HOLO: Holistic Lightweight Optimization for Scene Understanding with Auto-Annotation and Multimodal Learning Supplementary Materials

Xiaoyun Hu*, Xiaohan Yan*, Nan Wang, Gang Wei, Zhicheng Wang†
Tongji University
{2432053, yxhop666, 2233017, weigang, zhichengwang}@tongji.edu.cn

A. Dataset Generation

A.1. SceneCap Generation

We construct the SceneCap dataset through systematic scene decomposition and caption generation. Specifically, we process a total of 1,080 scene scans from ScanNet V2 and ScanNet++, subsampling the original point clouds into localized sub-scenes via image-based projection. For each sub-scene, we generate a corresponding scene-level description, resulting in 5,886 high-quality scene–caption pairs. Each pair supports a single-turn, instruction-based captioning task: given a local point cloud and a textual prompt as input, the model is expected to generate a coherent and detailed description of the scene.

To help readers better understand our automated annotation pipeline, Table 1 presents a concrete example. The first row of the table displays three components: the original full-scene point cloud rendering, five sampled multi-view images obtained through frame skipping, and the projected local scene point cloud, which is derived from the original point cloud using image-guided decomposition. These five images are grouped and fed into the InternVL2-40B model to generate a descriptive caption. The output serves as the reference description for the corresponding localized scene. Through this image-guided decomposition and caption generation process, we construct a high-quality, instruction-driven dataset tailored for 3D scene understanding.

A.2. Scalable Automatic Annotation Framework

We constructed SceneCap using ScanNet v2 and ScanNet++, and applied the same processing pipeline to the S3DIS dataset. While S3DIS does contain multi-view RGB images and corresponding camera parameters, in our current setup we deliberately focused only on the raw scene-level point clouds, treating it as a pure geometric dataset. To enable the use of our multi-view framework, we simulated virtual cameras around each scene to generate synthetic images. Although this strategy allowed us to apply the same annotation pipeline, the resulting image quality was relatively low, leading to less optimal outcomes compared to those achieved on ScanNet-based datasets.

Despite these challenges, the proposed lightweight automatic annotation framework demonstrates strong scalability and can be easily extended to various types of 3D scene data. In future work, we plan to improve this component to better support datasets with limited or low-quality image inputs and explore the adaptation of our framework to outdoor environments, paving the way for the construction of large-scale, diverse scene understanding datasets.

B. Training

As shown in Table 2, we adopt a three-stage training strategy. The table provides a detailed overview of the training data, parameter scale, and training time used in each stage. We use Phi-2 as our causal LLM backbone and fine-tune it to enhance its ability to attend to relevant scene regions accurately. For optimization, we adopt the AdamW optimizer with a weight decay of 0.05. The learning rate follows a linear warm-up cosine decay schedule, starting from an initial learning rate of 1×10^{-5} and decaying to 1×10^{-6}, with a warm-up learning rate set to 1×10^{-6}.

C. Qualitative Results

To better demonstrate the effectiveness of NIUMO-LLM on both scene-level captioning and object-level understanding tasks, we conduct extensive comparisons and visualizations covering scene descriptions, object recognition, and classification performance.

C.1. 3D Scene Understanding

As shown in Table 3, 4, we present additional qualitative results of NIUMO-LLM on the scene captioning task.For scene point clouds, it is capable of providing detailed de-

*Equal contribution.
†Corresponding author.

scriptions from multiple aspects. NIUMO-LLM demonstrates a robust ability to accurately identify a diverse set of objects within complex 3D scenes, recognize their spatial relationships, and infer the overall atmosphere and stylistic characteristics of the environment. Beyond basic object recognition, the model exhibits a fine-grained understanding of functional and aesthetic arrangements, allowing it to produce detailed, coherent, and contextually grounded scene descriptions.

C.2. 3D Object Understanding

As shown in Figure 1, we present additional visualization results for object-level recognition and captioning tasks. In this setting, NIUMO-LLM takes as input only the raw 3D point cloud of an object along with a simple textual instruction—such as "Caption this 3D model in detail" or "What is this?".

Despite this limited input, NIUMO-LLM demonstrates a remarkable ability to extract and interpret meaningful geometric and color-related cues from the point cloud data. It accurately identifies object shapes, surface structures, and chromatic patterns, and further leverages these low-level visual cues to infer higher-level semantic attributes. For instance, the model is able to recognize stylized or cartoon-like aesthetics, infer the potential function or usage of everyday objects (e.g., "a bottle used in animation or games"). Another representative example is the model's interpretation of a steepled structure as a church. Based solely on its geometric form and predominant white color, NIUMO-LLM not only correctly identifies the object as a religious building but also makes reasonable inferences about its interior function—such as its use for religious ceremonies or communal gatherings—even though the interior is not visible in the input point cloud. This illustrates the model's ability to extrapolate plausible functional roles based on learned priors and spatial reasoning, going beyond surface-level recognition.

These results highlight NIUMO-LLM's capacity for abstract reasoning, contextual imagination, and semantic grounding purely from 3D point cloud data. By bridging the gap between geometric perception and natural language generation, NIUMO-LLM sets a solid foundation for object-level understanding in multimodal and vision-language applications, particularly in low-supervision scenarios where image references or class labels are unavailable.

Figure 1. Results of 3D object-level recognition and captioning tasks. Given a 3D point cloud and an instruction, NIUMO-LLM generates a corresponding text response. The correct and detailed description words are emphasized in **bold**.

Table 1. **An example of data generation with InternVL2-40B.** The input consists of grouped multi-view images and a carefully designed system prompt. The output includes a detailed scene-level caption generated from the multi-view image inputs.

	
Scene_ID	scene0008_00_group_6
System Prompt	Below are images of the same scene taken from different perspectives: {*image_descriptions*}. Using the information from all the images, craft a detailed and cohesive description of the entire scene. Begin by setting the scene with an overall introduction to the atmosphere and main features of the room. Then, guide the reader through the room by describing each area and item in a logical order, starting from one side. Elucidate the purpose and characteristics of specific areas or items to provide a clear understanding of their function and design. Conclude by highlighting the room's functionality and the types of activities it is suitable for, tying together the elements into a comprehensive portrayal of the space.
Output	"The scene depicted in the images showcases a modern, well-lit waiting area, likely within a corporate or professional setting. The atmosphere is calm and inviting, with a minimalist design that emphasizes functionality and comfort. The room is spacious, with a high ceiling and ample natural light streaming in through large windows, which contribute to a bright and airy ambiance. The walls are painted in a neutral, warm tone, enhancing the welcoming feel of the space. The primary feature of the room is the seating arrangement, which includes several plush, white armchairs with curved, ergonomic designs. These chairs are arranged in a semi-circular pattern, facing a central point, suggesting that the room is intended for group discussions or meetings. The chairs are positioned on a carpeted floor, which adds to the comfort and softness of the space. In the foreground, there is a small, dark wooden side table with a green glass top, which serves as a convenient place to set down items such as books, magazines, or electronic devices. Adjacent to the side table is a floor lamp with a white shade, providing additional lighting and contributing to the room's overall aesthetic. The lamp is positioned to cast a soft, diffused light, enhancing the cozy atmosphere. The back wall features a large, open doorway that leads to another room, suggesting that this waiting area is part of a larger facility. The doorway is framed by a wooden door, which adds a touch of warmth and complements the room's overall design. The room is equipped with a small, round, brown ottoman, which can be used as additional seating or as a footrest. This ottoman is placed near the side table, making it easily accessible for anyone seated in the area. The overall design of the room is modern and functional, with a focus on comfort and convenience. The seating arrangement, lighting, and decor all work together to create a space that is both inviting and professional. This waiting area is suitable for a variety of activities, including meetings, interviews, or simply waiting for an appointment. The ergonomic chairs and comfortable seating options make it an ideal space for individuals to relax and wait comfortably. The room's design and layout also suggest that it is intended for use by a diverse group of people, with the flexibility to accommodate different needs and preferences. In conclusion, the waiting area depicted in the images is a well-designed, inviting space that combines modern aesthetics with practical functionality. The combination of comfortable seating, thoughtful lighting, and thoughtful decor creates an environment that is both professional and welcoming. This space is well-suited for a variety of activities, making it an essential part of the overall functionality of the facility."

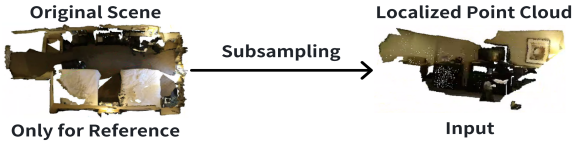
Table 2. Training setup and computational overhead for each stage.

Training Stage	Dataset Type	Dataset Scale	Init_lr & Min_lr	Trainable Parameters (M)	Training Time
Stage I	Brief Caption SceneCap	660k + 6k	3e-5, 1e-5	14.40	4:18:40
Stage II	Brief Caption Detailed Caption Conversation SceneCap	660k + 70k + 6k	1e-5, 1e-6	59.67	7:12:06
Stage III	Detailed Caption Conversation SceneCap	70k + 6k	5e-6, 1e-6	13.39	0:42:17

Table 3. Comparison of different models in 3D scene understanding.

Sample Scenes	 	 
Prompt	Please craft a detailed and cohesive description of the entire scene.	
Ground Truth	The scene depicted in the images showcases a cozy and well-organized living room that exudes a warm and inviting atmosphere. The room is characterized by its light wooden flooring, which adds a natural and airy feel to the space. The walls are..... Overall, the living room is designed to be both functional and comfortable, suitable for a variety of activities such as relaxing, dining, and socializing. The combination of comfortable furniture, warm lighting, and thoughtful decor creates a welcoming and inviting space that is perfect for both casual and formal gatherings.	The scene depicted in the images showcases a cozy and well-organized kitchen and dining area, blending functionality with a touch of personal style. The atmosphere is warm and inviting, with a blend of modern appliances and traditional decor elements. Starting from the left side of the room, the kitchen area is..... The overall design of the room is a harmonious blend of functionality and comfort, with a focus on practicality and a touch of personal flair. The kitchen is equipped with all the necessary appliances and storage solutions, while the dining area provides a comfortable space for family gatherings and meals. The room is suitable for a variety of activities, including cooking, dining, and socializing, making it a versatile and inviting space for everyday living.
MiniGPT-3D	the scene depicted in the images showcases a cozy and well-organized living room, characterized by a warm and inviting atmosphere. the room is furnished with a combination of modern and traditional elements, creating a harmonious blend of styles. the focal point of the room is a large, plush, beige sofa..... overall, the living room is designed to be both functional and comfortable, suitable for a variety of activities such as relaxation, socializing, and entertainment. the combination of comfortable seating, a central focal point, and thoughtful design elements creates a space that is both inviting and practical.	the scene depicted in the images showcases a cozy and well-organized living room, characterized by a warm and inviting atmosphere. the room is furnished with a comfortable sofa, a matching armchair, and a small coffee table..... the room is designed to be both functional and aesthetically pleasing, with a focus on creating a comfortable and inviting space for relaxation and socializing. the arrangement of the furniture and the thoughtful use of lighting and decor elements contribute to a sense of order and balance, making the room suitable for a variety of activities, including reading, watching television, or hosting small gatherings. overall, the living room is a well-designed space that combines comfort, functionality, and aesthetic appeal, making it an ideal setting for both relaxation and socializing.
NIUMO-LLM	**the scene depicted in the images showcases a cozy and functional living room, characterized by its warm and inviting atmosphere. the room is well-lit, with natural light streaming in from large windows, which are adorned with sheer curtains that allow soft, diffused light to fill the space..... the room is designed to be both functional and comfortable, with ample seating and a central gathering point. the combination of comfortable seating, decorative elements, and practical furniture makes it suitable for a variety of activities, including watching television, reading, and socializing with friends and family. the room's design and layout create a welcoming and inviting atmosphere, making it an ideal space for relaxation and entertainment.**	**the scene depicted in the images showcases a cozy and functional kitchen, characterized by its warm and inviting atmosphere. the kitchen is well-organized, with a focus on practicality and comfort. the overall ambiance is enhanced by the natural light streaming in through the large windows, which are adorned with sheer curtains that allow for a soft, diffused light. the kitchen features a variety of appliances and storage solutions, all designed to maximize space and efficiency. starting from the left side of the room, the first area is the dining area..... overall, the kitchen is designed to be both functional and aesthetically pleasing. it is suitable for a variety of activities, including cooking, dining, and entertaining. the combination of practical storage solutions and comfortable seating makes it an ideal space for everyday use.**

Table 4. Comparison of different models in 3D scene understanding. The blue magenta and underline highlight detailed description words. The original scene (left) is shown as a reference, while the right-hand side displays the local point cloud.

Sample Scene	
Prompt	Please craft a detailed and cohesive description of the entire scene.
Ground Truth	The room depicted in the images exudes a cozy and inviting atmosphere..... The overall ambiance is warm and welcoming, with soft lighting ... The room is well-organized and thoughtfully designed to cater to the needs of its occupants. Starting from the left side of the room, we see a large window dressed with sheer curtains..... Overall, the room is designed to be both comfortable and functional, catering to the needs of travelers or guests who require a space for both relaxation and productivity. The thoughtful arrangement of furniture and the use of lighting create a harmonious and inviting environment, making it suitable for a variety of activities, from working to unwinding after a long day.
MiniGPT-3D	the scene depicted in the images showcases a cozy and inviting living room, characterized by its warm and welcoming atmosphere. the room is well-lit, with natural light streaming in from a large window dressed with sheer curtains..... the walls are painted in a light, neutral color..... the focal point of the room is a large, plush, beige sofa..... the room is designed for relaxation and socializing, with ample seating and a comfortable atmosphere. the combination of the plush sofa, armchairs, and coffee table makes it an ideal spot for watching television, reading, or simply enjoying a quiet moment with friends and family. the room's layout and design elements work together to create a harmonious and functional space that is both comfortable and aesthetically pleasing.
NIUMO-LLM	**the room depicted in the images exudes a cozy and functional atmosphere, designed to cater to both relaxation and productivity. the overall ambiance is warm and inviting, with soft lighting that casts a gentle glow across the space. the room is furnished with a comfortable bed, a small desk, and a comfortable chair, each serving a specific purpose. starting from the left side of the room, the bed is..... overall, the room is designed to be both functional and comfortable, suitable for a variety of activities such as sleeping, working, and relaxing. the combination of practical furniture and thoughtful design elements creates a space that is both inviting and efficient.**