[Image and Vision Computing 137 (2023) 104758](https://doi.org/10.1016/j.imavis.2023.104758)

Contents lists available at [ScienceDirect](http://www.sciencedirect.com/science/journal/02628856)

Image and Vision Computing

journal homepage: [www.elsevier.com/locate/imavis](https://www.elsevier.com/locate/imavis)

Chat 3D: Interactive 3D Reconstruction With the Assistance of Large Language Model Mingang

a *Division of Intelligent Robot,*

b *Department of Computer*

A R T I C L E I N F O

*Keywords:*

3D reconstruction

Large-language model

Deep learning

Image synthesis

A B S T R A C T

Increasing computing power and the expansion of the metaverse market are driving demand for 3D mapping and modeling. Creating a 3D model is a complex process that requires a long production time and a high level of expertise. So, various companies have attempted to address this issue by creating 3D generation models. However, current 3D generation models have limitations in generating objects due to two main reasons. Firstly, it can be difficult for practitioners to precisely depict the objects they want to generate. Secondly, these models tend to synthesize unnecessary objects during the image generation process for 3D reconstruction or may fail to fully construct specific parts.

In this paper, we introduce a Chat3D model that effectively captures the user's intended 3D entity using Named Entity Recognition (NER) and LLM models. Also, we employed specific datasets fit in the target domain for better results. The model is divided into three parts: text processing, image generation, and 3D reconstruction. In text processing, we use NER to analyze the object categories in user-proposed sentences and recommend attributes not described by those objects. We used the LLM to generate sentences that identify and suggest missing attributes in a given sentence. Through fine-tuning the LLM using accurately structured sentences, the LLM becomes capable of presenting attributes specific to the category of each object. In the image generation part, we trained a single object with the background removed to generate a suitable image for 3D generation. In the 3D reconstruction part, the model uses multi-view images reconstructed from a single view.

## Introduction

Recently, the enhanced computing power has made Graphics Processing Units (GPUs) to render a plentiful object. Moreover, the significance of 3D content is progressively growing in various sectors such as medicine and the metaverse markets. 3D digital content is used in a variety of fields, including gaming, entertainment, architecture, engineering, and many others. Nonetheless, manually creating 3D assets can be time-consuming for practitioners and requires a lot of underlying knowledge.

Companies have endeavored to develop 3D generative models, such as Dreamfusion [2] and Magic123 [4], to tackle the issues with the above manual creation. Nevertheless, these current generative models have not been able to achieve adequate success for application in real-world circumstances due to two restrictions. Firstly, practitioners find it challenging to describe the specific objects they wish to generate. Secondly, the accuracy of the generated objects is affected by the lack of personalized learning with target data.

The process of 3D modeling commonly includes the creation of a prototype mesh with low polygon count, which is subsequently transformed into a high-poly mesh to generate the desired structure. Through the development of this prototype, non-technical individuals can directly employ the model without the need for complex commercial 3D design software. Even if the resulting 3D model does not precisely match one's preferences, it can still be used for low-poly prototyping purposes. Furthermore, this approach involving a prototype enables professionals to significantly reduce the amount of time spent.

In this study, we extracted attribute information from household data and provided recommendations related to the missing attributes. Through the interaction with assistance based on LLM, the model devised various illustrative sentences to support practitioners in accurately describing objects. Image generation model was trained using a solitary household dataset, in which background elements were eliminated to ensure that the generated objects do not include undescribed elements. Moreover, our model uses the Controlnet [23] module to modify the intermediate image during the image generation stage for 3D generation. Users can regenerate the image through an additional description text prompt. This process helps to enhance the descriptive capabilities of the objects to be generated.

We have developed a sample framework that produces 3D models possessing substantial descriptive capacity and minimal erroneous information derived from six classes of furniture data which pertain to Beds, Chairs, Dressers, Tables, Lamps, and Sofas. This dataset consists of a single object and a white background that is suitable for generating 3D objects. Our model targeting household data is an example for the methodology introduced in this paper. The model's objective can be altered based on the requirements of the person who is developing the model. Through their own property table and datasets, developers can make their own models. We present a methodology for developing interactive 3D generative models aimed at a specific target.

\* Corresponding author

*E-mail addresses:* mkjang0531@dgist.ac.kr (M. Jang), kky57389@sunmoon.ac.kr (G. Kim), seginp98@gmail.com (S. Park), mahyun91@sunmoon.ac.kr (H. Lee)

1 These authors equally contributed to this paper.

텍스트, 스크린샷, 도표, 멀티미디어 소프트웨어이(가) 표시된 사진

자동 생성된 설명

|  |
| --- |
|  |
| Figure 1. The overview of our text-to 3D reconstruction model (Chat3D). |

**2. Related works**

Recent research on 3D reconstruction using implicit representation is introduced in this section. Models such as Neural Radiance Fields (NeRF) [1] represent 3D space as a function, producing detailed texture and accurate results. Various works related to NeRF are presented, emphasizing the comparisons with previous studies. The strengths and limitations of implicit representation-based 3D reconstruction are discussed.

Dreamfusion [2] introduces a new text-to-3D synthesis approach using a pretrained text-to-image diffusion model. Due to the lack of 3D data, Dreamfusion proposed the Score Distillation Sampling (SDS) loss based on probability density distillation. This proposal enabled the use of a 2D diffusion model as a prior for the optimization of a parametric image generator. Using SDS loss, a randomly initialized 3D model (NeRF) undergoes gradient descent optimization to achieve low loss in its 2D renderings from random angles. This process yields a flexible 3D model capable of being viewed or composed into various environments from any angle. This approach demonstrates the effectiveness of pretrained image diffusion as a prior and SDS loss for text-to-3D synthesis. However, the limited spatial information available in 2D images leads to flat geometry and the Janus problem.

Realfusion [3] proposed a method for a full 360-degree reconstruction of any object from a single view image with the use of a 2D diffusion model. The proposed method synthesized a prompt to generate additional view of the object and also employed a NeRF to reconstruct the object's appearance and geometry faithfully. The researchers showed the possibility of achieving a full 360-degree reconstruction based on a single viewpoint image. Since this approach relies on 2D images, it may generate objects with flat surfaces due to the lack of spatial information of the images. This limitation could potentially lead to the Janus problem.

Magic123 [4] suggested an approach to generate a textured 3D mesh from a single unposed image by using 2D and 3D priors simultaneously. They set the trade-off parameter between using 2D and 3D priors as a hyperparameter and disclosed the optimal hyperparameter found through experiments. However, due to the structure containing 2D and 3D diffusions, the complexity of the 3D generation model increases and requires a lot of computational costs. Dependency on preprocessed segmentation and monocular depth estimation models introduces the risk of errors in these modules affecting the overall quality of the generated content.

ShapE [5] is a conditional generative model designed for creating 3D assets. It directly generates the parameters of implicit functions, enabling the generation of textured meshes and NeRF. The model uses a two-stage training methodology. Initially, an encoder is used to map 3D assets to the parameters of an implicit function. Subsequently, a conditional diffusion model is trained using the encoder outputs. When compared to another generative model for 3D representations, ShapE shows a faster convergence and achieves similar or superior sample quality, despite modeling a higher-dimensional, multi-representation output. ShapE is a powerful method, however, its results can lack fine details and appear rough, indicating the need for enhanced encoders to improve generation quality. This constraint is partially due to the insufficient amount of paired training data, which hinders the model's capacity to create high-quality samples.

**3. Methods**

The key components of this paper include LLM guided prompt recommendation (or text processing) and a user-centric semi-automated system that proceeds through user intervention. The input prompts are complemented by LLM looking for missing representations of style, size, etc. The complemented prompts are then submitted into a text-to-image diffusion prior, which generates numerous images for user selection. Lastly, the selected image is converted into a 3D object using an off-the-shelf 3D reconstruction model.

Using the above mentioned architecture, we developed and experimented with text-to-3D techniques, which will be discussed in more detail below.

**3.1. Text processing**

To enhance the precision of the process of describing the objects, we employ a Large Language Model (LLM) [14] to generate prompts that offer recommendations. Furthermore, we use NER [13] to detect missing tags in sentences. The fine-tuning of the language model is accomplished through the application of instruction fine-tuning techniques [8].

Instruction fine-tuning is a technique for enhancing a language model's capacity to learn zero shots. The foundational principle underlying instruction fine-tuning is to fine-tune a pre-trained language model on a set of tasks by providing explicit instructions to the model during the training process. For instance, if the task is translation, the prompt might consist of the sentence along with additional instruction such as "translate this sentence". Subsequently, the model is trained to generate a response to the query, based on the provided instruction.

**3.1.1. Description Supplementary**

To facilitate the training of the models necessary for text processing, we used the pretrained GPT3.5-turbo [24] to generate sentences pertaining to furniture with multiple recommendations. These sentences were used as the dataset fine-tuning the Flan-T5. To construct the NER dataset, we generated sentences by combining random words related to attributes such as color, material, and furniture type. Furthermore, we generated sentences that described random furniture for the dataset of LLM and subsequently produced additional representations of them.

In this paper, the NER model to detect missing tags is similar to DistilBERT [9]. This NER model was distilled from the BERT model [13] using Knowledge Distilling (KD) [15]. Also, the language models that were used in text processing are based on Flan-T5 [8].

The recommendation model bears resemblance to Transformer [10], albeit with the exclusion of layer normalization bias, the relocation of the residual path beyond layer normalization, and the utilization of distinct position embeddings that are relative positional embeddings [12]. The formulae for the relative position embedding are as follows:

The above formulae are employed to acquire the relative position representations for each relative position within a designated clipping distance, denoted as . The maximum relative position considered is limited to a maximum absolute value of , whereas the clip function guarantees that the relative position falls within this designated range. Subsequently, the acquired representations are used to capture valuable information with respect to the disparities in relative position between the input elements. The is added to the attention, and is added to the output of the sublayer.

The above generation model is instruction finetuned of the model according to architecture described above. The fine-tuning enhances the performance of the unseen task [8]. We used this model for the text processing part because these features are well-suited for handling unseen tasks.

|  |
| --- |
| 텍스트, 스크린샷, 폰트, 라인이(가) 표시된 사진 |
| Figure 2. Example of a prefixed sentence. |

**3.2. Diffusion Methods**

Chat3D is a two-stage coarse-to-fine framework that uses efficient scene models that enable high-resolution text-to-3D synthesis.

**3.2.1 Background**

Diffusion models provide a powerful approach to creating targeted probability distributions by systematically inserting and denoising controlled noise to images. This diffusion process is characterized by a variance schedule , resulting in a sequence of images that shift from their original state to their noisy versions. At each step , noise is injected to the image, and as the process advances, the image shifts from its original state to a noisy version. The noisy image at time is denoted as , where is a sample drawn Gaussian distribution with the same dimensionality as the image , , and . Meanwhile, earning a denoising neural network  is a critical step in diffusion models. This network takes the noisy image ​ as well as the corresponding noise level as inputs and aims to predict the noise component .

To sample from the distribution , one must first sample . Afterwards, the image is progressively denoised by iteratively applying at a predetermined sampling sequence [16, 17, 18] until the termination with sampled from .  
 Modern diffusion models are trained on large collections, , of images by minimizing the following loss:

This model can be easily extended to draw samples from a distribution conditioned on a prompt . Conditioning on the prompt is obtained by adding as an additional input of the network , and the strength of conditioning can be controlled via a classifier-free guidance [19].

**3.2.2. SDS Loss**

Dreamfusion [2] proposed a pioneering text-to-3D approach. The core idea is to implement 3D reconstruction or synthesis by optimizing NeRF using an existing latent diffusion model , rather than training a new diffusion model, which is costly. Specifically, Dreamfusion created a reference image using a text conditioning diffusion model and added noise to the NeRF rendered image when optimizing NeRF so that the loss of the diffusion model could be used. Here, the rendered image is derived via the rendering function , where the arguments of the function are opacity, color, and viewpoint, respectively. This loss is called the Score Distillation Sampling (SDS) loss and is denoted as follows:

where is the NeRF rendered image, is denoising neural network, and is the coefficients at time step . The difference between Eq. (4) and the standard diffusion model objective optimization lies in the absence of the Jacobian term for . Nevertheless, removing this term can enhance the quality of generated output while reducing computational and memory requirements.

**3.2.3. Image generation using Diffusion**

Our challenge is to generate a standalone image without a background that fits the category. The configuration of the model is similar to StableDiffusionXL (SDXL) [20]. The model consists of an encoder part to connect the text prompt with the diffusion prior and a Diffusion Unet part to generate the latent vector that becomes the image matrix. The model also consists of a refiner part to upscale and modify the generated image.

The difference with the existing SDXL is that we used a special image set to generate a suitable image for 3D representation.

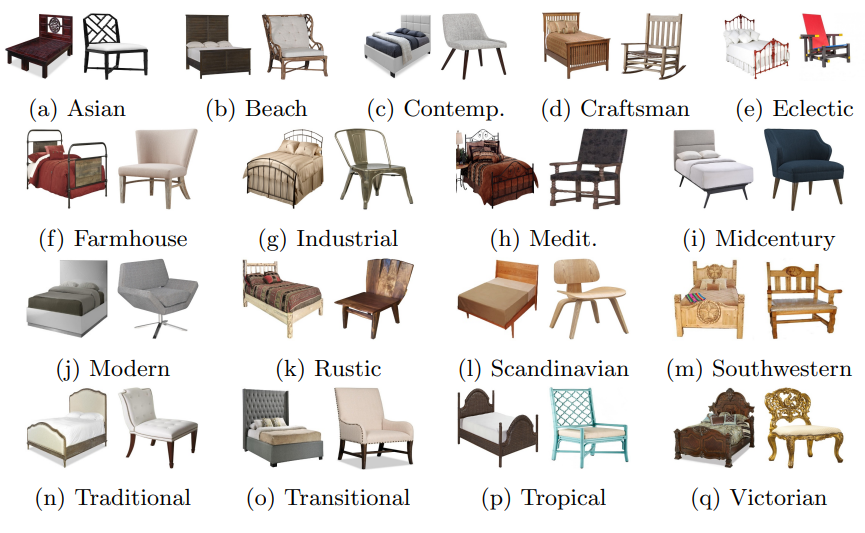


Figure 3. Bonn furniture styles dataset [22].

A typical 3D representation of furniture requires standalone objects. Traditional image diffusion models are trained on photos of the room's foreground and furniture at the same time, so there is a lot of unnecessary information in the generated image. The Bonn furniture styles dataset (Figure 3), published by the International University of Singapore, provides some form of category, style, and tag information for each image. The data used for training consists of 90,298 furniture data divided into 6 categories (Beds, Chairs, Dressers, Tables, Lamps, and Sofas) and 17 styles.

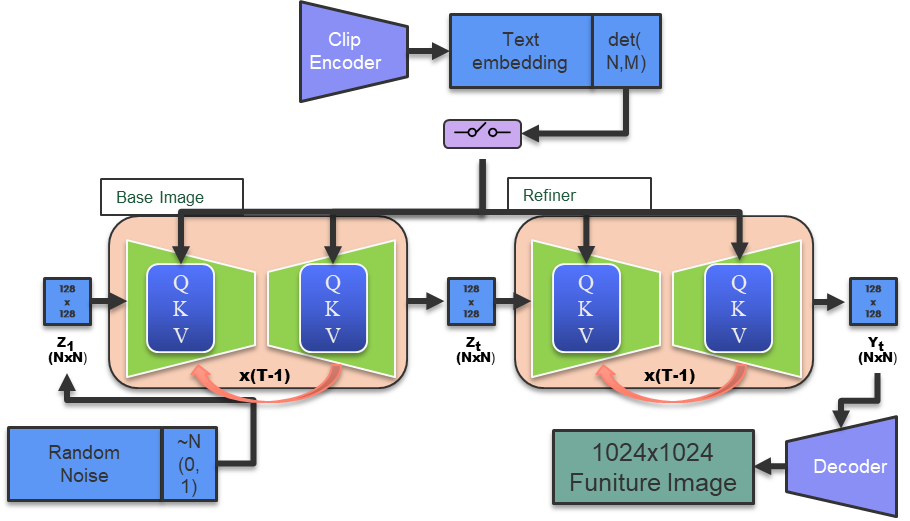


Figure 4. Image generation mechanism.

Free 3D employs an image generation algorithm similar to the diffusion model outlined in SDXL [20].

Initially, the text prompt modified in a previous step is conditioned by an encoder. Following that, a 128 x 128 base image is generated using the aforementioned diffusion algorithm. The base image might lack details regarding the object at hand. In such circumstances, the model uses the refiner to enhance the image, resulting in an accurate standalone image. The refiner processes an image with a latent vector value of 128x128. This is then upscaled in resolution to create a furniture image with dimensions of 1024x1024 through the image decoder.

To provide the user with an interactive experience, the generated image can be modified with a separate prompt via a custom language model. By putting the generated image and a separate prompt into the Controlnet [23] once more, we made modifying the image afterwards possible.

**3.3. 3D Reconstruction**

The 3D reconstruction phase is performed almost identically to the single image reconstructing process presented in Realfusion [3]. Briefly speaking, Realfusion performs 3D reconstruction by optimizing the NeRF model through frozen diffusion and SDS loss [2]. Frozen diffusion is used along with image augmentation to fill in missing spatial (multi-view) information in a single image. To obtain the robust NeRF, the first training stage is optimized with only half of the Instant-NGP hash grid, and the second training stage is optimized with the full hash grid.

Using the images generated in the previous step, Instant-NGP is optimized with the objective function. The key point here is that the objective function is composed by combining the reconstruction and prior losses.

The calculation of both SDS loss and normal vector regularization is conducted for the prior objective. The reconstruction objective entails the calculation of both image and mask losses.

**4. Experiment**

**4.1. Text processing**

In this paper, a NER model was employed to scrutinize user sentences and discern missing tags. Sentences were scrutinized with respect to the object "furniture," and the used tags encompass "color," "size," "style," "material," and "furniture." However, it should be noted that these tags solely represent a subset of furniture characteristics, which may vary when examining sentences associated with other objects. Instances of these tags can be located in Appendix Table 1.

Additionally, LLM was used to generate a suggested sentence based on user input. Since there are no suitable data sets, we manually collected and modified over 500 sentences that describe the household data using GPT 3.5. Sentences that provide supplemental information were used more than those that duplicate input data. For the purpose of fine-tuning instructions, the prefix shown in Figure. 2 was appended to the outset of the input data. Finally, to improve the quality of the results, we compared the finetuned model and the model that included examples in the input data. The results and the example are presented in Appendix Figure 1 and Table 2.

**4.2. Text to Image**

We focus on comparing our model with Stabel Diffusion XL [20] and Midjourney V4 [21] on 50 different text prompt describing furniture. The text prompt is entered into each model with four tags: furniture, no humans, solo, and white background with negative prompts: people, illustration, and low quality.

Since SDXL and our model have an output resolution of 1024 by 1024 while Midjourney V4 is 512 by 512, we upscaled Midjourney V4's output image to 1024 by 1024 overall.

|  |  |  |  |
| --- | --- | --- | --- |
| Text | Generated Image | | |
| Ours | StableDiffusionXL | Midjourney V 4.1  (Upscale) |
| A black leather armchair with a wooden frame | 가구, 의자, 팔걸이이(가) 표시된 사진  자동 생성된 설명 | 가구, 의자, 바닥, 실내이(가) 표시된 사진  자동 생성된 설명 |  |
| A sleek and modern sofa with a plush cushioned seat | 가구, 소파, 푸톤 패드, 스튜디오 소파이(가) 표시된 사진  자동 생성된 설명 | 가구, 소파, 스튜디오 소파, 푸톤 패드이(가) 표시된 사진  자동 생성된 설명 |  |
| A minimalist table with a clean and simple design |  |  |  |
| A simple drawer with modern design |  | KOPPANG Chest of 5 drawers, white, 90x114 cm - IKEA |  |
| A minimalist desk with clean lines |  |  |  |
| A modern white bed with comfortable cushion |  |  |  |
| Figure 5. Qualitative comparison of our model with Stable Diffusion XL & Midjourney | | | |

The model provides accurate image generation for a single piece of furniture with the background removed, giving you the ability to modify the image to more closely match the object you want to create.

**4.3. Image to 3D**

For 3D evaluation, chamfer distance is used to measure the fidelity of reconstructed 3D models to ground truth data. Moreover, the evaluation metrics of PSNR and LPIPS [7] are considered. Specifically, PSNR is used to evaluate fidelity in terms of pixel-wise similarity, while LPIPS is employed to evaluate perceptual similarity considering human visual perception.

In our work, the 3D reconstruction model is borrowed from

Realfusion [3], so the quantitative evaluation metrics are not

much different from the ones in previous works. Therefore, our attention is directed towards the evaluation of improvement resulting from the introduction of text processing by carrying out a qualitative analysis.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Input Text** | **Recommended Text** | **Text to Image** | **Image to 3D** | | | |
| “An armchair” | “A black velvet armchair with gold legs” |  | **가구, 의자, 팔걸이, 푸톤 패드이(가) 표시된 사진  자동 생성된 설명** | **의자, 가구, 팔걸이, 실내이(가) 표시된 사진  자동 생성된 설명** | **의자, 가구, 스케치, 팔걸이이(가) 표시된 사진  자동 생성된 설명** | **의자, 스케치, 가구, 예술이(가) 표시된 사진  자동 생성된 설명** |
| “An elegant armchair with a plush cushioned seat and backrest” |  | **가구, 의자, 팔걸이, 클럽 의자이(가) 표시된 사진  자동 생성된 설명** | **의자, 가구, 팔걸이, 클럽 의자이(가) 표시된 사진  자동 생성된 설명** | **가구, 의자, 팔걸이, 실내이(가) 표시된 사진  자동 생성된 설명** | **의자, 가구, 팔걸이, 흑백이(가) 표시된 사진  자동 생성된 설명** |
| “A black leather armchair with a wooden frame” |  | **가구, 팔걸이, 의자, 가죽이(가) 표시된 사진  자동 생성된 설명** | **가구, 실내, 미니어처, 바닥이(가) 표시된 사진  자동 생성된 설명** | **가구, 스케치, 의자이(가) 표시된 사진  자동 생성된 설명** | **가구, 의자, 오토만, 실내이(가) 표시된 사진  자동 생성된 설명** |
| “A sofa” | “A sleek and modern sofa” |  | **가구, 소파, 스튜디오 소파, 푸톤 패드이(가) 표시된 사진  자동 생성된 설명** | **가구, 소파, 푸톤 패드, 실내이(가) 표시된 사진  자동 생성된 설명** | **가구, 소파, 푸톤 패드, 실내이(가) 표시된 사진  자동 생성된 설명** | **스케치, 흑백, 예술이(가) 표시된 사진  자동 생성된 설명** |
| “A single sofa with a plush cushioned seat” |  | **가구, 소파, 스튜디오 소파, 푸톤 패드이(가) 표시된 사진  자동 생성된 설명** | **소파, 푸톤 패드, 가구, 쿠션이(가) 표시된 사진  자동 생성된 설명** | **스케치, 예술이(가) 표시된 사진  자동 생성된 설명** | **스케치이(가) 표시된 사진  자동 생성된 설명** |
| “A luxurious leather sofa with tufted backrest" |  | **가구, 소파, 스튜디오 소파, 푸톤 패드이(가) 표시된 사진  자동 생성된 설명** | **가구, 푸톤 패드, 오토만, 스툴이(가) 표시된 사진  자동 생성된 설명** | **소파, 가구이(가) 표시된 사진  자동 생성된 설명** | **박격포, 오토만이(가) 표시된 사진  자동 생성된 설명** |
| “A wooden chair” | “A wooden armchair with intricate carvings” |  | **가구, 의자, 팔걸이, 실내이(가) 표시된 사진  자동 생성된 설명** | **가구, 의자, 팔걸이, 실내이(가) 표시된 사진  자동 생성된 설명** | **의자이(가) 표시된 사진  자동 생성된 설명** | **스케치, 그림, 의자, 예술이(가) 표시된 사진  자동 생성된 설명** |
| “A wooden chair with a plush cushioned seat” |  | **가구, 의자, 팔걸이이(가) 표시된 사진  자동 생성된 설명** | **가구, 의자, 실내이(가) 표시된 사진  자동 생성된 설명** | **스케치, 의자, 흑백, 예술이(가) 표시된 사진  자동 생성된 설명** | **스케치, 의자, 가구, 예술이(가) 표시된 사진  자동 생성된 설명** |
| “A rustic wooden chair” |  | **가구, 의자이(가) 표시된 사진  자동 생성된 설명** | **가구, 의자이(가) 표시된 사진  자동 생성된 설명** | **가구, 의자이(가) 표시된 사진  자동 생성된 설명** | **가구, 의자이(가) 표시된 사진  자동 생성된 설명** |
| Figure 6. **Qualitative results of 3D reconstruction.** Our text-to-3D model can reconstruct various objects from simple prompt such as “A wooden chair” through the support of LLM. | | | | | | |

**5. Conclusion**

In this study, we present the concept of text-to-3D reconstruction. Through the use of a LLM, we can identify missing tags and supplement prompts that describe objects. This approach leads to a more user-friendly image generation in the text-conditioning diffusion model.

Three-dimensional objects have a more complex structure than 2D images. As of now, generative models are in their infancy, and generating complex objects is challenging. The training dataset needs to be large and sophisticated, and it needs to be able to describe the object well.

At this stage, it's difficult for a general purpose 3D model to satisfy these conditions. We solved this problem by targeting household data as an example. We created an interactive

environment with users through an LLM-based assistance to describe objects well.

We improved the quality of the dataset by using a dataset specialized in 3D generation.

The targets introduced in this paper are household data, but the methodology can be applied to any target for 3D generation.

**Acknowledgements**

This research was supported by the MISP (Ministry of Science, ICT & Future Planning), Korea, under the National Program for Excellence in SW supervised by the IITP (Institute for Information & communications Technology Promotion).

**Appendix**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table1. NER Tag Examples. | | | | |
| Size | Material | Design | Color | Furniture |
| Tiny | Marble top metal base | Ladder-style frame | Royal blue | Rocking bench |
| Large shade | Wool | Rustic elements | Magenta | Headboard |
| Mini | Wrought iron | Luxurious atmosphere | Coral | Built-in keyboard tray |
| minimalist | Frosted glass | Artistic look | Vibrant colors | Bookshelf |
| Smaller | Marble | Texture | Stunning ruby red | Desk lamp |
| Moderate-sized | Acrylic | Luxury to the space | Mint green color | Stool |
| Wide | Ottoman | Relaxation | Cobalt blue | Cabinets |
| Oversized | Hammered metal finish | Flexibility | Red | Armchair |
| Compact | Plush velvet | Vintage-inspired design | Sky blue | Wardrobe closet |
| Lilliputian side | Velvet | Cozy outdoor seating area | Soft cream-colored | Dining table |
| Petite | Fabric | Elegant vibe | Serene shade | Shoe rack |
| Slender | UV-resistant fabric | Warm | Aqua | Chandelier |
| Standard | Sustainable bamboo | Charming vintage | Gray | Sideboard |
| Small storage | Sleek stainless steel | Floral-patterned | Olive green | Wall shelves |
| Average-sized | Mahogany wood | Comfortable sleep | Yellow | Tables |
| Enough space | Glass | Whimsy | Deep navy | Bench |
| Expansive views | Rattan | Extendable leaves | Colorful | Lounge chairs |
| Medium-sized | Round marble | Modern touch | Pops of color | LED lighting |
| Miniature | Teak wood | Industrial feel | White | Dresser |
| Generous | Textured fabrics | Eco-friendly | Purple | Closet |
| Large | Crystal | Artistic focal point | Teal | Desk |
| Fitting perfectly | Wood | Warm glow | Plum | Sofa |
| Thin | Metal | Combining style | Dark | Ottoman |
| Spacious | Bamboo | Modern look | Baby blue | Coffee table |
| Sizable | Soft velvet fabric | Cozy | Black | Hammock |
| Various Sizes | Stain-resistant fabric | Charm | Ivory | Wine rack |
| Additional storage | Tempered glass | Unique design element | Navy blue | Floor lamp |
| Sizable sectional | Made of oak wood | Contemporary touch | Cyan | Chair |
| Proportional | Solid oak | Bold floral print | Pink | Bar table |
| Different heights | Bronze | Natural elements | Blue | Bed |

|  |
| --- |
| 텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진  자동 생성된 설명**1** |
| Figure 7. The prefix used to finetune the model. |

|  |  |
| --- | --- |
| Table 2. Comparison of the model applying prompt engineering and the finetuning model. | |
| Input Text | |
| Apply Prompt Engineering | Apply Fine-tuning |
| A white chair with handles the size of a large chair. | |
| A white chair with handles the size of a large chair.. The chair is in a living room.. The chair is in a dining room.. | White Office Chair with Adjustable Handles, White Dining Chair with Upholstered Handles, White Rocking Chair with Adjustable Handles and Footrest |
| Iron legged chair with white handles. | |
| A chair with white handles and an iron leg.. The chair is a classic design.. The chair is a classic design.. | White Iron Legged Chair with Adjustable Handles, White Iron Legged Chair with Chrome Legs, White Iron Legged Chair with Wooden Legs |
| White oversized iron legged chair. | |
| A white oversized iron legged chair. It has a curved back and a curved seat. It is a great chair for a living room or a dining room. | White Oversized Iron Legged Chair with Adjustable Height, White Oversized Iron Legged Chair with Metal Frame, White Oversized Iron Legged Chair with Wooden Legs |
| White chair with large handles. | |
| A white chair with large handles.. It is a classic design.. It is a classic design.. It is a classic design.. It is a classic design.. It is a classic design.. It | White Office Chair with Large Handles, White Dining Chair with Large Handles, White Rocking Chair with Large Handles and Armrests. |
| Chair with large white handles. | |
| A chair with large white handles.. The chair is a white chair.. The chair is a white chair.. The chair is a white chair.. | White Office Chair with Large White Handles, White Dining Chair with Large White Handles, White Rocking Chair with Large White Handles and Armrests |
| A white chair with a handle in the middle. | |
| A white chair with a handle in the middle. The chair is a modern design. The chair is a modern design. The chair is a modern design. | White Office Chair with Adjustable Handles, White Dining Chair with Wooden Legs, White Rocking Chair with Adjustable Handles and Footrest |
| Small white chair for children. | |
| A small white chair for children. It is a good chair for children. It is a good chair for children. It is a good chair for children. | Small White Chair for Children with Adjustable Height, Small White Chair for Children with Storage Compartment, Small White Chair for Children with Built-in Storage |
| White small iron legged chair. | |
| A small white chair with a small iron leg.. The chair is in a living room.. The chair is in a small living room. | White Small Iron Legged Chair with Adjustable Height, White Small Iron Legged Chair with Metal Frame, White Small Iron Legged Chair with Wooden Legs |
| Brown large chair without handles. | |
| A large brown chair without handles. It is a modern design. It is a modern design. It is a modern design. It is a modern design. | Brown Large Chair without Handles, Brown Large Chair without Handles, Brown Large Chair without Handles with Adjustable Height and Backrests, Brown Large Chair without Handles with Built-in Storage |
| Brown large iron-legged chair without handles. | |
| A brown large iron-legged chair without handles. It is a modern design. It is a modern design. It is a modern design. | Brown Large Iron-Legged Chair without Handles, Brown Large Iron-Legged Chair without Handles, Brown Large Iron-Legged Chair without Handles |
| Brown large chair without handles. | |
| A large brown chair without handles. It is a modern design. It is a modern design. It is a modern design. It is a modern design. | Brown Large Chair without Handles, Brown Large Chair without Handles, Brown Large Chair without Handles with Adjustable Height and Backrests, Brown Large Chair without Handles with Built-in Storage |
| Iron-legged chair with brown middle handle. | |
| A chair with a brown middle handle and iron legs. The chair is a modern design. It is a chair. The chair is a modern design. | Iron-legged Chair with Brown Middle Handle, Iron-legged Chair with Brown Middle Handle, Iron-legged Chair with Brown Middle Handle |
| Small brown chair with iron legs. | |
| A small brown chair with iron legs.. It is a good chair for a living room.. It is a good chair for a living room.. | Small Brown Leather Chair with Iron Legs, Small Brown Dining Chair with Iron Legs, Small Brown Office Chair with Iron Legs and Upholstery |
| Large black iron legged chair with handles. | |
| A large black iron legged chair with handles. It has a black leather seat. It has a black leather back. It has a black leather seat. | Large Black Iron Legged Chair with Adjustable Handles, Large Black Iron Legged Chair with Adjustable Handles, Large Black Iron Legged Chair with Adjustable Handles |
| Black iron-legged chair with large handles. | |
| A black iron-legged chair with large handles. It is a classic design. It is a chair. It is a chair. | Black Iron-Legged Chair with Large Handles, Black Iron-Legged Chair with Adjustable Handles, Black Iron-Legged Chair with Large Handles |
| Chair with large black handles. | |
| A chair with large black handles. The chair is a black armchair. The chair is a sleek and minimalist black chair designed with clean lines and a contemporary aesthetic. | Black Leather Executive Chair with Adjustable Handles, Black Rocking Chair with Adjustable Handles, Black Dining Chair with Adjustable Handles and Footrest |
| Black large armless chair with iron legs. | |
| A black armless chair with iron legs. It is a classic design. It is a large chair. It is a large chair. | Black Large Armless Chair with Iron Legs, Black Large Armless Chair with Iron Legs, Black Large Armless Chair with Iron Legs |
| Iron legged chair with black middle handle. | |
| A chair with a black middle handle and iron legs. It is a classic design. It is a chair. It is a chair. | Iron Legged Chair with Black Middle Handle, Iron Legged Chair with Black Middle Handle, Iron Legged Chair with Black Middle Handle |
| Small black chair with handles. | |
| A small black chair with handles. The chair is a classic design. It is a classic design. It is a classic design. | Small Black Leather Chair with Adjustable Handles, Small Black Dining Chair with Upholstered Handles, Small Black Office Chair with Adjustable Handles |
| Large oak chair without handles. | |
| A large oak chair without handles. It is a classic design. It is a great chair for any room. It is a great chair for any room. | Large Oak Dining Chair without Handles, Large Oak Office Chair without Handles, Large Oak Dining Chair without Handles and Armrests. |
| A chair with large oak handles. | |
| A chair with large oak handles.. The chair is a modern design.. The chair is a modern design.. The chair is a modern design.. The chair is a modern design.. The chair is a | A chair with large oak handles, A chair with large oak handles with a curved backrest, A chair with large oak handles with a curved seat |
| Oak chair with checkered pattern. | |
| The chair is made of oak and has a checkered pattern.. It is a classic design.. It is a chair. | Oak Chair with Checkered Pattern, Oak Dining Chair with Checkered Pattern, Oak Lounge Chair with Checkered Pattern and Upholstery. Oak Dining Chair with Checkered Pattern and Upholstery. |
| Oak iron legged chair without large handles. | |
| The chair is made of oak and iron. It has a large back and a small seat. It has a large back and a small seat. | Oak Iron Legged Chair without Large Handles, Oak Iron Legged Chair without Large Handles, Oak Iron Legged Chair without Large Handles |
| Iron-legged chair with no handles in the middle of the oak color. | |
| A chair with no handles in the middle of the oak color. The chair is made of iron. The chair is made of oak. The chair is made of iron. | Oak-colored Iron-Legged Chair with No Handles, Oak-colored Iron-Legged Chair with Adjustable Armrests, Oak-colored Iron-Legged Chair with Wooden Legs |
| White large iron legged desk. | |
| A white large iron legged desk. It has a white top and a white base. It has a white top and a white base. | White Large Iron Legged Desk with Adjustable Height, White Large Iron Legged Desk with Built-in Shelves, White Large Iron Legged Desk with Built-in Drawers |
| A large white desk. | |
| A large white desk with a white top.. The desk is in a modern style.. The desk is in a modern style. | White Writing Desk with Built-in Drawers, White Standing Desk with Adjustable Height, White Standing Desk with Built-in Lighting and USB Ports |
| White large iron legged desk. | |
| A white large iron legged desk. It has a white top and a white base. It has a white top and a white base. | White Large Iron Legged Desk with Adjustable Height, White Large Iron Legged Desk with Built-in Shelves, White Large Iron Legged Desk with Built-in Drawers |
| White desk with big iron legs. | |
| The desk is white with big iron legs. It has a big desk with big iron legs. It has a big desk with big iron legs. | White Writing Desk with Big Iron Legs, White Standing Desk with Big Iron Legs, White Standing Desk with Big Iron Legs and Storage Drawers |
| Middle white leg iron desk. | |
| The desk is made of iron and has white legs. It has a white top and a black bottom. The desk is in the middle of the room. | White Iron Desk with Middle Legs, White Writing Desk with Iron Legs, White Standing Desk with Iron Legs and Keyboard Trays |
| White middle desk with iron legs. | |
| The desk is white with iron legs. It is a middle desk. The desk is a white middle desk with iron legs. The desk is a white middle desk with iron legs. | White Middle Desk with Iron Legs, White Middle Desk with Iron Legs and Storage, White Middle Desk with Iron Legs and Adjustable Height |
| White small desk for children. | |
| The desk is small and white. It is for children. It is a desk for children. It is a small desk for children. | White Small Desk for Children with Adjustable Height, White Small Desk for Children with Built-in Drawers, White Small Desk for Children with Built-in Storage |

**References**

1. Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R. (2021). Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, *65*(1), 99-106.
2. Poole, B., Jain, A., Barron, J. T., & Mildenhall, B. (2022). Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*.
3. Melas-Kyriazi, L., Laina, I., Rupprecht, C., & Vedaldi, A. (2023). Realfusion: 360deg reconstruction of any object from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8446-8455).
4. Qian, G., Mai, J., Hamdi, A., Ren, J., Siarohin, A., Li, B., ... & Ghanem, B. (2023). Magic123: One image to high-quality 3D object generation using both 2D and 3D diffusion priors. *arXiv preprint arXiv:2306.17843*.
5. Jun, H., & Nichol, A. (2023). Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*.
6. Müller, T., Evans, A., Schied, C., & Keller, A. (2022). Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, *41*(4), 1-15.
7. Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 586-595).
8. Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., ... & Le, Q. V. (2021). Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.
9. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108.*
10. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30.*
11. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research, 21(1),* 5485-5551
12. Shaw, P., Uszkoreit, J., & Vaswani, A. (2018). Self-attention with relative position representations.*arXiv preprint arXiv:1803.02155.*
13. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding.*arXiv preprint arXiv:1810.04805.*
14. Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223.
15. Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network.*arXiv preprint arXiv:1503.02531.*
16. Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models.*Advances in Neural Information Processing Systems, 33,* 6840-6851.
17. Liu, L., Ren, Y., Lin, Z., & Zhao, Z. (2022). Pseudo numerical methods for diffusion models on manifolds.*arXiv preprint arXiv:2202.09778.*
18. Song, J., Meng, C., & Ermon, S. (2020). Denoising diffusion implicit models.*arXiv preprint arXiv:2010.02502.*
19. Dhariwal, P., & Nichol, A. (2021). Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems, 34,* 8780-8794.
20. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., ... & Rombach, R. (2023). Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952 .*
21. Midjourney. [<https://www.midjourney>.co](https://www.midjourney.co)m/
22. Aggarwal, D., Valiyev, E., Sener, F., & Yao, A. (2019). Learning style compatibility for furniture. In *Pattern Recognition: 40th German Conference, GCPR 2018, Stuttgart, Germany, October 9-12, 2018, Proceedings 40* (pp. 552-566). Springer International Publishing.
23. Zhang, L., Rao, A., & Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 3836-3847).
24. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877-1901.