PRESERVING PRODUCT FIDELITY IN LARGE SCALE IMAGE RECONTEXTUALIZATION WITH DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

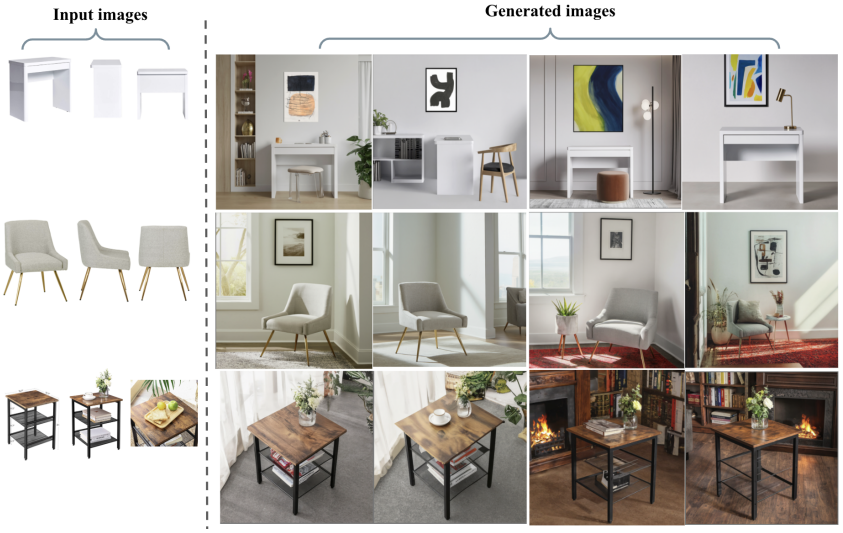


Figure 1: Given a few input images of a real world product, our system can generate images that not only maintain high fidelity to the original product, but also recontextualize it in novel settings beyond background changes: from showcasing it in a new perspective, adding object occlusions, to creating different and realistic lighting conditions.

ABSTRACT

We present a framework for high-fidelity product image recontextualization using text-to-image diffusion models and a novel data augmentation pipeline. This pipeline leverages image-to-video diffusion, in/outpainting negatives to create synthetic training data, addressing limitations of real-world data collection for this task. Our method improves the quality and diversity of generated images by disentangling product representations and enhancing the model's understanding of product characteristics. Evaluation on the ABO dataset and a private product dataset, using automated metrics and human assessment, demonstrates the effectiveness of our framework in generating realistic and compelling product visualizations, with implications for applications such as e-commerce and virtual product showcasing.

1 INTRODUCTION

Placing products seamlessly into new environments in images holds immense value for e-commerce, VR/AR, content creation among others. Photorealistic recontextualization demands both visual appeal and, crucially, accurate product detail preservation. The most common method that preserves product details, namely background swapping, does not provide realistic product relighting, novel perspectives, or natural occlusions. These shortcomings heavily limit how natural-looking and

creative the resulting images can be. While recent advancements in text-to-image diffusion models such as DreamBooth (1) and InstructPix2Pix (2) offer promise, directly applying them to diverse product images often compromises fidelity. We observe that these existing methods struggle with intricate details, consistent appearance, and product background disentanglement, especially when dealing with complex textures, reflections, and occlusions in the product images. Scaling to a wide variety of products amplifies these failure modes, motivating a new recontextualization approach.

We present a novel framework specifically designed to address these limitations and achieve high-fidelity product recontextualization across a wide variety of products. Compared with background replacement and other methods, our system not only preserves product details at scale, but also enables the three capabilities key to blending products seamlessly into new settings: 1) realistic relighting; 2) object occlusions; 3) novel product viewpoints.

2 METHOD

Existing methods (1) (3) that perform well on subject driven personalization commonly use few shot samples of the product. While simply collecting more samples would improve product fidelity as shown in (1), it's not feasible to do so at scale. Instead, we propose augmenting the existing few shot samples through synthetic augmentation to address common failure modes, and finetune the model using LoRA (4), in a similar setup to DreamBooth (1). This method provides simplicity and scalability for a large variety of products.

2.1 SYNTHETIC DATA AUGMENTATION

Our method addresses the limitations of existing methods by augmenting the training data with synthetically generated examples. We employ a data pipeline consisting of novel view generation, background/object disentanglement, and the inclusion of negative examples (Visualization of the pipeline in Appendix A.2.1). Each stage targets specific failure modes observed in existing methods.

2.1.1 NOVEL VIEW GENERATION

Increasing the number of input images improves fidelity (1), but simple repetition leads to overfitting. We generate novel viewpoints using image-to-video diffusion models (5), augmenting the training data with diverse product perspectives (Appendix A.1 Table 2). While 3D reconstruction models (6; 7) offer an alternative, we prioritize video frame interpolation for better prompt adherence and product fidelity. Figure 2 in Appendix A.1 demonstrates an example.

2.1.2 NEW CONTEXT IMAGES

To reduce overfitting to the background elements in input images, we used masked outpainting to change the backgrounds of the base, and novel view images. These images serve as positives while helping disentangle the product from the backgrounds during finetuning. We used an LLM to generate new prompts for each image, then created a segmentation mask using (8). The mask and new context prompt was used to outpaint the image to change the background. To reduce hallucinations we curate and cache the context prompts. The caching strategy is described in Appendix subsection A.2.2.

2.1.3 AUGMENT WITH NEGATIVES

While (9) uses a class prior preservation loss, we observed that adding negative images has a similar effect. We additionally added counterfactuals - images with the same background as the ground truth images, but with different objects inpainted into the product's masked region. We observed this gave the following benefits: (1) preserve class priors in the base generation model, (2) accurately render non-product objects in the background/foreground (3) render products with reduced artifacts in the image. To generate a negative image we took the caption of a fine tuning image, and generated a new image using the base diffusion model. We observe that a 2:1 positive:negative image ratio produced higher image quality and fidelity at the cost of diversity.

2.2 CAPTIONING

We used an internally trained captioning model based on a Vision Language Model (VLM) (10) that is provided the base set of images with an instruction asking for fine grained details of the image, including various attributes like color, position, lighting, and other fine grained product details. We found that fine grained captioning greatly improves subject recontextualization quality.

2.3 TRAINING DATA FILTERING

While generating data using the approaches mentioned above greatly increases the amount of samples, some images may be higher quality than others. For example, it is possible for outpainting to produce hallucinations that extend the products or add new attributes. We developed techniques to reduce errors from outpainting in two ways to improve the quality of our finetuning training set. Particularly, 1) Use automated metrics to evaluate each additional outpainted image and sample the highest rated images. Specifically, we use CLIP (11) (both CLIP-I and CLIP-T) as well as segmented CLIP Embedding for metrics; and 2) Compare the segmentation mask of the outpainted image and the reference image, and filter out outpainted images whose IoUs (Intersection over Unions) for the two masks fall below an experimentally determined threshold.

2.4 MODEL FINETUNING

We chose to adopt the approach described in (1), with three adjustments: Class prior preservation loss is omitted in favor of more negatives during finetuning, we trained on a higher learning rate and for longer and we swept over possible synthetic tokens to be used for each product. We observed that some tokens perform significantly better than the others. Tokens that can be confused as product names are avoided and we chose to sample from an existing list of rare tokens that have plausible association with the product being finetuned. We find that using LoRA finetuning produced better results compared to finetuning the whole model.

2.5 POST FINETUNING RANKING

We re-use the evaluation metrics for images to rank and pick the top N images above an experimentally determined threshold. We use CLIP-T, CLIP-I, DINO (12) embeddings to rank images on product fidelity. We pick the top N images per product after ranking them by the aggregate sum of these metrics. This leads to a relative 30.9% increase in image pass rate. Our analysis also showed a 0.4 Pearson correlation between these automatic metrics and human ratings. As a future direction, a ranking/classification model can be learned to maximize correlation with human ratings.

3 RESULTS

We evaluated our proposed framework on the Amazon Berkeley Objects (ABO) (13) dataset, creating a challenging dataset due to its diverse range of product categories and complex background scenes. We also used a private set of consumer products, similar to the ABO dataset in it's difficulty and range of product types. We quantitatively assess the fidelity of our generated images using established metrics, including CLIP and DINO based image similarity scores. We also conduct a qualitative analysis through human evaluation, comparing our approach to an existing baseline of DreamBooth + LoRA.

3.1 QUALITATIVE RESULTS

We showcase results on products in the ABO dataset in Figure 6. Our method preserves the product details with high fidelity while seamlessly integrating the product into a realistic lifestyle scene. We observe that the generated images exhibit diverse and plausible arrangements of furniture and decor, consistent with the provided prompt. Our method is able to achieve the following (1) object occlusions, (2) novel view generation and (3) varied realistic lighting conditions, particularly foreground relighting. In Figure 6 we demonstrate the high quality scene composition of the method, and exemplify its advantage over background replacement systems, which do not exhibit these features.

We also showcase additional qualitative results from ablation studies focusing on LoRA rank and training steps in section A.5

3.2 Quantitative Results

We provide human evaluation results as well as automated metrics on our approach. For evaluation, we selected a closed set of real world objects distinct from the ABO dataset. A small closed set of 100 products were used to account for the cost of human evaluation, reduce any effects of data contamination and demonstrate the generality of our approach to other object datasets. We compare our method to a baseline of using just the method from (1) and LoRA.

3.2.1 Human Evaluation Results

For human evaluation, we showed raters the source and generated images and asked them to accept or reject the generated images. A series of 8 questions all needed to pass and we took a majority vote over 3 raters (Additional details in Appendix A.3). We report the pass rate per image, and per product (defined as the % of products with at least one passing image). Human evaluations show that our method performs significantly better than a baseline method of DreamBooth + LoRA (Rank=64) on a closed evaluation set of various furniture products. Results are reported in Table 1.

Dataset	Overall Per Image pass rate ↑	Per Product pass rate ↑
Furniture (Ours)	**17.40%**	**45.5%**
Furniture (DreamBooth)	10.00%	24.00%

Table 1: Human Evaluation results on multiple datasets created by mixing several product types in a closed evaluation set. Due to variations in generation, we see that per product pass rate is often much higher than the per image pass rate, demonstrating the difficulty of the task.

3.2.2 Automatic Evaluation Metrics

Due to the cost of human evaluation, we also evaluated our method using automatically computed metrics (Table 2). Following (1) we use the CLIP-I and DINO metrics to compare generated images against the reference images to evaluate product fidelity. We also used CLIP-T to evaluate text alignment as from (1). We also used SegCLIP-I (and Segmented DINO) for more localized cosine similarity as suggested in (14). We achieved consistent performance on metrics on the ABO dataset and our private evaluation set, demonstrating the ability to scale to a wide variety of products.

	ABO	Private Set (Ours)	Private Set (DreamBooth + LoRA Baseline)
CLIP - I ↑	0.89	0.80	**0.81**
CLIP - T ↑	0.24	0.20	**0.25**
DINO - I ↑	0.98	**0.94**	0.92
Seg CLIP - I ↑	0.94	**0.90**	0.88
Seg CLIP - T ↑	0.21	**0.24**	0.22
Seg DINO - I ↑	0.99	**0.99**	0.98

Table 2: Image and Text alignment metrics for our process on ABO dataset and a private evaluation set. The segmented metrics are higher than the non-segmented versions, indicating that the products in reference and generated images are similar, while the background of the images varies.

4 Conclusion

We demonstrate an improved data pipeline and an alternate strategy to subject driven generation, particularly for product recontextualization, to millions of products at scale. Experimental observations and human evaluation show that our method performs significantly better than existing approaches while requiring no model surgery and expensive finetuning.

REFERENCES

[1] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023.

[2] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2023.

[3] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022.

[4] 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.

[5] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Yuanzhen Li, Tomer Michaeli, et al. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*, 2024.

[6] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19697–19705, 2023.

[7] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023.

[8] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[9] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation, 2023.

[10] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Rames, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and et al. Learning transferable visual models from natural language supervision, 2021.

[12] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021.

[13] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, Matthieu Guillaumin, and Jitendra Malik. Abo: Dataset and benchmarks for real-world 3d object understanding. *CVPR*, 2022.

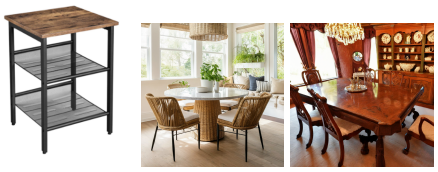
[14] Chenyang Zhu, Kai Li, Yue Ma, Chunming He, and Li Xiu. Multibooth: Towards generating all your concepts in an image from text, 2024.

A Appendix / supplemental material

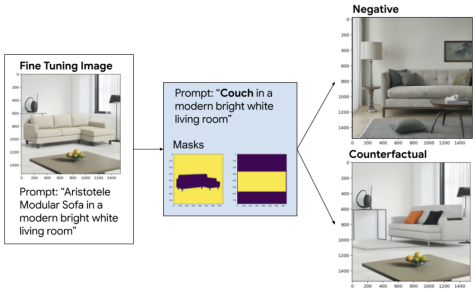
A.1 Synthetic Augmentation Techniques

Figure 2: Novel view generated using image-to-video diffusion. Left: Input image. Right: Generated view.

Figure 3: Samples of image positives for a given table, along with it's object mask used for background replacement/outpainting.



(a) Samples of image negatives for a dining table. Left most is the original unmodified image, while the rest are in-class negatives. The prompts used are described in Appendix A.4



(b) The process for creating counterfactuals and negatives from a starting sample.

Figure 4: Samples of counterfactuals and negatives used to augment product images.

A.2 System Design Choices

A.2.1 Finetuning Data Pipeline

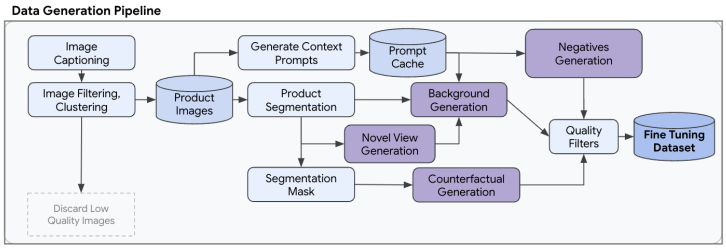


Figure 5: Figure illustrating our scalable finetuning data generation approach.

A.2.2 Caching Context Prompts

Creating context prompts to generate images on the fly, and at scale can be prohibitively expensive and difficult to control for quality. Moreover, while outpainting resulted in novel background images for the same product, generating new prompts for each image was prone to hallucinations. To address this issue, we cached context prompts before running the data pipeline and curated the prompts to remove unrealistic prompts.

To solve these issues, we instead do the following: for each type of product (chairs, phone cases, etc), we generate many possible context prompts using an LLM, and save these prompts in an offline file. We refer to this as a cached "prompt bank". For each individual product, we classify the type of product (either as part of product details available in the dataset metadata or using a VLM), and retrieve prompts from the prompt bank. We then use these prompts for synthetic data augmentations (e.g outpainting) and generating target generation prompts. Using the prompt bank allows us to save compute and improve our prompts over time, while still producing novel contexts for the product.

A.3 HUMAN EVALUATION GUIDELINES

Human evaluation guidelines focused on a fine grained evaluation across multiple categories and failure modes. Particularly, 8 questions were posed to human evaluators on a 4 point scale (yes, maybe, no, unclear) - focusing on product fidelity, logo generation fidelity, realistic product use, product size, presence of background/foreground hallucinations, product placement, image safety and likelihood to use the image as a business owner. A 'yes' was considered a pass, and a majority vote over 3 human raters was taken to consider an image to pass human rating.

A.4 GENERATED SAMPLES

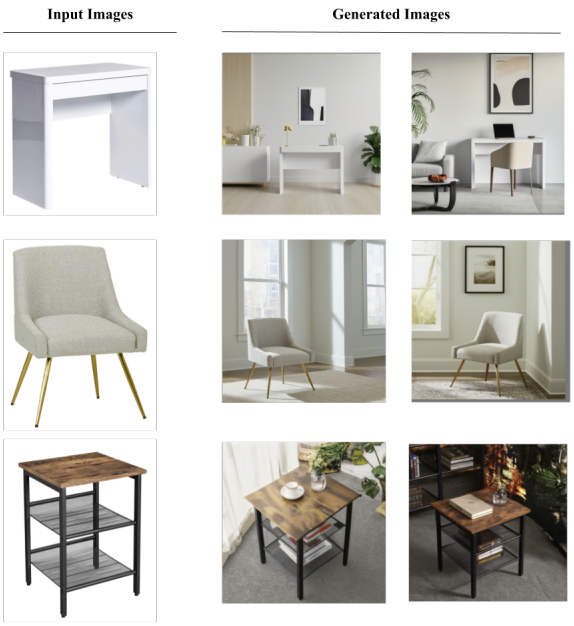


Figure 6: Recontextualization results on the ABO dataset, demonstrating high fidelity, object occlusions, novel perspectives, and realistic lighting. Object occlusions are present for the desk and table and all the products demonstrate novel view synthesis, where the angle the object is placed at in the generated images is different than the provided image. The chair highlights the ability of the model to apply realistic and diverse lighting on an object.

A.5 LORA RANK ABLATION QUALITATIVE RESULTS

In addition to the results above, we also showcase qualitative results from ablations on the LoRA rank used during fine tuning and number of training steps in Figure 8. We uran our process for the same set of input images of the desk and then finetune a model using LoRA rank=1 and rank=64. As shown in Figure 8, as both models were trained for more steps, they improved in the ability to generate an accurate depiction of the desk, and they both lost some amount of complexity and diversity in the scenes they generate.

The rank 1 model learned the appearance of the desk much quicker than than rank 64 model (the samples display high product fidelity at 700 steps as opposed to 1800 steps 8). When trained for the same number of steps as the rank 64 model, the rank 1 model showed clear signs of overfitting, in some cases even recreating the training data regardless of the prompt given to the model. The rank 1 model also began to generate very simple scenes in step 700, even once it gained the ability to faithfully render the desk.

In contrast, the rank 64 model took much longer to learn the specifics of the desk's appearance, but once it did, we note that the model maintains the ability to generate diverse and realistic scenes much better than the rank 1 model. The scenes from the rank 64 model show the desk in more unique

(a) "A round table with a marble top, placed in a sun-drenched breakfast nook with wicker chairs and potted herbs on the windowsill."

(b) "An antique table with ornate carvings and claw feet, displayed in a grand dining room with a crystal chandelier and antique china on display."

Figure 7: Prompts used to generate model negative samples.

viewpoints and in more complex environments, suggesting the rank 1 model lost this ability while it quickly learned the appearance of the desk. However, even the rank 64 model began to generate less diverse and and more simple scenes over time as it learns the appearance of the desk.

This ablation demonstrates the trade off between product fidelity and generating realistic and diverse scenes. As such, choosing the optimal rank and training steps is equivalent to finding the balance point between product fidelity and generating diverse images. Using a higher LoRA rank allows the model to learn more slowly and better maintain its ability to generate complex and diverse scenes, making it easier to find the balance point for this tradeoff.

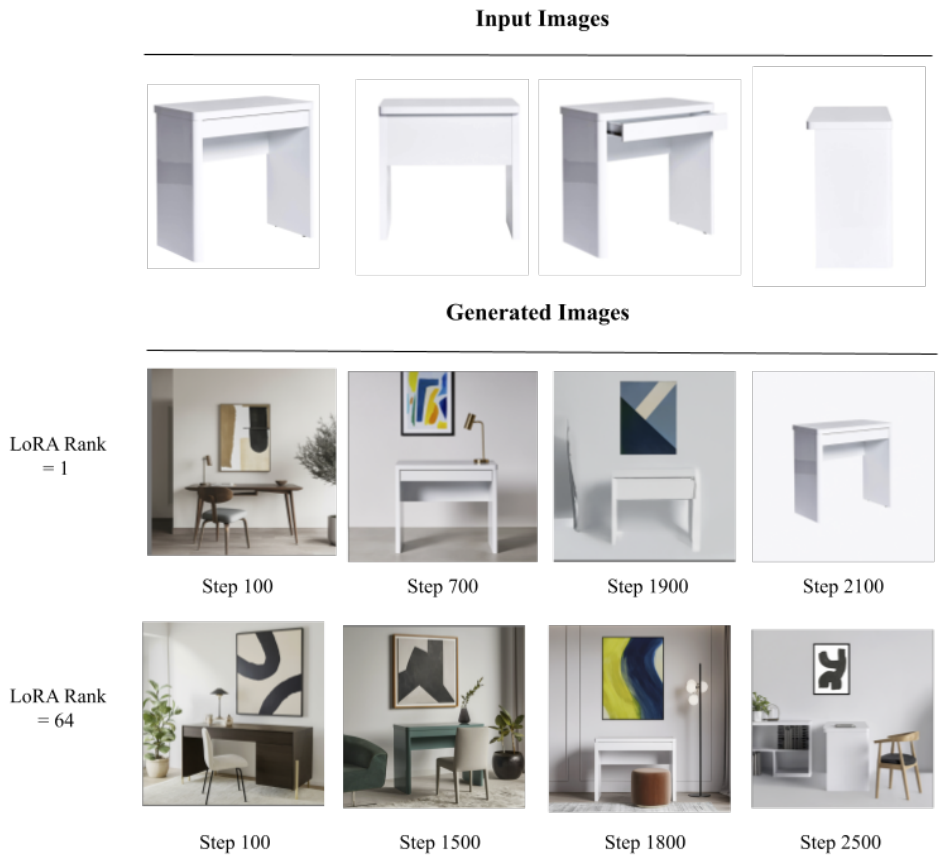


Figure 8: Recontextualization results for a desk with varied LoRA rank values and number of steps the model is trained for. These results showcase the tradeoff between object fidelity and generating novel and realistic scenes containing the products. LoRA rank 1 achieves high product fidelity in much fewer training steps but generates less diverse scenes, before eventually overfitting and beginning to reproduce training data regardless of prompt. LoRA rank 64 takes much longer to learn the product, but generates more varied scenes utilizing less common perspectives of the object.