Supplemental Material

Zhenbang Du[1,2*†], Wei Feng[2*], Haohan Wang[2], Yaoyu Li[2], Jingsen Wang[2], Jian Li[2], Zheng Zhang[2], Jingjing Lv[2], Xin Zhu[2], Junsheng Jin[2], Junjie Shen[2], Zhangang Lin[2], and Jingping Shao[2]

[1] School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China
dzb99@hust.edu.cn
[2] Retail Platform Operation and Marketing Center, JD, Beijing, China
{fengwei25, wanghaohan1, liyaoyu1, wangjingsen, lijian21, zhangzheng11, lvjingjing1, zhuxin3, jinjunsheng1, shenjunjie, linzhangang, shaojingping}@jd.com

In this supplemental material, we first describe the construction of the RF1M dataset (Sec. 1). The Recurrent Generation is described in Sec. 2. We then detail the human-involved experiments (Sec. 3) and address our responsibilities towards human subjects (Sec. 4). The detailed configuration for image generation is outlined in Sec. 5, followed by additional visualization results in Sec. 6. Finally, we discuss our ethical concerns (Sec. 7) and future work (Sec. 8).

1 Dataset Construction

1.1 Annotation Guidance

During the **RF1M** annotation process, annotators are provided with the original product images, product captions, and generated images along with the following instructions:

> Based on the product image and caption, determine the category of the advertising image:
>
> 1. **Space Mismatch.** Image where the product and background have inappropriate spatial relations contrary to physical laws, such as a part of the product is floating.
> 2. **Size Mismatch.** Discrepancies between the product size and its background violate common sense, *e.g.*, a massage chair appears smaller than a cabinet.
> 3. **Indistinctiveness.** Image where the product fails to stand out due to background complexity or color similarities, and you cannot clearly figure out which or where the product is. Additionally, this confusion might lead to misconceptions about the product itself, mistakenly suggesting the inclusion of extra freebies.

[*] Equal contribution
[†] Work done while interning at JD.com

4. **Shape Hallucination.** Background that erroneously extends the product shape, adding elements like pedestals or legs, and giving you incorrect perceptions about the product.
5. **Available.** Image deemed available for advertising purposes, not falling into any of the above categories.

* If the image belongs to several types simultaneously, *e.g.*, the product has an inappropriate size and extended shape, annotate it with the most significant issue.

And some examples to be annotated are shown in Fig. 1

Fig. 1: Some examples shown to annotators. The translations of Chinese captions are in the brackets.

1.2 More Examples in Dataset

All samples in the RF1M dataset are generated using the approach outlined in paper, created with the original diffusion model before fine-tuning. We showcase additional examples of both available and unavailable generated images in Figs. 2 and 3. For brevity, auxiliary modalities such as product captions, depth images, salience images, and prompts are not shown. The dataset includes a broad variety of products and background scenes with a strong visual appeal.

2 Recurrent Generation

We show our Recurrent Generation strategy in Algorithm 1.

Algorithm 1: Recurrent Generation

Input : Depth estimation model $Depth()$, Salience detection model $Salience()$, $RFNet()$, product image I_o and caption Cap, denoising process $Denoise()$, max attempts K

attempt $k \leftarrow 0$, initial noise $x_T \sim \mathcal{N}(0, \boldsymbol{I})$;
while $k < K$ **do**
  $x_0 = Denoise(x_T)$; $I_g \leftarrow x_0$; # latent to image
  $y_{pre} = RFNet(I_o, I_g, Depth(I_g), Salience(I_g), Cap)$;
  if y_{pre} == *"Available"* **then**
    $I_{ava} = I_g$;
    Break;
  else
    $I_{ava} = None$;
  end
  $k \leftarrow k + 1$;
end
Output: I_{ava}

3 Human-involved Experiments

For **human availability inspection**, the annotators who involved in annotating the RF1M dataset are engaged. They followed the same instructions and data format as during dataset construction (Sec. 1.1) to mark whether the generated images were available for advertising use.

For **human preference assessment**, we generated advertising images using different approaches with the same seed. We combined them with the product image in a random sequence to prevent visual fatigue, as shown in Fig. 4. Each advertising practitioner ranked them based on personal preference, providing a final ranking like "1423".

4 Responsibility to Human Subjects

We hired annotators and practitioners to obtain annotations and feedback. We have conducted a thorough review to ensure that the dataset did not include personally identifiable information or offensive content.

5 Detailed Configuration

We detail our configuration for image generation in Table 1.

6 More Visualization Results

6.1 Image Collapse

In Fig. 5, we display collapsed images generated by ReFL. Despite varying product images and prompts, the outputs are uniformly degraded and filled with

Table 1: Image generation configuration. "RF Interval" denotes the steps interval to fine-tune the diffusion model.

Image Size	512×512	Clip Skip	2
Sampler	DDIM	Total Steps	40
LoRA Scale	0.8	ControlNet Scale	1.0
Guidance Scale	7.5	RF Interval	30-40
Prompt Example: "An item sitting on a wooden table, (outdoor background: snowy mountain environment, heavy snow, snowflakes), depth of field, close-up, best quality, rich detail, 8k."			
Negative Prompt: "text, username, logo, (low quality, worst quality:1.4), (bad anatomy), (inaccurate limb:1.2), bad composition, inaccurate eyes, extra digit, fewer digits, (extra arms:1.2), watermark, multiple moles, mole on body, drawing, painting, crayon, sketch, graphite, impressions."			

strange textures, similar to adversarial samples [1, 2]. Thus the direct gradient backpropagation without regularization can be interpreted as training the diffusion model to produce images that deceive the RFNet.

6.2 Generalization Results

Fig. 6 shows the consistency of background impressions while the available rate improves with different LoRA/diffusion models.

6.3 Integration with Other Feedback

We present further examples with different feedback combination strategies in Fig. 7. Integrating F_{IR} without L_{CC} leads to overly detailed backgrounds. Combining F_{IR} and L_{CC} preserves background aesthetics, highlighting the role of L_{CC} to maintain the text prompt conditions in backgrounds.

6.4 Comparison with Other Refining Approaches

Fig. 8 and Fig. 9 showcase advertising images generated by different approaches. Our approach yields images with a high available rate for advertising purposeS while maintaining appealing visuals.

7 Ethical Concerns

Regarding image manipulation and automated advertising image generation, there's a risk of producing content that could be unethical or illegal, such as violating individual portrait rights or creating discriminatory content. Thus, these technologies require stricter oversight. In our scenario, we carefully select all product images to ensure neutrality and compliance with advertising standards. The prompts for background generation are also meticulously crafted to produce content that is neutral and meets regulatory requirements.

As for our RF1M dataset, all product information is sourced from an e-commerce platform and has undergone pre-review by both the merchants and the platform. The prompts for generating advertising images are created by professional advertising practitioners. During the generation process, we use Stable Diffusion's official safety checker to filter out NSFW content. Finally, professionals review and screen the dataset, ensuring it is free from bias or offensive content and complies with relevant local laws.

In terms of future applications, while these technologies offer significant benefits in efficiency and personalization, they also pose risks related to misinformation. It's vital to maintain high transparency levels to address these concerns. This includes clearly labeling AI-generated images, allowing consumers to differentiate between AI-generated and human-created content. Furthermore, developing and adhering to business standards and ethical guidelines is crucial to ensure AI's use in advertising upholds truthfulness and honesty. The automation of creative tasks could also alter the job market in creative industries. Rather than seeing AI as a replacement for human creativity, we envision a collaborative approach where AI serves as a tool for creative professionals, helping the workforce stay competitive and ensuring human creativity remains vital in advertising.

8 Future Work

Our research currently emphasizes the availability of generated images. Moving forward, we plan to incorporate additional types of feedback to refine diffusion models. For example, click-through rate (CTR) of different images reveals the customers' preferences. Our future work will explore utilizing the CTR to enhance the appeal of generated advertising images.

Additionally, we have found that conventional metrics for evaluating generated images are infeasible for our specific use case. Given that the central focus of advertising image is the identical product, traditional metrics like FID do not provide an accurate assessment of image quality, *i.e.*, different approaches yield similar FIDs despite varying levels of visual attractiveness. For example, ReFL scores an FID of 18.52 compared to our approach's 16.26 in Sec. 6.4. Therefore, we intend to develop a more appropriate evaluation metric for assessing the quality of generated backgrounds.

Fig. 2: Some available generated images in the dataset, encompass various products in multiple scenarios.

Fig. 3: Some unavailable generated images in the dataset, encompass four bad cases.

Fig. 4: Sample shown to advertising practitioners to assess their preferences.

Fig. 5: Collapsed images generated by ReFL.

Fig. 6: Comparison of the original model and our fine-tuned model using different LoRA and diffusion models.

Fig. 7: Comparison of different feedback combination strategies.

Fig. 8: Generated advertising images using different approaches.

Fig. 9: Generated advertising images using different approaches.

References

1. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: ICLR (2015)
2. Kurakin, A., Goodfellow, I.J., Bengio, S.: Adversarial examples in the physical world. In: ICLR Workshop (2017)