizer, while the target model uses the Adam optimizer. Each of the initial learning rate is separately set to 5e-3, which decays to 1e-3 after 15 epochs, with a total of 30 epochs of training.

All experiments were conducted on a single RTX 3090 with 24GB of VRAM.

B. Quantitative Results

Given the 3D-VQA task, primary approaches can be categorized into two groups: (1) models trained solely on the ScanQA or SQA3D dataset, including ScanQA [8], 3D-VisTA (scratch) [16], Multi-CLIP(scratch) [59], our method, and works that have demonstrated good performance in 2D-VQA and have been transferred to 3D, such as RandomImage+MCAN [7], VoteNet+MCAN [7] and ScanRefer+MCAN(e2e) [7]; (2) models pre-trained on other large 3D datasets and fine-tuned on the ScanQA or SQA3D dataset, including jin-Context-aware [15], CLIP [14], Multi-CLIP (pre-trained) [59], 3D-VisTA [16] and BridgeQA [49].

Comparison results of our method with other methods on the ScanQA test set with/without objects are presented in Tables II and III. The best and second-best methods are indicated by bold and underline, respectively. During the pre-training phase, the pre-trained methods (marked with *) utilize either 2D images rendered from 3D data or textual descriptions of the entire scene to gain a comprehensive understanding of

TABLE II

COMPARISON WITH STATE-OF-THE-ART METHODS ON THE SCANQA TEST SET WITH OBJECTS. BEST PERFORMANCE IS MARKED BOLD AND SECOND-BEST IS UNDERLINED. (*) DENOTES PRE-TRAINED METHODS. D IN INPUT DENOTES TEXT DESCRIPTION OF THE SCENE

Method	Input	EM@1	EM@10	BLEU-1	BLEU-4	ROUGE	METEOR	CIDEr
Methods only trained on ScanQA:								
RandomImage+MCAN(2022) [7]	3D	22.31	53.11	26.66	**14.26**	31.27	12.13	60.37
VoteNet+MCAN(2022) [7]	3D	19.71	50.76	29.46	6.08	30.97	12.07	58.23
ScanRefer+MCAN(e2e)(2022) [7]	3D	20.56	52.35	27.85	7.46	30.68	11.97	57.36
ScanQA(2022) [8]	3D	23.45	**56.51**	31.56	12.04	34.34	13.55	67.29
3D-VisTA(scratch)(2023) [16]	3D	**25.20**	55.20	-	10.50	35.50	13.80	68.60
Multi-CLIP(scratch)(2023) [59]	3D	22.76	-	31.08	13.31	33.84	13.28	65.81
Bias3D-VQA(Ours)	3D	24.28	56.25	**33.85**	13.75	**35.98**	**14.41**	**70.71**
Pre-trained methods with additional data:								
jin-Context-aware*(2023) [15]	3D+d	24.58	55.97	33.15	11.23	35.97	14.16	70.18
CLIP*(2023) [14]	3D+2D+d	23.92	-	32.72	14.64	35.15	13.94	69.53
Multi-CLIP(pre-trained)*(2023) [59]	3D+2D+d	24.02	-	32.63	12.65	35.46	13.97	68.70
3D-VisTA*(2023) [16]	3D+d	27.00	57.90	-	16.00	38.60	15.20	76.60
BridgeQA*(2024) [49]	3D+2D	31.29	-	34.49	24.06	43.26	16.51	83.75

TABLE III

COMPARISON WITH STATE-OF-THE-ART METHODS ON THE SCANQA TEST SET WITHOUT OBJECTS

Method	Input	EM@1	EM@10	BLEU-1	BLEU-4	ROUGE	METEOR	CIDEr	Inference time(it/s)↑
Methods only trained on ScanQA:									
RandomImage+MCAN(2022) [7]	3D	20.82	51.23	26.29	9.66	29.23	11.54	55.64	-
VoteNet+MCAN(2022) [7]	3D	18.15	48.56	29.63	7.10	29.12	11.68	53.34	-
ScanRefer+MCAN(e2e)(2022) [7]	3D	19.04	49.70	26.98	7.82	28.61	11.38	53.41	-
ScanQA(2022) [8]	3D	20.90	54.11	30.68	10.75	31.09	12.59	60.24	4.38
3D-VisTA(scratch)(2023) [16]	3D	20.40	51.50	-	8.70	29.60	11.60	55.70	-
Multi-CLIP(scratch)(2023) [59]	3D	20.71	-	31.22	11.49	31.35	12.80	60.75	-
Bias3D-VQA(Ours)	3D	**21.63**	**55.13**	**33.58**	**12.92**	**33.09**	**13.70**	**64.83**	4.59
Pre-trained methods with additional data:									
jin-Context-aware*(2023) [15]	3D+d	21.56	53.89	31.48	15.84	31.79	13.13	63.40	3.69
CLIP*(2023) [14]	3D+2D+d	21.37	-	32.70	11.73	32.41	13.28	62.83	4.58
Multi-CLIP(pre-trained)*(2023) [59]	3D+2D+d	21.48	-	32.69	12.87	32.61	13.36	63.20	-
3D-VisTA*(2023) [16]	3D+d	23.00	53.50	-	11.90	32.80	12.90	62.60	3.77
BridgeQA*(2024) [49]	3D+2D	30.82	-	34.41	17.74	41.18	15.60	79.34	0.12

the scenes. Subsequently, fine-tuning is conducted for various downstream tasks, especially for the 3D-VQA implemented on the ScanQA dataset. While a few pre-trained methods exhibit better results in terms of classification accuracy and text generation, the cost includes the complexity of computation and the consumption of resources, which are absent in methods without pre-training. Considering pre-trained methods' reliance on additional data, for fairness, direct comparison is not made. However, they are listed for reference, demonstrating that our method is comparable or even surpasses some pre-trained models in some aspects.

As shown in Table II, our method achieves significant performance across multiple evaluation metrics compared to other methods on the ScanQA test set with objects. While our method ranks second in EM@1 and EM@10, trailing by only 0.26% in EM@10, 3D-VisTA performs slightly better in EM@1 but drops behind by 1.05% in EM@10. Moreover, our method outperforms others significantly in text generation, showing substantial advantages in BLEU-1, ROUGE, METEOR, and CIDEr, with leads of 2.29%, 1.64%, 0.86%, and 3.43%, respectively. These leading indicators highlight the consistency in vocabulary and overlap with reference answers, as well as the ability to capture and utilize key details in 3D scenes for generating informative and creative answers.

Notably, our method achieves the fastest inference time, measured by the number of iterations per second, outperforming the baseline ScanQA model and all pre-trained models in this metric. Additionally, our method achieves comparable or even superior EM@1 and EM@10 to some pre-trained models, while also demonstrating different degrees of superiority in text generation quality compared to pre-trained models, except for 3D-VisTA and BridgeQA. Although these two methods show improvement across various metrics, they rely heavily on extensive pre-training with 2D data or text description to compensate for the lack of 3D scene information, because of which, they have a large number of parameters during the training process, as shown in Table IV. Additionally, we also include comparisons of FPS and FLOPs to further analyze computational efficiency. The advantage across these three metrics demonstrates that our method achieves faster processing speed and requires fewer computational resources compared to pre-trained methods. On the other hand, since our method is an improvement in training strategy, it can

TABLE IV

COMPARISON OF PARAMETERS, FPS, AND FLOPS WITH PRE-TRAINED METHODS ON THE SCANQA DATASET AND THE EFFECT OF GRAFTING TO PRE-TRAINED METHODS FOR PARTIAL EPOCHS ON THE TEST SET WITH OBJECTS. (#) DENOTES WE TRAIN THE METHOD FOR A FEW EPOCHS

Method	Parameters(M)↓	FPS↑	FLOPs(G)↓	EM@1	EM@10
3D-VisTA(pre-trained)* [16]	133.81	60.32	24.7	27.00	57.90
BridgeQA(pre-trained)* [49]	697.34	1.92	4047.1	31.29	-
Bias3D-VQA(Ours)	**26.45**	**73.44**	**12.0**	24.28	56.25
BridgeQA# [49]	-	-	-	24.68	46.22
BridgeQA# + Ours	-	-	-	**24.85**	**47.58**

TABLE V
COMPARISON WITH STATE-OF-THE-ART METHODS ON THE SQA3D TEST SET

Method	Input	EM@1
Methods only train on SQA3D fellow:		
ScanQA(2022) [8]	3D	46.58
3D-VisTA (scratch)(2023) [16]	3D	46.70
Multi-CLIP(scratch)(2023) [59]	3D	47.38
Bias3D-VQA(Ours)	3D	**47.53**
Pre-trained methods with additional data fellow:		
3D-VisTA*(2023) [16]	3D+d	48.50
Multi-CLIP(pre-trained)*(2023) [59]	3D+2D+d	48.02
BridgeQA*(2024) [49]	3D+2D	52.91

potentially be integrated with pre-trained approaches. As shown in Table IV, we compare the answer generation accuracy of the original BridgeQA method with the version incorporating our strategy after training for a few epochs. The results indicate that our strategy effectively enhances the performance of other methods to a certain extent.

The comparison results of our method with other methods on the ScanQA test set without objects are presented in Table III. Our method outperforms others consistently across all metrics, maintaining the top position. Specifically, our method leads by 0.73% and 1.02% in EM@1 and EM@10, respectively, in terms of answer accuracy. Regarding text generation quality, our method outperforms the second-best by 2.36%, 1.43%, 1.74%, 0.90%, and 4.08% in BLEU-1, BLEU-4, ROUGE, METEOR, and CIDEr, respectively. Compared to pre-trained models using additional data, our method surpasses almost all pre-trained models on most metrics, with only a slight shortfall on EM@1 compared to 3D-VisTA. As aforementioned, due to the large amounts of parameters demonstrated in Table IV from using of pre-trained models with quantities of data, BridgeQA serves merely as a reference point in this context. To sum up, these comparative results demonstrate that our method not only outperforms models trained on the ScanQA dataset but also surpasses most pre-trained models while saving computational resources.

To demonstrate the effectiveness of our proposed method, we conducts tests on the SQA3D dataset, with the results presented in Table V. Similar to the findings on the ScanQA dataset mentioned above, although our method slightly lags behind models pre-trained on additional data, it outperforms other methods trained solely on the SQA3D dataset, achieving the top ranking in terms of EM@1, indicating the highest answer accuracy.

C. Qualitative Results

We compare our proposed method with the ScanQA method on the ScanQA test set with objects. Partial visualization results are shown in Figure 4, covering various types of questions, including color, location, and classification. Since the ScanQA dataset does not provide ground truth answers for the test set, the accuracy of the visualized results is determined manually. Each example is evaluated by five individuals, with a consensus required for correctness.

The comparison results of various scenes in Figure 4 (a), (b), (c), (d), (e) and (f) demonstrate the superior performance of our method compared to the ScanQA method. In most scenes, our approach accurately predicts the correct answer when the object bounding boxes perfectly capture the target objects. The superior results across different scenarios further demonstrate that our method is an effective strategy for handling various bias-related challenges. In contrast, the ScanQA method often makes mistakes in the answers and fails to locate the correct objects, even if the answer is somewhat accurate. Given that our method utilizes the ScanQA architecture as its backbone, the superior performance of our approach demonstrates that bias learning effectively addresses the impact of biases on 3D-VQA performance. We observe that even when the top-1 answer is not accurate (EM@1), there is a high probability of the top-10 being correct (EM@10) when object detection is accurate. Incorrect answers often result from failed localization or detecting the wrong object, such as in Figure 4 (g) and (h) where our method fails to locate the backpack or chair mentioned in the questions, leading to incorrect answers. This observation aligns with the findings in [8], indicating that the performance of 3D-VQA is directly influenced by object detection accuracy.

TABLE VI
THE COMPARISON OF 4-FOLD CROSS-VALIDATION ON EM@1 METRIC ON THE SCANQA TRAINING SET. THE TRAINING SET IS DIVIDED INTO 4 SUBSETS (P1, P2, P3, P4). M1, M2, M3 AND M4 MEANS CHOOSING P1, P2, P3 AND P4 RESPECTIVELY AS THE VALIDATION SET AND THE REMAINING SUBSETS AS THE TRAINING SETS

Method	M1	M2	M3	M4	Overall
ScanQA(2022) [8]	14.56	14.53	14.67	13.93	14.42
Bias3D-VQA(Ours)	15.48	15.01	14.93	15.06	15.12

D. Validation of the Effectiveness

1) Stability of Adversarial Training: To assess training stability, we plotted the convergence curves of the generator and discriminator losses during the adversarial training of the bias model. As shown in Figure 5, the x-axis represents the training process over 30 epochs, corresponding to more than 90,000 iterations, while the y-axis indicates the loss values. The declining trends of both generator and discriminator losses demonstrate that, after multiple iterations, the losses gradually converge and oscillate within a stable, narrow range. This convergence behavior provides clear evidence that the adversarial training process remains relatively stable throughout.

2) Cross Validation: To further validate the effectiveness of the proposed method, we conducted K-fold cross-validation. Since the ground truth answers for the test set are not available, we manually performed a random K-fold split on the training set (with K =4 in this experiment), using one subset as the validation set and the remaining subsets for training in each iteration. Given that our method aims to improve answer accuracy by learning biases inherent in the corresponding dataset and model, we performed cross-validation on the ScanQA dataset, comparing our method with the original ScanQA approach, as shown in Table VI. We select the EM@1 metric as the evaluation criterion, which measures answer accuracy. The results demonstrate that our method significantly outperforms the original model, further validating

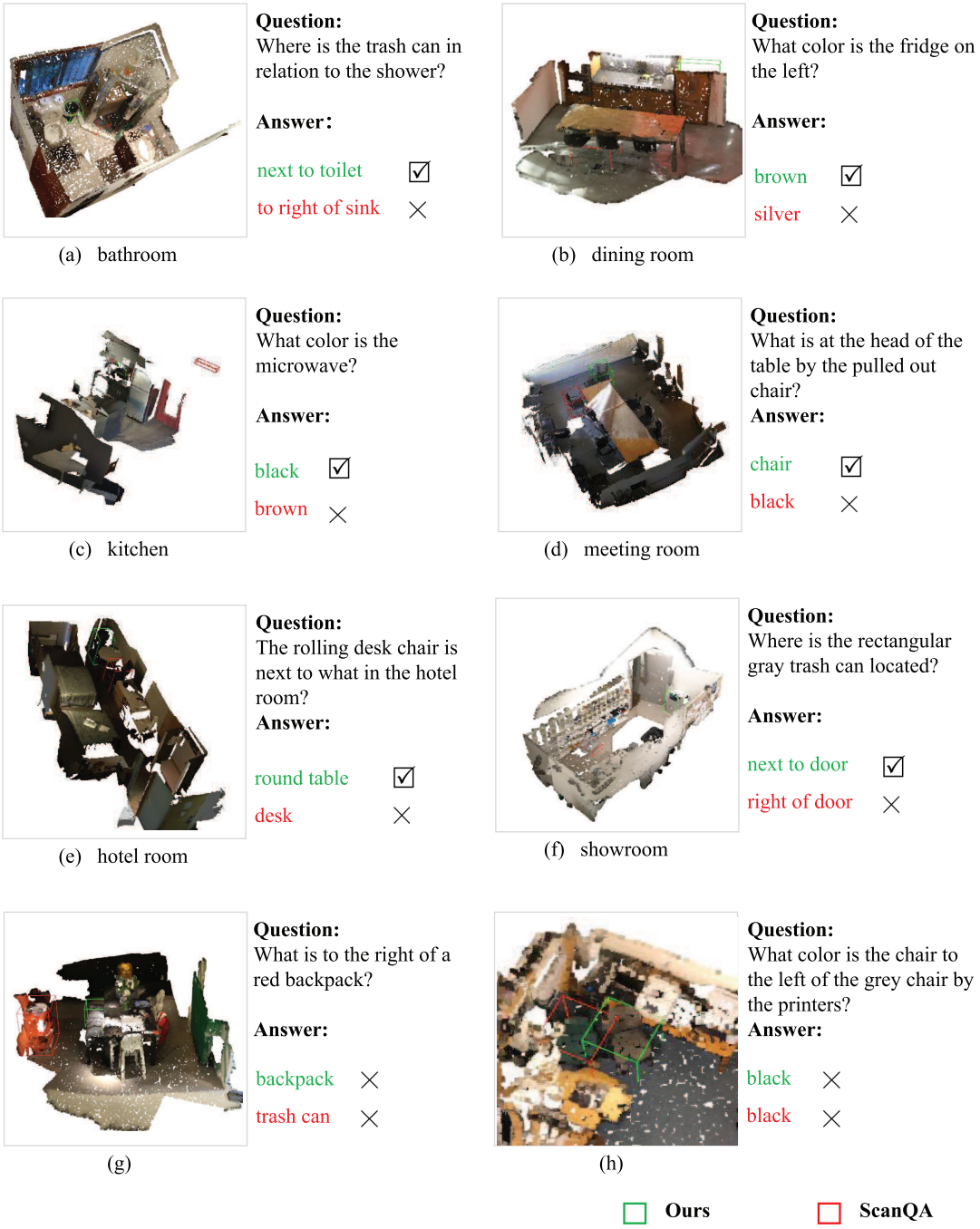

Fig. 4. Qualitative results on the ScanQA test set. Green boxes and text stand for the predictions from our method, and red for the ScanQA method. For our method, (a), (b), (c), (d), (e), (f) depict correct examples in various scenes, while the results in the subfigures (g) and (h) demonstrate errors due to inaccuracy in object detection.

the effectiveness of bias learning. By explicitly accounting for biases, our approach mitigates their impact during model training, ultimately enhancing the overall performance of 3D-VQA.

3) Results on Another Diverse Dataset: We have previously conducted experiments on two widely used 3D-VQA datasets ScanQA and SQA3D. To further assess the generalizability of our approach, we conduct additional experiments on another dataset "3DQA" [60], as shown in Table VII. This dataset consists of fully human-annotated question-answer pairs, covering complex question types such as functional reasoning and navigation, with answers expressed in free-form natural language. In terms of answer accuracy metrics EM@1 and

TABLE VII
RESULTS ON ANOTHER DIVERSE DATASET 3DQA

Method	EM@1	EM@10
3D-VisTA(scratch) [16]	49.30	88.60
Bias3D-VQA(Ours)	**50.12**	**89.26**

TABLE VIII
RESULTS OF SETTINGS OF DIFFERENT COMPONENTS ON THE SCANQA TEST SET WITH OBJECTS

Setting	EM@1	EM@10	BLEU-1	BLEU-4	ROUGE	METEOR	CIDEr
q-net+att+mul-task	21.93	55.51	30.64	12.42	32.93	13.03	63.37
p-net+att+mul-task	24.11	**56.47**	32.27	13.53	35.11	13.98	65.48
q-net+p-net+mul-task	22.83	56.05	31.39	11.54	34.19	13.62	65.80
Ours(all)	**24.28**	56.25	**33.85**	**13.75**	**35.98**	**14.41**	**70.72**

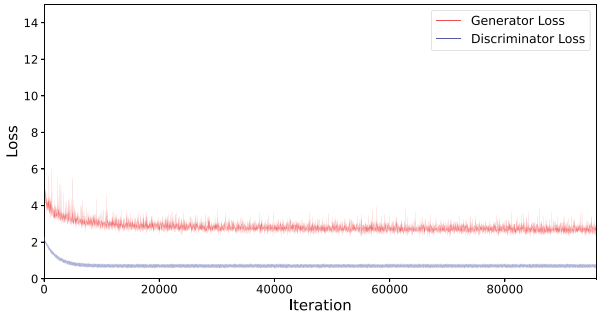

Fig. 5. Convergence Curves of Generator and Discriminator Losses during Adversarial Training.

EM@10, Bias3D-VQA demonstrates a notable advantage. On the diverse dataset, we compare our method against 3D-VisTA [16] trained from scratch. The superior performance of Bias3D further confirms the robustness and effectiveness of our bias-learning strategy.

4) Visualization of Answer Correction: Since our method aims to improve the accuracy of answer generation in 3D-VQA tasks by learning biases present in both the dataset and the model, we visualize the generated answers for selected scenes, as shown in Figure 6. The top-5 answers demonstrate that, compared to the original ScanQA baseline, our bias learning strategy effectively reduces the impact of biases on answer generation. Given that the test set is not publicly available, we conduct our visualization analysis on validation set samples, selecting samples whose two ground truth answers are identical. We analyze the three types of biases introduced in Sec. III-D individually. For biases inherent in the dataset's language modality, as shown in subfigure (a), chairs are generally perceived as brown both in the dataset and in common human knowledge. Our method successfully corrects the ScanQA's incorrect bias-driven prediction of "brown". For biases arising from multi-task joint training, the answer distribution corrections in subfigures (b) and (c) demonstrate that our method effectively captures and mitigates such biases. Regarding manually computed dataset distribution biases, since this bias is related to dataset categorization, we visualize the three most frequent question categories from Figure 3, corresponding to subfigures (b), (c), and (a). The clear improvements in answer predictions further validate the effectiveness of our bias learning strategy.

E. Ablation Studies

1) Ablation for Different Components: Our proposed method for addressing the 3D-VQA task can be divided into four main components: q-net (BiLSTM), p-net (VoteNet), att (attention modules, reflected in Transformer Encoder and Decoder Layers), and mul-task (multi-task learning module, including answering, object classification, and object localization). It is crucial to note that our approach requires the bias model, apart from the generator, to be identical to the target model in all our aspects. As a result, the architecture of the bias model changes with the target model.

Table VIII summarizes the results of different combinations of these components on the ScanQA test set with objects. The performance of our method, which integrates all components, achieves the best results on all metrics except for EM@10, where it trails by only 0.22%, falling within the error range. As shown in the first row, the method lacking the VoteNet component exhibits the poorest performance, underlining the importance of VoteNet for feature extraction and localization, which has been confirmed to significantly impact VQA performance in [8]. The effectiveness of the BiLSTM component is demonstrated in the second row, showing its contextual sensitivity and ability to capture long-distance dependencies, enabling better understanding of sentence structure and meaning. In the third row, methods lacking attention modules show disadvantages in all metrics, as the original point cloud contains much irrelevant information, and the Transformer's Encoder-Decoder structure extracts question-guided point cloud features, providing superior features for subsequent multi-task learning stages. Notably, the multi-task component is an integral part of our approach and remains consistently present throughout our ablation experiments. This component is indispensable for 3D-VQA tasks and closely related to the bias extraction in our method.

2) Ablation for Capture of Bias: Our method captures biases that may exist in the target model, including language modality biases, biases dynamically generated within various parts during training, and manually supplemented statistical distribution biases. Table IX shows the results of ablation experiments for various biases. The baseline without any bias is located in the first row. The second row, focusing only on static language bias, demonstrates the existence of significant biases in the language modality of the ScanQA dataset. In the setting of the third row, it refers to learning and reducing the influence of bias during the dynamic training process due to multi-task joint learning. It represents that in addition to the language modality bias, the classification and localization biases are also mitigated. In the training of the bias model, the optimization process includes losses from object classification and localization, enabling the bias model to capture biases accumulated in these sub-tasks (classification and localization biases). Subsequently, during the training of the target model,

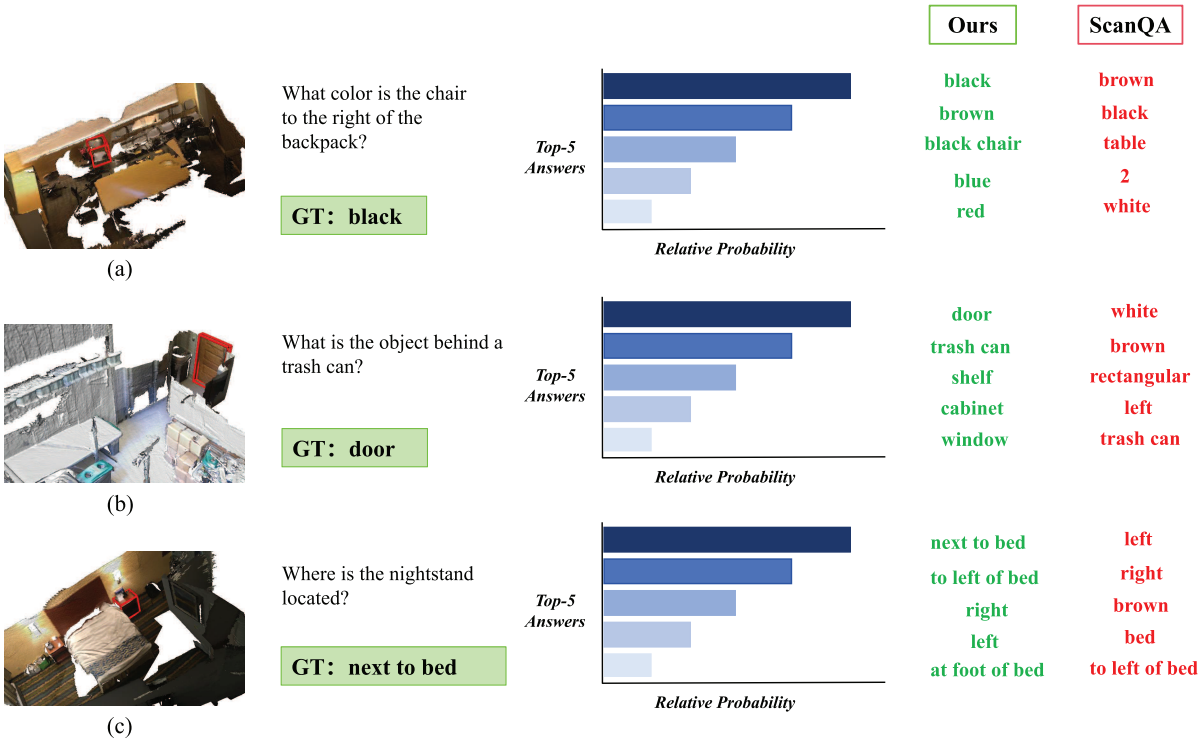

Fig. 6. Visualization of answer correction on the ScanQA validation set. Top-5 answers are listed for both our method and the original ScanQA model to compare their alignment with the ground truth. Sample pairs in (a), (b), and (c) are selected to illustrate the effectiveness of bias learning.

TABLE IX
RESULTS OF CAPTURE OF DIFFERENT BIASES ON THE SCANQA TEST SET WITH OBJECTS

Bias source	EM@1	EM@10	BLEU-1	BLEU-4	ROUGE	METEOR	CIDEr	ΔGap(EM@1)
None	21.93	55.51	30.64	12.42	32.93	13.03	63.37	-
Static(from language)	22.29	54.24	30.85	13.53	33.37	13.31	65.48	+0.36
Dynamic(from multi-task learning)	23.53	55.53	31.47	12.03	34.46	13.62	66.85	+1.24
Ours(all)	**24.28**	**56.25**	**33.85**	**13.75**	**35.98**	**14.41**	**70.72**	+0.75

the corresponding components in the debiased loss function are designed to reduce the influence of these biases on answer generation. Finally, as shown in the last row, our method achieves the best results on all metrics. Our approach of learning biases incorporates manually statistical language biases which are based on the question type (B_i in Sec. III-F), demonstrating that dynamically captured language biases and static language biases are not completely overlapping but complementary.

Each successive inclusion of bias can be seen to improve answer accuracy, as reflected in the EM@1 metric. The results highlight that learning dynamic biases from the multi-task joint training process provides the most significant improvement in answer accuracy for 3D-VQA tasks. In general, the results prove that as the model captures a wider variety of biases, our method becomes more effective in mitigating the impact of these biases through bias learning. We note that in the method without any bias, we use the BCE loss function to compute losses.

3) Ablation for Debiased Loss Function: In our method, the loss formulation in Eq. (9) does not impose any specific constraint on the debiased loss, theoretically allowing the use of any loss function capable of correcting biases. To explore this, as shown in Table X, we conducted ablation experiments within our framework for different debiased loss functions, including BCE, RUBi [34], GGE [38], and LMH [35]. It is clearly observed that LMH achieves the best results across all metrics. Although RUBi performs excellently in 2D-VQA, its performance in the 3D domain is very poor. BCE, which does not consider any bias, unexpectedly performs well in EM@10 and BLEU-4, especially in BLEU-4. This may be because the answer selection in 3D-VQA is essentially a classification problem rather than a text generation task, and the excellent classification ability of BCE leads to its good performance. The effectiveness of LMH underlines its

TABLE X
RESULTS OF DIFFERENT DEBIASED LOSS FUNCTION ON THE SCANQA TEST SET WITH OBJECTS

Debiased loss	EM@1	EM@10	BLEU-1	BLEU-4	ROUGE	METEOR	CIDEr
BCE	21.93	55.51	30.64	12.42	32.93	13.03	63.37
RUBi [34]	21.20	52.59	29.24	10.80	31.56	12.36	61.38
GGE [38]	23.31	55.02	32.44	11.96	34.46	13.72	67.45
LMH [35]	**24.28**	**56.25**	**33.85**	**13.75**	**35.98**	**14.41**	**70.72**

TABLE XI
COMPARISON WITH LLM-BASED APPROACHES ON THE SCANQA TEST SET WITH OBJECTS

Method	EM@1	BLEU-1	BLEU-4	ROUGE	METEOR	CIDEr
3D-LLM(2023) [61]	19.1	**38.3**	11.6	35.3	14.9	69.6
Agent3D-Zero(2024) [62]	21.3	31.4	5.1	**39.3**	**16.9**	**77.5**
Bias3D-VQA(Ours)	**24.28**	33.85	**13.75**	35.98	14.41	70.72

TABLE XII
COMPARISON WITH ZERO-SHOT GPT-4V AGENT ON THE SCANQA VALIDATION SET

Method	EM@1↑	EM@10↑	BLEU-1↑	BLEU-2↑	BLEU-3↑	BLEU-4↑	ROUGE↑	METEOR↑	CIDEr↑
GPT-4V [63]	18.01	18.01	24.49	10.68	4.64	1.63	**33.43**	**14.23**	58.32
Bias3D-VQA(Ours)	**20.11**	**49.82**	**30.59**	**19.39**	**13.06**	**7.58**	32.85	12.71	**62.24**

suitability for addressing biases in the 3D-VQA field, leading to improved overall performance on all metrics.

V. DISCUSSION

Recent advancements in large language models (LLMs) have significantly impacted multi-modal tasks, including 3D-VQA. Table XI shows the comparison of our method with two LLM-based approaches on the ScanQA test set with objects: 3D-LLM [61], and Agent3D-Zero [62]. Although our method lags behind on some metrics in text generation, it achieves the highest answer accuracy. This advantage can be attributed to the fact that LLMs were originally designed for text generation, whereas 3D-VQA is fundamentally a classification problem. Consequently, our method excels in the classification process within the candidate answer set. Besides, employing LLMs involves larger parameter sizes compared to the pre-trained models discussed earlier. In addition, we also compare our method and a typical LLM-based agent, GPT-4V Agent [63], on the ScanQA validation set, as shown in Table XII. Notably, GPT-4V operates in a zero-shot manner, where a scene-specific object list is extracted from the ground truth. As shown, our method outperforms the GPT-4V Agent on most metrics, demonstrating that our bias learning strategy, which mitigates bias effects during training, is more effective than directly relying on LLMs for answer prediction. In general, our method not only achieves superior answer accuracy but also conserves computational resources, making it a more efficient choice for 3D-VQA tasks.

While our method shows significant advancements in 3D-VQA, it still has some limitations. Performance may decline in complex scenes, and reliance on accurate object detection could affect overall results. If object detection fails, it may introduce irrelevant object feature information while missing the core object information required to answer the question. Such errors can result in misalignment between the visual modality and the linguistic modality of the question, ultimately affecting the accuracy of object-oriented question answering. We devised a human evaluation protocol to compare the object detection and localization accuracy of our method against the baseline ScanQA. For the 707–805 scene in the test set (99 scenes in total), we randomly selected one question per scene for evaluation, and asked human annotators to manually determine whether the predicted object bounding box was correct. For each method, we calculated the percentage of scenes where the human annotators deemed the bounding box prediction to be correct. This percentage served as our primary object detection and localization accuracy metric. As shown in Table XIII, compared with ScanQA, our bias learning strategy produced more accurate object detections by accounting for classification and localization biases.

TABLE XIII
COMPARISON OF OBJECT DETECTION AND LOCALIZATION ACCURACY ON 99 SCENES FROM THE SCANQA TEST SET

Method	Accuracy(707-805 scenes)(%)
ScanQA [8]	21.21
Bias3D-VQA(Ours)	78.79

While our framework performs well on indoor datasets like ScanQA and SQA3D, extending it to outdoor or dynamic environments presents challenges. Outdoor scenes often involve irregular spatial arrangements and less defined object boundaries, which can affect point cloud representation and object detection. Additionally, dynamic settings with moving objects require spatial and temporal reasoning, together with continuous tracking, which were not considered in this work. Also, in principle, our approach can be integrated into downstream tasks involving robust 3D vision-language alignment, with emerging paradigms like cognitive-driven spatial intelligence accelerating such integration, though fundamental challenges remain open for investigation. Future research could focus on incorporating these aspects to handle more diverse and unstructured environments.

VI. CONCLUSION

In this paper, we propose an innovate integrated debiasing method called Bias3D-VQA based on GANs for the 3D-VQA task. The entire framework consists of a target model, a bias model, and a generative adversarial component. Specifically, throughout the alternating training process, in each iteration, the bias model generates fake point cloud data and adjusts it through random noise in the generative adversarial component, effectively accumulating various biases, including those related to language modality. Subsequently, the debiased loss function is constructed leveraging the output of the bias model to optimize the target model, thereby improving its performance in both classification and text generation tasks.

Importantly, Bias3D-VQA proposed in this paper represents the first attempt to introduce the concept of a bias model in the field of 3D-VQA, addressing the crucial aspect of model robustness in this domain. Through comprehensive experimental comparisons, our approach demonstrates superior performance across almost all metrics when compared to methods trained solely on a single dataset. Moreover, our method realizes these performance improvement while also conserving computational resources, making it both effective and efficient for real-world applications.

REFERENCES

[1] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg, "ReferItGame: Referring to objects in photographs of natural scenes," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 787–798.

[2] P. Achlioptas, A. Abdelreheem, F. Xia, M. Elhoseiny, and L. Guibas, "Referit3D: Neural listeners for fine-grained 3D object identification in real-world scenes," in *Proc. 16th Eur. Conf. Comput. Vis.*, Glasgow, U.K. Cham, Switzerland: Springer, 2020, pp. 422–440.

[3] P. Anderson et al., "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6077–6086.

[4] D. Z. Chen, A. Gholami, M. Nießner, and A. X. Chang, "Scan2Cap: Context-aware dense captioning in RGB-D scans," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3192–3202.

[5] 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 Trans. Circuits Syst. Video Technol.*, vol. 34, no. 3, pp. 1322–1338, Mar. 2024.

[6] Y. Bi, H. Jiang, Y. Hu, Y. Sun, and B. Yin, "See and learn more: Dense caption-aware representation for visual question answering," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 2, pp. 1135–1146, Feb. 2024.

[7] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, "Deep modular co-attention networks for visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6281–6290.

[8] 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.

[9] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Deep compositional question answering with neural module networks," in *Proc. CVPR*, Nov. 2015, p. 3.

[10] M. Malinowski, M. Rohrbach, and M. Fritz, "Ask your neurons: A neural-based approach to answering questions about images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1–9.

[11] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2019, pp. 13 –23.

[12] C. R. Qi, Y. Li, H. Su, and L. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2017, pp. 5105–5114.

[13] Z. Zhu et al., "Move to understand a 3D scene: Bridging visual grounding and exploration for efficient and versatile embodied navigation," pp. 1–15, 2025, *arXiv:2507.04047*.

[14] M. Parelli et al., "CLIP-guided vision-language pre-training for question answering in 3D scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 5607–5612.

[15] Z. Jin, M. Hayat, Y. Yang, Y. Guo, and Y. Lei, "Context-aware alignment and mutual masking for 3D-language pre-training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 10984–10994.

[16] 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 *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 2899–2909.

[17] J. Lee, H. Park, B.-U. Lee, and K. Joo, "HUSH: Holistic panoramic 3D scene understanding using spherical harmonics," in *Proc. Comput. Vis. Pattern Recognit. Conf.*, Jun. 2025, pp. 16599–16608.

[18] A. Agrawal, D. Batra, D. Parikh, and A. Kembhavi, "Don't just assume; look and answer: Overcoming priors for visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4971–4980.

[19] Y. Bi, H. Jiang, Y. Hu, Y. Sun, and B. Yin, "Fair attention network for robust visual question answering," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 9, pp. 7870–7881, Sep. 2024.

[20] J. Liu, J. Xie, F. Zhou, and S. He, "Question type-aware debiasing for test-time visual question answering model adaptation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 11, pp. 10805–10816, Nov. 2024.

[21] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 2672–2680.

[22] X. Ma et al., "SQA3D: Situated question answering in 3D scenes," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, May 2023, pp. 1–24.

[23] J. W. Cho, D.-J. Kim, H. Ryu, and I. S. Kweon, "Generative bias for robust visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 11681–11690.

[24] S. Antol et al., "VQA: Visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2425–2433.

[25] 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. 6904–6913.

[26] D. A. Hudson and C. D. Manning, "GQA: A new dataset for real-world visual reasoning and compositional question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 6700–6709.

[27] J. Johnson, B. Hariharan, L. Van Der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick, "CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 2901–2910.

[28] H. Noh, P. H. Seo, and B. Han, "Image question answering using convolutional neural network with dynamic parameter prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 30–38.

[29] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 21–29.

[30] L. Li et al., "Multi-granularity relational attention network for audio-visual question answering," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 8, pp. 7080–7094, Aug. 2024.

[31] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Jun. 2017, pp. 5998–6008.

[32] A. Mao, Z. Yang, K. Lin, J. Xuan, and Y.-J. Liu, "Positional attention guided transformer-like architecture for visual question answering," *IEEE Trans. Multimedia*, vol. 25, pp. 6997–7009, 2022.

[33] S. Ramakrishnan, A. Agrawal, and S. Lee, "Overcoming language priors in visual question answering with adversarial regularization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, Jan. 2018, pp. 1541–1551.

[34] R. Cadene, C. Dancette, M. Cord, and D. Parikh, "RUBi: Reducing unimodal biases for visual question answering," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 841–852.

[35] C. Clark, M. Yatskar, and L. Zettlemoyer, "Don't take the easy way out: Ensemble based methods for avoiding known dataset biases," 2019, *arXiv:1909.03683*.

[36] C. Jing, Y. Wu, X. Zhang, Y. Jia, and Q. Wu, "Overcoming language priors in VQA via decomposed linguistic representations," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 34, no. 7, 2020, pp. 11181–11188.

[37] R. R. Selvaraju et al., "Taking a HINT: Leveraging explanations to make vision and language models more grounded," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2591–2600.

[38] X. Han, S. Wang, C. Su, Q. Huang, and Q. Tian, "Greedy gradient ensemble for robust visual question answering," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 1564–1573.

[39] L. Chen, X. Yan, J. Xiao, H. Zhang, S. Pu, and Y. Zhuang, "Counterfactual samples synthesizing for robust visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10800–10809.

[40] U. Demir and G. Unal, "Patch-based image inpainting with generative adversarial networks," 2018, *arXiv:1803.07422*.

[41] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7354–7363.

[42] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," 2018, *arXiv:1809.11096*.

[43] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.

[44] I. Ilievski and J. Feng, "Generative attention model with adversarial self-learning for visual question answering," in *Proc. Thematic Workshops ACM Multimedia*, Oct. 2017, pp. 415–423.

[45] L. Zhao et al., "Towards explainable 3D grounded visual question answering: A new benchmark and strong baseline," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 6, pp. 2935–2949, Jun. 2023.

[46] S. Ye, D. Chen, S. Han, and J. Liao, "3D question answering," 2021, *arXiv:2112.08359*.

[47] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[48] G. Tevet, B. Gordon, A. Hertz, A. H. Bermano, and D. Cohen-Or, "MotionCLIP: Exposing human motion generation to CLIP space," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Jan. 2022, pp. 358–374.

[49] W. Mo and Y. Liu, "Bridging the gap between 2D and 3D visual question answering: A fusion approach for 3D VQA," 2024, *arXiv:2402.15933*.

[50] 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. 9277–9286.
[51] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.
[52] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
[53] 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.
[54] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "ScanNet: Richly-annotated 3D reconstructions of indoor scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5828–5839.
[55] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2002, pp. 311–318.
[56] C.-Y. Lin and E. Hovy, "Manual and automatic evaluation of summaries," in *Proc. Workshop Autom. Summarization*, vol. 4, 2002, pp. 45–51.
[57] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop Intrinsic Extrinsic Eval. Measures Mach. Transl. Summarization*, 2005, pp. 65–72.
[58] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4566–4575.
[59] A. Delitzas et al., "Multi-CLIP: Contrastive vision-language pre-training for question answering tasks in 3D scenes," 2023, *arXiv:2306.02329*.
[60] S. Ye, D. Chen, S. Han, and J. Liao, "3D question answering," *IEEE Trans. Visualizat. Comput. Graph.*, vol. 30, no. 3, pp. 1772–1786, Mar. 2024.
[61] Y. Hong et al., "3D-LLM: Injecting the 3D world into large language models," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2023, pp. 20482–20494.
[62] S. Zhang et al., "Agent3D-zero: An agent for zero-shot 3D understanding," 2024, *arXiv:2403.11835*.
[63] S. Singh, G. Pavlakos, and D. Stamoulis, "Evaluating zero-shot GPT-4V performance on 3D visual question answering benchmarks," 2024, *arXiv:2405.18831*.

Aihua Mao received the B.Eng. degree from Hunan University in 2002, the M.Sc. degree from Sun Yat-sen University in 2005, and the Ph.D. degree from The Hong Kong Polytechnic University in 2009. He is a Professor with the School of Computer Science and Engineering, South China University of Technology (SCUT), China. His research interests include 3D vision and computer graphics.

Shuyi Wen received the B.S. degree in computer science and technology from Hunan Normal University in 2023. She is currently pursuing the M.S. degree in computer science with South China University of Technology, Guangzhou, China. Her current research interests include computer vision and 3D point cloud.

Feng Chen received the B.S. degree in computer science and technology from China Agricultural University in 2021. He is currently pursuing the M.S. degree in computer science with South China University of Technology, Guangzhou, China. His current research interests include machine learning and 3D point cloud.

Ran Yi received the B.Eng. and Ph.D. degrees from Tsinghua University, China, in 2016 and 2021, respectively. She is an Associate Professor with the School of Computer Science, Shanghai Jiao Tong University. Her research interests include computer vision, computer graphics, and computational geometry.

Yong-Jin Liu (Senior Member, IEEE) received the B.Eng. degree from Tianjin University, China, in 1998, and the Ph.D. degree from The Hong Kong University of Science and Technology, Hong Kong, China, in 2004. He is a Professor with the Department of Computer Science and Technology, Tsinghua University, China. His research interests include computer graphics and computer-aided design.