Towards Reliable Advertising Image Generation Using Human Feedback

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

Abstract. In the e-commerce realm, compelling advertising images are pivotal for attracting customer attention. While generative models automate image generation, they often produce substandard images that may mislead customers and require significant labor costs to inspect. This paper delves into increasing the rate of available generated images. We first introduce a multi-modal Reliable Feedback Network (RFNet) to automatically inspect the generated images. Combining the RFNet into a recurrent process, Recurrent Generation, results in a higher number of available advertising images. To further enhance production efficiency, we fine-tune diffusion models with an innovative Consistent Condition regularization utilizing the feedback from RFNet (RFFT). This results in a remarkable increase in the available rate of generated images, reducing the number of attempts in Recurrent Generation, and providing a highly efficient production process without sacrificing visual appeal. We also construct a Reliable Feedback 1 Million (RF1M) dataset which comprises over one million generated advertising images annotated by human, which helps to train RFNet to accurately assess the availability of generated images and faithfully reflect the human feedback. Generally speaking, our approach offers a reliable solution for advertising image generation. Our dataset and code are available at https://github.com/ZhenbangDu/Reliable_AD.

Keywords: Diffusion Model · Human Feedback · E-commerce

1 Introduction

An attractive advertising image is essential for e-commerce success, as it can lead to a higher click-through rate (CTR) [51]. Manual creation requires significant

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

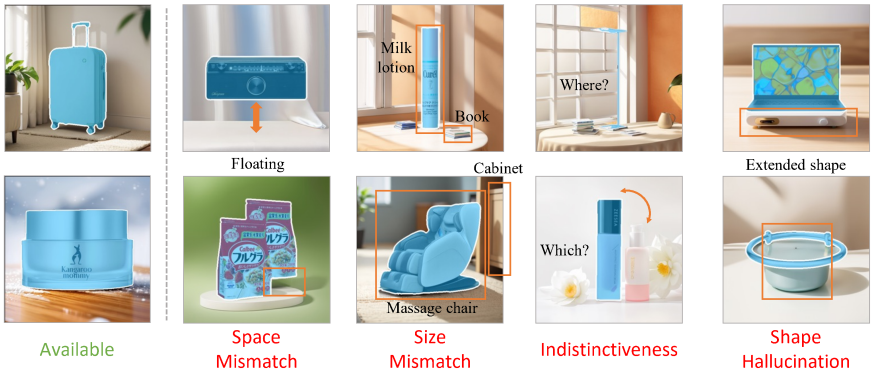

Fig. 1: The available generated advertising images and different types of bad cases. The products are highlighted by blue masks. Bad cases bring misleading information, *e.g.*, the unrealistic sizes or shapes of products, and customers may have difficulty discerning the products in images.

labor costs, therefore the demand for automatic advertising image production is on the rise. However, previous approaches [4, 5, 32, 53] often result in visual mismatches between the product and background. Advanced diffusion models [17] offer a remedy. A combination with ControlNet [58] has shown promise in producing harmonious backgrounds for products while keeping the details of products identical for advertising use [26, 49].

Despite the potential of generative models to create appealing backgrounds, we have observed a frequent production of substandard advertising images as depicted in Fig. 1, which encompass various cases such as space and size mismatches, indistinctiveness, and shape hallucination. These flawed images can lead to customer misunderstandings about products and bring a subpar shopping experience, therefore considerable labor is necessitated to inspect the generated images. Such drawbacks limit the broader application of generative models in advertising image production. So the core problem is *a low rate of available images*. It poses a new challenge to us how to establish a reliable advertising image generation pipeline capable of producing images with a high available rate.

A natural solution is to generate images repeatedly until an available image is obtained (Recurrent Generation) due to the randomness in generation. To substitute labor inspection in this repetitive process, a novel Reliable Feedback Network (**RFNet**) is introduced to act as a human inspector, assessing the availability of generated advertising images. As simply depending on a single generated image, the model can not effectively obtain the knowledge pivotal to the precise inspection, *e.g.*, what the product is and how the product appears in the background. So RFNet integrates multiple auxiliary modalities to provide critical information contributing to the judgment of different unavailable cases. Meanwhile, we construct a large-scale dataset called Reliable Feedback 1 Million dataset (**RF1M**), which includes over one million elaborate generated advertis-

ing images with rich human annotations, playing a crucial role in training RFNet to mirror human feedback accurately.

While Recurrent Generation greatly increases the number of available images, multiple attempts will dramatically prolong the generation process owing to the inherent poor ability of generated models. Using human feedback to enhance the capability of diffusion models [34,45] provides a viable option, which has successfully improved the visual quality of generated images [13, 47, 56]. However, the visual quality and availability of the generated images show a trade-off relationship, *e.g.*, products with repetitive and simple backgrounds gain a high available rate yet low aesthetics. To tackle this, we propose a novel loss term, Consistent Condition (CC) regularization L_{CC} to counteract uniformity and degradation of the generated backgrounds, circumventing the adversarial nature of conventional Kullback-Leibler (KL) regularization [13, 47]. Utilizing this regularization term, the feedback on generated images' deviation from the available type assessed by RFNet is directly back-propagated to fine-tune the diffusion model (**RFFT**). Our approach significantly enhances the available rate of generated advertising images without altering their aesthetics, offering a comprehensive solution to the challenge of reliable advertising image generation.

Our main contributions include:

- An advertising image generation solution Recurrent Generation alongside the novel multi-modal model, **RFNet**, which simulates human feedback and effectively utilizes various modalities to help attribute fine-grained issue types.
- A straightforward and effective approach, **RFFT**, to refine the diffusion model using human feedback, along with an innovative Consistent Condition regularization to prevent collapse.
- A large-scale multi-modal dataset, **RF1M**, comprising over one million generated advertising images with rich annotations.

2 Related Work

In this section, we review prior works on advertising image generation and the use of human feedback to refine diffusion models.

2.1 Advertising Image Generation

Automatic advertising image generation offers significant convenience for e-commerce. Current approaches generally fall into two folds, template-based and generation-based. Template-based approaches [4, 5, 32, 53] typically utilized predefined templates to assemble various elements into images. However, such approaches failed to achieve aesthetic harmony between the product and its background, resulting in a discernible disconnection. Additionally, designing a variety of templates was costly. To counter this, generation-based approaches [23, 26, 49] employed GANs [15, 21] or advanced text-to-image diffusion models [17, 39, 43, 44] to produce backgrounds that coordinate with the appearance of products, *e.g.*,

Wang *et al*. [49] made use of the inpainting technique [31] with product masks to produce personalized backgrounds. Although elaborate and coordinated advertising images were obtained by these approaches, the bad cases that happened in the generation process could result in confusion and complaints of customers. We tackle this issue with a Recurrent Generation strategy, complemented by the accurate inspection capability of RFNet. Additionally, our RFFT enhances the reliability of the diffusion model and further promotes the efficiency of the production process.

2.2 Refining Diffusion Models with Human Feedback

Reinforcement Learning from Human Feedback (RLHF) employs human-derived feedback to fine-tune models, aiming for outcomes that human prefers [3, 6, 20, 29, 34, 45, 60]. In aligning diffusion models with human preferences, RLHF plays a crucial role [12, 24, 25, 37, 59]. Works such as DDPO [2] and DPOK [13] treated the denoising process as a multi-step Markov decision process. They used a Policy Gradient [42] approach to update model parameters based on feedback from pre-trained reward models. Similarly, D3PO [57] and Diffusion-DPO [47] enhanced diffusion models using human comparison data, avoiding the need for reward model training. Contrary to the reinforcement learning strategy, ReFL [56] and DRaFT [7] directly fine-tuned diffusion models using the gradient of differentiable rewards in an efficient end-to-end manner [48]. Our work pioneers in using human feedback to establish a reliable advertising image generation solution. Additionally, our approach addresses the collapse during fine-tuning, resulting in a high available rate of generated images without compromising their appearance.

3 Dataset

The Reliable Feedback 1 Million (**RF1M**) dataset, constructed through meticulous human feedback, serves as a pivotal resource for inspecting and improving the generation of advertising images. Compared to prevailing large-scale image generation dataset Laion-5B [41], DiffusionDB [52] and Pick-a-Pic [22], RF1M is specifically designed for the advertising domain, addressing the acute need for expansive data resources in this field. Here's an in-depth look at its composition, annotation, and potential impact on the community.

Composition. The dataset is generated using a collection of extensive products from JD.com. It encompasses 1,058,230 samples, each consisting of a variety of components aimed at providing a comprehensive understanding of advertising image generation:

- The generated advertising image with corresponding transparent background product image, and carefully designed prompts by professional designers.
- Depth and salience images created by the dense prediction transformer [38] and U2-Net [36] trained on e-commerce data, along with product caption, which assists in inspecting the availability of generated advertising image.

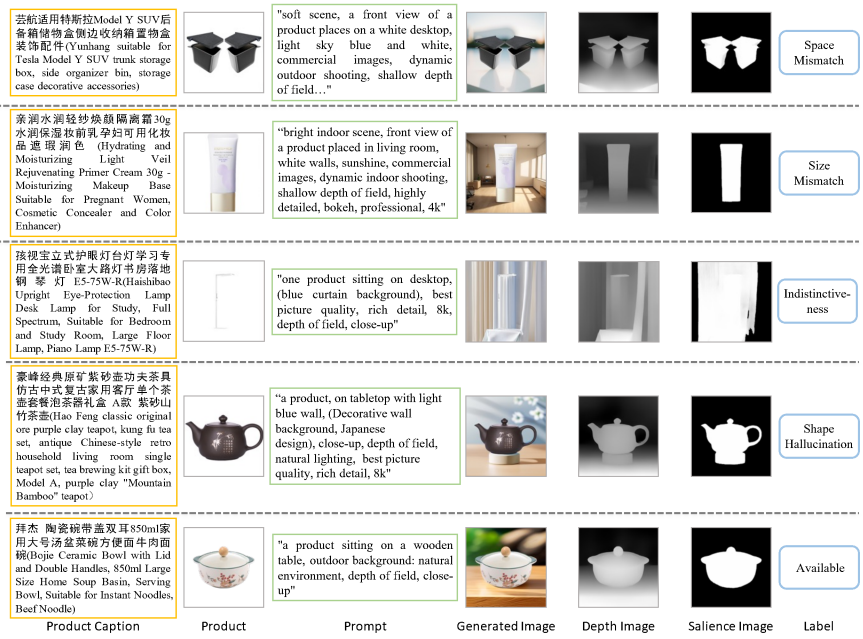

Fig. 2: Some examples in RF1M. Each comprises rich annotations. The translations of Chinese captions are in the brackets.

– Human-annotated label indicating the availability of image for advertising use.

These elements collectively offer a rich foundation for analyzing the generated advertising images. Some examples are exhibited in Fig. 2.

Annotation. The annotators involved are well-versed in advertising and possess a deep understanding of the standards for advertising images. Within the dataset, samples have been further classified into five fine-grained categories as illustrated in Fig. 1:

– **Available.** Images deemed suitable for advertising purposes.
– **Space Mismatch.** Images where the product and background have inappropriate spatial relations, such as a part of the product is floating.
– **Size Mismatch.** Discrepancies between the product size and its background, *e.g.*, a massage chair appears smaller than a cabinet.
– **Indistinctiveness.** Images where the product fails to stand out due to background complexity or color similarities.
– **Shape Hallucination.** Backgrounds that erroneously extend the product shape, adding elements like pedestals or legs.

Potential Impact. RF1M, with its multi-modal design and comprehensive features, is the cornerstone of the RFNet training and RFFT, and would further

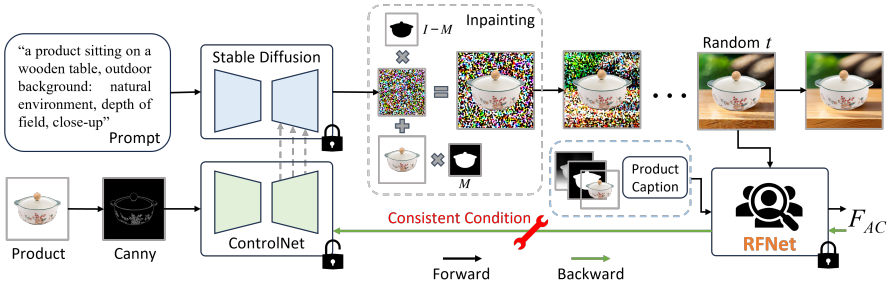

Fig. 3: An overview of image generation-inspection pipeline. The advertising image is generated using product image and prompt by inpainting. And the feedback F_{AC} provided by the RFNet is used to fine-tune the ControlNet with Consistent Condition regularization.

significantly impact the realm of e-commerce advertising and beyond, featuring three highlights:

1. **Large scale.** With its vast array of product categories and image types, RF1M surpasses previous advertising datasets BG60k [49], PPG30k [26] and human feedback dataset ImageRewardDB [56], RichHF-18K [27], while has comparable size to Pick-a-Pic [22], providing a robust base for RFNet to mirror human feedback accurately across the diverse advertising image generation tasks.
2. **Scalability.** The multi-modal nature of RF1M provides sufficient information to RFNet to make judgments precisely. Beyond generating reliable advertising images, it also supports tasks such as advanced image understanding and image matting. This flexibility ensures the dataset meets the changing demands and can be applied in numerous areas within and beyond advertising.
3. **Visual appeal.** The prompts and generation models are carefully designed for the products' characteristics, so the images have satisfying aesthetics and can attract customers' attention. We conducted a one-week online A/B test done in JD.com, resulting in a 2.2% increase in CTR from over 60 million exposures, underscoring the high quality of these images, which accurately captured user preferences.

4 Methodology

4.1 Preliminaries

Our approach for generating advertising images is depicted in Fig. 3. We start with a text prompt describing the desired background and a product image I_o with a transparent background. The prompt is input into Stable Diffusion [39],

and I_o is pre-processed with canny edge detection before being fed into the ControlNet [58]. We adopt DDIM [43] as our denoising schedule, the latent representation x at step t is calculated as following:

$$x_t = \sqrt{\bar{\alpha}_t}\frac{x_{t+1} - \sqrt{1-\bar{\alpha}_{t+1}}\epsilon_\theta\left(x_{t+1}, t+1\right)}{\sqrt{\bar{\alpha}_t + 1}} + \sqrt{1-\bar{\alpha}_t}\epsilon_\theta\left(x_{t+1}, t+1\right), \tag{1}$$

where ϵ_θ represents the model [39, 40] that predicts added noise, and $\{\bar{\alpha}\}$ is a set of coefficients controlling the forward noise-adding process. To preserve the product's integrity and ensure a cohesive background, we employ an inpainting technique [31]. The latent representation x_t is processed by:

$$x_t = (\boldsymbol{I} - M) \otimes x_t + M \otimes x_o, \tag{2}$$

where x_o is the latent of I_o, M is product mask, and $\otimes$ denotes the element-wise multiplication. After the x_0 is obtained, this latent is converted to the generated image I_g.

4.2 Recurrent Generation with RFNet

Due to the inherent randomness, repeated generation can significantly expand the number of available images. To automate the inspection process and eliminate human participation in this process, we introduce a multi-modal model, RFNet, to determine whether the generated image is available precisely, as illustrated in Fig. 4. In addition to I_o and I_g, RFNet combines information from auxiliary modalities:

- The depth image I_d of I_g produced by a depth estimation model, highlights the product's position relative to the background;
- The salience image I_s of I_g, created by a salience detection model, outlines the product;
- The product caption Cap which provides insight into the product's attributes.

I_o, I_g, I_d and I_s are fed into an image encoder to acquire respective image embeddings $\{\boldsymbol{e_o}, \boldsymbol{e_g}, \boldsymbol{e_d}, \boldsymbol{e_s}\}$. Concurrently, the Cap is input into BERT [9] to obtain the text embedding $\boldsymbol{e_c}$, aiding in recognizing the attributes of the product.

Since product captions often contain excessive information, *e.g.*, brands, we focus on distilling the vision-related attributes from the caption. So N_1 Feature Filter Module (FFM)s are first employed, each consisting of a cross-attention layer and several convolution layers. The output of FFM is formulated as:

$$\boldsymbol{e_f} = \mathrm{Conv}\left(\mathrm{Conv}(\mathrm{CrossAttention}(\boldsymbol{e_o}, \boldsymbol{e_c})) \otimes \mathrm{Conv}(\boldsymbol{e_o})\right) + \boldsymbol{e_o}, \tag{3}$$

where $\boldsymbol{e_o}$ serves as $Query$ with $\boldsymbol{e_c}$ acting as both Key and $Value$ in cross-attention layer, Conv() denotes convolution layer with 1×1 kernels, and $\otimes$ signifies element-wise multiplication. This process ensures that critical information

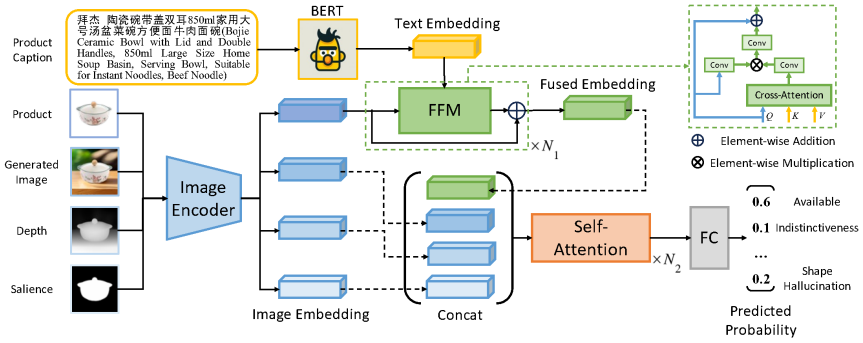

Fig. 4: The proposed RFNet. Multiple auxiliary modalities contribute to the final inspection. The translation of the Chinese caption is in the brackets.

from the caption is effectively integrated with the image embeddings, augmenting the model's understanding of the product.

With the fused embedding $\boldsymbol{e_f}$, different features are further integrated through N_2 self-attention layers [46],

$$\boldsymbol{f} = \text{SelfAttention}(\text{Concat}(\boldsymbol{e_f}, \boldsymbol{e_g}, \boldsymbol{e_d}, \boldsymbol{e_s})), \tag{4}$$

where Concat() stands for concatenation. These stacked layers capture the critical features across the embeddings. Finally, a fully-connected classifier determines each case's probability of the generated image.

Trained on large-scale RF1M, RFNet assesses the availability of generated advertising images accurately by considering a comprehensive set of visual and textual features, and providing nuanced feedback. This capability, combined with the Recurrent Generation strategy (illustrated in supplemental material), significantly increases the number of available generated images for advertising use in an automatic manner.

4.3 RFFT with Consistent Condition regularization

Although Recurrent Generation could produce more available images in total, the inherited poor ability of the generative model results in a prolonged and inefficient production process, posing a great challenge to the application. Our end-to-end generation-inspection pipeline allows feedback gradients from RFNet to directly fine-tune the diffusion model, enhancing its capability. Specifically, our proposed RFFT selects a random step t among last 10 steps during the 40-step denoising process to generate the $\hat{x}_0^t = \frac{x_{t+1} - \sqrt{1-\bar{\alpha}_{t+1}}\epsilon_\theta(x_{t+1}, t+1)}{\sqrt{\bar{\alpha}_{t+1}}}$ [35, 43, 56]. The resulting $\hat{x}_0^t$ is post-processed to $\hat{I}_g^t$ then inspected by RFNet to determine its availability, with feedback calculated as follows:

$$F_{AC} = -\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{y}_d \log(\hat{\boldsymbol{o}}_{\boldsymbol{i}}), \tag{5}$$

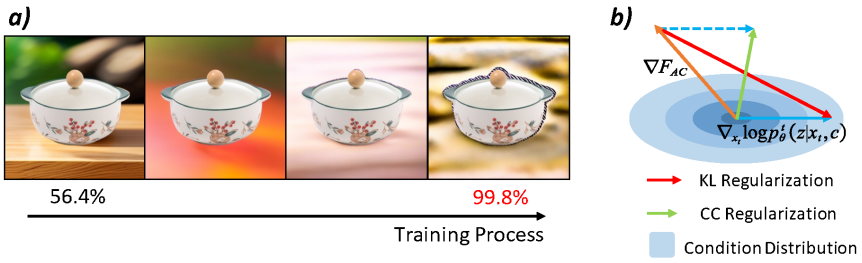

Fig. 5: a) Background destruction happens in direct feedback backpropagation, where a high available rate (99.8%) achievement with collapsed background. **b)** Comparison between KL regularization and our proposed CC regularization.

where $\boldsymbol{y}_d$ is the one-hot vector representing the desired "Available" category, the vector $\hat{\boldsymbol{o}}_{\boldsymbol{i}}$ holds the probability for each generated image case and N is the total number of samples. The gradient ∇F_{AC} is then back-propagated to steer the model towards producing images with a higher probability of being available.

Nevertheless, the training goal of making the model reliable runs counter to the aesthetics of the images generated by the model, *e.g.*, repetitive and simple backgrounds can perfectly circumvent the bad cases we mentioned before. So as illustrated in Fig. 5.a, as training progresses, the model achieves an extremely high available rate, yet yields homogeneous and aesthetically collapsed outputs. So a training approach that keeps the aesthetics of the image stable while increasing its availability rate is what we need. A prevalent solution involves the Kullback-Leibler (KL) regularization [13, 47], a loss term that ensures the modified model does not diverge significantly from the desired distribution, thus maintaining diversity and preventing the convergence to sub-optimal, repetitive results, and this loss term can be formulated as:

$$L_{KL} = \mathrm{KL}(p_\theta(\hat{x}_0^t|x_t, z, c)\|(p_{ref}(\hat{x}_0^t|x_t, z, c)), \tag{6}$$

where c and z are image and text control conditions, p_θ and p_{ref} represent the distributions of current and reference models. However, as the feedback gradient endeavors to steer the image generation towards a higher available rate, the KL regularization strives to maintain the generated image unchangeable. This opposition mirrors the principles of adversarial training [14, 15] and poses a challenge to achieve a win-win solution [18].

Instead of focusing on unchanged images, we aim to maintain visual quality. For text-to-image generation, the visual output is closely linked to the input text condition z [54]. In a classifier-free manner, we derive text guidance from the model's implicit classifier [10, 18, 33] by

$$\nabla_{x_t} \log p_\theta^t(z|x_t, c) \approx -\frac{1}{\sqrt{1-\bar{\alpha}_t}}\Big(\epsilon_\theta(x_t, z, c) - \epsilon_\theta(x_t, c)\Big), \tag{7}$$

which indicates the direction where the text condition influences image generation. To ensure improvements in image availability do not compromise the core

Table 1: Inspection performance of different models.

Model	Precision	Recall	F1	AP
ResNet50	74.87	73.66	74.26	77.29
ResNext50	77.73	76.88	77.30	79.62
HRNet	72.89	73.12	73.01	73.07
ViT	75.59	78.33	76.93	79.31
Ours	**86.45**	**85.23**	**85.83**	**87.58**

Table 2: Ablation study of RFNet.

I_o	I_g	I_d	I_s	Cap	AP
	✓	✓	✓	✓	81.17
✓		✓	✓	✓	82.06
✓	✓		✓	✓	85.31
✓	✓	✓		✓	83.91
✓	✓	✓	✓		84.53
✓	✓	✓	✓	✓	**87.58**
Coarse-grained					82.06

conditions, we introduce a Consistent Condition (CC) regularization term L_{CC}, as follows:

$$L_{CC} = \|\nabla_{x_t} \log p_\theta^t(z|x_t, c) - \nabla_{x_t} \log p_{ref}^t(z|x_t, c)\|_2. \tag{8}$$

Fig. 5.b illustrates the advantage of L_{CC} over L_{KL}. While L_{KL} acts to limit updates from ∇F_{AC} potentially leading to rigidity, L_{CC} offers a win-win approach. It maintains the direction of the condition, allowing for the model to be fine-tuned towards generating more available images. Thus the final feedback to fine-tune the diffusion model in RFFT is:

$$F_{total} = F_{AC} + \beta L_{CC}, \tag{9}$$

where β is a hyper-parameter.

5 Experiments

5.1 Implementation Details

For RFNet, we employ a ResNet50 [16], pre-trained on ImageNet [8], as the image encoder. RoBERTa [9, 28], fine-tuned on Chinese product descriptions, extracts text embeddings from product captions. We resize all images to 384×384 before encoding. The FFM and Self-Attention contain blocks of width 384 with 8 attention heads, and we set N_1 and N_2 to 1 and 3, respectively. The training spans 10 epochs, starting with a learning rate of 1e-4, which is reduced by a factor of 10 at epoch 5.

For diffusion model fine-tuning, we utilize 8 NVIDIA A100 GPUs, a local batch size of 4, and 4 gradient accumulation steps. We opt for AdamW [30] with a learning rate of 1e-5. Unless specified otherwise, MajicmixRealistic_v7 (Maji_v7)[3] serves as the diffusion model, complemented by ControlNet V1.1 [58][4]. During fine-tuning, only ControlNet is trained, while we freeze the remaining parameters. A 40-step DDIM with the last 10 steps chosen for fine-tuning is used [56].

[3] `https://civitai.com/models/43331/majicmix-realistic`
[4] `https://github.com/lllyasviel/ControlNet`

Table 3: Availability evaluation (%) of different approaches using one attempt RG.

Approach	Ava (↑)	Human Ava (↑)
Ori	56.4	70.1
PromptEng	62.9	73.2
PPO	65.9	74.9
DPO	57.3	71.8
ReFL	84.7	84.9
Ours	**85.5**	**86.3**

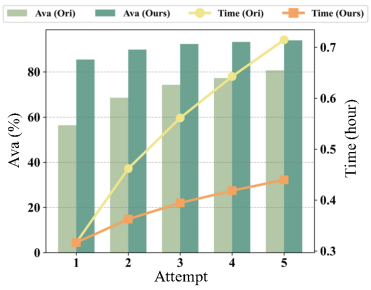

Fig. 6: Available rate (%) and time (hour) for Ori and Ours with different attempts RG.

5.2 Advertising Image Inspection Performance

SOTA Approaches. To validate the superiority of our proposed RFNet, we implement prevailing models ResNet50 [16], ResNeXt50 [55], HRNet [50] and Vision Transformer (ViT) [11] to inspect the generated images.
Evaluation Metrics. Precision, recall, F1 score, and average precision (AP) are used to evaluate the prediction results of different models. We conduct the test on 1,000 images. It is noteworthy that since we are only concerned about the availability of the generated images in Recurrent Generation, we focus on whether the model is able to accurately identify the available images.
Results. Table 1 shows that RFNet outperforms across all metrics, highlighting the benefits of integrating multi-modal information and its effective structure. We conducted extensive experiments to evaluate the impact of various components within RFNet, presented in Table 2. Our experiments demonstrate the significant impact of each component within RFNet on the final AP, especially the crucial role of product images, whose ablation leads to a notable 6.41 drop in AP. We further trained the model using coarse-grained labels (Available/Unavailable), the decrease showcased that a fine-grained label helped to attribute the issues precisely, leading to improved performance.

5.3 Advertising Image Reliability Performance

SOTA Approaches. To evaluate the effectiveness of refining diffusion models, our approach is compared with Ori (using the original model), Prompt Engineering (PromptEng) (using modified prompts)[5], and SOTA RLHF approaches PPO [2, 13, 25, 59], DPO [37, 47, 57], and ReFL [7, 56].
Evaluation Metrics. For availability evaluation, we conduct the test on 1,000 products. The available rate is defined as the ratio of images deemed available by our RFNet to the total number of images inspected, and denoted by "Ava".

[5] Add "irregular shape, extended shape, floating, table legs, pedestal, improper position, improper size, indistinct background" to the negative prompt [1].

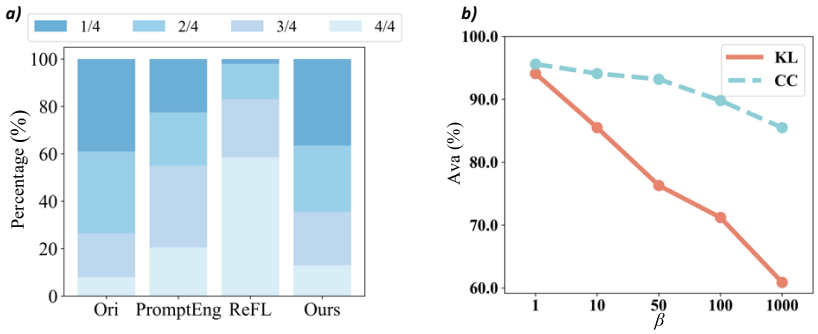

Fig. 7: a) Human rankings of different approaches; **b)** Available rate (%) with different β for L_{KL} (KL) and L_{CC} (CC).

To counter potential biases introduced by the RFNet, we also performed a human inspection which involved experienced advertising annotators assessing the available rate, denoted as "Human Ava". We calculate the total consuming time (hour) using an NVIDIA H800 GPU. For aesthetic assessment, we randomly select 200 available images with identical products, prompts, and generation seed for different approaches. 50 experienced advertising practitioners are employed to rank the images based on their preferences. We count the percentage of different rankings for distinct approaches.

Results. The outcomes highlight our approach's superior performance. Several conclusions could be made:

1. From the result in Table 3, our RFFT could get an extremely high available rate against other approaches. The same trend of "Ava" and "Human Ava" further demonstrates that RFNet reflects human feedback faithfully.
2. As shown in Fig. 6, RG could greatly increase the rate of available images with multiple attempts. Thanks to the inherited strong ability of our model, it needs a shorter production time, shedding light on our RFFL to provide a reliable and efficient solution.
3. The preference assessment in Fig. 7.a demonstrates that RFFT could achieve a relative aesthetic quality compared to the original model, and gain a lot over ReFL owing to the utilization of CC regularization.

These results underscore our approach's enhanced capability to increase the available rate and the producing efficiency while keeping the visual performance stable. Some bad cases solved by our approach are shown in Fig. 8.

5.4 Comparison with KL Regularization

We explore the impact of the hyper-parameter β on both KL regularization L_{KL} and our proposed CC regularization L_{CC}. The results, illustrated in Fig. 7.b, reveal a notable trend: as β increases, the available rate significantly decreases

Fig. 8: Comparison of generated advertising images by Ori and Ours.

under L_{KL}, highlighting its adversarial nature. Conversely, L_{CC} demonstrates resilience to increased β values, maintaining a higher available rate. This contrast underscores the effectiveness of L_{CC} in circumventing adversarial effects.

5.5 Generalization Performance

To assess the flexibility of our RFFT, we examine the generalized capability of the fine-tuned ControlNet when integrated with various LoRAs [19] and diffusion model weights. This is critical as retraining the network for each new LoRA or diffusion model weights combination would be impractical. As depicted in Table 4, experiments demonstrate that the ControlNet, once refined, significantly enhances the available rate across different LoRAs and diffusion model weights, including Maji_v6[6] and SD_v1.5[7]. This generalized ability reduces the need for repetitive training, and underscores broader application of RFFT.

5.6 Integration with Other Feedback

We further incorporate ImageReward [56] whose original goal is to improve aesthetics as an additional feedback, F_{IR}, during the RFFT process. The outcomes, both quantitative and qualitative, are presented in Table 5 and Fig. 9, respectively. The following observations were made:

- The inclusion of F_{IR}, and the combination of F_{IR} and F_{AC}, significantly enhance aesthetic appeal. However, these often result in excessive detail in the backgrounds, leading to obscured product features, which are undesirable for advertising purposes. Moreover, the combination of F_{IR} and F_{AC} results in collapsed textures despite achieving a high available rate.
- Both $F_{AC} + L_{CC}$ and $F_{IR} + F_{AC} + L_{CC}$ configurations achieve high available rates. However, comparing these two, F_{IR} does not significantly affect background aesthetics. This suggests that while F_{IR} attempts to modify the

[6] https://civitai.com/models/43331/majicmix-realistic
[7] https://huggingface.co/runwayml/stable-diffusion-v1-5

Fig. 9: Comparison of different feedback combination strategies.

background condition, L_{CC} aims to maintain it. This side-by-side corroborates that our proposed L_{CC} keeps the background visual quality unchanged, which is particularly important for our RFFL process.

Table 4: Available rate of different diffusion model weights/LoRA.

Approach	Maji_v7		Maji_v6	SD_v1.5
	LoRA_1	LoRA_2		
Ori	56.4	55.0	65.2	68.3
Ours	**85.5**	**79.7**	**84.1**	**84.0**

Table 5: Available rate of different combination strategies with F_{IR}.

F_{IR}	F_{AC}	L_{CC}	Ava
✓			31.5
✓	✓		87.6
	✓	✓	85.5
✓	✓	✓	81.4

6 Conclusion

In conclusion, our study addresses the critical need for reliable advertising image generation in e-commerce, where visual appeal directly influences customer engagement and sales. We introduce Recurrent Generation with the multi-modal RFNet, which faithfully reflects human feedback from our extensive RF1M dataset, establishing a foundation for creating a substantial number of available images for advertising. Our RFFT significantly enhances the generated images available rate, thereby boosting production efficiency. Moreover, our innovative Consistent Condition regularization strikes a balance between a high available rate and aesthetic quality. Ultimately, our work automates creative processes, reduces costs, and improves the customer experience, underscoring the potential of AI-driven tools to revolutionize e-commerce.

References

1. Armandpour, M., Zheng, H., Sadeghian, A., Sadeghian, A., Zhou, M.: Re-imagine the negative prompt algorithm: Transform 2d diffusion into 3d, alleviate janus problem and beyond. arXiv preprint arXiv:2304.04968 (2023)
2. Black, K., Janner, M., Du, Y., Kostrikov, I., Levine, S.: Training diffusion models with reinforcement learning. In: ICLR (2024)
3. Casper, S., Davies, X., Shi, C., Gilbert, T.K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., et al.: Open problems and fundamental limitations of reinforcement learning from human feedback. arXiv preprint arXiv:2307.15217 (2023)
4. Chen, J., Ge, T., Jiang, G., Zhang, Z., Lian, D., Zheng, K.: Efficient optimal selection for composited advertising creatives with tree structure. In: AAAI. vol. 35, pp. 3967–3975 (2021)
5. Chen, J., Xu, J., Jiang, G., Ge, T., Zhang, Z., Lian, D., Zheng, K.: Automated creative optimization for e-commerce advertising. In: WWW. pp. 2304–2313 (2021)
6. Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S., Amodei, D.: Deep reinforcement learning from human preferences. In: NeurIPS. vol. 30 (2017)
7. Clark, K., Vicol, P., Swersky, K., Fleet, D.J.: Directly fine-tuning diffusion models on differentiable rewards. In: ICLR (2024)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255 (2009)
9. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT. pp. 4171–4186 (2019)
10. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. In: NeurIPS. vol. 34, pp. 8780–8794 (2021)
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
12. Fan, Y., Lee, K.: Optimizing DDPM sampling with shortcut fine-tuning. In: ICML. vol. 202, pp. 9623–9639 (2023)
13. Fan, Y., Watkins, O., Du, Y., Liu, H., Ryu, M., Boutilier, C., Abbeel, P., Ghavamzadeh, M., Lee, K., Lee, K.: Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. In: NeurIPS (2023)
14. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., Lempitsky, V.: Domain-adversarial training of neural networks. JMLR **17**(59), 1–35 (2016)
15. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NeurIPS. vol. 27 (2014)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
17. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
18. Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: NeurIPS Workshop (2021)
19. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: ICLR (2022)

20. Ibarz, B., Leike, J., Pohlen, T., Irving, G., Legg, S., Amodei, D.: Reward learning from human preferences and demonstrations in atari. In: NeurIPS. vol. 31 (2018)
21. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR. pp. 1125–1134 (2017)
22. Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., Levy, O.: Pick-a-pic: An open dataset of user preferences for text-to-image generation. In: NeurIPS. vol. 36 (2024)
23. Ku, Y.N., Kuznetsov, M., Mishra, S., de Juan, P.: Staging e-commerce products for online advertising using retrieval assisted image generation. arXiv preprint arXiv:2307.15326 (2023)
24. Lee, K., Liu, H., Ryu, M., Watkins, O., Du, Y., Boutilier, C., Abbeel, P., Ghavamzadeh, M., Gu, S.S.: Aligning text-to-image models using human feedback. arXiv preprint arXiv:2302.12192 (2023)
25. Lee, S.H., Li, Y., Ke, J., Yoo, I., Zhang, H., Yu, J., Wang, Q., Deng, F., Entis, G., He, J., et al.: Parrot: Pareto-optimal multi-reward reinforcement learning framework for text-to-image generation. arXiv preprint arXiv:2401.05675 (2024)
26. Li, Z., Li, F., Feng, W., Zhu, H., Liu, A., Li, Y., Zhang, Z., Lv, J., Zhu, X., Shen, J.J., Lin, Z., Shao, J., Yang, Z.: Planning and rendering: Towards end-to-end product poster generation. arXiv preprint arXiv:2312.08822 (2023)
27. Liang, Y., He, J., Li, G., Li, P., Klimovskiy, A., Carolan, N., Sun, J., Pont-Tuset, J., Young, S., Yang, F., et al.: Rich human feedback for text-to-image generation. arXiv preprint arXiv:2312.10240 (2023)
28. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
29. Losey, D.P., Bajcsy, A., O'Malley, M.K., Dragan, A.D.: Physical interaction as communication: Learning robot objectives online from human corrections. The International Journal of Robotics Research **41**(1), 20–44 (2022)
30. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
31. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Gool, L.V.: Repaint: Inpainting using denoising diffusion probabilistic models. In: CVPR. pp. 11451–11461 (2022)
32. Mishra, S., Verma, M., Zhou, Y., Thadani, K., Wang, W.: Learning to create better ads: Generation and ranking approaches for ad creative refinement. In: CIKM. pp. 2653–2660 (2020)
33. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In: ICML (2021)
34. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. In: NeurIPS. vol. 35, pp. 27730–27744 (2022)
35. Prabhudesai, M., Goyal, A., Pathak, D., Fragkiadaki, K.: Aligning text-to-image diffusion models with reward backpropagation (2023)
36. Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O., Jagersand, M.: U2-net: Going deeper with nested u-structure for salient object detection. Pattern Recognition **106**, 107404 (2020)
37. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. In: NeurIPS (2023)
38. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: ICCV. pp. 12179–12188 (2021)

39. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
40. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., III, W.M.W., Frangi, A.F. (eds.) MICCAI. Lecture Notes in Computer Science, vol. 9351, pp. 234–241 (2015)
41. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. In: NeurIPS. vol. 35, pp. 25278–25294 (2022)
42. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
43. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021)
44. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: ICLR (2021)
45. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
46. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS. vol. 30 (2017)
47. Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., Ermon, S., Xiong, C., Joty, S., Naik, N.: Diffusion model alignment using direct preference optimization. arXiv preprint arXiv:2311.12908 (2023)
48. Wallace, B., Gokul, A., Ermon, S., Naik, N.V.: End-to-end diffusion latent optimization improves classifier guidance. In: ICCV. pp. 7246–7256 (2023)
49. Wang, H., Feng, W., Lu, Y., Li, Y., Zhang, Z., Lv, J., Zhu, X., Shen, J., Lin, Z., Bo, L., et al.: Generate e-commerce product background by integrating category commonality and personalized style. arXiv preprint arXiv:2312.13309 (2023)
50. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., Xiao, B.: Deep high-resolution representation learning for visual recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence **43**, 3349–3364 (2019)
51. Wang, S., Liu, Q., Ge, T., Lian, D., Zhang, Z.: A hybrid bandit model with visual priors for creative ranking in display advertising. In: WWW. pp. 2324–2334 (2021)
52. Wang, Z.J., Montoya, E., Munechika, D., Yang, H., Hoover, B., Chau, D.H.: Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. In: ACL (2023)
53. Wei, P., Liu, S., Yang, X., Wang, L., Zheng, B.: Towards personalized bundle creative generation with contrastive non-autoregressive decoding. In: SIGIR. pp. 2634–2638 (2022)
54. Witteveen, S., Andrews, M.: Investigating prompt engineering in diffusion models. arXiv preprint arXiv:2211.15462 (2022)
55. Xie, S., Girshick, R.B., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: CVPR. pp. 5987–5995 (2016)
56. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: Imagereward: Learning and evaluating human preferences for text-to-image generation. In: NeurIPS (2023)
57. Yang, K., Tao, J., Lyu, J., Ge, C., Chen, J., Li, Q., Shen, W., Zhu, X., Li, X.: Using human feedback to fine-tune diffusion models without any reward model. arXiv preprint arXiv:2311.13231 (2023)

58. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV. pp. 3813–3824 (2023)
59. Zhang, Y., Tzeng, E., Du, Y., Kislyuk, D.: Large-scale reinforcement learning for diffusion models. arXiv preprint arXiv:2401.12244 (2024)
60. Ziegler, D.M., Stiennon, N., Wu, J., Brown, T.B., Radford, A., Amodei, D., Christiano, P., Irving, G.: Fine-tuning language models from human preferences. arXiv preprint arXiv:1909.08593 (2019)