CTR-Driven Advertising Image Generation with Multimodal Large Language Models

Anonymous Author(s)†

Abstract

In web data, advertising images are crucial for capturing user attention and improving advertising effectiveness. Most existing methods generate background for products primarily focus on the aesthetic quality, which may fail to achieve satisfactory online performance. To address this limitation, we explore the use of Multimodal Large Language Models (MLLMs) for generating advertising images by optimizing for Click-Through Rate (CTR) as the primary objective. Firstly, we build targeted pre-training tasks, and leverage a large-scale e-commerce multimodal dataset to equip MLLMs with initial capabilities for advertising image generation tasks. To further improve the CTR of generated images, we propose a novel reward model to fine-tune pre-trained MLLMs through Reinforcement Learning (RL), which can jointly utilize multimodal features and accurately reflect user click preferences. Meanwhile, a product-centric preference optimization strategy is developed to ensure that the generated background content aligns with the product characteristics after fine-tuning, enhancing the overall relevance and effectiveness of the advertising images. Extensive experiments have demonstrated that our method achieves state-of-the-art performance in both online and offline metrics. We will release our code and weights upon acceptance of the paper.

Keywords

CTR-Driven, Advertising Image Generation, Online Advertising, Multimodal Large Language Models

ACM Reference Format:

Anonymous Author(s). 2024. CTR-Driven Advertising Image Generation with Multimodal Large Language Models. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 14 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Advertising images play a pivotal role in attracting user attention and boosting advertising efficacy [18, 30]. Recent advancements in image generation techniques, particularly the integration of Stable Diffusion [36] and ControlNet [49], have enabled the creation of harmonious and realistic backgrounds for product images. However, most existing advertising image generation approaches [9, 42, 50] primarily focus on offline metrics, such as image quality or semantic consistency, without fully considering the critical connection between visual content and online performance metrics like Click-Through Rate (CTR). This results in a notable discrepancy between the generated advertising images and the ideal images that align with actual user preferences.

Inspired by recent approaches [21, 44, 47] that incorporate Reinforcement Learning from Human Feedback (RLHF) [8, 29, 40] to align with human preferences, we can adopt a two-stage method to better capture online user preferences. The first stage involves collecting and analyzing online user feedback to train a Reward Model (RM) that accurately simulates user preferences in the e-commerce domain. In the second stage, we employ Reinforcement Learning (RL) algorithms to fine-tune the generation model, with the RM providing rewards to guide the optimization process. A critical aspect of this pipeline is the RM's ability to accurately reflect users' click preferences for images. However, previous methods that incorporate visual content for CTR prediction face two major limitations: First, applying different backgrounds to the same product can lead to significantly different CTR outcomes, as illustrated in Figure 1 (a). Existing methods [10, 23, 41, 47] often rely on models with limited image understanding capabilities, such as CNNs, vision transformers, or embedding-based methods. To compensate for this deficiency, these methods typically require incorporating numerous auxiliary tasks, such as object detection and OCR, which leads to additional annotation costs and labor-intensive data preparation processes. Second, integrating diverse yet crucial features from multiple modalities (such as product titles and attributes) is of paramount importance, as these significantly influence product CTR. Nevertheless, current methods primarily focus on dense visual features and require additional complex modules to fuse different types of features, potentially limiting the model's adaptability to the rapidly changing online advertising environments. For instance, products from distinct categories, such as water bottles and office chairs illustrated in Figure 1 (a), exhibit remarkably different baseline CTR due to their disparate nature and associated consumer behavior patterns.

To address these issues, leveraging the advanced multimodal understanding and representation capabilities of MLLMs [25, 26, 46] offers a promising solution. On the one hand, these models excel in zero-shot visual analysis, encompassing image representation [15, 27], object detection [22, 48], and various visual tasks without requiring task-specific training. On the other hand, by transforming sparse features (such as categories, tags, or other attributes) into natural language descriptions, MLLMs can process and reason about this textual information alongside visual data, offering a simpler paradigm for integrating multimodal information. While the introduction of MLLMs can effectively guide generation models to produce backgrounds with higher CTR, it is crucial to consider the relationship between the background and the product in advertising image generation. Existing RL algorithms [20, 21, 45] focus solely on optimizing rewards, neglecting the crucial balance between visual appeal and contextual appropriateness. This oversight can result in disharmonious backgrounds that mislead users and lead to poor shopping experiences. As illustrated in Figure 1 (b), while dynamic, sports-oriented backgrounds might boost CTR for athletic shoes, the model might erroneously apply similar backgrounds to unrelated products like cosmetics, compromising visual harmony and product relevance.

Figure 1: (a) Example of the impact of different backgrounds on product CTR. While visual features play a crucial role, other modalities such as textual caption and product attributes also have a significant influence on CTR. (b) Examples of product-background mismatches using existing reinforcement learning algorithms.

In this work, we propose a novel method called **C**TR-driven **A**dvertising **I**mage **G**eneration (CAIG), which leverages the MLLMs as core components to generate advertising images that are both CTR-optimized and coherent with product characteristics. As illustrated in Figure 2, we first design targeted pre-training tasks that utilize a large-scale e-commerce multimodal dataset to equip MLLMs with comprehensive e-commerce domain knowledge for generating advertising images. To further optimize the CTR of generated images, we propose a novel RM that transforms the traditional CTR prediction task into a binary classification problem, enabling the selection of positive and negative samples in subsequent RL process. By focusing on the relative performance between image pairs, our method can effectively mitigate the impact of absolute CTR variations across different product categories. Lastly, to avoid generating background-irrelevant advertisement images, we develop a Product-Centric Preference Optimization (PCPO) strategy. This strategy uses multimodal information of the product as the sole variable and constructs additional preference pairs, forcing the MLLM to generate background content that aligns with the product's characteristics during the RL process. To the best of our knowledge, this is the first work that utilizes MLLMs for CTR-driven advertising image generation.

We summarize our contributions as three-folds:

- We design targeted pre-training tasks using a large-scale e-commerce multimodal dataset to equip MLLMs with comprehensive domain knowledge, providing them with foundational capabilities for downstream tasks.
- We propose a two-branch RM that combines the powerful image understanding capabilities of MLLMs with multimodal product information fusion to effectively simulate human click preferences in e-commerce scenarios.
- We develop a product-centric preference optimization strategy, compelling the model to focus on the product's intrinsic information to generate both visually appealing and contextually consistent advertising images.

Extensive experiments on both public and commercial datasets demonstrate that our method achieves state-of-the-art performance across multiple key metrics, significantly improving online CTRs in real-world e-commerce scenarios.

2 Related Works

2.1 Advertising Image Generation

The primary goal of advertising image generation is to create natural and contextually relevant images while preserving the integrity and identity of the original product. Initially, template-based methods [6, 30, 43, 43] were employed for assembling advertising images, offering high efficiency but lacking personalization and flexibility. With the advent of generative adversarial networks (GANs) [12], researchers began exploring more flexible and automated approaches to advertising image creation. Ku et al. [18] introduced a novel approach of using GAN models as retrieval-assisted techniques for enhancing product images in advertising contexts. More recently, diffusion models have shown promise in producing high-quality, realistic ad images. InsertDiffusion [31] introduced a training-free diffusion architecture that effectively embeds objects into images while preserving their structural and identity features. Recognizing that ad quality involves multiple aspects such as aesthetics and text-image consistency, researchers have begun exploring multi-stage optimization methods [5, 21]. A notable example is VirtualModel [5], which employs a multi-branch structure to enhance the credibility of human-object interactions and ensure consistency in generation quality. Unlike previous methods primarily focusing on visual quality or text-image consistency, our method uniquely leverages MLLMs to generate CTR-optimized contextual descriptions, guiding diffusion models to produce visually appealing and product-specific advertising images.

2.2 Click-Through Rate Prediction

Click-Through Rate (CTR) prediction plays a crucial role in online advertising and recommendation systems, directly impacting user experience and revenue generation. In the context of CTR-driven advertising image generation, precise CTR estimation enables more effective selection and positioning of visual content, thereby enhancing the overall performance of online advertising campaigns.

The advent of deep learning has revolutionized traditional CTR prediction [16, 17, 19], enabling models to automatically learn hierarchical feature representations from raw input data. This paradigm shift not only improved the performance of textual or numerical-based CTR prediction methods [16, 19] but also paved the way for incorporating visual elements into the prediction process. For instance, Wang et al. [41] proposed a hybrid bandit approach that integrates visual priors with a dynamic ranking mechanism, demonstrating the potential of incorporating visual information in CTR prediction models. Recognizing that real-world advertisements are inherently multimodal, comprising text, visuals, and other data types, researchers have begun to explore methods that can effectively integrate these diverse modalities. CG4CTR [47] leveraged a multi-head self-attention module to jointly process textual and visual information from multimodal advertisements, extracting rich features for more accurate CTR estimation. However, these approaches often struggle with complex image understanding tasks and fail to effectively integrate multimodal information. Therefore, it is imperative to explore a more robust CTR estimation method that can seamlessly interpret visual content and harmoniously fuse information from multiple modalities.

2.3 Learning from Human Feedback

Reinforcement Learning from Human Feedback (RLHF) [4, 32, 52] involves collecting human feedback on model outputs. This feedback is then used to optimize the generation model using reinforcement learning algorithms such as PPO[37] or DPO [35]. For example, Lee et al. [20] proposed a three-stage fine-tuning method to improve text-image alignment in text-to-image (T2I) models using human feedback and reward-weighted likelihood maximization. Wu et al.[45] introduced a human preference score derived from a classifier trained on human-curated image choices, which is then utilized to adapt T2I models. Parrot [21] proposed a multi-reward RL approach that jointly optimizes the T2I model and prompt expansion network to improve image quality. However, current preference optimization methods for image generation, while showing promise in text-to-image (T2I) tasks, face significant challenges when applied to scenarios with strict visual requirements, such as advertising background generation. These methods often focus solely on optimizing specific metrics, neglecting the contextual relevance and visual harmony of the generated content. Therefore, our method emphasizes exploring optimization techniques that enable the model to effectively integrate multimodal information to generate diverse and coherent background descriptions that better align with user preferences.

3 Method

3.1 Overview

In this work, we introduce a novel method called **C**TR-Driven **A**dvertising **I**mage **G**eneration (CAIG), designed to generate compelling advertising images that capture user interest, as shown in Figure 2. We first pre-train the MLLM on a large-scale multimodal e-commerce dataset, injecting domain-specific knowledge into the model. This serves as the foundation for our Prompt Model (PM) and Reward Model (RM). Then, we initialize the RM from the pre-trained MLLM and further train it on extensive multimodal online user click data, enabling the RM to simulate human feedback. Finally, we introduce a CTR-driven preference optimization stage, which adopts Product-Centric Preference Optimization (PCPO) as its core strategy, detailed in Algorithm 1. This stage uses the RM's feedback to fine-tune the PM, ultimately generating advertising images that balance attractiveness and relevance.

Algorithm 1 CTR-Driven Preference Optimization

Input: N – Number of training epochs
M – Number of products
PM_Θ – Pre-trained Prompt Model
RM_Θ – Pre-trained Reward Model

for *epoch* $i = 1$ to N **do**
  $P \leftarrow \emptyset$ Initialize set for positive and negative sample pairs
  for *product* $j = 1$ to M **do**
    $(I_o, C) \leftarrow$ Get product image and instruct prompt
    $(y_1, y_2) \leftarrow \text{PM}_\Theta(I_o, C)$ Generate two background descriptions
    $(I_1, I_2) \leftarrow$ Generate advertising images using Stable Diffusion and ControlNet with I_o and (y_1, y_2)
    $(p_1, p_2) \leftarrow \text{RM}_\Theta([I_1; I_2], C)$ Predict relative CTR by RM
    $(y^+, y^-) \leftarrow (y_1, y_2)$ if $p_1 > p_2$ else (y_2, y_1)
    $P \leftarrow P \cup \{(I_o, C, y^+, y^-)\}$ Insert preference pair
  end for
  Update PM_Θ using P with $\mathcal{L}_{\text{DPO}} + \mathcal{L}_{\text{PCPO}}$ (Equation 12)
end for

Output: PM_Θ – Fine-tuned Prompt Model

3.2 E-commerce Knowledge Pre-training

To address the challenge of efficient and scalable advertising creative generation, we leverage the power of MLLMs by injecting domain-specific e-commerce knowledge through pre-training on a large-scale multimodal e-commerce dataset comprising 1.2M samples from major e-commerce platforms, as shown in Figure 2 (a). Specifically, the pre-training tasks involve three main tasks:

(1) **Image Understanding:** Describing the products or backgrounds based on product images.
(2) **Multimodal Content Comprehension:** Describing product background or generating product titles based on multimodal product information (e.g., titles, categories, tags).
(3) **Prompt Generation:** Generating or rewriting description prompts based on multimodal product information.

To facilitate the model's understanding of product information, we design an instruction function that elegantly integrates diverse product attributes into a unified, semantically rich description. Formally, this can be expressed as:

$$C = f_{\text{instruct}}(Q, i_1, i_2, ..., i_n), \tag{1}$$

where C is the instruct prompt constructed by the instruct function $f_{instruct}$ from n individual product attributes $\mathbf{i} = [i_1, i_2, ..., i_n]$ (such as title, category, price, etc.), Q is the task-specific question. For example, an instruct statement for a specific product might be formulated as: *"Generate a suitable product background description based on the following attributes: Product Title: 'Wireless Bluetooth*

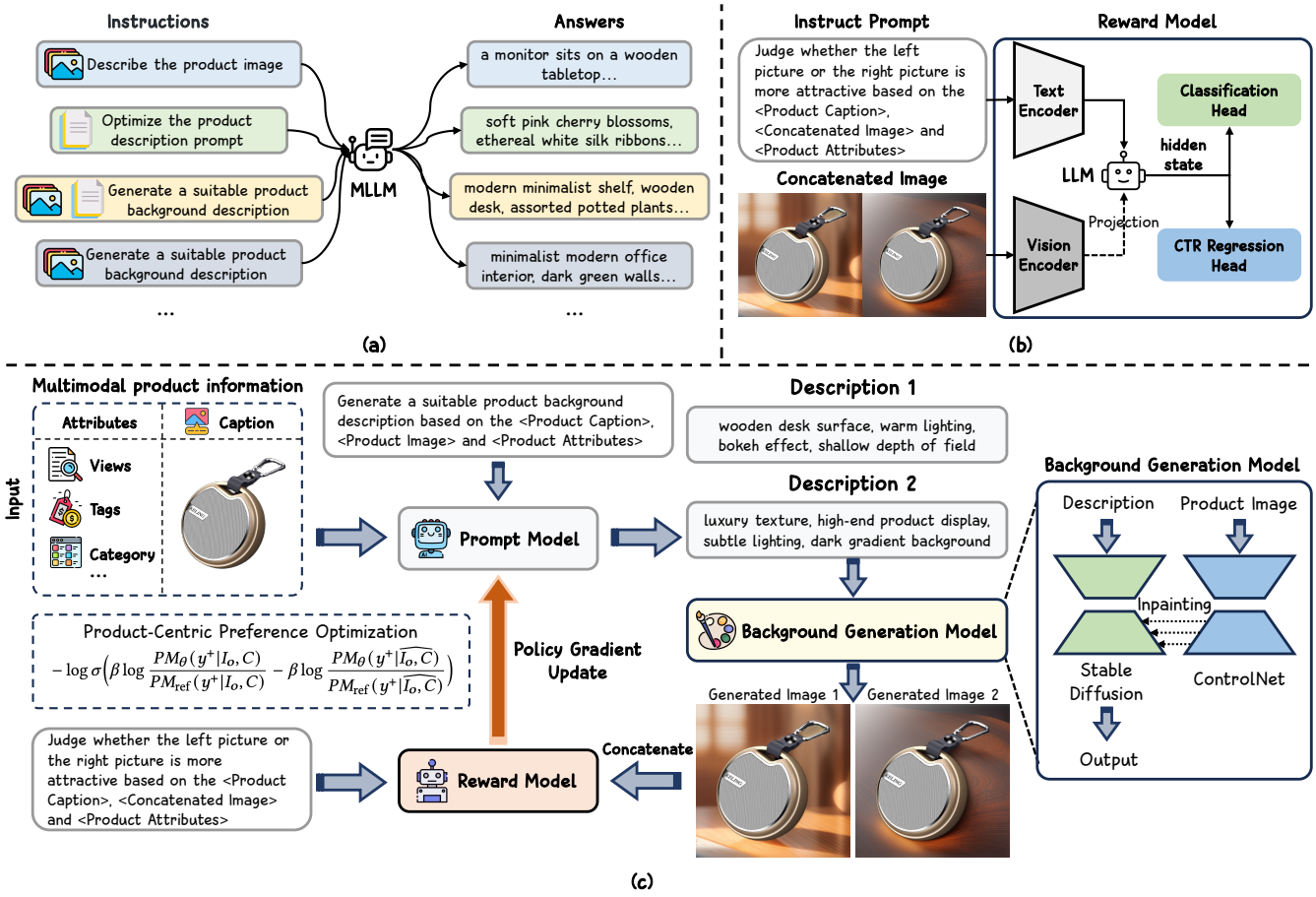


Figure 2: (a) E-commerce knowledge pre-training. The MLLM is pre-trained on a large-scale multimodal e-commerce dataset to incorporate domain-specific knowledge. (b) The Structure of RM. The RM integrates multimodal product features using visual and textual encoders, with dual branches to estimate CTR and identify appealing ad images. (c) CTR-driven preference optimization stage. The PM generates background descriptions for background generation model to create product images with various backgrounds. The RM then estimates the CTR for these images, simulating human feedback to optimize the PM.

Earbuds', Product Category: 'Electronics', Price: '$49.99', Customer Rating: '4.5 stars', Color Options: 'Black, White, Blue'".

By leveraging the power of MLLMs and our specialized pre-training tasks, our MLLM gains a deep understanding of e-commerce products and their attributes. This understanding lays a solid foundation for vision-based CTR prediction and advertising image generation in subsequent tasks, enabling the creation of more relevant and engaging visual content for e-commerce advertising.

3.3 Reward Model based on MLLM

To optimize the alignment between generated advertising images and online user preferences, we leverage user feedback data to train a RM for fine-tuning the advertising image generation pipeline. First, our method utilizes the strong visual representation capabilities and flexible multimodal input of the MLLM which is pre-trained with e-commerce knowledge to extract robust product features. Furthermore, to mitigate the impact of absolute CTR variations across different product categories, we reformulate the CTR regression task into a relative comparison task between pairs of images, as illustrated in Figure 2 (b).

Specifically, we construct pair-wise samples from user click data, where each pair contains two advertising images for the same product with their corresponding CTRs. The images are concatenated and combined with their associated instruct prompt to form a multimodal input. Formally, let $(I_1, I_2, \mathbf{i})$ represent a pair of images I_1 and I_2 with their shared product attributes $\mathbf{i} = [i_1, i_2, ..., i_n]$. The hidden state extraction process is described as:

$$H = \text{LLM}([f_{\text{vision}}([I_1; I_2]); f_{\text{text}}(C_{RM})]), \tag{2}$$

$$C_{RM} = f_{\text{instruct}}(Q_{\text{RM}}, i_1, i_2, ..., i_n). \tag{3}$$

Here, f_{vision} and f_{text} are vision and text encoders respectively. The instruct prompt C_{RM} is built by the the RM-specific question Q_{RM}. The Large Language Model (LLM) generates the hidden state $H \in \mathbb{R}^{d\times l}$, where d is the hidden dimension and l is the sequence length.

To obtain a compact representation of the hidden state, we apply a pooling operation:

$$h = f_{\text{pool}}(H), \tag{4}$$

where $h \in \mathbb{R}^d$ is the globally pooled hidden state.

Subsequently, we transform the CTR regression task into a binary classification problem that directly compares the relative CTR performance between the left and right images in each pair. Thus, a classification head FC_{cls} is employed to map the hidden state h which represents the multimodal feature extracted and pooled by the MLLM to a two-dimensional probability distribution $p \in \mathbb{R}^2$:

$$p = \text{softmax}(FC_{\text{cls}}(h)). \tag{5}$$

To train the RM, we initialize it with pre-trained weights infused with e-commerce domain knowledge and utilize the binary cross-entropy loss function for training. The loss function is defined as:

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{N} [t_i^T \log(p_i)], \tag{6}$$

where N is the number of training samples, $t_i \in \{[1,0],[0,1]\}$ indicates whether the left or right side of the concatenated image has a higher CTR, and p_i is the predicted probability distribution.

Additionally, to enable the model to predict the CTR of the left and right images in a composite image with fine-grained accuracy, we introduce a point-wise loss using a separate CTR regression branch:

$$\mathcal{L}_{\text{Point}} = \frac{1}{N}\sum_{i=1}^{N} ||FC_{\text{ctr}}(h_i) - \hat{t_i})||_2^2, \tag{7}$$

where FC_{ctr} represents the fully connected layer for CTR regression, $FC_{\text{ctr}}(h_i)$ represents the predicted CTR values for the i-th image pair, and $\hat{t_i} \in \mathbb{R}^2$ corresponds to the true CTRs for the left and right images in the pair.

The final loss function for training the RM is a combination of the binary cross-entropy loss and the PointLoss:

$$\mathcal{L}_{\text{reward}} = \lambda_1 \mathcal{L}_{\text{CE}} + \lambda_2 \mathcal{L}_{\text{Point}}, \tag{8}$$

where λ_1 and λ_2 are hyperparameters that balance the contribution of each loss component. This combined design of two components enables the model to learn the relative CTR of comparative advertising images during the training phase while incorporating absolute CTR as an auxiliary input. During the inference stage, we utilize the comparison results from the classification head as the basis for comparing CTR.

3.4 Product-Centric Preference Optimization

We formulate the task of generating higher CTR advertising images as a preference selection problem, encouraging the advertising generation model to choose higher attractive positive images I^+ and reject less attractive negative images I^-. This process involves two key steps: (1) generating image pairs and comparing their CTR with RM, (2) fine-tuning the generation model based on the feedback from the RM, as illustrated in Algorithm 1. For advertising image generation, we utilize the background description y generated by our PM as input to Stable Diffusion [36], along with the original product image I_o. We employ ControlNet [49] and inpainting techniques [28] to seamlessly integrate the product into the generated background. The process uses DDIM [39] as the denoising schedule, where the latent representation x_t at step t is calculated as :

$$x_t = \sqrt{\bar{\alpha}_t}\frac{x_{t+1} - \sqrt{1-\bar{\alpha}_{t+1}}\epsilon_\theta(x_{t+1}, y)}{\sqrt{\bar{\alpha}_{t+1}}} + \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_{t+1}, y) \tag{9}$$

where $\epsilon_\theta(x_{t+1}, y)$ represents the noise predicted by the model [36, 49], and $\bar{\alpha}$ is a set of coefficients controlling the forward noise-adding process. The latent representation x_t is processed by:

$$x_t = (\boldsymbol{I} - \boldsymbol{M}) \otimes x_t + \boldsymbol{M} \otimes x_o, \tag{10}$$

where x_o is the latent of I_o, $\boldsymbol{I}$ represents an identity matrix, $\boldsymbol{M}$ is product mask and $\otimes$ denotes the element-wise multiplication. The final latent x_0 is then converted to the generated image I_g.

Considering that collecting real CTR feedback is time-consuming and resource-intensive, we leverage the RM to distinguish in real-time between more attractive I^+ and less attractive I^- generated images to fine-tune the generation pipeline. Similar to Parrot [21], we empirically find that fine-tuning the background generation model has a much smaller impact on the image content compared to changing the background description. Therefore, to enhance training efficiency, we focus solely on fine-tuning the PM to choose higher attractive background descriptions y^+ and reject less attractive ones y^-. The Direct Preference Optimization (DPO) [35] is then adopted as our fundamental strategy due to its simplicity and efficiency. Specifically, given an optimization policy model PM_θ and a reference model PM_{ref}, the DPO objective is:

$$\mathcal{L}_{\text{DPO}} = -\log\sigma\Big(\beta\log\frac{PM_\theta(y^+|I_o,C)}{PM_{\text{ref}}(y^+|I_o,C)} - \beta\log\frac{PM_\theta(y^-|I_o,C)}{PM_{\text{ref}}(y^-|I_o,C)}\Big), \tag{11}$$

where (I_o, C) represent the original product image and corresponding instruct prompt. σ is the sigmoid activation function, β is a regularization parameter. During the DPO process, the reference model PM_{ref} is frozen to optimize the policy model PM_θ.

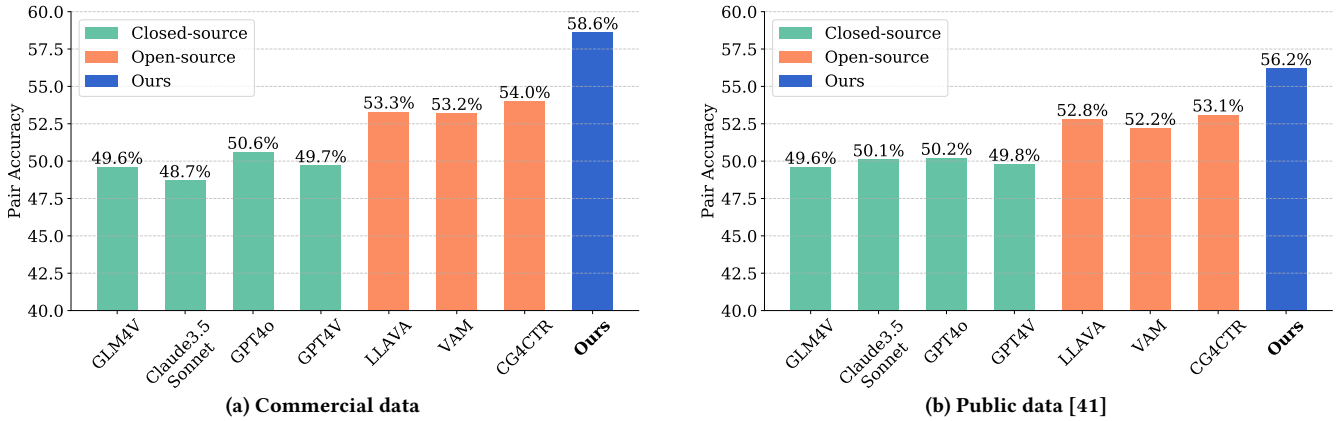
It is worth noting that excessive focus on CTR optimization during DPO training may ignore the product information in preference data, causing a mismatch between the foreground and background in the generated image. Therefore, we introduce the Product-Centric Preference Optimization (PCPO). The core mechanism of PCPO is to control product information as the sole variable during the training process and construct additional preference data pairs, thereby encouraging the model to generate background descriptions that match the product characteristics. Specifically, given a matched (y^+, I_o, C) and a mismatched $(y^+, \widehat{I_o, C})$, the PCPO objective is formulated as:

$$\mathcal{L}_{\text{PCPO}} = -\log\sigma\Big(\beta\log\frac{PM_\theta(y^+|I_o,C)}{PM_{\text{ref}}(y^+|I_o,C)} - \beta\log\frac{PM_\theta(y^+|\widehat{I_o,C})}{PM_{\text{ref}}(y^+|\widehat{I_o,C})}\Big). \tag{12}$$

We consider two strategies to construct a product information $(\widehat{I_o, C})$ that mismatches y^+: (1) visual-aware optimization: randomly masking 75% of input product images. (2) textual-aware optimization: randomly selecting and replacing textual information from other products. These strategies are designed to create hard negative samples that are not fully compatible with y^+ while retaining some common features with the original input. The total objective is a combination of the standard DPO and PCPO:

$$\mathcal{L}_{\text{opt}} = \mathcal{L}_{\text{DPO}} + \mathcal{L}_{\text{PCPO}}. \tag{13}$$

Finally, we utilize the fine-tuned PM to generate background descriptions for products. These descriptions are then fed into the background generation model to create product advertising images suitable for online environments.



(a) Commercial data (b) Public data [41]

Figure 3: Comparison of Pair Accuracy across different methods on commercial and public datasets.

4 Experiments

4.1 Experimental Setup

Datasets: For training and validating our RM, we conduct experiments on both public and commercial datasets. The public dataset [41] covers 500K product samples with 1.2M unique advertising images. The CTR data for public dataset is collected over an average period of 10 days for each creative placement. Our commercial dataset, collected from a well-known e-commerce platform, contains 1M product samples with 3.4M unique advertising images. For commercial dataset, the CTR data is collected over a one-month period. It is worth noting that our commercial dataset contains more detailed product information, including titles, categories, tags, and other relevant attributes. To ensure the quality and reliability of our training and test data, we apply specific criteria to both datasets. To improve the confidence of CTR estimates, we require each image to have a minimum exposure threshold (E). Additionally, to ensure distinguishable CTR differences within pairs, we require the relative CTR difference between paired images to exceed a certain threshold (D). For the training set, we set E = 50 and D = 1%, while for the test set, we apply more stringent criteria with E = 1,000 and D = 5%. After applying these preprocessing steps, the public dataset yields 890K training pairs and 1,034 test pairs, while our commercial dataset contains 1.15M training pairs and 1,528 test pairs. In the CTR-driven preference optimization stage, we fine-tune our generation pipeline using a dataset of 32K samples, which includes original product images along with multimodal information. These product samples are also collected from the same e-commerce platform, ensuring consistency in data source and characteristics.

Models: We employ the *LLaVA-v1.6-7B* [26] as our foundation MLLMs, which utilizes *Vicuna-7B* [7] as the text encoder and *CLIP-ViT-L/14-336* [34] as the vision encoder. For our background generation model, we use *MajicmixRealistic-v7*[1] as the base model, enhanced with *ControlNet-v1.1* [49][2] to provide finer control over the generated images.

Implementation Details: For the pre-training task of MLLMs, we perform full model fine-tuning. We optimize the learning process over 10 epochs using a cosine learning rate scheduler with an initial learning rate of 2e-6. The pre-training task takes approximately 5 days to complete. We then initialize the RM with the pre-trained weights and employ the same learning strategy to train on massive user click data, simulating user feedback. The hyperparameters λ_1 and λ_2 are set to 1 and 0.5, respectively. Finally, in the CTR-driven preference optimization stage, we utilize the frozen RM to drive the proposed advertising image generation pipeline. We employ LoRA [14] fine-tuning with a learning rate of 2e-5. This phase consists of 5 epochs and takes about 20 hours to complete. All experiments are conducted on a machine equipped with 8 NVIDIA A100 GPUs.

4.2 Analysis on Reward Model

4.2.1 Evaluation Metric. To evaluate the performance of our RM, we introduce the Pair Accuracy metric, defined as:

$$\text{Pair Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}(\text{argmax}(p_i) == y_i), \tag{14}$$

where N is the total number of image pairs, p_i is the predicted probability distribution for the i-th pair, y_i is the ground truth label, and $\mathbb{1}$ is the indicator function. It is worth noting that the task of CTR comparison is highly challenging, where even small improvements can lead to significant economic benefits.

4.2.2 Comparison with State-of-the-Art Methods. We conduct extensive experiments on both commercial and public datasets, comparing our method with various state-of-the-art open-source and closed-source models based on MLLMs, as shown in Figure 3. The open-source models are fine-tuned on the corresponding datasets to ensure a fair comparison. For closed-source models, we provide them with the same instructions and image pairs as our RM, then convert their textual responses into predicted labels. From the results, we can observe that existing closed-source models (GLM4V [11], Claude3.5 Sonnet [2], GPT4o [3], and GPT4V [1]) lack the ability to effectively compare the CTR of advertising images, as evidenced by their near-random performance (around 50% Pair Accuracy). This suggests that these models, despite their general capabilities, are not specifically tuned for CTR regression tasks in advertising contexts. Open-source models like VAM [41] and

[1] https://civitai.com/models/43331/majicmix-realistic
[2] https://github.com/lllyasviel/ControlNet

E-commerce pre-training	Classification head	Product caption	Additional information	Pointloss	Pair Accuracy (%)
✗	✗	✗	✗	✗	53.3
✓	✗	✗	✗	✗	54.4 (+1.1%)
✓	✓	✗	✗	✗	56.4 (+2.0%)
✓	✓	✓	✗	✗	57.3 (+0.9%)
✓	✓	✓	✓	✗	58.2 (+0.9%)
✓	✓	✓	✓	✓	58.6 (+0.4%)

Table 1: Ablation study for the reward model.

CG4CTR [47], while showing slight improvements, still demonstrate limited performance due to their weak visual representation capabilities and inability to effectively integrate multimodal information. In contrast, our proposed method, which leverages MLLM, achieves state-of-the-art performance on both commercial and public datasets. By effectively combining visual and textual modalities of product information, our method demonstrates superior ability in predicting relative CTR performance between image pairs. Specifically, our method achieves a Pair Accuracy of 58.6% on commercial data and 56.2% on public data, significantly outperforming all baseline models.

4.2.3 Ablation Study. To further analyze the contribution of each component in our proposed RM, we conduct a detailed ablation study on the commercial dataset, with results shown in Table 1. We start with the base *LLaVA-v1.6-Vicuna-7B* [26] and progressively add key components to observe their impact on model performance. First, we observe that incorporating a pre-training step with e-commerce domain knowledge increases the Pair Accuracy from 53.3% to 54.4%. This suggests that domain-specific pre-training provides a good starting point for the model, enhancing its baseline understanding of e-commerce concepts and product characteristics. A more substantial improvement is seen when replacing the original output layer with a dedicated classification head, which boosts the Pair Accuracy to 56.4%. This notable increase can be attributed to the classification head's ability to enable the model to learn explicit classification boundaries, thereby reducing the ambiguity often associated with natural language outputs in the original model architecture. Incorporating product captions and additional product information further enhances the model's accuracy to 58.2%. This demonstrates the importance of supplementary product data in improving the model's ability to compare advertising image attractiveness. Finally, we add an extra CTR regression branch and introduce the point loss, which improve the final performance to 58.6%. This enhancement demonstrates that by directly incorporating CTR values into the training objective, the model can more accurately capture subtle differences in CTR, thereby further improving prediction accuracy.

4.3 Analysis of Product-Background Matching

4.3.1 Evaluation Metric. Existing preference optimization methods focus solely on optimizing rewards, which may neglect the crucial balance between visual appeal and contextual appropriateness. To quantify the impact of different optimization methods on the compatibility between foreground products and generated backgrounds, we introduce the match rate metric. We calculate the match rate by randomly selecting 1,000 products and generating backgrounds for each using the generation models under evaluation.

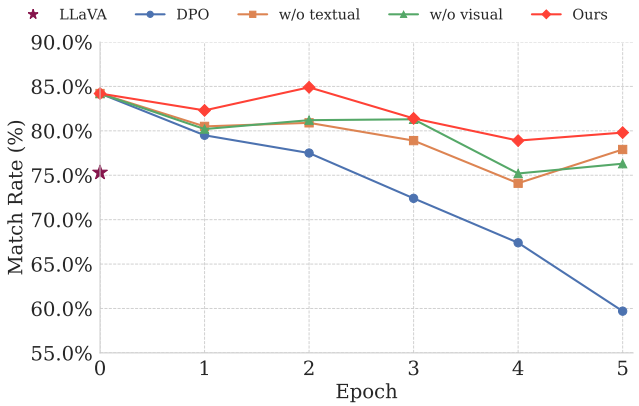


Figure 4: Comparison of Match Rate across different preference optimization strategies over training epochs.

Experienced advertising professionals then assess whether the foreground and background are compatible based on comprehensive product information, considering factors such as style consistency, color harmony, and contextual appropriateness. Detailed annotation guidelines and criteria are provided in Appendix A.2.

4.3.2 Comparison with Standard DPO. To ensure a fair comparison, we evaluate PCPO against standard DPO during preference optimization, using identical RM for CTR feedback and equal training epochs. Figure 4 illustrates the performance of both methods over training epochs. Notably, the standard DPO experiences a significant drop in match rate, declining from 0.842 to 0.597 after 5 epochs of training. In contrast, our PCPO demonstrates a more gradual decline in match rate, maintaining a higher value of 0.798 at the 5th epoch, which represents a 33.7% relative improvement over DPO at the same stage of training. Additionally, we showcase several examples in Figure 5 where the standard DPO produces images with mismatched foreground and background elements, highlighting the effectiveness of our PCPO in preserving product-context coherence throughout the optimization process.

4.3.3 Effectiveness of E-commerce Knowledge Pre-training. As shown in Figure 4, we compare the performance of our pre-trained model with the original LLaVA model without pre-training (indicated by asterisks). The results demonstrate a significant improvement in match rate after injecting e-commerce knowledge through pre-training, with our model achieving a score of 0.842 compared to LLaVA's 0.753. This performance gap highlights that our pre-training strategy provides a strong initialization point for the subsequent preference optimization process, underscoring the importance of domain-specific knowledge in MLLMs.

4.3.4 Ablation Studies. To further validate the effectiveness of our method, we conduct ablation studies on two key components of PCPO: PCPO without textual-aware optimization (w/o textual) and PCPO without visual-aware optimization (w/o visual), as illustrated in Figure 4. Both ablation variants show improvements over standard DPO but fall short of the full PCPO method. The "w/o textual" and "w/o visual" variants highlight the importance of both textual and visual components in our method. These results emphasize

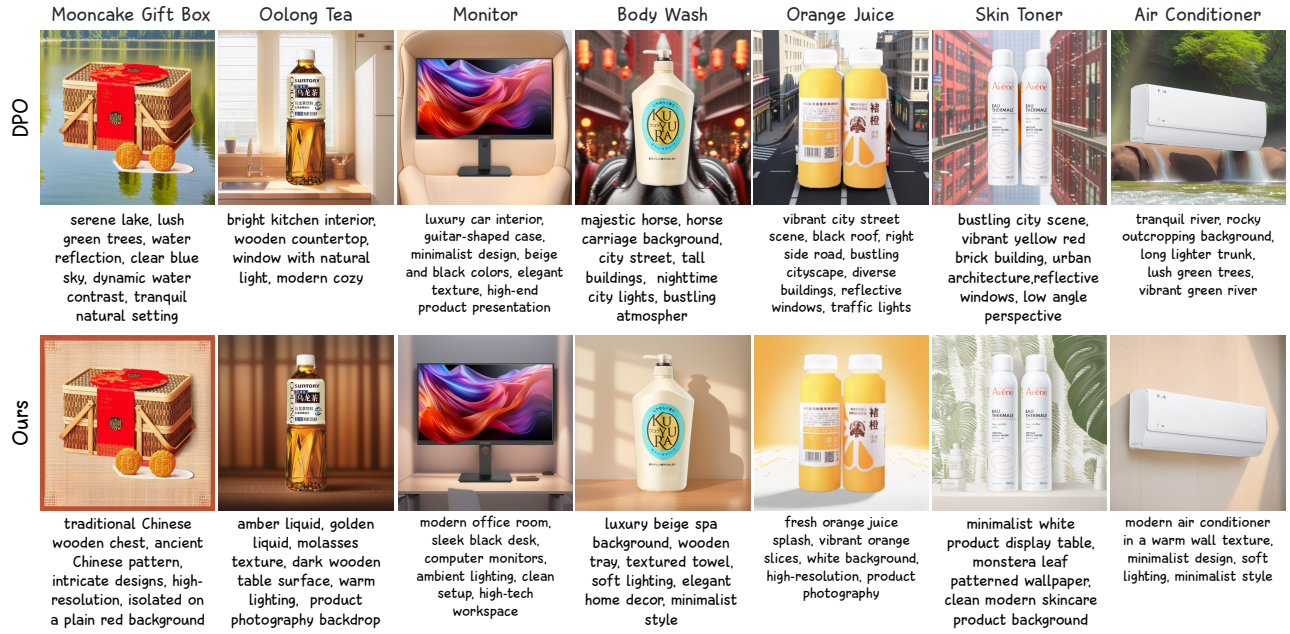


Figure 5: Comparison between DPO and the proposed PCPO. The first line shows the name of the product, followed by the generated results for each method, including the generated image and corresponding background prompt.

Methods	All	Beauty	Fashion	Home Appliances	Digital	Computers
VAM [41]	3.9	5.1	3.8	4.2	2.3	4.5
CG4CTR [47]	3.7	4.8	3.2	4.9	2.1	3.7
DPO [35]	5.7	3.7	3.2	6.2	**5.8**	4.7
Ours	**7.4**	**9.5**	**7.6**	**7.8**	3.2	**5.4**

Table 2: Online CTR improvement compared with the baseline of using the pre-trained MLLM with percentage.

that controlling the textual or visual modality input as the sole variable and constructing less relevant product information as negative samples during training can effectively prevent the model from generating contextually mismatched advertisement images. This strategy enhances the model's focus on the multimodal information of the products themselves, leading to more accurate and relevant product descriptions. The full PCPO strategy, which combines both textual and visual perturbations, is the most effective in optimizing the model's performance for product-centric tasks.

4.4 Online Results

To validate the effectiveness of our proposed CAIG in enhancing the CTR of generated advertising images, we conduct a one-week online experiment in a well-known e-commerce platform. We use different methods to generate two images for each product in 44 categories, which almost cover all common products, greatly exceeding the previous method [47] scope of only five categories. It is worth noting that to enhance user experience, we engage professional advertising practitioners to ensure that the images displayed online are front and background-matched. This experiment accumulates over 10 million impressions to validate the reliability and statistical significance of the CTR results. We use a multi-armed bandit based model as online display strategy.

We report the results of different methods in all categories and five common categories in Table 2, where the improvement of CTR is compared to directly using pre-trained MLLM. To demonstrate the superiority of our RM, we use different RMs during the CTR-drive preference optimization phase. Our RM outperforms previous methods [41, 47] in all categories and five common categories, demonstrating that more accurate CTR prediction can drive the generative model to produce images with higher CTR. We also compare using only DPO [35] as the optimization algorithm, and the results show that using our PCPO can enable the generated model to focus on product characteristics, resulting in an increase in CTR. We further conduct an online A/B test to verify the attractiveness of our generated images, and the results show that adding these images improves 2% in CTR with over 60 million impressions.

5 Conclusion

In this paper, we present an innovative **C**TR-Driven **A**dvertising **I**mage **G**eneration (CAIG) method, leveraging the powerful capabilities of Multimodal Large Language Models (MLLMs) to successfully address the limitations in optimizing online performance metrics. Our comprehensive framework, comprising targeted pre-training tasks, an MLLM-based two-branch reward model, and a product-centric preference optimization strategy, enables the generation of visually appealing and product-relevant advertising images. Extensive experiments demonstrate that CAIG achieves state-of-the-art performance in both online and offline metrics, significantly improving CTR in real-world e-commerce scenarios. This work not only advances the field of advertising image generation but also opens up new possibilities for applying MLLMs to complex multimodal tasks in e-commerce and digital advertising, laying a solid foundation for future research in this domain.

References

[1] 2023. GPT-4V(ision) System Card. https://openai.com/index/gpt-4v-system-card/

[2] 2024. Claude3.5-sonnet. https://www.anthropic.com/news/claude-3-5-sonnet

[3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[4] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862* (2022).

[5] Binghui Chen, Chongyang Zhong, Wangmeng Xiang, Yifeng Geng, and Xuansong Xie. 2024. VirtualModel: Generating Object-ID-retentive Human-object Interaction Image by Diffusion Model for E-commerce Marketing. *arXiv preprint arXiv:2405.09985* (2024).

[6] J Chen, J Xu, G Jiang, T Ge, Z Zhang, D Lian, and K Zheng. [n. d.]. Automated Creative Optimization for E-Commerce Advertising. arXiv 2021. *arXiv preprint arXiv:2103.00436* ([n. d.]).

[7] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality. https://lmsys.org/blog/2023-03-30-vicuna/

[8] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems* 30 (2017).

[9] Zhenbang Du, Wei Feng, Haohan Wang, Yaoyu Li, Jingsen Wang, Jian Li, Zheng Zhang, Jingjing Lv, Xin Zhu, Junsheng Jin, et al. 2024. Towards Reliable Advertising Image Generation Using Human Feedback. *arXiv preprint arXiv:2408.00418* (2024).

[10] Tiezheng Ge, Liqin Zhao, Guorui Zhou, Keyu Chen, Shuying Liu, Huimin Yi, Zelin Hu, Bochao Liu, Peng Sun, Haoyu Liu, et al. 2018. Image matters: Visually modeling user behaviors using advanced model server. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 2087–2095.

[11] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, et al. 2024. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. *arXiv preprint arXiv:2406.12793* (2024).

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM* 63, 11 (2020), 139–144.

[13] Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. 2024. Optimizing prompts for text-to-image generation. *Advances in Neural Information Processing Systems* 36 (2024).

[14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).

[15] Jitesh Jain, Jianwei Yang, and Humphrey Shi. 2024. Vcoder: Versatile vision encoders for multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 27992–28002.

[16] Chen Jie-Hao, Li Xue-Yi, Zhao Zi-Qian, Shi Ji-Yun, and Zhang Qiu-Hong. 2017. A CTR prediction method based on feature engineering and online learning. In *2017 17th International Symposium on Communications and Information Technologies (ISCIT)*. IEEE, 1–6.

[17] Yuchin Juan, Yong Zhuang, Wei-Sheng Chin, and Chih-Jen Lin. 2016. Field-aware factorization machines for CTR prediction. In *Proceedings of the 10th ACM conference on recommender systems*. 43–50.

[18] Yueh-Ning Ku, Mikhail Kuznetsov, Shaunak Mishra, and Paloma de Juan. 2023. Staging e-commerce products for online advertising using retrieval assisted image generation. *arXiv preprint arXiv:2307.15326* (2023).

[19] Rohit Kumar, Sneha Manjunath Naik, Vani D Naik, Smita Shiralli, VG Sunil, and Moula Husain. 2015. Predicting clicks: CTR estimation of advertisements using logistic regression classifier. In *2015 IEEE international advance computing conference (IACC)*. IEEE, 1134–1138.

[20] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. 2023. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192* (2023).

[21] Seung Hyun Lee, Yinxiao Li, Junjie Ke, Innfarn Yoo, Han Zhang, Jiahui Yu, Qifei Wang, Fei Deng, Glenn Entis, Junfeng He, et al. 2024. Parrot: Pareto-optimal multi-reward reinforcement learning framework for text-to-image generation. *arXiv preprint arXiv:2401.05675* (2024).

[22] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. LLaVA-OneVision: Easy Visual Task Transfer. *arXiv preprint arXiv:2408.03326* (2024).

[23] Kaiyi Lin, Xiang Zhang, Feng Li, Pengjie Wang, Qingqing Long, Hongbo Deng, Jian Xu, and Bo Zheng. 2022. Joint Optimization of Ad Ranking and Creative Selection. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2341–2346.

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 740–755.

[25] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. https://llava-vl.github.io/blog/2024-01-30-llava-next/

[26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems* 36 (2024).

[27] Yanqing Liu, Kai Wang, Wenqi Shao, Ping Luo, Yu Qiao, Mike Zheng Shou, Kaipeng Zhang, and Yang You. 2023. Mllms-augmented visual-language representation learning. *arXiv preprint arXiv:2311.18765* (2023).

[28] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11461–11471.

[29] James MacGlashan, Mark K Ho, Robert Loftin, Bei Peng, Guan Wang, David L Roberts, Matthew E Taylor, and Michael L Littman. 2017. Interactive learning from policy-dependent human feedback. In *International conference on machine learning*. PMLR, 2285–2294.

[30] Shaunak Mishra, Manisha Verma, Yichao Zhou, Kapil Thadani, and Wei Wang. 2020. Learning to create better ads: Generation and ranking approaches for ad creative refinement. In *Proceedings of the 29th ACM international conference on information & knowledge management*. 2653–2660.

[31] Phillip Mueller, Jannik Wiese, Ioan Craciun, and Lars Mikelsons. 2024. InsertDiffusion: Identity Preserving Visualization of Objects through a Training-Free Diffusion Architecture. *arXiv preprint arXiv:2407.10592* (2024).

[32] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.

[33] Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. 2024. Personalizing Reinforcement Learning from Human Feedback with Variational Preference Learning. *arXiv preprint arXiv:2408.10075* (2024).

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[35] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems* 36 (2024).

[36] 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.

[37] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[38] Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. 2023. Distributional preference learning: Understanding and accounting for hidden context in RLHF. *arXiv preprint arXiv:2312.08358* (2023).

[39] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).

[40] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems* 33 (2020), 3008–3021.

[41] Shiyao Wang, Qi Liu, Tiezheng Ge, Defu Lian, and Zhiqiang Zhang. 2021. A hybrid bandit model with visual priors for creative ranking in display advertising. In *Proceedings of the web conference 2021*. 2324–2334.

[42] Shiyao Wang, Qi Liu, Yicheng Zhong, Zhilong Zhou, Tiezheng Ge, Defu Lian, and Yuning Jiang. 2022. CreaGAN: An Automatic Creative Generation Framework for Display Advertising. In *Proceedings of the 30th ACM International Conference on Multimedia*. 7261–7269.

[43] Penghui Wei, Shaoguo Liu, Xuanhua Yang, Liang Wang, and Bo Zheng. 2022. Towards personalized bundle creative generation with contrastive non-autoregressive decoding. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2634–2638.

[44] Penghui Wei, Xuanhua Yang, Shaoguo Liu, Liang Wang, and Bo Zheng. 2022. CREATER: CTR-driven advertising text generation with controlled pre-training and contrastive fine-tuning. *arXiv preprint arXiv:2205.08943* (2022).

[45] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. 2023. Better aligning text-to-image models with human preference. *arXiv preprint arXiv:2303.14420* 1, 3 (2023).

[46] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671* (2024).

[47] Hao Yang, Jianxin Yuan, Shuai Yang, Linhe Xu, Shuo Yuan, and Yifan Zeng. 2024. A New Creative Generation Pipeline for Click-Through Rate with Stable Diffusion Model. In *Companion Proceedings of the ACM on Web Conference 2024*. 180–189.

[48] Yuhang Zang, Wei Li, Jun Han, Kaiyang Zhou, and Chen Change Loy. 2024. Contextual object detection with multimodal large language models. *International Journal of Computer Vision* (2024), 1–19.

[49] 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.

[50] Kang Zhao, Xinyu Zhao, Zhipeng Jin, Yi Yang, Wen Tao, Cong Han, Shuanglong Li, and Lin Liu. 2024. Enhancing Baidu Multimodal Advertisement with Chinese Text-to-Image Generation via Bilingual Alignment and Caption Synthesis. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2855–2859.

[51] Zhanhui Zhou, Jie Liu, Chao Yang, Jing Shao, Yu Liu, Xiangyu Yue, Wanli Ouyang, and Yu Qiao. 2023. Beyond one-preference-for-all: Multi-objective direct preference optimization. *arXiv preprint arXiv:2310.03708* (2023).

[52] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593* (2019).

A Appendices

This supplementary material provides:

(1) **Section A.1.** Visualization and analysis of the multi-task pre-training effectiveness.
(2) **Section A.2.** Detailed annotation guidelines and criteria used for evaluating generated images.
(3) **Section A.3.** Extensive visual examples demonstrating the capabilities of CAIG across various product categories.
(4) **Section A.4.** The composition of pre-training tasks and design of instruction sets.
(5) **Section A.5.** Discussion on current limitations of the method and proposed directions for future research.
(6) **Section A.6.** Analysis of potential social impacts, ethical considerations, and safeguards implemented.

A.1 Visualization of Pre-training Model

To validate the effectiveness of our proposed e-commerce knowledge injection pre-training method, we directly utilize the model pre-trained at this stage as a prompt model to generate a set of product images, as illustrated in Figure 6. The generated images demonstrate the model's ability to capture key visual attributes and styles commonly found in e-commerce product photography, such as visual focus with the product as the main subject. This visual evidence suggests that our pre-training method successfully incorporated domain-specific knowledge, resulting in a model capable of generating contextually relevant and visually coherent product images. Furthermore, the quality and diversity of the generated images indicate that the pre-trained MLLM provides a well-initialized distribution space for the subsequent preference optimization phase based on the CTR objective.

A.2 Annotation Guidance

During the match rate evaluation stage of the generation process, annotators are provided with the original product image, product title, and the generated image, along with the following strict guidelines regarding mismatches:

(1) **Scale Mismatch.** Images where the relative size of the product and background elements are disproportionate, such as a washing machine next to an oversized laundry detergent bottle.
(2) **Scene Mismatch.** Images where the product is placed in a setting that contradicts its intended use or cultural context, such as winter coats displayed in a tropical beach scene.
(3) **Color Mismatch.** Images exhibiting stark color conflicts between the product and background, creating visual discomfort or detracting from the product's appeal.
(4) **Available.** Images deemed suitable for advertising purposes, not falling into any of the aforementioned categories.

Additionally, Figure 7 illustrates some examples identified by the annotators.

A.3 More Visual Examples

As illustrated in Figure 8, we present an extensive array of additional examples showcasing our proposed CAIG method. These diverse visual results demonstrate the remarkable versatility and effectiveness of our method across a wide spectrum of product categories. From electronics to fashion items, and from household goods to specialty products, our method consistently generates varied and contextually appropriate backgrounds. This comprehensive set of examples not only highlights the robustness of CAIG in handling diverse product types but also underscores its ability to create visually appealing and relevant contextual environments.

A.4 Pre-training Tasks and Instruction Set

Our pre-training method for high-quality MLLMs in advertising background generation encompasses diverse tasks and instruction sets, as illustrated in Table 3. We utilize a mix of public and proprietary datasets: <Product Images> and <Product Caption> from our e-commerce knowledge pre-training dataset, <Prompt> from both Promptist [13] and our dataset, and COCO Caption [24] for unconstrained background description generation. All target outputs are generated by GPT4V [1] and subsequently reviewed by experienced annotators to ensure quality and relevance. Additionally, we design diverse instruction sets for the PM and RM, as shown in Table 4. These guide the models in generating and evaluating advertising backgrounds from various perspectives, leveraging multimodal product information. The PM set contains 8 distinct prompts for background creation, while the RM set includes 13 distinct prompts for the CTR comparison task.

A.5 Limitations and Future Work

A key limitation of this work is that our CTR optimization is based on aggregated data from all users, which may overlook the preferences of minority user groups or niche market segments. This lack of personalization could result in suboptimal experiences for diverse user segments. In future work, we plan to explore personalized RLHF [33, 38, 51] to better capture and integrate individual user preferences. By doing so, we aim to develop more inclusive and tailored advertising strategies that cater to a wider range of user needs and behaviors.

A.6 Social Impact

Regarding image processing and automatic advertisement image generation, there are risks of producing unethical or illegal content, such as infringing on personal portrait rights or creating discriminatory content. Therefore, these technologies require stricter regulation. During the generation process, we use Stable Diffusion's official safety checker to filter out inappropriate content. We also ensure that the generated images do not contain portraits or other elements that may infringe on privacy. Finally, professionals review and screen the generated images to ensure they are free from bias or offensive content and comply with relevant laws. To ensure the ethical use of AI in advertising, we maintain transparency by clearly labeling AI-generated images and adhering to established commercial and ethical guidelines. The automation of creative tasks may alter the job market in the creative industry. We should view AI as an auxiliary tool for creative professionals rather than a replacement to maintain the important role of human creativity in advertising.

Figure 6: Advertising images generated by directly using the e-commerce knowledge-injected MLLM as PM. For each product, we display the original transparent background product image in the first column, along with three different background images generated through random repetition.

Figure 7: Some match and mismatch examples identified by annotators.

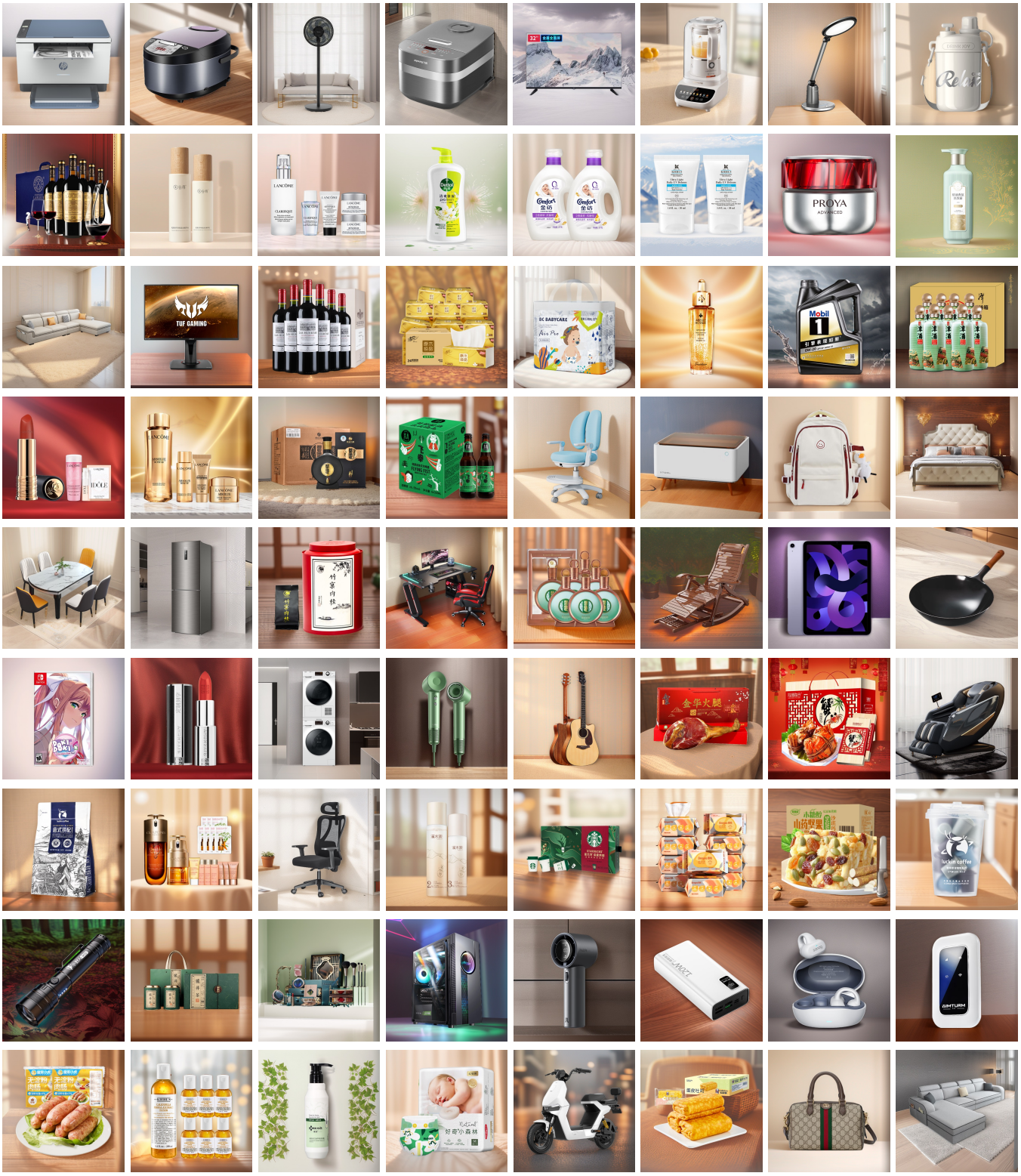


Figure 8: Extensive visual examples of our CAIG method applied to diverse product categories.

Pre-training Task	Input	Target	Nums
Image Understanding[3]	<Product Images>	<Image Description>	5w
	<Product Images>	<Product Background Description>	5w
Multimodal Content Understanding[3]	<Product Images><Product Caption>	<Image Description>	5w
	<Product Images><Product Caption>	<Product Background Description>	15w
Prompt Generation[1,2,3]	<None>	<Prompt>	5w
	<Prompt>	<Optimized prompt>	5w
	<Product Images>	<Prompt>	10w
	<Product Caption>	<Prompt>	10w
	<Product Images><Product Caption>	<Prompt>	61w
Total			121w

Table 3: Overview of pre-training tasks, input-target pairs, and data volume for our multimodal model. The tasks include image understanding, multimodal content understanding, and prompt generation, utilizing both public datasets COCO Caption[1] [24], Promptist[2] [13] and our e-commerce knowledge pre-training dataset[3].

Model	Instructions
Prompt Model	*Design a concise Stable Diffusion prompt that takes the product caption '{}' and product image as inspiration to generate an appealing advertising image background for this product.*
	Produce a short diffusion prompt considering the critical information in product caption '{}' and product image to create an advertisement background.
	Generate a short Stable Diffusion prompt that leverage the product caption '{}' and the visual elements of the product image to output an advertising background that underscores the product's attractiveness.
	Provide a succinct text to background diffusion model prompt suitable for this product according to caption '{}' and product image.
	Develop a compact prompt for Stable Diffusion to craft a background for an ad, using the product caption '{}' and the accompanying product image as creative influences.
	Draft a distilled text to image prompt to fabricate an ad background, infusing the essence of '{}' from the product caption and the picture to accentuate the product's features.
	Based on this product image and product caption '{}', formulate a brief diffusion prompt to synthesize a background tailored for advertising purposes.
	Make use of product title '{}' along with its image, assemble a terse stable diffusion model prompt to render an ad background that complements and highlights the product.
Reward Model	*Comparing the left part and right part of this image, which part is more suitable for the product '{}'?*
	The left and right part of this image is one advertising image for the product '{}', respectively, which is preferred by the user?
	Which part will bring more click-through rate in this image for product '{}'?
	Between the left and right sections of this image, which one is more appropriate for showcasing the product '{}'?
	Considering the left and right halves of this image, which side better represents the product '{}'?
	Which side of this image, left or right, is more effective for advertising the product '{}'?
	For the product '{}', which part of the image, left or right, is more appealing?
	When looking at the left and right portions of this image, which part is more suitable for promoting the product '{}'?
	In this image, which side, left or right, is preferred by users for the product '{}'?
	Which half of this image, left or right, is more likely to attract user preference for the product '{}'?
	Which section of this image, left or right, is expected to generate a higher click-through rate for the product '{}'?
	For the product '{}', which side of the image do users find more engaging, left or right?
	Between the left and right sides of this image, which one is anticipated to drive more clicks for the product '{}'?

Table 4: Instruct directives for Prompt and Reward Models.