See It All: Rethinking 3D Dense Captioning with Caption-Centric Attention

Anonymous ACL submission

Abstract

3D dense captioning is a task to localize objects in a 3D scene and generating descriptive sentences for each object. Recent approach in 3D dense captioning have adopted transformer encoder-decoder frameworks from object detection to build an end-to-end pipeline without hand-crafted components. However, these approaches struggle with the object-centric design, since the query attention has to view both the object regions and its contextual environment versatilely. This leads to a trade-off between object localization accuracy and context sensitivity. To overcome this challenge, we introduce a new pipeline that engages to 3D dense captioning with a caption-centric approach called SIA (See-It-All). Our transformer encoder-decoder architecture simultaneously decodes two sets of queries—context query and instance query—where the context query directly captures the region-of-interest per each caption and the instance query focuses on localization and attribute descriptions. We design a novel aggregator to generate captions based on the conTextual surroundings, and Global environment, and object Instances (i.e., TGI-Aggregator) to generate a fully informed contextual caption. Extensive experiments on two widely-used datasets, ScanRefer and Nr3D, demonstrates that our proposed method achieves a significant improvement over prior methods.

1 Introduction

3D dense captioning has been defined in former works (Chen et al., 2021, 2022; Yuan et al., 2022; Wang et al., 2022; Jiao et al., 2022; Cai et al., 2022; Chen et al., 2023a) as the task to localize all the objects in a 3D scene (i.e., object detection) and generate descriptive sentences for each object (i.e., object caption generation). Early works incorporated a two-stage "detect-then-describe" pipeline, where we first detect all the object proposals then generate the captions for each objects (Chen et al., 2021; Jiao et al., 2022; Wang et al., 2022; Zhong et al., 2022; Cai et al., 2022; Chen et al., 2022; Yuan et al., 2022). While being intuitive, the sequential design and hand-crafted components of these works have posed as a serious bottleneck in terms of both performance and efficiency. Succeeding work (Chen et al., 2023a) have emulated the transformer encoder-decoder pipeline from object detection (Carion et al., 2020) to alleviate these issues and fashioned an end-to-end pipeline.

This attention-based architecture (Chen et al., 2023a) introduces several key innovations to address 3D dense captioning: i) it conceptualizes the 3D dense captioning task as a set prediction challenge, where it directly translates features into sets of *objects*, pinpointing their locations and relevant captions through parallel prediction heads. ii) the vote decoder reformulates the object queries in 3DETR (Misra et al., 2021) into the format of the vote query, establishing the connection with VoteNet (Qi et al., 2019) while improving localization accuracy and training efficiency.

However, this object-centric design, where each query is linked to a specific object localization, is not compatible with dense captioning. As shown in Figure 1, dense captioning involves generating captions that either independently describe an object's attributes or describe the object within its contextual environment. Object-centric 3D dense captioning necessitates that the query's attention sufficiently covers both the object and its environment. This creates a challenge: focusing attention on the object enhances localization and detailed attribute description but reduces sensitivity to the surrounding context. Conversely, spreading attention to include the context can improve understanding of the environmental description but at the cost of localization accuracy. This behaves as a major hurdle towards optimal performance.

In this paper, we reexamine the role of transformer attention on the realm of dense captions.

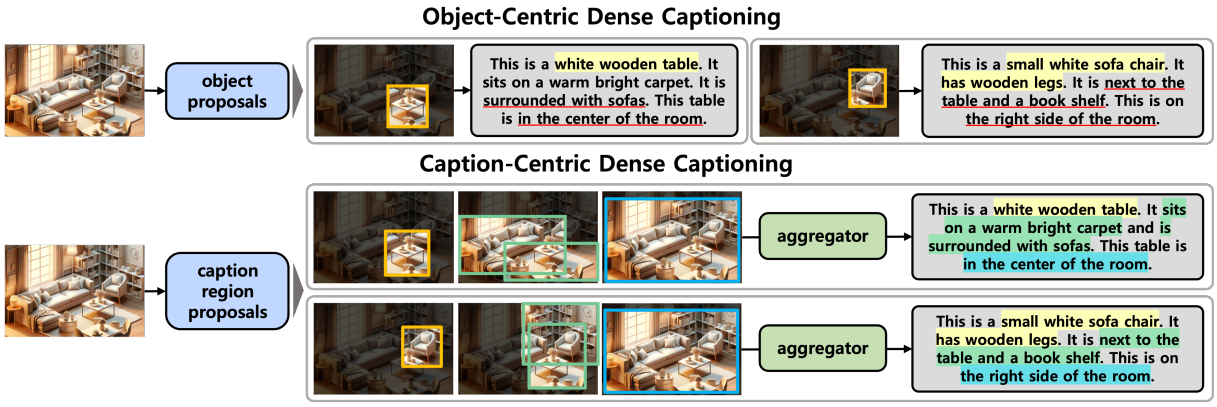

Figure 1: Conceptual illustration of object-centric and caption-centric dense captioning. Unlike object-centric methods that generate captions from a single feature and struggle to capture various contexts within one query, such as those highlighted by red underlines, our caption-centric approach addresses this issue by attending to and aggregating proposed caption regions, effectively focusing on local-global regions while generating captions.

To address the above issues, we propose a novel pipeline engaging *caption-centric* attention called **SIA (i.e., See It All)**. Instead of designating a query *per object* then integrating the contextual scene afterwards, our transformer encoder-decoder architecture leverages a separate set of context query that are decoded to capture the context region *per caption*. The two queries are parallely decoded and fed to our novel aggregator named TGI-Aggregator to generate captions based on the con**T**extual surroundings, **G**lobal descriptor, and **I**nstance features. While the fully informed feature from TGI-Aggregator generates a contextual caption, each instance feature generates an attribute caption for individual objects along with its localization. The two caption results are afterwards concatenated to consist the final caption prediction. This disentangled inference structure enables our model to more effectively focus on the region-of-interest when generating each caption.

Extensive experiments on two benchmarks, ScanRefer (Chen et al., 2020a) and Nr3D (Achlioptas et al., 2020), show that our method surpasses prior approaches by a large margin.

2 Preliminary

In this preliminary, we start with a basic transformer-based end-to-end 3D dense captioning pipeline (Chen et al., 2023a). The caption heads are attached on top of the existing 3D object detection pipeline (Qi et al., 2019; Misra et al., 2021) with vote queries that establish captions for each pinpointed object throughout the scene in an object-by-object manner. Afterward, we discuss why this object-centric application of transformer attention is not suitable for 3D dense captioning.

2.1 End-to-End 3D Object Detection

3D object detection aims to identify and localize objects in 3D scenes. VoteNet (Qi et al., 2019) incorporates an encoder-decoder architecture where the bounding boxes are predicted by aggregating the *votes* for the center coordinates. 3DETR (Misra et al., 2021) generates *object queries* by uniformly sampling seed points from a 3D scene. Vote2Cap-DETR (Chen et al., 2023a) uses *vote queries* that connects the object queries in 3DETR to VoteNet, resulting in better localization and improved training efficiencies.

2.2 Extension to 3D Dense Captioning

The goal of 3D dense captioning is to localize objects in a 3D scene and generate informative natural language descriptions per object. An intuitive extension from object detection to 3D dense captioning is to simply apply a captioning head for each object proposals (Chen et al., 2023a). Given an input indoor 3D scene as a point cloud $PC = [p_{\text{in}}; f_{\text{in}}] \in \mathbb{R}^{N\times(3+F)}$, where $p_{\text{in}} \in \mathbb{R}^{N\times 3}$ is the absolute locations for each point and $f_{\text{in}} \in \mathbb{R}^{N\times F}$ is additional input features for each point (Chen et al., 2020a, 2021), the objective of 3D dense captioning is to generate a set of box-caption pairs $(\hat{B}, \hat{C}) = \{(\hat{b_1}, \hat{c_1}), ..., (\hat{b_K}, \hat{c_K})\}$, representing an estimation of K distinctive objects in this 3D scene. Captions are generated in *parallel* with bounding box prediction using a caption head. Since the

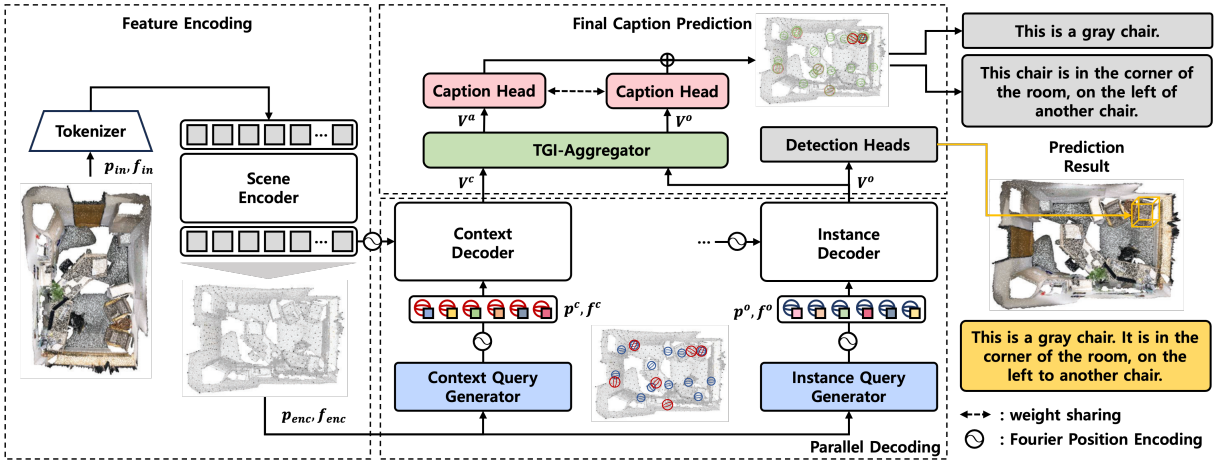

Figure 2: Overall architecture of the transformer encoder-decoder 3D dense captioning pipeline that capitalizes on our proposed caption-centric attention.

aforementioned vote queries (i.e., p_{vote}) fail to provide adequate attributes and spatial relations for informative caption generation, the contextual information is leveraged through a separate lightweight transformer (Chen et al., 2023a).

2.3 Retrospect on Object-Centric Captioning

Current 3D dense captioning benchmarks requires the model to generate multiple captions for each detected object. Therefore, it seems natural to approach this task in an *object-centric* manner, where we generate captions per each object proposal. However, unlike object detection, dense captioning requires an extensive understanding of the scene including the attributes of each object and the relative information between objects and the global scene. Therefore, designating the queries per object requires a single query attention to versatilely encompass both the individual object and its surrounding elements, failing to effectively concentrate on the local element it should describe. To address this issue and successfully incorporate contextual scene information, we propose a novel *caption-centric* approach for 3D dense captioning.

3 Method

In this section, we introduce our *caption-centric* transformer encoder-decoder pipeline for 3D dense captioning. Unlike previous approaches where the object proposals are fed to the caption head to generate captions *per-object*, we establish context queries that capture the context *per-caption* then integrate the relevant objects. The overall pipeline is illustrated in Figure 2.

3.1 Encoder

Following previous work (Chen et al., 2023a), our system also adopts 3DETR (Misra et al., 2021) encoder as our scene encoder. Given the input point cloud as $PC = [p_{\text{in}}; f_{\text{in}}] \in \mathbb{R}^{N\times(3+F)}$, the point cloud is first tokenized by a set-abstraction layer of PointNet++ (Qi et al., 2017). The tokenized output is inputted into a masked transformer encoder with a set-abstraction layer, which is then followed by two additional encoder layers. The final encoded scene tokens are denoted as $[p_{\text{enc}}; f_{\text{enc}}] \in \mathbb{R}^{1,024\times(3+256)}$.

3.2 Context Query and Instance Query

To disentangle the multiple captions that are bind to a single object, we designate a separate *context query* that captures a unique region per caption within the 3D scene. While the context query captures the local-global regions that are capable of a caption, the *instance query* generates standard object localization and attribute-related caption prediction for each object. The two queries are decoded in parallel and afterwards aggregated to consist the final caption.

Context Query. Given the encoded scene tokens $[p_{\text{enc}}; f_{\text{enc}}]$, we sample 512 context points p^c_{seed} with farthest point sampling (FPS) on p_{enc}. Then, the context query (p^c, f^c) is represented as:

$$(p^c, f^c) = \left(p^c_{\text{seed}}, \text{SA}_c(p_{\text{enc}}, f_{\text{enc}})\right), \qquad (1)$$

where SA_c denotes the set-abstraction layer (Qi et al., 2017) with a radius of 1.2 and samples 64 points for p^c.

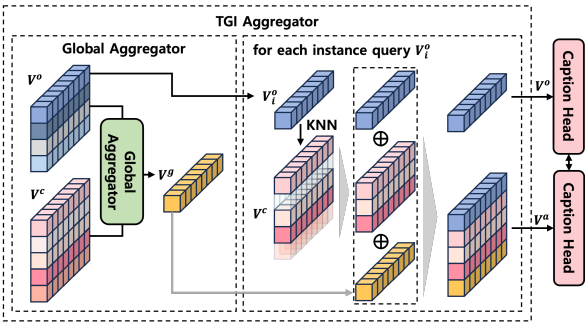

Figure 3: Conceptual illustration of our TGI-Aggregator. The Global Aggregator $G(\cdot)$ aggregates the decoded context query V^o and instance query V^c to construct a global descriptor V^g. Then, the instance feature V_i^o, the nearest neighbor feature in V^c, and the global descriptor V^g are concatenated to construct V^a.

Instance Query. As the context query captures the contextual information for each caption, the instance query is decoded to perform standard 3D object detection and generate captions for the individual attributes of each object. Likewise, the instance query (p^o, f^o) is written as:

$$(p^o, f^o) = \left(p^o_{\text{seed}} + \Delta p^o_{\text{vote}}, \text{SA}_o(p_{\text{enc}}, f_{\text{enc}})\right), \quad (2)$$

where $\Delta p^o_{\text{vote}} = \text{FFN}_o(f_{\text{enc}})$ is an offset that learns to shift the encoded points to object's centers spatially by a feed-forward network FFN_o, following (Chen et al., 2023a). SA_o denotes the set-abstraction layer with a radius of 0.3 and samples 16 points for p^o.

3.3 Caption-Centric Decoding

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 regions per caption, and the localization and attribute description for the participating objects is performed by the Instance Decoder. Given the decoded context query V^c and instance query V^o, we first construct a global descriptor V^g and feed it to our TGI-Aggregator. While the decoded instance query V^o is fed to the caption head to generate final dense captions for i) the attributes of the object itself, the output of the TGI-Aggregator V^a is fed to the caption head to generate ii) the relationship with surrounding objects, and iii) the relationship within the global scene.

TGI-Aggregator. To leverage sufficient contextual information and spatial relations for informative captions, we generate a global feature that encompasses the scene in whole. Fig. 3 shows a conceptual illustration of our TGI-Aggregator. Given the all decoded context queries V^c and instance queries V^o, we deploy a clustering based aggregator (Arandjelovic et al., 2016) $G(\cdot)$. As a result, we obtain a global descriptor $V^g \in \mathbb{R}^{256}$ by $V^g = G(V^c, V^o)$. Then, we concantenate this global descriptor V^g to each decoded instance query V_i^o and its K nearest feature within V^c, resulting with an aggregated feature V^a that contains a comprehensive information of con**T**ext, **G**lobal, and **I**nstance. We set K to 16.

Contextual Captions. Given the output of our TGI-Aggregator V^a, we generate captions with a fully informed feature including features for contextual surroundings, global context, and instance information. Based on this feature, SIA generates descriptions for relationships between multiple objects (e.g., the chair is sitting next to a book shelf) or global relationships (e.g., the table sits in the middle of the room).

Localization & Attribute Captions. To predict localization and attribute descriptions for the participating instances, we feed the decoded instance query V^v and parallely feed it into the detection head and (shared) caption head. For localization, we follow 3DETR (Misra et al., 2021), reformulating the box corner estimation as offset estimation from a query point to an object's center, and box size regression. All subtasks are implemented by FFNs. The object localization head is shared throughout the decoder layers, following the implementation details of DETR (Carion et al., 2020).

3.4 Training SIA

In this section, we include details of how to train our end-to-end transformer encoder-decoder pipeline with caption-centric attention, SIA (i.e., See-it-All).

Instance Query Loss. Identical to previous work (Chen et al., 2023a), we borrow the vote loss from VoteNet (Qi et al., 2019) to help the Instance Query Generator learn to shift points p_{enc} to an object's center. 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 written as:

$$\mathcal{L}^o = \frac{1}{M} \sum_{i=1}^{M} \sum_{j=1}^{N_{gt}} \|p_i^o - \text{cnt}_j\|_1 \cdot \mathbb{I}(p_{enc}^i), \quad (3)$$

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 a 3D scene, M is the size of p^o, and cnt_j is the center of j-th instance I_j.

Detection Loss. Following the implementation details of previous works (Misra et al., 2021; Chen et al., 2023a), we use the same Hungarian algorithm to assign each proposal with a ground-truth label. The detection loss $\mathcal{L}_{\text{det}}$ is written as:

$$\mathcal{L}_{\text{det}} = \alpha_1 \mathcal{L}_{\text{giou}} + \alpha_2 \mathcal{L}_{\text{cls}} + \alpha_3 \mathcal{L}_{\text{cnt}} + \alpha_4 \mathcal{L}_{\text{size}}, \quad (4)$$

where $\alpha_1 = 10, \alpha_2 = 1, \alpha_3 = 5, \alpha_4 = 1$ are set heuristically (Chen et al., 2023a). The detection loss is applied across all decoder layers for better convergence.

Caption Loss. Following the standard protocol for image captioning, we train our caption head with standard cross-entropy loss (MLE training) first, then fine-tune it with Self-Critical Sequence Training (SCST) (Rennie et al., 2017). In the MLE training, the model learns to predict 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 is established for a sentence with length T is defined as follows:

$$\mathcal{L}_{c_n} = \sum_{t=1}^{T} \mathcal{L}_{c_n}(t) = -\sum_{t=1}^{T} \log \hat{P}\left(c_n^{t+1} | \mathcal{V}, c_n^{[1:t]}\right), \quad (5)$$

Once the caption head is trained with word-level supervision, it is further refined using SCST. In this phase, the model produces multiple captions $\hat{c}_{1,\ldots,k}$ with a beam size of k, and an additional $\hat{g}$ using greedy search as a baseline. The loss function for SCST is formulated as follows:

$$\mathcal{L}_{c_n} = -\sum_{i=1}^{k} (R(\hat{c}_i) - R(\hat{g})) \cdot \frac{1}{|\hat{c}_i|} \log \hat{P}(\hat{c}_i | \mathcal{V}). \quad (6)$$

The reward function $R(\cdot)$ is based on the CIDEr (Vedantam et al., 2015) metric for caption evaluation, and the logarithmic probability of the caption $\hat{c}_i$ is normalized by caption length $|\hat{c}_i|$, promoting equal importance to captions of varying lengths by the model.

Final Loss for SIA. Given the voting loss for instance query $\mathcal{L}^o$, the detection loss for the instances 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 $\mathcal{L}$ for SIA is written as:

$$\mathcal{L} = \beta_1 \mathcal{L}^o + \beta_2 \sum_{i=1}^{n_{\text{dec-layer}}} \mathcal{L}^i_{\text{det}} + \beta_3 \mathcal{L}_{\text{cap}}, \quad (7)$$

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

4 Experiments

4.1 Datasets and Metrics

Datasets. In our studies, we focus on 3D dense captioning and employ two established datasets: ScanRefer (Chen et al., 2020a) and Nr3D (Achlioptas et al., 2020). These datasets are rich in human-generated descriptions, with ScanRefer providing $36,665$ descriptions for $7,875$ objects across 562 scenes, and Nr3D offering $32,919$ descriptions for $4,664$ objects in 511 scenes. For the purpose of training, these descriptions and objects are derived from the ScanNet (Dai et al.) database, which comprises $1,201$ 3D scenes. For evaluation, we use $9,508$ descriptions from ScanRefer and $8,584$ from Nr3D, corresponding to $2,068$ and $1,214$ objects across 141 and 130 scenes, respectively, from the 312 3D scenes in the ScanNet validation set.

Metrics. We evaluate the model using four types of performance metrics: CIDEr (Vedantam et al., 2015), BLEU-4 (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and ROUGE-L (Chin-Yew, 2004). Following the previous studies (Chen et al., 2021; Cai et al., 2022; Jiao et al., 2022; Wang et al., 2022; Chen et al., 2023a), we first employ Non-Maximum Suppression (NMS) to eliminate redundant object predictions from the object proposals. Each proposal is represented as a pair consisting of a predicted bounding box $\hat{b}_i$ and a generated caption $\hat{c}_i$. To assess both the model's ability to accurately locate objects and generate captions, we employ the $m@k$IoU, setting the IoU threshold k at 0.25 and 0.5 for our experiments, following (Chen et al., 2021):

$$m@kIoU = \frac{1}{N} \sum_{i=1}^{N} m\left(\hat{c}_i, C_i\right) \cdot \mathbb{I}\left\{\text{IoU}\left(\hat{b}_i, b_i\right) \geq k\right\}, \quad (8)$$

where N is the number of all annotated instances in the test set, and m represents the captioning metrics CIDEr, BLEU-4, METEOR, and ROUGE-L.

4.2 Implementation Details

Our training phase is structured into three stages, following the approach of (Chen et al., 2023a). Initially, we pre-train our network excluding the caption head on the ScanNet (Dai et al.) dataset for $1,080$ epochs. The batch size is 8. The loss function is minimized using AdamW optimizer (Loshchilov and Hutter, 2017), for which the initial learning rate is 5×10^{-4} and decreasing to 10^{-6} according to a

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	MLE	-	-	-	-	42.77	23.60	22.05	45.13	-	-	-	-	46.11	25.47	22.64	45.96
REMAN		-	-	-	-	-	-	-	-	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
Ours		**78.68**	**43.25**	**29.21**	**63.06**	**73.22**	**40.91**	**28.19**	**60.46**	**78.05**	**42.16**	**28.74**	**61.70**	**69.86**	**37.89**	**27.04**	**57.33**
Scan2Cap		-	-	-	-	-	-	-	-	-	-	-	-	48.38	26.09	22.15	44.74
D3Net		-	-	-	-	-	-	-	-	-	-	-	-	62.64	35.68	25.72	53.90
χ-Tran2Cap	SCST	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		-	-	-	-	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
Ours		**89.72**	**44.56**	**28.96**	**62.13**	**83.14**	**42.17**	**27.92**	**59.44**	**89.71**	**45.31**	**29.06**	**62.11**	**79.84**	**40.84**	**27.28**	**57.54**

Table 1: Experimental results on the ScanRefer (Chen et al., 2020a) dataset.

cosine annealing schedule. We also apply a weight decay of 0.1 and a gradient clipping of 0.1, as suggested by (Misra et al., 2021). Afterward, starting from the pre-trained weights, we jointly train the entire model with the standard cross-entropy loss for an additional 720 epochs on the ScanRefer (Chen et al., 2020a) and Nr3D (Achlioptas et al., 2020) datasets, fixing the detector's learning rate at 10^{-6} and reducing the caption head's from 10^{-4} to 10^{-6} to prevent overfitting (about 20/17 hours for ScanRefer/Nr3D). In the SCST (Rennie et al., 2017) phase, 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} (about 22/18 hours for ScanRefer/Nr3D). In the experimental setup that use additional 2D data, as shown in Table 1, we employ the pre-trained ENet (Chen et al., 2020b) to extract 128-dimensional multiview features from 2D ScanNet images, as in the (Chen et al., 2021). All experiments of our SIA are conducted with one Titan RTX GPU on PyTorch (Paszke et al., 2019). More details for implementation can be checked through the code to be released.

4.3 Comparison with Existing Methods

In this part of our study, we benchmark our performance against the state-of-the-art methods using the metrics CIDEr (Vedantam et al., 2015), METEOR (Banerjee and Lavie, 2005), BLEU-4 (Papineni et al., 2002), and ROUGE-L (Chin-Yew, 2004), denoted as **C**, **B-4**, **M**, and **R**, respectively. We apply IoU thresholds of 0.25 and 0.5 for ScanRefer (Chen et al., 2020a) as shown in Table 1 and an IoU threshold of 0.5 for Nr3D (Achlioptas et al., 2020) indicated in Table 3.

We compare our approach with eleven 3D dense captioning methods: Scan2Cap (Chen et al., 2021), D3Net (Chen et al., 2022), SpaCap3D (Wang et al., 2022), MORE (Jiao et al., 2022), 3DJCG (Cai et al., 2022), Contextual (Zhong et al., 2022), REMAN (Mao et al., 2023), 3D-VLP (Jin et al., 2023), χ-Tran2Cap (Yuan et al., 2022), Vote2Cap-DETR (Chen et al., 2023a), and Unit3D (Chen et al., 2023b). For the baselines, we present the evaluation results reported in the original papers, and "-" in Table 1 and Table 3 means that such results have not reported in the original paper or follow-up study.

ScanRefer. Table 1 summaries the results on the ScanRefer dataset. Our method significantly surpasses current methods in 3D dense captioning across all input data settings and IoU threshold configurations. We attribute this enhanced performance to our caption-centric attention mechanism. Unlike previous object-centric methods where surrounding information is binded to the center of object proposals, SIA directly targets the region-of-interest for contextual captions with the context query, while simultaneously addressing object localization and its attribute captions with the instance query.

Nr3D. Nr3D dataset was designed to evaluate how well the model performs when interpreting natural language descriptions of objects spoken by humans. Our SIA quantitatively shows state-of-the-art performance consistently across all evaluation measures (see Table 3) while being able to generate a wide variety of descriptions for an object (see Figure 4).

Model	Training	IoU=0.25						IoU=0.50					
		C↑	B-4↑	M↑	R↑	mAP↑	AR↑	C↑	B-4↑	M↑	R↑	mAP↑	AR↑
Vote2Cap-DETR	MLE	71.45	39.34	28.25	59.33	65.05	87.40	61.81	34.46	26.22	54.40	45.56	67.77
SIA with V^o	MLE	74.37	41.55	28.17	60.49	65.66	87.08	64.75	36.76	26.23	55.63	46.15	65.85
SIA with V^o, V^c	MLE	77.24	42.64	28.27	59.89	69.26	88.04	**73.38**	37.92	26.14	54.74	48.74	67.04
SIA with V^o, V^c, V^a (Ours)	MLE	**78.68**	**43.25**	**29.21**	**63.06**	**72.20**	**88.09**	73.22	**40.91**	**28.19**	**60.46**	**49.69**	**69.08**
Vote2Cap-DETR	SCST	84.15	42.51	28.47	59.26	65.05	87.40	73.77	38.21	26.64	54.71	45.56	67.77
SIA with V^o	SCST	85.10	44.06	28.30	60.44	65.66	87.08	73.80	39.06	26.33	55.50	46.15	65.85
SIA with V^o, V^c	SCST	89.34	43.30	28.70	60.33	69.26	88.04	81.45	41.19	26.22	55.34	48.74	67.04
SIA with V^o, V^c, V^a (Ours)	SCST	**89.72**	**44.56**	**28.96**	**62.13**	**72.20**	**88.09**	**83.14**	**42.17**	**27.92**	**59.44**	**49.69**	**69.08**

Table 2: Ablation study on the ScanRefer (Chen et al., 2020a) dataset. The core components of our model: instance query feature V^o, context query feature V^c, and TGI-Aggregator feature V^a, respectively.

Model	Training	C@0.5 ↑	B-4@0.5 ↑	M@0.5 ↑	R@0.5 ↑
Scan2Cap	MLE	27.47	17.24	21.80	49.06
D3Net	MLE	33.85	20.70	23.13	53.38
SpaCap3d	MLE	33.71	19.92	22.61	50.50
3DJCG	MLE	38.06	22.82	23.77	52.99
Contextual	MLE	35.26	20.42	22.77	50.78
REMAN	MLE	34.81	20.37	23.01	50.99
Vote2Cap-DETR	MLE	43.84	26.68	25.41	54.43
Ours	MLE	**56.39**	**30.87**	**27.54**	**60.36**
D3Net	SCST	38.42	22.22	24.74	54.37
χ-Tran2Cap	SCST	33.62	19.29	22.27	50.00
Contextual	SCST	37.37	20.96	22.89	51.11
Vote2Cap-DETR	SCST	45.53	26.88	25.43	54.76
Ours	SCST	**59.48**	**32.60**	**27.99**	**61.08**

Table 3: Experimental results on the Nr3D (Achlioptas et al., 2020) dataset.

Method	C@0.5 ↑	B-4@0.5 ↑	M@0.5 ↑	R@0.5 ↑
Contexts V^c	72.82	38.46	26.76	56.71
Single Instance V_i^o & Contexts V^c	72.89	38.21	27.04	57.33
Instances V^o & Contexts V^c	**73.22**	**40.91**	**28.19**	**60.46**

Table 4: Ablation for how the context features V^c and instance feature V^o participate in the Global Aggregator $G(\cdot)$ on the ScanRefer (Chen et al., 2020a) dataset. $\cdot_i$ denotes a single i-th decoded query feature.

4.4 Ablation Study and Discussion

In our ablation study, we validate that each and every component of our proposed SIA positively contributes to the final performance. The core components of our model are the Instance Decoder that generates instance features V^o, Context Decoder that generates contextual features V^c, and the TGI-Aggregator that generates a fully informed aggregated feature V^a. In Table 2, we set our baseline performance with Vote2Cap-DETR (Chen et al., 2023a), an object-centric transformer encoder-decoder architecture for comparison. It can be observed that the performance improves for both the parallel Context Decoder and the TGI-Aggregator.

Localization Performance. In previous object-centric approaches, better contextual caption generation came at a cost of degraded localization. To validate whether our caption-centric approach alleviates this issue, we report the localization score (i.e., mAP and AR score) for IoU threshold 0.25 and 0.50 in Table 2. The localization score for the MLE training (top rows) and SCST (Rennie et al., 2017) training (bottom rows) is identical since the detector is fixed during the SCST training phase. While using only the instance query V^o results in performance comparable to Vote2Cap-DETR, adding the context features (i.e., V^c and V^a) leads to an increase in mAP and AR, with gains of 6.54/3.54 and 1.01/3.23 at IoU thresholds of 0.25 and 0.50, respectively. We conjecture that the disentangling the attentions for local objects and contextual information successfully contributed to the final performance.

Global Feature Construction. Table 4 shows our ablation study on how we aggregate the decoded context query V^c and instance query V^o to construct the global feature V^g that is afterwards fed to our TGI-Aggregator (recall Figure 3). The scenarios include i) aggregating all context features V^c, ii) gathering one instance feature V_i^o with all context features V^c, and iii) aggregating all instance and context features V^o and V^c to extract a global feature. It can be observed that aggregating all instance and context features to create a global feature results in the best performance. This implies that it is better to reflect all instances and contexts when representing the entire scene.

4.5 Qualitative Analysis

To demonstrate the effectiveness of the proposed method, we provide qualitative results with the state-of-the-art models: Scan2Cap (Chen et al., 2021), SpaCap3D (Wang et al., 2022), D3Net (Chen et al., 2022), 3DJCG (Cai et al., 2022), and Vote2Cap-DETR (Chen et al., 2023a). As shown in

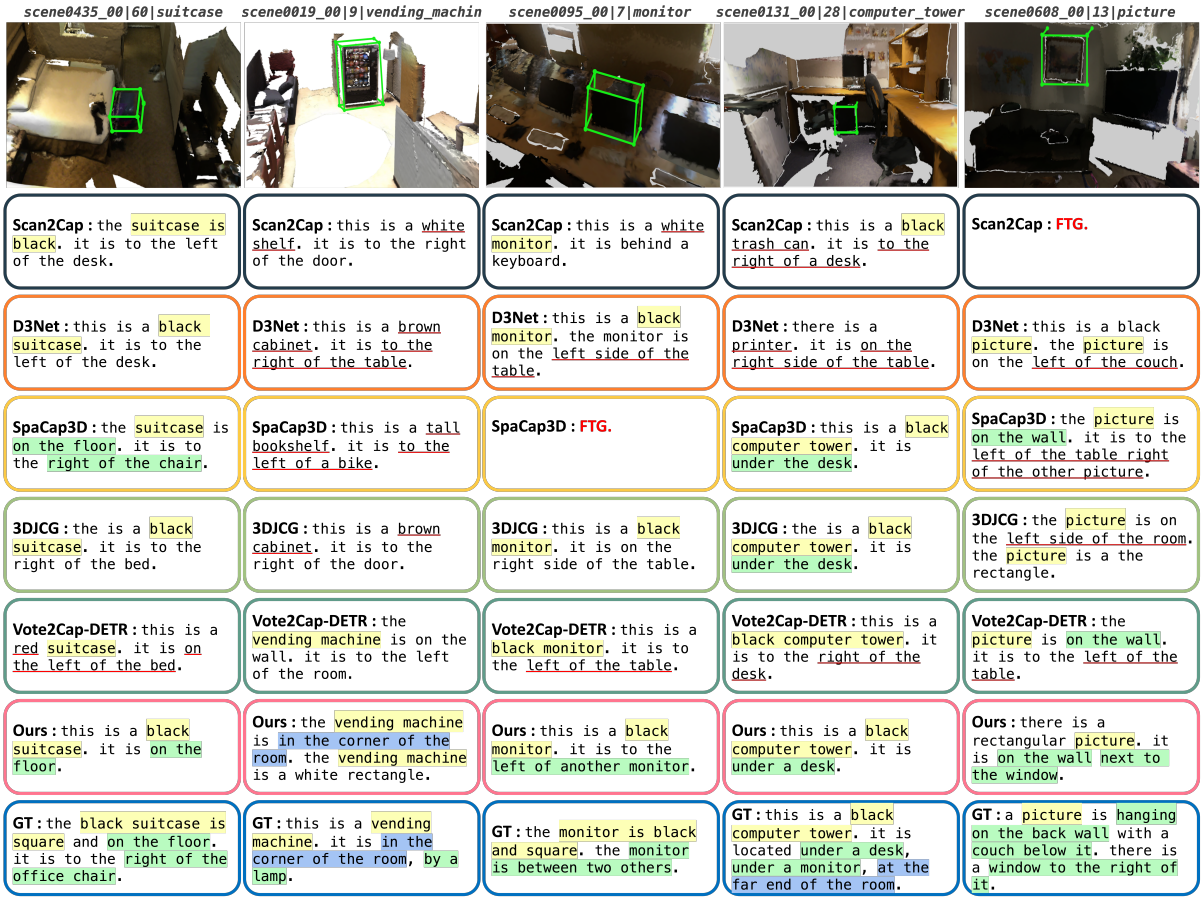

Figure 4: Qualitative results on the ScanRefer (Chen et al., 2020a). The yellow-highlighted sections show information specific to the object itself, the green-highlighted sections describes the relationships between objects, and the blue-highlighted sections depict the spatial position of the object in the 3D scene. Captions underlined in red indicate incorrect descriptions. **FTG.** represent failures in caption generation due to low IoU.

Figure 4, given a 3D scene as input, the model localizes objects and generates captions that describe the objects, which are then evaluated for performance with the ground truth for each object. The ground-truth includes descriptions of the intrinsic properties of the object (e.g., the monitor is black and square.), explanations using the relationships between objects (e.g., a picture is hanging on the back wall with a couch below it.), and descriptions of the object in the context of the entire space (e.g., at the far end of the room.). Existing models, focusing on objects, generate captions limited to the object and its immediate relations in a fixed format (e.g., this is a black suitcase.). Our model can handle not only object queries but also context queries, allowing it to generate sentences in various formats (e.g., the vending machine is a white rectangle.) and create descriptions of the entire space (e.g., the vending machine is in the corner of the room.). These results emphasize the importance of an integrated understanding of the object, its surroundings, and the overall space in captioning.

5 Conclusion

In the 3D dense captioning task, the description of an object in a 3D scene includes not only the characteristics of the object itself, but also the relationship with surrounding objects and the positional relationship of the object with respect to the entire space. In this work, we propose a novel approach that generates instance and context queries independently, and aggregates these features to handle a global description of the entire space. Extensive experiments on benchmark datasets demonstrate that our method significantly improves 3D dense captioning compared to previous approaches, thereby proving the importance of an integrated understanding of the object, its surroundings, and the entire space for caption generation.

6 Limitations

We propose SIA, a 3D dense captioning method that shifts from previous object-centric approaches to a caption-centric approach. However, it still poses the limitation of previous set prediction architectures, where we have to heuristically set the number of each instance query and context query. Future works might include methods where the number of query for instances and contexts can be dynamically adjusted based on how busy the scene is. Moreover, a 3D scene encompasses numerous objects, and although our model is capable of generating a variety of captions by leveraging spatial relationships and interactions between objects, it faces a limitation owing to the limited number of provided ground-truth captions. This limitation makes it difficult to assess every potential scenario comprehensively. Addressing this challenge is an important future direction in the field of captioning.

7 Ethical Considerations

Our proposed method is for 3D dense captioning, and does not raise any ethical concerns. All experiments in this study were carried out using publicly available datasets. Our captioning models are trained on datasets of 3D scenes with English dense captions.

References

Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. 2020. 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.

Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. 2016. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5297–5307.

Satanjeev Banerjee and Alon Lavie. 2005. 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*, pages 65–72.

Daigang Cai, Lichen Zhao, Jing Zhang, Lu Sheng, and Dong Xu. 2022. 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.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer.

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

Dave Zhenyu Chen, Qirui Wu, Matthias Nießner, and Angel X Chang. 2022. D3net: A unified speaker-listener architecture for 3d dense captioning and visual grounding. In *European Conference on Computer Vision*, pages 487–505. Springer.

Jintai Chen, Biwen Lei, Qingyu Song, Haochao Ying, Danny Z Chen, and Jian Wu. 2020b. 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)*, pages 392–401.

Sijin Chen, Hongyuan Zhu, Xin Chen, Yinjie Lei, Gang Yu, and Tao Chen. 2023a. End-to-end 3d dense captioning with vote2cap-detr. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11124–11133.

Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. 2021. Scan2cap: Context-aware dense captioning in rgb-d scans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3193–3203.

Zhenyu Chen, Ronghang Hu, Xinlei Chen, Matthias Nießner, and Angel X Chang. 2023b. Unit3d: A unified transformer for 3d dense captioning and visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18109–18119.

Lin Chin-Yew. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out, 2004*.

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/CVF Conference on Computer Vision and Pattern Recognition*.

Yang Jiao, Shaoxiang Chen, Zequn Jie, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. 2022. More: Multi-order relation mining for dense captioning in 3d scenes. In *European Conference on Computer Vision*, pages 528–545. Springer.

Zhao Jin, Munawar Hayat, Yuwei Yang, Yulan Guo, and Yinjie Lei. 2023. Context-aware alignment and

mutual masking for 3d-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10984–10994.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Aihua Mao, Zhi Yang, Wanxin Chen, Ran Yi, and Yongjin Liu. 2023. Complete 3d relationships extraction modality alignment network for 3d dense captioning. *IEEE Transactions on Visualization and Computer Graphics*.

Ishan Misra, Rohit Girdhar, and Armand Joulin. 2021. An end-to-end transformer model for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2906–2917.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

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

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

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7008–7024.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4566–4575.

Heng Wang, Chaoyi Zhang, Jianhui Yu, and Weidong Cai. 2022. Spatiality-guided transformer for 3d dense captioning on point clouds. *arXiv preprint arXiv:2204.10688*.

Zhihao Yuan, Xu Yan, Yinghong Liao, Yao Guo, Guanbin Li, Shuguang Cui, and Zhen Li. 2022. 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*, pages 8563–8573.

Yufeng Zhong, Long Xu, Jiebo Luo, and Lin Ma. 2022. Contextual modeling for 3d dense captioning on point clouds. *arXiv preprint arXiv:2210.03925*.