Dynamic Product Image Generation and Recommendation at Scale for Personalized E-commerce

Ádám Tibor Czapp
adam-tibor.c@taboola.com
Taboola Budapest
Budapest, Hungary

Mátyás Jani
matyas.j@taboola.com
Taboola Budapest
Budapest, Hungary

Bálint Domián
balint.d@taboola.com
Taboola Budapest
Budapest, Hungary

Balázs Hidasi
balazs.h@taboola.com
Taboola Budapest
Budapest, Hungary

ABSTRACT

Coupling latent diffusion based image generation with contextual bandits enables the creation of eye-catching personalized product images at scale that was previously either impossible or too expensive. In this paper we showcase how we utilized these technologies to increase user engagement with recommendations in online retargeting campaigns for e-commerce.

CCS CONCEPTS

• **Information systems → Recommender systems**.

KEYWORDS

stable diffusion, contextual bandit, recommender systems, ctr

ACM Reference Format:
Ádám Tibor Czapp, Mátyás Jani, Bálint Domián, and Balázs Hidasi. 2024. Dynamic Product Image Generation and Recommendation at Scale for Personalized E-commerce. In *18th ACM Conference on Recommender Systems (RecSys '24), October 14–18, 2024, Bari, Italy*. ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3640457.3688045

1 INTRODUCTION

The engagement of users with recommended products is greatly influenced by their presentation [3], which is second only to their relevance. This is especially true in online advertisement where the users' primary focus is not on the recommendations. Creatives of ad campaigns are often designed with great care, but this approach does not scale for product level ad campaigns (e.g. retargeting, Dynamic Product Ads (DPA)) where each item of the product catalog is subject to be recommended on any of the ad placements with different aspect ratios. The common approach is to show the original product image, optionally with additional design elements. We improve upon this by using image generation methods and place the products in appropriate environments. These more eye-catching creatives increase user engagement. This solution is also useful for enhancing user generated product photos (e.g. on marketplaces) that might have been taken in less appealing environments.

RecSys '24, October 14–18, 2024, Bari, Italy

ACM ISBN 979-8-4007-0505-2/24/10
https://doi.org/10.1145/3640457.3688045

Figure 1: Mild and extreme artifacts produced by inpainting.

Figure 2: Examples of color variations through conditioning.

2 GENERATING PRODUCT IMAGES

We designed a novel feature for our recommender system that creates eye-catching product images in the given size by generating the surrounding background. Background generation utilizes Stable Diffusion [5] – a popular diffusion [6] based image generation model – through the diffusers [7] package. The model is prompted with predefined prompts that describe environments appropriate for the product category. The product itself is not modified in any way.

AI generated images in commercial applications: Even with the rapid improvement of the technology, generative models – by their nature – are prone to create imperfect images. Fortunately, the application domain is somewhat permissive to smaller imperfections. Advertisements are rarely the main focus of the users, thus minor issues are likely overlooked and forgotten. But it is still crucial to build a pipeline where the chance of critical failures is low. The naive approach for this task is inpainting: mask the product and generate the background around it. However, this approach often produces artifacts by extending the product with virtual parts (see Figure 1). A better solution is to utilize ControlNet [8], which allows for the injection of additional constraints into the image generation process. Our pipeline uses the edges of the product as the constraint in ControlNet. While this approach does draw under the product mask, it creates an object similar to the product that is later replaced by the real product. This significantly reduces the artifacts and "floating product" images as well. We further improved this solution by conditioning the generation on the product. This technique reduces visible outlines and allows for subtly adapting the color palette and the composition without changing the prompt (see Figure 2).

The main steps of the generation pipeline are shown on Figure 3:

(1) Object detection and masking.

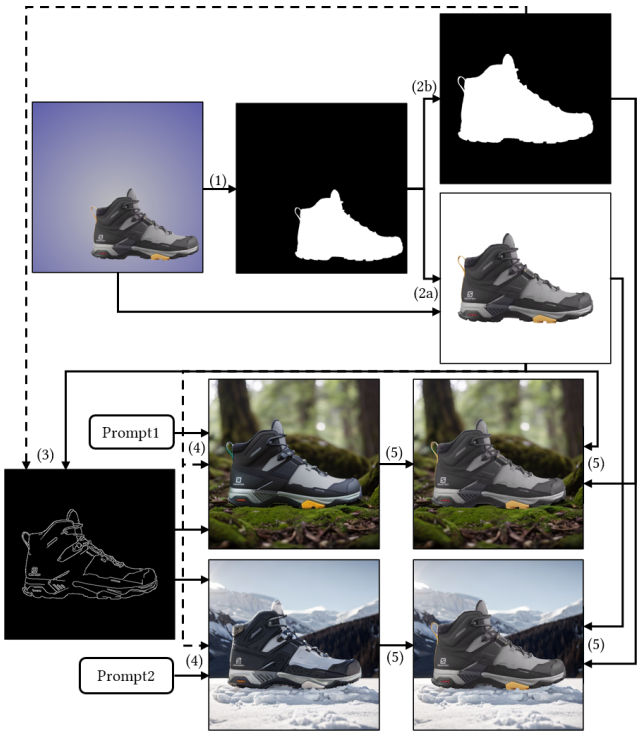

Figure 3: Main steps of the background generation pipeline.

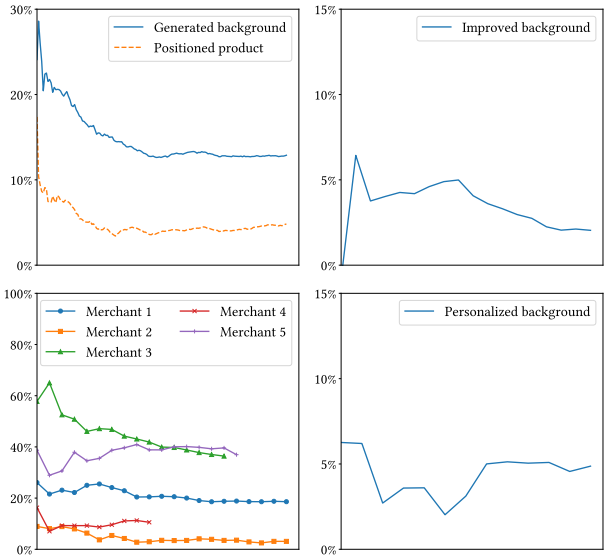

Figure 4: Relative CTR gains. (top left) phase I: positioned product on white & generated background vs. original image; (top right) phase II: improved vs. old pipeline; (bottom left) phase II: generated image (improved pipeline) vs. original image; (bottom right) phase III: personalized vs. non-personalized background.

(2) Position & scale the product (and mask) according to the placement. Remove original background, if applicable.
(3) Edge detection (optional: reinforce contours using the mask).
(4) Image generation using the edges as constraints (optional: also condition on the product to increase quality).
(5) Cut back the original product onto the generated image.

Production considerations: The speed of image generation is too slow for the strict response time requirements of recommender systems, therefore the first request for a given (image, prompt, size) triplet only puts the request on a queue and the recommendation is served with the original image. Once the image is ready, the service calls back the recommender system that caches the generated image for future use. Costs are kept down by utilizing additional caches (e.g. for masks) and grouping similar aspect ratios together. As no machine learning method is flawless, there are certain points where humans can intervene if necessary.

Personalizing the experience: Having a pretty background is not enough, it has to be appealing to the user that views it. Most products look good in multiple environments. Different users find different variations attractive and preferences might be influenced by where the recommendation is shown. Individually personalizing for every user–item–placement triplet is not possible due to the sparseness of the data. We deployed a solution similar to [1] and use the LinUCB [2, 4] contextual bandit algorithm to select the prompt that has the best estimated CTR in the given context – defined by user, item and placement features – from the predefined prompt pool belonging to the product's category.

3 ONLINE TESTS

Our approach was validated by online A/B tests. Multiple independent tests were performed over different timeframes and product catalogs. The sizes of the product catalogs were in the range of a few thousands to several tens of thousands items, and most of the products were in the apparel category (clothing, footwear, accessories, etc.). The primary metric was click-through rate (CTR), because creatives have direct effect on clicks only. However, we observed that other metrics (e.g. number of conversions, cost per action (CPA)) also improved as an indirect effect of driving more users to the merchant's site. The test was conducted in three phases. Figure 4 summarizes the results. All CTR gains are statistically significant at $p < 0.05$.

Phase I focused on examining the added value of **generated backgrounds**, and was executed from September 2023 to February 2024. The performance of the generated images was compared to that of the original product images. Since the positioning and the size of the product can influence CTR, the output of step (2a) from Figure 3 was also included in the comparison as a separate group. The test validated both assumptions: (1) ~ 5% improvement can be gained by simply paying attention to product positioning/scaling; (2) products with generated backgrounds performed even better with ~ 15% gain over the baseline.

Phase II validated that the **improvements we made on the pipeline** resulted in additional CTR gains. This phase was also used to check the performance of the improved pipeline against the original product images **in multiple experiments**. While generated images always outperform the baseline, the relative CTR gain

varies in a wide range ($\sim 4-40\%$ in these experiments). The exact number mainly depends on the product catalog of the merchant, the ad placements, and the composition and quality of the original product images.

Phase III investigated the added benefit of **personalizing the backgrounds**. Three appropriate prompts were defined for each product category. In the treatment group, prompts are selected by the LinUCB algorithm based on context and user features, while the control group is served with images based on randomly selected prompts. Results suggest that even this lightweight personalization can further improve the performance of the system by $\sim 5\%$.

AUTHOR BIOS

Ádám Tibor Czapp and **Bálint Domián** are Machine Learning Engineers working on recommender systems & algorithms. **Mátyás Jani** is a Senior Machine Learning Software Engineer with 5+ years of experience in putting research results into live production systems, designing and implementing data pipelines and services around algorithms. **Balázs Hidasi** is a Leading Research Scientist with 15+ years of experience in machine learning and 10+ in leading machine learning and data science teams; he conducts research, directs research projects and oversees the machine learning related initiatives of Taboola Budapest (formerly Gravity R&D).

REFERENCES

[1] Fernando Amat, Ashok Chandrashekar, Tony Jebara, and Justin Basilico. 2018. Artwork personalization at netflix. In *Proceedings of the 12th ACM Conference on Recommender Systems* (Vancouver, British Columbia, Canada) *(RecSys '18)*. Association for Computing Machinery, New York, NY, USA, 487–488. https://doi.org/10.1145/3240323.3241729

[2] Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. 2011. Contextual Bandits with Linear Payoff Functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 15)*, Geoffrey Gordon, David Dunson, and Miroslav Dudík (Eds.). PMLR, Fort Lauderdale, FL, USA, 208–214. https://proceedings.mlr.press/v15/chu11a.html

[3] Anjan Goswami, Naren Chittar, and Chung H Sung. 2011. A study on the impact of product images on user clicks for online shopping. In *Proceedings of the 20th international conference companion on World wide web*. 45–46.

[4] Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. 2010. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web (WWW '10)*. ACM. https://doi.org/10.1145/1772690.1772758

[5] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

[6] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. (2021).

[7] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. 2022. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers.

[8] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3836–3847.