This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Situational Awareness Matters in 3D Vision Language Reasoning

Yunze Man Liang-Yan Gui Yu-Xiong Wang
University of Illinois Urbana-Champaign
{yunzem2,lgui,yxw}@illinois.edu

Abstract

Being able to carry out complicated vision language reasoning tasks in 3D space represents a significant milestone in developing household robots and human-centered embodied AI. In this work, we demonstrate that a critical and distinct challenge in 3D vision language reasoning is the situational awareness, which incorporates two key components: (1) The autonomous agent grounds its self-location based on a language prompt. (2) The agent answers open-ended questions from the perspective of its calculated position. To address this challenge, we introduce SIG3D, an end-to-end Situation-Grounded model for 3D vision language reasoning. We tokenize the 3D scene into sparse voxel representation, and propose a language-grounded situation estimator, followed by a situated question answering module. Experiments on the SQA3D and ScanQA datasets show that SIG3D outperforms state-of-the-art models in situational estimation and question answering by a large margin (e.g., an enhancement of over 30% on situation accuracy). Subsequent analysis corroborates our architectural design choices, explores the distinct functions of visual and textual tokens, and highlights the importance of situational awareness in the domain of 3D question-answering. Project page is available at https://yunzeman.github.io/situation3d.

1. Introduction

Humans learn knowledge efficiently through the interactions with the 3D world and the integration of multi-modal information, such as verbal guidance or instructions. Similarly, introducing language guidance into the visual comprehension task can greatly enhance models' learning efficiency [3, 42]. Nonetheless, despite considerable advancements in linguistic understanding [7, 11, 31, 58] and vision-language integration [3, 37, 54, 61], current methodologies remain deficient in accurately perceiving and rationalizing within real-world 3D environments, which is largely attributed to the lack of 3D situational reasoning capabilities.

Compared with machine learning models, humans put themselves inside the 3D world and then perceive and interact with the surrounding environment from their ego-perspective (Figure 1). Such situational awareness is a crucial difference between 2D and 3D visual understanding, and a key to achieve seamless understanding of spatial concepts in more complex real-world environments. Several existing methods recognize the lack of positional understanding in 3D and propose new benchmarks and joint optimization functions [44], or positional embedding methods [26] to enhance the overall reasoning performance.

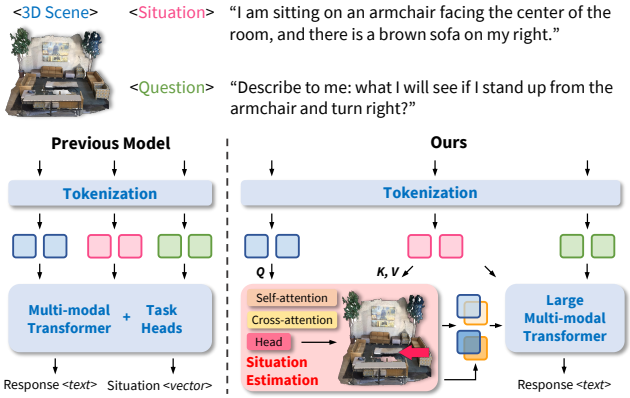

Figure 1. Previous methods perform direct 3D vision language reasoning without modeling the situation of an embodied agent in the 3D environment. Our method, SIG3D, grounds the situational description in the 3D space, and then re-encodes the visual tokens from the agent's intended perspective before vision-language fusion, resulting in a more comprehensive and generalized 3D vision language (3DVL) representation and reasoning pipeline. Q, K, V stands for query, key, and value, respectively.

However, the lack of an *explicit* situation modeling and situation-grounded 3D reasoning method restricts them from obtaining a generalizable and consistent 3D vision-language (VL) representation. As shown in Figure 2, the situation prediction of the state-of-the-art method [44] (in blue) diverges significantly from the ground truth vectors (in red) in almost all scenes in the dataset [13]. Moreover, our pilot study in Section 3 also reveals that situational understanding, despite being very crucial in comprehending the context of questions, only plays a minor role in the final

Figure 2. Situational estimation in existing methods [44] fails in most scenarios, indicating the missing registration between the situational description and 3D embeddings. Red: Ground truth (GT) vector. Blue: Estimated vector.

question-answering performance of existing methods.

In this work, we propose SIG3D, a novel approach designed to *precisely model and estimate an embodied agent's ego-location and orientation from a textual description*, before performing multi-modal QA tasks from the agent's egocentric perspective, as shown in Figure 1. Specifically, we leverage large-scale pretrained language and visual encoders to process the input text and 3D data, and fuse the tokens with attention modules to predict a situational vector. Previous attempts to directly predict the ego-situation are hindered by the expansive search space inherent in 3D environments. To address this challenge, we re-conceptualize the task as an *anchor-based classification*, where visual tokens are regarded as anchor points, and a likelihood of position together with a set of rotation parameters are concurrently regressed for each visual token. After obtaining the situational estimation, we propose a situational alignment and a situation-guided token re-encoding strategy, to perceive the environment from the agent's intended perspective. These strategies enhance the visual token with more accurate situational awareness for subsequent QA tasks.

Experiments on two challenging 3D visual question-answering (VQA) datasets [5, 44] demonstrate the significant improvement on situational estimation and QA tasks of our model. In particular, we improve the situational estimation accuracy by over 30%, and the subsequent QA performance by up to 3%. Further qualitative and quantitative analysis verifies our design choices and highlights the significance of situational awareness in 3D reasoning tasks.

To sum up, our paper has the following contributions: (1) We recognize the lack of situational awareness as a significant oversight in existing research. To address this, we introduce SIG3D, a situation-grounded 3D VL reasoning architecture, specifically designed to fill this void. (2) We propose an anchor-based approach to situational estimation, which effectively narrows the extensive search space in 3D environments for precise grounding of 3D positions and orientations with textual descriptions. Additionally, we investigate situational alignment and visual re-encoding mechanisms to leverage situational awareness for enhanced QA performance. (3) Our model demonstrates superior performance on two challenging datasets, SQA3D and ScanQA, surpassing the state of the art in both situational estimation and QA metrics. Ablation studies highlight the importance of situation-guided encoding, revealing its beneficial impact on general QA tasks.

2. Related Work

Vision Language Models (VLMs). Early transformer-driven [59] textual and visual encoders [17, 31] have facilitated great progress in recent vision language learning. Text-image contrastive models [30, 54] propose to align the feature space of two modalities with large-scale pretraining, fueling numerous downstream tasks from generalized open-vocabulary visual perception [21, 34, 36] to text-to-image generation [56]. Concurrently, some work uses text and vision encoders on separate modalities followed by feature fusion [18, 33] for multi-modal reasoning tasks. Since the emergence of Large Language Models (LLMs) [7, 58, 69], VLMs have experienced huge improvement with the help of LLMs as building blocks for multi-modal learning architectures. Specifically, recent work directly projects visual embeddings into language-space tokens as input to LLM [23, 40, 47, 70], or use the latent bottleneck structure for cross-modal visual decoding [3, 26, 37, 38], or treat LLM layers as encoder blocks for various visual tasks [49].

In the domain of visual question-answering (VQA) [4, 73], recent work has pushed the frontier towards video understanding [14, 28, 29, 40, 62], knowledge-based understanding [20, 22, 23, 41, 46, 57, 63], and commonsense reasoning [68]. Despite the outstanding performance in 2D image interpretation, most existing methods lack the capability to generalize to 3D scenarios. In contrast, our work studies the representation of visual information and its fusion with language embeddings in the 3D domain by targeting on the 3D situation-guided visual language interpretation.

Grounding Language in 3D Space. Compared with 2D images, knowledge such as spatial relationships, interactive exploration, and topological analysis, which only exists in the 3D world provides additional challenges and opportunities to develop better language models with stronger commonsense reasoning capability grounded in the real-world 3D scenarios. In this direction, early work seeks

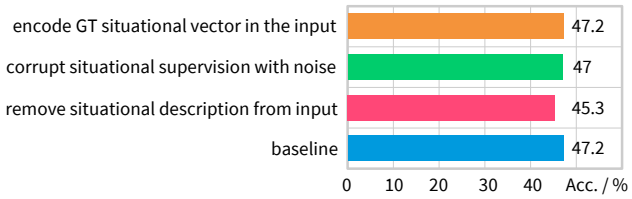

Figure 3. Results on variants of the representative SQA3D baseline method [44] demonstrate that situational understanding, despite being indispensable in perceiving the context of questions, makes negligible contribution in existing methods. This motivates a situation-guided 3D encoding mechanism in our model.

to ground isolated objects [1, 9] or objects within more complex scenes [2, 8, 19, 27] using natural language descriptions. Recently, with more collected 3D vision language datasets, several work starts to explore language-guided 3D visual interpretation and reasoning on a diverse set of datasets, including 3D scene captioning [10], open-vocabulary segmentation [16, 32, 51], and question-answering [5, 15, 25, 65, 74].

The success of LLMs also elicits the usage of them in 3D vision language reasoning for task decomposition [64], data generation, and multi-modal feature fusion [26]. Motivated by ScanQA [5], SQA3D [44] takes the first step into exploring the challenging 3D situational reasoning problem by developing a situated question-answering benchmark, and proposing the first joint learning baseline on the benchmark. Our work highlights the uniqueness and significance of situational awareness in the 3D vision language learning paradigm, leading to notably better 3D situational grounding and question-answering performance.

3. Pilot Study on Situated Reasoning

Despite pointing out the significance of situated understanding and reasoning, existing methods [44] fall short in providing effective situational estimation, as illustrated in Figure 2. This section delves into a pilot study examining the impact of situational understanding on downstream reasoning tasks. The SQA3D baseline [44] incorporates situational descriptions and uses ground truth (GT) situational vectors for supervision in a direct regression task. We investigate three variants of this baseline to assess the effect of situational understanding. In the first variant, we *remove the situational description and supervision* from the model, by passing in empty situational tokens. In another variant, we *corrupt the situation supervision* by introducing very large Gaussian noise to the GT vectors to effectively randomizing them. Finally, we try to *encode the GT situational vector in the input* with learnable multi-layer perceptron layers, which are then integrated with the visual and textual tokens.

Figure 3 demonstrates the results of this study, revealing negligible changes in performance across these variants. Notably, corrupting the GT situational information or directly incorporating it results in only marginal alterations in the QA outcomes. Omitting the situational description entirely from the input results in a mere 2% decline in accuracy. However, in the absence of this information, the model resorts to random guessing when determining the correct answer, as all responses depend on situational context. The findings from Figures 2 and 3 collectively indicate a deficiency in existing methods regarding situation vector estimation and the application of situational understanding in subsequent reasoning tasks. These unresolved challenges motivate the development of our proposed method.

4. Method

An overview of our approach SIG3D is illustrated in Figure 4. Our method begins with a set of points that represent a 3D scene, accompanied by a situational description and a question, which defines the overall context of the problem. We tokenize them into separate token embeddings (Section 4.1), and ground the textual description in the 3D scene with a vector comprising of location and orientation. We find direct single vector estimation to be challenging due to the vast and complex nature of the 3D search space, so we propose an anchor-based situational estimation strategy (Section 4.2). Subsequently, we re-encode the visual tokens from the perspective of situational vectors, enhancing the situational awareness for downstream reasoning tasks (Section 4.3). The finalized visual and textual tokens are fused by a transformer decoder to generate the final response.

4.1. Visual and Textual Tokenization

Leveraging input scene point clouds and textual prompts, our objective is to generate three distinct types of tokens: 3D visual tokens $z^{\mathrm{3D}} \in \mathbb{R}^{N_v \times C_v}$, situational tokens $z^{\mathrm{S}} \in \mathbb{R}^{N_s \times C_s}$, and question tokens $z^{\mathrm{Q}} \in \mathbb{R}^{N_q \times C_q}$. Each type of token is characterized by two primary components: N, representing the number of tokens, and C, encapsulating the feature embeddings. To tokenize and capture the feature embeddings for situational and question inputs, we employ a shared text tokenizer $\mathrm{E}^{\mathrm{TXT}}$ following prior methods [5, 44]. We assume that situational and question prompts are separated in the input data. If not, LLMs [7] can be used to parse the textual input without changing the semantic meaning of the sentences. However, there is a lack of consensus on a standard 3D visual tokenization method E^{3D} that is apt for the 3DVL reasoning task, prompting a more detailed exploration in the subsequent paragraphs.

Visual Tokenization. Given an input point cloud $\mathbf{p} \in \mathbb{R}^{N \times 3}$, most prior methods [5, 15, 44] adopt a VoteNet [53] detector to acquire object-level tokens $z^{\mathrm{3D}} \in \mathbb{R}^{N_{\mathrm{obj}} \times C_{\mathrm{obj}}}$ as the visual representation, where N_{obj} is the number of object proposals, and C_{obj} is the object-level feature embeddings. However, we point out several problems with

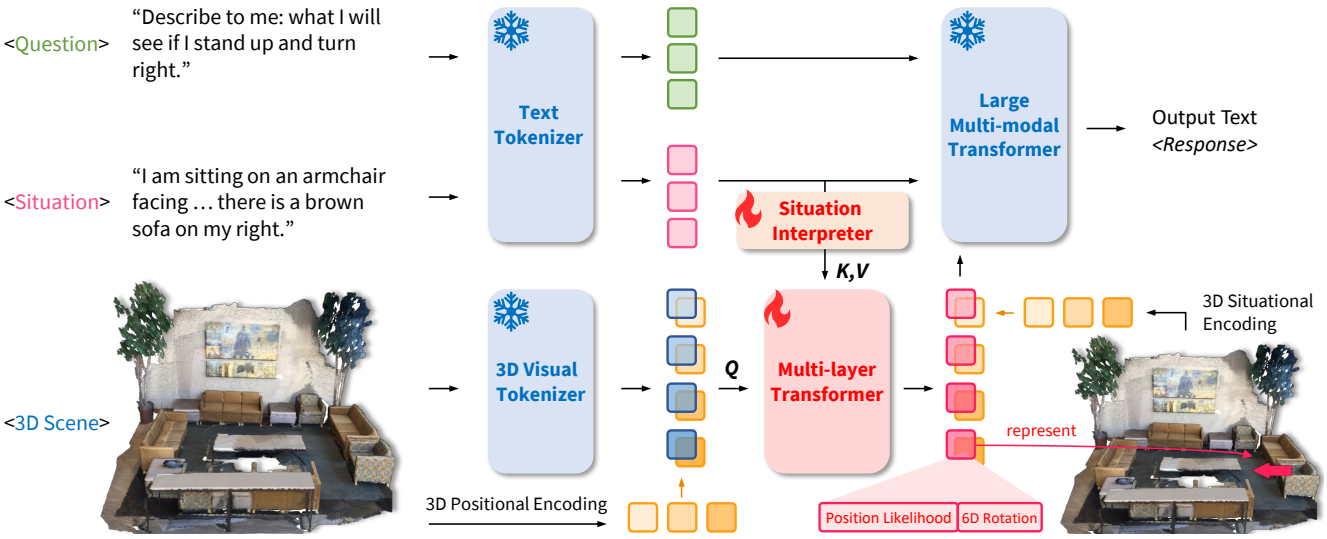

Figure 4. **Overview of our SIG3D model**, which includes 3D scene and text encoding, anchor-based situational estimation, situation-guided visual re-encoding, and multi-modal decoder modules. We tokenize the 3D scene into voxels, treat each token as an anchor point, and query the text tokens to predict a token-level position likelihood and rotation matrix to locate the situational vector associated with the text descriptions. Then we update the scene tokens with situational position encoding (PE), and finally perform the 3DVL reasoning task with a large transformer decoder.

this abstraction strategy: (1) A detection-based tokenization method tends to ignore the non-object regions in the scene, which can be indispensable in some reasoning scenarios (*e.g.*, carpets on the ground, ceiling, walls.) (2) After object-level abstraction, the visual representation losses the high-level information of the scene (*e.g.*, the shape of the living room, the corner of the kitchen.) (3) A supervised detector trained from scratch can only recognize objects within the training set (*e.g.*, only 20 categories for ScanNet [13]), meaning that the method does not have zero-shot capability to reason about novel unseen objects that are inevitably common in real-world scenarios.

In light of this, we adopt a pretrained open-vocabulary voxel-based tokenization method from OpenScene [51]. The scene is first discretized into regular small 3D voxels and fed into a visual encoder for feature extraction:

$$x^{\mathrm{3D}} = \mathrm{E}^{\mathrm{3D}}(\mathcal{V}(\mathbf{p})), \tag{1}$$

where $\mathcal{V}$ represents the voxelization process, and E^{3D} is a Minkowski sparse 3D convolutional network [12]. The sparse network is pretrained by distillation from CLIP [54] embeddings of rendered multi-view 2D images, resulting in a feature map with better language alignment and 3D awareness. We take the upsampled bottleneck-layer feature embeddings from the encoder network, and compute the mean average over the z-axis (vertical) to project the voxels onto the x-y plane and treat the feature grids in the resulting 2D feature map as our N_v visual tokens. We find that this bird's-eye-view projection results in a more compact representation and improves the final performance.

4.2. Situational Estimation

Given 3D visual tokens z^{3D} and situational tokens z^{S}, our objective is to estimate the situational vector $\vec{s}$ referred by the situational description, which comprises of a position component s^{pos} represented by (x, y, z) coordinates, and a rotation component s^{rot} represented by (θ, ψ, ϕ) Euler angles, where pitch angles ψ are always defined as 0, meaning that situational vectors are defined to be parallel with the ground plane. The prior method [44] utilizes a transformer block to calculate the cross-attention feature between visual and language tokens, and directly regress a final situational vector out of the averaged attention map. We find such strategy producing very inaccurate estimation, as shown in Figure 2, due to the large search space in the entire 3D volume. Inspired by the recent 3D object detection methods [45, 66, 71], we reduce the search space by *turning the localization problem into a classification problem*.

Positional Embedding and Feature Fusion. After the voxelization and 3D encoding process, each 3D token associates with a 3D position (x, y, z) representing the center of its voxel. We first provide positional information to the model by generating learnable positional embeddings (PE) using a two-layer perceptron for each of the N_v visual tokens, and add learnable positional embeddings to the token features z^{3D}. We use a situation interpreter [55] to extract situational information, and ask the updated visual tokens to attend to these situational tokens with several transformer layers to produce the joint feature embeddings.

Anchor-based Situational Estimation. We treat each

output token of the feature fusion module as an anchor point, and use it to predict a position likelihood $p \in [0, 1]$ and a rotation estimation. Since each token has an associated 3D position (x, y, z), the position likelihood p indicates how likely the situational vector locates at the center of this token (voxel). We define a soft ground truth for this classification task with a Gaussian kernel, meaning that the closer a token is to the actual situational vector s^{pos}, a higher ground truth probability p will be assigned to that token. In order to counteract the sparse supervisory signal and increase the positive supervision around the vector position, we adopt the peak enlarging technique in CenterPoint [66], where the size of the Gaussian kernel is increased (meaning that the σ is increased) to allow denser supervision around the vector position. Furthermore, we explore different rotation representation and find that compared with quaternion and $(\sin\theta, \cos\theta)$ representations, 6D vector proposed by [72] achieves the best performance. Hence, we adopt a situational estimation head with MLP layers to output 7-dimensional vector for each of the tokens, where the first channel represents the position likelihood and the other six channels represent the 6D rotation matrix. We take the center of the token with the peak position likelihood as our estimated s^{pos}, and convert its corresponding 6D rotation vector as our estimated s^{rot}. The estimation can be equivalently represented as a rotation matrix R and a translation matrix T. More discussion about the architecture and design choices is in Section 5.3.

4.3. Situation-guided Visual Encoding

After obtaining the situational estimation, we investigate a better approach to enhancing the downstream response generation, inspired by human cognitive processes. Intuitively, humans typically comprehend their immediate 3D environment by first interpreting their own situation in space, and then discerning their surroundings from an appropriate viewpoint. Our model is designed to emulate this natural strategy. Utilizing the situational vector $\vec{s}$, we adjust the coordinate system by repositioning the origin at s^{pos}, and reorienting the axes according to s^{rot} such that the new y-axis is aligned with the indicated direction. We keep the z-axis vertically oriented, and project the situational vectors onto the x-y plane. This is in line with the format of the dataset [44], where situational vectors are assumed to be parallel with the ground plane. Subsequently, we compute a new situation-guided PE for each of the N_v visual tokens, akin to the learnable 3D PE outlined in Section 4.2. They allow the model to grasp positional interrelations from the perspective of the current situation. These situational embeddings are added to the output embeddings of the situational estimation module, which consists of blocks featuring self-attention layers for visual tokens, succeeded by cross-attention layers that bridge visual and situational information. This structure allows for the re-encoding of visual tokens under the influence of situational and question context, guiding the model to assign higher weights to situation-related and question-related visual tokens. The output, termed *situation-guided visual tokens*, embodies this re-contextualized understanding.

4.4. Question Answering Head

We follow existing methods [26] to use a large vision-language decoder to fuse the final visual and textual tokens and generate textual response to the input question. We explore both auto-regressive response generation, and classification-based answer prediction [5, 44]. For classification, we predict a vector $v^{\text{ans}} \in \mathbb{R}^{n_a}$ for the n_a answer candidates in the training set following [5].

5. Analysis in 3D VQA Task

We evaluate SIG3D for 3DVL reasoning on two challenging benchmarks, addressing both visually-oriented situational estimation and textually-focused QA tasks. We present a detailed examination of the implementation strategies adopted, the datasets employed, and the metrics applied in our research. For an exhaustive understanding, implementation and training details, and other additional information are available in the **supplementary materials**.

Datasets. We evaluate our method on SQA3D [44] and ScanQA [5], two challenging indoor 3D VQA datasets. Both datasets are derived from the ScanNet dataset [13], serving as the foundational source for their 3D scenes. SQA3D features over 33K question-answer pairs for the 3D VQA task and 26K unique situational descriptions for the situational estimation task. Each entry in this dataset includes a 3D scene point cloud, a situational description, a question, and pertinent annotations. ScanQA consists of over 41K question-answer pairs, without situational descriptions and situational annotations. We use it to demonstrate the generalizability of our method on general QA tasks. We use the splits provided by these datasets.

Evaluation Metrics. For SQA3D, in order to compare with baseline methods [44, 49, 74], we use a shallow transformer decoder task head to perform answer classification task, and evaluate the performance with exact matches (EM@1), which is equivalent to Top-1 answer accuracy. We also provide EM@1 on a breakdown of question types, including "What," "Is," "How," "Can," "Which," and "Other," based on the first word in the question sentence. Additionally, we evaluate situational estimation performance with localization accuracy and orientation accuracy. In both tasks, we use accuracy within different distance or angle thresholds as our metrics. For example, "Acc@0.5m" means accuracy of location estimation when positive threshold is set to 0.5 meter. For ScanQA,

Model	Question Breakdown						Overall
	What	Is	How	Can	Which	Other	
GPT-3 [7]	**39.7**	46.0	40.5	45.6	36.1	38.4	41.0
ClipBERT [35]	30.2	60.1	38.7	63.3	42.5	42.7	43.3
MCAN [67]	28.9	59.7	44.1	68.3	40.7	40.5	43.4
ScanQA [5]	28.6	65.0	47.3	66.3	43.9	42.9	45.3
SQA3D [44]	33.5	66.1	42.4	69.5	43.0	46.4	47.2
Multi-CLIP [15]	-	-	-	-	-	-	48.0
LM4Vision [49]	34.3	67.1	48.2	68.3	48.9	45.6	48.1
3D-LLM [26]	36.5	65.6	47.2	68.8	48.0	46.3	48.1
3D-VisTA [74]	34.8	63.3	45.4	69.8	47.2	**48.1**	48.5
SIG3D (Ours)	35.6	**67.2**	**48.5**	**71.4**	**49.1**	45.8	**52.6**

Table 1. Our proposed method SIG3D achieves state-of-the-art performance on SQA3D benchmark [44]. We perform the best on "Is", "How", and "Can" breakdown types of questions, as well as the average accuracy. The results are reported on test set.

Model	Localization		Orientation	
	Acc@0.5m	Acc@1.0m	Acc@15°	Acc@30°
Random	7.2	25.8	8.4	16.9
SQA3D [44]	9.5	29.6	8.7	16.5
SQA3D (*separate*)	10.3	31.4	17.1	22.8
3D-VisTA [74]	11.7	34.5	16.9	24.2
SIG3D (Ours)	**27.4**	**59.1**	**28.7**	**42.5**

Table 2. Our proposed method SIG3D performs significantly better than prior method [44] in situational estimation task. "Acc@0.5m" stands for localization accuracy with 0.5m threshold. "Acc@15°" represents orientation accuracy with 15° threshold. *separate* means disabling other tasks to let the model focus on situational estimation only.

Model	BLEU-1	BLEU-4	ROUGE	METEOR	CIDEr
BLIP2 [38]	29.7	5.9	26.6	11.3	45.7
Flamingo [3]	25.6	8.4	31.1	11.3	55.0
VN+MCAN [67]	28.0	6.2	29.8	11.4	54.7
SR+MCAN [67]	26.9	7.9	30.0	11.5	55.4
ScanQA [5]	30.2	10.1	33.3	13.1	64.9
3D-LLM [26]	39.3	12.0	35.7	**14.5**	**69.4**
SIG3D	**39.5**	**12.4**	**35.9**	13.4	68.8

Table 3. Performance of SIG3D on ScanQA dataset [5] is on-par with the state-of-the-art with large-scale text-3D pre-training. VN and SR stand for VoteNet and ScanRefer, respectively. 3D-LLM [74] leverages pretrained 2DVL foundation models and LLM models [3, 7, 37, 38], and is pretrained on a large-scale held-in 3D-text dataset before the finetuning on ScanQA.

we perform auto-regressive answer generation with large transformer decoder [26], and evaluate with BLEU [50], ROUGE [39], METEOR [6], and CIDEr [60] metrics.

5.1. Situated Question Answering

Baselines. Our study involves a comparative analysis with a range of representative baselines on the SQA3D dataset. In particular, we evaluate against **GPT-3** [7], **ClipBERT** [35], and **MCAN** [67], which are, as reported in prior work [44], baselines focused on language-only, 2D video, and 2D image QA, respectively. For GPT-3, we follow SQA3D [44] to convert the visual input into a caption using Scan2Cap [10] for LLMs to process. **ScanQA** [5] represents a 3D QA baseline that ignores the situational input. Both **SQA3D** [44] and **Multi-CLIP** [15] employ situational descriptions and annotations for direct regression tasks. **LM4Vision** [49] utilizes LLMs as visual and textual encoders. Additionally, **3D-VisTA** [74] undergoes a pretraining procedure on their large-scale 3D scene-text dataset, ScanScribe, prior to the finetuning on this dataset.

Situational Estimation. As shown in Table 2, our work performs significantly better than the state of the art [44, 74] in both localization and orientation estimation tasks. For 3D-VisTA [74], we use a pretrained model and finetune a new situation head with the SQA3D dataset following [44]. We also report a *random* baseline, in which we randomly sample position and orientation from a uniform distribution as a lower-bound performance. Note that original SQA3D performs only marginally better than the random baseline, meaning that it does not acquire any situational awareness, despite having the situational estimation loss. Disabling the QA task and asking the model to exclusively focus on the situational estimation task results in a slight better performance. Our method, with the anchor-based position likelihood estimation, results in a much better understanding of the 3D situational relationship. Our method also outperforms 3D-VisTA, which is pretrained on a large-scale 3D-text dataset, indicating that large pretraining alone is not enough to address the situational awareness problem. Note that we do not include the random baseline performance reported in [44], because each value is obtained by generating three random values and taking the *closest* one to the ground true, and thus it does not reflect a true "random" baseline.

Situated Question Answering. SIG3D outperforms prior

(a) Number of Visual Tokens

	Acc@1.0m	Acc@30°	EM@1
128	48.9	38.2	49.2
256	**59.1**	**42.5**	**50.9**
512	57.8	42.1	50.7

(b) Voxel size (*in meters*)

	Acc@1.0m	EM@1
0.01	54.1	49.5
0.02	**59.1**	**50.9**
0.05	47.3	48.8

(c) Rotation representation

	Acc@30°	EM@1
Quaternion	31.4	50.0
6D vector	42.5	**50.9**
$\sin\theta, \cos\theta$	**42.6**	50.6

Table 4. Ablation study validates that our various design choices improves the performance. "Acc@1.0m", "Acc@30°", and "EM@1" are accuracy (%) for localization estimation, orientation estimation, and QA tasks. Our settings are marked in gray.

	Acc@1.0m	Acc@30°	EM@1
3D Vision Encoders			
Text-only (*no vision input*)	-	-	47.5
VoteNet [53]	37.4	28.2	49.1
3DETR [48]	47.2	29.1	49.4
OpenScene - OpenSeg [51]	57.5	41.6	50.2
OpenScene - LSeg [51]	**59.1**	**42.5**	**50.9**
Language Tokenizer / Encoders			
GloVe + LSTM [24, 52]	44.3	30.9	48.7
SBERT - MiniLM [55]	56.1	38.6	49.4
SBERT - MPNet [55]	55.9	40.6	49.7
SBERT - MPNet (*finetune*)	**59.1**	**42.5**	**50.9**

Table 5. Performance of SIG3D improves with stronger visual and language encoders. We find that open-vocabulary point encoder and MPNet-based sentence BERT leads to best performance. "Acc@1.0m" and "Acc@30°" stands for localization and orientation accuracy in situational estimation task. "EM@1" demonstrate exact match metric in QA task.

methods in most question breakdown categories and overall accuracy, as shown in Table 1. Our work achieves leading results *without large-scale pretraining* (compared with 3D-VisTA) and LLM (compared with GPT-3), indicating its superiority in situational awareness. Note that the LLM baseline GPT-3 achieves the best performance on the "What" category, suggesting the potential of stronger language encoder in interpreting the complicated question prompt.

5.2. General Question Answering on ScanQA

Baselines. We compare with 2D image VQA **MCAN**-based baselines [67], **ScanQA** [5], **3D-LLM** [26] which leverages large-scale pretrained 2D VLMs and LLMs as backbone models, and **3D-VisTA** [74] pretrained on their proposed large-scale 3D-text dataset.

Question Answering. As shown in Table 3, despite that the questions do not explicitly require situational understanding to answer in ScanQA, SIG3D achieves comparable results with state-of-the-art methods without the large-scale 3D-text pretraining and powerful 2D VLM and LLM backbone models. Our work pretrained on SQA3D [44] leads to higher performance on BLEU-1, BLEU-4, and ROUGE metrics, showing its generalizability on general 3D QA scenarios.

	Acc@1.0m	Acc@30°	EM@1
Baseline (*joint optimization*)	29.5	23.1	47.7
How to achieve better situational estimation			
+ 3D PE	38.8	23.6	47.8
+ 6D Representation	38.5	27.4	47.7
+ Anchor-based Estimation	58.8	41.9	48.2
How to utilize situational estimation for better QA			
+ 3D Situational PE	58.9	41.8	50.0
+ Visual Token Re-encoding	**59.1**	**42.5**	**50.9**
Oracle Models (Ground Truth Situation Information)			
Situation as direct input	100	100	47.7
Situation as intermediate input	100	100	53.9

Table 6. Ablation study verifies that our proposed modules leads to better situational estimation and better QA performance.

5.3. Ablation Study & Analysis

Vision & Language Encoders. We study the impacts of different visual and textual tokenizers in Table 5. It is observed that the open-vocabulary visual encoder (OpenScene) outperforms detection-based encoders (such as VoteNet and 3DETR) across all metrics. This superior performance of OpenScene is attributed to the limitations of 3D detectors, which are typically trained on a limited set of object categories, rendering them less effective in recognizing novel objects mentioned in textual prompts. Regarding language encoders, our findings indicate that a stronger backbone correlates with better performance, primarily due to its improved capability to interpret complex textual inputs. This leads to the suggestion of integrating LLM with our method to potentially further enhance performance, an avenue we intend to explore in future research.

Situated Awareness. In Table 6 we verify the crucial role of situational awareness in 3DVL task. Firstly, we show that 3D PE, 6D rotation estimation, and anchor-based position estimation all leads to much better position and orientation estimation performance. We further establish that situational PE and visual token re-encoding modules lead to better utilization of the predicted situational vector for QA task. Additionally, We design two oracle models under the assumption of having access to the ground truth situational vector as an input. The outcomes from these models reveal a critical insight: the model fails to effectively interpret situational information when it is directly incorporated into the input visual embeddings. This underlines the necessity

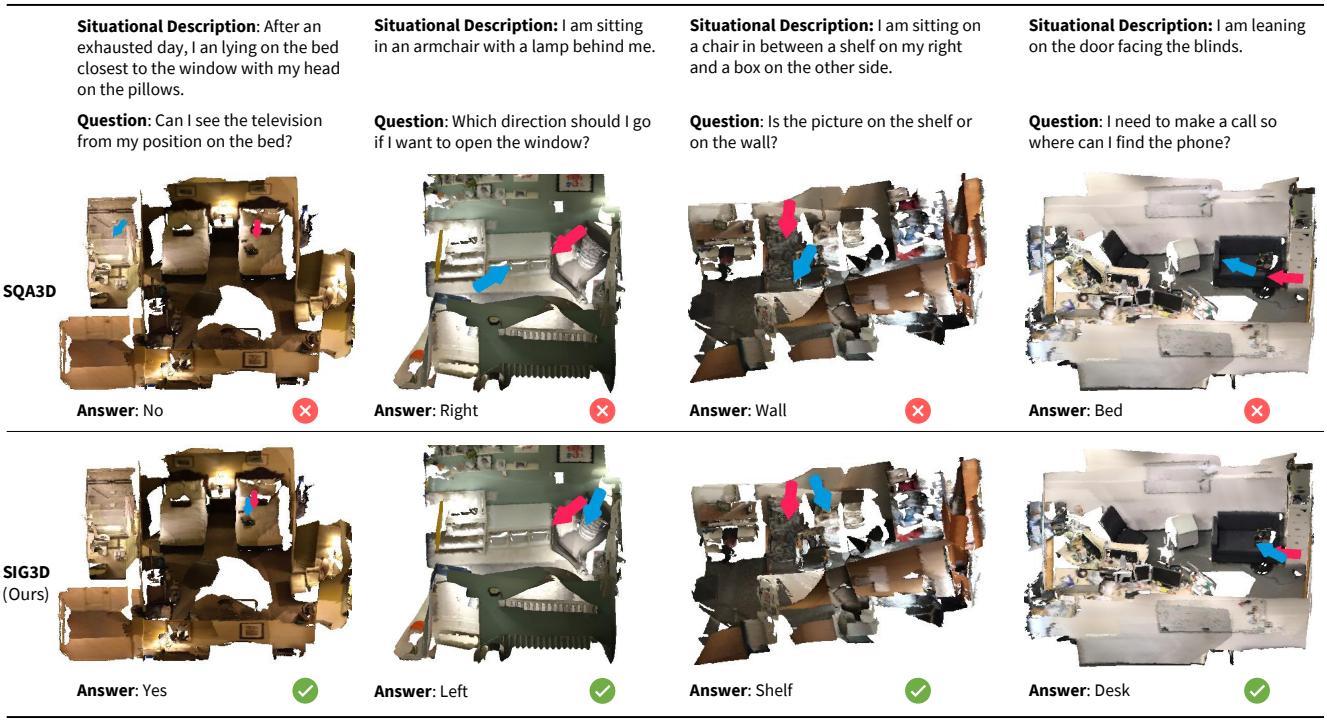

Figure 5. Qualitative results demonstrate significant improvement of SIG3D over prior method. The first row is results from SQA3D [44], and the second row is results from our methods. In the 3D scene, red: Ground truth (GT) vector, and blue: Estimated vector.

of the intermediate representation and encoding mechanism we have proposed, affirming its importance in achieving optimal 3DVL task performance.

Architectural Design. We explore different architectural design choices of our model in Table 4. We find that the number of visual tokens sampled from the visual feature embeddings affects the performance of both situational estimation and QA tasks. Sampling fewer visual tokens increases the risk of missing the region of significance, while sampling more does not lead to a better performance as well. We study the size of voxels and find 0.02m to be the most effective choice, as OpenScene [51] backbone is pretrained with the same voxel size. We also find $\sin\theta, \cos\theta$ and 6D vector representation performs a lot better than quaternion in rotation estimation task. This is consistent with the finding reported in [72].

5.4. Qualitative Analysis

Finally, we demonstrate some qualitative results of our SIG3D in Figure 5. We show the ground truth and estimated situational vectors in red and blue colors, respectively, in their corresponding 3D scenes. We also print the answers with red cross or green checkmark indicating the correctness. It is clear that our method performs significantly better in situational estimation task, resulting in vectors very close to the ground truth in both position and orientation perspectives. The better situational awareness also aids the complicated embodied navigation and common sense QA activities. This also demonstrate great potential of our method in the development of indoor robotics and/or conversational agents.

Supplementary Material. The supplementary section offers an extensive analysis, encompassing a detailed examination of the *3D visual token activation changes* pre and post situational re-encoding. Additionally, it includes a comprehensive collection of *positive and negative samples*, an insightful *failure case analysis*, and a forward-looking discussion on *limitations and future work*.

6. Conclusion

In this paper, we introduce SIG3D, a situation-aware vision language model for 3D reasoning tasks. We propose to represent 3D scene as feature tokens, treat tokens as anchors points to estimate a situational vector from a text description, and use the estimated situation as guidance to align and re-encode the visual tokens to enhance the features for reasoning tasks. We observe consistent and significant performance gain on both situational estimation and question answering tasks.

Acknowledgement. This work was supported in part by NSF Grant 2106825, NIFA Award 2020-67021-32799, the Jump ARCHES endowment, and the IBM-Illinois Discovery Accelerator Institute. This work used NVIDIA GPUs at NCSA Delta through allocations CIS220014, CIS230012, and CIS230013 from the ACCESS program.

References

[1] Panos Achlioptas, Judy Fan, Robert Hawkins, Noah Goodman, and Leonidas J Guibas. ShapeGlot: Learning language for shape differentiation. In *ICCV*, 2019. 3

[2] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3D: Neural listeners for fine-grained 3D object identification in real-world scenes. In *ECCV*, 2020. 3

[3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *NeurIPS*, 2022. 1, 2, 6

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *ICCV*, 2015. 2

[5] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. ScanQA: 3D question answering for spatial scene understanding. In *CVPR*, 2022. 2, 3, 5, 6, 7

[6] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005. 6

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 2020. 1, 2, 3, 6

[8] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3D object localization in rgb-d scans using natural language. In *ECCV*, 2020. 3

[9] Kevin Chen, Christopher B Choy, Manolis Savva, Angel X Chang, Thomas Funkhouser, and Silvio Savarese. Text2shape: Generating shapes from natural language by learning joint embeddings. In *ACCV*, 2019. 3

[10] Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. In *CVPR*, 2021. 3, 6

[11] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 1

[12] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4D Spatio-Temporal ConvNets: Minkowski Convolutional Neural Networks. In *CVPR*, 2019. 4, 1

[13] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3D reconstructions of indoor scenes. In *CVPR*, 2017. 1, 4, 5

[14] Samyak Datta, Sameer Dharur, Vincent Cartillier, Ruta Desai, Mukul Khanna, Dhruv Batra, and Devi Parikh. Episodic memory question answering. In *CVPR*, 2022. 2

[15] Alexandros Delitzas, Maria Parelli, Nikolas Hars, Georgios Vlassis, Sotirios Anagnostidis, Gregor Bachmann, and Thomas Hofmann. Multi-CLIP: Contrastive Vision-Language Pre-training for Question Answering tasks in 3D Scenes. In *BMVC*, 2023. 3, 6

[16] Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. PLA: Language-Driven Open-Vocabulary 3D Scene Understanding. In *CVPR*, 2023. 3

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 2

[18] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. An empirical study of training end-to-end vision-and-language transformers. In *CVPR*, 2022. 2

[19] Mingtao Feng, Zhen Li, Qi Li, Liang Zhang, XiangDong Zhang, Guangming Zhu, Hui Zhang, Yaonan Wang, and Ajmal Mian. Free-form description guided 3D visual graph network for object grounding in point cloud. In *ICCV*, 2021. 3

[20] François Gardères, Maryam Ziaeefard, Baptiste Abeloos, and Freddy Lecue. Conceptbert: Concept-aware representation for visual question answering. In *EMNLP*, 2020. 2

[21] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *ECCV*, 2022. 2

[22] Liangke Gui, Borui Wang, Qiuyuan Huang, Alex Hauptmann, Yonatan Bisk, and Jianfeng Gao. Kat: A knowledge augmented transformer for vision-and-language. In *NAACL*, 2022. 2

[23] Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven Hoi. From Images to Textual Prompts: Zero-shot Visual Question Answering with Frozen Large Language Models. In *CVPR*, 2023. 2

[24] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 1997. 7

[25] Yining Hong, Chunru Lin, Yilun Du, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. 3D Concept Learning and Reasoning from Multi-View Images. In *CVPR*, 2023. 3

[26] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3D-LLM: Injecting the 3D World into Large Language Models. *NeurIPS*, 2023. 1, 2, 3, 5, 6, 7

[27] Pin-Hao Huang, Han-Hung Lee, Hwann-Tzong Chen, and Tyng-Luh Liu. Text-guided graph neural networks for referring 3D instance segmentation. In *AAAI*, 2021. 3

[28] Baoxiong Jia, Yixin Chen, Siyuan Huang, Yixin Zhu, and Song-Chun Zhu. LEMMA: A Multi-view Dataset for LE arning M ulti-agent M ulti-task A ctivities. In *ECCV*, 2020. 2

[29] Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. Egotaskqa: Understanding human tasks in egocentric videos. In *NeurIPS*, 2022. 2

[30] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom

Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 2

[31] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*, 2019. 1, 2

[32] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *iccv*, 2023. 3

[33] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, 2021. 2

[34] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2

[35] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *CVPR*, 2021. 6

[36] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven Semantic Segmentation. In *ICLR*, 2022. 2

[37] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 1, 2, 6

[38] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ICML*, 2023. 2, 6, 1

[39] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 2004. 6

[40] Xudong Lin, Simran Tiwari, Shiyuan Huang, Manling Li, Mike Zheng Shou, Heng Ji, and Shih-Fu Chang. Towards fast adaptation of pretrained contrastive models for multi-channel video-language retrieval. In *CVPR*, 2023. 2, 1

[41] Yuanze Lin, Yujia Xie, Dongdong Chen, Yichong Xu, Chenguang Zhu, and Lu Yuan. Revive: Regional visual representation matters in knowledge-based visual question answering. In *NeurIPS*, 2022. 2

[42] Zhiqiu Lin, Samuel Yu, Zhiyi Kuang, Deepak Pathak, and Deva Ramanan. Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models. In *CVPR*, 2023. 1

[43] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[44] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. SQA3D: Situated question answering in 3D scenes. In *ICLR*, 2023. 1, 2, 3, 4, 5, 6, 7, 8

[45] Yunze Man, Liang-Yan Gui, and Yu-Xiong Wang. BEV-Guided Multi-Modality Fusion for Driving Perception. In *CVPR*, 2023. 4

[46] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*, 2019. 2

[47] Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. Linearly mapping from image to text space. *ICLR*, 2023. 2

[48] Ishan Misra, Rohit Girdhar, and Armand Joulin. An end-to-end transformer model for 3D object detection. In *ICCV*, 2021. 7

[49] Ziqi Pang, Ziyang Xie, Yunze Man, and Yu-Xiong Wang. Frozen Transformers in Language Models Are Effective Visual Encoder Layers. *arXiv preprint arXiv:2310.12973*, 2023. 2, 5, 6

[50] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002. 6

[51] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3D scene understanding with open vocabularies. In *CVPR*, 2023. 3, 4, 7, 8, 1

[52] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014. 7

[53] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3D object detection in point clouds. In *ICCV*, 2019. 3, 7

[54] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 2, 4

[55] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *EMNLP*, 2019. 4, 7, 1

[56] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2

[57] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *ECCV*, 2022. 2

[58] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1, 2

[59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 2, 1

[60] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015. 6

[61] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *arXiv preprint arXiv:2305.11175*, 2023. 1

[62] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. In *NeurIPS (Datasets and Benchmarks Track)*, 2021. 2

[63] Jialin Wu, Jiasen Lu, Ashish Sabharwal, and Roozbeh Mottaghi. Multi-modal answer validation for knowledge-based vqa. In *AAAI*, 2022. 2

[64] Jianing Yang, Xuweiyi Chen, Shengyi Qian, Nikhil Madaan, Madhavan Iyengar, David F Fouhey, and Joyce Chai. LLM-Grounder: Open-Vocabulary 3D Visual Grounding with Large Language Model as an Agent. *arXiv preprint arXiv:2309.12311*, 2023. 3

[65] Shuquan Ye, Dongdong Chen, Songfang Han, and Jing Liao. 3D question answering. *IEEE Transactions on Visualization and Computer Graphics*, 2022. 3

[66] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3D object detection and tracking. In *CVPR*, 2021. 4, 5

[67] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *CVPR*, 2019. 6, 7, 1

[68] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, 2019. 2

[69] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 2

[70] Yaqi Zhang, Di Huang, Bin Liu, Shixiang Tang, Yan Lu, Lu Chen, Lei Bai, Qi Chu, Nenghai Yu, and Wanli Ouyang. MotionGPT: Finetuned LLMs are General-Purpose Motion Generators. *arXiv preprint arXiv:2306.10900*, 2023. 2

[71] Brady Zhou and Philipp Krähenbühl. Cross-view Transformers for real-time Map-view Semantic Segmentation. In *CVPR*, 2022. 4

[72] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, 2019. 5, 8

[73] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *CVPR*, 2016. 2

[74] Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3D-VisTA: Pre-trained Transformer for 3D Vision and Text Alignment. In *ICCV*, 2023. 3, 5, 6, 7