Data-Efficiently Learn Large Language Model for Universal 3D Scene Perception

Zehan Wang[1*] , **Haifeng Huang**[1*], **Yang Zhao**[2], **Ziang Zhang**[1], **Tao Jin**[1], **Zhou Zhao**[1†]
[1]Zhejiang University; [2]ByteDance

Abstract

3D scene understanding has gained significant attention due to its wide range of applications. However, existing methods for 3D scene understanding are limited to specific downstream tasks, which hinders their practicality in real-world applications. This paper presents Chat-3D, which combines the 3D visual perceptual ability of pre-trained 3D representations and the impressive reasoning and conversation capabilities of advanced LLMs to achieve the first universal dialogue systems for 3D scenes. Specifically, we align 3D representations into the feature space of LLMs, thus enabling LLMs to perceive the 3D world. Given the scarcity of 3D scene-text data, we propose a three-stage training strategy to efficiently utilize the available data for better alignment. To enhance the reasoning ability and develop a user-friendly interaction scheme, we further construct a high-quality object-centric 3D instruction dataset and design an associated object-centric prompt. With limited data, Chat-3D achieves a 82.2% relative score compared with GPT-4 on the constructed instruction dataset, and comparable performance to state-of-the-art LLM-based methods.

1 Introduction

3D vision is an important way for robots to perceive the rich semantic and spatial information of the real world. 3D scene understanding (Azuma et al., 2022; Ma et al., 2022; Chen et al., 2020; Achlioptas et al., 2020; Chen et al., 2021) has garnered increasing attention in recent years, owing to its broad range of applications in human-robot interaction, metaverse, robotics, and embodied intelligence. However, current methods (Wang et al., 2023a,b; Yang et al., 2021; Jiao et al., 2022; Yuan et al., 2022; Parelli et al., 2023) are limited in addressing specific downstream tasks, such as captioning and question answering, while lacking the ability to engage in general dialogue regarding a 3D scene, restricting their practicality in various real-world tasks. A universal dialogue system for 3D scenes is an imperative component of high-level intelligent robots.

The general dialogue system for 3D scenes requires two kinds of abilities: 3D perception and reasoning. Recently, several studies (Yu et al., 2022; Pang et al., 2022; Wang et al., 2021; Zhang et al., 2022; Xue et al., 2023; Liu et al., 2023b) on pre-trained 3D representations shows impressive performance in 3D perception. However, the reasoning ability for the 3D world remains constrained owing to the scarcity of reasoning and describing data for 3D scenes.

Large language models (LLMs) (Chiang et al., 2023; OpenAI, 2023; Touvron et al., 2023; Chowdhery et al., 2022), on the other hand, exhibit remarkable prowess in complex reasoning and open-domain conversations. Moreover, recent methods (Li et al., 2023b; Liu et al., 2023a; Zhao et al., 2023; Zhang et al., 2023a; Zhu et al., 2023) attempt to extend LLMs to image and video fields. These works typically adopt a two-stage training scheme: Firstly, the visual representations are aligned into the word embedding space of LLMs by leveraging large-scale image-text and video-text datasets (Lin et al., 2014; Sharma et al., 2018; Changpinyo et al., 2021; Schuhmann et al., 2021, 2022; Bain et al., 2021; Miech et al., 2019; Xu et al., 2016). Secondly, they enhance the reasoning capabilities of LLMs regarding visual concepts by fine-tuning on the instruction datasets.

Despite the success of image and video understanding fields, introducing LLMs to perceive 3D scenes faces two challenges: 1) Compared to the millions or even billions of image-text and video-text data (Sharma et al., 2018; Changpinyo et al., 2021; Schuhmann et al., 2021, 2022; Bain et al., 2021), the 3D scene-text data (Achlioptas et al.,

---

[*] Equal Contribution.
[†] Corresponding author

Findings of the Association for Computational Linguistics:
NAACL 2025, pages 313–333
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

2020; Chen et al., 2020) is limited. Consequently, in the low-resource scenarios, the commonly used two-stage training scheme in previous multi-modal LLMs is less effective in aligning pre-trained 3D representations to the feature space of LLMs. 2) 3D scenes always encompass a greater number of objects compared to an image or a video clip. Thus, the common questions or instructions in images and videos are more susceptible to ambiguity in 3D scenes. Consider a simple question like "What is in front of this chair?" on a 3D scene that contains multiple chairs. The dialogue model cannot understand which specific chair the user is asking about, and uniquely describing an object (the chair) in question is often difficult and user-unfriendly due to the complex object relations.

In this paper, we propose Chat-3D, the first attempt to extend the reasoning and conversation capabilities of LLMs to 3D scene understanding. We employ a three-stage training scheme to more efficiently utilize the limited data. Specifically, in the first stage, we directly align the features of 3D objects with the word embeddings of their class names. In the second stage, we learn a 3D object relation module via 3D scene-text data to capture semantic information about the whole 3D scene. Finally, in the third stage, we further tune the model with a high-quality instruction dataset. To further enhance the reasoning ability of Chat-3D, we construct the instruction dataset via an object-centric scheme, which means all instructions are related to a specific object. Combining our object-centric prompt, users can effortlessly select the object in the scene they want to engage in a dialogue about, without the need to uniquely describe the specific object in their instructions.

In summary, our contributions can be summarized as follows:

(1) We build the first universal dialogue system for 3D scenes, leveraging the advanced visual perception capabilities of 3D pre-trained models, in conjunction with the powerful reasoning and open-domain conversational abilities of LLMs.

(2) We introduce a new three-stage training scheme for multi-modal LLM, enabling the model to progressively transition from learning individual object attributes to capturing complex spatial object relations. This approach effectively improves the quality of dialogue with limited available data.

(3) We construct a high-quality object-centric 3D instruction dataset including diverse dialogues about object attributes, positions, relationships, functionalities, placement suggestions, and detailed descriptions within 3D scenes. We propose a corresponding object-centric prompt approach to provide a user-friendly interaction method.

(4) Our experiments demonstrate that Chat-3D exhibits remarkable capabilities in universal dialogue and spatial reasoning based on 3D scenes. We also employ quantitative comparison to evaluate the effectiveness of our three-stage training scheme and instruction dataset.

2 Related Work

3D Representation Learning 3D point cloud is a fundamental visual modality. Recently, numerous attempts are made to learn discriminative and robust representations for point cloud objects. Point-BERT (Yu et al., 2022), Point-MAE (Pang et al., 2022), Transformer-OcCo (Wang et al., 2021), and point-m2ae (Zhang et al., 2022) employ self-supervised learning approaches to extract meaningful representations of 3D objects from unlabeled point cloud data. Another series of works aims to extend representation from other modalities to 3D. For instance, ULIP (Xue et al., 2023) and openshape (Liu et al., 2023b) construct (3D-image-text) triplets to align point clouds within the CLIP (Radford et al., 2021; Cherti et al., 2023) representation space, while I2P-MAE (Zhang et al., 2023b) and ACT (Dong et al., 2023) learn 3D representations from image pre-trained models (Dosovitskiy et al., 2020; He et al., 2016). These powerful 3D representations can effectively capture the detailed information of a 3D object. In Chat-3D, we segment the 3D scene into objects and extract features for each object, which yields a set of object features to represent the 3D scene and serves as a prerequisite for an object-centric interactive approach.

3D-Language Tasks The interaction between 3D point clouds and natural language has wild applications and has garnered increasing attention recently. 3D captioning (Chen et al., 2021, 2020; Achlioptas et al., 2020) focuses on generating descriptions of a specific object in a 3D scene. In 3D visual question answering (Azuma et al., 2022), the model is required to answer questions based on the visual content of the 3D scene, while the more complex 3D situated question answering (Ma et al., 2022) requires the model to understand agent's situation (position, orientation, etc.) in a 3D scene as described by text, reason about the surrounding environment. Different from vision-language tasks (Kazemzadeh et al., 2014; Krishna et al., 2017; Goyal et al., 2017; Antol et al., 2015; Lin et al., 2014; Grauman et al., 2022) and methods (Li et al., 2022, 2023a, 2021; Lin et al., 2022) based on images and videos, these 3D-language tasks and corresponding methods place more emphasis on spatial reasoning and the possible interaction between agents and scenes. Despite the significant progress made in this field, existing methods still focus on improving isolated task-specific models, without exploring a unified dialogue system.

Multi-modal Large Language Models Recently, Large Language Models showcase remarkable abilities in complex reasoning and conversational communication with humans. To extend the knowledge, reasoning, and conversation abilities acquired from vast amounts of text data to more modalities, some studies (Li et al., 2023b; Liu et al., 2023a; Zhao et al., 2023; Zhang et al., 2023a; Zhu et al., 2023; Huang et al., 2024) attempt to instruction tune LLMs for multimodal learning. Specifically, these works first use the caption learning objective to learn the aligning of visual features with pre-trained LLMs from large-scale vision-language paired data. Then, a high-quality instruction dataset is utilized to further enhance the LLMs' comprehension of the visual world. However, in the 3D-Language field, 3D scene-text pairs are scarce. Thus the indirect aligning method is unreliable and incomplete for 3D representations and pre-trained LLMs. To mitigate this issue, we propose a more data-efficient three-stage tuning scheme that establishes a more direct learning stage for alignment, reduces the annotation requirements, and provides a smooth learning curve.

3 Methods

3.1 Architecture

Chat-3D aims to create a universal dialogue system for 3D scenes by aligning 3D representations with pre-trained LLM (Touvron et al., 2023). The overall architecture is illustrated in Figure 1.

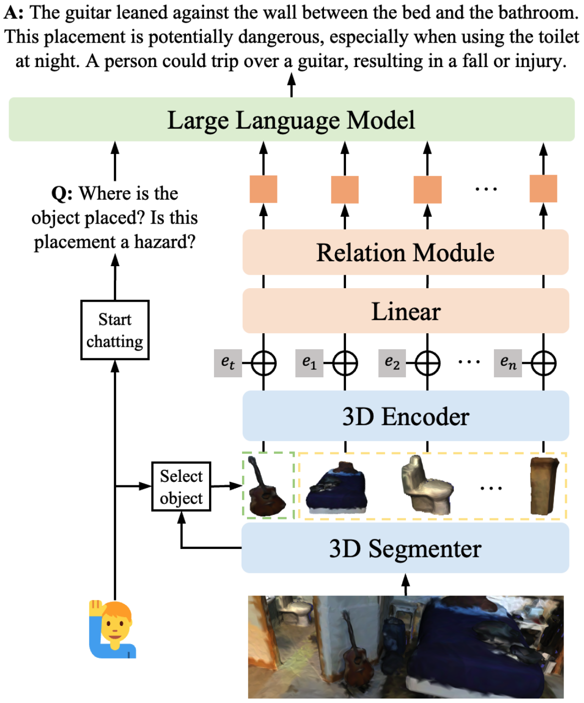

A: The guitar leaned against the wall between the bed and the bathroom. This placement is potentially dangerous, especially when using the toilet at night. A person could trip over a guitar, resulting in a fall or injury.
Large Language Model
Q: Where is the object placed? Is this placement a hazard?
Relation Module
Linear
Start chatting
e_t
e_1
e_2
e_n
3D Encoder
Select object
3D Segmenter

Figure 1: The overall architecture of Chat-3D.

For the input 3D scene S, we first use a 3D object segmentation model (Jiang et al., 2020; Misra et al., 2021; Qi et al., 2019) or ground truth annotations to segment it into objects. Then, users can select the specific object they want to engage in dialogue. The selected target object is denoted as o_t and other objects in the scene are represented as $O_s = [o_1, o_2, \ldots, o_{n_s}]$, where n_s is the number of other objects in the 3D scene. For each object, we use a pre-trained 3D point encoder $g(\cdot)$ to extract features, Besides, we further incorporate extra object attributes (e.g. color, size, location) into these object features by a projector $f_e(\cdot)$ to enrich semantic information. These 3D object features are projected to the word embedding space of pre-trained LLM via a projector $f_a(\cdot)$. The process of 3D object feature extraction and mapping can be expressed as:

$$\mathbf{z}_i = f_a(g(o_i)+\mathbf{e}_i), \text{ with } \mathbf{e}_i = f_e([\mathbf{c}_i;\mathbf{s}_i;\mathbf{l}_i]) \quad (1)$$

where $i \in [t, 1, 2, \ldots, n_s]$, and $\mathbf{c}_i, \mathbf{s}_i, \mathbf{l}_i \in R^3$ respectively represent the RGB value, bounding box size, and location for the i-th object. The extracted 3D features of target object and other objects are denoted as $\mathbf{z}_t$ and $\mathbf{Z}_s = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{n_s}]$.

Furthermore, we further introduce a relation module $t(\cdot)$ for capturing complex relations between objects. The features of objects then interact with each other to provide additional object relation information about the scene.

$$[\hat{\mathbf{z}}_t, \hat{\mathbf{z}}_1, \hat{\mathbf{z}}_2, \ldots, \hat{\mathbf{z}}_{n_s}] = r([\mathbf{z}_t, \mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{n_s}]) \quad (2)$$

The representations of a 3D scene are provided as $\hat{\mathbf{z}}_t \in R^d$, $[\hat{\mathbf{z}}_1, \hat{\mathbf{z}}_2, \ldots, \hat{\mathbf{z}}_{n_s}] \in R^{n_s \times d}$, and d is the dimension of hidden states in the pre-trained LLMs.

Lastly, to facilitate user-friendly interaction between our system and users, we design an object-centric prompt as: *###Human: [target]* $\hat{\mathbf{z}}_t$ *[/target], [scene]* $\hat{\mathbf{z}}_1, \hat{\mathbf{z}}_2, \ldots, \hat{\mathbf{z}}_{n_s}$ *[/scene], [instruction] ###Assistant:*. Through this prompt, the LLM can comprehend the specific object the user wants to discuss and generate responses based on the 3D visual information and the given instructions.

3.2 Three-stage Training

Previous multi-modal LLMs (Li et al., 2023b; Liu et al., 2023a; Zhao et al., 2023; Zhang et al., 2023a; Zhu et al., 2023) primarily follow a two-stage training scheme. In the first stage, LLMs take inputs from visual modality and learn to generate corresponding captions. The large-scale image- and video-text datasets allow comprehensive alignment between visual representations and the word embedding space of LLM. In the second stage, the model is fine-tuned with a high-quality instruction dataset, thereby further enhancing the perceptual and reasoning abilities.

However, in the 3D understanding field, the 3D scene-text data is significantly less than image- or video-text data. For example, the commonly used ScanRefer (Chen et al., 2020) dataset, which provides descriptions for ScanNet (Dai et al., 2017), only contains 36,655 captions for training. In contrast, the datasets used for the first stage training in previous multi-modal LLM methods are million-level or even billion-level, such as CC3M (Sharma et al., 2018), CC12M (Changpinyo et al., 2021), LAION-400M (Schuhmann et al., 2021), LAION-5B (Schuhmann et al., 2022) and WebVid-10M (Bain et al., 2021). Considering the scarcity of 3D scene-text data, we propose a more data-efficient three-stage training approach, which divides the process of aligning 3D features with the pre-trained LLM into two progressive stages: 3D object alignment and 3D scene alignment.

Stage 1: 3D Object Alignment The first stage is designed to learn the alignment between the representation of individual 3D objects and pre-trained LLM. Given a 3D object and its annotated category, the 3D object is encoded into a representation $\mathbf{z} \in R^d$ according to Equation 1. Its category name is encoded into a word embedding $\mathbf{y} \in R^d$ using the tokenizer of the pre-trained LLM. By maximizing the cosine similarity between the corresponding $\mathbf{z}$ and $\mathbf{y}$, we can learn projectors $f_e(\cdot)$ and $f_a(\cdot)$ that effectively inject the 3D object representations into the word embedding space of LLM.

The advantage of Stage 1 is three-fold: 1) Compared to learning alignment through captioning objective, maximizing the similarity between representations provides a more direct learning objective for alignment, which can achieve more efficient alignment in low-resource scenarios. 2) Stage 1 enables the utilization of 3D point cloud object classification datasets, such as ShapeNet (Chang et al., 2015), ScanObjectNN (Uy et al., 2019), and Objaverse (Deitke et al., 2023), which enhances the model's generalization performance on diverse real-world objects. 3) The introduction of Stage 1 offers a smoother learning curve for comprehending complex 3D scenes. The model progressively transitions from learning individual object attributes to capturing intricate spatial object relations.

Stage 2: 3D Scene Alignment After aligning individual 3D object feature with pre-trained LLM, Stage 2 takes a step further by integrating the entire 3D scene into LLM. The training data is sourced from the ScanRefer dataset, which provides annotations for objects in a scene primarily based on their spatial relationships. Considering a 3D scene, which can be segmented into object set $[o_1, o_2, \ldots, o_n]$, we sequentially select each object as target objects and construct the input for LLM according to the methodology discussed in Section 3.1. The instruction in prompts requests the model to generate a brief description of the target object within the 3D scene. The learning objective is to generate a description that aligns with the description provided by the ScanRefer dataset for the target object, and only the two projectors $f_e(\cdot)$, $f_a(\cdot)$ and the relation module $r(\cdot)$ are learnable in this stage.

Stage 3: Instruction Tuning For enhancing the reasoning ability about 3D world, we curate a high-quality instruction dataset which comprises rich and detailed instructions. By tuning Chat-3D on this dataset, we further enhance its capability to comprehend diverse instructions, generate imaginative and contextually appropriate responses, engage in intricate spatial reasoning, and effectively incorporate external knowledge into its responses.

4 Object-centric Instruction Dataset

The complex object relationships and intricate interactions between agents and scenes impose elevated demands on reasoning capabilities. To enhance the reasoning ability pertaining to 3D world, we construct a high-quality object-centric instruc-

Caption of the target object:
Descriptions: ["There is a single white armchair. placed next to the window of the room.", "The sofa chair is the corner chair. lying parallel to the wall. a small table with the lamp is present beside the chair.", "This is a white sofa chair. it is under a window.", "This is a white armchair. is next to a lamp.", "This is the corner sofa chair. a small table with a lamp can be seen near this chair."]
Categories and locations of target object and its 10 neighbors:
Described object: {sofa chair:[-1.31, 3.15, 0.59]}; Neighbor objects: {window:[-1.12, 4.12, 1.59], table:[0.86, 1.61, 0.38], doorframe:[-2.25, 0.67, 1.27], windowsill:[0.88, 3.97, 0.98], windowsill:[-1.32, 3.93, 0.91], sofa chair:[0.98, 3.35, 0.71], window:[1.16, 4.18, 1.73], pillow:[1.35, 0.29, 0.46], table:[-0.15, -2.66, 0.26], tv:[-2.2, -0.55, 1.52]}

Table 1: An example of textualizing an object in a 3D scene

You are an AI 3D visual assistant, and you are seeing an object in a 3D scene. What you see is provided with several sentences, describing the same object you are looking at, and the position of surrounding objects in the 3D scene to represent the content of the 3D scene. Based on these descriptions of this object and the location of surrounding objects in the 3D scene, summary and describe the placement, function of this object, and how a person can access this object in detail as if you are in the 3D scene.
Importantly, do not mention any specific spatial coordinate values. The description should be more than 150 words and less than 200 words.

Table 2: Prompt for descriptive object-centric captions.

tion dataset based on the annotations in ScanRefer. Specifically, we leverage the remarkable reasoning and summarizing capabilities of ChatGPT to automatically generate descriptive and detailed captions as well as diverse conversations centered around specific objects within 3D scenes.

Object-centric Descriptive Captions ScanRefer annotates multiple captions for objects in a 3D scene based on their attributes and spatial relationships. We employ ChatGPT to summarize and rewrite these short captions into imaginative paragraphs. To facilitate ChatGPT's comprehension of the 3D scene, we also textualize the 3D scene as shown in Table 1, providing the categories and XYZ coordinates of the target object and its ten nearest objects in the scene. Furthermore, we propose a prompt to request ChatGPT to focus on perceiving and reasoning about the object relations and agent interactions as exemplified in Table 2. During dataset construction, we initially manually annotated several examples and randomly selected two of them as in-context examples to guide the generation of ChatGPT. One example of the generated descriptive object-centric caption is shown in Table 7.

Object-centric Conversations To enhance the capability of handling diverse instructions and general conversations, we further require ChatGPT to autonomously generate multi-turn dialogues in a self-questioning and self-answering format based on the brief captions of the target object and the textualized 3D scene.

5 Experiments

5.1 Implementation Details

During the training phase, we directly use ground truth annotations (point cloud and extra attributes) of each object in the 3D scene for better training quality. We employ the pre-trained Point-Bind(Guo) model with Point-BERT(Yu et al., 2022) architecture as $g(\cdot)$ to extract features for each object. Meanwhile, we use a linear layer as $f_e(\cdot)$ to incorporate extra attributes (such as color, size, and location) into the extracted features. Then, a two-layer MLP serves as $f_a(\cdot)$ to map these 3D object features to the word embedding space of the pre-trained LLM, and the relation module $r(\cdot)$ is implemented using a one-layer vanilla transformer encoder. It is worth mentioning that the relation module is zero-initialized, thereby preserving the

Design a conversation between you and a person asking about this object in the 3D scene. The answers should be in a tone that a visual AI assistant is in the 3D scene and answering the question. Ask diverse questions and give corresponding answers.
Include questions asking about the visual content of this object, including the object types, object shape, object attribute, object functions, object locations, relative positions between objects, etc.
Only include questions that have definite answers:
(1) Questions whose contents can be confidently observed and answered based on the 3D scene.
(2) Questions whose absence from the 3D scene can be confidently determined.

Table 3: Prompt for object-centric conversations.

Training scheme	Training Data		Evaluate Set		Overall
	Conversation	Detailed Caption	Conversation	Detailed Caption	
Three-Stage			**88.2**	**76.2**	**82.2**
Two-Stage			84.8	65.9	75.4
Three-Stage		×	85.7	53.9	69.8
Three-Stage	×		85.1	69.2	77.2
Three-Stage	×	×	56.8	55.0	55.9

Table 4: Relative scores on the object-centric dataset (rated by GPT-4).

information learned in Stage 1 when Stage 2 begins. The chosen LLM for our experiment is a Vicuna 7B model(Chiang et al., 2023), which is fine-tuned from the LLaMA base model(Touvron et al., 2023).

5.2 Evaluation on Object-centric Dataset

Relative Score rated by GPT-4 In order to quantitatively evaluate the universal dialogue ability of Chat-3D and analyze the effect of the three-stage training scheme and our instruction dataset, we adopt GPT-4 (OpenAI, 2023) to measure the quality of our Chat-3D's generated responses following LLaVA (Liu et al., 2023a) and miniGPT4 (Zhu et al., 2023). Specifically, we randomly select 30 scenes from the ScanRefer validation set and randomly choose one object as the target object for each scene. We employ the instruction dataset construction methodology described in Section 4 and Chat-3D respectively to generate responses under the same scene and instruction inputs. After that, we input the textualized scene, instructions, and the two kinds of generated responses into GPT-4 and request GPT-4 to provide an overall score on a scale of 1 to 10 for each response based on its helpfulness, relevance, accuracy, and level of detail. A higher score indicates a higher quality of response.

In Table 4, we study the effectiveness of the instruction dataset and compare the Chat-3D trained via our three-stage training scheme and the two-stage training method used in previous methods (Li et al., 2023b; Liu et al., 2023a; Zhao et al., 2023; Zhang et al., 2023a; Zhu et al., 2023). First, our three-stage training scheme significantly outperforms the previous two-stage method by 6.8 points, demonstrating the data efficiency of our three-stage training method in the low-resource setting. Second, by comparing different combination settings of the instruction dataset, we observe that incorporating conversation data leads to a higher improvement in conversation tests, while integrating detailed caption data enhances performance in detailed caption tests. By utilizing all the data together, our model demonstrates proficiency in both conversation and detailed caption tasks, ultimately achieving the highest overall score.

Caption Score In Table 5, we evaluate the caption metrics BLEU, METEOR, and ROUGE-L under various conditions. This evaluation covers the entire dataset, including all conversations and detailed captions, offering a more comprehensive understanding of the significance of different components. It is clear from the table that excluding either conversation data or detailed caption data leads to a significant decrease in performance. Without stage three, performance deteriorates even more dramatically. These results highlight the effectiveness of our proposed three-stage training architecture.

Training Scheme	Training Data Conv.	Detail.	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
Three-Stage			**38.19**	**22.71**	**13.80**	**8.81**	**19.83**	**35.41**
Two-Stage			36.54	21.31	12.69	7.95	19.59	34.95
Three-Stage		×	10.63	6.79	4.51	3.12	9.84	34.57
Three-Stage	×		18.28	10.54	6.00	3.54	18.12	15.92
Three-Stage	×	×	1.31	0.76	0.39	0.21	3.89	17.71

Table 5: Caption scores on the object-centric dataset.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
VoteNet+MCAN	28.0	16.7	10.8	6.2	11.4	29.8	54.7
ScanRefer+MCAN	26.9	16.6	11.6	7.9	11.5	30.0	55.4
ScanQA	30.2	20.4	15.1	10.1	13.1	33.3	64.9
LLaVA (zero-shot)	7.1	2.6	0.9	0.3	10.5	12.3	5.7
3D-LLM (flamingo)	30.3	17.8	12.0	7.2	12.2	32.3	59.2
3D-LLM (BLIP2-flant5)	39.3	25.2	18.4	12.0	14.5	35.7	69.4
Chat-3D	29.1	16.3	10.1	6.4	11.9	28.5	53.2

Table 6: Evaluation results on ScanQA validation set.

5.3 Evaluation on ScanQA

In order to evaluate on the ScanQA dataset, we finetune the pretrained Chat-3D to fit the answer format of this dataset.

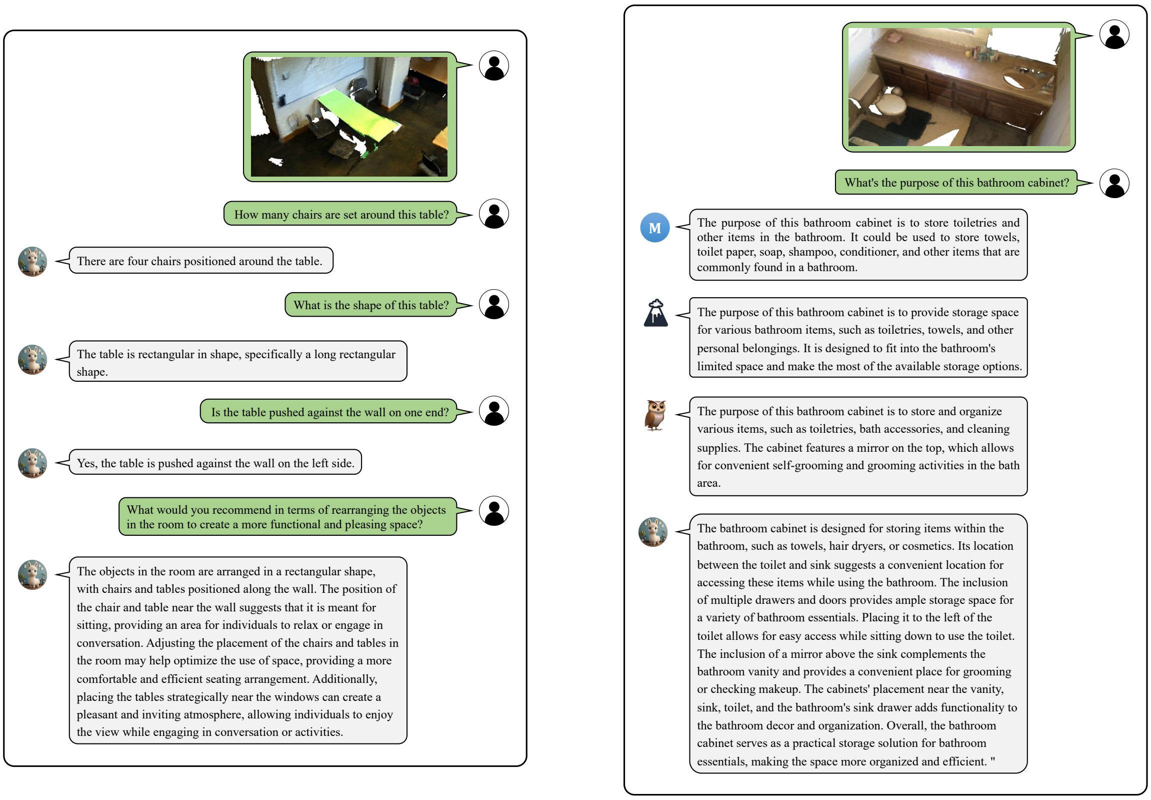
Baselines We include representative baseline models on the benchmark. Specifically, **ScanQA** (Azuma et al., 2022) utilizes VoteNet to generate object proposals and then integrates them with language embeddings. **ScanRefer+MCAN** (Chen et al., 2020) and **VoteNet+MCAN** (Ding et al., 2019) detect 3D objects and incorporate them into a standard VQA model known as MCAN (Yu et al., 2019). **LLaVA** (Liu et al., 2023a) proposes the visual instruction tuning method, which establishes a connection between a vision encoder and LLM to enable general-purpose visual and language understanding. **3D-LLM** (Hong et al., 2023) relies on 2D Vision-Language Models (VLMs) as their backbone, such as flamingo and BLIP-2. It extracts meaningful 3D features from rendered multi-view images, which serve as the input for the VLM.

Analysis Essentially, Chat-3D achieves competitive results compared to fully supervised methods such as ScanQA. The most intriguing aspect is the performance gap between 3D-LLM and Chat-3D. When changing the backbone from Flamingo to BLIP2-flant5, 3D-LLM experiences a significant performance boost across all metrics. This suggests that 3D-LLM heavily relies on the robust 2D VLM, which is pretrained on billion-level data. In contrast, Chat-3D solely utilizes 3D data for pretraining and fine-tuning, which is based on a much smaller data set. Nevertheless, it still manages to achieve competitive results compared to 3D-LLM (Flamingo), highlighting the effectiveness of our training architecture.

5.4 Qualitative Comparisons & Analysis

In section, we provide visualization examples of conversations about 3D scenes with Chat-3D. From these cases, we mainly study the perception, reasoning, and dialogue capabilities of Chat-3D. Besides, we further compare Chat-3D with 2D multimodal LLM methods such as MiniGPT-4 (Zhu et al., 2023), LLaVA (Liu et al., 2023a), and mPLUG-owl (Ye et al., 2023) to demonstrate the advantages and necessity of developing a specific multi-modal LLM for 3D scenes.

Perception, Reasoning and Dialogue We provide several examples of conversations with Chat-3D in Figure 2-7 covering various commonly-seen object types (*e.g.*, table, chair, and bed). In Figure 2(a), Chat-3D shows strong perception capabilities by accurately counting objects, recognizing shapes, and precisely localizing them within the 3D space. In Figure 4, Chat-3D demonstrates impressive reasoning capabilities by deducing the cabinet's purpose and evaluating its practicality based on its



How many chairs are set around this table?

There are four chairs positioned around the table.

What is the shape of this table?

The table is rectangular in shape, specifically a long rectangular shape.

Is the table pushed against the wall on one end?

Yes, the table is pushed against the wall on the left side.

What would you recommend in terms of rearranging the objects in the room to create a more functional and pleasing space?

The objects in the room are arranged in a rectangular shape, with chairs and tables positioned along the wall. The position of the chair and table near the wall suggests that it is meant for sitting, providing an area for individuals to relax or engage in conversation. Adjusting the placement of the chairs and tables in the room may help optimize the use of space, providing a more comfortable and efficient seating arrangement. Additionally, placing the tables strategically near the windows can create a pleasant and inviting atmosphere, allowing individuals to enjoy the view while engaging in conversation or activities.

(a) Example 1 of Chat-3D conversation.

What's the purpose of this bathroom cabinet?

The purpose of this bathroom cabinet is to store toiletries and other items in the bathroom. It could be used to store towels, toilet paper, soap, shampoo, conditioner, and other items that are commonly found in a bathroom.

The purpose of this bathroom cabinet is to provide storage space for various bathroom items, such as toiletries, towels, and other personal belongings. It is designed to fit into the bathroom's limited space and make the most of the available storage options.

The purpose of this bathroom cabinet is to store and organize various items, such as toiletries, bath accessories, and cleaning supplies. The cabinet features a mirror on the top, which allows for convenient self-grooming and grooming activities in the bath area.

The bathroom cabinet is designed for storing items within the bathroom, such as towels, hair dryers, or cosmetics. Its location between the toilet and sink suggests a convenient location for accessing these items while using the bathroom. The inclusion of multiple drawers and doors provides ample storage space for a variety of bathroom essentials. Placing it to the left of the toilet allows for easy access while sitting down to use the toilet. The inclusion of a mirror above the sink complements the bathroom vanity and provides a convenient place for grooming or checking makeup. The cabinets' placement near the vanity, sink, toilet, and the bathroom's sink drawer adds functionality to the bathroom decor and organization. Overall, the bathroom cabinet serves as a practical storage solution for bathroom essentials, making the space more organized and efficient. "

(b) Example 1 of comparison between Chat-3D and 2D Multi-modal LLMs.

Figure 2: Qualitative examples of Chat-3D. [1]

placement and spatial relationships with surrounding objects. Guided by the object-centric prompt outlined in Section 3.1, Chat-3D adeptly directs its attention to the specific target object indicated by the user. This enables Chat-3D to maintain focus on the intended subject without being diverted by other similar objects present in the scene. Moreover, the conversational exchanges consistently demonstrate the high-quality dialogue delivered by Chat-3D.

Comparisons with 2D Multi-modal LLMs We compare Chat-3D with MiniGPT-4 (Zhu et al., 2023), LLaVA (Liu et al., 2023a), and mPLUG-owl (Ye et al., 2023) in Figure 2, 8, 9 and 10. As illustrated in Figure 2(b), Chat-3D demonstrates its capability to provide a clear and meticulous analysis of the given question. In Example 2, depicted in Figure 8, we assess the model's spatial perception ability in determining whether both monitors are of identical size. Chat-3D showcases accurate identification, while the other 2D models provide incorrect answers due to their limitations in grasping depth and perspective relationships within the 2D image. In Example 3, presented in Figure 9, the constraints of 2D models become more apparent as they struggle to accurately identify the spatial relationships between the target object and its surrounding objects. Compared to 2D models, Chat-3D's analytical prowess shines brightly due to its remarkable aptitude for perceiving and comprehending concepts within the 3D space.

6 Conclusion

In this paper, we build the first universal dialogue system for 3D scenes, leveraging the advanced visual perception capabilities of 3D pre-trained models, in conjunction with the powerful reasoning and open-domain conversational abilities of LLMs. To overcome the challenge of limited 3D data availability, we introduce a three-stage training scheme for multi-modal LLMs to progressively transition from learning individual object attributes to capturing complex spatial object relations. Furthermore, we construct a high-quality object-centric 3D instruction dataset and propose a corresponding object-centric prompt approach to facilitate a user-friendly interaction method. Experimental results demonstrate that Chat-3D showcases remarkable capabilities in universal dialogue, spatial reasoning, and the enhancement of external knowledge based on 3D scenes.

---

[1] More examples can be found in the appendix.

7 Limitations

Universal dialogue in 3D scenes is still an emerging field, and there is currently a lack of standard and diverse test datasets for evaluating 3D universal dialogue capabilities. Therefore, Chat-3D can only be tested on some academic datasets, which may not fully demonstrate the generalized and general dialogue reasoning capabilities of Chat-3D. More diverse 3D universal evaluation datasets are important and promising research directions.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China under Grant No.624B2128.

References

Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. 2020. 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.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. 2022. Scanqa: 3d question answering for spatial scene understanding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19129–19139.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738.

Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. 2015. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568.

Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. 2020. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European conference on computer vision*, pages 202–221. Springer.

Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. 2021. Scan2cap: Context-aware dense captioning in rgb-d scans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3193–3203.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. 2023. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153.

Zhipeng Ding, Xu Han, and Marc Niethammer. 2019. Votenet: A deep learning label fusion method for multi-atlas segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22*, pages 202–210. Springer.

Runpei Dong, Zekun Qi, Linfeng Zhang, Junbo Zhang, Jianjian Sun, Zheng Ge, Li Yi, and Kaisheng Ma. 2023. Autoencoders as cross-modal teachers: Can pretrained 2d image transformers help 3d representation learning? In *The Eleventh International Conference on Learning Representations (ICLR)*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012.

Ziyu Guo. Point-bind: Align 3d point clouds with multi-modalities. `https://github.com/ZrrSkywalker/Point-Bind`.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 2023. 3d-llm: Injecting the 3d world into large language models. *arXiv preprint arXiv:2307.12981*.

Haifeng Huang, Yilun Chen, Zehan Wang, Rongjie Huang, Runsen Xu, Tai Wang, Luping Liu, Xize Cheng, Yang Zhao, Jiangmiao Pang, et al. 2024. Chat-scene: Bridging 3d scene and large language models with object identifiers. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. 2020. Pointgroup: Dual-set point grouping for 3d instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and Pattern recognition*, pages 4867–4876.

Yang Jiao, Shaoxiang Chen, Zequn Jie, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. 2022. More: Multi-order relation mining for dense captioning in 3d scenes. In *European Conference on Computer Vision*, pages 528–545. Springer.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023b. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2022. Swinbert: End-to-end transformers with sparse attention for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17949–17958.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yinhao Zhu, Xuanlin Li, Shizhong Han, Hong Cai, Fatih Porikli, and Hao Su. 2023b. Openshape: Scaling up 3d shape representation towards open-world understanding. *arXiv preprint arXiv:2305.10764*.

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

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640.

Ishan Misra, Rohit Girdhar, and Armand Joulin. 2021. An end-to-end transformer model for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2906–2917.

OpenAI. 2023. Gpt-4 technical report.

Yatian Pang, Wenxiao Wang, Francis EH Tay, Wei Liu, Yonghong Tian, and Li Yuan. 2022. Masked autoencoders for point cloud self-supervised learning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, pages 604–621. Springer.

Maria Parelli, Alexandros Delitzas, Nikolas Hars, Georgios Vlassis, Sotirios Anagnostidis, Gregor Bachmann, and Thomas Hofmann. 2023. Clip-guided vision-language pre-training for question answering in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5606–5611.

Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. 2019. Deep hough voting for 3d object detection in point clouds. In *proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9277–9286.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. 2019. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1588–1597.

Hanchen Wang, Qi Liu, Xiangyu Yue, Joan Lasenby, and Matt J Kusner. 2021. Unsupervised point cloud pre-training via occlusion completion. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9782–9792.

Zehan Wang, Haifeng Huang, Yang Zhao, Linjun Li, Xize Cheng, Yichen Zhu, Aoxiong Yin, and Zhou Zhao. 2023a. 3drp-net: 3d relative position-aware network for 3d visual grounding. *arXiv preprint arXiv:2307.13363*.

Zehan Wang, Haifeng Huang, Yang Zhao, Linjun Li, Xize Cheng, Yichen Zhu, Aoxiong Yin, and Zhou Zhao. 2023b. Distilling coarse-to-fine semantic matching knowledge for weakly supervised 3d visual grounding. *arXiv preprint arXiv:2307.09267*.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.

Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. 2023. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1179–1189.

Zhengyuan Yang, Songyang Zhang, Liwei Wang, and Jiebo Luo. 2021. Sat: 2d semantics assisted training for 3d visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1856–1866.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. 2022. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19313–19322.

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6281–6290.

Zhihao Yuan, Xu Yan, Yinghong Liao, Yao Guo, Guanbin Li, Shuguang Cui, and Zhen Li. 2022. X-trans2cap: Cross-modal knowledge transfer using transformer for 3d dense captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8563–8573.

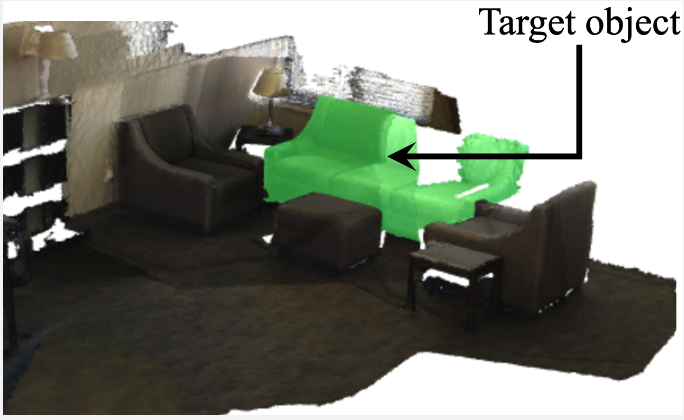
Hang Zhang, Xin Li, and Lidong Bing. 2023a. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Renrui Zhang, Ziyu Guo, Peng Gao, Rongyao Fang, Bin Zhao, Dong Wang, Yu Qiao, and Hongsheng Li. 2022. Point-m2ae: multi-scale masked autoencoders for hierarchical point cloud pre-training. *Advances in neural information processing systems*, 35:27061–27074.

Renrui Zhang, Liuhui Wang, Yu Qiao, Peng Gao, and Hongsheng Li. 2023b. Learning 3d representations from 2d pre-trained models via image-to-point masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21769–21780.

Yang Zhao, Zhijie Lin, Daquan Zhou, Zilong Huang, Jiashi Feng, and Bingyi Kang. 2023. Bubogpt: Enabling visual grounding in multi-modal llms. *arXiv preprint arXiv:2307.08581*.

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



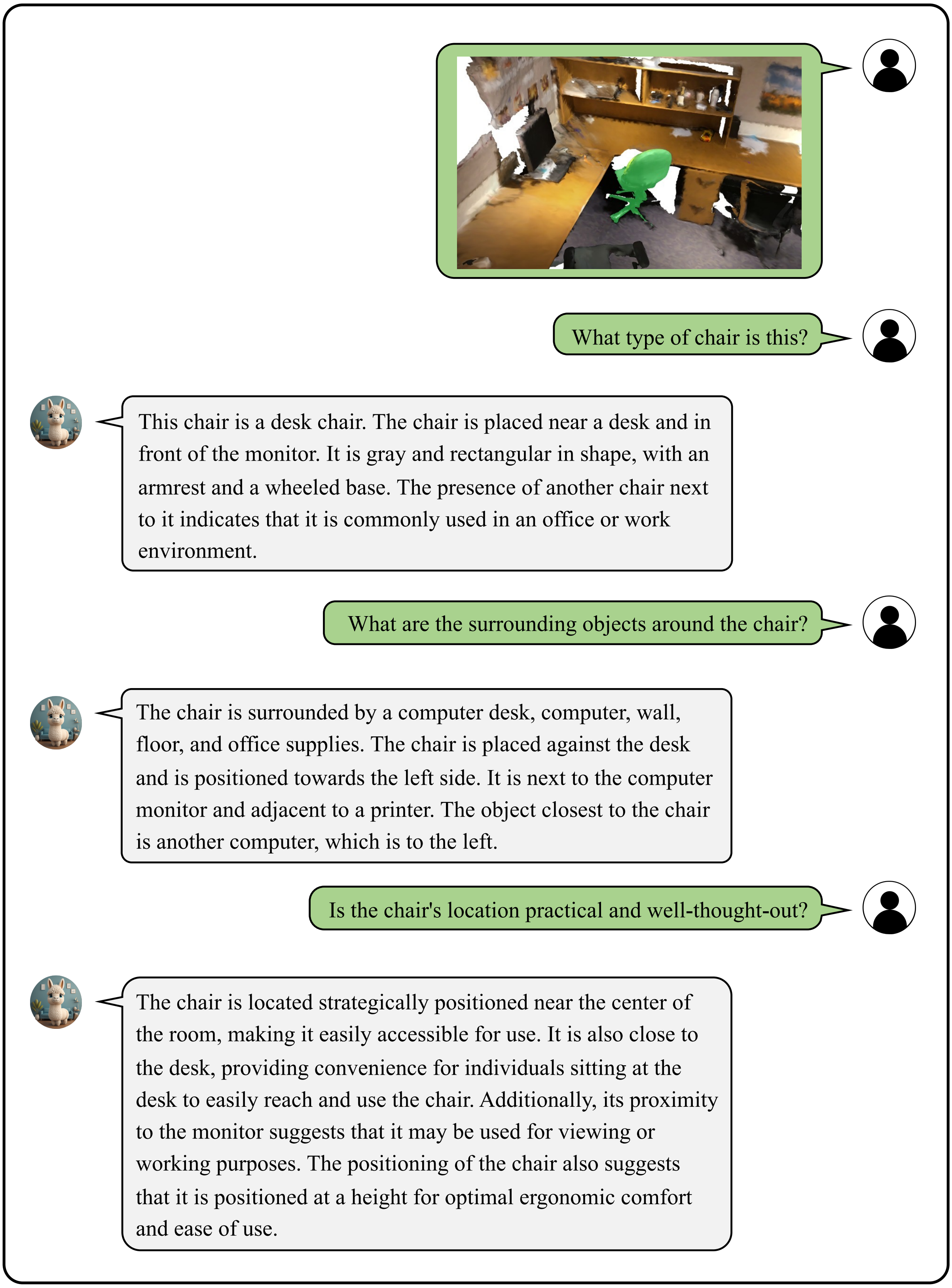
Detailed description: The object is a brown couch with three seats, positioned on the left corner of the room. It is next to a gold lamp, which serves as a source of lighting and complements the overall aesthetics of the area. The couch is accompanied by an ottoman placed in front of it, providing a comfortable footrest and enhancing the seating experience.

The presence of two armchairs nearby indicates that this space is designed for relaxation and socializing. The grey-colored couch with three cushions exudes a modern and stylish vibe, adding to the room's visual appeal. The strategic placement of the couch and ottoman encourages a cozy and inviting atmosphere, ideal for leisurely activities and gatherings.

This setup creates a comfortable seating arrangement, promoting interaction and conversation among individuals using the couch and armchairs. The nearby window adds natural light to the area, enhancing the ambiance. The presence of an end table with a lamp contributes to the functionality of the space by providing a convenient spot for placing items and lighting.

In summary, the brown couch with its ottoman and neighboring armchairs forms a functional and aesthetically pleasing seating arrangement. The combination of the couch's design, the presence of the ottoman, and the proximity to the end table with a lamp ensures a comfortable and enjoyable experience for individuals using this part of the room.

Table 7: Example of descriptive object-centric caption.



What type of chair is this?

This chair is a desk chair. The chair is placed near a desk and in front of the monitor. It is gray and rectangular in shape, with an armrest and a wheeled base. The presence of another chair next to it indicates that it is commonly used in an office or work environment.

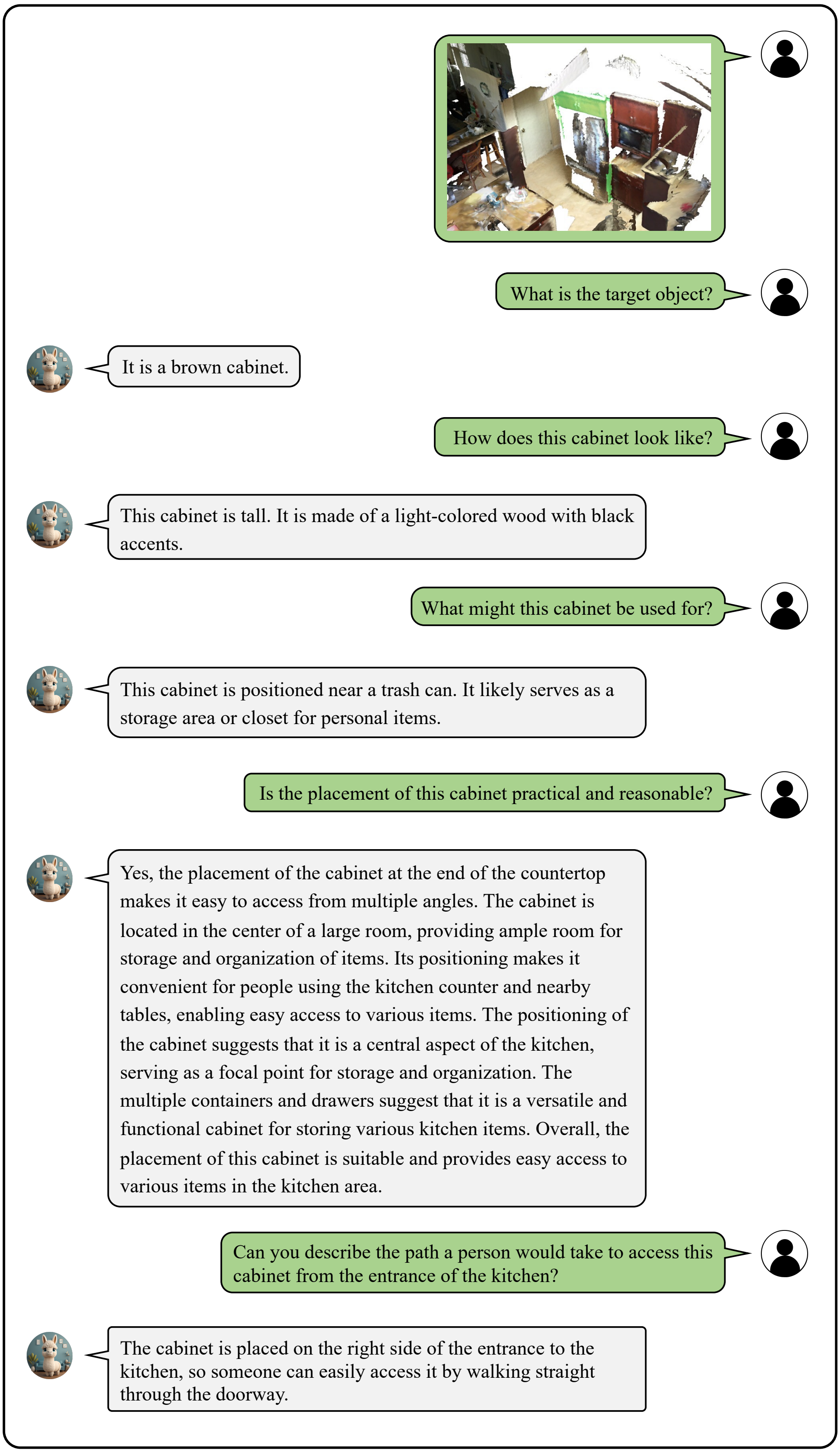
What are the surrounding objects around the chair?

The chair is surrounded by a computer desk, computer, wall, floor, and office supplies. The chair is placed against the desk and is positioned towards the left side. It is next to the computer monitor and adjacent to a printer. The object closest to the chair is another computer, which is to the left.

Is the chair's location practical and well-thought-out?

The chair is located strategically positioned near the center of the room, making it easily accessible for use. It is also close to the desk, providing convenience for individuals sitting at the desk to easily reach and use the chair. Additionally, its proximity to the monitor suggests that it may be used for viewing or working purposes. The positioning of the chair also suggests that it is positioned at a height for optimal ergonomic comfort and ease of use.

Figure 3: Example 2 of Chat-3D conversation.



What is the target object?

It is a brown cabinet.

How does this cabinet look like?

This cabinet is tall. It is made of a light-colored wood with black accents.

What might this cabinet be used for?

This cabinet is positioned near a trash can. It likely serves as a storage area or closet for personal items.

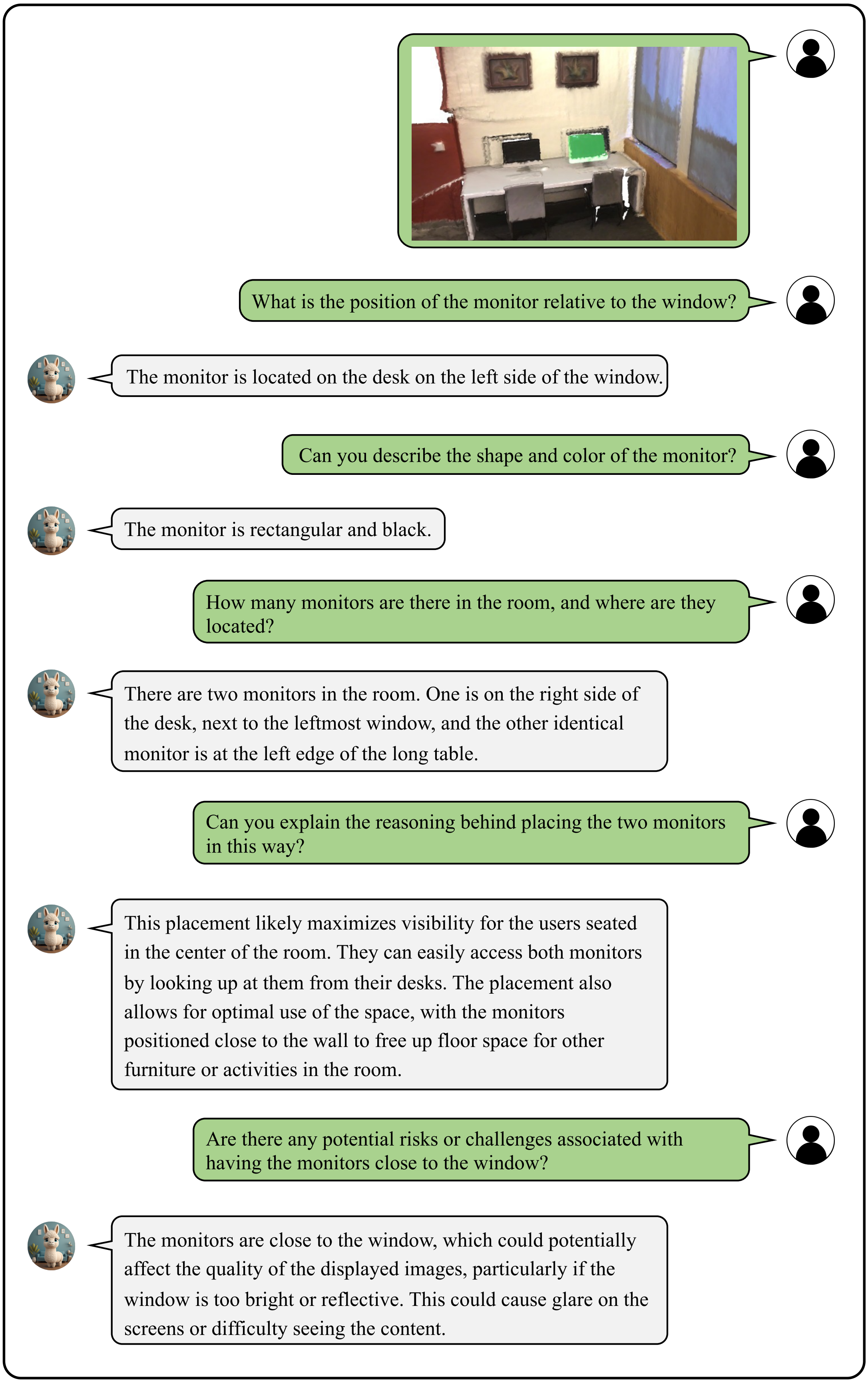
Is the placement of this cabinet practical and reasonable?

Yes, the placement of the cabinet at the end of the countertop makes it easy to access from multiple angles. The cabinet is located in the center of a large room, providing ample room for storage and organization of items. Its positioning makes it convenient for people using the kitchen counter and nearby tables, enabling easy access to various items. The positioning of the cabinet suggests that it is a central aspect of the kitchen, serving as a focal point for storage and organization. The multiple containers and drawers suggest that it is a versatile and functional cabinet for storing various kitchen items. Overall, the placement of this cabinet is suitable and provides easy access to various items in the kitchen area.

Can you describe the path a person would take to access this cabinet from the entrance of the kitchen?

The cabinet is placed on the right side of the entrance to the kitchen, so someone can easily access it by walking straight through the doorway.

Figure 4: Example 3 of Chat-3D conversation.



What is the position of the monitor relative to the window?

The monitor is located on the desk on the left side of the window.

Can you describe the shape and color of the monitor?

The monitor is rectangular and black.

How many monitors are there in the room, and where are they located?

There are two monitors in the room. One is on the right side of the desk, next to the leftmost window, and the other identical monitor is at the left edge of the long table.

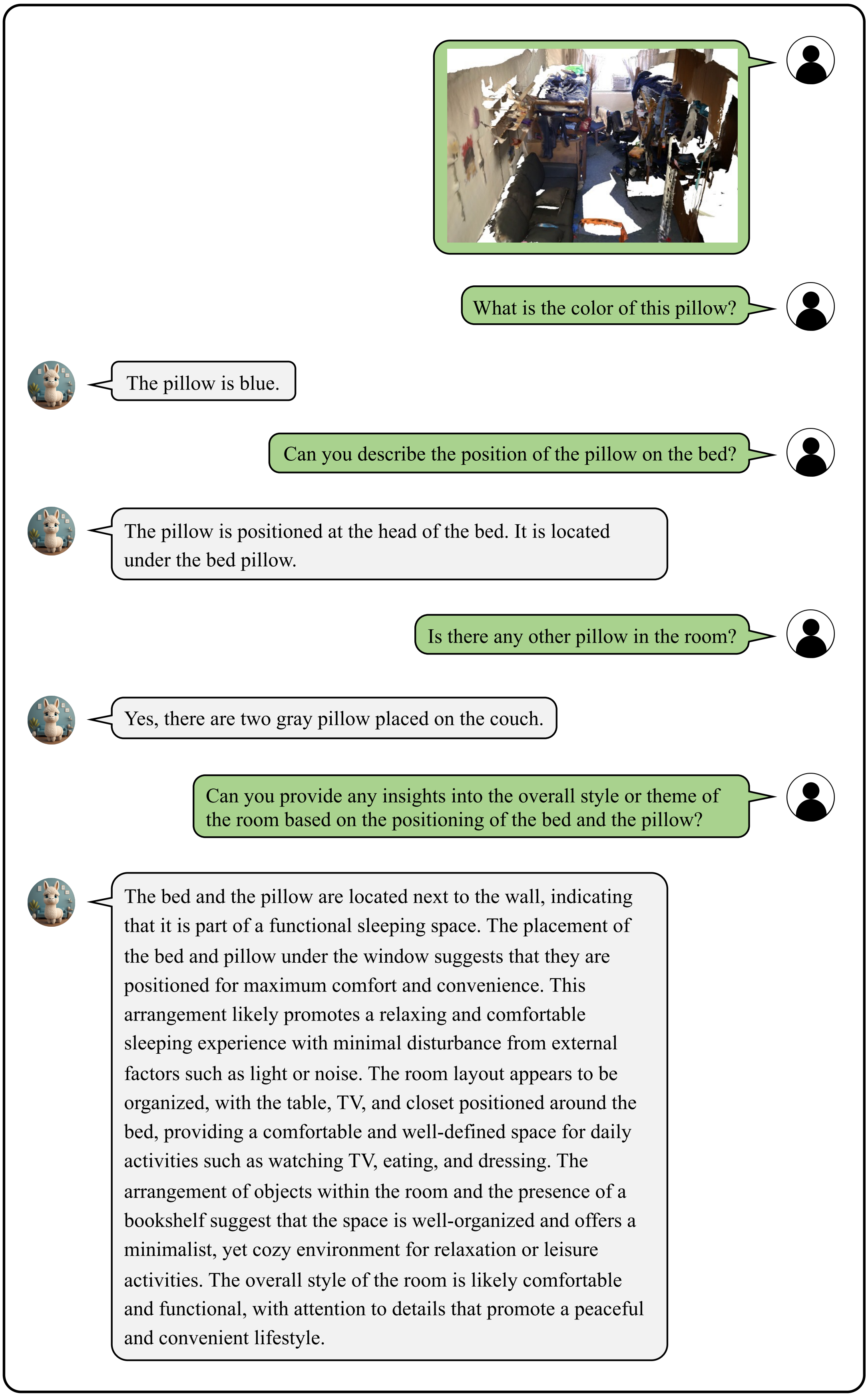
Can you explain the reasoning behind placing the two monitors in this way?

This placement likely maximizes visibility for the users seated in the center of the room. They can easily access both monitors by looking up at them from their desks. The placement also allows for optimal use of the space, with the monitors positioned close to the wall to free up floor space for other furniture or activities in the room.

Are there any potential risks or challenges associated with having the monitors close to the window?

The monitors are close to the window, which could potentially affect the quality of the displayed images, particularly if the window is too bright or reflective. This could cause glare on the screens or difficulty seeing the content.

Figure 5: Example 4 of Chat-3D conversation.



What is the color of this pillow?

The pillow is blue.

Can you describe the position of the pillow on the bed?

The pillow is positioned at the head of the bed. It is located under the bed pillow.

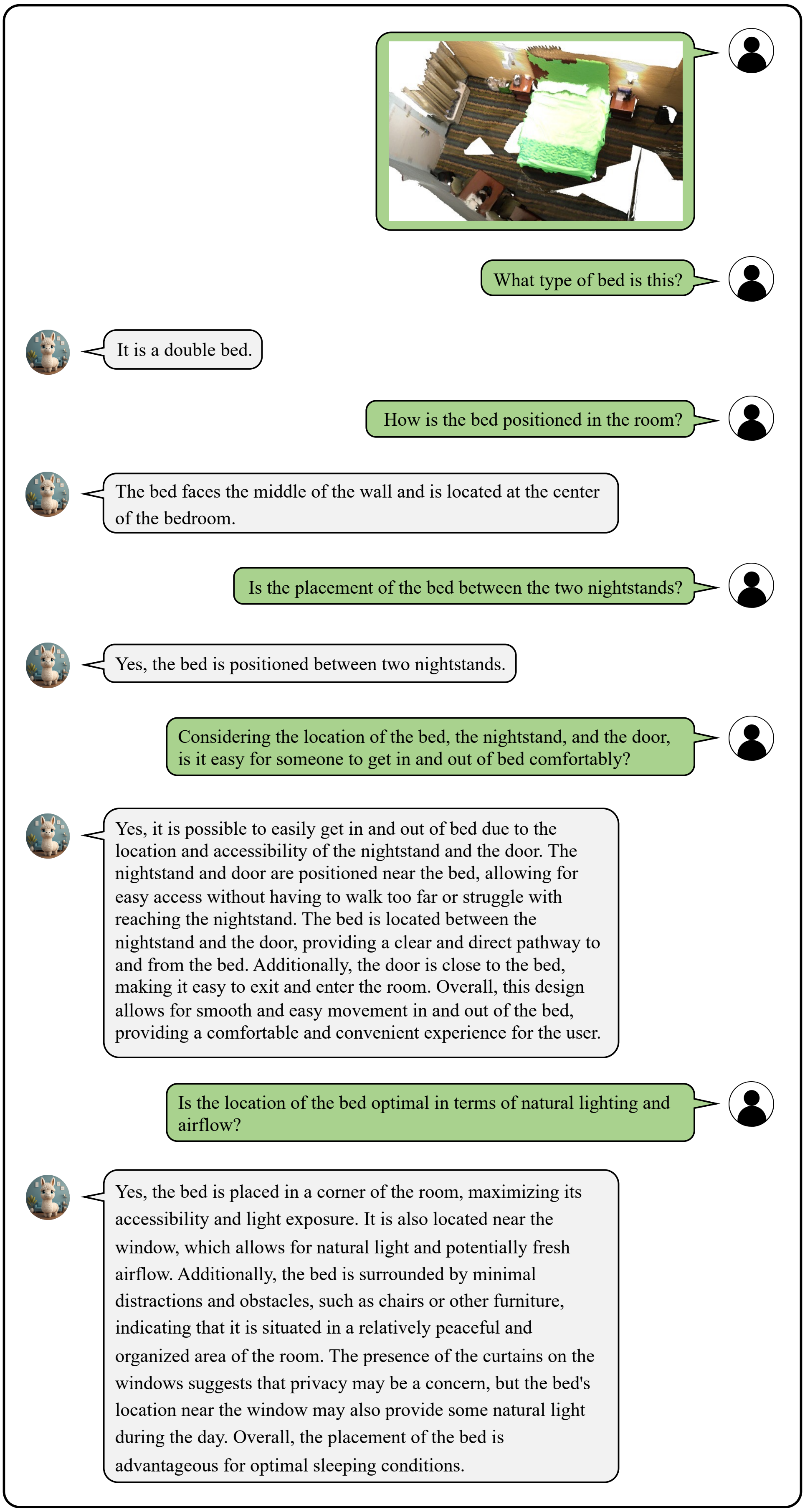
Is there any other pillow in the room?

Yes, there are two gray pillow placed on the couch.

Can you provide any insights into the overall style or theme of the room based on the positioning of the bed and the pillow?

The bed and the pillow are located next to the wall, indicating that it is part of a functional sleeping space. The placement of the bed and pillow under the window suggests that they are positioned for maximum comfort and convenience. This arrangement likely promotes a relaxing and comfortable sleeping experience with minimal disturbance from external factors such as light or noise. The room layout appears to be organized, with the table, TV, and closet positioned around the bed, providing a comfortable and well-defined space for daily activities such as watching TV, eating, and dressing. The arrangement of objects within the room and the presence of a bookshelf suggest that the space is well-organized and offers a minimalist, yet cozy environment for relaxation or leisure activities. The overall style of the room is likely comfortable and functional, with attention to details that promote a peaceful and convenient lifestyle.

Figure 6: Example 5 of Chat-3D conversation.



What type of bed is this?

It is a double bed.

How is the bed positioned in the room?

The bed faces the middle of the wall and is located at the center of the bedroom.

Is the placement of the bed between the two nightstands?

Yes, the bed is positioned between two nightstands.

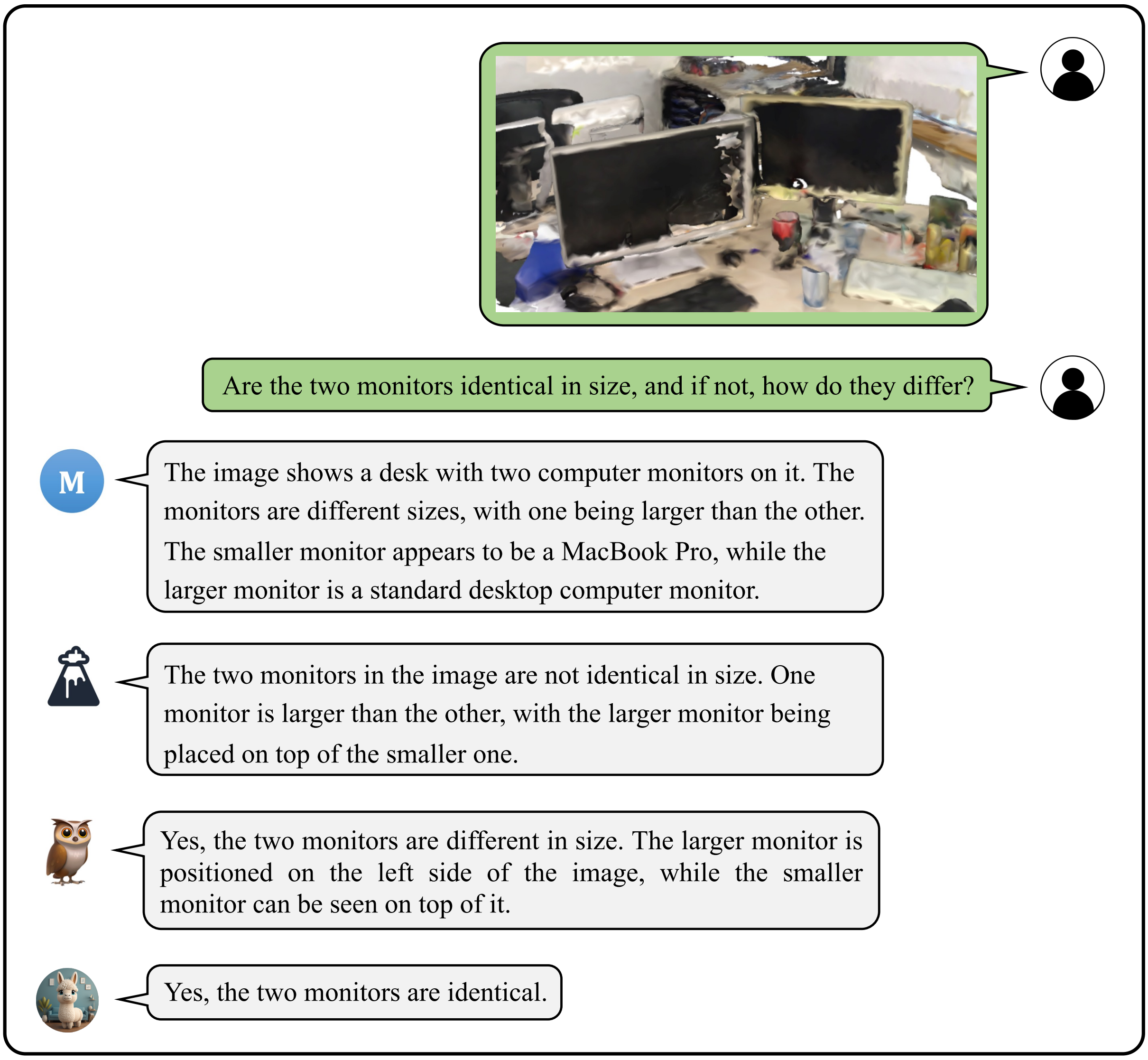
Considering the location of the bed, the nightstand, and the door, is it easy for someone to get in and out of bed comfortably?

Yes, it is possible to easily get in and out of bed due to the location and accessibility of the nightstand and the door. The nightstand and door are positioned near the bed, allowing for easy access without having to walk too far or struggle with reaching the nightstand. The bed is located between the nightstand and the door, providing a clear and direct pathway to and from the bed. Additionally, the door is close to the bed, making it easy to exit and enter the room. Overall, this design allows for smooth and easy movement in and out of the bed, providing a comfortable and convenient experience for the user.

Is the location of the bed optimal in terms of natural lighting and airflow?

Yes, the bed is placed in a corner of the room, maximizing its accessibility and light exposure. It is also located near the window, which allows for natural light and potentially fresh airflow. Additionally, the bed is surrounded by minimal distractions and obstacles, such as chairs or other furniture, indicating that it is situated in a relatively peaceful and organized area of the room. The presence of the curtains on the windows suggests that privacy may be a concern, but the bed's location near the window may also provide some natural light during the day. Overall, the placement of the bed is advantageous for optimal sleeping conditions.

Figure 7: Example 6 of Chat-3D conversation.



Are the two monitors identical in size, and if not, how do they differ?

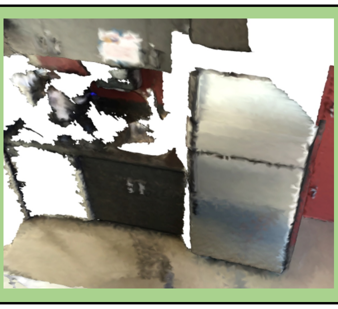
The image shows a desk with two computer monitors on it. The monitors are different sizes, with one being larger than the other. The smaller monitor appears to be a MacBook Pro, while the larger monitor is a standard desktop computer monitor.

The two monitors in the image are not identical in size. One monitor is larger than the other, with the larger monitor being placed on top of the smaller one.

Yes, the two monitors are different in size. The larger monitor is positioned on the left side of the image, while the smaller monitor can be seen on top of it.

Yes, the two monitors are identical.

Figure 8: Example 2 of comparison between Chat-3D and 2D Multi-modal LLMs.



How is the kitchen cabinet positioned concerning the refrigerator and the dishwasher?

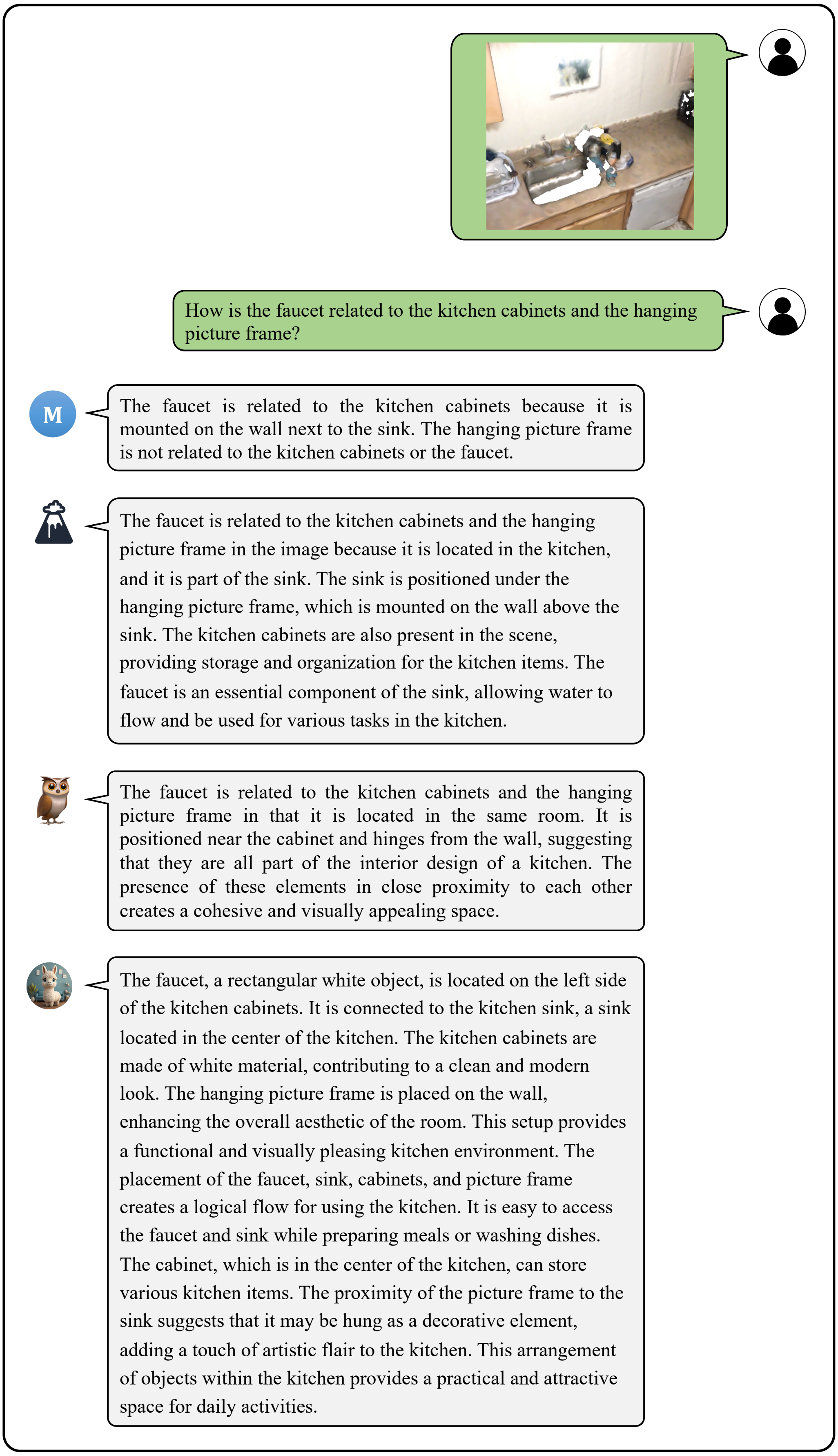
The kitchen cabinet is positioned next to the refrigerator and the dishwasher. The refrigerator is on the left side of the kitchen and the dishwasher is on the right side of the kitchen.

The kitchen cabinet is positioned between the refrigerator and the dishwasher.

The kitchen cabinet is positioned next to the dishwasher and in front of the refrigerator.

The kitchen cabinet is positioned to the left of the refrigerator and situated beneath the dishwasher.

Figure 9: Example 3 of comparison between Chat-3D and 2D Multi-modal LLMs.



How is the faucet related to the kitchen cabinets and the hanging picture frame?

The faucet is related to the kitchen cabinets because it is mounted on the wall next to the sink. The hanging picture frame is not related to the kitchen cabinets or the faucet.

The faucet is related to the kitchen cabinets and the hanging picture frame in the image because it is located in the kitchen, and it is part of the sink. The sink is positioned under the hanging picture frame, which is mounted on the wall above the sink. The kitchen cabinets are also present in the scene, providing storage and organization for the kitchen items. The faucet is an essential component of the sink, allowing water to flow and be used for various tasks in the kitchen.

The faucet is related to the kitchen cabinets and the hanging picture frame in that it is located in the same room. It is positioned near the cabinet and hinges from the wall, suggesting that they are all part of the interior design of a kitchen. The presence of these elements in close proximity to each other creates a cohesive and visually appealing space.

The faucet, a rectangular white object, is located on the left side of the kitchen cabinets. It is connected to the kitchen sink, a sink located in the center of the kitchen. The kitchen cabinets are made of white material, contributing to a clean and modern look. The hanging picture frame is placed on the wall, enhancing the overall aesthetic of the room. This setup provides a functional and visually pleasing kitchen environment. The placement of the faucet, sink, cabinets, and picture frame creates a logical flow for using the kitchen. It is easy to access the faucet and sink while preparing meals or washing dishes. The cabinet, which is in the center of the kitchen, can store various kitchen items. The proximity of the picture frame to the sink suggests that it may be hung as a decorative element, adding a touch of artistic flair to the kitchen. This arrangement of objects within the kitchen provides a practical and attractive space for daily activities.

Figure 10: Example 4 of comparison between Chat-3D and 2D Multi-modal LLMs.