

Style-based Text to 3D Furniture Object Generation

김건형, 주성민, 최대한

<https://github.com/edenkim9741/AI-System-Project.git>

Contents

1	Motivation	2
2	Related Works	2
2.1	Histogram-based Similarity Metrics	2
2.1.1	Chi-Square Distance	2
2.1.2	Bhattacharyya Distance	2
2.2	Text to 3D Object Generation	3
3	Dataset	3
3.1	Color Analysis Dataset	3
3.2	Size Prediction Dataset	3
4	Methods	3
4.1	Dominant Color Extraction by Styles	3
4.2	Text Finetuning for Furniture Generation	4
4.3	Color aware 3D Object Generation	4
4.4	Furniture Size Prediction	4
5	Experiments and Results	5
5.1	Dominant Color Extraction Method Comparison	5
5.2	Color Histogram Comparison by Styles	6
5.2.1	Color Histogram in RGB Space	6
5.2.2	Color Histogram in HSV Space	7
5.3	Text Finetuning Results	8
5.4	3D Object Generation	11
5.5	Size Prediction	11
6	Conclusion	11

1 Motivation

최근 1인 가구가 증가함에 따라 인테리어에 대한 관심도 높아지고 있다. 하지만 인테리어를 처음 접하는 사람들은 어떤 인테리어 스타일의 가구를 선택해야 할지, 어떤 색상의 가구를 배치해야 할지 결정하기 어렵다. 또한, 가구를 실제로 구매하기 전에는 가구가 어떻게 배치될지 상상하기 어렵다. 이러한 문제를 해결하기 위해, 본 프로젝트에서는 사용자가 입력한 텍스트를 기반으로 3D 가구 객체를 생성하는 방법을 제안한다. 이 방법은 사용자의 입력으로 통해 원하는 스타일의 가구를 3D 객체로 생성해줄뿐만 아니라, 각 스타일 별로 주로 사용되는 색을 분석하여 자동으로 가구들의 색상을 매치시켜준다. 또한, 생성된 3D 가구 객체와 함께 가구의 적절한 크기도 함께 생성되며, 해당 객체를 3D 그래픽 툴을 통해 사용자가 원하는 공간에 배치할 수 있도록 obj, mtl 등의 확장자로 저장할 수 있다.

2 Related Works

2.1 Histogram-based Similarity Metrics

2.1.1 Chi-Square Distance

카이제곱 거리는 두 히스토그램 H_1 과 H_2 간의 차이를 측정하는 거리 기반 유사도 지표로, OpenCV 라이브러리[1]에서는 다음과 같이 정의된다:

$$d(H_1, H_2) = \sum_I \frac{(H_1(I) - H_2(I))^2}{H_1(I)}$$

여기서 I 는 히스토그램의 모든 빈을 나타내며, $H_1(I)$ 과 $H_2(I)$ 는 각각 히스토그램 H_1 과 H_2 의 빈 I 에 해당하는 값이다. 0~1 사이의 값을 가지며, 값이 작을수록 두 히스토그램이 유사하다는 것을 의미한다. 색상 히스토그램의 경우, 히스토그램 간의 세부적인 차이를 민감하게 포착하기 때문에 색상 분포의 작은 변화도 감지할 수 있다.

2.1.2 Bhattacharyya Distance

바타차리야 거리는 두 분포간의 중첩 정도를 측정한다. OpenCV 라이브러리[1]에서는 다음과 같이 정의된다:

$$d(H_1, H_2) = \sqrt{1 - \frac{1}{\sqrt{\bar{H}_1 \bar{H}_2} N^2} \sum_I \sqrt{H_1(I) \cdot H_2(I)}}$$

여기서 \bar{H}_1 과 \bar{H}_2 는 각각 히스토그램 H_1 과 H_2 의 정규화된 값이며, N 은 히스토그램의 빈의 개수이다. 0~1 사이의 값을 가지며, 값이 작을수록 두 히스토그램이 유사하다는 것을 의미한다. 바타차리야 거리는 분포의 전반적인 형태와 겹치는 영역을 고려하기 때문에 이미지의 색상 분포가 부분적으로 일치하거나 위치가 미세하게 달라지는 경우에도 두 이미지를 유사하다고 판단할 수 있다.

2.2 Text to 3D Object Generation

최근에는 텍스트를 입력으로 받아 3D 객체를 생성하는 연구가 활발히 진행되고 있다. 대표적으로 Shap-E[2] 모델은 텍스트와 이미지로부터 3D 객체를 생성할 수 있는 모델로, 다양한 조건을 반영하여 3D 객체를 생성할 수 있다. Shap-E는 CLIP[3]을 이용하여 텍스트 또는 이미지를 임베딩하고, 이를 기반으로 3D 객체를 생성한다. CLIP을 통해 텍스트와 이미지의 조건을 반영하여 3D 객체를 생성할 수 있기 때문에, 본 프로젝트에서 사용자의 입력을 기반으로 가구를 생성하는데 적합하다.

3 Dataset

3.1 Color Analysis Dataset

각 인테리어 스타일 별로 주로 사용되는 색상을 분석하기 위해 다양한 인테리어 스타일의 이미지들을 수집하였다. 오늘의 집[4]의 스타일 카테고리를 참고하여 다음과 같은 7가지의 대표적인 인테리어 스타일을 선정하였다: Antique, Modern, Natural, Northern European, Romantic, Traditional Korean Style, Vintage. unsplash[5], pexels[6], pixabay[7]에서 각 스타일 별로 100장 이상의 이미지를 크롤링하였다.

3.2 Size Prediction Dataset

가구의 적절한 크기를 예측하기 위해 3D-Front[8] 데이터셋을 사용하였다. 3D-FRONT는 다양한 가구가 배치되어 있는 3D 공간을 제공하는 데이터셋으로, 각 가구의 크기 정보가 포함되어 있다. 또한 각 가구의 클래스 정보도 포함되어 있어, 가구의 종류에 따라 적절한 크기를 예측하는데 사용할 수 있다. 본 프로젝트에서는 3D-FRONT 데이터셋에서 제공하는 가구의 이미지와 크기 정보를 활용하여 가구의 크기를 예측하는 모델을 학습하였다.

4 Methods

4.1 Dominant Color Extraction by Styles

사용자의 입력으로부터 인테리어에 사용할 색상을 결정하기 위해 인테리어 스타일 별로 주로 사용되는 색상을 분석하였다. K-Means 클러스터링, Mean-Shift 클러스터링, DBSCAN의 방법으로 대표 색상 5개를 추출해보고 각 추출 방법의 결과를 정성적으로 비교하였다. 또한, RGB, HSV 색 공간에서의 Color Histogram을 비교하여 인테리어 스타일 간의 거리를 분석하였다.

$$\mathcal{S}^{style} = [\mathbf{c}_1^{style}, \mathbf{c}_2^{style}, \dots, \mathbf{c}_k^{style}] = \text{ExtractDominantColors}(\mathbf{I}^{style}, k), \quad (1)$$

여기서 \mathcal{S}^{style} 은 스타일 별로 추출된 대표 색상들, \mathbf{c}_i^{style} 은 i 번째 Dominant Color, \mathbf{I}^{style} 은 스타일 별로 수집된 이미지, k 는 추출할 대표 색상의 개수이다.

4.2 Text Finetuning for Furniture Generation

사용자의 입력을 기반으로 가구들에 대한 설명을 생성하기 위해 ollama에서 Llama3.2-vision:11b[9]를 사용하였다. 프롬프트 엔지니어링을 통해서 사용자가 입력한 텍스트를 기반으로 가구에 대한 설명을 생성하도록 모델을 파인튜닝하였다. 예를 들어, “A modern room contains round items”와 같은 입력이 주어지면, 모델은 해당 스타일과 색상을 반영한 가구를 생성하기 위한 텍스트를 아래와 같이 생성한다:

“Create a modern glass table with round shape”,
“Create a modern wooden chair with round backrest”,
...

이렇게 생성된 텍스트는 3D 객체 생성을 위한 입력으로 사용된다. 또한, 생성된 가구의 종류는 3D-FRONT 데이터셋에서 제공하는 가구의 클래스 정보와 매칭되어, 가구의 크기를 예측하는데 사용된다.

4.3 Color aware 3D Object Generation

3D 객체 생성을 위해 Shap-E[2] 모델을 사용하였다. Shap-E 모델은 Clip[3]을 이용하여 텍스트를 임베딩하고 이를 기반으로 3D 객체를 생성하기 때문에, 자연어로 표현된 색은 입력으로 사용될 수 있다. 하지만 사전에 추출된 스타일별 대표 색상은 컬러 코드로 표현되기 때문에 이를 자연어로 변환하는 과정이 필요하다. 예를 들어, RGB 색상 (255, 0, 0)은 “red”로 변환될 수 있다. 정확하게 일치하지 않는 대표 색상이 있을 수 있기 때문에, RGB 공간에서 가장 가까운 색상으로 변환하는 방법을 사용하였다. 본 프로젝트에서는 대표 색상을 자연어로 변환하기 위해 148개의 색상 이름과 RGB 색상 값을 매핑한 자료를 활용하였다. 해당 자료는 matplotlib에서 제공하는 CSS/XKCD 색상 이름을 기반으로 제작하였다.

대표 색상을 자연어로 변환하는 과정은 다음과 같이 표현할 수 있다:

$$\begin{aligned}\tilde{\mathbf{c}}_i^{style} &= \operatorname{argmin}_{\mathbf{c} \in \mathcal{C}} \|\mathbf{c} - \mathbf{c}_i^{style}\|_2, \\ \mathcal{C} &= \{\text{red, green, blue, } \dots\},\end{aligned}\tag{2}$$

여기서 $\tilde{\mathbf{c}}_i^{style}$ 는 i 번째 Dominant Color, \mathcal{C} 는 자연어로 표현된 색상의 집합이다. 이 과정을 통해 스타일 별로 추출된 대표 색상을 자연어로 표현된 색상 중 하나로 변환할 수 있다.

변환된 대표 색상 텍스트는 Shap-E 모델의 입력으로 사용되며, 사용자가 입력한 텍스트와 함께 3D 객체를 생성하는데 활용된다. 예를 들어, “A modern glass table with round shape, dominant colors are white, black”과 같은 입력이 주어지면, Shap-E 모델은 해당 스타일과 색상을 반영한 3D 의자 객체를 생성한다.

4.4 Furniture Size Prediction

가구의 적절한 크기를 예측하기 위해서는 가구를 설명하는 텍스트와 가구의 크기를 매칭하는 데이터셋이 필요하다. 하지만 그러한 데이터셋은 존재하지 않았기 때문에, Shap-E 모델이 입력으로 이미지와

텍스트를 받는다는 점을 활용하여, 3D-FRONT 데이터셋에서 제공하는 가구의 이미지와 크기 정보를 이용하여 가구의 크기를 예측하는 모델을 학습하였다.

크기 예측 모델은 Shap-E로부터 추출된 latent feature를 Fully Connected Layer를 통해 Projection 하고 가구의 종류를 one-hot 인코딩한 벡터를 concatenate하여 다시 한 번 Projection한 후, 가구의 크기를 예측하는 구조로 설계되었다. 수식으로 표현하면 다음과 같다:

$$\mathbf{s} = \text{FC}([\text{FC}(\mathbf{f}_{\text{latent}}), \mathbf{v}_{\text{class}}]), \quad (3)$$

여기서 \mathbf{s} 는 예측된 가구의 크기, $\mathbf{f}_{\text{latent}}$ 는 Shap-E로부터 추출된 latent feature, $\mathbf{v}_{\text{class}}$ 는 가구의 종류를 one-hot 인코딩한 벡터, FC는 fully connected layer를 의미한다.

5 Experiments and Results

5.1 Dominant Color Extraction Method Comparison

K-Means 클러스터링, Mean-Shift 클러스터링, DBSCAN의 방법으로 대표 색상 5개를 추출한 결과는 Fig 1과 같다

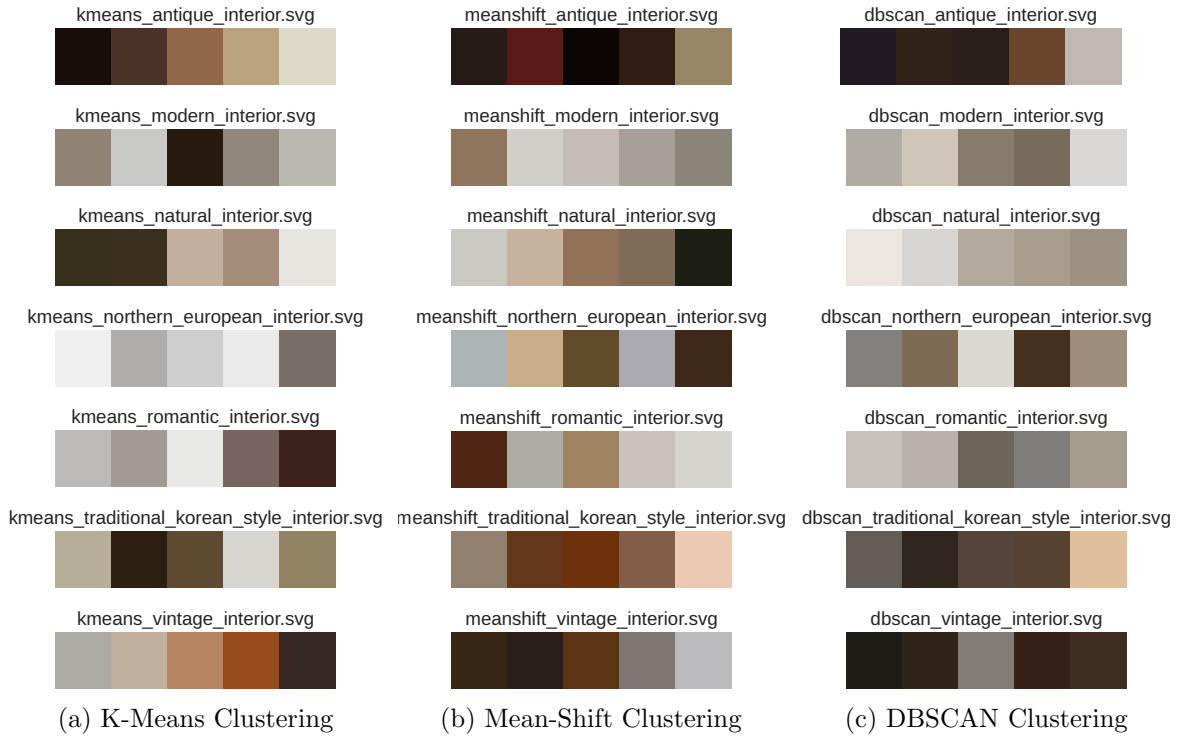


Figure 1: Dominant Color Extraction Results

각 색 추출 방법에 대해 정성적인 비교를 진행하였다. K-Means 클러스터링은 여러 색을 다양하게 추출하면서도 인테리어의 느낌을 잘 살리는 색상을 추출하였고, Mean-Shift 클러스터링은 전체적으로 더 강한 명도의 색을 추출하였다. DBSCAN의 경우에는 색의 다양성을 잘 살리지 못하고, 3 4개의 색상이 비슷하게 추출되는 경향을 보였다.

Fig 1 같은 결과를 토대로 본 프로젝트에서는 K-Means 클러스터링을 사용하여 대표 색상을 추출하기로 결정하였다.

5.2 Color Histogram Comparison by Styles

색 공간을 달리하며 색 히스토그램을 추출하고 카이제곱 거리와 바타차리야 거리를 통해 비교하였다.

5.2.1 Color Histogram in RGB Space

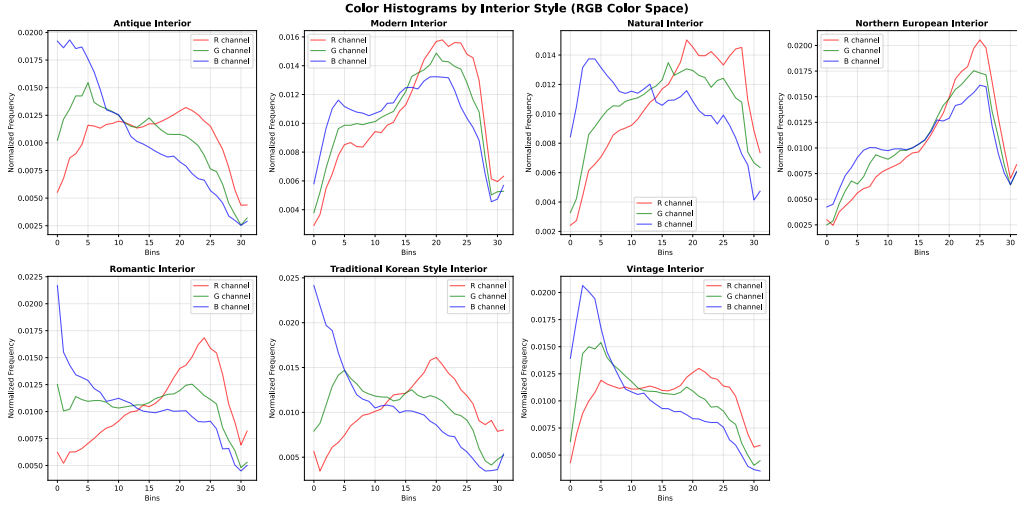
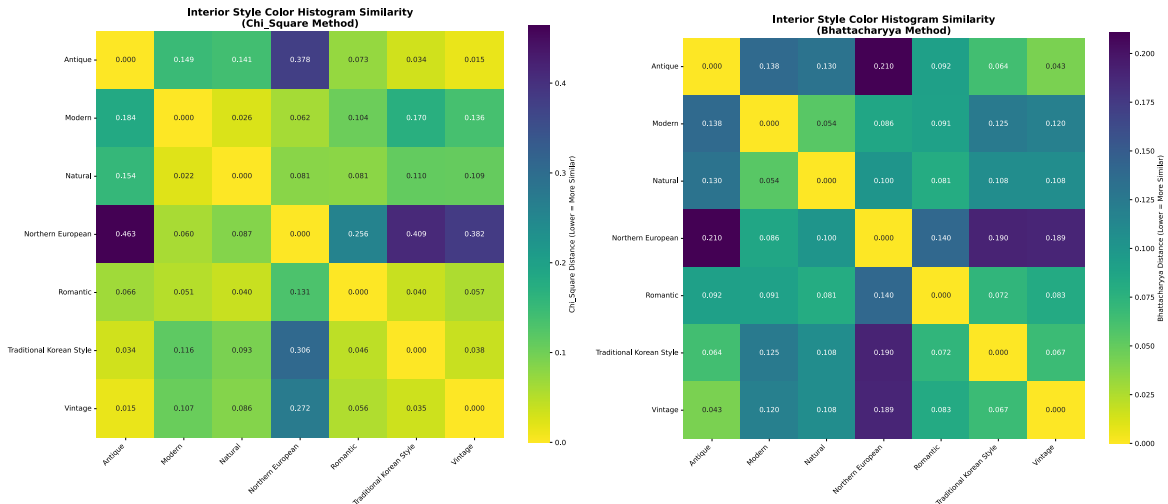


Figure 2: Color Histogram in RGB Space

RGB 색 공간에서의 색 히스토그램을 추출한 결과는 Fig 2와 같다. 이를 조금 더 정량적으로 비교하기 위해서 카이제곱 거리와 바타차리야 거리를 계산하였다. Fig 3.



(a) Chi-Square Distance in RGB Space

(b) Bhattacharyya Distance in RGB Space

Figure 3: Distance Comparison in RGB Space

Antique와 Vintage는 두 거리 계산 방법 모두에서 가장 유사한 것으로 나타났고, Antique와 Northern European은 두 거리 계산 방법 모두에서 가장 유사하지 않은 것으로 나타났다. 전반적으로 Romantic 이 다른 모든 스타일들과 대부분 유사한 것으로 나타났고, Northern European이 다른 스타일들과 가장

유사하지 않은 것으로 나타났다. 이외에도 Traditional Korean Style은 Vintage와 가장 유사한 것으로 나타났고, Modern과 Natural은 서로 가장 유사한 것으로 나타났다. 비대칭 카이제곱 거리를 사용했기 때문에 각 스타일 간의 거리가 대칭적이지 않다는 점에 유의해야 한다.

5.2.2 Color Histogram in HSV Space

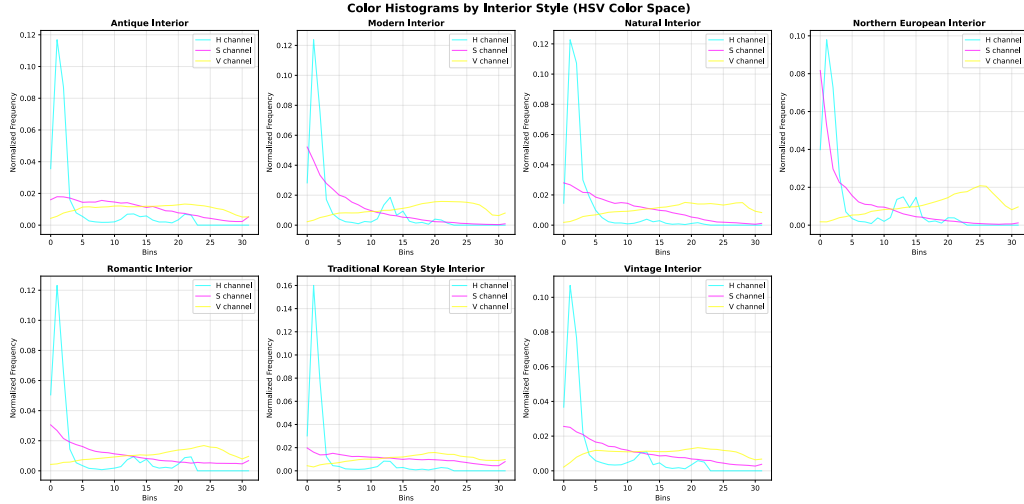
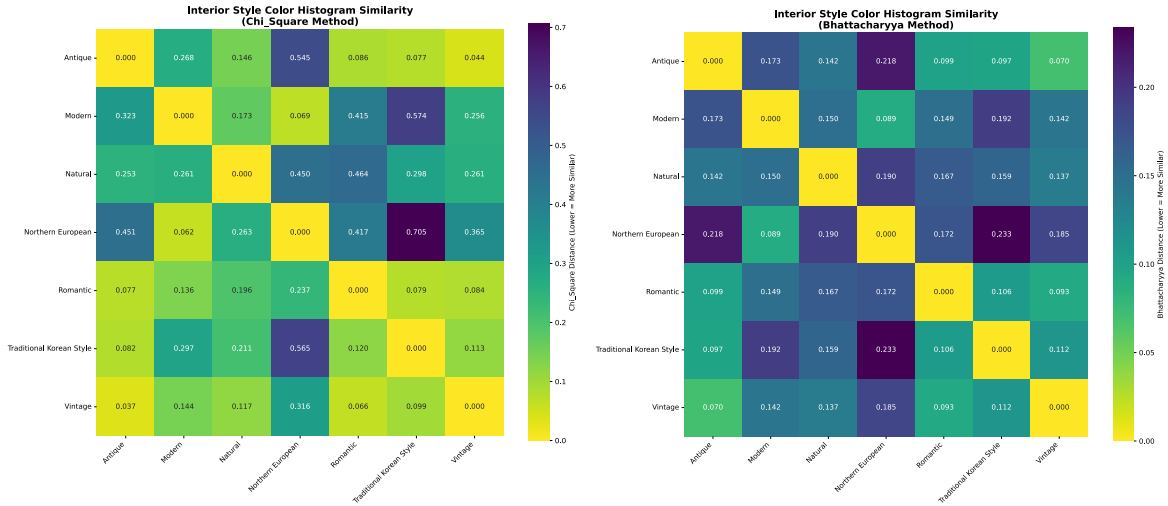


Figure 4: Color Histogram in HSV Space

HSV 색 공간에서의 색 히스토그램을 추출한 결과는 Fig 4와 같다. RGB 색 공간과 동일하게 카이제곱 거리와 바타차리야 거리를 계산하여 비교하였다. Fig 5.



(a) Chi-Square Distance in HSV Space

(b) Bhattacharyya Distance in HSV Space

Figure 5: Distance Comparison in HSV Space

RGB 색 공간과는 다르게 Northern European과 Traditional Korean Style이 가장 유사하지 않은 것으로 나타났고, Antique와 Vintage는 두 거리 계산 방법 모두에서 가장 유사한 것으로 나타났다. 전반적으로 Romantic이 다른 모든 스타일들과 대부분 유사한 것으로 나타났고, RGB 색 공간과는 다르게 Modern, Natural, Northern European이 다른 색들과 거리가 먼 것으로 나타났다.

5.3 Text Finetuning Results

사용자가 입력한 텍스트를 기반으로 가구에 대한 설명을 생성하기 위해 ollama에서 Llama3.2-vision:11b 모델을 사용하였다. 입력된 Query와 모델이 생성한 답변의 예시는 다음과 같다.

Q1. A modern room contains round items

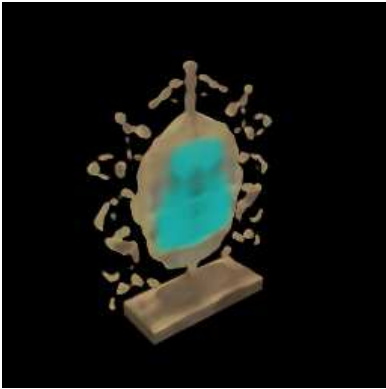
- A1.**
- Create a modern round coffee table.
 - Create a modern round floor lamp.
 - Create a modern round side table.
 - Create a modern round ottoman.
 - Create a modern round rug.
 - Create a modern round wall mirror.
 - Create a modern round pendant light.
 - Create a modern round console table.

Q2. Let me show luxurious room

- A1.**
- Create a plush velvet sofa.
 - Create a marble-topped coffee table.
 - Create a chandelier with crystal drops.
 - Create a floor-to-ceiling velvet curtain.
 - Create a solid wood, ornate four-poster bed.
 - Create a pair of matching, high-backed armchairs.
 - Create a console table with a metallic base.
 - Create a large, stone-encased fireplace.
 - Create a set of ornate, gold-framed mirrors.
 - Create a plush area rug with a subtle, metallic sheen.

Q3. Introduce some representative pieces of traditional Korean furniture

- A1.**
- Create a traditional hanji screen.
 - Create a traditional jeongja sofa.
 - Create a traditional jangdokdae low table.
 - Create a traditional ondol underfloor heating system.
 - Create a traditional hansot bed.
 - Create a bamboo floor.
 - Create traditional Korean lanterns.
 - Create traditional Korean artwork.



(a) prompt: Create an antique bronze clock. dominant colors are black, darkslategray, sienna, rosybrown, lightgray.



(b) prompt: Create a modern glass table. dominant colors are gray, silver, black, gray, silver.



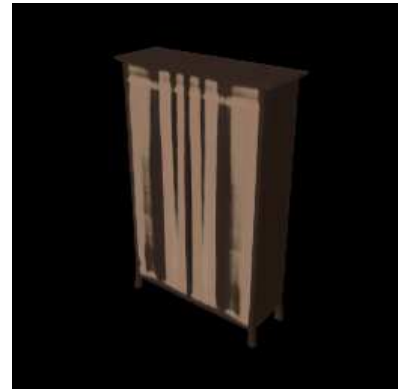
(c) prompt: Create a natural wooden table. dominant colors are darkslategray, darkslategray, tan, rosybrown, gainsboro.



(d) prompt: Create a northern european metallic lamp. dominant colors are whitesmoke, darkgray, lightgray, lavender, dimgray.



(e) prompt: Create a romantic fluffy bed. dominant colors are silver, darkgray, linen, dimgray, darkslategray.



(f) prompt: Create a korean traditional wooden closet. dominant colors are darkgray, black, darkolivegreen, lightgray, gray.



(g) prompt: Create a vintage wooden chair. dominant colors are darkgray, tan, peru, saddlebrown, darkslategray.

Figure 6: Generated 3D objects from prompts and dominant color guidance.

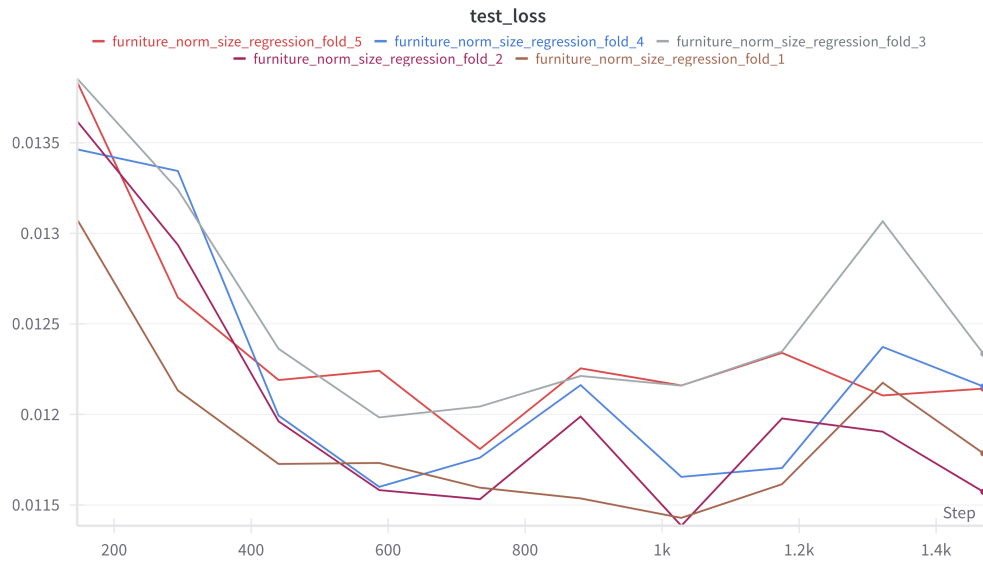


Figure 7: Size Prediction Results

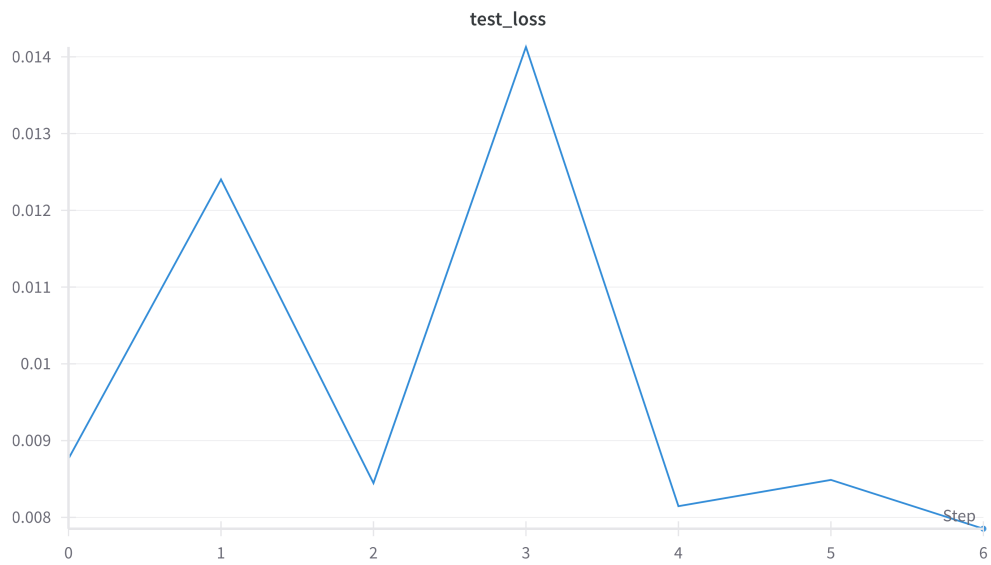


Figure 8: Size Prediction Results from Captioning

5.4 3D Object Generation

사용자가 입력한 텍스트를 기반으로 3D 가구 객체를 제대로 생성하는지 확인하기 위해서 정성적인 확인 방법을 사용하였다. Fig 6를 보면 대체적으로 색상을 잘 유지하면서 각 인테리어 스타일에 맞는 가구 객체를 생성하는 것을 확인할 수 있었다.

5.5 Size Prediction

가구의 크기를 예측하기 위해서 우회적인 방법을 사용했기 때문에 2가지에 대한 실험이 진행되었다. 첫번째는 image로부터 추출된 latent feature와 가구의 종류를 one-hot 인코딩한 벡터를 concatenate하여 가구의 크기를 예측하는 모델의 성능 평가이고, 두번째는 image를 텍스트로 설명하여 (captioning) latent feature를 추출했을 때에도 성능이 유지되는지 확인하는 것이다.

첫번째 실험은 3D-FRONT 데이터셋에서 제공하는 가구의 이미지와 크기 정보를 이용하여 가구의 크기를 예측하는 모델을 학습하였다. 실험 결과는 Fig 7와 같다. 2916개의 가구 사진을 5-fold로 나누어 크기(가로, 세로, 높이)를 예측하는 모델을 학습하였고, 평균 제곱 오차(MSE)를 사용하여 평가하였다. 최종적인 평균 MSE는 0.011998로 나타났다. 이는 가구의 크기를 예측하는데 있어 약 10%의 오차가 발생한다는 것을 의미한다.

두번째 실험으로 BLIP[10] 모델을 활용하여 이미지로부터 텍스트를 생성하고, 해당 텍스트로부터 Shap-E 모델을 통해 latent feature를 추출한 후, 첫번째 실험에서 학습한 모델을 통해 가구의 크기를 예측하는 테스트를 진행하였다. 약 100개의 가구 사진을 BLIP 모델에 입력하여 텍스트를 생성하고, 해당 텍스트로부터 Shap-E 모델을 통해 latent feature를 추출하였다. 실험 결과는 Fig 8와 같다. 각 텍스트로부터 추출한 latent feature를 통해 가구의 크기를 예측한 결과, 평균 MSE는 0.00975로 나타났다. 이는 BLIP 모델을 통해 생성된 텍스트로부터도 가구의 크기를 예측할 수 있음을 보여준다.

6 Conclusion

본 프로젝트에서는 사용자가 입력한 텍스트를 기반으로 3D 가구 객체를 생성하는 방법을 제안하였다. 이를 구현하기 위해 K-Means 클러스터링을 이용하여 인테리어 스타일 별로 주로 사용되는 색상을 분석하고, Shap-E 모델을 사용하여 3D 객체를 생성하였다. 또한, 3D-FRONT 데이터셋을 활용하여 가구의 크기를 예측하는 모델을 학습하였다.

생성된 3D 가구 객체는 3D 그래픽 툴을 통해 사용자가 원하는 공간에 배치될 수 있을뿐만 아니라, 3D 객체 Retrieval을 통해 유사한 가구를 인터넷에서 검색하고 구매를 유도하는 데에도 활용될 수 있을 것이라 기대한다.

References

- [1] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.
- [2] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions, 2023.

- [3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021.
- [4] 오늘의집. 오늘의 집. <https://ohou.se>. Accessed: 2025-06-18.
- [5] Unsplash. Unsplash. <https://unsplash.com>. Accessed: 2025-06-18.
- [6] Pexels. Pexels. <https://www.pexels.com>. Accessed: 2025-06-18.
- [7] Pixabay. Pixabay. <https://pixabay.com>. Accessed: 2025-06-18.
- [8] Huan Fu, Bowen Cai, Lin Gao, Lingxiao Zhang, Cao Li, Qixun Zeng, Chengyue Sun, Yiyun Fei, Yu Zheng, Ying Li, Yi Liu, Peng Liu, Lin Ma, Le Weng, Xiaohang Hu, Xin Ma, Qian Qian, Rongfei Jia, Binqiang Zhao, and Hao Zhang. 3d-front: 3d furnished rooms with layouts and semantics. *arXiv preprint arXiv:2011.09127*, 2020.
- [9] Hugo Touvron, Louis Martin, Kevin El Boukkouri, et al. Llama 3: Open foundation and instruction models. <https://ai.meta.com/llama/>, 2024. Accessed: 2025-06-18.
- [10] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.