Bi-directional Contextual Attention for 3D Dense Captioning

Minjung Kim[1,4,*], Hyung Suk Lim[3], Soonyoung Lee[2], Bumsoo Kim[2,†], and Gunhee Kim[1,†]

[1] Seoul National University, Seoul, Korea
minjung.kim@vision.snu.ac.kr, gunhee@snu.ac.kr
[2] LG AI Research, Seoul, Korea
{soonyoung.lee,bumsoo.kim}@lgresearch.ai
[3] Diquest, Seoul, Korea
hslim@diquest.com
[4] SNU-LG AI Research Center, Seoul, Korea

Abstract. 3D dense captioning is a task involving the localization of objects and the generation of descriptions for each object in a 3D scene. Recent approaches have attempted to incorporate contextual information by modeling relationships with object pairs or aggregating the nearest neighbor features of an object. However, the contextual information constructed in these scenarios is limited in two aspects: first, objects have multiple positional relationships that exist across the entire global scene, not only near the object itself. Second, it faces with contradicting objectives–where localization and attribute descriptions are generated better with tight localization, while descriptions involving global positional relations are generated better with contextualized features of the global scene. To overcome this challenge, we introduce BiCA, a transformer encoder-decoder pipeline that engages in 3D dense captioning for each object with Bi-directional Contextual Attention. Leveraging parallelly decoded instance queries for objects and context queries for non-object contexts, BiCA generates object-aware contexts, where the contexts relevant to each object is summarized, and context-aware objects, where the objects relevant to the summarized object-aware contexts are aggregated. This extension relieves previous methods from the contradicting objectives, enhancing both localization performance and enabling the aggregation of contextual features throughout the global scene; thus improving caption generation performance simultaneously. Extensive experiments on two of the most widely-used 3D dense captioning datasets demonstrate that our proposed method achieves a significant improvement over prior methods.

1 Introduction

3D dense captioning is a task that requires 1) determining the location of all objects and 2) generating descriptive sentences for each object detected within a 3D

* Work done during internship at LG AI Research
† Corresponding authors

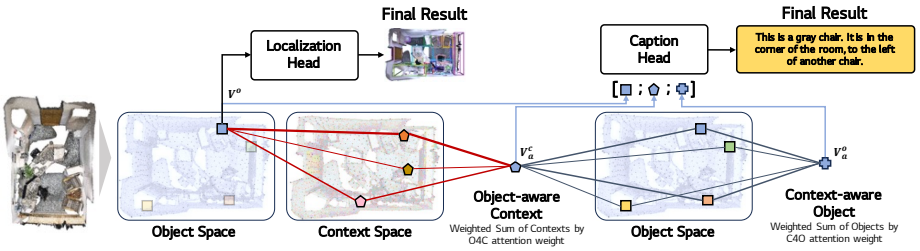

Fig. 1: Conceptual illustration of the multi-stage pipeline of BiCA (best viewed in color).

scene. Early approaches have utilized a two-stage process [4, 7, 11, 19, 33, 36, 37], where the objects are detected first, then captions are generated afterward by sequentially feeding the features of each detected object. Following work [9] has borrowed the end-to-end transformer encoder-decoder pipeline from object detection [5] to actively involve contextual information for 3D dense captioning. In these works, contextual information for predicting relationships is composed by combining object features (e.g., with dot product [4, 33] or transformer attention [9]) and aggregating local nearest neighbor visual features [9].

Despite their improved performance, the scope of contextual information in previous 3D dense captioning methods poses its own set of challenges. First, 3D dense captioning heavily involves various positional relationships throughout the global scene; thus, the context scope restricted to the combination of individual objects or their spatially nearest neighbors typically cannot provide sufficient information. Next, the endeavor to incorporate the contextual features of various positions into a single object encounters a conflicting objective since the object must encompass precise local features for localization while simultaneously capturing various positional relationships throughout the global context for caption generation. Clearly, these challenges limit the performance of 3D dense captioning, which raises an interesting research question– *"Can we design an architecture for 3D dense captioning that can effectively aggregate relevant context features without harming localization performance?"*

To this end, we propose a **Bi**-directional **C**ontextual **A**ttention (BiCA) network for 3D dense captioning. BiCA parallelly decodes a fixed set of *object* queries and *context* queries sampled from the geometric positions that do not overlap with each other. While the decoded object query performs standard object detection, the context queries are designed to distinctively capture *non-object* contexts and subsequently aggregate relevant contextual information for each object. Recognizing that non-object regions often lack sufficient visual features in 3D point clouds, BiCA incorporates a multi-stage pipeline where context features are enriched by relevant object features, thus implementing *bi-directional* attention: i) Contextual Attention of Objects for Context (O4C):

context representation is represented by the attention summarization of global objects. ii) Contextual Attention of Contexts for Object (C4O): the object representation is adjusted by summarizing the non-object contexts.

Then, localization is inferred with object queries while the captions are generated with object-aware context and context-aware object features. By disentangling the attention for localization and the attention for incorporating contextual features, BiCA successfully resolves the aforementioned contradicting objectives for local and global features. Moreover, since the contextual attentions retrieve globally relevant contexts and objects throughout the scene, it could accurately produce descriptions that transcend the spatial boundary of localization and its nearest neighbor. The conceptual overview of our BiCA is illustrated in Figure 1.

BiCA sets a new state-of-the-art standard for 3D dense captioning, while simultaneously improving 3D object detection performance. Extensive experiments on two widely used benchmarks in 3D dense captioning (i.e., ScanRefer [6] and Nr3D [1]) show that our proposed BiCA surpasses prior approaches by a large margin. The contribution of our paper can be summarized as:

- We propose a novel Bi-directional Contextual Attention network with multi-stage contextual attention for 3D dense captioning. This enables our model to capture relevant contexts throughout the global scene without being bound to single-object localization or their nearest neighbors.
- By performing localization with decoded object queries and caption generation by including object-aware context features and context-aware object correspondence features, BiCA can simultaneously improve the performance of localization and caption generation for 3D dense captioning.
- Our BiCA achieves state-of-the-art performances across multiple evaluation metrics on two widely used benchmarks for 3D dense captioning: ScanRefer and Nr3D datasets.

2 Related work

2.1 Image Captioning

Image captioning is a fundamental task in visual language creation that automatically generates descriptive sentences for images. The main goal is to improve 2D scene understanding and generate captions that accurately reflect the content and context of the image [17]. However, with the advancement of deep learning in image captioning, there has been a marked shift towards the development and adoption of more sophisticated models [16]. Image captioning methods using deep learning adopt an encoder-decoder architecture, where the decoder generates a sentence from the visual features extracted by the encoder. Attention-based methods for grid regions [34] and detected objects [2] focus on specific image regions and use graph neural networks [35] or transformer layers to capture relationships between objects [14].

2.2 3D Dense Captioning

3D dense captioning, an emerging field focused on achieving detailed object-level understanding of 3D scenes through natural language descriptions, has garnered significant interest in recent years. This task involves converting 3D visual data [15] into a consistent set of bounding boxes and generating appropriate natural language descriptions for each identified object. This task presents a considerable challenge due to the complexity of 3D scene information.

Most 3D dense captioning models employ an encoder-decoder architecture comprising three main components: a scene encoder, a relational module, and a feature decoder. The scene encoder in 3D dense captioning utilizes 3D object detection methods such as 3DETR [24], PointNet++ [28], VoteNet [27], and PointGroup [18] to extract object-level visual features and contextual details from 3D point cloud datasets. The relational module is crucial in 3D dense captioning, modeling complex connections and cross-modal interactions among objects within indoor scenes. Relational modeling approaches vary based on task requirements and can be categorized into graph-based [7, 11, 19], transformer-based [4, 9, 33], and knowledge distillation-based [36] methods. The feature decoder generates bounding boxes and captions for candidate objects by considering attributes and relational features. GRU-based decoders with attention mechanisms are used in Scan2Cap [11], MORE [19], and D3Net [7], while transformer-based decoders are employed in SpaCap3D [33], χ-Tran2Cap [36], 3DJCG [4], and Vote2Cap-DETR [9] to facilitate the caption generation process.

However, most existing methods use a single query set for object localization and caption generation, which often leads to conflicts in achieving the objectives of the two tasks. To address this, Vote2Cap-DETR++ [10] enhances task-specific feature capture by decoupling queries for localization and captioning, overcoming the limitations of a unified query system. Despite this improvement, challenges persist as its performance remains fundamentally bound by the precision of its object localization capabilities.

3 Method

Our goal is to take a 3D scene as input, identify all the objects within it, and generate captions for each object. Previous research has adopted a pipeline to identify objects in a 3D scene and simultaneously generate captions from the same features. However, while each caption may describe features of the object itself, it can also be based on the relationships with surrounding objects or information about the overall scene.

This section introduces a novel transformer encoder-decoder approach with bi-directional contextual attention for the 3D dense captioning task. This approach distinguishes between features for instances and context information between instances, indirectly supervising the context features to enable end-to-end learning with the instance features. Instance queries aim to localize instances within a 3D scene and capture the characteristics of instances. In contrast, each

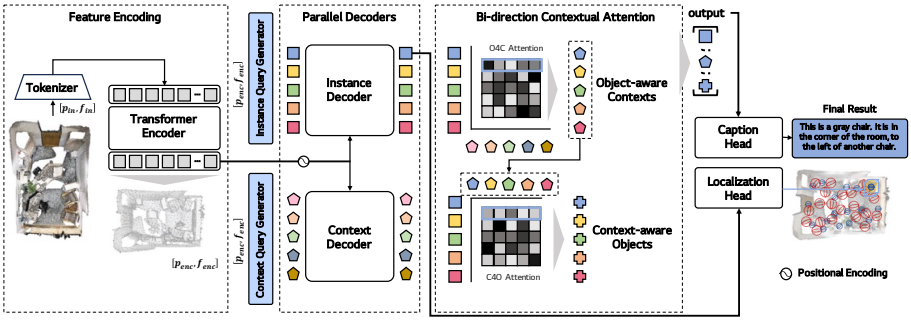

Fig. 2: The overall pipeline of BiCA. We parallelly generate and decode two sets of queries (i.e., Instance Query and Context Query) that encodes the instance features and the non-object context features throughout the global scene, respectively. The object-aware contexts are calculated per each object by the weighted sum of the context queries, where the weights are calculated by the attention between the decoded instance query and context query. Then, with the object-aware context feature, the context-aware object feature is obtained by the weighted sum of the instances, which is weighted by the attention between the object-aware contexts.

context query is designed to capture the non-object region within a scene, including positional relationships between instances in the 3D scene. The overall architecture of the newly proposed model is described in Figure 2.

3.1 Encoder

Following previous work [9], our model also adopts 3DETR [24] encoder as our scene encoder. This encoder is applied to transform the input point cloud into a set of encoded tokens that capture spatial, structural, and contextual information. Given the input point cloud $PC = [p_{\text{in}}; f_{\text{in}}] \in \mathbb{R}^{N\times(3+F)}$, it is initially tokenized by a set-abstraction layer of PointNet++ [28]. This tokenized output is subsequently fed into a masked transformer encoder incorporating the set-abstraction layer, followed by two additional encoder layers. The final encoded scene tokens are denoted as $p_{enc} \in \mathbb{R}^{1,024\times3}$ and $f_{enc} \in \mathbb{R}^{1,024\times256}$.

3.2 Query Generator

To separate captions bound to a single object from captions containing information relative to other objects or the global scene, we specify two separate *Instance Query* and *Context Query* from the encoded scene. While the instance query set is designed to capture individual features of objects bounded by their localization, the context query set is designed to encompass a broad area that consists of the *positional information* of a caption (e.g., next to, on the right of, is most far from, etc.).

Instance Query Generator. In the Instance Query Generator, instance queries are generated to detect all objects in the 3D scene and generate captions for each object's individual attributes. The instance query (p^o, f^o) is defined as follows:

$$[\Delta p_{\text{vote}}; \Delta f_{\text{vote}}] = \text{FFN}_o(f_{\text{enc}}), \tag{1}$$

$$(p^o, f^o) = \text{SA}_o\big(p_{\text{enc}} + \Delta p_{\text{vote}}, f_{\text{enc}} + \Delta f_{\text{vote}}\big), \tag{2}$$

where $[\Delta p_{\text{vote}}; \Delta f_{\text{vote}}] \in \mathbb{R}^{1,024\times(3+256)}$ is an offset that learns to shift the encoded points to object's centers spatially by a feed-forward network FFN_o, following [9]. SA_o represents the set-abstraction layer with a radius of 0.3 and samples 16 points for p^o. Since our Instance Query Generator extracts the features from the candidate coordinates after the voting, our instance queries are not focused on specific objects. However, they are distributed across the locations where objects are present throughout the 3D scene. As a result, 256 instance queries are extracted from our Instance Query Generator.

Context Query Generator. 3D scenes are composed of 3D points arranged in XYZ coordinates, forming specific objects, and the spaces between objects signify space. We recognize that obtaining context features should not merely involve interpolating the spaces between Instance Queries but also encompassing how these points are arranged and the composition of empty spaces to represent a scene. Inspired by this observation, we define *context information* as the geometric details in a 3D scene that illustrate the relationships between objects and between objects and the scene itself. We decide to extract structural information centered on specific points in the 3D scene and call them *Context Queries*. This approach is distinct from the Instance Query Generator. While the Instance Query Generator uses a voting network to obtain object center coordinates that do not exist in the 3D scene and extracts features such as identification from them, the proposed Context Query Generator extracts structural information without disrupting the arrangement of the 3D points.

Given the encoded scene tokens $(p_{\text{enc}}, f_{\text{enc}})$, we sample 512 points p^c_{seed} with farthest point sampling (FPS) [28] on p_{enc} that each represent a unique geometry within the global scene. The context query (p^c, f^c) is then defined as:

$$(p^c, f^c) = \text{SA}_c(p_{\text{enc}}, f_{\text{enc}}), \tag{3}$$

where SA_c denotes the set-abstraction layer [28] with a radius of 1.2, sampling 64 points for p^c. All hyper-parameters are determined experimentally.

3.3 Decoder

Instance Decoder and Context Decoder. Given the context queries (p^c, f^c) and instance queries (p^o, f^o), we build a parallel decoding pipeline where the Context Decoder designates contextual information between objects and the

Instance Decoder performs the localization and attribute description for the participating objects.

Each Instance Decoder and Context Decoder consists of a transformer decoder with 8 layers according to 3DETR [24], and Fourier positional encoding is applied to XYZ coordinates. Fourier positional encoding transforms the XYZ coordinates of each query into a Fourier position embedding [31] and then combines them with the corresponding query embeddings. This encoding transforms the input data into frequency functions, capturing invariant changes in complex shapes and structures. It aids the context queries in accurately identifying geometric shapes and the instance queries in detecting objects. The context queries V^c are afterwards contextualized for each instance query V^o, then retrieve relevant instance query features that are structurally involved.

3.4 Bi-directional Contextual Attention

Captions must be generated for each object in relation to the scene context, and through positional encoding, it becomes possible to understand directional relationships, such as "to the right" with surrounding objects. Here, we gain the insight that information about the relative objects closely related to these relationships is also necessary, and we design the model to proceed simultaneously with a multi-stage contextualization. With the decoded queries, our method proceeds in two stages: i) Contextual Attention of Objects for Context (O4C): context representation is represented by the attention summarization of global objects. ii) Contextual Attention of Contexts for Object (C4O): the object representation is adjusted by summarizing the non-object contexts. Figure 1 and Figure 2 show detailed illustrations of obtaining the object-aware context feature and the context-aware object feature.

Contextual Attention of Objects for Context (O4C). To retrieve the structural relationship between objects throughout the global scene, we first construct a Object-aware Context V_a^c per object by summating the attention of each context position feature with regard to each object. The attention result between the instance feature and the context feature is applied to the context feature, making a weighted summarization of the geometries. At this time, the ratio is adjusted using learnable gamma. We name this result V_a^c.

Contextual Attention of Contexts for Object (C4O). Most contexts not only interact with the reference object but also have other objects that react against them. To achieve this, we go through a Contextual Attention of Contexts for Object process. This could specify the information that can simply be said to be 'next to' as 'next to the red chair'. To this end, the attention result of the context feature and instance feature with the attention weight applied is applied to the instance query feature. Likewise, it is adjusted using a learnable parameter lambda. We name this result V_a^o. V_a^o results from finding instance combinations that have a meaningful relationship with the contextual region focused on a

specific instance. The concatenation of the instance token V^o, contextualized geometry V_a^c, and contextualized object V_a^o, denoted as V^a, is fed to the caption head to generate final captions.

3.5 3D Dense Captioning

Our final goal is to identify all objects within the input 3D scene and generate a caption for each object. We perform object detection and caption generation in parallel and group the results into one instance.

Localization. To localize instances in the 3D scene, we use the decoded instance query V^o. Through 5 MLP heads, we reformulate the box corner estimation as offset estimation from a query point to an object's center and box size regression. These localization heads are shared across the decoder layers.

Caption Generation. For caption generation, we use a transformer decoder-based caption head based on GPT-2 [29], following the methods of Vote2Cap-DETR [9] and SpaCap3D [33]. Our caption head comprises two transformer decoder blocks, sinusoid positional embedding, and a linear classification layer. When generating captions for a proposal, we substitute the standard Start Of Sequence ('SOS') prefix with V^a. During the inference process, our approach employs beam search to produce captions. The beam size is 5.

3.6 Training BiCA

Instance Query Loss. To train instance query generator to find an object's center by shifting points p_{enc}, we use the vote loss from VoteNet [27]. Given the generated instance query (p^o, f^o) and the encoded scene tokens (p_{enc}, f_{enc}), the vote loss $\mathcal{L}^o$ is denoted as:

$$\mathcal{L}^o = \frac{1}{M} \sum_{i=1}^{M} \sum_{j=1}^{N_{gt}} \|p_i^o - \mathrm{cnt}_j\|_1 \cdot \mathbb{I}(p_{enc}^i), \tag{4}$$

where $\mathbb{I}(x)$ is an indicator function that equals 1 when $x \in I_j$ and 0 otherwise, N_{gt} is the number of instances in the 3D scene, M is the number of p^o, and cnt_j is the center of j-th instance I_j.

Detection Loss. Following DETR [5], we use Hungarian matching [21] to assign each proposal with the ground truth. The detection loss $\mathcal{L}_{\mathrm{det}}$ is defined as:

$$\mathcal{L}_{\mathrm{det}} = \alpha_1 \mathcal{L}_{\mathrm{giou}} + \alpha_2 \mathcal{L}_{\mathrm{cls}} + \alpha_3 \mathcal{L}_{\mathrm{cnt}} + \alpha_4 \mathcal{L}_{\mathrm{size}}, \tag{5}$$

with $\alpha_1 = 10, \alpha_2 = 1, \alpha_3 = 5, \alpha_4 = 1$ set heuristically. This loss is applied across all decoder layers for better convergence.

Caption Loss. Following the standard image captioning protocol, we first train the caption heads using cross-entropy loss for Maximum Likelihood Estimation (MLE). During MLE training, the model predicts the $(t+1)$-th word c_n^{t+1} based on the first t words $c_n^{[1:t]}$ and the visual features $\mathcal{V}$. The loss function for a final caption of length T is defined as follows:

$$\mathcal{L}_{c_n} = \sum_{t=1}^{T} \mathcal{L}_{c_n}(t) = -\sum_{t=1}^{T} \log \hat{P}\Big(c_n^{t+1} | \mathcal{V}, c_n^{[1:t]}\Big). \tag{6}$$

After word-level supervision, the caption head is refined using Self-Critical Sequence Training (SCST) [30], where the model generates multiple captions $\hat{c}_{1,\ldots,k}$ with a beam size of k and an additional caption $\hat{g}$ via greedy search. The loss function for SCST is formulated as follows:

$$\mathcal{L}_{c_n} = -\sum_{i=1}^{k} (R(\hat{c}_i) - R(\hat{g})) \cdot \frac{1}{|\hat{c}_i|} \log \hat{P}(\hat{c}_i | \mathcal{V}). \tag{7}$$

The reward function $R(\cdot)$ is based on the CIDEr [32] metric for caption evaluation, and the logarithmic probability of the caption $\hat{c}_i$ is normalized by its length $|\hat{c}_i|$, ensuring equal importance to captions of different lengths.

Final Loss. Given the instance query Loss $\mathcal{L}^o$, the detection loss for the i-th decoder layer as $\mathcal{L}^i_{\text{det}}$, and the average of the caption loss $\mathcal{L}_{c_n}$ within a batch, denoted as $\mathcal{L}_{\text{cap}}$, the final loss $\mathcal{L}$ is formulated as:

$$\mathcal{L} = \beta_1 \mathcal{L}^o + \beta_2 \sum_{i=1}^{n_{\text{dec-layer}}} \mathcal{L}^i_{\text{det}} + \beta_3 \mathcal{L}_{\text{cap}} \ , \tag{8}$$

where $\beta_1 = 10$, $\beta_2 = 1$, and $\beta_3 = 5$.

4 Experiments

4.1 Datasets and Metrics

Datasets. We focus on 3D dense captioning, leveraging two benchmark datasets: ScanRefer [6] and Nr3D [1]. These datasets offer an extensive human-generated description of 3D scenes and objects. ScanRefer encompasses 36,665 descriptions covering 7,875 objects within 562 scenes, while Nr3D contains 32,919 descriptions for 4,664 objects across 511 scenes. Both datasets draw their training data from the ScanNet [15] database, which includes 1,201 3D scenes. For evaluation, we use 9,508 descriptions for 2,068 objects across 141 scenes from ScanRefer and 8,584 descriptions for 1,214 objects in 130 scenes from Nr3D, all sourced from the 312 3D scenes in the ScanNet validation set.

Metrics. We evaluate model performance using four metrics: CIDEr [32], BLEU-4 [25], METEOR [3], and ROUGE-L [13], denoted as **C**, **B-4**, **M**, and **R**, respectively. Following the previous studies [4,9,11,19,33], Non-Maximum Suppression (NMS) is initially applied to filter out duplicate object predictions among the proposals. Each proposal is represented as a pair consisting of a predicted bounding box $\hat{b}i$ and its associated caption $\hat{c}i$. To accurately evaluate the model's capacity for object localization and caption generation, we employ the metric $m@k$, setting the IoU thresholds at 0.25 and 0.5 for our experiments, following [11]:

$$m@k = \frac{1}{N} \sum_{i=1}^{N} m\left(\hat{c}_i, C_i\right) \cdot \mathbb{I}\left\{\mathrm{IoU}\left(\hat{b}_i, b_i\right) \geq k\right\}, \tag{9}$$

where N denotes the total number of annotated objects in the evaluation set, and m stands for the captioning metrics C, B-4, M, and R.

4.2 Implementation Details

Following the [9], our training comprises three stages. We pre-train our network without the caption head on the ScanNet dataset [15] for $1,080$ epochs, using a batch size of 8. We minimize the loss function with an AdamW optimizer [22], starting with a learning rate of 5×10^{-4} that reduces to 10^{-6} using a cosine annealing schedule, along with a weight decay of 0.1 and gradient clipping at 0.1 as per [24]. We then proceed to jointly train the model using standard cross-entropy loss for 720 epochs on both ScanRefer [6] and Nr3D [1], maintaining the detector's learning rate at 10^{-6} and reducing the caption head from 10^{-4} to 10^{-6} to avoid overfitting. In the SCST [30], we adjust the caption head using a batch size of 2 while keeping the detector fixed over a span of 180 epochs and maintain a constant learning rate of 10^{-6}. Additionally, for experiments incorporating 2D data, we use the pre-trained ENet [8] to extract 128-dimensional multi-view features from ScanNet images as outlined in Scan2Cap [11]. The model has 16.9M parameters, and the average inference time on ScanRefer [6] is 1.8ms. All experiments are conducted on a single Titan RTX GPU using PyTorch [26].

4.3 Comparison with Existing Methods

In this section, we evaluate our performance against state-of-the-art methods: Scan2Cap [11], D3Net [7], SpaCap3D [33], MORE [19], 3DJCG [4], Contextual [37], REMAN [23], 3D-VLP [20], χ-Tran2Cap [36], Vote2Cap-DETR [9], Unit3D [12], and Vote2Cap-DETR++ [10]. We apply IoU thresholds of 0.25 and 0.5 for ScanRefer [6] as shown in Table 1 and an IoU threshold of 0.5 for Nr3D [1] indicated in Table 2. For the baselines, we present the evaluation results reported in the original papers, and "-" in Table 1 and Table 2 indicates that such results have not been reported in the original paper.

Model	Training	w/o additional 2D data								w/ additional 2D data							
		IoU=0.25				IoU=0.50				IoU=0.25				IoU=0.50			
		C↑	B-4↑	M↑	R↑	C↑	B-4↑	M↑	R↑	C↑	B-4↑	M↑	R↑	C↑	B-4↑	M↑	R↑
Scan2Cap		53.73	34.25	26.14	54.95	35.20	22.36	21.44	43.57	56.82	34.18	26.29	55.27	39.08	23.32	21.97	44.78
D3Net		-	-	-	-	-	-	-	-	-	-	-	-	46.07	30.29	24.35	51.67
SpaCap3d		58.06	35.30	26.16	55.03	42.76	25.38	22.84	45.66	63.30	36.46	26.71	55.71	44.02	25.26	22.33	45.36
MORE		58.89	35.41	26.36	55.41	38.98	23.01	21.65	44.33	62.91	36.25	26.75	56.33	40.94	22.93	21.66	44.42
3DJCG		60.86	39.67	27.45	59.02	47.68	31.53	24.28	51.80	64.70	40.17	27.66	59.23	49.48	31.03	24.22	50.80
Contextual		-	-	-	-	42.77	23.60	22.05	45.13	-	-	-	-	46.11	25.47	22.64	45.96
REMAN	MLE	-	-	-	-	-	-	-	-	62.01	36.37	27.76	56.25	45.00	26.31	22.67	46.96
3D-VLP		64.09	39.84	27.65	58.78	50.02	31.87	24.53	51.17	70.73	41.03	28.14	59.72	54.94	32.31	24.83	51.51
Vote2Cap-DETR		71.45	39.34	28.25	59.33	61.81	34.46	26.22	54.40	72.79	39.17	28.06	59.23	59.32	32.42	25.28	52.38
Unit3D		-	-	-	-	-	-	-	-	-	-	-	-	46.69	27.22	21.91	45.98
Vote2Cap-DETR++		76.36	41.37	28.70	60.00	67.58	37.05	26.89	55.64	77.03	40.99	28.53	59.59	64.32	34.73	26.04	53.67
BiCA (Ours)		**78.42**	**41.46**	**28.82**	**60.02**	**68.46**	**38.23**	**27.56**	**58.56**	**78.35**	**41.20**	**28.82**	**59.80**	**66.47**	**36.13**	**26.71**	**54.54**
Scan2Cap		-	-	-	-	-	-	-	-	-	-	-	-	48.38	26.09	22.15	44.74
D3Net		-	-	-	-	-	-	-	-	-	-	-	-	62.64	35.68	25.72	53.90
χ-Tran2Cap		58.81	34.17	25.81	54.10	41.52	23.83	21.90	44.97	61.83	35.65	26.61	54.70	43.87	25.05	22.46	45.28
Contextual	SCST	-	-	-	-	50.29	25.64	22.57	44.71	-	-	-	-	54.30	27.24	23.30	45.81
Vote2Cap-DETR		84.15	42.51	28.47	59.26	73.77	38.21	26.64	54.71	86.28	42.64	28.27	59.07	70.63	35.69	25.51	52.28
Vote2Cap-DETR++		88.28	44.07	28.75	59.89	78.16	39.72	26.94	55.52	88.56	43.30	28.64	59.19	74.44	37.18	26.20	53.30
BiCA (Ours)		**89.72**	**44.97**	**28.96**	**60.69**	**80.14**	**40.16**	**27.76**	**56.10**	**89.34**	**44.56**	**28.74**	**59.33**	**76.34**	**37.34**	**26.60**	**54.00**

Table 1: Experimental results on the ScanRefer [6]. C, B-4, M, and R represent the captioning metrics CIDEr [32], BLEU-4 [25], METEOR [3], and ROUGE-L [13], respectively. A higher score for each indicates better performance.

Model	Training	C@0.5 ↑	B-4@0.5 ↑	M@0.5 ↑	R@0.5 ↑
Scan2Cap		27.47	17.24	21.80	49.06
D3Net		33.85	20.70	23.13	53.38
SpaCap3d		33.71	19.92	22.61	50.50
3DJCG		38.06	22.82	23.77	52.99
Contextual	MLE	35.26	20.42	22.77	50.78
REMAN		34.81	20.37	23.01	50.99
Vote2Cap-DETR		43.84	26.68	25.41	54.43
Vote2Cap-DETR++		47.08	27.70	25.44	55.22
BiCA (Ours)		**48.77**	**28.35**	**25.60**	**55.81**
D3Net		38.42	22.22	24.74	54.37
χ-Tran2Cap		33.62	19.29	22.27	50.00
Contextual		37.37	20.96	22.89	51.11
Vote2Cap-DETR	SCST	45.53	26.88	25.43	54.76
Vote2Cap-DETR++		47.62	28.41	25.63	54.77
BiCA (Ours)		**49.81**	**28.83**	**25.85**	**56.46**

Table 2: Experimental results on the Nr3D [1] with IoU threshold at 0.5.

ScanRefer. Descriptions in the ScanRefer include the target object's attributes and spatial relationships with surrounding objects. As Table 1 shows, our method outperforms existing methods in all data settings and IoU thresholds, thanks to our multi-stage contextual attention method.

Nr3D. The Nr3D dataset evaluates the model's proficiency in interpreting human-spoken, free-form object descriptions. Our method shows a notable performance improvement over existing models in generating diverse object descriptions, as indicated in Table 2.

4.4 Ablation Study and Discussion

The core components of our method are i) decomposition of the query set into Instance Query V^o and Context Query V^c, ii) Contextual Attention of Objects

Model	IoU=0.50					
	C↑	B-4↑	M↑	R↑	mAP↑	AR↑
Vote2Cap-DETR [9]	73.77	38.21	26.64	54.71	45.56	67.77
BiCA using only V^o	74.90	40.67	26.76	55.31	50.12	69.49
Vote2Cap-DETR++ [10]	78.16	39.72	26.94	55.52	55.48	70.89
BiCA using V^o, KNN(V^c)	79.03	41.36	27.31	56.76	55.95	69.62
BiCA using V^o, V_a^c	81.22	40.90	27.39	57.95	56.91	70.38
BiCA using V^o, V_a^c, V_a^o (Ours)	**85.14**	**42.27**	**27.98**	**59.37**	**57.58**	**72.68**

Table 3: Ablation study on the ScanRefer [6]. The core components of our BiCA are i) decomposition of the query set into Instance Query V^o and Context Query V^c, ii) the Object-aware Context feature V_a^c, and iii) the Context-aware object feature V_a^o.

for Context (O4C) ii) Contextual Attention of Contexts for Object (C4O). We demonstrate that all components of BiCA contribute positively to the final performance, as shown in Table 3.

Instance Query Generator. We define BiCA using only the Instance Query Generator (i.e., BiCA using only V^o in Table 3) as our baseline and compare it with Vote2Cap-DETR [9], an object-centric transformer encoder-decoder architecture. The major difference between our instance query generator and Vote2Cap-DETR is how we generate the query set for instances. Vote2Cap-DETR uses farthest point sampling (FPS) to generate queries before the query coordinates are adjusted through voting. Therefore, if the coordinates are mistakenly focused on a specific object after voting, features will be extracted from the same object. On the other hand, our instance query generator extracts the features from the candidate coordinates after the voting. This improves the number of matching candidates (e.g., for $2,068$ objects in our evaluation set, our method has $1,540$ matching proposals while Vote2Cap-DETR has $1,498$). This enhancement boosts localization performance in terms of mean Average Precision (mAP) and Average Recall (AR), which directly contributes to improving dense captioning performance.

Context Query Generator. To enable object features to focus on localization, we independently generate context and instance queries separately. In the Vote2Cap-DETR++ [10], the decoupled queries are projections of the object-centric queries and still entail the limitations of the object-centric design. In contrast, our context queries capture structural information from the entire 3D scene, effectively decoupling features from object localization, as shown in Table 3.

O4C module and C4O module. As shown in Table 3, utilizing the object-aware context feature V_a^c and the context-aware object feature V_a^o results in performance improvements across all aspects. Interestingly, we can see that performance improves even when surrounding context information is collected using KNN (See the results of BiCA using only V^o, KNN(V^c)). In the setting of BiCA

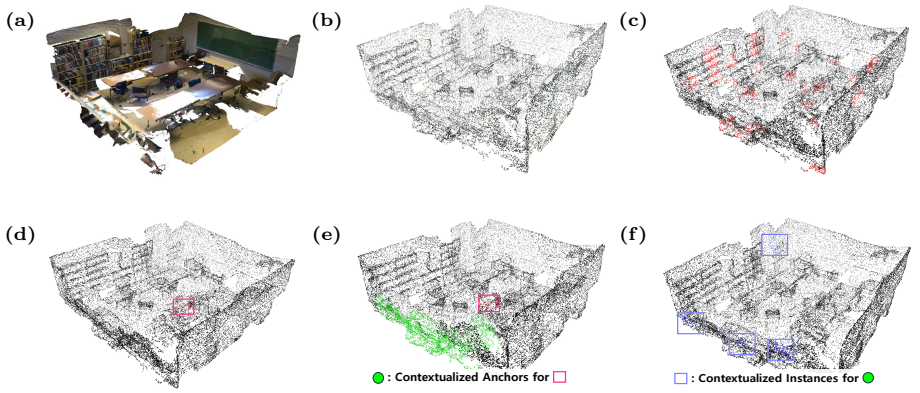

Fig. 3: Visualization of (a) the input 3D scene, (b) input point cloud without color, (c) visualization of instance queries, (d) one sampled instance query, (e) geometry contextualization of (d), and (f) instance contextualization of (e) on the ScanRefer [6].

using only V^o, KNN(V^c), we set $K = 16$. Additionally, aggregating contextualized object features improves the model performance significantly. It proves that when capturing context information for captioning, adding additional object information that matches the context is helpful.

4.5 Qualitative Analysis

Figure 3 shows the parts of the entire scene being focused on at each stage of the proposed model. In Figure 3c and Figure 3d, our instance queries are not concentrated on a specific object but are spread throughout the 3D scene and exist in the location where the object exists. Figure 3e shows the part of the context queries with the highest attention score. The green part of the Figure 3e indicates the surroundings of the object and the corners of the room further away, but it needs to be related to other instances or scenes to express the situation more clearly. The blue part of Figure 3f describes the part of related instance queries for the object-aware context feature selected in Figure 3e. Most instances are inside the green contextualized region with high scores of Figure 3e, and some instances on the other side are also included. This observation shows that our method can check all instances of the entire scene to generate captions for one object.

To demonstrate the effectiveness of the proposed method, we provide qualitative results with state-of-the-art models. SpaCap3D [33], 3DJCG [4], Vote2Cap-DETR [9], and Vote2Cap-DETR++ [10] have attempted to incorporate contextual information by modeling relationships with object pairs or aggregating nearest neighbor features. In the Figure 4, these methods generate captions limited to the object and its immediate relations in a fixed format (e.g., this is a white radiator. it is under right of the desk). Since their context is not collected based

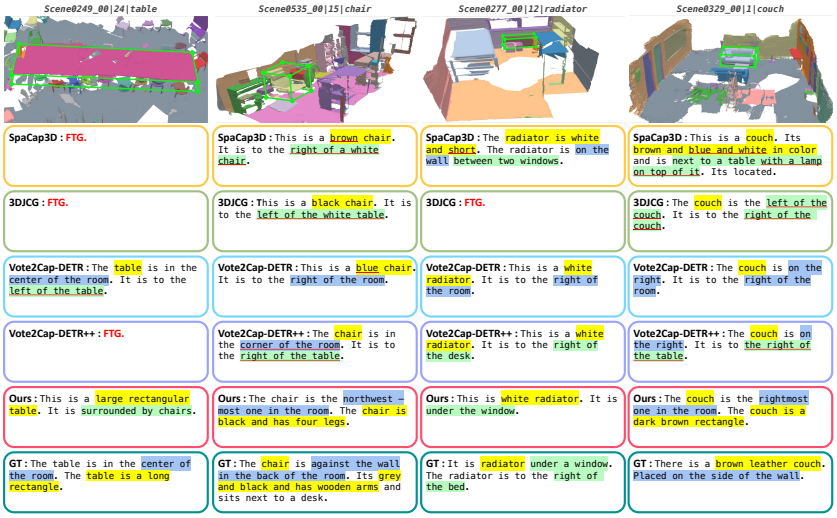

Fig. 4: Qualitative results on the ScanRefer [6]. The yellow-highlighted sections show information specific to the object itself, the green-highlighted sections describe the relationships between objects, and the blue-highlighted sections depict the spatial position of the object in the 3D scene. Captions underlined in red indicate incorrect descriptions. **FTG.** represent failures in caption generation due to low IoU.

on the entire scene, it is challenging to localize relationships with surrounding objects or locations in the scene. BiCA accurately predicts the attributes (i.e., can obtain precise representations for regions that have localization boundaries) and captions that require a sufficient understanding of structural information throughout the scene (i.e., can retrieve geometrically substantial information within the scene that exceeds the boundary of localization constraints) in various formats (e.g., the chair is the northwest – most one in the room.).

5 Conclusion

In this work, we propose BiCA, a novel end-to-end transformer encoder-decoder pipeline with multi-stage contextual attention for 3D dense captioning. Our BiCA parallelly decodes a fixed set of object queries that contain local features of individual objects and context queries which contain the non-object contexts in the 3D scene. This allows our model to capture relevant object-aware context and context-aware object features across the entire scene without being restricted to single object localization or their immediate surroundings. As the representation for localization and caption generation is disentangled, BiCA can improve both localization and contextual dense captioning performance. We validate the effectiveness of BiCA by showing that it outperforms the state-of-the-art across all metrics on two benchmarks for 3D dense captioning.

Acknowledgement

This work was supported by LG AI Research and Institute of Information & Communications Technology Planning & Evaluation (IITP) grant (No. RS-2019-II191082, No. RS-2022-II220156) funded by the Korea government (MSIT).

References

1. Achlioptas, P., Abdelreheem, A., Xia, F., Elhoseiny, M., Guibas, L.: 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. pp. 422–440. Springer (2020)
2. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6077–6086 (2018)
3. Banerjee, S., Lavie, A.: 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. pp. 65–72 (2005)
4. Cai, D., Zhao, L., Zhang, J., Sheng, L., Xu, D.: 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. pp. 16464–16473 (2022)
5. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
6. Chen, D.Z., Chang, A.X., Nießner, M.: Scanrefer: 3d object localization in rgb-d scans using natural language. In: European conference on computer vision. pp. 202–221. Springer (2020)
7. Chen, D.Z., Wu, Q., Nießner, M., Chang, A.X.: D3net: A unified speaker-listener architecture for 3d dense captioning and visual grounding. In: European Conference on Computer Vision. pp. 487–505. Springer (2022)
8. Chen, J., Lei, B., Song, Q., Ying, H., Chen, D.Z., Wu, J.: A hierarchical graph network for 3d object detection on point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 392–401 (2020)
9. Chen, S., Zhu, H., Chen, X., Lei, Y., Yu, G., Chen, T.: End-to-end 3d dense captioning with vote2cap-detr. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11124–11133 (2023)
10. Chen, S., Zhu, H., Li, M., Chen, X., Guo, P., Lei, Y., Gang, Y., Li, T., Chen, T.: Vote2cap-detr++: Decoupling localization and describing for end-to-end 3d dense captioning. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
11. Chen, Z., Gholami, A., Nießner, M., Chang, A.X.: Scan2cap: Context-aware dense captioning in rgb-d scans. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3193–3203 (2021)
12. Chen, Z., Hu, R., Chen, X., Nießner, M., Chang, A.X.: Unit3d: A unified transformer for 3d dense captioning and visual grounding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18109–18119 (2023)

13. Chin-Yew, L.: Rouge: A package for automatic evaluation of summaries. In: Proceedings of the Workshop on Text Summarization Branches Out, 2004 (2004)
14. Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R.: Meshed-memory transformer for image captioning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10578–10587 (2020)
15. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
16. Ghandi, T., Pourreza, H., Mahyar, H.: Deep learning approaches on image captioning: A review. arXiv preprint arXiv:2201.12944 (2022)
17. Hossain, M.Z., Sohel, F., Shiratuddin, M.F., Laga, H.: A comprehensive survey of deep learning for image captioning. ACM Computing Surveys (CsUR) **51**(6), 1–36 (2019)
18. Jiang, L., Zhao, H., Shi, S., Liu, S., Fu, C.W., Jia, J.: Pointgroup: Dual-set point grouping for 3d instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and Pattern recognition. pp. 4867–4876 (2020)
19. Jiao, Y., Chen, S., Jie, Z., Chen, J., Ma, L., Jiang, Y.G.: More: Multi-order relation mining for dense captioning in 3d scenes. In: European Conference on Computer Vision. pp. 528–545. Springer (2022)
20. Jin, Z., Hayat, M., Yang, Y., Guo, Y., Lei, Y.: Context-aware alignment and mutual masking for 3d-language pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10984–10994 (2023)
21. Kuhn, H.W.: The hungarian method for the assignment problem. Naval research logistics quarterly **2**(1-2), 83–97 (1955)
22. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
23. Mao, A., Yang, Z., Chen, W., Yi, R., Liu, Y.j.: Complete 3d relationships extraction modality alignment network for 3d dense captioning. IEEE Transactions on Visualization and Computer Graphics (2023)
24. Misra, I., Girdhar, R., Joulin, A.: An end-to-end transformer model for 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2906–2917 (2021)
25. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
26. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems 32, pp. 8024–8035. Curran Associates, Inc. (2019), http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
27. Qi, C.R., Litany, O., He, K., Guibas, L.J.: Deep hough voting for 3d object detection in point clouds. In: proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9277–9286 (2019)
28. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems **30** (2017)
29. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)

30. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7008–7024 (2017)
31. Tancik, M., Srinivasan, P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J., Ng, R.: Fourier features let networks learn high frequency functions in low dimensional domains. Advances in Neural Information Processing Systems **33**, 7537–7547 (2020)
32. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4566–4575 (2015)
33. Wang, H., Zhang, C., Yu, J., Cai, W.: Spatiality-guided transformer for 3d dense captioning on point clouds. arXiv preprint arXiv:2204.10688 (2022)
34. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning. pp. 2048–2057. PMLR (2015)
35. Yang, X., Tang, K., Zhang, H., Cai, J.: Auto-encoding scene graphs for image captioning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10685–10694 (2019)
36. Yuan, Z., Yan, X., Liao, Y., Guo, Y., Li, G., Cui, S., Li, Z.: X-trans2cap: Cross-modal knowledge transfer using transformer for 3d dense captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8563–8573 (2022)
37. Zhong, Y., Xu, L., Luo, J., Ma, L.: Contextual modeling for 3d dense captioning on point clouds. arXiv preprint arXiv:2210.03925 (2022)