DenseScan: Advancing 3D Scene Understanding with 2D Dense Annotation

Zirui Wang[1] **Tao Zhang**[2]
[1]University of Illinois at Urbana-Champaign [2]Wuhan University
ziruiw3@illinois.edu zhang_tao@whu.edu.cn

Abstract

3D understanding is a key capability for real-world AI assistance. High-quality data plays an important role in driving the development of the 3D understanding community. Current 3D scene understanding datasets often provide geometric and instance-level information, yet they lack the rich semantic annotations necessary for nuanced visual-language tasks. In this work, we introduce DenseScan, a novel dataset with detailed multi-level descriptions generated by an automated pipeline leveraging multi-view 2D images and multimodal large language models (MLLMs). Our approach enables dense captioning of scene elements, ensuring comprehensive object-level descriptions that capture context-sensitive details. Furthermore, we extend these annotations through scenario-based question generation, producing high-level queries that integrate object properties, spatial relationships, and scene context. By coupling geometric detail with semantic richness, DenseScan broadens the range of downstream tasks, from detailed visual-language navigation to interactive question answering. Experimental results demonstrate that our method significantly enhances object-level understanding and question-answering performance in 3D environments compared to traditional annotation pipelines. We release both the annotated dataset and our annotation pipeline to facilitate future research and applications in robotics, augmented reality, and beyond. Through DenseScan, we aim to catalyze new avenues in 3D scene understanding, allowing researchers and practitioners to tackle the complexities of real-world environments with richer, more contextually aware annotations.

1 Introduction

Understanding 3D world is a crucial in various applications such as robotics and autonomous driving, where agents are expected to know about the surrounding environment and carry out complex tasks based on human's instructions. However, current 3D MLLMs [15, 42, 41, 14] perform poorly in these practical applications and there are still significant gaps between their capabilities and the requirements of real-world applications. The scarcity of high-quality 3D multimodal data severely limits the development of 3D MLLMs.

Although there are many multimodal datasets [9, 3, 19, 41] in the current 3D community available for training 3D MLLMs, these datasets such as ScanRefer [9], ReferIt3D [3], and Reason3D [19] only focus on localizing 3D objects based on short, direct or implicit text referring. Besides, recent advances in 3D Question Answering datasets [28, 18, 6] emphasize integrating point clouds, images, and textual data to form rich multimodal representations, but LLM-based QA-pairs generation often introduce hallucinated information and contextual misalignment. As a result, they cannot effectively inject rich 3D knowledge into models during training, nor can they effectively benchmark how far current MLLMs are from real-world applications. Unlike prior referring setups optimized for short

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Workshop on Space in Vision, Language, and Embodied AI.

disambiguation, our scenario-driven segmentation requires models to parse multi-sentence context, inter-object relations beyond the target, and functional cues to produce precise masks.

In the past, 3D multimodal data relied on manual annotation, which was extremely costly and made it difficult to produce a rich diversity and massive quantity of annotations to drive rapid community development. Some recent works [19, 41, 14] have attempted to design automated pipelines to generate 3D multimodal data, such as using GPT-like models to refine existing textual descriptions rather than generating new ones from direct scene analysis. These models, though effective in producing fluent text, lack a deep understanding of spatial relationships, object affordances, and occlusions, as they operate in a predominantly text-driven space without strong visual grounding.

Currently, 2D MLLMs such as GPT-4o [2], Gemini, Qwen2.5 VL [39], and InternVL 2.5 [10] demonstrate powerful capabilities, including fine-grained understanding of visual signals and robust reasoning abilities. The responses generated by these models even far exceed the quality of data produced by most skilled human annotators. However, the quality of 3D multimodal datasets is significantly lower than that in the 2D community, and the intelligence level of 3D MLLMs lags far behind 2D MLLMs.

In this paper, we aim to design a systematic and efficient pipeline that leverages 2D MLLMs to produce multi-level (including object-level and scene-level) high-quality 3D data, in order to further advance the 3D MLLM community. Specificly, we propose an automated pipeline to annotate ScanNet [11] with detailed scene-level descriptions.

Our automated annotation system is organized into four distinct stages: 1) we initiate the process with multi-level object captioning, where state-of-the-art 2D MLLMs generate detailed captions for objects in multi-view 2D images, effectively capturing rich semantic information; 2) These initial captions undergo a rigorous phase of filtering and consistency checking, along with style adaptation, to transform them into natural-flow referring expressions that are coherent and contextually appropriate; 3) leverages multi-level object captions, we generate scenario-driven questions by prompting the MLLM annotator to draw on its world knowledge to suggest multiple frequent events or interactions that are typical for similar scenes. It is important to note that these event cues are not necessarily accurate depictions of what is happening in the specific scene, but rather informed hypotheses based on general knowledge; they are then seamlessly combined with object descriptions to form complex scenario-driven referring expressions that reflect intricate contextual relationships; 4) the generated annotations are subjected to LLM verification, augmented by human-assisted filtering, to ensure their accuracy and establish robust validation benchmarks. The resulting dataset, DenseScan, comprises 1,513 scenes and 20,113 object instances from ScanNet [11], featuring 38,765 dense referring expressions and 37,483 scenario-based questions. Compared to existing referring datasets, DenseScan stands out by offering significantly richer linguistic diversity and more precise object annotations while maintaining a comparable scale.

Along With the DenseScan dataset, we introduce a novel task called scenario-driven segmentation, which requires models to locate objects and generate segmentation masks based on detailed, long-text descriptions. Compared to traditional referring segmentation, which typically involves short referring expressions tied to specific objects, scenario-driven segmentation presents a greater challenge by incorporating complex contextual cues, spatial relationships, and functional attributes within a scene. This task pushes models to move beyond simple object identification and short referring expression learning, requiring deeper semantic understanding and reasoning about object interactions within diverse environments.

To tackle the challenge of complex scenario-driven reasoning in 3D point clouds, we propose Dense3D, a novel LLM-based framework that uniquely embeds multi-modal information rather than relying solely on 2D image inputs. Our model consists of three key components: a Point Encoder, which applies voxelization and a Sparse 3D U-Net backbone to extract point-wise features, further refined through a superpoint pooling layer for computational efficiency; Multi-Modal LLMs, which integrate both 2D multi-view images and 3D point cloud representations (from depth map and camera paramters) to enhance spatial and semantic understanding; and a 3D Query Decoder, a transformer-based module that translates high-level textual cues into precise segmentation masks. By fusing 3D point clouds, 2D multi-views images, and textual descriptions—Dense3D achieves deeper contextual reasoning and more precise segmentation, setting a new standard for multi-modal 3D scene understanding.

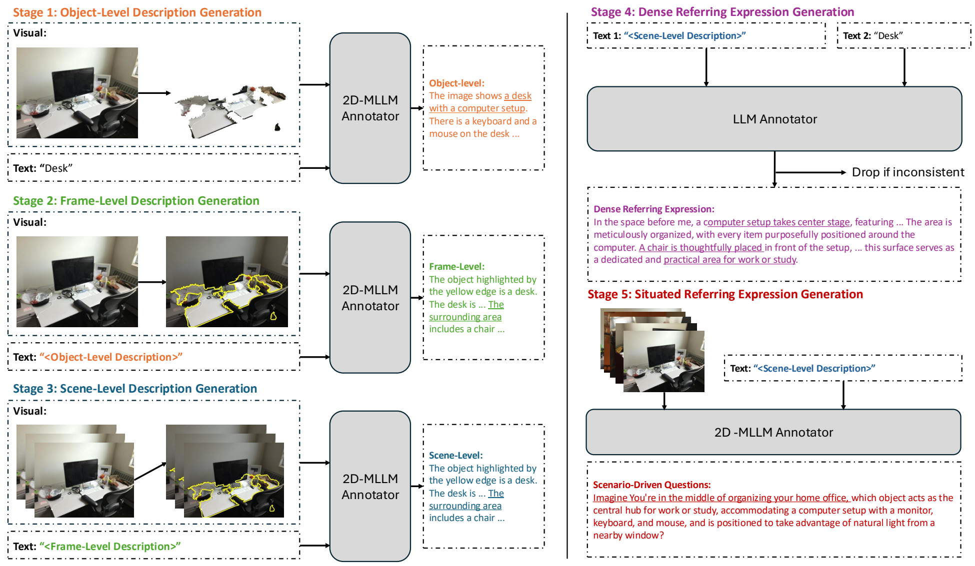

Figure 1: **Overview of DenseScan Data Generation Pipeline**. *Stage 1:* crop the target object and generate object-level description; *Stage 2:* highlight target object in a single frame and generate frame-level description to capture spacial dependencies with the surroundings; *Stage 3:* use multiple frames with target object highlighted to compose scene-level descriptions; *Stage 4:* Raw scene-level description need to go through a LLM for consistency checking, and in-consistent description will be eliminated; *Stage 5:* Adopt MLLM annotator to generate scenario-driven questions and verified by LLM and human before release to benchmark.

In summary, our contributions are threefold:

1. **Automated Annotation System.** We have developed an automated annotation system that leverages state-of-the-art 2D MLLMs to generate reliable and high-quality 3D multimodal data. Based on this pipeline, we have created the DenseScan dataset, which features rich, multi-level annotations—including detailed object-level captions and comprehensive scene-level descriptions—that better capture complex spatial relationships and contextual cues.
2. **Scenario-Driven Segmentation.** We have manually reviewed and refined a more challenging benchmark that incorporates a novel task: *scenario-driven segmentation*, with requires models to generate precise segmentation masks guided by long, context-rich descriptions, pushing the boundaries of traditional referring segmentation and better reflecting real-world application demands.
3. **Dense3D Framework.** We have introduced Dense3D, a 3D MLLM framework that effectively fuses 3D point cloud data, 2D multi-view images, and textual information. Through a Point Encoder, a Multi-Modal LLMs, and a 3D Query Decoder, we achieve deeper contextual reasoning and more accurate segmentation, thus taking a step further in multi-modal 3D scene understanding.

2 Related Work

Datasets for 3D Scene Understanding. Advances in 3D computer vision are deeply intertwined with the availability of large-scale, high-quality datasets[29, 5, 8, 7, 13]. In the realm of 3D instance segmentation, large-scale real-world datasets are limited. Early benchmarks like ScanNet[11] and S3DIS[5] provided a solid foundation by offering detailed scans of real-world indoor environments captured using RGB-D cameras or Matterport systems. Subsequently, researchers have increasingly built upon these public datasets by incorporating diverse, richly detailed annotations [9, 3, 40, 19, 21, 41]. Notably, ScanRefer[9] provides a set of natural language referring expressions for objects in indoor 3D scenes. Subsequent efforts, such as ReferIt3D[3], refines the annotation process through more fine-grained object categorization and the inclusion of multiple object instances per scene, and

Multi3DRefer[40], expands the task scope by accommodating descriptions that reference zero, one, or multiple objects, which better mirrors the complexities of real-world scenarios. More recently, several efforts have pushed the boundary even further by designing implicit object descriptions that require higher-level reasoning, such as Reason3D[19], ScanReason[41], Instruct3D[14] and ReasonSeg3D[21]. Despite these advancements, our work develop a denser and longer text description for objects, capturing detailed semantic attributes, spatial context, and subtle visual cues that are often overlooked in earlier datasets. Besides, while previous reasoning-based questions tend to be abstract and focus on broad, often decontextualized logic or inference, scenario-like questions are tightly grounded in the spatial and semantic context of a specific scene. Together, the new dataset provide a more comprehensive framework for advancing 3D scene understanding and segmentation.

3D Multi-Modal Large Language Models. Inspired by the powerful reasoning abilities of Large Language Models, researchers have injected LLMs into vision domain[4, 26, 24, 43, 27], namely multi-modal large language models (MLLMs). Upon prior works [26, 4] at integrating visual context with language models, recent efforts has been made to develop MLLMs that seamlessly integrating diverse capabilities for instruction-based tasks. For instance, VisionLLM[35] offers a vision-centric interface through instruction tuning, though it doesn't support advanced reasoning. Another novel line of research has emerged with the debut of LISA[23], alongside several subsequent studies [31, 32]that have significantly advanced the field of multimodal large language models (MLLMs) in 2D space. Building on the advancements in 2D multimodal language models, researchers are now venturing into the realm of 3D MLLMs, aiming to enhance spatial understanding and unlock novel applications in complex volumetric environments. 3D-LLM[16] leverage 2D foundation models to inject 3D spatial understanding into language models. PointLLM[38] build on top of LLaVA [27] to train LLM with 3D point cloud representations.

Grounding MLLMs. Recent work has aimed to fully leverage the reasoning capabilities of large language models for addressing 3D downstream tasks such as segmentation and localization. In particular, SegPoint[14] and Reason3D[19] focus on interpreting complex textual descriptions of individual objects, whereas ReasonSeg3D[21] targets multi-object referring expression segmentation with accompanying text explanations. ScanReason[41] developed comprehensive and hierarchical 3D reasoning grounding benchmark that assess fundamental-understanding of 3D world to high-level reasoning skills. Our work distinguishes itself from previous efforts by incorporating denser text descriptions extracted from detailed 2D visuals, while our model leverages both point cloud and 2D visual information to enhance reasoning capabilities.

3 DenseScan

Existing 3D indoor scene datasets [9, 3, 40] typically rely on brief, one-sentence annotations or fragmented short-phrase labels that fail to capture the full semantic and spatial complexity of diverse environments. Moreover, traditional annotation pipelines often overlook the rich contextual details available from 2D RGB-D frames, thereby missing the fine-grained visual cues essential for robust scene understanding. Although more recent datasets [14, 19, 41] employ large language models such as GPT-4 [2] to generate diverse referring expressions, they still struggle to incorporate the inherent visual context of 2D spaces. ReasonSeg3D [21] leverages GPT4o—which supports visual input—by providing a single scene image with its ground-truth segmentation to enhance 3D spatial comprehension; however, this approach is constrained by the reliance on reconstructed 3D scene images that do not fully reflect the true attributes of the objects. DenseScan addresses these deficiencies by offering extended, detailed text descriptions that comprehensively depict each scene and by integrating an advanced generation pipeline that harnesses intricate details from 2D scans. This enhanced approach not only enriches the descriptive quality of the dataset but encourage more accurate and nuanced 3D scene analysis.

DenseScan build upon a widely used 3D scene understanding dataset ScanNet [11] by directly using the multi-view 2D RGB-D video frames along with its semantic annotations for rich context extraction. Specifically, DenseScan enrich each 3D scans with dense object-level captions and scenario-specific questions, providing a richer context for downstream tasks such as referring expression segmentation, visual grounding and question answering. An example sample of DenseScan and the detailed statistics and distribution of the dataset is shown in Table 1 and Figure 2, respectively.

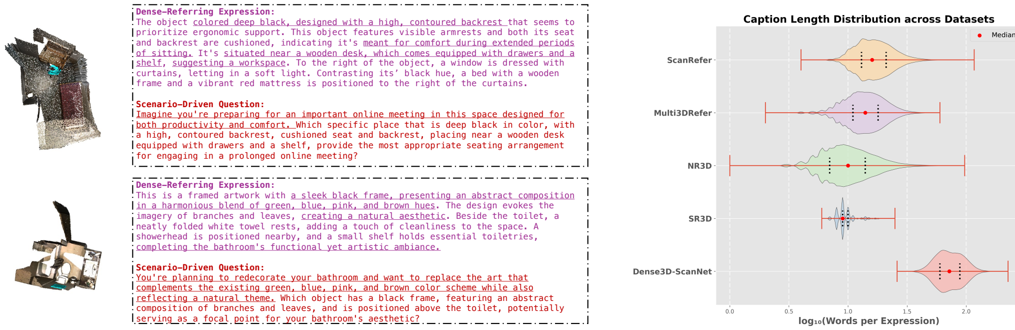

Figure 2: **DenseScan Dataset.** (a) Two objects sampled from ScanNet[11] with dense-referring expressions and scenario-driven questions. Objects shown are "office chair" and "pictures", both highlighted in the point cloud. (b) The distribution of description lengths, with dotted lines marking quartiles. The x-axis is scaled logarithmically (base-10) to handle the long-tail distribution.

3.1 Data Annotation Pipeline

We carefully designed an automatic annotation pipeline to generate dense scene-level descriptions and scenario-based referring object text expressions for ScanNet dataset [11], as illustrated in Fig. 1. The pipeline leverages the InterVL2-76B [10] multimodal large language model (MLLM) as the multimodal data annotator and Qwen2 [39] as our LLM assistant for consistency checking,filtering out unidentified objects and conflicting object descriptions.

Object-Level Description. To obtain the object-level dense caption rich is instance-specific semantics information, we first select the single video frame with the largest object area from the video frames and mask out non-object pixels. In this way, object isolation minimizes the interference from the background or nearby objects and directs the MLLM annotator to focus solely on annotated object's features. The cropped object image is then fed into InterVL2-76B to generate detailed descriptions.

Frame-Level Description. Building upon the detailed single-object descriptions, we enrich the single-object text description by incorporating spatial and relational information from the selected video frame. Specifically, we enhance the visual prominence of the annotated object using yellow contours, which serves as a visual cue to isolate the object from its surroundings while preserving the contextual detials of the scene. The annotated image is fed into the MLLM annotator along with a carefully designed prompt. The prompt is tailored to instruct the model to focus on generating a caption that not only describes the highlighted object but also captures its spatial relationship and interactions with adjacent elements.

Scene-Level Description. To acquire a comprehensive description of the scene, we uniformly sampled 8 frames from the video, ensuring that diverse perspectives is captured. We also apply a yellow contour to consistently highlight the target objects. These frames, along with frame-level descriptions, are processed by InterVL2-76B though a text prompt to generate scene-level object captions. The scene-level caption is further processed by an LLM annotator to transform into natural-flow referring expressions that are coherent, named *dense referring expression*.

Scenario-Like Questions. In the final stage of our pipeline, we aim to generate scenario-like questions that target the unique aspects of the selected object. Using the scene-level caption generated from target object, combing with a comprehensive list of all other objects present in the 3D environment, we carefully prompt the LLM annotator to craft scenario-like questions that specifically refer only to the target object. These questions are designed to evoke contextual reasoning and encourage nuanced interpretations that highlight the role, functionality, and significance of the object within realistic scene scenarios.

Benchmark Quality Control. Before releasing scenario-driven questions for benchmakring, we apply additional quality checks to the existing scenario-driven questions. Specifically, we provided these questions to a LLM and prompt it to identify the referring objects-along with unique characteristics for the objects to verify that the description accurately represents the object. Any descriptions that are inconsistent are eliminated. Additionally, following [9], we manually eliminate under-sampled

	# of Scenes	# of Descriptions	Annotation Method
ScanRefer[9]	703	51,583	human labeling
Sr3D[3]	707	83,572	human labeling
Nr3D[3]	707	41,503	human labeling
Multi3DRefer[40]	800	61,926	human + ChatGPT
Reason3D[19]	-	2,484	GPT-4[2]
Instruct3D[14]	280	2,565	-
ScanReason[41]	1,456	12,929	GPT-4[2]
ReasonSeg3D[21]	1,513	20,113	GPT-4V[1]
DenseScan (Ours)	**1,513**	**76,248**	InterVL2.5[10] + Qwen2[39]

Table 1: **Comparison of various 3D referring expression datasets**. showing the number of scenes, number of descriptions, and annotation methods used. The proposed DenseScan features 1,513 scenes, 76,248 descriptions (including 38,765 dense-referring expressions and 37,483 scenario-based questions), and an automated annotation pipeline, surpassing previous datasets in scale.

objects such as "stick", and structural objects such as "ceiling", "wall" to keep a balanced distribution of commonly referenced objects. Finally, we perform manual reviews on a random subset of the remaining descriptions to ensure overall quality and consistency.

3.2 Dataset Statistics

We quantitatively compare DenseScan with existing 3D scene benchmarks along two key dimensions: data scale and referring expression length.

Data Scale. DenseScan comprises 1,513 scenes and 20,113 object instances in total from ScanNet [11], with 38,765 dense-referring expressions and 37,483 scenario-based questions. Following ScanNet [11], we partition it into a training set of 1201 scans and a validation set of 312 scans. As can be seen from Table 1, DenseScan provides comparable number of referring expressions than existing datasets like ScanRefer [9], Multi3DRefer [40]. This is achieved through the special design of automated data collection pipeline with MLLM annotator, which is more robust and time-efficient compared to labor-intensive labeling.

Expression Length. Caption length is another important feature of DenseScan is its emphasis on generating detailed, instance-specific captions. In our dataset, each object is described with referring expressions that are notably longer and more descriptive than those in other benchmarks. While traditional datasets often offer brief captions that focus on basic attributes, DenseScan delivers captions that, on average, contain a higher word count—reflecting complex semantic details and contextual relationships. This extended caption length facilitates deeper semantic parsing, enabling models to capture intricate object properties and inter-object relationships that are critical for high-level reasoning in 3D scenes.

3.3 Evaluation Metrics

Following traditional 3D segmentation methods [20, 36, 14, 22, 34], we adopt mean intersection over union (mIoU), which quantifies the overall alignment between the predicted and ground-truth point cloud by averaging the Intersection over Union scores over all 3D point clouds. We also use Accuracy (Acc) to measures the percentage of segmentation predictions that achieve an IoU above a specified threshold $k = \{0.25, 0.5\}$

4 3D Scenario-Driven Segmentation

4.1 Task Definition

Scenario-driven segmentation extends traditional segmentation by incorporating rich, scenario-specific semantic cues, capturing the intended real-world scenario depicted by the annotations. Specifically, 3D scenario-driven segmentation task involves generating a 3D segmentation map $\mathcal{M}$ from a given point cloud representation $\mathcal{P}$ of a scene, along with a long-context scenario-like

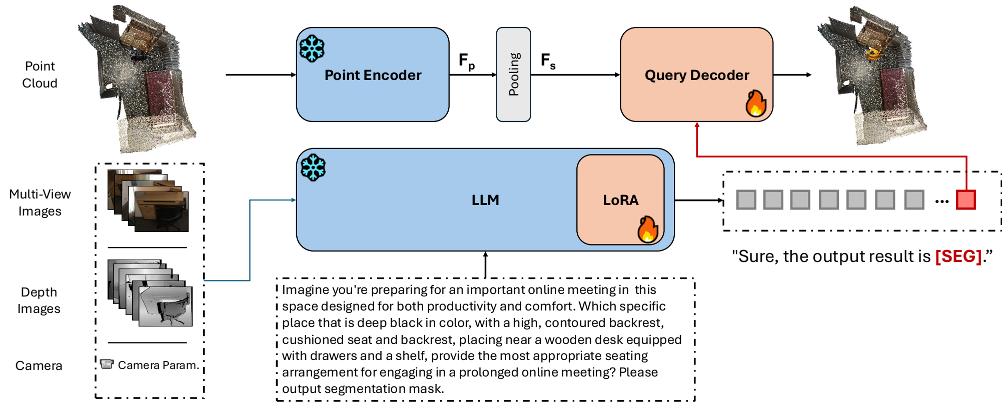

Figure 3: **Dense3D Model Architecture**. Given the 3D scene and language description, the model first reprocess 3D scene into multimodal 2D data, including RGB frames, depth map and camera poses. Depth map and camera poses composes the 3D positional embedding, along with the RGB frames and text descriptions to be send into the LLM. Output from LLM include special [SEG] token that is crucial to guide the Query Decoder for mask generation.

question $\mathcal{X}_{txt}$, extending beyond traditional referring expression segmentation tasks with short text descriptions [20, 36].

4.2 Vanilla Baseline

We design a simple LLM-based framework for this task to for complex scenario-driven reasoning within 3D point clouds. The overall architecture of Dense3D is illustrated in Figure 3.

Point Encoder. We first apply a novelization operation to the raw point cloud to discretize the spacial data. A Sparse 3D U-Net backbone [12] is employed to extract point-wise features, represented as $F_p \in \mathbb{R}^{N_p \times C_p}$, where N is the number of sampled points and C denotes the feature dimension. To mitigate computational complexity, these features are further processed through a superpoint pooling layer that leverages pre-computed superpoints [25]. Aggregating point-wise features using average pooling results in superpoint features $F_s \in \mathbb{R}^{N_s \times C_p}$, where N_s represents the number of superpoints.

Multi-Modal LLMs. To fuse 2D information in to 3D point cloud to allow model to perform better understanding capabilities, we instruct LLM to learn detail semantics leverage the strong 2D understanding priors of 2D MLLMs. We adopt pre-trained LLaVA-like models as our 2D multi-modal LLM backbone, which comprises a visual encoder, a visual projection layer, and an LLM. The visual encoder processes 2D multi-view videos of scanned 3D scenes enriched with depth and pose information—to extract robust visual features. These features are then mapped into visual tokens through the projection layer. The visual tokens, in conjunction with the corresponding text tokens, are subsequently fed into the LLM to generate textual predictions.

To enable prediction of mask features, we follow prior works [23, 14, 19, 21] and expand LLM's vocabulary with a special [SEG] token, indicating a request for a 3D segmentation mask.

3D Query Decoder. In the final stage, we route the [SEG] token into a decoder module which built on transformer-based architecture to decode a segmentation mask directly from the superpoint features F_s. By integrating the semantic directive provided by [SEG] with the detailed point-level features, the decoder effectively translates high-level textual cues into precise segmentation outputs for the target 3D scene regions.

Training Objectives. Our model jointly optimize both the multi-modal language generation and the 3D segmentation capabilities. The overall loss is formulated as a weighted combination of the three components as follows:

$$\mathcal{L} = \lambda_{LLM}\mathcal{L}_{LLM} + \lambda_{BCE}\mathcal{L}_{BCE} + \lambda_{DICE}\mathcal{L}_{DICE}$$

where $\mathcal{L}_{LLM}$ ensure coherent and accurate text generation, and $\mathcal{L}_{BCE}$ and $\mathcal{L}_{DICE}$ help refine high-quality segmentation masks.

	Acc@0.25	**Acc@0.50**	**mIOU**
3D-STMN[37]	20.8	13.2	14.4
MDIN[36]	21.2	10.8	15.1
Dense3D* (Ours)	34.3	20.1	23.2
Dense3D (Ours)	**35.3**	**20.9**	**24.0**

Table 2: Quantitative results on the 3D scenario-driven segmentation task on **DenseScan** benchmark among Dense3D (ours) and existing methods. * represents removing DenseScan dataset from the instruction tuning dataset.

	Acc@0.25	**Acc@0.50**	**mIOU**
TGNN[20]	38.6	32.7	28.8
X-RefSeg3D[30]	40.2	33.5	30.6
3D-STMN[37]	54.6	**39.8**	39.5
SegPoint[14]	-	-	**41.7**
Dense3D* (Ours)	**56.8**	33.7	36.6
Dense3D (Ours)	56.5	34.3	37.8

Table 3: Quantitative results on the **ScanRefer** benchmark among Dense3D (ours) and exisiting methods (including specialist methods such as 3D-STMN[37] and LLM generalist such as SegPoint [14]. * represents removing DenseScan dataset from the instruction tuning dataset.

4.3 Experiments

Instruction Tuning Datasets. Our training dataset contains the following three types of datasets: 1) semantic segementation datasets include ScanNet200 [33]; 2) referring expression datasets includes ScanRefer [9], Multi3DRefer[40] and ReferIt3D [3] as short-text referring expression datasets and DenseScan as long-text referring expression dataset; 3) question answering dataset include ScanQA.

Implementation Detail. The model is trained on 8 NVIDIA A100 GPUs. We adopt the AdamW optimizer with a learning rate of 3e-4 and use a learning rate scheduler WarmupDecay LR with the warmup steps of 100. The total batch size is set to be 16. The loss weight parameters λ_{TXT} is set to 1.0, and weight parameters for segmentation λ_{DICE} and λ_{BCE} is set to 1.0 and 1.0 respectively. We adopt the pretrained point cloud encoder following Sparse 3D-Unet [20] as the 3D visual encoder. We initialize our model with the weight of LLaVA-3D [42], and durring training, we use LoRA[17] to efficiently finetune the Large Language Model to reduce the computation costs while preserving the original 3D scene understanding capability.

Evaluation on 3D Scenario-Driven Segmentation. The performance of 3D Scenario-Driven Segmentation is evaluated on the validation set of DenseScan, and it is shown in Table 2. We performed a comparison between our baseline model to existing methods of various types, inncluding LLM-based methods and non-MLLM methods. Besides, to better validate the performance of our model and ensure a fair comparison, we removed DenseScan dataset from the training data, and denoted Dense3D*. From Table 1, we can see that LLM-based method generally out-perform non-LLM methods by a large amount, demonstrating its strong 3D scene understanding and reasoning capabilities over long-text referring expressions, and our baseline Dense3D achieve competitive performance on DenseScan, outperforming most existing baselines. We also show qualitative results in Figure 4.

Does Dense Object Descriptions help 3D segmentation? Long-text referring expressions provide extensive details that enhance a model's generalization by capturing nuanced attributes and maintaining long-range dependencies to differentiate similar objects. An interesting question, however, is whether dense object descriptions also benefit traditional short-context referring expressions. As shown in Table 3 and Table 4, training on such rich descriptions enables our model to learn fine-grained associations between language and visual features, ultimately improving its generalization. Notably, on ScanRefer[9], despite being trained on long-text data rather than optimized for short, structured expressions like specialist models (e.g., 3D-STMN), our model achieves comparable performance in identifying referents in complex real-world scenarios. This demonstrates the effective-

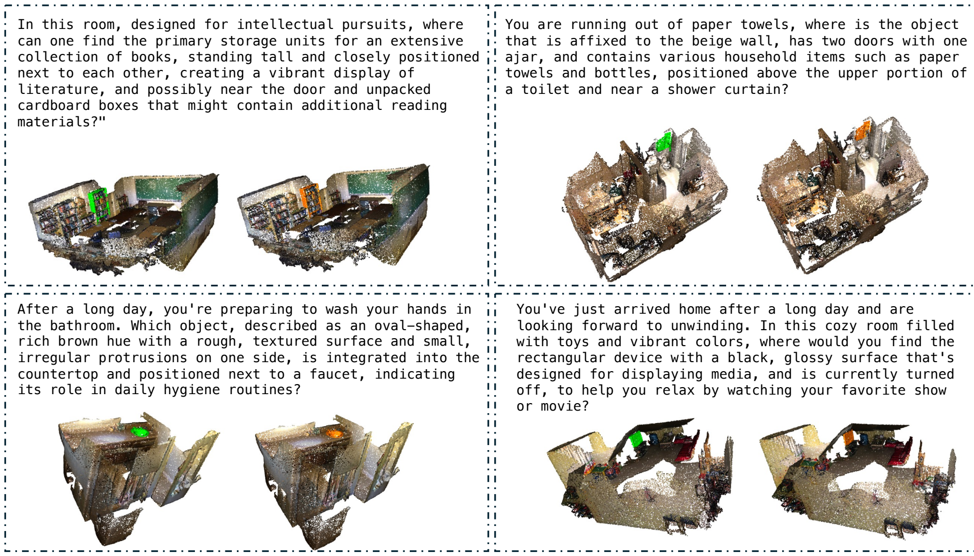

Figure 4: **Visual results for 3D Scenario-Driven Segmentation Task**. Each visual output presents a textual scenario-driven question, along with ground truth point cloud in green regions and predicted point cloud in orange regions.

	Acc@0.25	Acc@0.50	mIOU
M3DRef-CLIP[40]	55.7	37.5	37.4
SegPoint[14]	-	-	36.1
Dense3D* (Ours)	57.1	36.0	38.4
Dense3D (Ours)	**57.7**	**37.5**	**39.2**

Table 4: Quantitative results on the **Multi3DRefer** benchmark among Dense3D (ours) and existing methods. * represents removing DenseScan dataset from the instruction tuning dataset.

ness of leveraging long-text multimodal training, suggesting that our approach is not only competitive with specialized methods but also more adaptable to diverse and realistic tasks.

5 Conclusion

In this paper, we propose a pipeline that leverages state-of-the-art 2D MLLMs to generate high-quality, multi-level 3D annotations, culminating in the DenseScan dataset, which significantly enhances linguistic diversity and contextual richness in 3D scene descriptions. Additionally, we introduce *3D scenario-driven segmentation*, a benchmark that challenges models to reason about object interactions and spatial relationships beyond simple object identification. We also present Dense3D, a 3D MLLM that integrates 2D multi-view images, 3D point clouds, and textual descriptions to achieve deeper semantic understanding and more precise segmentation. We hope our work takes a step toward bridging the gap between 3D MLLMs and real-world applications, demonstrating that high-quality data and well-structured benchmarks are essential for advancing the field.

References

[1] Gpt-4v(ision) system card. 2023.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[3] 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.

[4] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[5] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1534–1543, 2016.

[6] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[7] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9297–9307, 2019.

[8] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017.

[9] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. *16th European Conference on Computer Vision (ECCV)*, 2020.

[10] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, 2024.

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

[12] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9224–9232, 2018.

[13] Timo Hackel, Nikolay Savinov, Lubor Ladicky, Jan D Wegner, Konrad Schindler, and Marc Pollefeys. Semantic3d. net: A new large-scale point cloud classification benchmark. *arXiv preprint arXiv:1704.03847*, 2017.

[14] Shuting He, Henghui Ding, Xudong Jiang, and Bihan Wen. Segpoint: Segment any point cloud via large language model. In *ECCV*, 2024.

[15] 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. *NeurIPS*, 2023.

[16] 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.

[17] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[18] 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. *arXiv preprint arXiv:2311.12871*, 2023.

[19] Kuan-Chih Huang, Xiangtai Li, Lu Qi, Shuicheng Yan, and Ming-Hsuan Yang. Reason3d: Searching and reasoning 3d segmentation via large language model. In *International Conference on 3D Vision (3DV)*, 2025.

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

[21] Xueying Jiang, Lewei Lu, Ling Shao, and Shijian Lu. Multimodal 3d reasoning segmentation with complex scenes, 2024.

[22] Maxim Kolodiazhnyi, Anna Vorontsova, Anton Konushin, and Danila Rukhovich. Oneformer3d: One transformer for unified point cloud segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20943–20953, 2024.

[23] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023.

[24] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024.

[25] Loic Landrieu and Martin Simonovsky. Large-scale point cloud semantic segmentation with superpoint graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4558–4567, 2018.

[26] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[28] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. *arXiv preprint arXiv:2210.07474*, 2022.

[29] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[30] Zhipeng Qian, Yiwei Ma, Jiayi Ji, and Xiaoshuai Sun. X-refseg3d: Enhancing referring 3d instance segmentation via structured cross-modal graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4551–4559, 2024.

[31] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13009–13018, 2024.

[32] Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26374–26383, 2024.

[33] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[34] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d: Mask transformer for 3d semantic instance segmentation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8216–8223. IEEE, 2023.

[35] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems*, 36:61501–61513, 2023.

[36] Changli Wu, Yihang Liu, Jiayi Ji, Yiwei Ma, Haowei Wang, Gen Luo, Henghui Ding, Xiaoshuai Sun, and Rongrong Ji. 3d-gres: Generalized 3d referring expression segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7852–7861, 2024.

[37] Changli Wu, Yiwei Ma, Qi Chen, Haowei Wang, Gen Luo, Jiayi Ji, and Xiaoshuai Sun. 3d-stmn: Dependency-driven superpoint-text matching network for end-to-end 3d referring expression segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5940–5948, 2024.

[38] Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. In *European Conference on Computer Vision*, pages 131–147. Springer, 2024.

[39] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

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

[41] Chenming Zhu, Tai Wang, Wenwei Zhang, Kai Chen, and Xihui Liu. Empowering 3d visual grounding with reasoning capabilities. *arXiv e-prints*, pages arXiv–2407, 2024.

[42] Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness. *arXiv preprint arXiv:2409.18125*, 2024.

[43] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract is constructed based on the main contribution and scope of the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [No]

Justification: Limitation is not discussed in the paper.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper does not include any theoretical results so there's no proof attached.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper explains implementation details, data annotation pipeline and the architecture in section 4.3 of the main paper.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Currently we are not planning to release the code.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Experimental Settings and detailed information about design choice are listed in the main paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The experiment in the paper is a proof of concept and further experiment on statistical significance will be carried out in the future development.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Details of hardware compute resources is mentioned in the section 4.3 of the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: I have reviewed the Code of Ethics and strictly follow the rules.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have mentioned in the introduction section where the current work has great potential in the field of robotics.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

Justification: The paper does not possess such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: The paper and its codebase is developed from scratch by the author without using existing assets.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not relase new assests so it's not applicable.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.