

# GroundFlow: A Plug-in Module for Temporal Reasoning on 3D Point Cloud Sequential Grounding

Zijun Lin<sup>1,2</sup> Shuting He<sup>3</sup> Cheston Tan<sup>2</sup> Bihan Wen<sup>1</sup>

<sup>1</sup>Nanyang Technological University <sup>2</sup>Centre for Frontier AI Research, A\*STAR  
<sup>3</sup>MoE Key Laboratory of Interdisciplinary Research of Computation and Economics,  
 Shanghai University of Finance and Economics

## Abstract

Sequential grounding in 3D point clouds (SG3D) refers to locating sequences of objects by following text instructions for a daily activity with detailed steps. Current 3D visual grounding (3DVG) methods treat text instructions with multiple steps as a whole, without extracting useful temporal information from each step. However, the instructions in SG3D often contain pronouns such as “it”, “here” and “the same” to make language expressions concise. This requires grounding methods to understand the context and retrieve relevant information from previous steps to correctly locate object sequences. Due to the lack of an effective module for collecting related historical information, state-of-the-art 3DVG methods face significant challenges in adapting to the SG3D task. To fill this gap, we propose GroundFlow — a plug-in module for temporal reasoning on 3D point cloud sequential grounding. Firstly, we demonstrate that integrating GroundFlow improves the task accuracy of 3DVG baseline methods by a large margin (+7.5% and +10.2%) in the SG3D benchmark, even outperforming a 3D large language model pre-trained on various datasets. Furthermore, we selectively extract both short-term and long-term step information based on its relevance to the current instruction, enabling GroundFlow to take a comprehensive view of historical information and maintain its temporal understanding advantage as step counts increase. Overall, our work introduces temporal reasoning capabilities to existing 3DVG models and achieves state-of-the-art performance in the SG3D benchmark across five datasets.

## 1. Introduction

To understand and perform human instructions in the real world, robots should precisely identify the referred objects to complete complex tasks through natural language. 3D Visual Grounding (3DVG) [13, 42, 48, 49, 53] is a widely

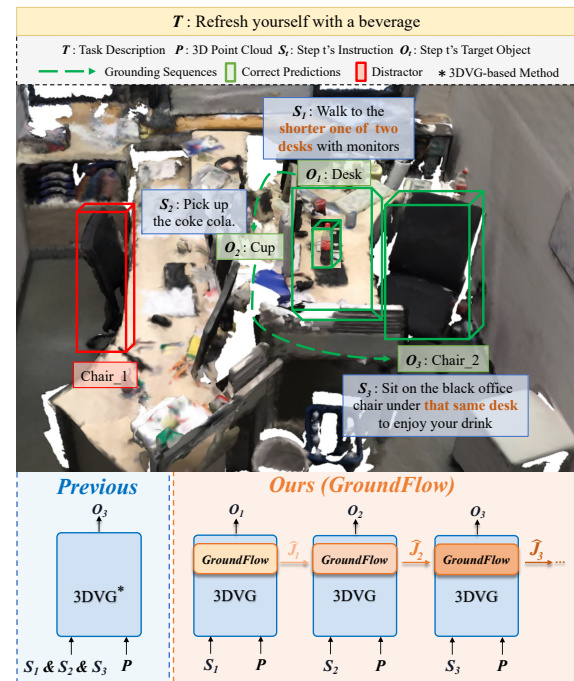


Figure 1. An example of SG3D task (above) and a comparison between previous visual grounding framework (bottom left) and our recurrent framework (bottom right) integrated with **GroundFlow** plug-in module to enhance temporal reasoning when localizing the target object in  $S_3$ . As shown, GroundFlow module’s output  $\hat{J}_t$  will be treated as input in the next step  $t + 1$ .

studied task that requires the agent to locate the target objects in 3D scenes based on a text instruction describing the object attributes in detail such as color, shape and spatial relationships. Recently, to advance grounding tasks toward more realistic applications, a new task called Sequential Grounding in 3D point clouds (SG3D) [52] has been introduced. Although this task shares similarities with 3DVG in grounding language information into the visual field, SG3D distinguishes itself by two characteristics — *sequen-*

*tial grounding* and *task-driven*. As illustrated in the example in Figure 1, unlike 3DVG task, which provides detailed object information and grounds the target in a single step, SG3D requires agents to locate a sequence of objects in the correct order over multiple steps based on task-oriented instructions. This mirrors how humans instruct robots to carry out daily activities using step-by-step instructions, making the SG3D task more applicable to real-life scenarios. Therefore, properly addressing the SG3D problem is crucial for realizing practical embodied AI applications.

Current visual grounding models [9, 20, 23, 27, 38] show promising performance on many 3DVG datasets, such as ScanRefer [6, 11], Sr3D, Nr3D [2], and Multi3DRefer [50], but have poor generalization on the sequential grounding task [52]. The main reason for the huge performance gap between the two tasks is that current 3DVG methods are not designed to reason over historical information. As shown in Figure 1, 3DVG methods typically process all text instructions as a single, undifferentiated input, which works for traditional visual grounding tasks but is not suitable for the task-driven nature of SG3D. In SG3D, instructions rely heavily on contextual pronouns such as “it”, “here”, “the same”, which are rarely seen in 3DVG tasks. These pronouns highlight the importance of effectively retrieving relevant information from past steps as the instructions progress. Simply concatenating all previous instructions to localize current step’s target makes it difficult for the model to differentiate important history information from irrelevant details. Additionally, there is a trend of incorporating large language models into the 3D field due to their excellent reasoning capabilities and vast pre-trained knowledge [22, 24, 54]. While 3D LLMs achieve state-of-the-art results in various 3D tasks, they still face significant difficulty adapting to the complex SG3D problem [52].

To address this, we design a recurrent framework, depicted in Figure 1. This framework sequentially takes each step instruction and processes only the current step instruction as input rather than handling all prior text instructions simultaneously. In addition, we propose GroundFlow module, which can be built on top of the existing 3DVG methods to perform temporal fusion with previous step embeddings, improving the comprehension of history instructions.

Furthermore, in the SG3D tasks, some step instructions depend on information from multiple previous steps. Take the task in Figure 1 as an example, there are multiple chairs in the scene and to accurately identify the “chair under that same desk” for the last step, the model needs to retrieve relevant information about the “shorter one of two desks” mentioned in the first step. Attending only to the immediate previous step is insufficient and the problem would be more pronounced as the number of steps increase in the task, requiring the model to gather information from a longer timeline. Thus, carefully extracting useful information from

both the short-term and long-term context is essential.

Humans tend to retrieve past information by first accessing recent activity and then integrating earlier details based on their relevance to the current context [44]. Accordingly, in the design of the GroundFlow module, we implement a memory structure mimicking how humans recall past information, selectively fusing short-term and long-term embeddings based on their correlation to the current embeddings. This approach allows the model to maintain perspective across time, efficiently leveraging both immediate and distant context as needed.

In summary, we make the following contributions:

- We propose the GroundFlow module with a recurrent framework, which can be integrated into previous 3DVG baselines and introduce important temporal reasoning capabilities to them. Evaluated on SG3D benchmark, this approach results in significant improvements of up to 6.3% and 10.2% in task accuracy, as well as 3.8% and 7.5% in step accuracy for dual-stream and query-based models, which are two classic 3DVG methods.
- We draw inspiration from how human retrieves the past information and effectively extract relevant short-term and long-term instructions based on current needs. This allows GroundFlow to improve the 3DVG baseline methods consistently as task steps increases.
- We achieve state-of-the-art performance on SG3D benchmark across five datasets without using large language model, outperforming 3D LLM by 2% in both step accuracy and task accuracy.

## 2. Related Work

### 2.1. 3D Visual Grounding

3D visual grounding (3DVG) aims to build connections between natural language and the 3D physical world [13, 19, 38, 42, 43, 45, 46, 49]. Specifically, the 3DVG task requires agents to localize related target object(s) based on a single text instruction. In this emerging domain, several methods have been proposed to enhance agent’s grounding capabilities in different aspects. For example, Zhang et al. [50] use M3DRef-CLIP, a CLIP-based [35] approach with contrastive learning, to ground text instructions to multiple 3D objects. Cai et al. [4] introduce a unified framework that jointly improves performance for dense captioning and visual grounding. Additionally, Guo et al. [17] leverage the large language model to enrich input texts and adopt inter-view attention to address view discrepancy issues in 3DVG. Although these methods focus on different aspects, they can be broadly categorized into dual-stream models [2, 4, 8, 17, 45], which process text and scene information separately before fusing them for visual grounding, and query-based models [25, 46, 57], which initialize a query that sequentially incorporates text and scene infor-

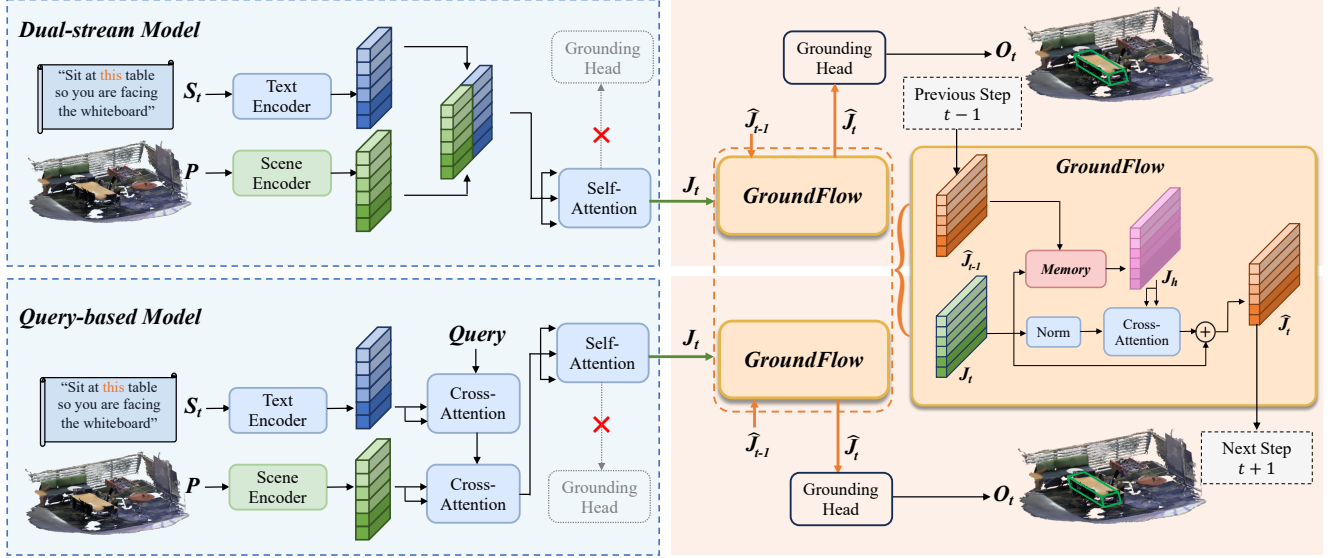


Figure 2. The overview of two 3DVG baseline models (blue background) integrated with our proposed plug-in temporal fusion module — **GroundFlow** (orange background). Unlike baseline methods that directly use  $J_t$  for grounding, GroundFlow integrates previous step information to produce  $\hat{J}_t$ , enhancing context-awareness in a recurrent framework.

mation before the grounding stage. However, these methods are primarily designed for object-centric 3DVG tasks and face significant challenges when applied to sequential grounding tasks. In this work, we introduce a plug-in module, GroundFlow, which could be easily implemented on top of both dual-stream and query-based models, allowing existing 3DVG methods to transition smoothly to sequential grounding tasks.

## 2.2. Temporal Context Learning

Understanding and reasoning about the temporal information allow intelligent agents to retrieve useful past information and make logical predictions about current or future scenarios. This ability is thus widely explored in many research fields [29, 32, 41, 55]. Recurrent Neural Networks (RNNs) [14] are commonly employed for temporal learning as they are designed to handle sequential data by maintaining a hidden state that captures dependencies over time. Chen et al. [7] adopt History Aware Multimodal Transformer (HAMT) for the task of Vision-and-language navigation (VLN) to learn the temporal dynamics across panoramas of the navigation history. In video retrieval task, Shao et al. [37] propose temporal context aggregation to incorporate temporal information between frame-level features to better capture long-range dependency. Furthermore, in spatial-temporal video grounding, Gu et al. [16] calculate the confidence score for each video frame in temporal refinement to filter out irrelevant instance context by a confidence threshold. In our work, we focus on referencing relevant short-term and long-term contextual information based

on current-step instructions and introduce a recurrent framework that dynamically adapts to the sequential grounding and task-driven nature of the complex SG3D task.

## 3. GroundFlow

In the SG3D task, we are provided with a 3D point cloud of an indoor scene  $\mathcal{P} \in \mathbb{R}^{N \times (3+C)}$  with ground-truth object masks, where  $N$  denotes the number of points and  $C$  represents the feature channels per point. The step-by-step instructions  $\mathcal{S} = \{s_1, \dots, s_n\}$  is given as textual input, where  $n$  denotes the number of steps. The model is required to predict the object sequences  $\mathcal{O} = \{o_1, \dots, o_n\}$  based on the 3D point cloud and step instructions, meaning that the model needs to learn the mapping  $f : (\mathcal{P}, \mathcal{S}) \rightarrow \mathcal{O}$ .

The challenges of the SG3D task, compared to other visual grounding tasks in 3D scenes, lie in grounding a sequence of targets given the implicit task-driven step instructions. Effectively retrieving relevant previous information is essential for solving this problem, yet current visual grounding methods lack a design for temporal fusion. In this section, we describe the technical details of our proposed plug-in module — GroundFlow. Different from the general reasoning capabilities of large language models, GroundFlow specifically aims to equip current 3DVG models with temporal reasoning abilities.

### 3.1. Recurrent Framework with Temporal Fusion

Existing approaches in 3D visual grounding treat the entire text as a single input, which works well for most 3DVG datasets. However, these methods struggle to generalize

in sequential grounding tasks, where the instructions are implicit and task-oriented, requiring the model to extract relevant information from previous steps conditioned on the current step instruction. Additionally, the average text length for a task in SG3D is 70.5 words, significantly longer than the 10 to 15 words typical in various 3DVG datasets. Simply concatenating all previous step instructions as input would introduce a substantial amount of redundant or irrelevant information, potentially degrading the model’s performance. Therefore, a recurrent framework with temporal fusion is proposed in GroundFlow to better capture the information at each step.

The proposed framework shown in Figure 2 is implemented on top of two types of 3DVG models — dual-stream and query-based. Technically, the dual-stream models [2, 4, 5, 17, 45] process the scene and text information separately and fuse them using a self-attention module, whereas the query-based models [8, 25, 46, 57] sequentially process the scene and text embeddings to perform cross-attention with an initialized learnable query. Although the mechanisms for processing the input differ between the two models, both of them generate the joint embeddings  $J_t$ , which are intended to collectively represent the scene and text information in current step  $t$ .

As shown in Figure 2, our proposed recurrent framework takes the current step’s joint embedding  $J_t$  and the previous step’s GroundFlow module output  $\hat{J}_{t-1}$  as input, where the joint embedding with the hat ‘^’ denotes the output after performing temporal fusion in GroundFlow. Consequently, the module’s output at the current step  $\hat{J}_t$  serves as input for the next step’s inference. Since the joint embedding  $J_t$  right before the grounding head is the only information required from 3DVG baselines to generate the corresponding  $\hat{J}_t$  with important temporal information, our proposed recurrent framework is expected to adapt to a broader range of 3DVG model categories.

Equation 1 shows how the module output  $\hat{J}_t$  grasps the prior relevant information from history trajectory. Cross-attention is used to inject critical history information  $J_h$  into the current step embeddings  $J_t$ . Notably,  $J_h$  comprehensively integrates both short-term and long-term information through a novel memory structure, which will be further discussed in section 3.2.

$$\begin{cases} \hat{J}_t = J_t + \text{softmax}\left(\frac{J_t J_h^T}{\sqrt{H}}\right) J_h, & t \in [2, n] \\ \hat{J}_t = J_t, & t = 1 \end{cases}, \quad (1)$$

where  $J_t \in \mathbb{R}^{T \times H}$  contains  $T$  tokens and  $H$  dimension for each token. After the cross-attention,  $\hat{J}_t$  captures both the current step’s scene-text relationship and the information from the relevant previous steps. For the first step, since there is no historical information,  $\hat{J}_t$  directly takes the value of  $J_t$ ’s similar to traditional 3DVG methods.

Afterwards,  $\hat{J}_t$  is fed into the grounding head to predict the target object  $O_t$  for current step  $t$ . The structure of the grounding head in 3DVG models typically consists of several MLP layers. Hence, our recurrent framework introduces the temporal reasoning capabilities for previous visual grounding models without hindering their predictions on the 3DVG task.

### 3.2. Memory Structure

Directly applying the immediate previous instruction  $\hat{J}_{t-1}$  to fuse temporally with the current embedding  $J_t$  only allows the output  $\hat{J}_t$  to contain relevant information from the immediate last step without the consideration of long-term information. Earlier information tends to fade as the step evolves, but there are cases where important information is mentioned in initial step instructions. How to effectively retrieve instructions for both short-term and long-term memory conditioned on current needs is challenging but crucial for GroundFlow to achieve more robust temporal reasoning.

To address this, we further implement a memory structure in GroundFlow. A detailed illustration of the memory component is shown in Figure 3. The memory takes the current step’s joint embedding  $J_t$  and the last step’s GroundFlow output  $\hat{J}_{t-1}$  as input. All previous step information is processed as long-term information within the memory structure to generate the history embedding  $J_h$ :

$$\begin{cases} J_h = \text{Memory}(\hat{J}_{t-1}, J_t), & t \in [3, n] \\ J_h = \hat{J}_{t-1}, & t = 2 \end{cases}. \quad (2)$$

It is worth noting that the memory component only starts functioning after the second step, as there is no long-term instruction prior to it. For each task step, tokens are padded to the maximum length, and a padding mask is created to help the model ignore padding tokens during processing.

In the memory structure, the recent GroundFlow output  $\hat{J}_{t-1}$  is regarded as short-term memory, while  $\hat{J}_m$  is considered long-term memory, which aggregates all the processed joint embeddings from the first step up to two steps prior:

$$\hat{J}_m = \sum_{t=1}^{t-2} \hat{J}_t, \quad (3)$$

where  $\hat{J}_m \in \mathbb{R}^{T \times H}$  contains  $T$  tokens, with each token having a dimension  $H$ , matching the shape of  $\hat{J}_t$ . As indicated by the dotted arrow in Figure 3,  $\hat{J}_m$  is updated (i.e.,  $\hat{J}_m = \hat{J}_m + \hat{J}_{t-1}$ ) for next step after completing the current memory inference.

Given that not all the previous step’s long-term information is relevant to the current prediction, naively using entire history embeddings indiscriminately is not an optimal solution. Therefore, we adopt cosine similarity to extract the most relevant information in  $\hat{J}_m$  based on the current joint

embedding  $J_t$ :

$$\mathbf{A}_{i,j} = \textcircled{S}(\hat{J}_m, J_t), \quad (4)$$

where  $\textcircled{S}$  denotes cosine similarity and  $\mathbf{A}_{i,j} \in \mathbb{R}^{T \times T}$ . Each row of  $\mathbf{A}_{i,j}$  indicates the importance of each token in  $\hat{J}_m$  to the current joint embedding  $J_t$ . The value of a specific row is expected to be higher if the corresponding tokens in  $\hat{J}_m$  contain relevant information for  $J_t$ .

After obtaining the importance score for each token in the long-term memory  $\hat{J}_m$ , we average all the scores to get a single value for each token for subsequent calculations. This gives each token in  $\hat{J}_m$  a relative importance score with respect to  $J_t$ :

$$\alpha = \frac{1}{T} \sum_{j=1}^T \mathbf{A}_{i,j}, \quad (5)$$

where  $\alpha \in \mathbb{R}^{T \times 1}$ . This approach allows the model to treat each history token differently based on its relevance to the current information.

The ultimate history embedding  $J_h$  combines both short-term and long-term information. We achieve this by summing the original short-term memory  $\hat{J}_{t-1}$  with the weighted long-term memory  $\hat{J}_m \odot \gamma(\alpha)$  together. The approach is inspired by the human memory retrieval process, which intuitively reflects on the most recent activity and the most relevant past experiences. The final equation for obtaining the history information  $J_h$  is shown as follows:

$$J_h = \hat{J}_{t-1} + \hat{J}_m \odot \gamma(\alpha), \quad (6)$$

where  $\odot$  is the Hadamard product (i.e., element-wise multiplication) and  $\gamma$  is the broadcasting function.

Subsequently, cross-attention is applied between the current joint embedding  $J_t$  and the history information  $J_h$ , as shown in Equation 1, yielding  $\hat{J}_t$  as the final representation before the grounding head. This representation incorporates the current step embedding with rich contextual understanding. Overall, the memory structure provides the model with a comprehensive view of the prior trajectory and extract relevant historical information, further enhancing the temporal reasoning capabilities of GroundFlow.

### 3.3. Training Objective

Following the SG3D benchmark [52], we use the same cross-entropy loss to optimize the dual-stream model and the query-based model. As defined in Equation 7, the loss compares the predicted object score  $f(\mathcal{P}, \mathcal{S})$  and the ground truth score  $\mathcal{O}$ . For the 3D LLM LEO [24], which is state-of-the-art method in SG3D benchmark, we follow the original approach. In addition to the loss of token predictions when pre-trained on other datasets, an extra cross-entropy loss is incorporated to fine-tune the model on SG3D data.

$$\mathcal{L}_{\text{grd}} = \mathbb{E}_{(\mathcal{P}, \mathcal{S}, \mathcal{O}) \sim \mathcal{D}} \text{CrossEntropy}(f(\mathcal{P}, \mathcal{S}), \mathcal{O}). \quad (7)$$

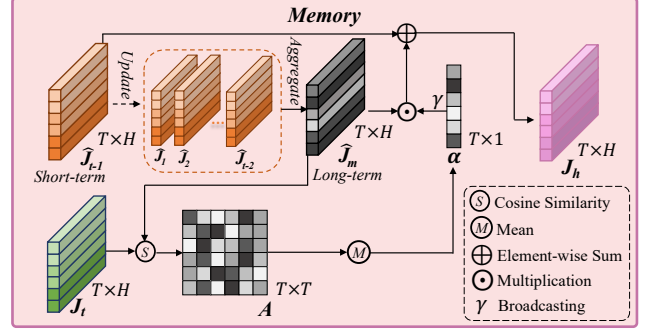


Figure 3. Detailed illustration of Memory component in GroundFlow, which enables the module to extract relevant information of both **short-term** ( $\hat{J}_{t-1}$ ) and **long-term** ( $\hat{J}_m$ ) effectively. The shape of each tensor is labeled adjacent to it. The dotted arrow represents the update to the long-term memory for the next step, which will be performed **after** the current step  $t$  is completed.

## 4. Experiments

### 4.1. Dataset and Evaluation Metrics

**Dataset.** GroundFlow is evaluated on the newly proposed benchmark — SG3D [52], which is specifically designed for the task of sequential grounding in 3D point clouds. The benchmark utilizes real-world scenes from the SceneVerse [26], incorporating indoor scans from 5 different datasets — ScanNet [11], 3RScan [40], MultiScan [31], ARKitScenes [3] and HM3D [36]. Detailed statistics of the SG3D benchmark are provided in Table 4.

**Evaluation Metrics.** As defined in SG3D benchmark [52], all models’ grounding performances is evaluated based on two key metrics: step accuracy (s-acc) and task accuracy (t-acc). Step accuracy is calculated as the average accuracy of the model in predicting the target object across all individual steps  $S$ , while task accuracy refers to the grounding accuracy over the total number of tasks  $T$ . For task accuracy, a sample is considered correct if the predicted sequence of objects for each step matches the ground-truth sequence.

### 4.2. Implementation Details

For a fair comparison, all methods follow the same hyperparameter settings. The models are trained for 50 epochs with batch size of 32 and evaluated on the last epoch using evaluation split of the SG3D benchmark. AdamW optimizer is employed with the learning rate of  $1e-4$  and weight decay of 0.05 for optimization, while  $\beta_1$ ,  $\beta_2$  are set to 0.9 and 0.999, respectively. Due to GPU memory constraints, the batch size for LEO is reduced to 16. All other training details for the baselines strictly follow the original paper’s settings. All 3DVG methods integrated with GroundFlow (only 22M #params) are able to deploy on single NVIDIA 24GB A5000 GPU. Kindly refer to Table 5 for comparisons of computational efficiency.

Model Type	Method	ScanNet		3RScan		MultiScan		ARKitScenes		HM3D		Overall	
		s-acc	t-acc	s-acc	t-acc	s-acc	t-acc	s-acc	t-acc	s-acc	t-acc	s-acc	t-acc
LLM-based	GPT4 + PointNet++ (Zero-shot) [52]	42.6	10.9	25.5	2.4	27.0	0.0	27.6	6.0	20.8	7.7	27.3	7.6
	LEO (3DLLM) [24]	61.2	25.7	55.8	16.0	52.7	7.6	69.6	41.5	61.5	35.7	62.8	34.1
Dual-stream	3D-VisTA [56]	60.1	24.7	52.7	13.5	47.6	7.0	68.4	37.8	57.5	30.6	60.3	28.8
	3D-VisTA+ <b>GroundFlow</b>	63.0	26.6	56.8	21.7	57.1	14.0	71.9	46.0	62.3	36.9	64.1	35.1
	MiKASA [5]	57.8	19.4	53.0	10.9	48.7	2.3	67.1	35.7	57.3	30.1	60.8	31.9
	MiKASA + <b>GroundFlow</b>	62.7	<b>28.9</b>	58.9	17.4	54.0	11.6	70.2	42.9	61.8	36.2	63.5	34.2
Query-based	PQ3D [57]	53.7	17.9	50.2	9.9	43.5	4.7	64.9	32.0	56.9	30.6	57.3	25.9
	PQ3D + <b>GroundFlow</b>	62.0	28.2	<b>60.1</b>	21.0	51.3	7.0	<b>73.0</b>	<b>48.1</b>	<b>63.6</b>	<b>38.0</b>	<b>64.8</b>	<b>36.1</b>
	Vil3DRel [8]	59.3	19.9	55.9	15.2	50.9	4.7	69.3	38.6	58.7	31.0	61.1	28.6
	Vil3DRel + <b>GroundFlow</b>	<b>63.1</b>	27.8	58.8	<b>22.5</b>	<b>57.6</b>	<b>20.9</b>	72.4	45.1	62.3	36.6	64.4	35.2

Table 1. Comparisons on SG3D benchmark across five datasets. The values for the metrics s-acc (step accuracy) and t-acc (task accuracy) are expressed as percentages (%). The methods with our proposed modules are highlighted in orange and the numbers in bold represent the best performance in given dataset and metrics.

### 4.3. Comparison on SG3D Benchmark

**Comparison with 3DVG methods.** We integrate GroundFlow into two classic 3DVG models – dual-stream and query-based. Multiple state-of-the-art 3DVG models (3D-VisTA [56] and MiKASA [5] for dual-stream, PQ3D [57] and Vil3DRel [8] for query-based) are selected to represent these two categories. To demonstrate the effectiveness of GroundFlow, the experimental results of these approaches before and after introducing GroundFlow will be compared. It is important to note that, beyond these 3DVG approaches, GroundFlow can serve as a plug-in module for a broader range of visual grounding methods, as it only requires the joint embedding right before the grounding head as input to perform temporal fusion within a recurrent framework.

Table 1 highlights that the previous 3DVG methods struggle to generalize to the SG3D benchmark. Their degraded performance is particularly reflected in their overall task accuracy, with three of the models are falling below 30%. These results are expected due to the absence of a temporal reasoning module. On the other hand, significant performance improvements can be observed when these models are integrated with GroundFlow, as shown in the rows highlighted in orange. Both the step accuracy and the task accuracy see substantial boosts in SG3D benchmark for all the 3DVG methods. Specifically, the overall performance in the SG3D benchmark of 3D-VisTA increases by 3.8% in step accuracy and 6.3% in task accuracy. For PQ3D, the improvements are even more pronounced, with gains of 7.5% and 10.2%, respectively and the task accuracy improves by 1.2 to 2.1 times across five datasets with the introduction of GroundFlow. These remarkable improvements validate our findings that effective temporal fusion is crucial for addressing the sequential grounding challenge.

**Comparison with LLM-based methods.** Following [52], an LLM with the 3D object classifier PointNet++ [34] is tested in a zero-shot manner. GPT-4 receives scene information, including each object’s ID, size, position, and the step instruction, while PointNet++, pre-trained on ScanNet, is used to predict the semantic labels for each ob-

ject. The results in the first row of the table demonstrate poor zero-shot performance. Furthermore, the state-of-the-art 3D large language model, LEO, after fine-tuning on the SG3D benchmark is also compared. It is pre-trained on an extensive range of 3D tasks, including object captioning [30, 51], object referring [1, 18, 48], 3D QA [12, 15, 47] and 3D navigation [28, 33], achieving top performance across various 3D point cloud benchmarks. In fine-tuning stage, LEO predicts a special  $[GRD]_t$  token at each step  $t$ , which is concatenated with object tokens and passed to the grounding head to predict the target object  $O_t$ . It is shown that LEO has decent results in SG3D benchmark and performs better than previous 3DVG methods due to its Internet-scale pre-trained knowledge.

However, the 3DVG methods combined with our proposed GroundFlow module outperform LEO across all five datasets, setting new state-of-the-art performance on SG3D benchmark. With the integration of GroundFlow, the step accuracy and task accuracy of the dual-stream model 3D-VisTA are 1.3% and 1% higher than LEO, respectively, while the query-based PQ3D surpasses LEO by 2% for both metrics. Therefore, LLM-based solutions excel at common-sense reasoning but may struggle with temporal knowledge, while GroundFlow focuses on modeling temporal dependencies in sequential grounding and selectively integrates historical information, leading to the better performance compared to 3D LLM.

Overall, GroundFlow, as a plug-in module, successfully enhances the capability of 3DVG models to handle temporal information, enabling them to seamlessly transition to sequential grounding tasks.

### 4.4. Ablation Study

**Comparisons of different temporal fusion methods.** To evaluate the effectiveness of our proposed temporal fusion module, GroundFlow, we compare its results on 3D-VisTA and PQ3D with other established approaches such as LSTM [21], GRU [10] and Transformer [39].

As shown in Table 2, GroundFlow demonstrates more effective improvements compared to classic temporal fusion

Models	Temporal Fusion Methods	s-acc	t-acc	$\Delta$ s-acc	$\Delta$ t-acc
3D-VisTA	LSTM	61.4	29.5	+1.1	+0.7
	GRU	62.0	28.8	+1.7	+0.0
	Transformer	62.9	33.5	+2.6	+4.7
	<b>GroundFlow</b>	<b>64.1</b>	<b>35.1</b>	<b>+3.8</b>	<b>+6.3</b>
PQ3D	LSTM	63.1	30.8	+5.8	+4.9
	GRU	63.8	30.7	+6.5	+4.8
	Transformer	63.4	33.6	+6.1	+7.7
	<b>GroundFlow</b>	<b>64.8</b>	<b>36.1</b>	<b>+7.5</b>	<b>+10.2</b>

Table 2. Comparison of temporal fusion methods for 3D-VisTA and PQ3D. The  $\Delta$  improvement in the last two columns is relative to the original 3DVG baselines.

methods. This advantage could stem from the limitations of existing methods: LSTM or GRU tends to forget long-term information. While transformers use attention mechanisms to capture long-range dependencies, they treat each step indiscriminately. As a result, they may still lose crucial historical information that is important at current step  $t$  but not relevant at previous step  $t - 1$ . Since previous step embeddings do not attend to this lost information, it cannot be carried forward to subsequent steps, even if it is essential for the current prediction. This issue is common in SG3D, where the current instruction often refers to objects from multiple past steps while intermediate instructions are irrelevant, as illustrated in the second example of Figure 5.

To address these limitations, the memory component in GroundFlow computes similarity scores to selectively retrieve and integrate context-specific past information based on its relevance to the current step, leading to superior performance enhancements, especially in task accuracy.

**Improvements in t-acc over task steps.** In the SG3D benchmark, the step count of a task ranges from 2 to 10. To investigate whether GroundFlow can consistently improve task accuracy of the 3DVG methods in more challenging scenarios with a high number of steps, we create seven subsets from the evaluation split. These subsets contain tasks with step counts ranging from 2 to 7. Tasks with step counts greater than 7 are combined into one subset, as creating separate subsets for them would result in fewer than 100 tasks, which may not be sufficient to reliably reflect performance.

As shown in Figure 4, the introduction of GroundFlow improves the performance of all subsets with different step counts, with the query-based method PQ3D showing greater performance gains compared to the dual-stream method 3D-VisTA. Notably, task accuracy improvements for PQ3D exceed 10% across all subsets with step counts greater than six, further highlighting GroundFlow’s effectiveness in retrieving long-term information in the challenging high-step-count subsets.

**Impact of different memory configurations.** In GroundFlow, the history embeddings  $J_h$  accumulate rich information from both short-term and long-term sources by summing them up, as described in Equation 6. To explore the impact of different memory configurations, we compare the

Models	Short-term	Long-term	s-acc	t-acc	$\Delta$ s-acc	$\Delta$ t-acc
3D-VisTA	$\hat{J}_{t-1}$	N.A.	63.4	34.2	+3.1	+5.4
	N.A.	$\hat{J}_{t-1} + \hat{J}_m$	64.1	34.6	+3.8	+5.8
	$J_{t-1}$	$\hat{J}_m$	64.0	34.7	+3.7	+5.9
	$\hat{J}_{t-1}$	$J_m$	64.1	34.6	+3.8	+5.8
	$\hat{J}_{t-1}$	$\hat{J}_m$	<b>64.1</b>	<b>35.1</b>	<b>+3.8</b>	<b>+6.3</b>
PQ3D	$\hat{J}_{t-1}$	N.A.	64.3	35.7	+7.0	+9.8
	N.A.	$\hat{J}_{t-1} + \hat{J}_m$	63.6	34.5	+6.3	+8.6
	$J_{t-1}$	$\hat{J}_m$	63.3	33.3	+6.0	+7.4
	$\hat{J}_{t-1}$	$J_m$	63.4	34.6	+6.1	+8.7
	$\hat{J}_{t-1}$	$\hat{J}_m$	<b>64.8</b>	<b>36.1</b>	<b>+7.5</b>	<b>+10.2</b>

Table 3. Comparison of different short-term and long-term memory settings for 3D-VisTA and PQ3D. Aligned with section 3.2,  $\hat{J}_m$  denotes the aggregation of the GroundFlow output  $\hat{J}_t$  from the first step to two steps prior while  $J_m$  denotes the summation of original joint embeddings  $J_t$ , (i.e.,  $\hat{J}_m = \sum_{t=1}^{t-2} \hat{J}_t$ ,  $J_m = \sum_{t=1}^{t-2} J_t$ ).

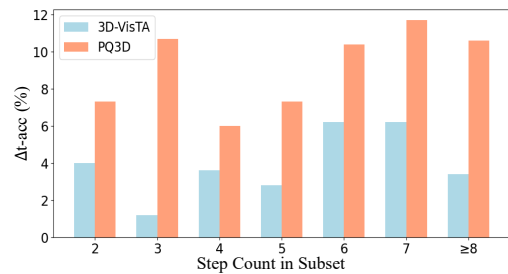


Figure 4. Improvements after GroundFlow module is integrated in terms of task accuracy of 3D-VisTA and PQ3D across different step count subsets.

various settings of short-term and long-term information within the memory component. In Table 3, the performance without one of the memory parts is presented in the first and second rows. Additionally, we assess whether embeddings before or after the application of GroundFlow yield better results, with their outcomes shown in the third and fourth rows. The final configuration adopted in GroundFlow, highlighted in orange, is shown in the fifth row for both models.

Results in Table 3 show that both short-term and long-term memory are crucial for effective memory retrieval, as the absence of either term results in a performance downgrade. Furthermore, the memory component using the joint embeddings after performing temporal fusion in GroundFlow yields better results, demonstrating that the temporal reasoning capabilities provided by GroundFlow enable the memory structure to capture important temporal information. Therefore, we use  $\hat{J}_{t-1}$  as short-term memory and  $\hat{J}_m$  as long-term memory in our final implementation, leading to the best performance.

#### 4.5. Qualitative Visualization

Figure 5 shows the visualization results of one of our baseline methods PQ3D, alongside its integration with GroundFlow. In the first example, to localize the correct target for

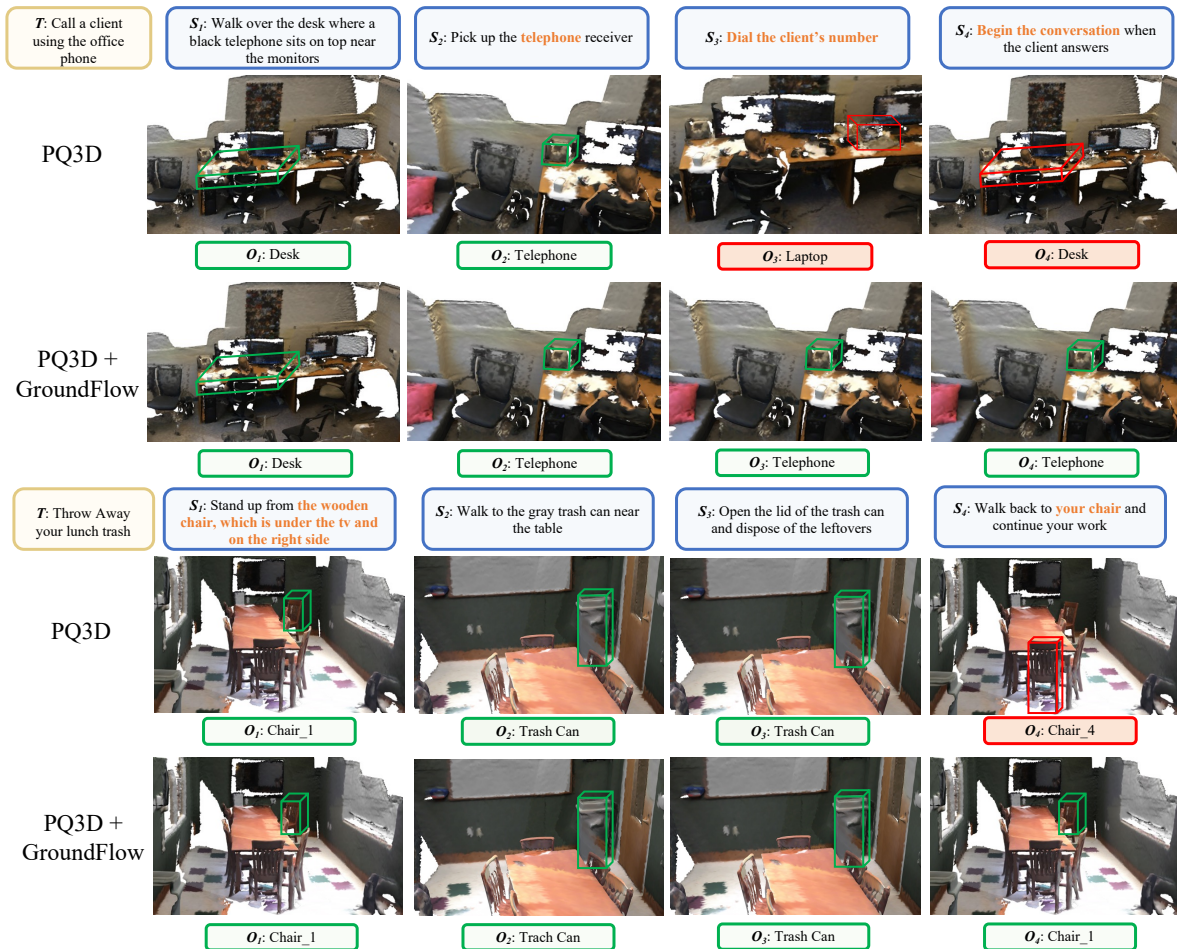


Figure 5. Visualization results from PQ3D and PQ3D+GroundFlow.  $T$  represents the task description,  $S_t$  and  $O_t$  denote the step instruction and corresponding referred target object in step  $t$ . Red are wrong predictions and green are correct predictions.

the last two step instructions “Dial the client’s number” and “Begin the conversation”, models need to understand the context that the telephone has been picked up in previous steps. It is shown that PQ3D fails to correctly choose the target “Telephone”, while PQ3D+GroundFlow makes the correct predictions of “Telephone” for both steps. Therefore, with the help of our proposed GroundFlow module, 3DVG methods can capture contextual information and exhibit essential temporal reasoning capabilities, generalizing to challenging task-oriented sequential grounding tasks.

In the second example, there are multiple chairs in the scene and the last step instruction, “Walk back to your chair”, refers to the chair that is “under the tv and on the right side” mentioned in the first step. This requires the model to retrieve earlier information beyond just short-term memory. Unlike PQ3D, which incorrectly selects another chair, PQ3D integrated GroundFlow consistently identifies the correct chair referenced in the first step. These results highlight that the memory component in GroundFlow en-

ables the model to retain important context over time, allowing it to accurately retrieve and apply information from both recent and earlier steps. Kindly refer to Figure 8 for more qualitative comparisons.

## 5. Conclusion

We propose GroundFlow, a plug-and-play module with a recurrent framework that introduces temporal reasoning capabilities to existing 3D visual grounding models, enabling them to seamlessly adapt to complex and challenging sequential grounding tasks. The memory structure within the GroundFlow module effectively retrieves relevant short-term and long-term historical information based on current needs. The extensive experiments show that, with the integration of GroundFlow, the performance of both classic types of 3DVG models — dual-stream and query-based, improve significantly. Their results surpass the 3D large language model pre-trained in various 3D tasks, achieving consistent state-of-the-art performance on SG3D benchmark.

## Acknowledgements

Zijun Lin is supported by the Agency for Science, Technology and Research (A\*STAR) Computing and Information Science (ACIS) Scholarship. Shuting He is sponsored by Shanghai Pujiang Programme 24PJD030 and Natural Science Foundation of Shanghai 25ZR1402138. This research is supported in part by A\*STAR SERC CRF funding to C.T., and in part by A\*STAR IAF-ICP Programme I2501E0041 and the Schaeffler-NTU Corporate Lab (SHARE@NTU). The work was done at Rapid-Rich Object Search (ROSE) Lab, School of Electrical & Electronic Engineering, Nanyang Technological University.

## 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*, pages 3941–3950, 2022. 6
- [2] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 422–440. Springer, 2020. 2, 4
- [3] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021. 5
- [4] Daigang Cai, Lichen Zhao, Jing Zhang, Lu Sheng, and Dong Xu. 3djcg: A unified framework for joint dense captioning and visual grounding on 3d point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16464–16473, 2022. 2, 4
- [5] Chun-Peng Chang, Shaoxiang Wang, Alain Pagani, and Didier Stricker. Mikasa: Multi-key-anchor & scene-aware transformer for 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14131–14140, 2024. 4, 6
- [6] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European conference on computer vision*, pages 202–221. Springer, 2020. 2
- [7] Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. History aware multimodal transformer for vision-and-language navigation. *Advances in neural information processing systems*, 34:5834–5847, 2021. 3
- [8] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Language conditioned spatial relation reasoning for 3d object grounding. *Advances in neural information processing systems*, 35:20522–20535, 2022. 2, 4, 6
- [9] Zhenyu Chen, Ronghang Hu, Xinlei Chen, Matthias Nießner, and Angel X Chang. Unit3d: A unified transformer for 3d dense captioning and visual grounding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 18109–18119, 2023. 2
- [10] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 6
- [11] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 2, 5
- [12] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–10, 2018. 6
- [13] Chun Feng, Joy Hsu, Weiyu Liu, and Jiajun Wu. Naturally supervised 3d visual grounding with language-regularized concept learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13269–13278, 2024. 1, 2
- [14] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*. MIT press Cambridge, 2016. 3
- [15] Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. Iqa: Visual question answering in interactive environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4089–4098, 2018. 6
- [16] Xin Gu, Heng Fan, Yan Huang, Tiejian Luo, and Libo Zhang. Context-guided spatio-temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18330–18339, 2024. 3
- [17] 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*, pages 15372–15383, 2023. 2, 4
- [18] 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*, pages 2344–2352, 2021. 6
- [19] Shuting He and Henghui Ding. Refmask3d: Language-guided transformer for 3d referring segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8316–8325, 2024. 2
- [20] Shuting He, Henghui Ding, Xudong Jiang, and Bihan Wen. Segpoint: Segment any point cloud via large language model. In *European Conference on Computer Vision*, pages 349–367. Springer, 2024. 2
- [21] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 6
- [22] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *Advances*

- in *Neural Information Processing Systems*, 36:20482–20494, 2023. 2
- [23] Joy Hsu, Jiayuan Mao, and Jiajun Wu. Ns3d: Neuro-symbolic grounding of 3d objects and relations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2614–2623, 2023. 2
- [24] Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024. 2, 5, 6
- [25] Ayush Jain, Nikolaos Gkanatsios, Ishita Mediratta, and Katerina Fragkiadaki. Bottom up top down detection transformers for language grounding in images and point clouds. In *European Conference on Computer Vision*, pages 417–433. Springer, 2022. 2, 4
- [26] Baoxiong Jia, Yixin Chen, Huangyue Yu, Yan Wang, Xuesong Niu, Tengyu Liu, Qing Li, and Siyuan Huang. Sceneverse: Scaling 3d vision-language learning for grounded scene understanding. In *European Conference on Computer Vision*, pages 289–310. Springer, 2024. 5
- [27] Zhao Jin, Munawar Hayat, Yuwei Yang, Yulan Guo, and Yinjie 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*, pages 10984–10994, 2023. 2
- [28] Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, pages 104–120. Springer, 2020. 6
- [29] Lili Liang, Guanglu Sun, Jin Qiu, and Lizhong Zhang. Neural-symbolic videoqa: Learning compositional spatio-temporal reasoning for real-world video question answering. *arXiv preprint arXiv:2404.04007*, 2024. 3
- [30] Tiange Luo, Chris Rockwell, Honglak Lee, and Justin Johnson. Scalable 3d captioning with pretrained models. *arXiv preprint arXiv:2306.07279*, 2023. 6
- [31] Yongsan Mao, Yiming Zhang, Hanxiao Jiang, Angel Chang, and Manolis Savva. Multiscan: Scalable rgbd scanning for 3d environments with articulated objects. *Advances in neural information processing systems*, 35:9058–9071, 2022. 5
- [32] Yue Meng, Rameswar Panda, Chung-Ching Lin, Prasanna Sattigeri, Leonid Karlinsky, Kate Saenko, Aude Oliva, and Rogerio Feris. Adafuse: Adaptive temporal fusion network for efficient action recognition. *arXiv preprint arXiv:2102.05775*, 2021. 3
- [33] Steven D Morad, Roberto Mecca, Rudra PK Poudel, Stephan Liwicki, and Roberto Cipolla. Embodied visual navigation with automatic curriculum learning in real environments. *IEEE Robotics and Automation Letters*, 6(2):683–690, 2021. 6
- [34] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 6
- [35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2
- [36] Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-matterport 3D dataset (HM3d): 1000 large-scale 3D environments for embodied AI. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. 5
- [37] Jie Shao, Xin Wen, Bingchen Zhao, and Xiangyang Xue. Temporal context aggregation for video retrieval with contrastive learning. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3268–3278, 2021. 3
- [38] Xiangxi Shi, Zhonghua Wu, and Stefan Lee. Aware visual grounding in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14056–14065, 2024. 2
- [39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 6
- [40] Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Nießner. Rio: 3d object instance re-localization in changing indoor environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7658–7667, 2019. 5
- [41] Jianren Wang, Haiming Gang, Siddarth Ancha, Yi-Ting Chen, and David Held. Semi-supervised 3d object detection via temporal graph neural networks. In *2021 International conference on 3D Vision (3DV)*, pages 413–422. IEEE, 2021. 3
- [42] Yuan Wang, Yali Li, and Shengjin Wang. G<sup>3</sup>-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*, pages 13917–13926, 2024. 1, 2
- [43] Zehan Wang, Haifeng Huang, Yang Zhao, Linjun Li, Xize Cheng, Yichen Zhu, Aoxiong Yin, and Zhou Zhao. Distilling coarse-to-fine semantic matching knowledge for weakly supervised 3d visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2662–2671, 2023. 2
- [44] Xiaoqian Xiao, Qi Dong, Jiahong Gao, Weiwei Men, Russell A Poldrack, and Gui Xue. Transformed neural pattern reinstatement during episodic memory retrieval. *Journal of Neuroscience*, 37(11):2986–2998, 2017. 2
- [45] Can Xu, Yuehui Han, Rui Xu, Le Hui, Jin Xie, and Jian Yang. Multi-attribute interactions matter for 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17253–17262, 2024. 2, 4

- [46] Li Yang, Ziqi Zhang, Zhongang Qi, Yan Xu, Wei Liu, Ying Shan, Bing Li, Weiping Yang, Peng Li, Yan Wang, et al. Exploiting contextual objects and relations for 3d visual grounding. *Advances in Neural Information Processing Systems*, 36:49542–49554, 2023. [2](#), [4](#)
- [47] Licheng Yu, Xinlei Chen, Georgia Gkioxari, Mohit Bansal, Tamara L Berg, and Dhruv Batra. Multi-target embodied question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6309–6318, 2019. [6](#)
- [48] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1791–1800, 2021. [1](#), [6](#)
- [49] Zhihao Yuan, Jinke Ren, Chun-Mei Feng, Hengshuang Zhao, Shuguang Cui, and Zhen Li. Visual programming for zero-shot open-vocabulary 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20623–20633, 2024. [1](#), [2](#)
- [50] 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*, pages 15225–15236, 2023. [2](#)
- [51] Ziqi Zhang, Yaya Shi, Chunfen Yuan, Bing Li, Peijin Wang, Weiming Hu, and Zhengjun Zha. Object relational graph with teacher-recommended learning for video captioning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13275–13285, 2020. [6](#)
- [52] Zhuofan Zhang, Ziyu Zhu, Junhao Li, Pengxiang Li, Tianxu Wang, Tengyu Liu, Xiaojian Ma, Yixin Chen, Baoxiong Jia, Siyuan Huang, and Qing Li. Task-oriented sequential grounding and navigation in 3d scenes, 2025. [1](#), [2](#), [5](#), [6](#)
- [53] Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3dvg-transformer: Relation modeling for visual grounding on point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2928–2937, 2021. [1](#)
- [54] Duo Zheng, Shijia Huang, Lin Zhao, Yiwu Zhong, and Liwei Wang. Towards learning a generalist model for embodied navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13624–13634, 2024. [2](#)
- [55] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 803–818, 2018. [3](#)
- [56] Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3d-vista: Pre-trained transformer for 3d vision and text alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2911–2921, 2023. [6](#)
- [57] Ziyu Zhu, Zhuofan Zhang, Xiaojian Ma, Xuesong Niu, Yixin Chen, Baoxiong Jia, Zhidong Deng, Siyuan Huang, and Qing Li. Unifying 3d vision-language understanding via promptable queries. In *European Conference on Computer Vision*, pages 188–206. Springer, 2024. [2](#), [4](#), [6](#)