DSPNet: Dual-vision Scene Perception for Robust 3D Question Answering

Jingzhou Luo[1] Yang Liu[1,5*] Weixing Chen[1] Zhen Li[2] Yaowei Wang[3,4] Guanbin Li[1,4,5] Liang Lin[1,4,5]
[1]Sun Yat-sen University, China [2]The Chinese University of Hong Kong, Shenzhen
[3]Harbin Institute of Technology, Shenzhen [4]Peng Cheng Laboratory
[5]Guangdong Key Laboratory of Big Data Analysis and Processing
{luojzh5,chenwx228}@mail2.sysu.edu.cn, {liuy856,liguanbin}@mail.sysu.edu.cn,
lizhen@cuhk.edu.cn, yaoweiwang@gmail.com, linliang@ieee.org

Abstract

*3D Question Answering (3D QA) requires the model to comprehensively understand its situated 3D scene described by the text, then reason about its surrounding environment and answer a question under that situation. However, existing methods usually rely on global scene perception from pure 3D point clouds and overlook the importance of rich local texture details from multi-view images. Moreover, due to the inherent noise in camera poses and complex occlusions, there exists significant feature degradation and reduced feature robustness problems when aligning 3D point cloud with multi-view images. In this paper, we propose a **Dual-vision Scene Perception Network (DSPNet)**, to comprehensively integrate multi-view and point cloud features to improve robustness in 3D QA. Our Text-guided Multi-view Fusion (TGMF) module prioritizes image views that closely match the semantic content of the text. To adaptively fuse back-projected multi-view images with point cloud features, we design the Adaptive Dual-vision Perception (ADVP) module, enhancing 3D scene comprehension. Additionally, our Multimodal Context-guided Reasoning (MCGR) module facilitates robust reasoning by integrating contextual information across visual and linguistic modalities. Experimental results on SQA3D and ScanQA datasets demonstrate the superiority of our DSPNet. Codes will be available at https://github.com/LZ-CH/DSPNet.*

1. Introduction

Recently, 3D Question Answering (3D QA), the task of answering questions about 3D scenes, has emerged as a significant research area in artificial intelligence [22]. Unlike traditional 2D Question Answering (2D QA), which relies on flat images, 3D QA offers the potential for richer spatial comprehension and immersive interaction. The expansion of QA tasks from 2D to 3D also broadens the scope of cross-domain applications, such as visual language navigation [2, 36], embodied agents [14, 31] , and autonomous driving [4, 28, 37]. However, compared with the 2D VQA task, 3D QA introduces unique challenges that extend beyond planar visual understanding. These challenges primarily manifest in the necessity to accurately perceive complex geometric relations among multiple scene entities and effectively reason about spatially semantic dependencies through natural language interactions in 3D environments.

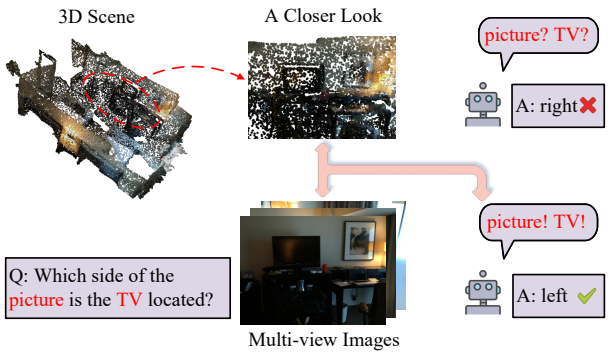

Figure 1. Comprehensive scene perception with dual-vision (point clouds and multi-view images).

Many efforts have been made to address the challenges of 3D QA. For example, ScanQA [3] introduced a 3D perception-based model to fuse 3D and language information. To capture rich high-level semantic relations among objects, 3DGraphQA [38] proposed a Graph Transformer-based model for intra-graph and inter-graph feature fusion. However, most of these methods predominantly rely on 3D point clouds as the primary source of visual information, overlooking the critical role of multi-view images for comprehensive 3D scene perception and reasoning. For example, consider the question given in Fig. 1, "Which side of the picture is the TV located?" not only requires recognizing entities in geometric scenes but also understanding the

*Corresponding Author

complex semantic and spatial relations between scene entities and questions. However, it is difficult for existing 3D QA models to accurately identify some flat and small objects (*e.g.*, TV, picture, carpet, phone, etc.) by relying solely on point cloud information, while multi-view images can make up for this with rich local texture details [8, 12].

To take advantage of multi-view images, a naive approach inspired by 3DMV [9] is to back-project the multi-view image features into point cloud coordinates, pool the features from multiple views for aggregation, and then simply concatenate them with the point cloud features. However, experiments conducted by ScanQA [3] demonstrated that this straightforward approach is ineffective for their 3D QA tasks. We believe this is due to the inherent limitations of the back-projection. As shown in Fig. 2(a), the weights of each view remain fixed when aggregating multi-view features, though ideally, the importance of features from different views should vary based on the specific question. Furthermore, as shown in Fig. 2(b), inherent noise in camera poses, the absence of certain views, and complex occlusions lead to unavoidable feature degradation during back-projection from multi-view images to 3D point cloud space. This reduces feature reliability, especially at the edges of the field of view and in occluded regions.

To address the aforementioned issues, we propose DSPNet, a novel dual-vision scene perception network designed to comprehensively integrate multi-view and point cloud features, adaptively fuse visual information, and perform more effective context-guided reasoning for robust 3D QA. To prioritize view images more closely aligned with the textual content, we introduce a Text-Guided Multi-View Fusion (TGMF) module to integrate back-projected multi-view features by weighting them according to the learnable importance of each image relative to the question. Facing the inherent limitations of feature degradation in back-projection, we design an Adaptive Dual-vision Perception (ADVP) module to adaptively fuse back-projected image features with point cloud features into a unified point-level visual representation by filtering high-confidence point features and suppressing low-confidence point features. To achieve efficient and detailed vision-language interaction, we propose a Multimodal Context-guided Reasoning (MCGR) module with L layers. This module mitigates the high computational cost and feature redundancy of direct cross-modal attention on dense visual features, as well as the semantic loss caused by downsampling, while preserving spatial fidelity and semantic granularity through context-guided reasoning.

Our DSPNet is validated on the ScanQA [3] and SQA3D datasets [25]. The results demonstrate that our DSPNet achieves state-of-the-art performance on these benchmarks. Our main contributions can be summarized as follows:

- To achieve comprehensive scene perception and reasoning for 3D QA, we propose a novel Dual-vision Scene Perception Network (DSPNet) based on point clouds and multi-view images. Extensive experiments demonstrate that our DSPNet outperforms all baseline methods on SQA3D and ScanQA datasets.
- We introduce a Text-guided Multi-view Fusion (TGMF) module to integrate multi-view image features, allowing the model to prioritize views that are more closely aligned with the text content.
- We design an Adaptive Dual-vision Perception (ADVP) module that adaptively fuses back-projected image features with point cloud features into a unified visual representation, coupled with a Multimodal Context-guided Reasoning (MCGR) module for comprehensive 3D scene reasoning with cross-modal contextual interaction.

2. Related Work

2.1. 3D Question Answering

The current 3D question answering (QA) methods primarily focus on two key task settings [17]: 3D visual question answering (3D VQA) [3] and 3D situated question answering (3D SQA) [25]. 3D VQA focuses on question answering tasks in complex scenes, requiring the model to have strong spatial perception and reasoning capabilities. In contrast, 3D SQA introduces a contextual description of the agent's position and orientation in the task setting, requiring the agent to perceive scene and answer questions from a first-person perspective. Inspired by 2D VQA models (*e.g.*, MCAN [40], GraghVQA [21]), researchers have attempted to design similar architectures in the 3D QA domain to effectively fuse features from 3D point clouds and text.

For 3D VQA, ScanQA [3] designed a fusion module composed of Transformer [32] layers and a fusion layer [40] integrate contextualized word representations of question with object proposal features, followed by a classification layer for answer prediction. For 3D SQA, SQA3D [25] built upon ScanQA [3] by utilizing a shared-parameter text encoder for additional situation description encoding. It sequentially fuses 3D object proposal features with situation description and question features through Transformer [32]. To capture rich high-level semantic relations among objects, 3DGraphQA [38] proposed a graph-based 3D QA method, which consists of a graph transformer model for intra-graph feature fusion and a bilinear graph neural network for inter-graph feature fusion.

Following the recent success of 2D VLM pre-training in various downstream tasks, researchers have begun exploring the pre-training paradigms in 3D scene understanding. These methods aim to obtain universal 3D vision-language representations through knowledge transfer from 2D VLMs or large-scale data pre-training. Multi-CLIP [11] utilized contrastive learning to align 3D scene representations with

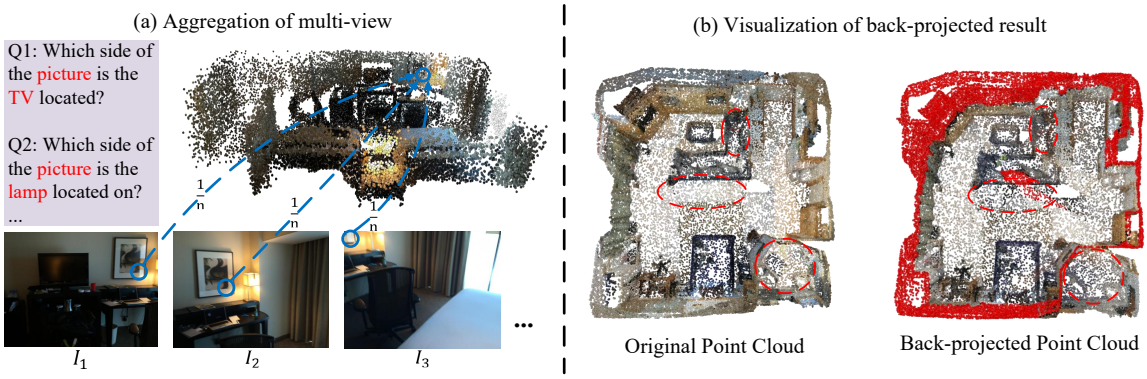

Figure 2. Inherent limitations of back-projection illustrated with a sample from the ScanQA dataset. (a) When aggregating features for coordinates from n mapped multi-view images, each view's weight remains constant at $\frac{1}{n}$, regardless of the question context. (b) Feature degradation occurs during back-projection from multi-view images to 3D point cloud space. Red color points indicate points missed during back-projection, and red ellipses highlight areas with noticeable degradation compared to the original point cloud features.

corresponding text embeddings and multi-view image embeddings in the feature space of CLIP [29] , transferring knowledge from CLIP to enhance 3D vision-language understanding capabilities. 3D-VisTA [42] pre-trained on the large-scale scene-text paired dataset ScanScribe [42], employing masked language modeling, masked object modeling, and scene-text matching strategies. During fine-tuning, 3D-VisTA efficiently adapted to various downstream tasks by adding lightweight task-specific head structures, without requiring additional auxiliary losses or task-specific optimization techniques.

However, most of the existing methods overlook the significance of multi-view images in comprehensive scene perception and reasoning. To address the existing limitations in multi-view image feature fusion and to enhance comprehensive 3D scene reasoning through cross-modal contextual interaction, we propose a novel Dual-vision Scene Perception Network (DSPNet) for robust 3D question answering.

2.2. 3D Visual Grounding

3D visual grounding(3D VG) [1, 6] is a 3D language object localization task. ScanRefer [6] proposed a novel approach for 3D object localization using natural language and presents the first large-scale scene-language dataset (i.e., ScanRefer dataset). Subsequently, the ReferIt3D dataset [1] was introduced to support fine-grained 3D object identification in real-world scenes, providing detailed multi-instance labels to facilitate distinguishing instances within the same object class. Based on the two datasets, a large amount of research [5, 7, 41] have been dedicated to the 3D VG tasks. Additionally, these datasets have also been explored for pre-training in 3D QA (*e.g.*, 3D-VisTA [42], Multi-CLIP [11]). 3D VG focuses on identifying and localizing specific objects in 3D scenes based on natural language descriptions, while 3D QA extends beyond localization to include spatial reasoning and scene understanding to answer questions about the environment.

2.3. Multi-view Based 3D Perception

Recently, fusing point cloud and multi-view images in 3D detection has attracted increasing interest. Early works [8, 9, 19, 39] projected 3D queries to multi-view images for collecting useful semantics. While PointPainting [33] and PointAugmenting [34] directly decorated raw 3D points with 2D semantics. 3D-CVF [39] performed multi-modal fusion at both point and proposal levels. However, since 3D points are inherently sparse, a hard association approach wastes the dense semantic information in 2D features. Recently, multi-modal 3D detectors [15, 18, 20, 35] back-projected dense 2D seeds to 3D space for learning the 2D-3D joint representation in a shared space. However, most of them overlook the limitation of feature degradation in back-projection from 2D to 3D. Moreover, they neglect the influence of textual semantics in multi-view feature fusion.

3. Methodology

In this section, we provide a detailed introduction to our DSPNet, which employs dual-vision comprehensive scene perception to address the task of 3D question answering.

3.1. Overall Architecture

As illustrated in Fig. 3(a), the point cloud P of the 3D scene, the question T, and the multi-view images I are served as input for DSPNet, which aims to predict correct answer vector $\alpha \in \mathbb{R}^{N_\alpha}$ for the N_α answer candidates. DSPNet first encodes the text via a text encoder, encodes the multi-view images via a frozen image encoder, and processes the point cloud via a 3D encoder. Then, we introduce a Text-guided Multi-view Fusion (TGMF) module to fuse the features from multi-view and design an Adaptive Dual-vision Perception (ADVP) module to adaptively perceive the vision

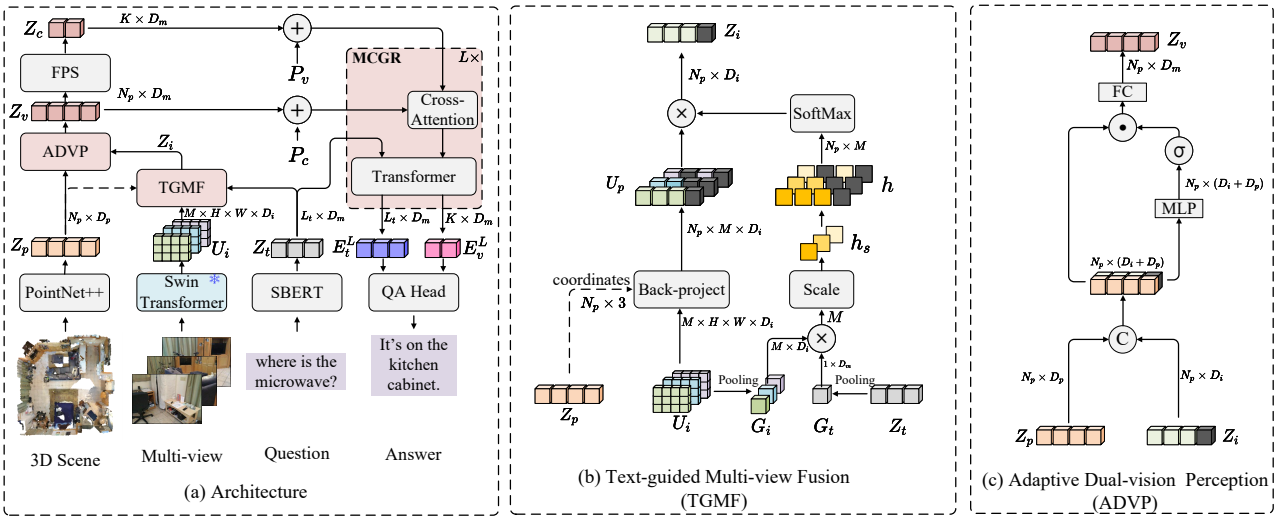

Figure 3. (a) The overall architecture of the DSPNet: it takes the 3D scene, multi-view images, and question as the inputs, ultimately output answers to questions. (b) The Text-guided Multi-view Fusion (TGMF) module aims to fuse the multi-view features. (c) The Adaptive Dual-vision Perception (ADVP) module aims to adaptively perceive the vision information derived from point cloud and multi-view images.

information derived from point cloud and multi-view images. Finally, we incorporate a Multimodal Context-guided Reasoning (MCGR) module that facilitates efficient cross-modal interaction between visual and language for comprehensive scene reasoning and question answering.

3.2. Modal Encoder

3D Encoder. Given an RGB-colored input point cloud $p \in \mathbb{R}^{N\times 6}$, where N represents the number of points, most prior methods [3, 11, 25] adopt a pre-trained VoteNet [27] detector to acquire object-level tokens $Z_p \in \mathbb{R}^{N_o\times D_o}$ as the visual representation, where N_o is the number of object proposals, and D_o is the dimension of object-level feature. However, these methods exhibit several limitations: (1) Detection based feature extraction approaches often overlook non-object areas within the scene, which are essential in some reasoning scenarios (*e.g.*, carpets on the floor, pictures on the wall, hanging lamps on the ceiling). (2) After object-level abstraction, high-level information of the scene (*e.g.*, the layout of a bedroom, the corners of a kitchen) is lost in the visual representation. (3) Joint optimization of detection and reasoning tasks requires careful balancing of their respective loss functions, potentially diverting focus from the primary objective of scene reasoning.

In light of these, we adopt a pre-trained PointNet++ [26] from VoteNet [27] (i.e., VoteHead of VoteNet is discarded) as our 3D encoder. Specifically, the 3D encoder comprises several set abstraction layers and feature propagation (upsampling) layers with skip connections. It processes the input point cloud and outputs a subset of points, referred to as seed points. These seed points are represented by their XYZ coordinates and an enriched feature vector $Z_p \in \mathbb{R}^{N_p\times D_p}$, where N_p denotes the number of seed points, and D_p represents the dimension of the point-level features.

Image Encoder. Given M multi-view images, we employ a pre-trained Swin Transformer [23] to extract multi-view image features $U_i \in \mathbb{R}^{M\times H\times W\times D_i}$, where M denotes the number of images, D_i represents the feature dimension, and $H \times W$ indicates the spatial resolution of the feature maps.

Text Encoder. To robustly capture both local and global features of the situation description and question, we adopt a pre-trained Sentence-BERT (SBERT) [30] to extract context word-level features $Z_t \in \mathbb{R}^{L_t\times D_m}$, where L_t denotes the sequence length and D_m represents the feature dimension. To unify two task settings of 3D VQA and 3D SQA, in 3D SQA, we directly concatenate the situation description and question as the input of the text encoder as in [42].

3.3. Text-guided Multi-view Fusion

To fuse M multi-view image features $U_i \in \mathbb{R}^{M\times H\times W\times D_i}$ from the image encoder, we design a Text-guided Multi-view Fusion (TGMF) module that performs back-projection [9] and text-guided fusion to merge these features.

As illustrated in Fig. 3(b), we initially back-project U_i into 3D coordinates space of Z_p by leveraging the known camera intrinsic and extrinsic parameters associated with each image, obtaining multi-view back-projected features $U_p \in \mathbb{R}^{N_p\times M\times D_i}$ based on point-to-pixel correspondences. Since feature representations of the same entity often vary across different views, especially for entity relations in first-person perspective, we introduce an attention mechanism to learn context-specific importance weights $s \in \mathbb{R}^{N_p\times M}$ of multi-view for each point location. These weights are then

used to preferentially aggregate multi-view information:

$$Q = G_i W_q, \quad K = G_t W_k, \quad h_s = \frac{QK^T}{\sqrt{d_k}} \tag{1}$$

$$s = \text{SoftMax}(h, dim = 1), \quad Z_i = sU_p \tag{2}$$

where $G_i \in \mathbb{R}^{M \times D_i}$ is the global pooling feature of multi-view image features U_i, $G_t \in \mathbb{R}^{1 \times D_m}$ is the global pooling feature of contextualized word-level text features Z_t extracted by the text encoder, $W_q \in \mathbb{R}^{D_i \times d_k}$ and $W_k \in \mathbb{R}^{D_m \times d_k}$ are learnable weights that project G_i and G_t into a same latent space, $h \in \mathbb{R}^{N_p \times M}$ is derived by duplicating $h_s \in \mathbb{R}^{M}$ across all valid back-projected points, $Z_i \in \mathbb{R}^{N_p \times D_i}$ denotes the weighted aggregated feature from multi-view back-projected features.

3.4. Adaptive Dual-vision Perception

Given the back-projected features Z_i and point cloud features Z_p, we aim to adaptively fuse the texture-rich back-projected features with spatially-informative point cloud features into a unified visual representation.

Inspired by SENet [13], we design an Adaptive Dual-vision Perception (ADVP) module to point-wise and channel-wise filter high-confidence features and suppress low-confidence ones. As illustrated in Fig. 3(c), after concatenating the back-projected features Z_i and point cloud features Z_p, we utilize a MLP : $\mathbb{R}^{D_i+D_p} \rightarrow \mathbb{R}^{D_i+D_p}$ and sigmoid function (σ) to learn the importance of each feature channel at each point. Let $Z_h \in \mathbb{R}^{N_p \times (D_i+D_p)}$ denote the refined features, which are calculated as follows:

$$Z_h = \sigma(\text{MLP}([Z_i, Z_p])) \odot [Z_i, Z_p] \tag{3}$$

where $\sigma(\text{MLP}([Z_i, Z_p])) \in \mathbb{R}^{N_p \times (D_i+D_p)}$, and $\odot$ denotes element-wise multiplication. Finally, a fully connected layers (FC), are employed to map and obtain refined point-level visual features $Z_v \in \mathbb{R}^{N_p \times D_m}$:

$$Z_v = \text{FC}(Z_h) \tag{4}$$

3.5. Multimodal Context-guided Reasoning

After acquiring refined point-level visual features Z_v and the contextualized word-level text features Z_t extracted by the text encoder, we aim to derive cross-modal representations through vision-language interaction in the shared semantic space. Therefore, we introduce a Multimodal Context-guided Reasoning (MCGR) module with L layers, which ensures computational efficiency while mitigating the semantic information loss caused by downsampling.

Initially, we apply farthest point sampling (FPS) to sample K points from the dense point-level visual features Z_v, resulting in sparse candidate features $Z_c \in \mathbb{R}^{K \times D_m}$. We then get the position embeddings $P_{v(c)}$ by passing the corresponding coordinates $p_{v(c)}$ through a learnable $\text{MLP}_{v(c)}$: $\mathbb{R}^3 \rightarrow \mathbb{R}^{D_m}$. The position embeddings are added to Z_v and Z_c to form the dense visual embeddings E_v and sparse visual embeddings E_c, respectively:

$$E_{v(c)} = Z_{v(c)} + \text{MLP}_{v(c)}(p_{v(c)}) \tag{5}$$

We send E_v, E_c and text features $Z_t \in \mathbb{R}^{L_t \times D_m}$ to MCGR module, each layer of which contains a cross-attention and transformer sub-layer. Inside the i-th layer, the cross-attention sub-layer is first applied. The query are the fused visual output E_c^{i-1} (where $E_c^0 = E_c$ initially) of the (i-1)-th layer, and context vectors are the dense visual embeddings E_v:

$$h_c^i = \text{CrossAtt}(E_c^{i-1}, E_v) \tag{6}$$

This interactive process can capture essential point features from dense point visual features under context guidance. And then h_c^i interacts with the fused text feature E_t^{i-1} (where $E_t^0 = Z_t$ initially) through a transformer sub-layer:

$$[E_c^i, E_t^i] = \text{Transformer}([h_c^i, E_t^{i-1}]) \tag{7}$$

where $[\cdot, \cdot]$ denotes the concatenation operation along the sequence dimension.

3.6. Training Objective

Question Answering Head. Following [3], we feed the fused text features output $E_t^L \in \mathbb{R}^{L_t \times D_m}$ and the fused visual features output $E_v^L \in \mathbb{R}^{K \times D_m}$ into a modular co-attention network (MCAN) [40] to predict answer $\alpha \in \mathbb{R}^{N_\alpha}$ for the N_α answer candidates.

3D VQA task. We model the final loss as a linear combination of answer classification loss L_{ans}, object classification loss L_{cls} and reference object center localization loss L_{loc}. For L_{loc}, the location closes to a ground truth object center (within 0.3 meters) is considered a ground truth location, as in [3, 27]. We consider the above three training objectives as multi-label classification problems. Since the labels are noisy (*e.g.*, some samples' answers do not include all the correct answers with different expressions in the candidate set, and some samples do not have the ground truth labels of the reference objects), we use soft-ranked cross entropy loss as the loss function for multi-class classification:

$$L(y, p) = -\log\left(\sum_{i=1}^{N} y_i p_i\right) \tag{8}$$

where $y_i \in \{0, 1\}$ is the target label, $p_i \in [0, 1]$ is the predicted label confidence through softmax normalization. The final loss is computed as:

$$L_{\text{3DVQA}} = L_{ans} + \lambda_1 L_{cls} + \lambda_2 L_{loc} \tag{9}$$

where λ_1 and λ_2 are the weighting factors. In our experiments, we set all these hyper-parameters to 1.

3D SQA task. We use the answer classification loss for training (i.e., $L_{\text{3DSQA}} = L_{ans}$) as in [42].

Method	Pre-trained	Test set						Avg.
		What (1,147)	Is (652)	How (465)	Can (338)	Which (351)	Other (566)	
ClipBERT [16]	×	30.2	60.1	38.7	63.3	42.5	42.7	43.3
MCAN [40]	×	28.9	59.7	44.1	68.3	40.7	40.5	43.4
ScanQA [3]	×	28.6	65.0	47.3	66.3	43.9	42.9	45.3
SQA3D [25]	×	33.5	**66.1**	42.4	69.5	43.0	46.4	47.2
Multi-CLIP [11]	√	-	-	-	-	-	-	48.0
3D-VisTA [42]	×	32.1	62.9	47.7	60.7	45.9	48.9	46.7
3D-VisTA [42]	√	34.8	63.3	45.4	**69.8**	47.2	48.1	48.5
3DGraphQA [38]	×	36.4	64.7	46.1	**69.8**	**47.6**	48.2	49.2
DSPNet (ours)	×	**38.2**	66.0	**51.2**	66.6	42.5	**51.6**	**50.4**

Table 1. The question answering accuracy on the SQA3D dataset. In the test set column: the brackets indicate the number of samples for each type of question. The best results are in bold, and the second-best ones are underlined.

4. Experiments

In this section, we validate our DSPNet on two 3D question answering tasks. The tasks for evaluation are (a) 3D situated question answering on the SQA3D dataset [25] and (b) 3D visual question answering on the ScanQA dataset [3].

4.1. Experimental Setup

Dataset. The ScanQA dataset [3] contains 41,363 diverse question-answer pairs and 3D object localization annotations for 800 indoor 3D scenes of the ScanNet dataset [10]. The SQA3D [25] dataset is designed for embodied scene understanding by integrating situation understanding and situated reasoning. It consists of 6.8k unique situations based on 650 ScanNet scenes, accompanied by 20.4k descriptions and 33.4k diverse reasoning questions for these situations. ScanNet [10] is a large-scale annotated 3D mesh reconstruction dataset for indoor spaces, where each scene contains the raw RGB-D sequences.

Implemetation Details. We begin by uniformly sampling multi-view images from the original video at a 0.1 ratio for each scene. Subsequently, we select 20 multi-view images with a resolution of 224×224 as input, utilizing random sampling during training and uniform sampling during inference. Additionally, for each scene, we sample 40,000 points from the raw point cloud as input, using random sampling during training and farthest point sampling during inference. We use the pointnet++ from pre-trained VoteNet [27], the pre-trained Swin Transformer [23] and the MPNet-based pre-trained Sentence-BERT (SBERT) [30] while training other modules randomly initialized from scratch following end-to-end manners. The K value is set to 256 in the FPS stage preceding the MCGR module, and the hidden size of the MCGR module is set to 768. The network is trained using AdamW [24] optimizer with β_1 = 0.9, β_2 = 0.999 and a weight decay of $1e^{-5}$. We use 4 GPUs with 12 training samples on each to train the model for 12 epochs. The learning rate schedule includes a 500-step warm-up phase, linearly increasing from $5e^{-5}$ to $1e^{-4}$, followed by cosine decay back to $5e^{-5}$. The text encoder use a 0.1× smaller learning rate. We implement DSPNet in Pytorch and train it with NVIDIA GeForce RTX 3090.

Evaluation Metrics. On the ScanQA dataset [3], we employ the same evaluation metrics as [3], which include EM@1 and EM@10, where EM denotes the exact match and EM@K represents the percentage of predictions that exactly match any ground truth answer among the top K predicted answers. Meanwhile, we utilize BLEU-4, ROUGE, METEOR, and CIDEr as the sentence-level evaluation metrics. On the SQA3D dataset [25], we adopt the answer accuracy under different types of questions.

4.2. Results on SQA3D Dataset

Baseline. We perform a comparison evaluation with several representative baselines on the SQA3D dataset. In particular, we evaluate against ClipBERT [16] and MCAN [40] which are, as reported in prior work [25] baselines focused on egocentric video and bird-eye view (BEV) image QA. ScanQA [3] represents a 3D QA baseline that ignores the situational input. SQA3D [25] allows location descriptions and questions to interact separately with object proposal features. Multi-CLIP [11] and 3D-VisTA [42] are pre-trained on external 3D-Text paired datasets before being fine-tuned on this dataset. 3DGraphQA [38] is trained on SQA3D Pro dataset by incorporating multi-view images to complement the image information of the first-person perspective situations in the SQA3D dataset.

Results Analysis. Our DSPNet leverages dual-vision scene perception and reasoning, utilizing multi-view images that contain rich local texture details to achieve a more nuanced understanding of scene intricacies. As shown in Tab. 1, we achieve the best results on *What*, *How* and *Other* questions and outperforms other methods including those pre-trained on external 3D-Text paired datasets in terms of average accuracy. This validates that our DSPNet has competitive question reasoning capability. In contrast, SQA3D, 3D-VisTA, and 3DGraphQA exhibit better performance on sim-

Method	Pre-trained	EM@1	EM@10	BLEU-4	ROUGE	METEOR	CIDEr
Image+MCAN [3]	×	22.3 / 20.8	53.1 / 51.2	14.3 / 9.7	31.3 / 29.2	12.1 / 11.5	60.4 / 55.6
ScanRefer+MCAN [3]	×	20.6 / 19.0	52.4 / 49.7	7.5 / 7.8	30.7 / 28.6	12.0 / 11.4	57.4 / 53.4
ScanQA [3]	×	23.5 / 20.9	56.5 / 54.1	12.0 / 10.8	34.3 / 31.1	13.6 / 12.6	67.3 / 60.2
Multi-CLIP [11]	√	24.0 / 21.5	- / -	12.7 / 12.9	35.4 / 32.6	14.0 / 13.4	68.7 / 63.2
3D-VisTA [42]	×	25.2 / 20.4	55.2 / 51.5	10.5 / 8.7	35.5 / 29.6	13.8 / 11.6	68.6 / 55.7
3D-VisTA [42]	√	**27.0** / 23.0	57.9 / 53.5	**16.0** / 11.9	38.6 / 32.8	15.2 / 12.9	76.6 / 62.6
3DGraphQA [38]	×	25.6 / 22.3	58.7 / **56.1**	15.1 / 12.9	36.9 / 33.0	14.7 / 13.6	74.6 / 62.9
DSPNet (ours)	×	26.5 / **23.8**	**58.8** / **56.1**	15.4 / **15.7**	**39.3** / **35.1**	**15.7** / **14.3**	**78.1** / **69.6**

Table 2. Answer accuracy on ScanQA. Each entry denotes "test w/ object" / "test w/o object". The best results are marked bold, and the second-best ones are underlined.

TGMF	ADVP	MCGR	ScanQA	SQA3D
×	×	×	22.35	49.33
√	×	×	22.69	49.58
√	√	×	22.80	49.87
√	×	√	23.23	49.77
√	√	√	**23.47**	**50.36**

Table 3. Ablation study of components in our method.

pler questions with fewer answer options, such as *Is*, *Can*, and *Which*, where answers can often be inferred correctly based on only the question without relying on 3D scenes. However, these methods exhibit limited capabilities in fine-grained perception and reasoning when answering complex and open-ended questions like *What* and *How*. These validate that our DSPNet can comprehensively understand the 3D scene and infer the correct answer.

4.3. Results on ScanQA Dataset

Baseline. We further perform a comparison evaluation with several representative baselines on the ScanQA dataset. In particular, we compare with 2D image VQA MCAN [40] based baselines [3], ScanQA [3], Multi-CLIP [11], 3D-VisTA [42] and 3DGraphQA [38].

Results Analysis. In Tab. 2, our method outperforms existing representative approaches on most evaluation metrics, especially in CIDEr, ROUGE and METEOR, where it significantly surpasses other methods. Specifically, our method's high CIDEr score reflects its effective capture of relevant semantic content in the answers, the elevated ROUGE score indicates comprehensive coverage of key information, and the high METEOR score demonstrates close alignment with reference answers in both vocabulary and structure. Additionally, on the EM@1 metric, our method requires no additional pre-training but achieves performance comparable to 3D-VisTA, which is pre-trained on a external large-scale 3D-Text paired dataset.

4.4. Ablation Studies

We conducted ablation studies on the validation split of the ScanQA dataset and the test split of the SQA3D dataset

Methods	ScanQA	SQA3D
w/o 2D	22.26	49.05
DSPNet (ours)	**23.47**	**50.36**

Table 4. Ablation study on the effectiveness of using 2D modality.

to evaluate the effectiveness of our proposed components. Starting from a baseline model without any of our modules, we incrementally added the Text-guided Multi-view Fusion (TGMF), Adaptive Dual-vision Perception (ADVP), and Multimodal Context-guided Reasoning (MCGR) modules to assess their individual and combined contributions.

Baseline. The baseline model excludes TGMF, ADVP, and MCGR. It uses average pooling to aggregate back-projected multi-view features and employs simple concatenation to combine dual visual features. In the reasoning process, it removes the cross-attention sub-layer, performing only basic interactions between text features and candidate visual features sampled from dense point-level features. As shown in Tab. 3, the baseline achieves EM@1 scores of 22.35% on ScanQA and 49.33% on SQA3D.

Text-guided Multi-view Fusion. Incorporating the TGMF module improves the baseline performance to 22.69% on ScanQA and 49.58% on SQA3D. This indicates that the TGMF, which prioritizes view images aligned with the textual content, is essential for multi-view feature fusion and contributes to overall QA performance.

Adaptive Dual-vision Perception. After building upon the model with TGMF, adding ADVP further improves performance to 22.80% on ScanQA and 49.87% on SQA3D. The results demonstrate the effectiveness of the ADVP module in adaptively fusing back-projected image features with point cloud features.

Multimodal Context-guided Reasoning. Alternatively, adding MCGR to the model with TGMF improves performance to 23.23% on ScanQA and 49.77% on SQA3D. This emphasizes the importance of enhanced multimodal reasoning provided by the MCGR module, as it incorporates dense visual features into contextual interactions via a cross-attention mechanism, significantly preserving scene

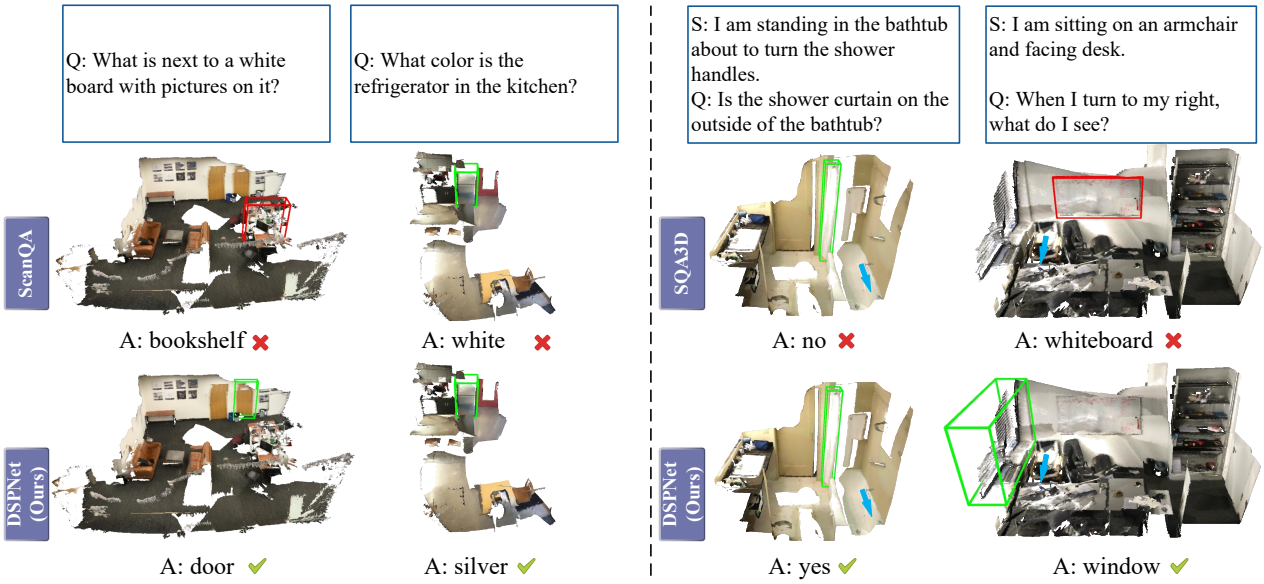

Figure 4. The qualitative comparison of our method with ScanQA and SQA. Our method achieves higher answer accuracy for questions that directly or indirectly involve some challenging entities, such as those with flat shapes and rich local texture details.

(a) Number of Views

	ScanQA	SQA3D
10	22.87	49.73
15	23.04	49.99
20	**23.47**	**50.36**

(b) Depth of MCGR

	ScanQA	SQA3D
2	23.04	49.79
4	**23.47**	**50.36**
6	22.48	49.30

Table 5. Ablation study of various design choices. Our settings are marked in gray.

information and enhancing the model's contextual reasoning capabilities.

Full Model. Combining all three modules, TGMF, ADVP, and MCGR, into our full model yielded the highest performance, achieving EM@1 scores of **23.47%** on ScanQA and **50.36%** on SQA3D. The results confirm that the refined features from TGMF and ADVP enhance MCGR's contextual reasoning, which in turn more effectively utilizes these integrated features, leading to optimal overall performance.

Architectural Design. We compare our full model with a variant, "w/o 2D", which removes all multi-view images and only adopts the 3D point cloud as visual information input. Tab. 4 shows that incorporating local texture details from multi-view images for 3D scene perception and reasoning can bring large improvements to both 3D QA tasks, which proves the necessity of dual vision in our method. In addition, we find that the number of views affects the performance on both 3D QA datasets, especially on ScanQA. As shown in Tab. 5(a), the more views incorporated, the better the performance, as additional views provide richer scene features. We also study the effect of the depth of Multimodal Context-guided Reasoning module by varying the number of layers. As shown in Tab. 5(b), using 4 layers achieves the best performance and simply adding more layers does not help. This is because under the condition of limited-scale 3D QA datasets, deeper networks have stronger representation capabilities but are also more prone to overfitting, so a balance needs to be struck here.

More ablation studies and analyses are provided in the supplementary material.

4.5. Qualitative Analysis

We qualitatively compare our method with ScanQA and SQA3D on the 3D VQA and 3D SQA tasks, respectively. As shown in Fig. 4, our DSPNet performs well in perceiving and reasoning about some challenging entities, such as those with flat shapes and rich local texture details that are difficult to identify based on point cloud geometry alone. Furthermore, DSPNet can distinguish subtle color differences, such as between white and silver, thus enhancing its robustness in identifying fine-grained visual distinctions.

5. Conclusion

In this paper, we propose DSPNet, a dual-vision network for 3D QA. DSPNet integrates multi-view image features via a Text-guided Multi-view Fusion module. It adaptively fuses image and point cloud features into a unified representation using an Adaptive Dual-vision Perception module. Finally, a Multimodal Context-guided Reasoning module is introduced for comprehensive 3D scene reasoning. Experimental results have demonstrated that DSPNet outperforms existing methods with better alignment and closer semantic structure between predicted and reference answers.

Acknowledgement

This work is supported in part by the National Key R&D Program of China under Grant No.2021ZD0111601, in part by the National Natural Science Foundation of China under Grant NO. 62436009, NO. 62002395 and NO. 62322608, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant NO. 2025A1515011874 and No.2023A1515011530, and in part by the Guangzhou Science and Technology Planning Project under Grant No. 2023A04J2030.

References

[1] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 422–440. Springer, 2020. 3

[2] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3674–3683, 2018. 1

[3] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19129–19139, 2022. 1, 2, 4, 5, 6, 7

[4] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 1

[5] Daigang Cai, Lichen Zhao, Jing Zhang, Lu Sheng, and Dong Xu. 3djcg: A unified framework for joint dense captioning and visual grounding on 3d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16464–16473, 2022. 3

[6] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European conference on computer vision*, pages 202–221. Springer, 2020. 3

[7] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Language conditioned spatial relation reasoning for 3d object grounding. *Advances in Neural Information Processing Systems*, 35:20522–20535, 2022. 3

[8] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1907–1915, 2017. 2, 3

[9] Angela Dai and Matthias Nießner. 3dmv: Joint 3d-multi-view prediction for 3d semantic scene segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 452–468, 2018. 2, 3, 4

[10] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 6

[11] 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 *Proc. of the British Machine Vision Conference (BMVC)*, 2023. 2, 3, 4, 6, 7

[12] Deng-Ping Fan, Ge-Peng Ji, Peng Xu, Ming-Ming Cheng, Christos Sakaridis, and Luc Van Gool. Advances in deep concealed scene understanding. *Visual Intelligence*, 1(1):16, 2023. 2

[13] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. 2018. 5

[14] Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024. 1

[15] Yang Jiao, Zequn Jie, Shaoxiang Chen, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. Msmdfusion: Fusing lidar and camera at multiple scales with multi-depth seeds for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21643–21652, 2023. 3

[16] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7331–7341, 2021. 6

[17] Yinjie Lei, Zixuan Wang, Feng Chen, Guoqing Wang, Peng Wang, and Yang Yang. Recent advances in multi-modal 3d scene understanding: A comprehensive survey and evaluation. *arXiv preprint arXiv:2310.15676*, 2023. 2

[18] Yanwei Li, Yilun Chen, Xiaojuan Qi, Zeming Li, Jian Sun, and Jiaya Jia. Unifying voxel-based representation with transformer for 3d object detection. *Advances in Neural Information Processing Systems*, 35:18442–18455, 2022. 3

[19] Ming Liang, Bin Yang, Shenlong Wang, and Raquel Urtasun. Deep continuous fusion for multi-sensor 3d object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 641–656, 2018. 3

[20] Tingting Liang, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao Tang, Bing Wang, and Zhi Tang. Bevfusion: A simple and robust lidar-camera fusion framework. *Advances in Neural Information Processing Systems*, 35:10421–10434, 2022. 3

[21] Weixin Liang, Yanhao Jiang, and Zixuan Liu. Graghvqa: Language-guided graph neural networks for graph-based visual question answering. In *Proceedings of the Third Work-*

shop on Multimodal Artificial Intelligence, pages 79–86. Association for Computational Linguistics, 2021. 2

[22] Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. Aligning cyber space with physical world: A comprehensive survey on embodied ai. *arXiv preprint arXiv:2407.06886*, 2024. 1

[23] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 4, 6

[24] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[25] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. In *The Eleventh International Conference on Learning Representations*, 2022. 2, 4, 6

[26] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 4

[27] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9277–9286, 2019. 4, 5, 6

[28] Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4542–4550, 2024. 1

[29] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[30] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2019. 4, 6

[31] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9339–9347, 2019. 1

[32] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 2

[33] Sourabh Vora, Alex H Lang, Bassam Helou, and Oscar Beijbom. Pointpainting: Sequential fusion for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4604–4612, 2020. 3

[34] Chunwei Wang, Chao Ma, Ming Zhu, and Xiaokang Yang. Pointaugmenting: Cross-modal augmentation for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11794–11803, 2021. 3

[35] Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, et al. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19757–19767, 2024. 3

[36] Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6629–6638, 2019. 1

[37] Yizhou Wang, Longguang Wang, Qingyong Hu, Yan Liu, Ye Zhang, and Yulan Guo. Panoptic segmentation of 3d point clouds with gaussian mixture model in outdoor scenes. *Visual Intelligence*, 2(1):10, 2024. 1

[38] Zizhao Wu, Haohan Li, Gongyi Chen, Zhou Yu, Xiaoling Gu, and Yigang Wang. 3d question answering with scene graph reasoning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1370–1378, 2024. 1, 2, 6, 7

[39] Jin Hyeok Yoo, Yecheol Kim, Jisong Kim, and Jun Won Choi. 3d-cvf: Generating joint camera and lidar features using cross-view spatial feature fusion for 3d object detection. In *Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part XXVII 16*, pages 720–736. Springer, 2020. 3

[40] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6281–6290, 2019. 2, 5, 6, 7

[41] Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3dvg-transformer: Relation modeling for visual grounding on point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2928–2937, 2021. 3

[42] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2911–2921, 2023. 3, 4, 5, 6, 7