arXiv:2506.14495v1 [cs.CV] 17 Jun 2025

I Speak and You Find: Robust 3D Visual Grounding with Noisy and Ambiguous Speech Inputs

Yu Qi, Lipeng Gu, Honghua Chen, Liangliang Nan, and Mingqiang Wei, *Senior Member, IEEE*

***Abstract*—Existing 3D visual grounding methods rely on precise text prompts to locate objects within 3D scenes. Speech, as a natural and intuitive modality, offers a promising alternative. Real-world speech inputs, however, often suffer from transcription errors due to accents, background noise, and varying speech rates, limiting the applicability of existing 3DVG methods. To address these challenges, we propose SpeechRefer, a novel 3DVG framework designed to enhance performance in the presence of noisy and ambiguous speech-to-text transcriptions. SpeechRefer integrates seamlessly with existing 3DVG models and introduces two key innovations. First, the Speech Complementary Module captures acoustic similarities between phonetically related words and highlights subtle distinctions, generating complementary proposal scores from the speech signal. This reduces dependence on potentially erroneous transcriptions. Second, the Contrastive Complementary Module employs contrastive learning to align erroneous text features with corresponding speech features, ensuring robust performance even when transcription errors dominate. Extensive experiments on the SpeechRefer and SpeechNr3D datasets demonstrate that SpeechRefer improves the performance of existing 3DVG methods by a large margin, which highlights SpeechRefer's potential to bridge the gap between noisy speech inputs and reliable 3DVG, enabling more intuitive and practical multimodal systems.**

***Index Terms*—SpeechRefer, Speech-guided 3D Visual Grounding, Audio-Visual Learning, 3D Scene Understanding**

I. Introduction

INTERACTING with visual data using natural language is a rapidly advancing field [1]–[7] with real-world applications such as robotics and AR/VR. Within this domain, 3D visual grounding, commonly referred to as text-guided 3D visual grounding (T-3DVG), focuses on locating specific objects based on precise text prompts within 3D scenes.

Recent research [8]–[16] has focused on designing sophisticated architectures to achieve state-of-the-art performance in 3DVG. For example, [9] introduces a framework that explicitly models geometry-aware visual representations while generating fine-grained, language-guided object queries. [10] learns the multi-attribute interactions to refine the intra-modal and inter-modal grounding cues. Although accurate text features offer structured word-level context for strong performance in T-3DVG tasks, they often fail to address challenges in complex real-world environments and overlook the actuality that precise text cannot appear out of thin air.

Considering the real-world environments, speech stands out as a natural modality compared to text, offering an intuitive way for humans to interact with AI systems. For example, you can directly ask a robot to locate "*the brown guitar next to the bed*", which reflects natural human communication in real-world scenarios. Logically, speech recognition models can serve as a bridge between speech and existing T-3DVG methods to achieve speech-guided 3D visual grounding (S-3DVG). Despite advances in speech recognition, complex real-world environments inevitably lead to a significant challenge during transcriptions: errors and uncertainties that arise due to variations in speech quality, such as accents, background noise, or speech rate. For instance, as illustrated in Figure 1, the input speech with Chinese accent "*There is a grey office chair*" might be transcribed incorrectly as "*There is a grain office chair*". This misinterpretation leads the model to search for a non-existent object (attribute) and fails to ground the correct object. Such transcription errors are common and have cascading effects, severely impairing the system's ability to ground objects accurately. Unfortunately, existing T-3DVG methods rely heavily on precise text inputs to perform effectively, which reveals a significant gap between T-3DVG methods and real-world scenarios with complexity and noise.

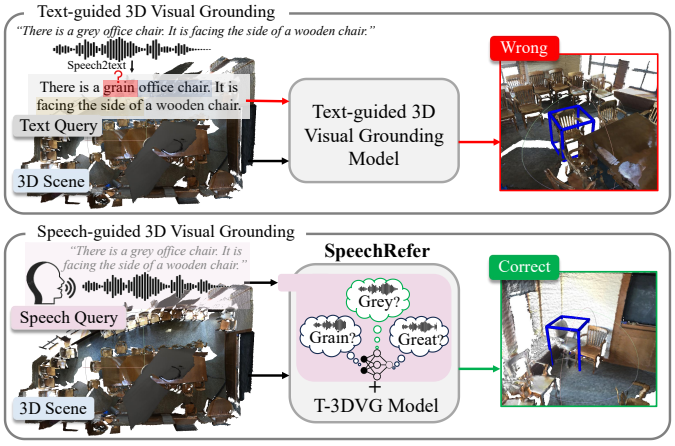


Fig. 1. **Illustration of SpeechRefer.** SpeechRefer captures acoustic similarities between phonetically related words, which acts as a complement to potential transcription errors, reducing dependence on potentially erroneous transcriptions.

To address this challenge, we propose **SpeechRefer**, a novel 3DVG framework that seamlessly integrates with existing T-3DVG methods, leveraging the features of raw speech to complement text derived from the speech and mitigate the impact of transcription errors. As shown in Figure 1, the speech features space preserves acoustic similarities between phonetically similar words (e.g. *grey, grain*) while capturing

Y. Qi, L. Gu, H. Chen and M. Wei are with the School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China (e-mail: qiyuyu@nuaa.edu.cn; glp1224@163.com; chenhonghuacn@gmail.com; mingqiang.wei@gmail.com).

L. Nan is with the Urban Data Science Section, Delft University of Technology, Delft, Netherlands (e-mail: liangliang.nan@tudelft.nl).

phonetic nuances often lost during speech-to-text conversion. This allows the system to infer the correct target object by considering alternative interpretations of the user's intent, rather than relying solely on potentially erroneous transcription.

Specifically, we propose a speech complementary module comprising two components: a phonetic-aware refinement module and a confidence-based complementary module. The phonetic-aware refinement module inherently captures acoustic similarities between phonetically related words while preserving subtle speech nuances. Then, the confidence-based complementary module explicitly generates complementary proposal scores based on speech features, effectively complementing the transcribed text scores and reducing reliance on potentially erroneous transcription. By incorporating complementary contextual information, the system identifies "*grey*" as the most likely intended attribute, even when the transcription incorrectly suggests "*grain*". To address scenarios where erroneous text dominates and renders speech complementary scores ineffective, we further propose a contrastive complementary module. This module uses contrastive learning to align erroneous text features with accurate speech features, ensuring that speech features effectively complement text derived from the speech even in challenging situations. By integrating speech features as a complement to textual features, SpeechRefer effectively addresses the challenges of imperfect speech-to-text transcriptions, which enables more accurate interpretation of user intent in uncertain conditions

Our contributions include:

- SpeechRefer is the first and novel 3DVG framework designed to enhance model robustness in the presence of noisy and ambiguous transcriptions, which can integrate with existing T-3DVG methods and improve its performance by a large margin.
- The speech complementary module retains acoustic similarities while distinguishing subtle differences, avoiding being constrained by potential transcription errors. The contrastive complementary module corrects misaligned text features using speech features, ensuring robust performance even in transcription-dominated scenarios.
- We introduce the first speech datasets that reflect complex real-world environments, representing a promising step toward more intuitive 3D visual grounding.

II. Related Work

A. Speech Recognition

Speech recognition is a technology that converts speech into text and is one of the most critical advancements in AI-driven interaction. Its applications have permeated various aspects of daily life and production, such as voice assistants in mobile phones. After years of research and development, speech recognition has become quite advanced [17], [18], particularly for English. For instance, the Whisper [17] model achieves a word error rate (WER) as low as 4.1% on the English category of the Fleurs dataset. However, even with state-of-the-art speech recognition models, transcription errors are still common due to variations in speech quality, such as accents, background noise, or speech rate. Potential misinterpretation leads the model to search for a non-existent object (attribute) and fails to ground the correct object. SpeechRefer was proposed to address this challenge.

B. Language and Speech in 3D Interaction

Recent works in 3D vision-language understanding have advanced tasks like object grounding, dense captioning, and question answering in 3D scenes. Mao et al. [19] proposed a modality alignment network for 3D dense captioning, capturing both local and global spatial relationships to improve semantic alignment. Ye et al. [20] introduced 3DQA, extending traditional VQA into 3D environments by jointly modeling geometry, appearance, and language inputs . These efforts highlight the potential of language as a powerful interface for 3D scene understanding.

Meanwhile, audio-visual learning has been increasingly explored in 3D tasks such as 3D facial animation [21] and embodied interaction, where speech serves as a driving signal for motion generation or behavior control. This line of work highlights the growing interest in leveraging speech as a natural modality for enhancing realism, interactivity, and user engagement in 3D environments.

Inspired by these directions, our work explores speech-guided 3D visual grounding under noisy conditions—a novel yet promising task. By enhancing robustness to speech ambiguities, we aim to support more intuitive, speech-based 3D interaction in real-world applications such as AR/VR.

C. 3D Visual Grounding

Since the introduction of the T-3DVG task by ScanRefer [22] and ReferIt3D [23], it has developed rapidly and found applications across various fields. Existing T-3DVG methods are typically categorized into two-stage and one-stage frameworks, with the majority adopting a two-stage framework. Two-stage methods [24]–[29] first extract text features from the query language using language models [30], [31] and generate proposal boxes using a pre-trained detector [32], [33]. These features are then fused to select the best-matched object. Among these, ConcreNet [34] presents four novel stand-alone modules designed to enhance performance in challenging repetitive instances. MA2TransVG [10] learns the multi-attribute interactions to refine the intra-modal and inter-modal grounding cues. Conversely, one-stage methods [35]–[37] directly localize the target in a single step. Notably, 3D-SPS [35], the first one-stage model, utilizes text features to guide visual keypoint selection, facilitating progressive object grounding. EDA [36] introduces a text decoupling module to generate textual features for each semantic component, significantly enhancing grounding performance, and even surpassing most existing two-stage frameworks.

In general, most methods focus on designing sophisticated architectures to achieve state-of-the-art grounding performance, but they overlook the challenges posed by complex real-world environments and practical applications. Our SpeechRefer is the first and novel 3DVG framework that could integrate with existing T-3DVG methods, bridging the

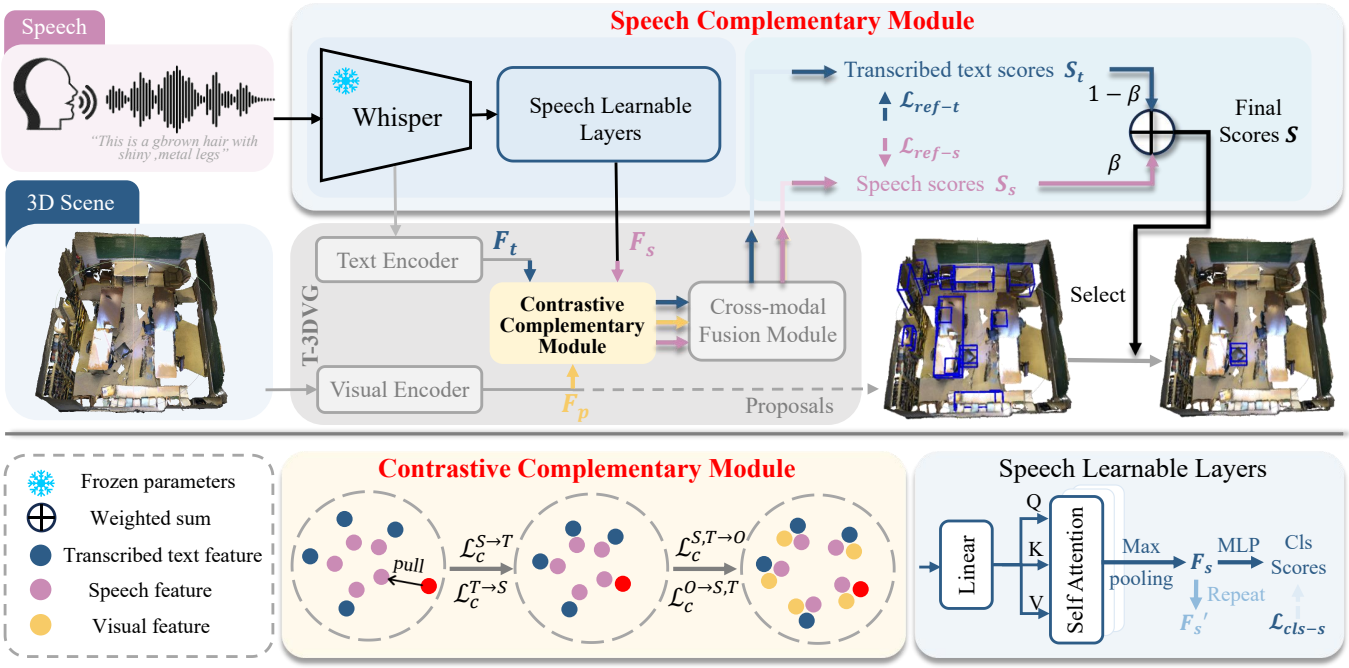


Fig. 2. **Overview of SpeechRefer.** The framework builds on existing T-3DVG methods, incorporating two novel modules highlighted here: (1) **Speech complementary module**: This module encodes speech features F_s and generates speech scores S_s to explicitly complement the transcribed text. By integrating speech features, it helps identify the correct object and mitigates the impact of potential transcription errors. (2) **Contrastive complementary module**: This module aligns potentially erroneous text features with corresponding speech features by contrastive learning. This alignment ensures robust performance even when erroneous text features dominate. Key components are color-coded, where gray represents pre-existing modules (e.g., visual and text encoders). S, T, and O denote the global speech feature, sentence-level transcribed text feature, and the mean of object features of all target objects associated with the same description, respectively.

gap between T-3DVG methods and real-world scenarios with complexity and noise.

III. Method

A. Overview

As shown in Figure 2, our network has two inputs: one is point cloud $P \in \mathbb{R}^{N\times(3+K)}$ that represents the whole 3D scene by 3D coordinates and K-dimensional auxiliary feature (e.g., RGB, normal vectors, or the pre-trained multi-view features [22]). Another input is speech, which is a human voice describing a target object in the 3D scene. The output is the predicted target object, which is an axis-aligned bounding box with the center $\mathbf{c} = [c_x, c_y, c_z] \in \mathbb{R}^3$ in the world coordinate, and the size $\mathbf{s} = [s_x, s_y, s_z] \in \mathbb{R}^3$. Our SpeechRefer is a novel S-3DVG framework that could seamlessly integrate with existing T-3DVG methods, so the visual encoder, language encoder, and cross-modal fusion module are consistent with the T-3DVG models.

Our SpeechRefer features two key innovations. First, the **speech complementary module** (III-C) captures acoustic similarities between phonetically related words while distinguishing subtle distinctions. This module generates complementary proposal scores from speech signals, reducing dependence on potentially erroneous transcriptions. **Contrastive complementary module** (III-D) aligns erroneous text features with corresponding correct speech features through contrastive learning, ensuring robust performance even when errors dominate. Finally, Section III-E describes the total loss.

B. Baseline: Text-guided 3D Visual Grounding via Speech-to-Text

To clarify the challenges addressed by our approach, we first describe the baseline T-3DVG method that relies on speech-to-text transcription. The speech recognition model serves as a bridge between speech and T-3DVG by converting speech to text, which is then processed by any T-3DVG model to locate objects. However, even advanced speech recognition models frequently introduce critical errors due to factors like accents, background noise, and speech rate, reflecting real-world conditions. Figure 3 illustrates common transcription errors when the input speech contains uncertainties such as accents, where the original T-3DVG model often fails to ground the correct objects. Among these errors, inaccuracies in transcribing target object names are particularly frequent and problematic. Prior studies [36] have confirmed that grounding results significantly degrade when object names are omitted. Additionally, errors in attributes, such as misinterpreting "*white*" as "*wide*" can also mislead the model by overlooking crucial contextual cues, increasing its susceptibility to distractors within the same category and ultimately impairing overall performance.

C. Speech Complementary Module

To address transcription errors, we propose a speech complementary module that incorporates phonetic features directly from speech. It draws on the acoustic similarities and nuances of input speech as a complement to mitigate the negative effects of potentially erroneous transcriptions. It mainly com-

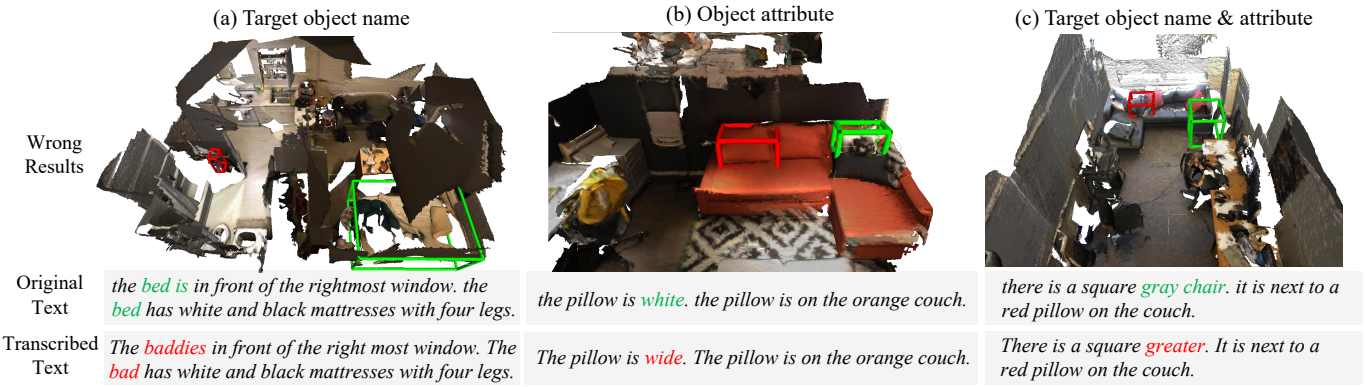


Fig. 3. Incorrectly transcribed text may include errors such as misinterpreting the target object name or its attributes. These mistakes can mislead the model into searching for a non-existent object or attribute, ultimately failing to correctly identify and ground the intended object.

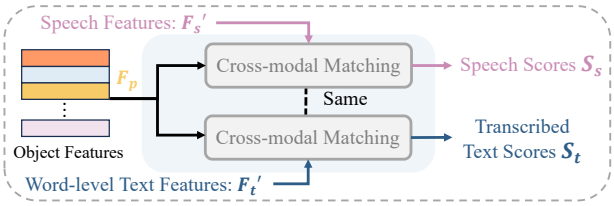


Fig. 4. Cross-modal Fusion Module. The module consists of two identical cross-modal matching modules, which is the specific matching module of existing T-3DVG models.

prises two submodules.

Phonetic-aware Refinement Module. First, we utilize the Whisper [17] as our speech encoder, which inherently preserves information about acoustic similarities that are often lost during the speech-to-text conversion. Whisper is a general-purpose speech recognition model trained on a large, diverse speech dataset. As a multitask model, it can also perform language identification and speech translation, effectively exploiting contextual clues in speech, which makes it an excellent source of initial speech features for a wide range of downstream networks. Before the raw speech is fed into the network, it's resampled to 16,000 Hz, and an 80-channel log-magnitude Mel spectrogram representation $M_s \in \mathbb{R}^{80\times L}$ is computed on 25 millisecond windows with a stride of 10 milliseconds [17]. The speech log-magnitude Mel spectrogram M_s is encoded into speech feature $W_s \in \mathbb{R}^{N_s\times D_s}$ through the Whisper encoder, where N_s denotes the temporal sequence length of speech, and D_s denotes semantic feature dimension.

To enhance the pre-trained Whisper model, we design learnable layers and introduce classification loss to fine-tune the speech features. This process emphasizes the speech features of the target object class, which enables the model to better distinguish subtle differences. The initial speech feature W_s is the input of learnable layers, which can be formulated as:

$$W_s' = SelfAttention(Linear(W_s)) \tag{1}$$

where Linear(·) denotes the linear projection, SelfAttention(·) denotes the multi-head self-attention mechanism [38]. Then, the global speech feature is extracted via max pooling, denoted as $F_s \in \mathbb{R}^{1\times D_s}$, and the features are repeated and stacked, generating the final speech feature representation F_s'. After that, a speech classification loss is included, which can be formulated as:

$$\mathcal{L}_{cls-s} = -\sum_{i=1}^{C} y_i log(\hat{y}_i) \tag{2}$$

where C is the number of classes, y_i is the ground truth one-hot encoding, $\hat{y}_i$ is classification scores after a classifier head. This enables the model to dynamically capture acoustic similarities in speech while distinguishing subtle distinctions by emphasizing the features of the target object name, further improving the expressiveness of speech features.

Confidence-based Complementary Module. Based on the former, the confidence-based complementary module explicitly generates complementary scores for proposals based on speech features to complement the textual score, mitigating the impact of potential transcription errors. Figure 4 illustrates the structure, the cross-modal matching component is the specific matching module used in existing T-3DVG models. We utilize the same network to fuse speech-visual and transcribed text-visual features respectively, ensuring feature consistency. Then the speech-visual fusion features are fed into an FFN layer to generate the speech scores S_s for the M generated proposals except for transcribed text scores S_t. This allows the model to incorporate multiple potential interpretations of the user's intent, leveraging speech scores to complement the transcribed text ones. This reduces the model's reliance on potentially erroneous transcriptions. The final scores $S = \beta S_s + (1-\beta)S_t$, where β is a weighting factor. Eventually, the bounding box with the highest final score will be considered as the final grounding result. Additionally, we introduce a reference loss to supervise the module, encouraging that the proposal with the highest score is closer to the ground truth, thereby promoting more accurate target object selection:

$$L_{ref} = \alpha_1 L_{ref-s} + \alpha_2 L_{ref-t} \tag{3}$$

$$L_{ref-s} = -\sum_{i=1}^{M} t_i log(s_{si}), L_{ref-t} = -\sum_{i=1}^{M} t_i log(s_{ti}) \tag{4}$$

where L_{ref-s} and L_{ref-t} are cross-entropy losses based on speech and transcribed text scores respectively, following the

strategy in ScanRefer [22], we set the label t_i for the i^{th} box that has the highest IoU score with the ground truth as 1 and others as 0.

D. Contrastive Complementary Module.

While utilizing speech information significantly mitigates the negative impact of potential transcription errors, situations still arise where erroneous text weighs heavily on proposal selection, leading to failures in identifying the target object despite the complementary speech scores. To address this, we propose a contrastive complementary module. This module builds on the observation that most objects (or attributes) incorrectly generated by transcription, such as "*bat*", "*grain*", etc., do not exist in the input scene. At the feature level, even if the textual features represent "*grain*", they should align with the speech features of "*grey*" to correctly infer the target object. To achieve this, we employ contrastive learning to align incorrect textual features with the correct speech features, ensuring more accurate target object inference:

$$\mathcal{L}_c^{T\to S} = -\sum_{i=1}^{N} log\frac{exp(s(T_i, S_i)/t)}{\sum_{j=1}^{n} exp(s(T_i, S_j)/t)} \tag{5}$$

where N is the number of samples in one batch, t is a temperature parameter, and $s(\cdot)$ is the cosine similarity. S is the global speech feature (i.e. F_s), and T is the sentence-level transcribed text feature (i.e. F_t).

Additionally, to align with the corresponding visual features, we introduce additional contrastive loss between language and visual features:

$$\begin{aligned}\mathcal{L}_c^{S,T\to O} = -(&\sum_{i=1}^{N} log\frac{exp(s(S_i, O_i)/t)}{\sum_{j=1}^{n} exp(s(S_i, O_j)/t)} + \\ &\sum_{i=1}^{N} log\frac{exp(s(T_i, O_i)/t)}{\sum_{j=1}^{n} exp(s(T_i, O_j)/t)})\end{aligned} \tag{6}$$

O is the mean of object features of all target objects paired with the same description. We symmetrize these losses by adding the analogous terms, $\mathcal{L}_c^{S\to T}$, $\mathcal{L}_c^{O\to S,T}$. The total contrastive loss is the average of these contrastive losses. By aligning text features with corresponding speech features, the model can more effectively correct transcription errors, ensuring robust performance even when errors dominate.

E. Total Loss

We train the network end-to-end with the total loss $\mathcal{L} = \mathcal{L}_{det} + \gamma_1\mathcal{L}_c + \gamma_2\mathcal{L}_{ref} + \gamma_3\mathcal{L}_{cls}$. $\mathcal{L}_{det}$ varies with different integration methods. For example, in 3DVG-Transformer, $\mathcal{L}_{det} = 10\mathcal{L}_{vote-reg} + \mathcal{L}_{objn-cls} + \mathcal{L}_{sem-cls} + 10\mathcal{L}_{box}$, and this object detection loss exactly follows the loss used in [33] for the ScanNet dataset [39]. To better adapt the total loss function to different methods, we adjust $\gamma_1, \gamma_2, \gamma_3$ to ensure that all the loss components are balanced and remain in the same order of magnitude.

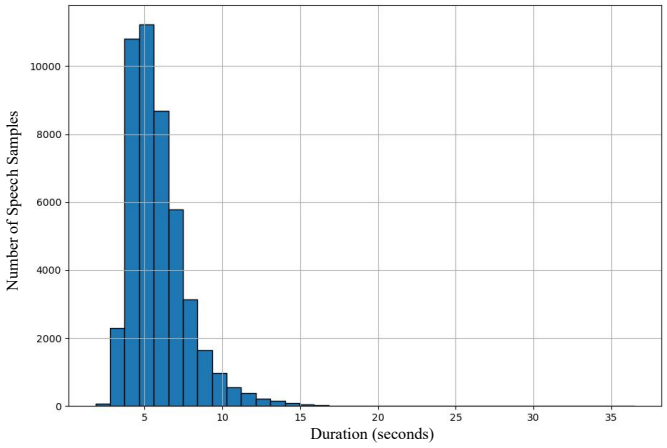


Fig. 5. Speech Duration Distribution of SpeechRefer dataset.

IV. Experiments

A. Experimental Setting

SpeechRefer and SpeechNr3D datasets. We introduce new speech datasets designed to simulate real-world variations, such as accents, speech rates, and background noise, based on ScanRefer [22] and Nr3D [23]. To incorporate accents, we use the off-the-shelf text-to-speech tool[1] to generate speech with a **C**hinese **a**ccent, creating the SpeechRefer_CA and SpeechNr3D_CA datasets. For background noise, we overlay one of five common sounds, i.e., birds chirping, rain, cars, people talking, and piano, onto the speech recordings. Datasets with normal rate and without accent and background noise are denoted as SpeechRefer_VF and SpeechNr3D_VF (**V**ariations-**f**ree) accordingly. Our SpeechRefer dataset matches the size of ScanRefer and classifies scenes as "unique" (containing a single object of the class) or "multiple" (containing multiple objects). Similarly, the SpeechNr3D dataset mirrors Nr3D and is categorized as "Easy" or "Hard" based on the number of distractors, and as "View-dependent" or "View-independent" depending on the reliance on the speaker's viewpoint. The distribution of speech durations in the SpeechRefer dataset is illustrated in Figure 5, showing a prominent peak in the 4–8 second range, which represents the most frequent duration span. Overall, durations vary from 1 to 35 seconds, though instances at the extremes—particularly those shorter than 3 seconds or longer than 15 seconds—are comparatively rare. The average duration across all samples is 5.89 seconds. To further evaluate the robustness of SpeechRefer, we manually recorded an additional set of speech data comprising 699 training samples and 352 validation samples, designed to reflect real-world usage scenarios more accurately. The corresponding evaluation results are presented in IV-B.

Evaluation metrics. We adopt the same metrics as other T-3DVG methods: Acc@0.25IoU and Acc@0.5IoU, which measure the percentage of correctly predicted bounding boxes with IoU scores exceeding 0.25 and 0.5. For SpeechNr3D, we evaluate the percentage of successful matches between

[1]https://pypi.org/project/pyttsx3/

TABLE I
RESULTS ON THE SPEECHREFER_CA DATASET (I.E., SPEECHREFER DATASET WITH CHINESE ACCENT) AND COMPARISON RESULTS AGAINST BASELINE. NOTABLY, SPEECHREFER EXHIBITS PROMISING PERFORMANCE WHEN FINE-TUNED ON THE SPEECHREFER_VF DATASET (VARIATIONS-FREE SPEECHREFER DATASET).

Methods	Setting	Modality	Unique(19%)		Multiple(81%)		Overall	
			0.25	0.5	0.25	0.5	0.25	0.5
ScanRefer	Baseline	3D	67.93	46.87	30.99	21.39	38.15	26.34
3D-VisTA	Baseline	3D	79.19	72.42	42.63	38.34	49.40	44.64
3DVG-Transformer	Baseline	3D	75.93	55.21	37.27	27.02	44.77	32.49
	+SpeechRefer (ours)		**77.13**	**57.19**	**39.74**	**28.73**	**47.00**	**34.25**
	Improvements		**+1.20**	**+1.98**	**+2.47**	**+1.71**	**+2.23**	**+1.76**
	Fine-tune on SpeechRefer_VF		79.02	57.58	39.88	29.14	47.47	34.66
M3DRefCLIP	Baseline	3D+2D	84.60	**74.50**	42.00	35.20	50.30	42.80
	+SpeechRefer (ours)		**85.20**	74.10	**44.80**	**37.30**	**52.60**	**44.40**
	Improvements		**+0.60**	-0.40	**+2.80**	**+2.10**	**+2.30**	**+1.60**
	Fine-tune on SpeechRefer_VF		85.90	75.30	46.10	38.00	53.80	45.20

TABLE II
RESULTS ON SPEECHNR3D_CA DATASET SET WITH GT BOXES AND COMPARISON RESULTS WITH THE BASELINE.

Methods	Setting	Modality	Overall	Easy	Hard	View Dep	View Indep.
M3DRefCLIP	Baseline	3D+2D	42.64	48.78	36.98	36.26	46.07
	+SpeechRefer (ours)		**46.24**	**52.62**	**40.34**	**39.80**	**49.69**
	Improvements		**+3.60**	**+3.84**	**+3.36**	**+3.54**	**+3.62**
	Fine-tune on SpeechNr3D_VF		47.70	54.25	41.66	39.81	51.73

predicted and ground truth bounding boxes, following previous work [23].

Implementation details. We train the model end-to-end on a single NVIDIA RTX4090. The optimizer and learning rate are consistent with the T-3DVG methods integrated with SpeechRefer. For 3DVG-Transformer, we train for 200 epochs with a batch size of 2, pairing each scene with 32 sentences. For M3DRefCLIP, the network is trained for 60 epochs with a batch size of 4, pairing each scene with 8 sentences. Other training configurations align with standard T-3DVG methods.

B. Experimental Results

Tabel I presents the results of 3DVG-Transformer [40] and M3DRefCLIP [41] integrated with our SpeechRefer on the SpeechRefer_CA dataset, alongside baseline results from four methods [22], [36], [40], [41].

The baseline performance suffers significantly when compared to results using precise text inputs due to errors in transcribed text. In contrast, SpeechRefer shows marked improvements. For 3DVG-Transformer, SpeechRefer achieves gains of 2.23%/1.76% in overall accuracy. For M3DRefCLIP, the overall accuracy improves by more than 2.30%/1.60%, with a notable 2.80%/2.10% gain on the "multiple" subset, demonstrating SpeechRefer's ability to effectively distinguish distractors. Further fine-tuning on the variations-free SpeechRefer_VF dataset yields even better results than training directly on it (please refer to the results in supplementary materials). This highlights SpeechRefer's robustness in handling variations. Quantitative results demonstrate that our SpeechRefer can reliably localize target objects from speech descriptions and can easily integrate with existing T-3DVG methods, significantly enhancing performance and robustness. For the SpeechNr3D_CA dataset, as shown in Table II, our results outperform the baseline across all subsets, with an overall accuracy improvement of 3.60%. Compared to the SpeechRefer dataset, SpeechNr3D is more sensitive to transcription errors due to its shorter descriptions (11.4 words on average, compared to ScanRefer's 20.27 words). The shorter descriptions make transcription errors more impactful, but SpeechRefer mitigates these adverse effects effectively, demonstrating robust performance.

Furthermore, we evaluate SpeechRefer on the aforementioned real-world dataset, which closely mirrors practical application scenarios, with the results summarized in Table III. The experiments show that SpeechRefer can effectively adapt to real-world speech inputs with a relatively small amount of fine-tuning. Specifically, it's fine-tuned on 699 manually recorded training samples and evaluated on 352 validation samples. For comparison, the results in the first row of Table III are obtained using the same number of synthesized validation samples, which are drawn from the introduced SpeechRefer_CA dataset, further demonstrating the model's robustness and generalization ability in realistic settings.

C. Ablation Studies

Speech rate and background noise. We evaluate SpeechRefer's robustness under varied noisy conditions by introducing two additional variations: speech rate and background noise. For speech rate, we tested slower (100) and faster (200) rates. Results in Table IV show improved perfor-



Fig. 6. Qualitative results. We utilize green, red, and blue boxes to represent the ground truth, the prediction of baseline, and the prediction of SpeechRefer based on 3DVG-Transformer, respectively.

TABLE III
PERFORMANCE ON REAL SPEECH DATA. BOTH EVALUATIONS—ON SYNTHETIC AND HUMAN-RECORDED SPEECH—WERE CONDUCTED USING THE SAME 352-SAMPLE VALIDATION SET. THE SYNTHETIC SPEECH SAMPLES WERE TAKEN DIRECTLY FROM THE SPEECHREFER_CA DATASET, WHILE THE REAL-WORLD SPEECH CONSISTED OF MANUALLY RECORDED AUDIO COLLECTED FROM HUMAN SPEAKERS.

Methods	Unique	Data
3DVG-T	75.75/58.18	Synthetic speech
+SpeechRefer (ours)	89.69/67.27	Real-world speech

TABLE IV
ABLATION STUDY OF SPEECHREFER'S ROBUSTNESS UNDER DIFFERENT CONDITIONS, INCLUDING VARYING SPEECH RATE AND BACKGROUND NOISE, BASED ON 3DVG-TRANSFORMER.

Type of variations		Unique 0.25	Multiple 0.25	Overall 0.25
Speech Rate	100	78.44	40.43	47.81
	200	77.13	39.74	47.00
Background Noise		77.48	40.50	47.67

mance at slower rates, where reduced speed enhances semantic clarity. Even at faster rates, SpeechRefer maintains strong performance, showcasing its resilience. Under background noise conditions, SpeechRefer consistently outperforms baseline, which indicates its robustness in real-world noisy scenarios.

Module contributions. Table V provides ablation results for different module combinations based on 3DVG-Transformer. The second row highlights the significant gains achieved by the speech learnable layers, which capture subtle differences in addition to acoustic similarities, thereby enhancing speech feature expressiveness. From the final row, we can see that the

TABLE V
ABLATION STUDY OF DIFFERENT MODULE COMBINATIONS IN SPEECHREFER: SLL (SPEECH LEARNABLE LAYERS), CBM (CONFIDENCE-BASED COMPLEMENTARY MODULE), CCM (CONTRASTIVE COMPLEMENTARY MODULE). EVALUATED ON SPEECHREFER_CA DATASETS.

SLL	CBM	CCM	Multiple		Overall	
			0.25	0.5	0.25	0.5
			37.27	27.02	44.77	32.49
✓			38.52	27.72	45.92	33.32
✓		✓	38.33	28.13	45.56	33.47
✓	✓	✓	**39.74**	**28.73**	**47.00**	**34.25**

TABLE VI
PERFORMANCE CROSS ALIGNMENT TYPES (T REPRESENTS TEXT FEATURES, O REPRESENTS VISUAL FEATURES, S REPRESENTS SPEECH FEATURES). ROWS 1–4: SPEECHREFER + 3DVG-TRANSFORMER; ROWS 5–6: SPEECHREFER + M3DREFCLIP.

T&O	S&O	T&S	Unique	Multiple	Overall
			75.84/55.78	38.29/27.63	45.57/33.09
✓			75.64/54.51	36.74/25.98	44.29/31.51
✓		✓	76.17/54.98	36.33/26.37	44.06/31.92
✓	✓	✓	77.13/57 19	39.74/28.73	47.00/34.25
✓			84.60/74.50	42.00/35.20	50.30/42.80
✓	✓	✓	85.20/74.10	44.80/37.30	52.60/44.40

combination of the confidence-based complementary module and the contrastive complementary module further improves performance. This validates their roles in mitigating transcription errors and enhancing robustness.

Alignment types in Contrastive Complementary Module. Table VI illustrates performance across alignment types (T: text features; O: visual features; S: speech features). When aligning only speech and visual features or only text and visual features, the performance is not optimal.

Only Speech. While transcribed text improves overall accuracy, it is not an ideal solution for the speech-guided 3D visual grounding task, as incorporating text adds complexity to the network and training process. Therefore, we conducted a preliminary exploration of a simplified SpeechRefer architecture that relies solely on the speech and point cloud modality. The results, presented in Table VII, show that our approach significantly outperforms AP-Refer [42], the first and only 3DVG network relies solely on the speech and point cloud. This demonstrates the effectiveness and robustness of our SpeechRefer framework.

We also explore the impact of different speech encoders. Table VIII presents the results of the speech-only network using Imagebind and Whisper as speech encoders, respectively. The results indicate that when Imagebind is used as the encoder, the performance is notably lower compared to Whisper. We attribute this discrepancy to the fact that, while Imagebind is designed to learn joint embeddings across six modalities, it primarily integrates these modalities through image-paired data. As a result, when applied to point cloud-centric tasks, a gap remains between the modalities. In contrast, Whisper is an encoder specifically optimized for speech processing, enabling it to more effectively capture the phonetic similarities

TABLE VII
COMPARISON RESULTS OF AP-REFER [42] AND SIMPLIFIED SPEECHREFER INTEGRATE WITH 3DVG-TRANSFORMER (THE SECOND ROW) AND M3DREFCLIP (THE FINAL ROW).

Methods	Unique	Multiple	Overall
AP-Refer	48.62/29.59	16.94/9.96	23.09/13.77
Ours	76.04/54.44	35.93/25.62	43.72/31.22
	84.00/73.90	41.60/34.50	49.50/42.20

TABLE VIII
ABLATION STUDY OF SPEECH ENCODER. RESULTS ON DIFFERENT SPEECH ENCODERS (IMAGEBIND [43] AND WHISPER [17]), EVALUATING ON 3DVG-TRANSFORMER.

Encoder	Unique	Multiple	Overall
Imagebind	55.10/39.98	22.63/16.38	28.93/20.96
Whisper	76.04/54.44	35.93/25.62	43.72/31.22

within speech and guide the network toward better results.

Ablation study of confidence score weight. Table IX presents the results from various weight combinations assigned to the speech confidence score S_a and the text confidence score S_t in computing the final confidence output S. Among the configurations tested, the best performance was achieved with equal weights of 0.5 for both modalities. In contrast, relying solely on either the text or speech confidence score resulted in significantly lower performance. This highlights the importance of both speech and text information in guiding the network, demonstrating that a balanced integration of these modalities yields the most effective results.

D. Visualization and Limitations

To further demonstrate the effectiveness of SpeechRefer, Figure 6 illustrates qualitative comparisons between our method and the baseline on 3DVG-Transformer. Transcription errors often lead the baseline to misidentify objects, while SpeechRefer overcomes these challenges to successfully identify the target objects.

Despite the promising performance, our method still has limitations. As shown in Figure 7, significant transcription errors, such as consecutive mistranscribed words caused by accents or background noise, can cause the textual features to deviate too much, leading the model to misidentify targets.

TABLE IX
ABLATION STUDY ON CONFIDENCE SCORE WEIGHTING WITH VARIOUS COMBINATIONS OF THE SPEECH SCORE S_s AND TRANSCRIBED TEXT SCORE S_t. FIVE DIFFERENT WEIGHT CONFIGURATIONS WERE TESTED. EVALUATED ON SPEECHREFER_CA DATASET.

S_s	S_t	Unique	Multiple	Overall
0.0	1.0	75.84/55.78	38.29/27.63	45.57/33.09
0.2	0.8	77.01/56.75	39.52/28.61	46.80/34.07
0.5	0.5	**77.13/57.19**	**39.74/28.73**	**47.00/34.25**
0.8	0.2	75.32/55.96	38.28/27.68	45.46/33.17
1.0	0.0	73.15/54.66	36.70/26.60	43.78/32.05



Fig. 7. Visualization of a failure case in the SpeechRefer_CA dataset. The green boxes and blue boxes represent the ground truth object and the wrong prediction of our SpeechRefer based on 3DVG-Transformer, respectively.

While text features improve overall accuracy, reliance on text is not ideal for speech-guided 3D visual grounding tasks. Future work will explore improved extraction and utilization of semantic information directly from speech, enabling reliable reasoning without text dependency.

V. Conclusion

We propose SpeechRefer, a novel 3D visual grounding framework that bridges the gap between speech and 3DVG in practical noisy environments. By leveraging speech features as a complement to transcribed text features derived from speech input, our method addresses the challenges of imperfect speech-to-text transcriptions, enhancing both performance and robustness. Our method integrates seamlessly with existing T-3DVG methods, as demonstrated through extensive experiments on our SpeechRefer and SpeechNr3D datasets, where it consistently outperforms baselines by a large margin. More importantly, our work provides additional insights distinct from previous 3D visual grounding methods—ambiguous speech in practical noisy environments introduces more possibilities than text, which enables the system to better infer user intent amidst a variety of possibilities and choices, rather than being constrained by definitive text. Overall, as the first attempt to localize a target object using noisy and ambiguous speech input, SpeechRefer shows promising results with room for improvement, paving the way for more intuitive 3D visual grounding. Future work will focus on deeper speech learning and better speech-visual integration to robust performance without text reliance.

References

[1] Z. Yuan, J. Ren, C.-M. Feng, H. Zhao, S. Cui, and Z. Li, "Visual programming for zero-shot open-vocabulary 3d visual grounding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20 623–20 633.

[2] M. Hamilton, A. Zisserman, J. R. Hershey, and W. T. Freeman, "Separating the" chirp" from the" chat": Self-supervised visual grounding of sound and language," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 117–13 127.

[3] S. Chen, P.-L. Guhur, C. Schmid, and I. Laptev, "History aware multimodal transformer for vision-and-language navigation," *Advances in neural information processing systems*, pp. 5834–5847, 2021.

[4] Z. Lu, Y. Pei, G. Wang, P. Li, Y. Yang, Y. Lei, and H. T. Shen, "Scaneru: Interactive 3d visual grounding based on embodied reference understanding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 4, 2024, pp. 3936–3944.

[5] Z. Li, Z. Li, J. Zhang, Y. Feng, and J. Zhou, "Bridging text and video: A universal multimodal transformer for audio-visual scene-aware dialog," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2476–2483, 2021.

[6] Z. Chen, H. Liu, and Y. Wang, "Dialogmcf: Multimodal context flow for audio visual scene-aware dialog," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 753–764, 2023.

[7] S. Parekh, S. Essid, A. Ozerov, N. Q. Duong, P. Pérez, and G. Richard, "Weakly supervised representation learning for audio-visual scene analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 416–428, 2019.

[8] D. Z. Chen, Q. Wu, M. Nießner, and A. X. Chang, "D 3 net: A unified speaker-listener architecture for 3d dense captioning and visual grounding," in *European Conference on Computer Vision*. Springer, 2022, pp. 487–505.

[9] Y. Wang, Y. Li, and S. Wang, "Gˆ 3-lq: Marrying hyperbolic alignment with explicit semantic-geometric modeling for 3d visual grounding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 917–13 926.

[10] C. Xu, Y. Han, R. Xu, L. Hui, J. Xie, and J. Yang, "Multi-attribute interactions matter for 3d visual grounding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 17 253–17 262.

[11] A. Jain, N. Gkanatsios, I. Mediratta, and K. Fragkiadaki, "Bottom up top down detection transformers for language grounding in images and point clouds," in *European Conference on Computer Vision*. Springer, 2022, pp. 417–433.

[12] D. Cai, L. Zhao, J. Zhang, L. Sheng, and D. Xu, "3djcg: A unified framework for joint dense captioning and visual grounding on 3d point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 464–16 473.

[13] X. Shi, Z. Wu, and S. Lee, "Viewpoint-aware visual grounding in 3d scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 14 056–14 065.

[14] Z. Chen, R. Hu, X. Chen, M. Nießner, and A. X. Chang, "Unit3d: A unified transformer for 3d dense captioning and visual grounding," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 18 109–18 119.

[15] Z. Guo, Y. Tang, R. Zhang, D. Wang, Z. Wang, B. Zhao, and X. Li, "Viewrefer: Grasp the multi-view knowledge for 3d visual grounding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 372–15 383.

[16] S. Huang, Y. Chen, J. Jia, and L. Wang, "Multi-view transformer for 3d visual grounding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15 524–15 533.

[17] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, 2023, pp. 28 492–28 518.

[18] M. Bain, J. Huh, T. Han, and A. Zisserman, "Whisperx: Time-accurate speech transcription of long-form audio," in *24th Annual Conference of the International Speech Communication Association, Interspeech 2023, Dublin, Ireland, August 20-24, 2023*, 2023, pp. 4489–4493.

[19] A. Mao, Z. Yang, W. Chen, R. Yi, and Y.-j. Liu, "Complete 3d relationships extraction modality alignment network for 3d dense captioning," *IEEE Transactions on Visualization and Computer Graphics*, 2023.

[20] S. Ye, D. Chen, S. Han, and J. Liao, "3d question answering," *IEEE Transactions on Visualization and Computer Graphics*, vol. 30, no. 3, pp. 1772–1786, 2022.

[21] B. Li, X. Wei, B. Liu, Z. He, J. Cao, and Y.-K. Lai, "Pose-aware 3d talking face synthesis using geometry-guided audio-vertices attention," *IEEE Transactions on Visualization and Computer Graphics*, 2024.

[22] D. Z. Chen, A. X. Chang, and M. Nießner, "Scanrefer: 3d object localization in RGB-D scans using natural language," in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XX*, 2020, pp. 202–221.

[23] P. Achlioptas, A. Abdelreheem, F. Xia, M. Elhoseiny, and L. J. Guibas, "Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes," in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, 2020, pp. 422–440.

[24] H. Liu, A. Lin, X. Han, L. Yang, Y. Yu, and S. Cui, "Refer-it-in-rgbd: A bottom-up approach for 3d visual grounding in rgbd images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6032–6041.

[25] Z. Yuan, X. Yan, Y. Liao, R. Zhang, S. Wang, Z. Li, and S. Cui, "Instancerefer: Cooperative holistic understanding for visual grounding

on point clouds through instance multi-level contextual referring," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1791–1800.

[26] Z. Yang, S. Zhang, L. Wang, and J. Luo, "Sat: 2d semantics assisted training for 3d visual grounding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1856–1866.

[27] P. Huang, H. Lee, H. Chen, and T. Liu, "Text-guided graph neural networks for referring 3d instance segmentation," in *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, 2021, pp. 1610–1618.

[28] Z. Zhu, X. Ma, Y. Chen, Z. Deng, S. Huang, and Q. Li, "3d-vista: Pre-trained transformer for 3d vision and text alignment," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2911–2921.

[29] Z. Jin, M. Hayat, Y. Yang, Y. Guo, and Y. Lei, "Context-aware alignment and mutual masking for 3d-language pre-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 984–10 994.

[30] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[31] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[32] Z. Liu, Z. Zhang, Y. Cao, H. Hu, and X. Tong, "Group-free 3d object detection via transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2949–2958.

[33] C. R. Qi, O. Litany, K. He, and L. J. Guibas, "Deep hough voting for 3d object detection in point clouds," in *proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9277–9286.

[34] O. Unal, C. Sakaridis, S. Saha, and L. Van Gool, "Four ways to improve verbo-visual fusion for dense 3d visual grounding," in *European Conference on Computer Vision (ECCV)*, October 2024.

[35] J. Luo, J. Fu, X. Kong, C. Gao, H. Ren, H. Shen, H. Xia, and S. Liu, "3d-sps: Single-stage 3d visual grounding via referred point progressive selection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 2022, pp. 16 433–16 442.

[36] Y. Wu, X. Cheng, R. Zhang, Z. Cheng, and J. Zhang, "EDA: explicit text-decoupling and dense alignment for 3d visual grounding," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, 2023, pp. 19 231–19 242.

[37] T. Zhang, S. He, T. Dai, Z. Wang, B. Chen, and S.-T. Xia, "Vision-language pre-training with object contrastive learning for 3d scene understanding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, pp. 7296–7304.

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

[39] A. Dai, A. X. Chang, M. Savva, M. Halber, T. A. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 2017, pp. 2432–2443.

[40] L. Zhao, D. Cai, L. Sheng, and D. Xu, "3dvg-transformer: Relation modeling for visual grounding on point clouds," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2928–2937.

[41] Y. Zhang, Z. Gong, and A. X. Chang, "Multi3drefer: Grounding text description to multiple 3d objects," in *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, 2023, p. 15179.

[42] C. Zhang, Z. Cai, X. Chen, F. Da, and S. Gai, "3d visual grounding-audio: 3d scene object detection based on audio," *Neurocomputing*, vol. 611, p. 128637, 2025.

[43] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "Imagebind one embedding space to bind them all," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, 2023, pp. 15 180–15 190.