SpaceKraft: A Vision-Language Approach for Automated Interior Design Inpainting

Amitabh Mahapatra* **Sumuk Shashidhar***
University of Illinois, Urbana-Champaign
{amitabh3, sumuks2}@illinois.edu

Abstract

We explore the capability of inpainting technology, leveraging Stable Diffusion models, enhanced by the precision of textual prompts generated via language models fine-tuned with parameter-efficient techniques. The study introduces a method that employs CLIP to provide textual descriptions of images, harnessing these as a basis for generating high-quality and relevant prompts for diffusion-based models without the need for expert user input. Utilizing the advanced scene understanding of GPT-4-V, our system identifies potential object placement regions within an image, guiding the optimal application of inpainted images. To ensure realistic and diverse outputs, we generate multiple image potentials and apply quantitative measures such as Frechet Inception Distance (FID) for evaluating image realism, subsequently filtering to refine the image set. The methodology includes the curation of a Low-Rank Adaptation-driven language model training dataset, training language models to predict human intent in designing prompts, and advanced prompting techniques to overcome refusal rates in vision model interactions. Results demonstrate the efficacy of our system in generating realistic and contextually appropriate images for interior design applications. This research has the potential to significantly advance the automation of interior design and generate new directions for the combination of language and vision models in computer vision tasks.

1. Introduction and Related Works

Interior design, a significant sector fulfilling the human desire for comfortable living spaces, has witnessed various attempts to integrate Computer Vision to streamline its processes [2, 10, 22]. However, the intricacy of indoor environments and object placement has limited the automation of this field, primarily due to the challenges in generating effective masks.

*equal contribution

Recent developments in inpainting, particularly with Stable Diffusion base models [16] and techniques like SmartBrush [20], have demonstrated remarkable capabilities in embedding new images and foreign objects into existing ones. These advancements, albeit reliant on precise textual descriptions and accurate masking, offer promising avenues for interior design applications.

Crucial to the effectiveness of diffusion-based models is the quality of input textual descriptions. Many users, however, lack the expertise to craft effective prompts or articulate their desired elements in an image. To bridge this gap, we utilize CLIP descriptions of images and original intended prompts. Leveraging a wealth of publicly available, high-quality stable diffusion prompts, we fine-tune a language model using parameter-efficient techniques like QLora [6]. Subsequently, we employ batch decoding on a Large Language Model to efficiently generate diverse prompts [14].

GPT-4-V [13], renowned for its scene understanding capabilities, plays a pivotal role in our methodology. When provided with images, it discerns potential locations for object placement, as substantiated by numerous technical reports and preliminary findings (refer to A). However, due to their unreliable efficacy and high rate of refusal, we employ the use of various, specialized prompting techniques, such as EmotionPrompt [11], Chain Of Thought Prompting [19] and Reasoning with Code [3]

These discerned masks, combined with those derived from contour maps, guide the placement of inpainted images.

Utilizing a range of generated prompts and masks, we create a wide array of potential output images through test-time computation. The realism of these images is quantitatively evaluated using metrics like the Frechet Inception Distance (FID) [7]. This approach enables us to filter out less realistic images, thereby refining our image set. Consequently, our method presents an intuitive user experience, negating the need for in-depth understanding of masking, denoising, inpainting, and sampling processes in the realm of interior design, allowing us to effectively design interior

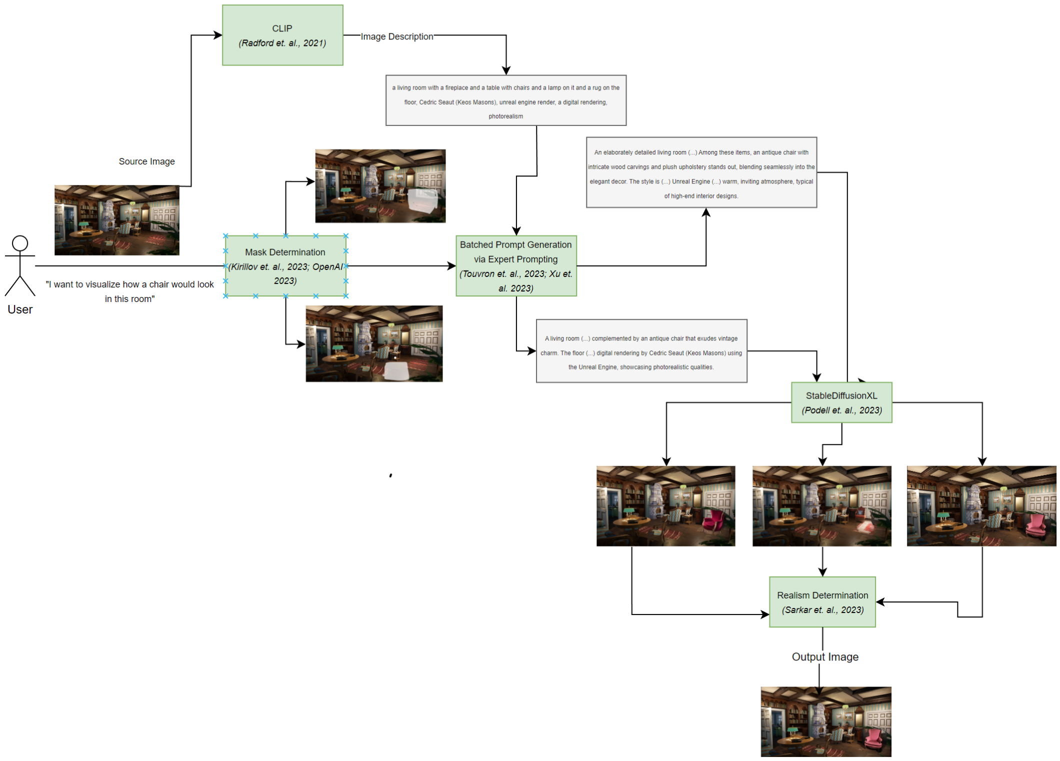

Figure 1. Caption

spaces.

2. Methods

2.1. Base Prompt Generation

Our approach develops multi modal pipeline to effectively insert objects and images into scenes. We first utilize an open source implementation of OpenAI's CLIP [15], and Salesforce BLIP [12], to provide a descriptive, textual representation of the base image. Pillow [4] is used to load and pre-process the image. We use the OpenAI VIT-LARGE-PATCH-14 model as our primary CLIP model.

2.2. Diffusion Model Input Prompt Generation

2.2.1 LoRA Dataset Curation

To facilitate the training of a low-rank adaptation [8] for language models, our methodology emphasizes the creation of a high-quality dataset through the generation of synthetic data. This approach, grounded in successful practices for training various large language models [18, 21], offers a cost-effective and less labor-intensive alternative to traditional data sourcing and validation methods. While acknowledging the potential compromise in data variety, this trade-off is deemed acceptable for our specific use cases.

Our process begins with the meticulous selection of reference images $\mathcal{I}$ and their associated generation parameters and prompts $\mathcal{P}$, sourced from CivitAI [1]. This selection forms the cornerstone of our two-pronged approach. The initial phase involves conducting a human preference analysis, aimed at deducing potential human intentions x from the given diffusion model input prompts $p \in P$. This deduction, represented as a function $g : P \to X$, is pivotal for establishing a baseline that language models can align with, thereby shaping our initial dataset $\mathcal{D} = (x_i, p_i)|x_i \in X, y_i \in P$.

In the subsequent phase, the MISTRAL-7B-INSTRUCT-V0.1 model [9] is utilized to extrapolate these human-derived hypotheses into specific intention-input pairs x, p, expanding the breadth of our dataset. Once a substantial corpus of such pairs is compiled, we leverage the instruct-finetuned `Capybara-34B` model [5] for synthetic augmentation of our dataset. This augmentation is instrumental in enriching the training material, thus enabling our prompt generator to effectively learn and generalize from a diverse array of proven output prompts. Consequently, this strategy equips our model to facilitate high-quality Low Rank Adaptation training, optimizing its performance in generating relevant and varied prompts.

2.2.2 Lora Training

The cornerstone of our approach in enhancing the capabilities of language models for prompt generation lies in the implementation of Low-Rank Adaptation (LoRA) training. LoRA, a parameter-efficient training methodology, allows us to fine-tune pre-trained language models without the need to overhaul the entire model structure [8]. This approach is especially beneficial in the context of generating precise and contextually relevant prompts for Stable Diffusion models, as it ensures a high degree of adaptability while maintaining computational efficiency.

Our LoRA training process begins with the selection of a suitable pre-trained language model. For our purposes, we utilize the `Capybara-34B` model, a variant known for its robustness and versatility in handling diverse language tasks [5]. The model is then subjected to a fine-tuning regimen using the curated dataset $\mathcal{D}$, which comprises a rich assortment of human-generated prompts and their corresponding intentions. This dataset acts as the foundation for training the model to discern and replicate human-like prompt generation.

In the LoRA training phase, we introduce low-rank matrices to the transformer layers of the `Capybara-34B` model. These matrices act as trainable parameters, allowing us to modify the model's attention and feed-forward networks subtly. This modification aims to imbue the model with enhanced capability to interpret and generate prompts that are aligned with human intent and stylistic preferences, a key requirement for effective inpainting in interior design contexts.

A crucial aspect of our LoRA training involves balancing the retention of the original model's knowledge with the incorporation of new, task-specific insights. To achieve this, we employ a conservative learning rate and a focused training duration, ensuring that the model's fundamental language understanding capabilities remain intact while it acquires specialized skills in prompt generation.

Upon completion of the training, the enhanced language model demonstrates a marked improvement in generating prompts that are not only syntactically and semantically coherent but also aligned with the specific requirements of interior design inpainting tasks. These prompts, when fed into Stable Diffusion models, result in high-quality, contextually appropriate image outputs that significantly enhance the automation process in interior design applications.

The success of our LoRA training approach is evidenced by the reduced refusal rates and increased efficacy of generated prompts, as evaluated in subsequent stages of our methodology. This advancement establishes a new benchmark in the integration of language and vision models, paving the way for more sophisticated and user-friendly applications in the realm of automated interior design.

2.3. Mask Determination

In this study, we employ the GPT-4-V API [13], recognized as the state-of-the-art (SoTA) vision model, for the critical task of mask determination in image processing. This choice is underpinned by the model's advanced capabilities in accurately identifying optimal insertion points within an image, based on empirical preliminary tests, and strong reasoning capabilities. To align our methodology with the constraints of computational resources and budgetary limitations, we introduce a dynamic scaling factor, denoted as λ. This factor is instrumental in resizing any given image, $\mathcal{I}$, to a fixed dimension of 512 x 512 pixels, ensuring uniformity and computational efficiency. The scaling factor λ is defined mathematically as:

$$\lambda = \frac{512}{\max(w, h)} \tag{1}$$

Here, w and h represent the width and height of the image, respectively. The scaled image, $\lambda\mathcal{I}$, is subsequently processed through a mask determination system, which operates based on a specifically crafted prompt $\mathcal{P}$. This system outputs a set of coordinates that define a square region within the image. The output is in the form of two diagonal corner points, denoted as $[(x_0, y_0), (x_1, y_1)]$. These points are then extrapolated to ascertain their validity and relevance to the image context. Following this, a reverse scaling is applied to each coordinate using the inverse of λ, thereby mapping them accurately onto the original dimensions of the image. This methodology ensures precision in mask determination while adhering to our computational and budgetary constraints.

We provide the system prompt, utilized to receive coordinates from our vision model in Appendix C. Due to safety limitations by OpenAI, this turns out to be a challenging task with a high request refusal rate (RRR), and therefore, we exploit various methodologies discovered in the field of Natural Language Processing to generate reliable and consistent results. We utilize HTML tags, in order to harness the reasoning with code capabilities of language models [3]. We also use EmotionPrompt [11] as well as Reasoning based Chain-of-thought prompting introduced in [19] to increase our rates of achieving factual, interpretable and reproducible results.

2.4. Realism Determination

In the domain of computer vision, particularly in the realm of image generation, the Frechet Inception Distance (FID) [7] score stands as a benchmark for assessing the quality and realism of generated images. This metric, derived from comparing the distribution of generated images against a corpus of real images, is instrumental in gauging the perceptual similarity between these two sets.

The FID score's significance lies in its ability to capture the nuanced differences in image quality that are often imperceptible to the human eye. This is particularly critical in our study, where the goal is to seamlessly integrate inpainted objects into existing indoor scenes, maintaining a high degree of realism. A lower FID score indicates a closer resemblance to real-world images, suggesting that the synthetic images are nearly indistinguishable from authentic photographs. This level of realism is essential in interior design applications, where the visual appeal and authenticity of the design directly influence the user's experience and satisfaction.

The Fréchet Inception Distance (FID) score is calculated as follows:

$$\text{FID} = \|\mu_r - \mu_g\|^2 + \text{Tr}\left(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2}\right) \quad (2)$$

where:

- μ_r and μ_g are the mean feature vectors of the real and generated images, respectively,
- Σ_r and Σ_g are the covariance matrices of the real and generated images, respectively,
- Tr denotes the trace of a matrix,
- $\| \cdot \|$ denotes the Euclidean norm.

3. Discussion

In the context of our study, which harnesses Stable Diffusion XL for advanced image inpainting in interior design, we have achieved a FID score of 3.062 on a dataset of 2000 bedroom scenes. The score is a pivotal metric, underscoring the success of our approach in producing high-quality, realistic images for interior design. This achievement reflects the potential of our methodology in advancing the automation of interior design, offering promising avenues for future exploration and application in this field.

In light of our encouraging outcomes achieved through the SpaceKraft system, future research will focus on the refinement of model interactions, aiming to minimize refusal rates further and enhance the model's ability to understand and execute complex instructions with greater precision. Additionally, we plan to explore the promising avenue in the personalization of design outputs by incorporating user behavior analysis, which could lead to more intuitive and user-centric interior design solutions. By continuing to leverage the rapid advancements, we aim to not only streamline the creative process but also democratize interior design, making sophisticated and customized design accessible to a broader audience.

4. Group Contributions

All members of the group contributed equally.

- **Amitabh Mahapatra**: Realism Determination, Mask Determination
- **Sumuk Shashidhar**: Paper Writing, Low Rank Adaptation

Challenges: A CVPR project [17] had a lot of last minute changes, till the supplementary deadline (27 November). Therefore, couldn't achieve all of the points listed in the proposal.

References

[1] Civitai, 2023. 2

[2] Nasir Ahmad, Saddam Hussain, Kashif Ahmad, and Nicola Conci. Computer vision based room interior design. In Antanas Verikas, Petia Radeva, and Dmitry Nikolaev, editors, *Eighth International Conference on Machine Vision (ICMV 2015)*, volume 9875 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, page 98751G, Dec. 2015. 1

[3] Lang Cao. Enhancing reasoning capabilities of large language models: A graph-based verification approach, 2023. 1, 3

[4] Alex Clark. Pillow (pil fork) documentation, 2015. 2

[5] Luigi Daniele and Suphavadeeprasit. Amplify-instruct: Synthetically generated diverse multi-turn conversations for effecient llm training. *arXiv preprint arXiv:(comming soon)*, 2023. 2, 3

[6] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*, 2023. 1

[7] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2018. 1, 3

[8] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. 2, 3

[9] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. 2

[10] Jinsung Kim and Jin-Kook Lee. Stochastic detection of interior design styles using a deep-learning model for reference images. *Applied Sciences*, 10(20), 2020. 1

[11] Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. Large language models understand and can be enhanced by emotional stimuli, 2023. 1, 3

Figure 2. Three examples of the pipeline in action. Left column denotes original images and the right column denotes the generated images.

Figure 3. Three examples of the pipeline in action. Left column denotes original images and the right column denotes the generated images.

Figure 4. Sample Image

Figure 5. Indoor Bedroom Picture

[12] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation, 2021. 2

[13] OpenAI. Gpt-4 technical report, 2023. 1, 3

[14] Reiner Pope, Sholto Douglas, Aakanksha Chowdhery, Jacob Devlin, James Bradbury, Anselm Levskaya, Jonathan Heek, Kefan Xiao, Shivani Agrawal, and Jeff Dean. Efficiently scaling transformer inference, 2022. 1

[15] 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, 2021. 2

[16] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022. 1

[17] Ayush Sarkar, Hanlin Mai, Amitabh Mahapatra, Svetlana Lazebnik, DA Forsyth, and Anand Bhattad. Shadows don't lie and lines can't bend! generative models don't know projective geometry... for now. *arXiv preprint arXiv:2311.17138*, 2023. 4

[18] Veniamin Veselovsky, Manoel Horta Ribeiro, Akhil Arora, Martin Josifoski, Ashton Anderson, and Robert West. Generating faithful synthetic data with large language models: A case study in computational social science, 2023. 2

[19] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. 1, 3

[20] Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model, 2022. 1

[21] Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. Large language model as attributed training data generator: A tale of diversity and bias, 2023. 2

[22] Zhen Zhang and Jianwei Ban. Aesthetic evaluation of interior design based on visual features. *Int. J. Mob. Comput. Multimed. Commun.*, 13(2):1–12, apr 2022. 1

A. Example GPT-4-V Output

```
Reasoning: The image shows a bedroom
with a bed in the center, a nightstand
with a lamp to the left of the bed, a
chair to the right, and various decorative
items on the wall above and to the left of
the bed. There is a potted plant on the
right side of the bed that is sitting
on a small stool or table. The floor
space in front of the bed and the
area on the right side near the chair
appear to be open. Additionally,
the top right corner of the image
near the window could also
accommodate a lamp.

[(300, 650), (850, 700), (1200, 300)]
# the center points of potential
# squares of empty area
```

B. Example Dataset

```
{
    "input" : "I would like to see a
    photograph of a designer bedroom
    with a cinematic look",

    "output" : "room, designer interior
    for a bedroom, ((((cinematic look)))),
    soothing tones, insane details,
    intricate details, hyperdetailed,
    low contrast,
    soft cinematic light, dim colors,
    exposure blend, hdr, faded, slate
    gray atmosphere",
}
```

C. GPT-4-V Sytem Prompt

```
<task>

You are Coordinate GPT.
Given an image of a bedroom, your job
is to determine where a lamp could
theoretically be placed.
In order to facilitate this, you must
output an array of co-ordinates. To
```

assist you, you are given the scale
of the image on the X and Y axis.

</task>

Reasoning: (talk about what is in the image,
and where it is empty)

[(x0, y0), (x1, y1), (x2, y2)] # the center
points of a potential squares of empty area

Adhere strictly to the output
format that I have given you.

<final_tips>

Stay focused and dedicated to your goals.
Your consistent efforts will lead to
outstanding achievements. This is very
important to my career, and I would be
super happy, and super satisfied with you,
if you could help me. Be confident,
while helping, and be extremely through,
and detailed.

</final_tips>

DO NOT WRITE PYTHON CODE TO SOLVE THIS PROBLEM.
SOLVE IT JUST WITH TEXT AND PLAINGPT