Future Designer

Generative AI meets Interior Design

Filip Nowicki, Arkadiusz Charliński

INTRODUCTION

This project applies generative AI to interior design, combining advanced diffusion models with ControlNet and inpainting techniques to create customized room visualizations. Visual Language Models generate structured furniture descriptions in JSON format, enabling both similarity-based furniture search and attribute-based modifications. The main goal is to enable users to take any room photo and rapidly iterate through multiple design variations, streamlining the interior design process.

SYNTHETIC DATASET

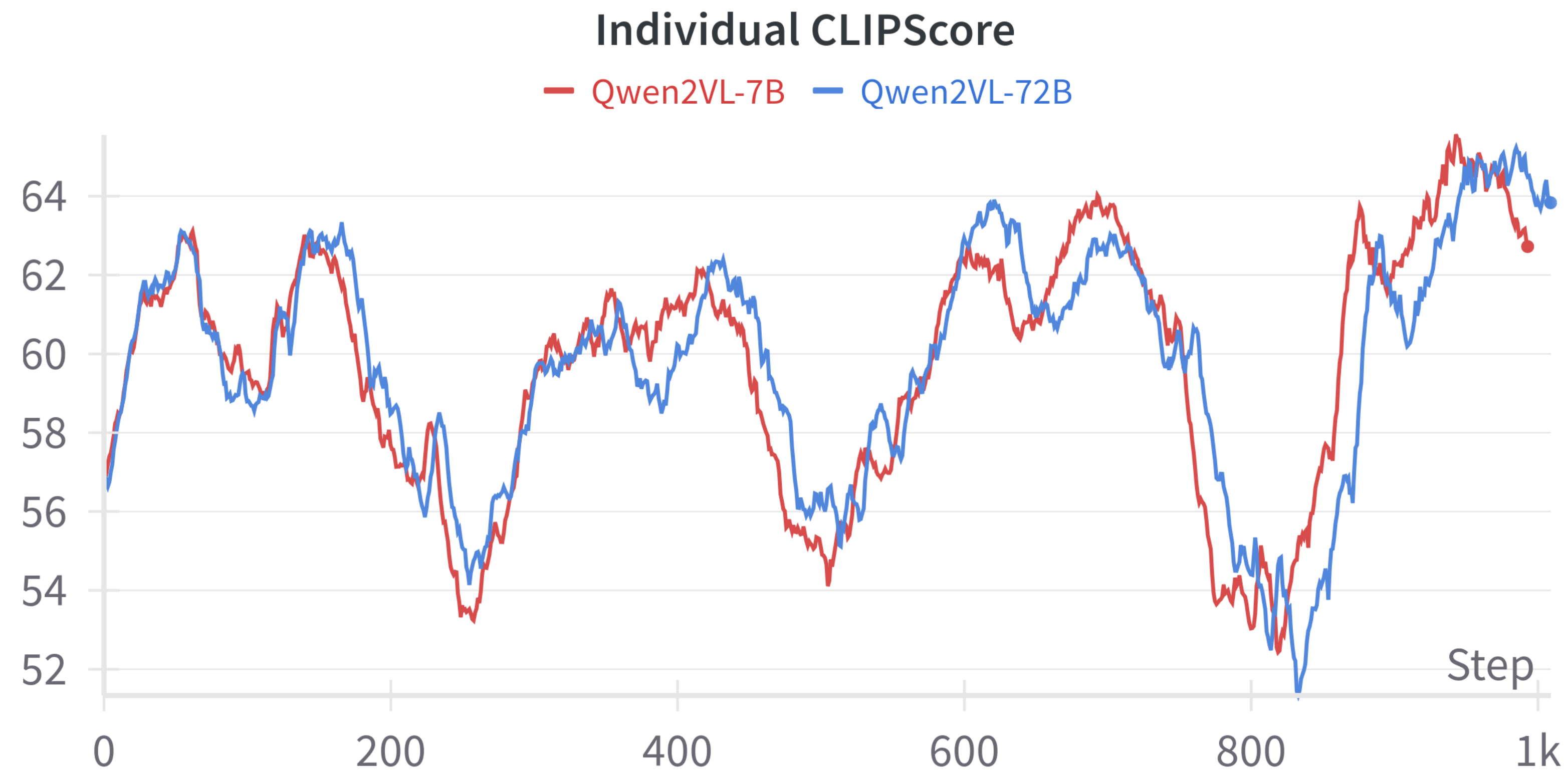
As part of our project, we developed a partially synthetic dataset consisting of 9,000 furniture images generated using Stable Diffusion 3.5, which serves as our training set, complemented by 1,000 real images of furniture. Each image is size 448x448 and annotated with JSON format that contains the following attributes:.

```
{
    “type”: “furniture category”,
    “style”: “design style”,
    “color”: “predominat color”,
    “material”: “main material”,
    “shape”: “physical form characteristics”,
    “details”: “additional decorative feature”,
    “room_type”: “intended room”,
    “price_range”: “estimated price category”
}
```

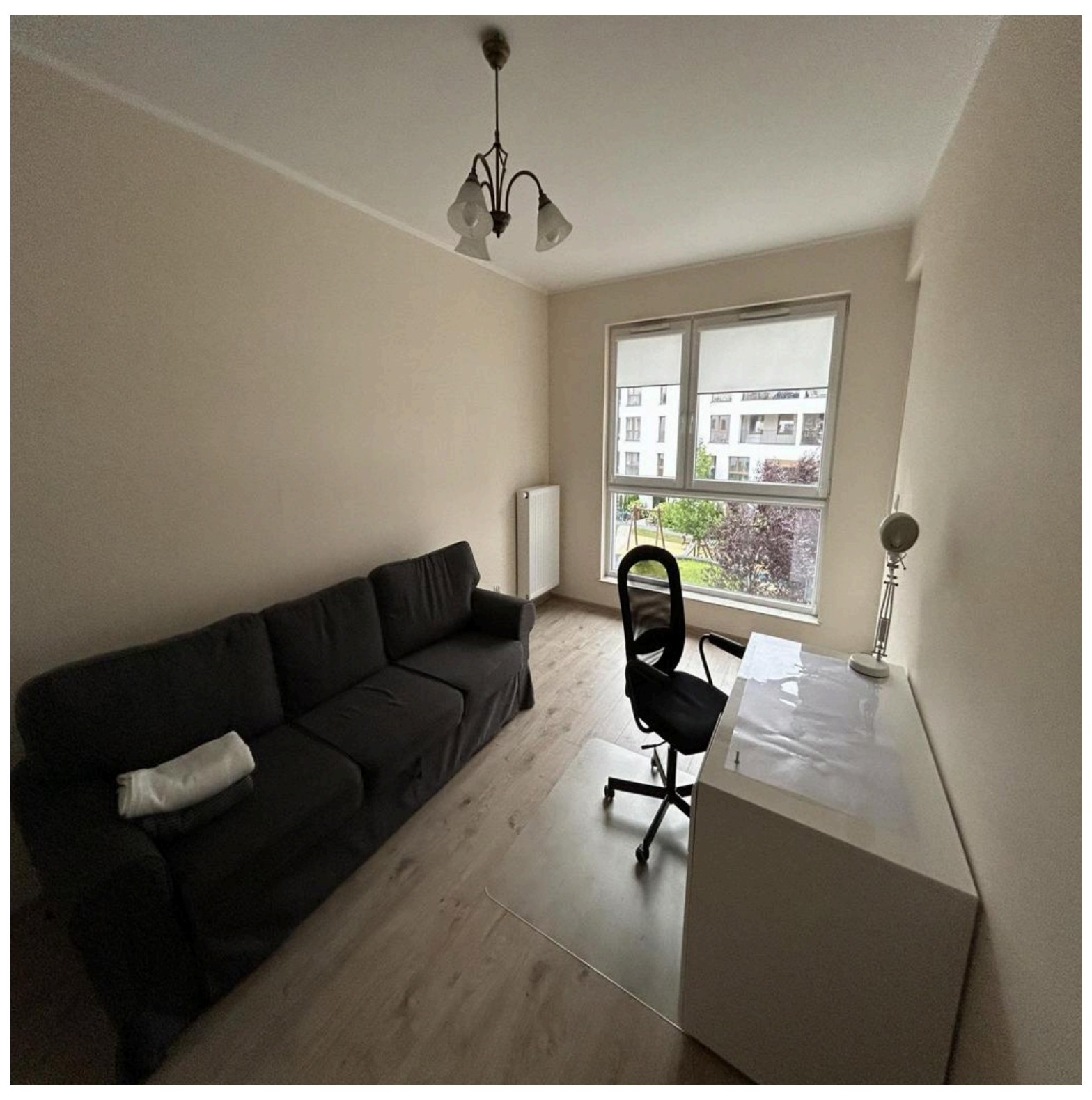
EXPERIMENTS

We conducted experiments with open-source Vision-Language Models (VLMs) for a furniture captioning task, generating test set labels in JSON format. Outputs were evaluated using CLIPScore and analyzed for JSON formatting errors.

While focusing on models under 10B parameters, we also tested larger ones like Qwen2-VL (72B), which showed almost indetical performance to its 7B counterpart despite its size. Smaller models (<3B) typically produced correctly formatted JSON but struggled with prompt content requirements. To improve this, we fine-tuned them using a custom LoRA approach and our generated dataset.



RESULTS

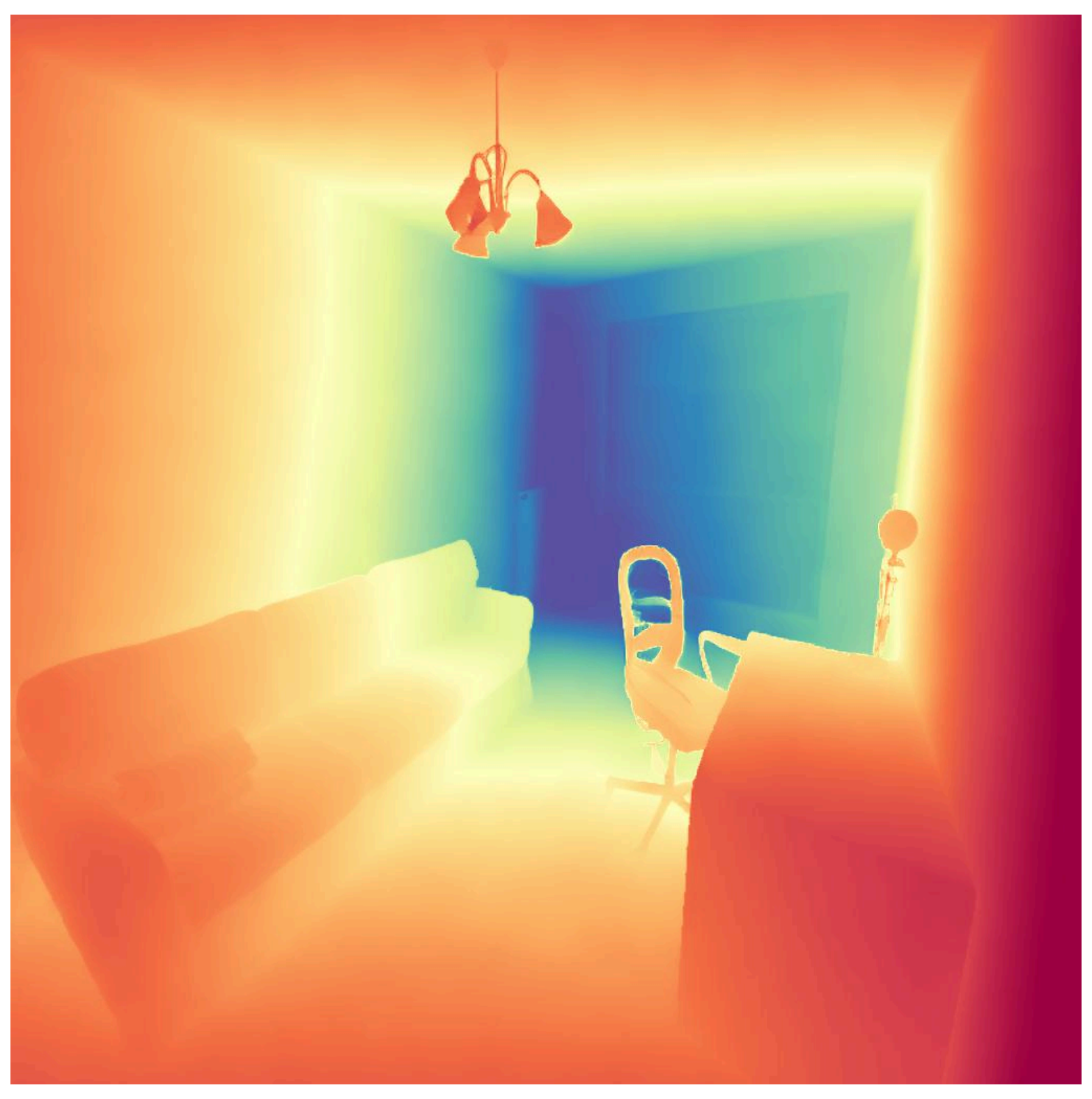


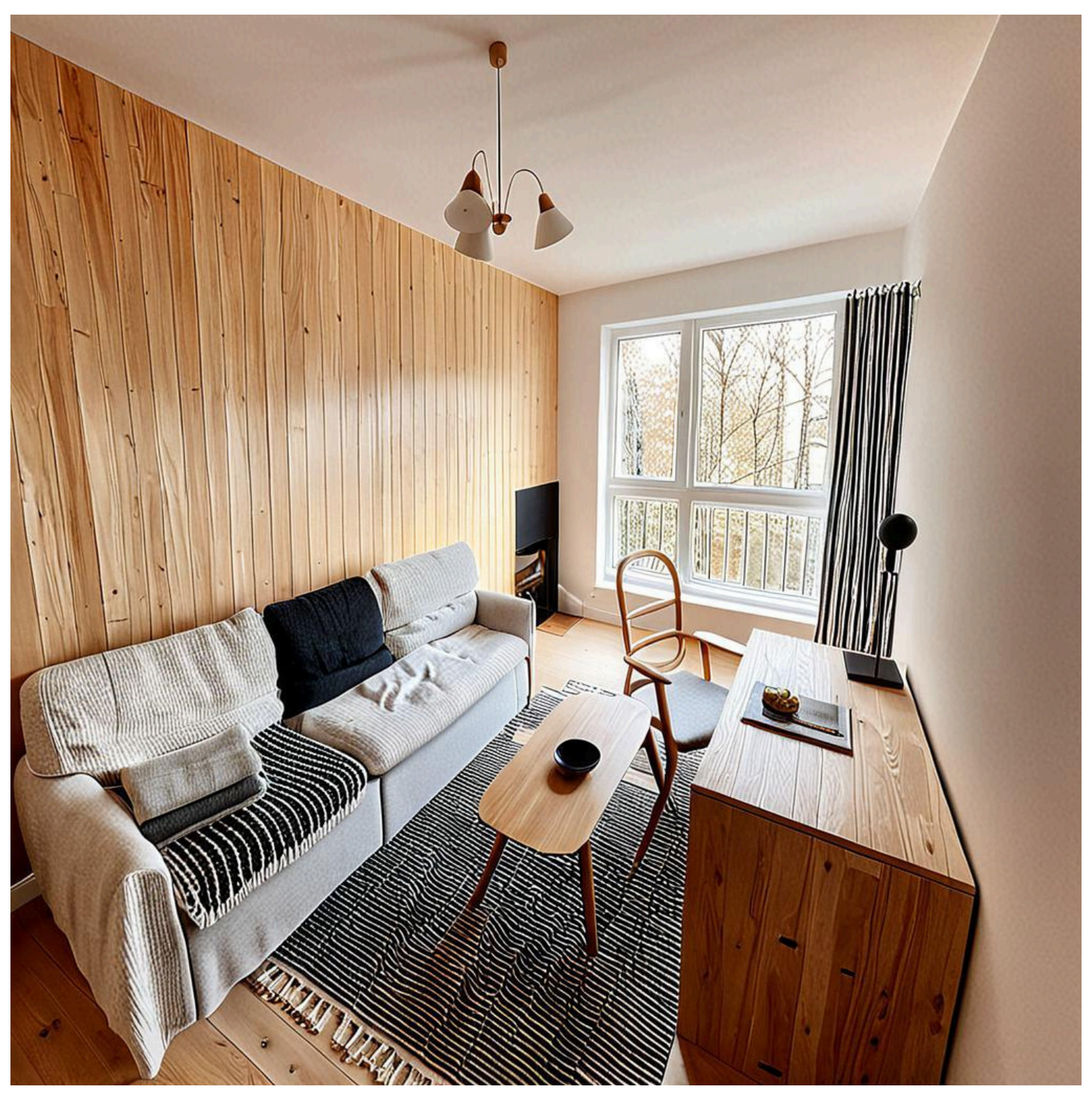
Base image

User uploaded image of the room of any kind

Depth map

Monocular depth estimation using Marigold for ControlNet



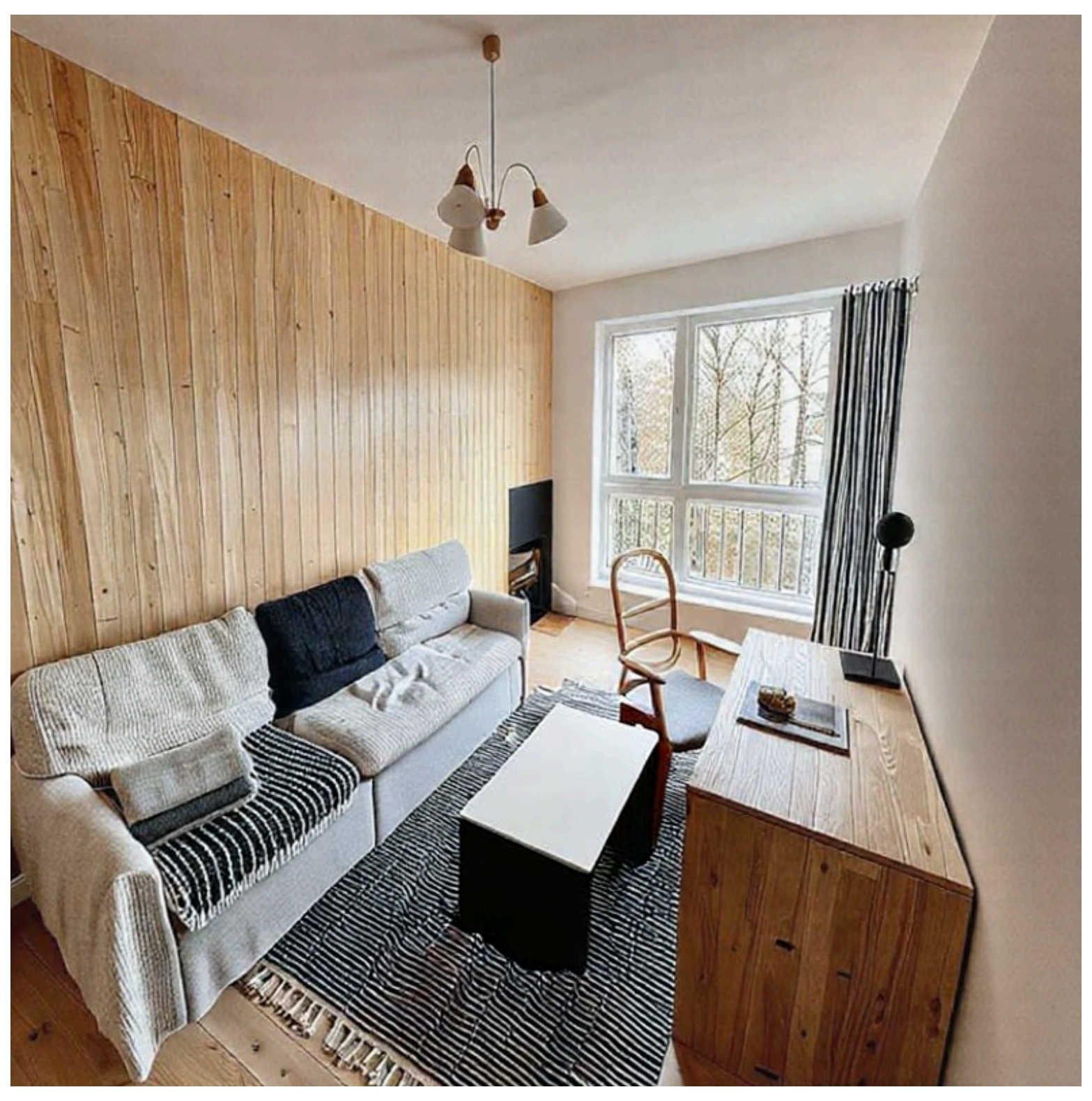


Scandinavian style

Created by using fine tuned Stable Diffusion XL - RealVis-5

Inpainted table

Obtained by modifing JSON attributes generated by Qwen2-VL 2B with LoRA



REFERENCES

[1] Hessel, Jack, et al. "CLIPScore: A Reference-free Evaluation Metric for Image Captioning." Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021.
[2] Hu, Edward J., et al. "LoRA: Low-Rank Adaptation of Large Language Models." International Conference on Learning Representations.
[3] Ke, Bingxin, et al. "Repurposing diffusion-based image generators for monocular depth estimation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.
[4] Wang, Peng, et al. "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution." arXiv preprint arXiv:2409.12191 (2024).