MiKASA: Multi-Key-Anchor & Scene-Aware Transformer for 3D Visual Grounding

Chun-Peng Chang, Shaoxiang Wang, Alain Pagani, Didier Stricker
DFKI Augmented Vision

Abstract

3D visual grounding involves matching natural language descriptions with their corresponding objects in 3D spaces. Existing methods often face challenges with accuracy in object recognition and struggle in interpreting complex linguistic queries, particularly with descriptions that involve multiple anchors or are view-dependent. In response, we present the MiKASA (Multi-Key-Anchor Scene-Aware) Transformer. Our novel end-to-end trained model integrates a self-attention-based scene-aware object encoder and an original multi-key-anchor technique, enhancing object recognition accuracy and the understanding of spatial relationships. Furthermore, MiKASA improves the explainability of decision-making, facilitating error diagnosis. Our model achieves the highest overall accuracy in the Referit3D challenge for both the Sr3D and Nr3D datasets, particularly excelling by a large margin in categories that require viewpoint-dependent descriptions. The source code and additional resources for this project are available on GitHub: https://github.com/birdy666/MiKASA-3DVG

1. Introduction

3D visual grounding serves as a crucial component in the intersection of natural language processing and computer vision. This task aims to identify and localize objects within a 3D space, using linguistic cues for spatial and semantic grounding. While existing research has made significant strides, challenges remain. Key issues include the lack of explainability in current models, limitations in object recognition within point cloud data, and the complexity of handling intricate spatial relationships.

Most existing 3D visual grounding models [2, 15, 16, 33, 41] consist of three parts: (1) object encoder, (2) text encoder, and (3) fusion model. The object encoder processes the provided point cloud and generates features in the embedding space. However, because the points in a point cloud are unordered and inconsistent in sparsity [26], it is not straightforward to apply the methodology typically used for 2D images. An additional challenge is that 3D point cloud datasets are not as extensive as those for 2D images [10, 25], which makes it difficult for the models to correctly recognize object categories. While enlarging the dataset could conceivably improve performance, we refrain from doing so to ensure a fair comparison with existing state-of-the-art methods. Existing works [6, 18, 27] mainly use different techniques such as noise addition, dropping out colors, and transformations to expand the sample space. Though these techniques may increase the stability of the produced object embeddings, the improvement is limited.

(a) Target category: "chair"

(b) "The chair in the front of the blue-lit monitor."

Figure 1. Our methodology utilizes a dual-prediction framework for 3D visual grounding. First, we assign a target category score based on object categorization, as detailed in Fig. 1a. Next, a spatial score is integrated according to the object's alignment with the textual description, as shown in Fig. 1b.

Inspired by previous works [20, 21, 39] which aims to solve object recognition problem, we leverage the fact that data availability on objects within a specific space can provide valuable insights into the characteristics and relationships of their surrounding entities. For instance, when we come across a cuboid-shaped object in a kitchen, we may naturally assume that it is a dishwasher. Conversely, if we spot the same shape in a bathroom, it is more plausible that it is a washing machine. Contextual information is crucial in determining the true identity of an object and gives us a nuanced understanding of our surroundings.

Therefore, by incorporating a scene-aware object encoder that considers all nearby objects, we demonstrate that the model's object classification accuracy improves before the data is fed into the fusion model.

Another important task is to represent the spatial relations between objects. Previous works [15, 16] have primarily focused on encoding the absolute locations of objects in world coordinates, a method we find suboptimal. Inspired by how humans use objects as anchors for spatial reasoning and thereby shift the perceptual focus to these anchor objects, our approach hypothesizes that encoding spatial relations relative to anchors enhances the model's ability to accurately identify objects. This hypothesis has been empirically validated in our experiments.

Therefore, we introduce the multi-key-anchor concept to enhance spatial understanding in 3D models. This approach translates the coordinates of potential anchors relative to a target object and explicitly evaluates the importance of nearby objects based on textual descriptions. Notably, models like PointNet++ [26] are designed for rotational invariance, often leading to directional ambiguity in object features. Our method mitigates this issue by leveraging the spatial context of key nearby objects, thereby implicitly suggesting the orientation of the target object. For example, a chair is typically placed facing a table or against a wall, and the presence of the table or the wall implicitly defines the direction of the chair. This approach provides a nuanced understanding of object orientation and spatial placement. We validate this claim experimentally, by showing a higher improvement in view-dependent cases compared to the overall accuracy.

Finally, instead of treating the information as input for a black-box fusion module, we introduce a new architecture that employs a more novel approach to the 3D visual grounding task. Inspired by human-like object searching behavior—for instance, when given the instruction "The chair in the front of the blue-lit monitor.", one would first identify all the chairs in the room before pinpointing the specific target, as illustrated in Fig. 1. Our model employs late fusion and generates two distinct output scores. The first score aims to identify the target object category, while the second assesses the location and language expression, informed by spatial data. These scores are designed to collaborate, mitigating the influence of objects that may superficially resemble the target or occupy positions that seem to fit the verbal description but are not the intended objects. We fuse these scores through a strategic fusion mechanism, which enables the final target to distinguish itself more clearly from distractors. By examining the two scores and the final result, one can better diagnose the types of errors the model may make, making the decision process more explainable.

The main contributions of our work can be summarized as follows:

- We introduce a scene-aware object encoder that considers the contextual information and increases models ability to understand the object category.
- We present the multi-key-anchor technique, which enhances spatial understanding. This approach redefines coordinates relative to target objects and explicitly assesses the importance of nearby objects through textual context. It addresses the directional ambiguity often found in rotationally invariant models like PointNet++ [26], by using spatial contexts to imply target object orientation.
- We develop a novel, end-to-end trainable and explainable architecture, that leverages late fusion to separately process distinct aspects of the data, thereby enhancing the model's accuracy and explainability.

2. Related Works

2.1. 3D Visual Grounding

3D visual grounding, intersecting computer vision and natural language processing, focuses on identifying objects in 3D spaces using language. Unlike 2D grounding that relies on images, 3D visual grounding utilizes point cloud data. This transition from 2D models to point clouds introduces a new layer of complexity, and marks a distinct shift from conventional 2D grounding models [17, 23, 35] due to the unique nature of point clouds [10]. Our experiments and comparative analyses are conducted using the Referit3D benchmark [2].

The methods for feature fusion for 3D visual grounding have recently evolved. Initially, graph-based algorithms were predominantly used [2, 14]. However, with the rise of transformer models, the focus has shifted towards these, given their effectiveness in multimodal data fusion [1, 12, 15, 16, 33, 41]. Notable among these are LAR [4], which synthesizes 2D clues from point clouds for 3D visual grounding, SAT [33] that leverages 2D image semantics in training, and 3DVG-Transformer [41], using a relation-aware approach with contextual clues for proposal generation. To enhance point cloud data representation, MVT [15] maps 3D scenes into multi-view spaces, aggregating positional information from various perspectives. ViewRefer [11] builds upon this by resolving view discrepancies through the integration of multi-view inputs and inter-view attention. Furthermore, ViewRefer utilizes GPT-3 [5] to generate multiple geometry-consistent descriptions from a single grounding text, thereby enriching the model's interpretation of 3D environments.

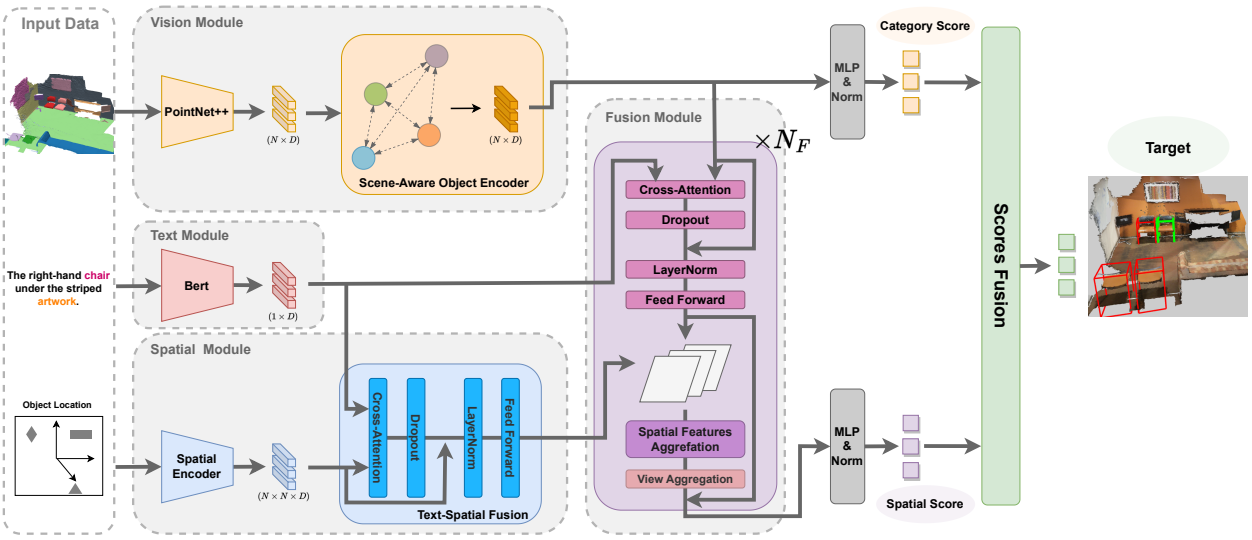


Figure 2. Architecture of our 3D Visual Grounding Model, which includes four main modules: a text encoder (Bert), a vision module with a scene-aware object encoder, a spatial module that fuses spatial and textual data, and a multi-layered fusion module. The fusion module combines text, spatial, and object features, employing a dual-scoring system for enhanced object category identification and spatial-language assessment.

2.2. Context-Aware Object Recognition

3D object recognition and segmentation using point clouds is a foundational task in computer vision. Most established methodologies [25, 26] predominantly utilize appearance features such as color and position to define objects, achieving impressive outcomes. However, in contexts where a comprehensive scene is involved, such as a room filled with diverse objects, methods that incorporate inter-object relationships and spatial context [9, 20, 21, 28, 34, 39] offer better performance. These models do not just consider individual object features, but also the relative placements and attributes of neighboring objects. This allows for a more sophisticated understanding of complex 3D spaces.

2.3. Multi-Modal Features Fusion

In multi-modal features fusion, two primary fusion approaches exist: early and late. Early fusion merges features from different modalities at the outset and trains a unified model, as commonly seen in 3D visual grounding work [2, 15, 16, 33, 41]. This approach benefits from direct inter-modality interactions but can be challenging to fine-tune and lacks transparency regarding its decision-making process. Late fusion, on the other hand, processes each modality separately and fuses the resulting logits or decision scores. Various techniques, from simple averaging to attention-based methods, are used in existing works [3,22,24,32,37,38]. The fusion strategy significantly impacts the system's robustness and accuracy, especially when modalities provide conflicting cues.

3. Method

Fig. 2 presents our novel architecture for the 3D visual grounding task, including key modules: a vision module, a text encoder (Bert [8]), a spatial module, and a fusion module. We maintain configurations for Bert and PointNet++ [26] as in MVT [15] for consistent comparison. The vision module, with our scene-aware object encoder, refines object features by considering surrounding objects. The spatial module encodes and merges spatial features with textual data from Bert using a transformer encoder. In the fusion module, comprising N_F layers, object-spatial features are progressively refined with text and spatial information, enhancing spatial features through refined anchor information. Our model concludes with a dual-scoring system, generating scores for object category identification and spatial feature assessment. This approach mitigates the influence of distractors and enhances the explainability of the model's decisions.

3.1. Data Augmentation

Given the limitations and scarcity of 3D point cloud datasets, data augmentation emerges as a crucial strategy. To improve the model against overfitting and enhance its ability to generalize essential features, we've adopted different data augmentation techniques. This includes the notable multi-view augmentation as presented in MVT [15], which has demonstrated effectiveness. Additionally, we place emphasis on augmenting color features, adjusting contrast, and introducing noise. More details can be found in the supplementary materials.

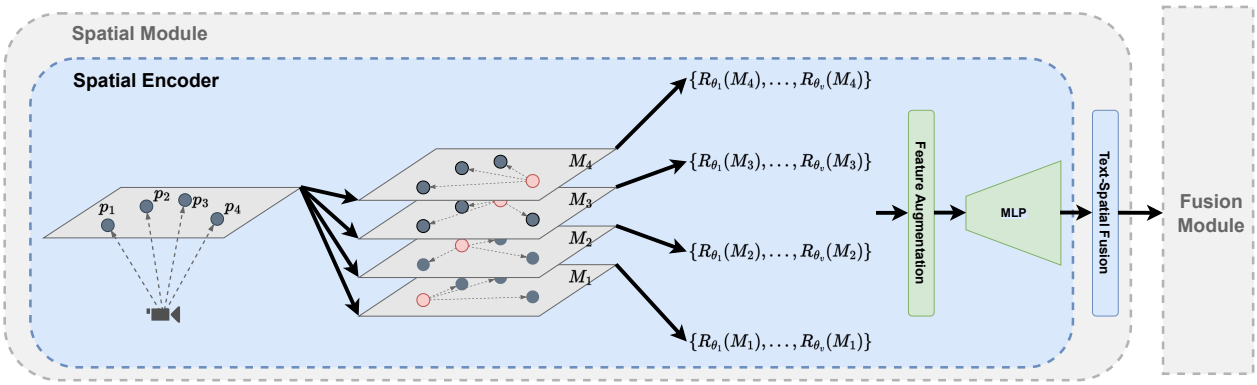


Figure 3. Our novel spatial module captures relative spatial information from a single viewpoint by treating each object in the scene as a potential anchor. This approach generates unique spatial maps, each offering a different perspective of the scene. These maps are then undergo feature augmentation, where distances and angles are calculated, followed by normalization and scaling. Subsequently, a MLP layer is employed to transform these low-dimensional features into higher-dimensional ones for effective fusion with textual data.

3.2. Scene-Aware Object Encoder

Traditional point cloud methods usually analyze objects separately, missing contextual clues from other objects in the same room. To overcome this, our model includes a scene-aware object encoder that gathers information from surrounding objects. This approach overcomes the limits of standard object encoders. By considering this additional context, our model better understands object categories, leading to improved accuracy and performance.

While graph-based algorithms such as DGCNN [31], GCN [19], and GAT [30] are often used for similar tasks, they present their own set of challenges, such as computational intensity and the complexity of defining a suitable distance metric for determining neighbors. As a result, we opted for a self-attention mechanism for feature aggregation. This choice offers a range of advantages, including computational efficiency, ease of training, and consistently superior performance, making it the most fitting solution for the objectives of our model. For N objects in the room, represented as $O = \{O_1, O_2, \ldots, O_N\}$ where each O_i is a feature vector with D dimensions, the scene-aware object features O_i^{sa} are computed using self-attention as defined in Eq. (1). In this process, $Q = W_Q \cdot O$, $K = W_K \cdot O$, and $V = W_V \cdot O$ represent the queries, keys, and values, respectively, each transformed by their respective learnable weight matrices W_Q, W_K, and W_V [29]. By aggregating this information, particularly through the weighted sum of values (V), each object feature O_i^{sa} becomes enriched with contextual data from its surroundings. This approach ensures that the object features capture more than just individual properties.

$$O_i^{sa} = \sum_{j=1}^{N} \frac{\exp(Q_i \cdot K_j)}{\sum_{k=1}^{N} \exp(Q_i \cdot K_k)} \cdot V_j \tag{1}$$

3.3. Spatial Features Encoding

Instead of relying on a single viewpoint for understanding a 3D space, our model employs a novel multi-anchor strategy to better comprehend spatial relationships, as shown in Fig. 3. For N objects in the room, this approach involves generating N feature maps, each representing relative positions from N unique local coordinate systems. We define a set of spatial maps for a given set of object coordinates, $P = \{p_1, p_2, \ldots, p_N\}$, as $M = \{M_1, M_2, \ldots, M_N\}$. Each map M_i, defined in Eq. (2), is composed of the relative positions of all other objects $A = \{a_j | a_j \in P \text{ and } j \neq i\}$ to a target object p_i.

$$M_i = \{(a_j - p_i) | a_j \in A\} \tag{2}$$

To enhance robustness against varying initial viewpoints, our model incorporates the viewpoint augmentation strategy similar to MVT [15]. We utilize a rotation matrix R_{θ_k} for each viewpoint, which rotates the entire map by an angle θ_k. This approach results in an augmented set M^{aug}, defined in Eq. (3), wherein each map M_i from the original set M is represented under v different rotated views.

$$M^{aug} = \{R_{\theta_k}(M_i) \mid M_i \in M,\ k \in \{1, 2, \ldots, v\}\} \tag{3}$$

Each element of a map then goes through feature augmentation to incorporate additional spatial features, including distance and angles. These augmented features are later on normalized and scaled to ensure stability. Normalization and scaling details are in the supplementary materials.

Subsequently, the map features undergo a dimensional transformation $\mathcal{T}$ to align with the dimension D, as described in Eq. (4). This process prepares the map features for effective fusion with other features at later stages of the model.

$$M^D = \mathcal{T}(M^{aug}, D) \tag{4}$$

3.4. Multi-Modal Feature Fusion

To create a multi-anchor spatial map, we combine features from various modalities. Specifically, we merge the object feature, spatial feature, and the text feature. This integrative approach generates a new detailed spatial feature that accurately represents an object's location within a room.

3.4.1 Text-Spatial Fusion

To optimize computational efficiency, we merge text and spatial features at an initial stage instead of in the multi-layer fusion module. We use a single cross-attention layer, represented by $\mathcal{A}$, followed by a subsequent feedforward layer, denoted as $\mathcal{F}$. This approach is employed instead of merging these features within the fusion module alongside additional features. With N spatial maps at our disposal, this method provides a computationally economical means of feature integration. The fusion of the spatial map M with the textual information T synchronizes spatial features with corresponding text data, as expressed in Eq. (5):

$$M_i^{text} = \mathcal{F}(\mathcal{A}(M_i, T)),\ \forall i \in \{1, 2, \ldots, N\} \qquad (5)$$

3.4.2 Fusion Module

Our fusion module's architecture consists of four key components in each layer: (1) Text-Object Fusion, (2) Object-Spatial Fusion, (3) Spatial Feature Aggregation, and (4) View Aggregation. As data progresses through these layers, the model incrementally adjusts anchor weightings and extracts relevant spatial information based on textual input. This progressive fusion method refines features within specific spatial maps while dynamically updating anchor features across the maps.

The effectiveness of this approach is particularly notable in complex scenarios. Take, for example, the instruction "Choose the suitcase that is in front of the bed near the window curtains." Here, the primary anchor ("bed") is contextualized by another reference point ("curtains"), underscoring the need for coherent interaction between different spatial maps. This scenario highlights the value of our fusion process in managing intricate spatial relations.

Text-Object fusion & Object-Spatial Fusion:
The Text-Object Fusion component employs an architectural approach akin to that of the Text-Spatial Fusion in Eq. (5), incorporating both a cross-attention layer and a feedforward neural network layer. Subsequently, the text-object features are integrated into each spatial map by addition and linear transformation. The augmented map now encapsulates not only the spatial information but also the characteristics of the anchor objects.

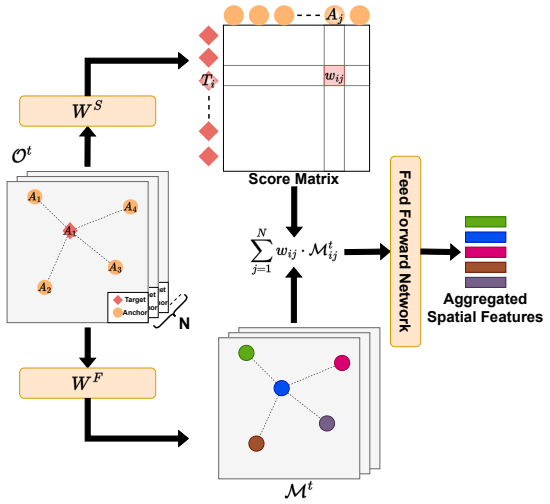


Figure 4. Our novel attention-based spatial feature aggregation. Each map designates a different object as the target, while treating all other objects as anchors. The importance of each anchor relative to the potential target object is represented in row i of the score matrix, indicating the relevance of each anchor in the context of the target, where W_S and W_F are learnable weight matrices.

Spatial Feature Aggregation & View Aggregation:
In our model, essential information is gathered from the fused feature maps through an attention-based aggregation stage, as depicted in Fig. 4. This stage employs attention mechanisms to determine the relevance of potential key anchors relative to the target object. The process effectively consolidates these weighted anchors into a single spatial feature for each object, enhancing the model's grasp of the spatial relationships of each object within the scene. Subsequently, to ensure robustness against the initial viewpoint, the features undergo further refinement by aggregating views from various perspectives, a method elaborated in Section 3.1.

Progressive Feature Enhancement:
The fusion module's effectiveness hinges on the iterative refinement of object-spatial features $\mathcal{O}^t$, as detailed in Eq. (6). In this process, w_j^i denotes the significance of anchor j within the fused feature map $\mathcal{M}_i^t$, as illustrated in Fig. 4. This approach integrates aggregated spatial features into object features, facilitating progressive enhancement across spatial maps. As a result, both anchor and spatial features undergo continuous refinement with each layer, progressively improving the model's ability to accurately represent the spatial dynamics of the scene.

$$\mathcal{O}_i^{t+1} = \mathcal{O}_i^t + \sum_{j=1}^{N} w_{ij} \cdot \mathcal{M}_{ij}^t, \quad \forall i \in \{1, 2, \ldots, N\} \qquad (6)$$

3.5. Multi-Modal Predictions Fusion

Previous works [2, 15, 33] treat the fusion process as a "black box", where information is aggregated in the early stages and then passed to a fusion module such as a transformer decoder to produce a single set of logits. While this method is effective, it offers limited insight into the intricacies of the fusion process. In contrast, our model adopts a segmented approach to the grounding problem, tackling it through two distinct subtasks. The first subtask determines the target category score by evaluating the object's correspondence with the text-described target category. Subsequently, a spatial score is computed to assess how well the object's spatial configuration aligns with the spatial description provided in the text.

In our methodology, directly merging logits, $f_1(X;\theta_1)$ and $f_2(X;\theta_2)$ from different modalities can lead to suboptimal results due to discrepancies in their scales. To mitigate this, we employ a normalization function, g to transform each logit to have zero mean and unit variance, facilitating a uniform scale. Subsequently, we assign weights, denoted as λ and μ, which are optimized as hyperparameters through experimentation, detailed in the supplementary materials. These weights are applied to the normalized logits, enabling a balanced integration of the modalities. The resulting final prediction $\mathcal{P}$ is represented as in Eq. (7).

$$\mathcal{P} = \lambda \cdot g(f_1(X;\theta_1)) + \mu \cdot g(f_2(X;\theta_2)) \tag{7}$$

3.6. Loss Function

The loss function $\mathcal{L}$ employed in our model is a composite of multiple terms, each aimed at a specific aspect of the 3D visual grounding task. Formally, the loss is defined as Eq. (8).

$$\mathcal{L} = L_{\text{ref}} + \alpha L_{\text{text}} + \beta L_{\text{obj}} + \gamma L_{\text{obj\_scene}} \tag{8}$$

where all terms are computed using cross-entropy loss. L_{ref} is the primary loss that evaluates the reference results, measuring how accurately the model identifies the correct target among distractors. The auxiliary losses serve to fine-tune different components of the model. Specifically, L_{text} evaluates the model's ability to correctly categorize the target object from a given sentence (e.g., recognizing "table" as the target category in the sentence "a table near the window"). L_{obj} evaluates the categorization of all objects in the 3D space. $L_{\text{obj\_scene}}$, which we might alternatively name "Scene-Aware Object Categorization Loss", assesses the categorization performance but does so after the scene-aware object encoder has been applied. The hyperparameters α, β, and γ control the relative contributions of these auxiliary losses and the choice of their value is discussed in the supplementary material.

4. Experiment

4.1. Datasets

Natural Reference in 3D (Nr3D) [2] :
The dataset consists of 41.5k human utterances describing 707 unique 3D indoor scenes from ScanNet [7]. It contains 76 fine-grained object classes. These utterances were collected during a reference game played by two individuals. One person acted as the speaker, selecting an object from a set of distractors (ranging from 1 to 6), while the other person identified the target object based on the speaker's instructions. Each scenario in the data can be categorized as easy/hard and view-dependent/view-independent, depending on the number of distractors and whether the utterances depend on a specific viewpoint.

Spatial Reference in 3D (Sr3D/Sr3D+) [2] :
Similar to Nr3D, Sr3D dataset contains 83.5k synthetic utterances describing the 3D indoor scenes from ScanNet [7]. Each utterance was generated from the template

$$target\_class + spatial\_relation + anchor\_class(es)$$

Sr3D+ expand the dataset by sampling that did not contain more than one distractor were added to Sr3D . This resulted in an increase to 114.5k total utterances.

4.2. Experimental Setup

4.2.1 Implementation Details

In the implementation of our proposed architecture, we utilize a pre-trained BERT [8] model as the text encoder, generating a 768-dimensional output. The object encoder is implemented using the PointNet++ [26] framework. To ensure a fair comparison with the MVT [15], the settings for both the text and object encoders are aligned with those used in MVT. Our fusion module is composed of three layers. Importantly, all modules—including the text encoder, object encoder, and fusion module—are trained end-to-end, negating the need to train each component separately. The optimization is carried out using the Adam optimizer with a batch size of 12. All experiments were conducted on an A100 GPU.

4.2.2 Evaluation Metrics

In our experiments, we focus on the datasets from Referit3D, namely Nr3D and Sr3D. Evaluation proposals are generated directly from the ground truth annotations. The primary metric for evaluation is accuracy, which gauges the model's ability to successfully identify the correct target among various distractors. A successful match is defined as the model accurately pointing out the designated target from a pool of distractors in the 3D space.

Method \Dataset	Sr3D					Nr3D				
	Overall	Easy	Hard	VD*	VI†	Overall	Easy	Hard	VD*	VI†
ReferIt3D [2] *ECCV 20*	40.8%	44.7%	31.5%	39.2%	40.8%	35.6%	43.6%	27.9%	32.5%	37.1%
InstanceRefer [36] *ICCV 21*	48.0%	51.1%	40.5%	45.4%	48.1%	38.8%	46.0%	31.8%	34.5%	41.9%
3DVG-Transf. [41] *ICCV 21*	51.4%	54.2%	44.9%	44.6%	51.7%	40.8%	48.5%	34.8%	34.8%	43.7%
SAT [33] *ICCV 21*	57.9%	61.2%	50.0%	49.2%	58.3%	49.2%	56.3%	42.4%	46.9%	50.4%
MVT [15] *CVPR 22*	64.5%	66.9%	58.8%	58.4%	64.7%	55.1%	61.3%	49.1%	54.3%	55.4%
BUTD-DETR [16] *ECCV 22*	67.0%	68.6%	63.2%	53.0%	67.6%	54.6%	60.7%	48.4%	46.0%	58.0%
NS3D [13] *CVPR 23*	62.7%	64.0%	59.6%	62.0%	62.7%	-	-	-	-	-
M3DRef [40] *ICCV 23*	-	-	-	-	-	49.4%	55.6%	43.4%	42.3%	52.9%
ViewRefer [11] *ICCV 23*	67.0%	68.9%	62.1%	52.2%	67.7%	56.0%	63.0%	49.7%	55.1%	56.8%
Ours	**75.2%**	**78.6%**	67.3%	**70.4%**	**75.4%**	**64.4%**	**69.7%**	**59.4%**	**65.4%**	**64.0%**
vs. BUTD-DETR [16] *ECCV 22*	+8.2%	+10.0%	+4.1%	**+17.4%**	+7.8%	+9.8%	+9.0%	+11.0%	**+19.4%**	+6.0%
vs. Ns3D [13] *CVPR 23*	+12.5%	+14.6%	+7.7%	**+8.4%**	+12.7%	-	-	-	-	-
vs. M3DRef [40] *ICCV 23*	-	-	-	-	-	+15.0%	+14.1%	+16.0%	**+23.1%**	+11.1%
vs. ViewRefer [11] *ICCV 23*	+8.2%	+9.7%	+5.2%	**+18.2%**	+7.7%	+8.4%	+6.7%	+9.7%	**+10.3%**	+7.2%

Table 1. Comparative accuracy on Sr3D and Nr3D Challenges. This table showcases the performance of various models across all subcategories in the Sr3D and Nr3D datasets, highlighting the MiKASA Transformer's enhancements. The increments in performance achieved by MiKASA over previous methods are detailed for each subcategory, underlining its superior accuracy and effectiveness. *View-Dependent, †View-Independent

4.2.3 Baseline Comparison

7Table 1 provides a comparative analysis of MiKASA against existing models in the Sr3D and Nr3D challenges [2]. MiKASA leads in overall accuracy for both Sr3D and Nr3D, achieving 75.2% and 64.4% respectively. Specifically, it demonstrates exceptional performance in the view-dependent category, achieving 70.4% in Sr3D and 65.4% in Nr3D, and significantly surpasses previous works. This performance underlines its capability to handle complex scenarios requiring changes in viewpoint, proving the effectiveness of our multi-key-anchor and features fusion strategy.

4.3. Ablation Studies

Effectiveness of Spatial Module:
In Table 2 we conduct a detailed ablation study on the spatial module. Removing the spatial encoder completely and use MLP for direct object location encoding significantly decreased accuracy to 45.0%. Removing the feedforward layer led to a 20% reduction in GPU memory usage, but caused a 2% accuracy drop to 62.4%. Omitting text-spatial fusion from our model caused a 1.8% decrease in overall performance, bringing it to 62.6%.

Spatial encoder	✓	-	✓	✓
Feedforward layer	✓	✓	-	✓
Text-Spatial fusion	✓	✓	✓	-
Overall acc	**64.4%**	45.0%	62.4%	62.6%

Table 2. Ablations of the spatial module, results highlighting the essentiality of each component.

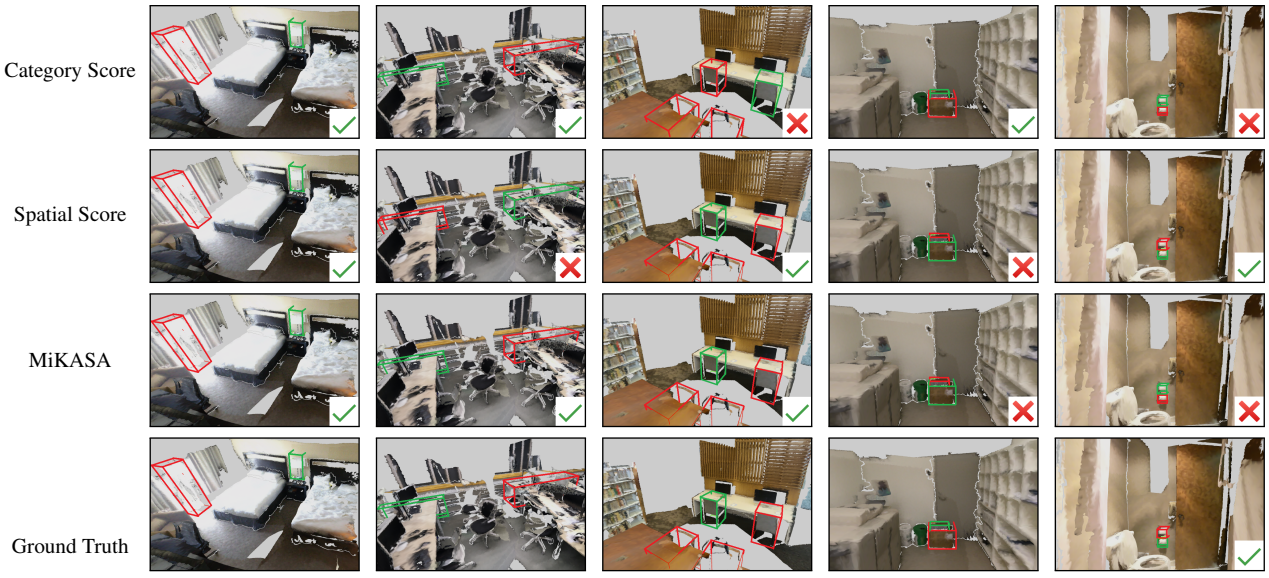
Object Encoder	Accuracy
PointNet++	63.8%
PointNet++ & GCN	65.5%
PointNet++ & Self-Attention Based	**70.8%**

Table 3. Comparison of object encoding strategies, presenting the object recognition accuracy achieved with different object encoding techniques. Showing the effectiveness of scene-aware object encoder.

Effectiveness of Scene-Aware Object Encoder:
Table 3 demonstrates how incorporating scene context enhances the performance of our object encoder, as compared to a standard PointNet++ encoder [26]. We introduced a scene-aware module after the PointNet++ layer in two variants: one employing Graph Convolutional Networks (GCN) [19] and the other using a self-attention mechanism [29].

For the GCN-based approach, we utilized Euclidean distance in the 3D space to determine neighborhood relations, specifically selecting the 10 nearest neighbors as the basis for graph construction. Our results show significant improvements in object classification accuracy. While the standard PointNet++ encoder achieved an accuracy of 63.8%, the GCN-based scene-aware encoder increased accuracy to 65.5%.

The self-attention-based scene-aware encoder further enhanced accuracy to 70.8%, showing the best performance. This highlights the effectiveness of scene-aware modules in improving object recognition by utilizing information about nearby objects.



(a) "The lamp between the two beds." (b) "The smaller radiator, close to the less cluttered desks." (c) "The chair facing wooden wall closest to the bookshelf." (d) "The smaller box closest to the door." (e) "The lower roll of toilet paper next to the toilet."

Figure 5. Visual representation of the model's decision-making process in diverse situations. Rows, from top to bottom, depict: (1) Choices determined by category score, (2) Choices determined by spatial score, (3) Our model's final selection after combining both scores, and (4) The established ground truth. Columns from left to right showcase varying scenarios. The green bounding box refers to the chosen object, and the red bounding box refers to the unchosen distractors.

Ablation Study on Spatial Features Aggregation: Table 4 shows our proposed method excels particularly in view-dependent scenarios. We substituted the attention-based aggregate layer with a simple mean function (I). This change led to a significant drop in accuracy, down to 33.9%. In contrast, employing max pooling (II) achieved a 61.7% accurac. It is worth noting that without the attention mechanism, accuracy in view-dependent scenarios falls significantly with mean and max pooling.

		Overall1	Easy	Hard	VD	VI
I	mean	33.9	42.7%	25.4%	32.8%	34.4%
		↓30.5%	↓27.0%	**↓34.0%**	**↓32.6%**	↓29.6%
II	max	61.4%	67.8%	55.2%	60.6%	61.7%
	pool	↓3.0%	↓1.9%	**↓4.2%**	**↓4.8%**	↓2.3%
	Ours	**64.4%**	**69.7%**	**59.4%**	**65.4%**	**64.0%**

Table 4. Ablations on fusion module

4.4. Multi-Modal Prediction

In Fig. 5, we illustrate MiKASA's decision-making in various scenarios. (a) shows accurate predictions where category and spatial scores align. (b) highlights accurate object identification with spatial relation challenges due to nearby objects with similar spatial features. (c) depicts scenarios where the model excels in spatial discernment, which is crucial in situations with multiple objects of the same category. (d) presents challenges in predictions where spatial cues are minimal, exemplified by two boxes near a door at a similar distance. Finally, (e) reveals cases where accurate spatial scoring is offset by inadequate category identification, such as with a poorly represented roll of toilet paper. The figure shows our model's decision-making is more explainable and facilitates easier diagnosis of errors. See the supplementary materials for more analysis.

5. Conclusion

In our study, we introduced the MiKASA (Multi-Key-Anchor Scene-Aware) Transformer, an innovative model designed to address the challenges in 3D visual grounding. This model uniquely combines a scene-aware object encoder with a multi-key-anchor technique, significantly enhancing object recognition and spatial understanding in 3D environments. The scene-aware object encoder effectively tackles object categorization issues, while the multi-key-anchor technique offers improved interpretation of spatial relationships and viewpoints. The results demonstrate that MiKASA outperforms current state-of-the-art models in both accuracy and explainability, underscoring its efficacy in advancing 3D visual grounding research. For future work, we suggest enhancing the model to explicitly preserve the directional information of objects post-encoding, aiming to further refine accuracy in view-dependent scenarios.

Acknowledgement: This research has been partially funded by EU project FLUENTLY (GA: Nr 101058680) and the BMBF project SocialWear (01IW20002).

References

[1] Ahmed Abdelreheem, Ujjwal Upadhyay, Ivan Skorokhodov, Rawan Al Yahya, Jun Chen, and Mohamed Elhoseiny. 3dreftransformer: Fine-grained object identification in real-world scenes using natural language. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3941–3950, January 2022. 2

[2] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas J. Guibas. ReferIt3D: Neural listeners for fine-grained 3d object identification in real-world scenes. In *16th European Conference on Computer Vision (ECCV)*, 2020. 1, 2, 3, 6, 7

[3] Firoj Alam, Ferda Ofli, and Muhammad Imran. Crisismmd: Multimodal twitter datasets from natural disasters. *CoRR*, abs/1805.00713, 2018. 3

[4] Eslam Mohamed Bakr, Yasmeen Youssef Alsaedy, and Mohamed Elhoseiny. Look around and refer: 2d synthetic semantics knowledge distillation for 3d visual grounding. In *Advances in Neural Information Processing Systems*. 2

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. 2

[6] Yunlu Chen, Vincent Tao Hu, Efstratios Gavves, Thomas Mensink, Pascal Mettes, Pengwan Yang, and Cees G. M. Snoek. Pointmixup: Augmentation for point clouds, 2020. 1

[7] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 6

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. 3, 6

[9] Carolina Galleguillos and Serge Belongie. Context based object categorization: A critical survey. *Computer Vision and Image Understanding*, 114(6):712–722, 2010. Special Issue on Multi-Camera and Multi-Modal Sensor Fusion. 3

[10] Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. Deep learning for 3d point clouds: A survey, 2020. 1, 2

[11] Zoey Guo, Yiwen Tang, Ray Zhang, Dong Wang, Zhigang Wang, Bin Zhao, and Xuelong Li. Viewrefer: Grasp the multi-view knowledge for 3d visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15372–15383, October 2023. 2, 7

[12] Dailan He, Yusheng Zhao, Junyu Luo, Tianrui Hui, Shaofei Huang, Aixi Zhang, and Si Liu. TransRefer3d: Entity-and-relation aware transformer for fine-grained 3d visual grounding. In *Proceedings of the 29th ACM International Conference on Multimedia*. ACM, oct 2021. 2

[13] Joy Hsu, Jiayuan Mao, and Jiajun Wu. Ns3d: Neuro-symbolic grounding of 3d objects and relations, 2023. 7

[14] Pin-Hao Huang, Han-Hung Lee, Hwann-Tzong Chen, and Tyng-Luh Liu. Text-guided graph neural networks for referring 3d instance segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(2):1610–1618, May 2021. 2

[15] Shijia Huang, Yilun Chen, Jiaya Jia, and Liwei Wang. Multi-view transformer for 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15524–15533, 2022. 1, 2, 3, 4, 6, 7

[16] Ayush Jain, Nikolaos Gkanatsios, Ishita Mediratta, and Katerina Fragkiadaki. Bottom up top down detection transformers for language grounding in images and point clouds, 2021. 1, 2, 3, 7

[17] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar, Oct. 2014. Association for Computational Linguistics. 2

[18] Sihyeon Kim, Sanghyeok Lee, Dasol Hwang, Jaewon Lee, Seong Jae Hwang, and Hyunwoo J. Kim. Point cloud augmentation with weighted local transformations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 548–557, October 2021. 1

[19] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017. 4, 7

[20] Miao Li, Shuying Zang, Bing Zhang, Shanshan Li, and Changshan Wu. A review of remote sensing image classification techniques: the role of spatio-contextual information. *European Journal of Remote Sensing*, 47(1):389–411, 2014. 1, 3

[21] Yong Liu, Ruiping Wang, Shiguang Shan, and Xilin Chen. Structure inference net: Object detection using scene-level context and instance-level relationships. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6985–6994, 2018. 1, 3

[22] Sreenivasulu Madichetty, Sridevi Muthukumarasamy, and P. Jayadev. Multi-modal classification of twitter data during disasters for humanitarian response. *Journal of Ambient Intelligence and Humanized Computing*, 12(11):10223–10237, jan 2021. 3

[23] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase

correspondences for richer image-to-sentence models, 2016. 2

[24] Raj Pranesh. Exploring multimodal features and fusion strategies for analyzing disaster tweets. In *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, pages 62–68, Gyeongju, Republic of Korea, Oct. 2022. Association for Computational Linguistics. 3

[25] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation, 2017. 1, 3

[26] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *arXiv preprint arXiv:1706.02413*, 2017. 1, 2, 3, 6, 7

[27] Guocheng Qian, Yuchen Li, Houwen Peng, Jinjie Mai, Hasan Hammoud, Mohamed Elhoseiny, and Bernard Ghanem. Pointnext: Revisiting pointnet++ with improved training and scaling strategies. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1

[28] Damien Teney, Lingqiao Liu, and Anton van den Hengel. Graph-structured representations for visual question answering. *CoRR*, abs/1609.05600, 2016. 3

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 4, 7

[30] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph Attention Networks. *International Conference on Learning Representations*, 2018. 4

[31] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (TOG)*, 2019. 4

[32] Junxiao Xue, Yabo Wang, Yichen Tian, Yafei Li, Lei Shi, and Lin Wei. Detecting fake news by exploring the consistency of multimodal data. *Inf. Process. Manage.*, 58(5), sep 2021. 3

[33] Zhengyuan Yang, Songyang Zhang, Liwei Wang, and Jiebo Luo. Sat: 2d semantics assisted training for 3d visual grounding. In *ICCV*, 2021. 1, 2, 3, 6, 7

[34] Jian Yao, Sanja Fidler, and Raquel Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 702–709, 2012. 3

[35] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions, 2016. 2

[36] Zhihao Yuan, Xu Yan, Yinghong Liao, Ruimao Zhang, Sheng Wang, Zhen Li, and Shuguang Cui. Instancerefer: Cooperative holistic understanding for visual grounding on point clouds through instance multi-level contextual referring, 2021. 7

[37] Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng. Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE Journal of Selected Topics in Signal Processing*, PP:1–1, 04 2020. 3

[38] Duoyi Zhang, Richi Nayak, and Md Abul Bashar. Exploring fusion strategies in deep learning models for multi-modal classification. In Yue Xu, Rosalind Wang, Anton Lord, Yee Ling Boo, Richi Nayak, Yanchang Zhao, and Graham Williams, editors, *Data Mining*, pages 102–117, Singapore, 2021. Springer Singapore. 3

[39] Mengmi Zhang, Claire Tseng, and Gabriel Kreiman. Putting visual object recognition in context. *CoRR*, abs/1911.07349, 2019. 1, 3

[40] Yiming Zhang, ZeMing Gong, and Angel X. Chang. Multi3drefer: Grounding text description to multiple 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15225–15236, October 2023. 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 (ICCV)*, pages 2928–2937, October 2021. 1, 2, 3, 7