ADSPROMPTER: BRIDGING USER AND DESIGNER PREFERENCES FOR MULTIMODAL ADVERTISING DESIGN GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative image modeling has shown remarkable advances recently, yet generating advertising designs with multimodal elements remains an enduring challenge. Existing works optimize either design aesthetics according to design criteria or user behaviors measured by click-through rates, overlooking the inherent discrepancy between user and designer preferences. We observe that visually appealing advertisements do not necessarily yield high user engagement, while solely maximizing user engagement may negatively affect brand perception. To bridge this gap, we introduce *AdsPrompter*, a new framework that aligns designer intent with user engagement for multimodal advertisement generation. Given a product image and taglines as inputs, *AdsPrompter* learns to create aesthetically pleasing advertising designs of varying sizes while faithfully preserving the product semantics, textual messages, and intended engagement goals. At the core of our framework is User-Designer Preference Optimization (UDPO), a coarse-to-fine learning strategy with complementary reward models for prompt refinement. The refined prompt guides the inputs to a diffusion-based model for preference-aligned advertisement generation. Furthermore, we introduce a large-scale advertisement dataset with 200k samples featuring designer and user preferences, along with a comprehensive benchmark. Experimental results show that *AdsPrompter* consistently outperforms existing methods in user engagement and design aesthetics.

1 INTRODUCTION

Multimodal advertising designs are of great importance in online commerce, acting as the primary role between products and customers (Lin et al., 2023; Li et al., 2023; Chen et al., 2025c). However, creating effective advertisements manually is labor-intensive, as it requires substantial design expertise to coherently integrate multimodal elements, including product images, background images, promotional text, and layout structures, in a professional way to attract user attention, deliver product messages, and guide user action. Recent advances in generative modeling (Dhariwal & Nichol, 2021), especially text-to-image generation (Zuo et al., 2025), have substantially improved the quality of generated content. Leveraging these models to automate the advertising design process has also attracted increasing attention (Lin et al., 2023; Chen et al., 2025c).

Towards this objective, existing works can be categorized into multi-stage pipelines and end-to-end generation methods. Multi-stage pipelines mainly decompose the advertising design generation problem into specialized subtasks, including background image generation (Chen et al., 2025c; Cao et al., 2024), layout generation (Yang et al., 2024; Seol et al., 2024; Lin et al., 2023), or visual text generation (Gao et al., 2025; Hsu et al., 2023). While effective for certain stages, these methods (Lin et al., 2023; Chen et al., 2025a) need to combine different models together to obtain final advertising designs, making the automation process vulnerable to error accumulation, cross-stage inconsistency, and weak global harmony. To alleviate these limitations, recent end-to-end methods (Hu et al., 2025; Lu et al., 2026) attempt to generate advertising designs within a unified framework. Despite simplifying the generation process, these methods typically optimize only one aspect of advertising effectiveness, i.e., either design aesthetics under design criteria (Lin et al., 2023) or user engagement measured by the click-through rate (CTR) (Chen et al., 2025c).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: Multimodal advertising design generation. Given the product image and taglines as inputs, existing methods (OpenAI, 2026; Zuo et al., 2025) often produce undesirable trade-offs between designer intent (DI) and user engagement (UE). In contrast, our method generates aesthetically pleasing advertisements while preserving faithful and effective user engagement signals.

In this work, we observe that such a formulation overlooks the inherent discrepancy between designer intent and user engagement. An advertisement that satisfies professional design principles, such as visual harmony, layout balance, and brand consistency, may still fail to attract user engagement. Conversely, optimizing solely for user engagement may favor visually aggressive designs that undermine brand perception (Zeng et al., 2021). Figure 1 illustrates this observation using advertisements generated by representative models (OpenAI, 2026; Zuo et al., 2025). We use an aesthetic score to represent designer intent and a CTR-based score for user engagement. The observed trade-off motivates us to raise a fundamental question: *How to bridge the gap between user and designer preferences for professional advertising design generation?*

To tackle this issue, we introduce AdsPrompter, a novel framework that aligns designer intent with user engagement for multimodal advertising design generation. Given a product image and taglines as inputs, AdsPrompter learns to create aesthetically pleasing advertising designs of varying sizes while faithfully conveying the intended engagement messages. At the core of our approach is User-Designer Preference Optimization (UDPO), a coarse-to-fine learning strategy with complementary reward models for prompt refinement. Specifically, we first perform supervised fine-tuning to generate coarse-level prompts that specify the overall composition, visual style, and arrangement of advertising elements. We then refine these prompts from both user and designer preferences via policy optimization using the proposed complementary reward models. The refined prompt guides the inputs to a diffusion-based model to produce preference-aligned advertising designs. To support the detailed analysis of the proposed framework, we further construct a large-scale advertisement dataset with 200k samples featuring designer and user preferences, along with a comprehensive benchmark. Experimental results demonstrate that AdsPrompter outperforms existing methods qualitatively and quantitatively, achieving a superior balance between user engagement and design aesthetics.

To sum up, we make the first attempt to explore multimodal advertising design generation that bridges the gap between designer and user preferences. We propose AdsPrompter, a novel framework with the UDPO training strategy for coarse-to-fine prompt optimization and advertisement generation. We also construct a large-scale advertisement dataset with designer and user preference annotations. Extensive evaluations show that our method consistently outperforms existing works in user engagement and design aesthetics. The code and dataset will be made publicly available.

2 Related Work

Designer Intent-aware Advertisement Generation. Many studies have explored how to incorporate design principles in advertising design generation, focusing on design-related factors, such as element placement, visual balance, typography arrangement, color compatibility, and style consistency. Layout-aware generation methods, for example, aim to improve the spatial organization of visual and textual elements (Seol et al., 2024; Hsu et al., 2023). Other works (Xu et al., 2023) study aesthetic scoring, poster-oriented planning, or unsupervised design quality assessment to provide signals for selecting or refining visually stronger results. Recent advances (Zhang et al., 2026) push the frontier further by leveraging powerful generative model architectures.

Despite these advances, existing methods mainly emphasize visual quality and design compliance, while paying less attention to whether the generated advertisement effectively delivers the product message and encourages user action, which is vital for online commerce applications. In contrast, our model considers multiple advertising factors by jointly optimizing designer and user preferences.

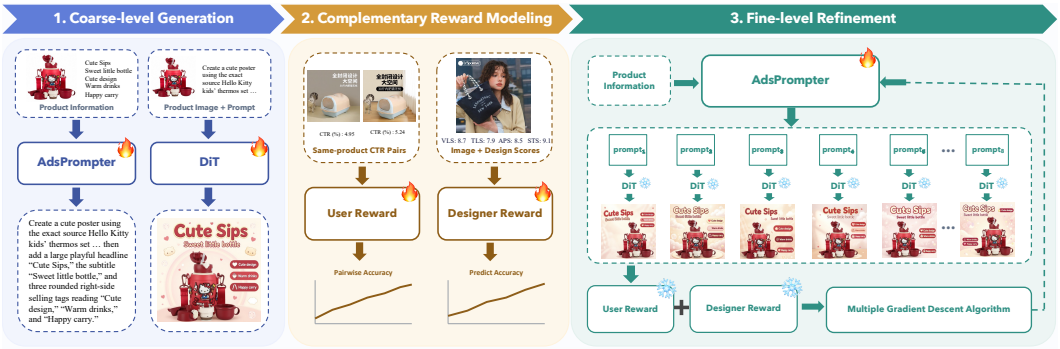

Figure 2: Overview of the AdsPrompter framework. The training pipeline consists of three stages, i.e., coarse-level generation, complementary reward modeling, and fine-level refinement. In the coarse-level generation stage, AdsPrompter generates initial prompts and candidate advertisements. In the complementary reward modeling stage, we train complementary user and designer reward models to evaluate user engagement and designer intent. In the fine-level refinement stage, AdsPrompter optimizes the prompt refinement policy with verifiable rewards.

User Engagement-aware Advertisement Generation. To evaluate whether an advertisement is likely to attract attention and encourage user engagement (Wang et al., 2021; Yang & Zhai, 2022), several works model click preference signals and use them to guide advertisement generation or ranking (Lu et al., 2026; Xu et al., 2026). CTR-driven advertising image generation (Chen et al., 2025c) optimizes product-centered backgrounds toward higher predicted click-through rates with multimodal large language models, while InnoAds-Composer (Qin et al., 2026) efficiently composes e-commerce posters by assembling reusable visual conditions.

Instead of optimizing user engagement alone, we jointly align designer intent and user engagement with UDPO, a coarse-to-fine learning strategy based on complementary reward models. Furthermore, we aim to generate professional advertising designs with multimodal elements, rather than focusing only on product-compatible background images.

Prompt Optimization for Generative Models. Prompt generation and optimization have become increasingly important in text-to-image and multimodal generation systems, as the quality of prompts strongly influences the fidelity, controllability, and stylistic properties of generated results. Prompt2Poster (Wang et al., 2024) creates artistic posters directly from textual prompts. PosterCraft (Chen et al., 2025b) further rethinks high-quality aesthetic poster generation within a unified, prompt-driven framework.

Different from existing prompt optimization methods that focus on instruction-following ability or generation quality for natural images, we formulate prompt refinement as a learnable process that can be optimized toward user engagement and design intent in advertising design generation.

3 Methodology

Given a product image and taglines, our goal is to generate a multimodal advertising design that preserves the product semantics and textual messages while improving user engagement. As shown in Figure 2, the training pipeline consists of three stages. In coarse-level refinement, AdsPrompter generates initial prompts and candidate advertisements. In complementary reward modeling, we train user and designer reward models to evaluate user engagement and designer intent. In fine-level refinement, AdsPrompter optimizes the prompt policy with verifiable rewards.

3.1 Coarse-level Generation

Prompt Generation. Given a product sample

$$x = (I_p, \mathcal{T}, S), \tag{1}$$

where I_p denotes the product foreground image, $\mathcal{T}$ denotes the taglines, and S denotes the target canvas size, we first construct a unified instruction condition

$$c = f_{\text{inst}}(x), \tag{2}$$

where $f_{\text{inst}}(\cdot)$ is an instruction formatting function that converts heterogeneous product information into a unified textual condition.

The instruction asks the model to act as a visual designer and prompt engineer and to generate a commercial advertisement for the product image while considering the candidate copy list. The model then produces a advertisement-oriented prompt y:

$$y \sim P_\phi(y \mid I_p, c), \tag{3}$$

where P_ϕ denotes the prompt generation model parameterized by ϕ. The training objective follows standard supervised fine-tuning:

$$\mathcal{L}_{\text{SFT}} = -\sum_{l=1}^{L} \log P_\phi(y_l \mid y_{<l}, I_p, c), \tag{4}$$

where L is the length of the target prompt y and y_l is its l-th token.

Image Generation. Unlike prior methods that mainly edit or replace product backgrounds, we synthesize multi-modal advertisements that jointly organize product presentation, textual promotion, and visual style, which requires adapting the renderer beyond generic image editing.

Given a product image I_p and a prompt y, the renderer produces an advertisement:

$$\hat{I} \sim G_\theta(\cdot \mid I_p, y), \tag{5}$$

where G_θ denotes DiT parameterized by θ. The prompt y specifies the advertising scene, selling points, and style, while I_p provides the visual identity to be preserved.

We perform full-parameter tuning rather than lightweight adaptation, since advertisement rendering requires product-aware scene construction, advertisement-style composition, and alignment with prompt semantics. This yields the renderer used in the subsequent reward-driven optimization.

The renderer is trained on paired data consisting of a product foreground image, a target advertisement, and the corresponding ground-truth advertisement caption. Let I_p denote the input foreground image, I^* the target advertisement, and y^* the target prompt. The supervised objective of the renderer is denoted as

$$\mathcal{L}_{\text{gen}} = \mathcal{L}_{\text{edit}}(G_\theta(I_p, y^*), I^*), \tag{6}$$

where $\mathcal{L}_{\text{edit}}$ denotes the training objective on paired advertisement data.

3.2 Complementary Reward Models

User Reward Model. Visual attractiveness alone is insufficient, since advertisements must also satisfy business objectives such as highlighting selling points and improving click-through potential. We construct a functional reward model that estimates the commercial effectiveness of generated advertisements. Given a rendered advertisement $\hat{I}$ and product condition x, the model outputs a functional reward score:

$$r_{\text{func}} = R_{\text{func}}(\hat{I}, x). \tag{7}$$

Instead of scoring prompts directly, the reward is computed on rendered advertisements conditioned on product information, evaluating each candidate advertisement's attractiveness and click potential.

Since CTR-related utility is better captured by relative preference than by absolute category-dependent values, we formulate reward learning as pairwise ranking. For a preference pair $(\hat{I}^+, \hat{I}^-)$ of the same product, where $\hat{I}^+$ is preferred over $\hat{I}^-$, the model is trained using

$$\mathcal{L}_{\text{func}} = -\log \sigma\Big(R_{\text{func}}(\hat{I}^+, x) - R_{\text{func}}(\hat{I}^-, x)\Big), \tag{8}$$

where $\sigma(\cdot)$ denotes the sigmoid function.

We train it with LoRA on all linear modules of Qwen3-VL-30B-A3B-Instruct, capturing relative utility differences among candidates while remaining robust to category-specific exposure and click variations. Training hyperparameters are in Appendix A.

Designer Reward Model. Beyond user utility, advertisements must also satisfy design requirements such as layout, visual attractiveness, and stylistic coherence. We construct an aesthetic reward model that predicts the visual quality of generated advertisements.

We adopt DINOv2 (Oquab et al., 2023) as the visual backbone and attach an ordinal MLP head. Given a generated advertisement $\hat{I}$, the head predicts M cumulative probabilities as:

$$r_{\text{aes}} = R_{\text{aes}}(\hat{I}) = \sum_{j=1}^{M} p_j(\hat{I}), \qquad u_{\text{aes}}(\hat{I}) = \sqrt{\sum_{j=1}^{M} p_j(\hat{I})\big(1 - p_j(\hat{I})\big)}. \tag{9}$$

The aesthetic supervision is derived from four design dimensions: appearance (s^{app}), visual layout (s^{vl}), textual layout (s^{tl}), and style coherence (s^{sty}). For advertisements rendered without text, s^{tl} is unassessable, so we neutralize this dimension rather than rewarding text-free outputs, which avoids creating an incentive to omit required promotional text.

We pre-compute the average textual layout score μ_{tl} from advertisements that contain text. The target aesthetic score $\tilde{s}(\hat{I})$ is then computed conditionally on the presence of text in $\hat{I}$:

$$\tilde{s}(\hat{I}) = \begin{cases} \frac{1}{4}\Big(s^{\text{app}}(\hat{I}) + s^{\text{vl}}(\hat{I}) + s^{\text{tl}}(\hat{I}) + s^{\text{sty}}(\hat{I})\Big), & \text{if } \hat{I} \text{ contains text,} \\ \frac{1}{4}\Big(s^{\text{app}}(\hat{I}) + s^{\text{vl}}(\hat{I}) + \mu_{\text{tl}} + s^{\text{sty}}(\hat{I})\Big), & \text{if } \hat{I} \text{ has no text.} \end{cases} \tag{10}$$

Let $z_j = \mathbf{1}[\tilde{s}(\hat{I}) > j]$ denote the ordinal target at threshold j. The aesthetic reward model is trained with a cumulative binary cross-entropy objective:

$$\mathcal{L}_{\text{aes}} = -\sum_{j=1}^{M} \Big[z_j \log p_j(\hat{I}) + (1 - z_j) \log\big(1 - p_j(\hat{I})\big) \Big]. \tag{11}$$

3.3 Fine-level Refinement

User-Designer Preference Optimization. Based on the supervised prompt model and the fixed renderer, we optimize the prompt policy to jointly improve functional utility and aesthetic quality. Since both reward models operate on rendered advertisements rather than prompts, optimization follows a prompt-to-image-to-reward pipeline. We denote the prompt policy as $\pi_\phi(y \mid I_p, c)$.

Conditioned on the product image I_p and the instruction condition c, the policy samples a group of candidate prompts

$$\{y_1, y_2, \ldots, y_K\} \sim \pi_\phi(\cdot \mid I_p, c), \tag{12}$$

where K denotes the group size. Each prompt is then rendered into a candidate advertisement:

$$\hat{I}_k \sim G_\theta(\cdot \mid I_p, y_k), \quad k = 1, \ldots, K. \tag{13}$$

Each rendered image is evaluated by the functional and aesthetic reward models. Let A_k^{func} and $\widetilde{A}_k^{\text{aes}}$ denote their separately group-standardized advantages, with the latter modulated by the aesthetic uncertainty. The fused advantage for the k-th sample is

$$A_k = w\, \widetilde{A}_k^{\text{aes}} + (1 - w) A_k^{\text{func}}, \quad k = 1, \ldots, K, \tag{14}$$

where $w \in [0, 1]$ is an adaptive global blend. It balances the two optimization objectives globally rather than imposing manually selected category-specific trade-offs: the product-conditioned URM and the visual DRM retain responsibility for modeling category-dependent preferences. Rather than receiving a direct gradient through group-centered rewards, w is periodically updated by MGDA from the two policy-gradient directions and smoothed by an exponential moving average.

For the periodic MGDA update, given the policy-gradient directions g_{aes} and g_{func} induced by the two streams, we solve the minimum-norm combination:

$$w^* = \text{clip}_{[0,1]}\left(\frac{\langle g_{\text{func}} - g_{\text{aes}},\, g_{\text{func}}\rangle}{\|g_{\text{aes}} - g_{\text{func}}\|_2^2}\right), \tag{15}$$

The update is synchronized across workers, and w^* is exponentially averaged with the previous weight to reduce batch-level noise.

The training objective is defined as:

$$\mathcal{L}_{\text{UDPO}} = -\frac{1}{K}\sum_{k=1}^{K} A_k \log \pi_\phi(y_k \mid I_p, c). \tag{16}$$

This objective encourages the policy to produce prompts that lead to higher-reward advertisements. We apply LoRA to the attention projection modules (q_proj, k_proj, v_proj, o_proj) of the prompt model and use diverse sampling to improve prompt diversity.

4 Dataset

We construct a large-scale advertisement dataset to address the scaricity of multimodal advertising design datasets with designer intent and user engagement annotations. The dataset contains approximately 200K product and advertisement pairs collected from real commercial campaigns, together with more than 10K product-controlled CTR preference pairs.

Data Source and Curation. Each corpus sample is represented as $(I_p, T, y, I_{\text{poster}})$, where I_p is the product image, T denotes OCR-extracted taglines, y is an advertisement description generated by a VLM under a domain-specific design protocol, and I_{poster} is the professionally designed advertisement. The design protocol describes the advertisement from multiple aspects, including scene context, product presentation, typography, decorative elements, depth, and overall style. We remove low-quality samples through automated filtering and human review to ensure reliable product visibility, readable text, and coherent advertising layouts.

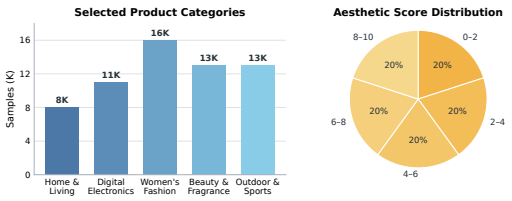

Figure 3: Dataset statistics including distributions over product categories and design aesthetics.

Designer Intent and User Engagement Annotations. Designer intent is annotated with four scores, i.e., visual layout, textual layout, appearance, and style. The mean of these four scores is used as the overall aesthetic target. User engagement is derived from one month of online logs by pairing creatives for the same product. We remove samples with fewer than 5K impressions and discard pairs whose CTR gap is below five percentage points. This product-controlled construction reduces product-popularity bias and provides reliable relative supervision for user engagement.

Benchmark Protocol. The benchmark contains 5,000 held-out product conditions and evaluates multimodal advertising design generation from user engagement and designer intent dimensions. We further use independent VLM judges, human studies, and online CTR measurements to validate whether the generated advertisements are visually credible and commercially effective. Please refer to the Appendix for the additional details of the dataset and the benchmark.

5 Experiments

5.1 Experimental Setup

Metrics. We evaluate generated advertisements from two perspectives, i.e., designer intent and user engagement. For **designer intent**, we use a large vision-language-model judge, Qwen3-VL-235B-A22B-Thinking (Qwen Team, 2025), to rate four aesthetic dimensions: Visual Layout Score (VLS) for composition balance, Textual Layout Score (TLS) for typography and readability, Appearance

Table 1: Quantitative Comparisons. We compare AdsPrompter with three types of existing models, i.e., background image generation models, multi-stage pipelines, and end-to-end generation models.

Method	Designer Intent					User Engagement		
	VLS ↑	TLS ↑	APS ↑	STS ↑	DRM ↑	Con ↑	Sen. Acc ↑	URM ↑
FLUX-Fill (Black Forest Labs, 2024)	8.217	-	7.894	7.951	0.621	0.893	-	0.512
CAIG (Chen et al., 2025c)	7.649	-	7.872	7.403	0.572	0.875	-	0.491
USO (Wu et al., 2025b)	8.357	-	7.914	8.216	0.668	0.914	-	0.536
AutoPoster (Lin et al., 2023)	8.192	7.160	7.685	8.077	0.634	0.751	0.788	0.518
Qwen-Image-Edit (Wu et al., 2025a)	8.156	7.023	7.857	8.124	0.657	0.908	0.814	0.553
Wan 2.7 (Team Wan et al., 2025)	7.986	6.958	7.559	7.822	0.596	0.824	0.753	0.478
Nano Banana Pro (Zuo et al., 2025)	8.298	7.215	7.812	8.156	0.692	0.921	0.849	0.601
GPT-Image-2 (OpenAI, 2026)	8.402	7.089	7.921	8.301	0.704	0.918	0.831	0.581
Ours	**8.561**	**7.356**	**7.971**	**8.469**	**0.821**	**0.929**	**0.870**	**0.742**

Score (APS) for visual clarity, and Style Score (STS) for color harmony. We also report the normalized Designer Reward Model (DRM) score as an overall measure of design quality. For **user engagement**, we report product Consistency (Con), computed by the cosine similarity between DINO features of generated and reference products, sentence-level text accuracy (Sen. Acc) for rendered-text correctness, online Click-Through Rate (CTR) from real-world A/B testing, and the normalized User Reward Model (URM) score for product-conditioned commercial appeal.

Compared Methods. To comprehensively evaluate our approach, we compare AdsPrompter with two types of previous SOTA models, i.e., multi-stage pipelines (AutoPoster (Lin et al., 2023)) and end-to-end generation methods, including Qwen-Image-Edit (Wu et al., 2025a), Wan 2.7 (Team Wan et al., 2025), GPT-Image-2 (OpenAI, 2026), and Nano Banana Pro (Zuo et al., 2025). In addition, we compare with product-centered background image generation methods, i.e., CAIG (Chen et al., 2025c), FLUX-Fill (Black Forest Labs, 2024), and USO (Wu et al., 2025b).

Implementation Details. All experiments run on 32×NVIDIA A100-80GB GPUs. We fine-tune Qwen3-VL-8B-Instruct (Bai et al., 2025; Qwen Team, 2025) for prompt refinement with the proposed UDPO learning strategy. The user reward model is built upon Qwen3-VL-30B-A3B-Instruct (Bai et al., 2025; Qwen Team, 2025) with LoRA adaptation, while the designer reward model adopts DINOv2 (Oquab et al., 2023) as the visual backbone with an ordinal head that provides both a score and an uncertainty estimate. Please refer to the Appendix for additional details.

5.2 Comparison with Prior Works

Quantitative Comparison. As shown in Table 1, our method consistently outperforms all baselines across designer intent and user engagement metrics. Although strong general-purpose generators such as USO and Qwen-Image-Edit demonstrate competitive visual generation capability, our method achieves the best product consistency and text accuracy, indicating stronger preservation of product semantics and promotional messages. GPT-Image-2 obtains solid appearance and style scores, but it achieves lower textual layout scores (TLS), as it often injects unrequested copy and renders text in overly small regions, which hurts both readability and sentence-level text accuracy (Sen. Acc). Nano Banana Pro renders text more faithfully, but it lags behind in appearance score (APS) and style score (STS), reflecting its recurring aesthetic and product-background mismatching issues. In contrast, our method achieves the strongest performance across all metrics, demonstrating that AdsPrompter more effectively bridges commercial utility and professional design aesthetics, a balance that existing methods struggle to achieve simultaneously.

Qualitative Comparison. As illustrated in Figure 4, our method demonstrates stronger multimodal advertising design quality in terms of text fidelity, layout hierarchy, product-background consistency, and overall professionalism. For the microphone example in the first row, competing methods such as AutoPoster and Qwen-Image-Edit can produce legible text, but their typography remains basic and lacks a clear design hierarchy. Wan 2.7 produces a less structured composition, resulting in a clunky layout that weakens the advertising message. The overall qualities of GPT-Image-2 and Nano Banana Pro are better, but they tend to hallucinate extra content or mismatched backgrounds, which harms readability and textual faithfulness. In contrast, our model generates aesthetically pleasing advertising designs while preserving the given product semantics and conveying the intended engagement message faithfully.



Figure 4: Qualitative Comparisons. Given the input taglines and the product image (1st column), we show the results from baselines (2nd to 5th columns) and our method (6th column), respectively.



Figure 5: Qualitative comparisons of background image generation.



Figure 6: Generalization ability of AdsPrompter across generators

Background Image Comparison. We further compare AdsPrompter with existing product-aware background image generation methods (Black Forest Labs, 2024; Chen et al., 2025c; Wu et al., 2025b). The results are shown in Figure 5. For the first example, our method constructs a serene and well-lit scene with relevant props, leading to a more appealing and commercially suitable presentation. By comparison, USO introduces a mismatched cartoon forest, while CAIG and FLUX-Fill produce simplistic dark surfaces that provide limited advertising value. These results suggest that AdsPrompter can synthesize backgrounds that are not only visually appealing but also semantically aligned with the product and suitable for real-world advertising scenarios.

Generalization of AdsPrompter across Generators. In addition, we evaluate whether AdsPrompter can generalize across different image generators. Given the same inputs, we compare two strategies in Figure 6. The first strategy directly feeds the original inputs to each image generator, while the second first refines the inputs with AdsPrompter and then sends the refined prompt to the same generator. Without AdsPrompter, Nano Banana and Wan 2.7 produce highly divergent advertisements, showing large variations in layout hierarchy, visual communication, and product placement. In comparison, these generators produce more similar advertising designs after AdsPrompter refinement, closely following the composition, visual hierarchy, and product emphasis. These results suggest that our model provides generator-agnostic design guidance, indicating that AdsPrompter learns transferable advertising plans rather than generator-specific prompt patterns.

5.3 Ablation Study

Component Analysis. To evaluate the contribution of each component in AdsPrompter, we conduct ablation studies with several variants that remove the coarse-to-fine learning strategy and individual preference rewards. In addition to the metrics mentioned above, we also conduct an online A/B test for CTR scores to evaluate whether AdsPrompter improves real user engagement. As shown in Table 2, the baseline model provides basic prompt generation ability but remains limited in product-relevant rendering and advertising alignment. Using coarse-level generation alone (i.e., w/o RL) improves these aspects, showing that advertising-specific instruction tuning is important for strong prompt initialization. Further optimization with both user and designer rewards brings consistent

Table 2: Quantitative ablation study of the proposed components.

Method	Designer Intent					User Engagement			
	VLS ↑	TLS ↑	APS ↑	STS ↑	DRM ↑	Con ↑	CTR (%) ↑	Sen. Acc ↑	URM ↑
Baseline	7.940	6.911	7.579	7.852	0.621	0.885	5.12	0.812	0.538
w/o RL	8.372	7.275	7.853	8.350	0.706	0.910	5.33	0.862	0.585
w/o Designer Reward	8.283	7.144	7.625	8.236	0.691	0.915	5.46	0.805	0.655
w/o User Reward	8.351	7.218	7.886	8.312	0.769	0.924	5.35	0.858	0.562
Ours (full)	**8.561**	**7.356**	**7.971**	**8.469**	**0.821**	**0.929**	**5.51**	**0.870**	**0.742**

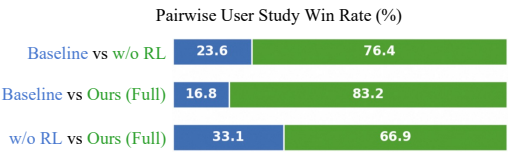

Figure 7: Pairwise win rates from the user study.



Figure 8: Effect of prompt strcture.

gains across all metrics, indicating that fine-level refinement improves the model beyond supervised imitation. Removing either reward causes a clear performance drop, with the user reward mainly affecting commercial appeal and the designer reward mainly affecting textual layout and stylistic harmony. Our full model achieves the best performance, confirming the effectiveness of the proposed pipeline in balancing commercial utility and design aesthetics.

User Study. We conduct a pairwise human evaluation to assess the perceived quality of generated advertisements, involving 6 professional designers and 24 novice users. We randomly sample 200 comparison pairs, covering more than 50 product categories. The study follows a double-blind protocol with randomized left and right placement and randomized presentation order. Participants are asked to select the advertisement that better satisfies advertising standards in terms of textual fidelity, visual quality, layout design, and commercial appeal. As shown in Figure 7, the two-sided binomial tests show significant preferences for our full model over the baseline (61.2% win rate, $p < 0.005$) and the coarse-level generation (63.5% win rate, $p < 0.001$).

Prompt Structure Analysis. To evaluate whether our structured description protocol captures the information for faithful advertisement reconstruction, we conduct a prompt structure analysis as follows. Specifically, we compare our full prompt protocol with three ablated variants for advertising design generation, i.e., w/o Structure that removes layout structure and inter-element interaction information, w/o Text that removes fine-grained typographic attributes, and w/o Visual that removes quantitative visual attributes. As shown in Figure 8, removing any part of the structured prompt leads to noticeably degraded reconstruction quality. These results show that structural, textual, and visual descriptions provide complementary information, and their joint use is necessary for high-fidelity advertisement generation with accurate layout, typography, and visual coherence.

6 Conclusion

In this paper, we address the gap between design aesthetics and user engagement in advertising design generation. We propose AdsPrompter, a new framework that aligns designer intent with user engagement for multimodal advertisement generation. The key idea is to refine advertising prompts with User-Designer Preference Optimization, a coarse-to-fine learning strategy with complementary reward models for prompt refinement. Experimental results show that AdsPrompter generates advertisements that are both visually appealing and commercially effective across quantitative metrics, human evaluation, and cross-generator generalization.

Although AdsPrompter works well across diverse advertising scenarios, it remains limited in cases that require extremely dense and complex textual rendering. In such cases, direct image generation may struggle to preserve fine-grained text layout, typography, and readability. A potential solution is to represent textual elements as executable HTML/CSS components guided by the structural descriptions produced by AdsPrompter.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing and grammatical correction. All research ideas, method design, experiments, and analyses were conceived and carried out by the authors, and all AI-assisted text has been reviewed by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our work studies advertising design generation using data collected from a commercial e-commerce platform. The click-through rate records used in this work contain no personally identifiable information about individual users. We have obtained IRB approval for the user study.

REPRODUCIBILITY STATEMENT

We describe the model architecture and all training objectives in Section 3, the dataset construction pipeline and annotation protocol in Section 4, and the evaluation protocol, baselines, and metrics in Section 5. Additional implementation details for the proposed framework are reported in Appendix A. As the first step towards this direction, we will release both the code and the dataset to facilitate future research in the community.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024. Accessed: 2024-10-27.

Tingfeng Cao, Junsheng Kong, Xue Zhao, Wenqing Yao, Junwei Ding, Jinhui Zhu, and Jiandong Zhang. Product2img: Prompt-free e-commerce product background generation with diffusion model and self-improved lmm. In *ACM MM*, 2024.

Hongyu Chen, Min Zhou, Jing Jiang, Jiale Chen, Yang Lu, Zihang Lin, Bo Xiao, Tiezheng Ge, and Bo Zheng. T-stars-poster: A framework for product-centric advertising image design. *arXiv preprint arXiv:2501.14316*, 2025a.

SiXiang Chen, Jianyu Lai, Jialin Gao, Tian Ye, Haoyu Chen, Hengyu Shi, Shitong Shao, Yunlong Lin, Song Fei, Zhaohu Xing, et al. Postercraft: Rethinking high-quality aesthetic poster generation in a unified framework. *arXiv preprint arXiv:2506.10741*, 2025b.

Xingye Chen, Wei Feng, Zhenbang Du, Weizhen Wang, Yanyin Chen, Haohan Wang, Linkai Liu, Yaoyu Li, Jinyuan Zhao, Yu Li, et al. Ctr-driven advertising image generation with multimodal large language models. In *WWW*, 2025c.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021. URL `https://arxiv.org/abs/2105.05233`.

Yifan Gao, Zihang Lin, Chuanbin Liu, Min Zhou, Tiezheng Ge, Bo Zheng, and Hongtao Xie. Postermaker: Towards high-quality product poster generation with accurate text rendering. In *CVPR*, 2025.

Gemini Team. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2025. URL `https://arxiv.org/abs/2312.11805`.

Hsiao Yuan Hsu, Xiangteng He, Yuxin Peng, Hao Kong, and Qing Zhang. Posterlayout: A new benchmark and approach for content-aware visual-textual presentation layout. In *CVPR*, 2023.

Xiwei Hu, Haokun Chen, Zhongqi Qi, Hui Zhang, Dexiang Hong, Jie Shao, and Xinglong Wu. Dreamposter: A unified framework for image-conditioned generative poster design. *arXiv preprint arXiv:2507.04218*, 2025.

Zhaochen Li, Fengheng Li, Wei Feng, Honghe Zhu, Yaoyu Li, Zheng Zhang, Jingjing Lv, Junjie Shen, Zhangang Lin, Jingping Shao, et al. Planning and rendering: Towards product poster generation with diffusion models. *arXiv preprint arXiv:2312.08822*, 2023.

Jinpeng Lin, Min Zhou, Ye Ma, Yifan Gao, Chenxi Fei, Yangjian Chen, Zhang Yu, and Tiezheng Ge. Autoposter: A highly automatic and content-aware design system for advertising poster generation. In *ACM MM*, 2023.

Shuo Lu, Haohan Wang, Wei Feng, Weizhen Wang, Shen Zhang, Yaoyu Li, Ao Ma, Zheng Zhang, Jingjing Lv, Junjie Shen, et al. One size, many fits: Aligning diverse group-wise click preferences in large-scale advertising image generation. *arXiv preprint arXiv:2602.02033*, 2026.

OpenAI. Gpt-image-2. `https://openai.com/zh-Hans-CN/index/introducing-chatgpt-images-2-0/`, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Yuxin Qin, Ke Cao, Haowei Liu, Ao Ma, Fengheng Li, Honghe Zhu, Zheng Zhang, Run Ling, Wei Feng, Xuanhua He, et al. Innoads-composer: Efficient condition composition for e-commerce poster generation. *arXiv preprint arXiv:2603.05898*, 2026.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL `https://arxiv.org/abs/2505.09388`.

Jaejung Seol, Seojun Kim, and Jaejun Yoo. Posterllama: Bridging design ability of language model to content-aware layout generation. In *ECCV*, 2024.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. TRL: Transformer Reinforcement Learning. `https://github.com/huggingface/trl`, 2020.

Shaodong Wang, Yunyang Ge, Liuhan Chen, Haiyang Zhou, Qian Wang, Xinhua Cheng, and Li Yuan. Prompt2poster: Automatically artistic chinese poster creation from prompt only. In *ACM MM*, 2024.

Shiyao Wang, Qi Liu, Tiezheng Ge, Defu Lian, and Zhiqiang Zhang. A hybrid bandit model with visual priors for creative ranking in display advertising. In *WWW*, 2021.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025a. URL `https://arxiv.org/abs/2508.02324`.

Shaojin Wu, Mengqi Huang, Yufeng Cheng, Wenxu Wu, Jiahe Tian, Yiming Luo, Fei Ding, and Qian He. Uso: Unified style and subject-driven generation via disentangled and reward learning. *arXiv preprint arXiv:2508.18966*, 2025b.

Chenchen Xu, Min Zhou, Tiezheng Ge, Yuning Jiang, and Weiwei Xu. Unsupervised domain adaption with pixel-level discriminator for image-aware layout generation. In *CVPR*, 2023.

Yexing Xu, Wei Feng, Shen Zhang, Haohan Wang, Yuxin Qin, Yaoyu Li, Ao Ma, Yuhao Luo, Lu Wang, Xudong Ren, Haoran Wang, Run Ling, Zheng Zhang, Jingjing Lv, Junjie Shen, Ching Law, Longguang Wang, and Yulan Guo. Design your ad: Personalized advertising image and text generation with unified autoregressive models. In *CVPR*, 2026.

Kota Yamaguchi. Canvasvae: Learning to generate vector graphic documents. *ICCV*, 2021.

Tao Yang, Yingmin Luo, Zhongang Qi, Yang Wu, Ying Shan, and Chang Wen Chen. Posterllava: Constructing a unified multi-modal layout generator with llm. *arXiv preprint arXiv:2406.02884*, 2024.

Yanwu Yang and Panyu Zhai. Click-through rate prediction in online advertising: A literature review. *Information Processing & Management*, 2022.

Eric Zeng, Tadayoshi Kohno, and Franziska Roesner. What makes a "bad" ad? user perceptions of problematic online advertising. In *CHI*, 2021.

Hui Zhang et al. Creatidesign: A unified multi-conditional diffusion transformer for creative graphic design. In *ICLR*, 2026.

Jialong Zuo, Haoyou Deng, Hanyu Zhou, Jiaxin Zhu, Yicheng Zhang, Yiwei Zhang, Yongxin Yan, Kaixing Huang, Weisen Chen, Yongtai Deng, et al. Is nano banana pro a low-level vision all-rounder? a comprehensive evaluation on 14 tasks and 40 datasets. *arXiv preprint arXiv:2512.15110*, 2025.

This appendix provides additional details on model implementation, dataset construction, and experimental analysis. In particular, we provide training configurations for each component in the pipeline, a comparison with related datasets, explanations of the dataset annotation protocol, independent validation, and additional results.

A Additional Model Details

This section provides supplementary implementation details for the full pipeline described in Section 3. To stay aligned with the organization in the main text, we present the additional details following the same coarse-stage and fine-stage formulation. Table 3 summarizes the main training configurations.

A.1 Coarse Stage

A.1.1 Prompt Generation Model

Following Section 3.1, the prompt generation model is initialized from Qwen3-VL-8B-Instruct (Bai et al., 2025; Qwen Team, 2025). Given a product sample $x = (I_p, T, S)$, where I_p denotes the product foreground image, T denotes the candidate tagline set, and S denotes the target canvas size, the model generates an image-oriented prompt conditioned on the multimodal product information. In addition to the product image and textual selling points, the canvas size S is explicitly included as part of the condition, so that the generated prompt can better adapt to different image aspect ratios and layout requirements. This makes the supplementary description fully consistent with the formulation in Equations 1 and 3 of the main paper.

During supervised fine-tuning, we freeze the visual encoder and optimize only the language-side parameters. The model is trained for 3 epochs with a cosine learning rate scheduler and a learning rate of 1×10^{-5}. The per-device batch size is set to 1, with gradient accumulation steps of 4. The maximum sequence length is set to 4096 tokens. All prompt generation experiments are conducted in bfloat16 precision.

A.1.2 Image Renderer

For image rendering, we adopt Qwen-Image-Edit-2511 (Wu et al., 2025a) as the base model and perform full-parameter fine-tuning, as described in Section 3.1. Different from conventional background replacement methods, our renderer is adapted to generate complete advertising images conditioned on the product foreground image and the image-oriented prompt.

In our implementation, the renderer is trained on 64 GPUs for 2 epochs. We use the Prodigy optimizer with a learning-rate coefficient of 1.0 and a cosine-with-warmup scheduler. The minimum learning-rate ratio is set to 0.1. The per-device batch size is 1, with gradient accumulation steps of 2. The total number of training steps is 5625, including 340 warmup steps. The maximum image resolution is set to 1024×1024, and the trainable module is the DiT component of Qwen-Image-Edit-2511. Gradient checkpointing is enabled during training. Although the maximum pixel budget is bounded during training, the renderer supports flexible canvas sizes and aspect ratios in practice, enabling the generation of square, portrait, and landscape advertising images under different layout requirements.

A.2 Fine Stage

A.2.1 User Reward Model

The user reward model is built upon Qwen3-VL-30B-A3B-Instruct (Bai et al., 2025; Qwen Team, 2025). It is trained as an image-level reward model with pairwise supervision derived from real CTR preference pairs. Given a rendered image and the corresponding product condition, the model predicts a click-oriented functional reward that reflects user-side commercial preference.

We adopt LoRA adaptation and train the model for 8 epochs with a learning rate of 2×10^{-6}. The per-device batch size is 1, and the gradient accumulation steps are set to 4. The LoRA rank and

Table 3: Training configurations of different components in the AdsPrompter pipeline.

Component	Backbone	Type	Ep.	LR	Optimization and Other Settings
Coarse Stage					
Prompt Generator	Qwen3-VL-8B-Instruct	SFT	3	$1e{-5}$	Freeze visual encoder; cosine scheduler. Input (I_p, T, S); cutoff length 4096; bf16; batch size 1; grad. acc. 4.
Image Renderer	Qwen-Image-Edit-2511	Full FT	2	1.0	Prodigy; cosine-with-warmup; 5625 steps; warmup 340. Batch size 1; grad. acc. 2; max resolution 1024×1024.
Fine Stage					
User RM	Qwen3-VL-30B-A3B-Instruct	LoRA reward	8	$2e{-6}$	Rank 256, alpha 512; all-linear; warmup ratio 0.1. Center reward coefficient 0.01; batch size 1; grad. acc. 4.
Designer RM	DINOv2 + ordinal head	Ordinal reward	–	–	Cumulative binary cross-entropy; outputs score and uncertainty; text-free images use neutralized textual-layout supervision.
GRPO	Qwen3-VL-8B-Instruct + adapter	LoRA policy	3	$1e{-6}$	MGDA+EMA adaptive reward fusion; 8 generations; temp. 1.0; top-p 0.85; rank 256, alpha 512; $q_{\text{proj}}/k_{\text{proj}}/v_{\text{proj}}/o_{\text{proj}}$; batch size 1; grad. acc. 8.

LoRA alpha are set to 256 and 512, respectively, and LoRA is applied to all linear layers. We use a warmup ratio of 0.1 and a center reward coefficient of 0.01 during training.

A.2.2 Designer Reward Model

The designer reward model adopts DINOv2 (Oquab et al., 2023) as the visual backbone with an ordinal head that produces both an aesthetic score and an uncertainty estimate. We choose DINOv2 because it provides strong and stable visual representations for image-level quality assessment. For results containing little or no text, we replace the unassessable textual-layout term with its corpus-level mean; this neutralization avoids penalizing text-free images without rewarding the omission of required text. The model is trained with a cumulative binary cross-entropy objective over ordinal thresholds and later provides the aesthetic reward and uncertainty signal for fine-stage optimization.

A.2.3 User-Designer Preference Optimization

We further optimize the supervised prompt model with GRPO (Shao et al., 2024). The optimization starts from the supervised prompt model checkpoint and applies LoRA adaptation to the attention projection layers, including q_proj, k_proj, v_proj, and o_proj. For each product input, the prompt model samples multiple candidate prompts, which are then rendered by the fixed Qwen-Image-Edit-2511 renderer into candidate images. Each generated image is scored by the user and designer reward models. Their separately group-standardized advantages are fused with an MGDA-updated global weight and EMA smoothing, consistent with Equation 14 of the main paper.

In our implementation, the GRPO stage is trained for 3 epochs with a learning rate of 1×10^{-6}. The per-device batch size is 1, and the gradient accumulation steps are set to 8. We sample 8 candidate prompts for each input. During generation, we set the sampling temperature to 1.0 and top-p to 0.85. We further enable dynamic sampling with a maximum resampling number of 3, together with overlong filtering to improve prompt quality and stability. The LoRA rank and alpha are set to 256 and 512, respectively.

A.3 Training Cost and Resources

AdsPrompter incurs one-time offline training costs. GRPO training takes 52 hours on two nodes of eight A100 GPUs, URM training takes 60 hours on four nodes of eight A100 GPUs, and renderer fine-tuning takes 39 hours on eight nodes of eight A100 GPUs. At serving time, a single image



Figure 9: Examples of prompt annotation.

is rendered in approximately 30 seconds on an H20 GPU; these offline training resources are not required for online inference.

B Additional Dataset Details

This section provides supplementary details for the dataset introduced in Section 4. Our dataset is designed to support multimodal advertising design generation under both designer and user preferences. In addition to the large-scale training corpus and preference annotations described in the paper, we further summarize the practical construction details, sample organization, and release protocol in this appendix.

B.1 Prompt Annotation

While the main paper briefly introduces the image-oriented prompt generation process, here we further clarify its annotation granularity. Instead of directly using generic captions, we adopt a domain-specific prompting protocol that encourages the VLM to describe the image from a professional e-commerce design perspective. The generated prompt typically contains the following aspects:

- Main subject description: product category, quantity, position, scale, and product-to-product relations;
- Scene environment: background type, surface material, lighting condition, color palette, and depth cues;



(a) Designer Preference Annotation **(b) User Preference Annotation**

Figure 10: Examples of preference annotation.

- Typography attributes: text placement, hierarchy, font style, outline, color, readability, and callout elements;
- Decorative graphics: icons, stickers, speech bubbles, rays, abstract shapes, and other supporting visual elements;
- Photographic characteristics: brightness, contrast, saturation, depth of field, and blur pattern;
- Overall style: category-specific advertising tone such as cartoon, premium, technological, or home-lifestyle style.

These prompts are intentionally much richer than ordinary image captions. Their role is twofold: first, they provide supervised targets for the prompt generation model in the coarse stage; second, they serve as high-quality conditional descriptions for the image renderer. As shown in Figure 9, such prompts preserve both holistic scene semantics and fine-grained design details, especially for typography-heavy advertising images.

B.2 Designer Preference Annotation

The main paper defines the designer preference score as the average of four sub-scores, namely VLS, TLS, APS, and STS. Here we further explain the practical meaning of these dimensions and their usage in data curation.

Visual Layout Score (VLS). This score focuses on subject prominence, compositional balance, foreground-background coordination, and whether the overall visual arrangement follows common commercial design principles. Typical failure cases include subject crowding, awkward cropping, and visually unstable composition.

Textual Layout Score (TLS). This dimension evaluates whether promotional texts are readable, well organized, and stylistically appropriate. It also reflects whether fonts, outlines, and stickers are professionally arranged and whether important text overlaps critical product regions.

Appearance Score (APS). This score measures low-level and mid-level visual quality, including sharpness, texture clarity, material realism, and lighting quality. It is particularly important for product-centric advertising because poor appearance quality can directly harm trustworthiness and purchase intent.

Style Score (STS). This dimension measures color harmony, atmosphere consistency, and whether the image style matches the product category and commercial positioning. For example, children's products often prefer playful and saturated styles, while home appliances may benefit from cleaner and more premium visual tones.

As shown in Figure 10(a), the four scores are stored at the sample level and are later aggregated into the overall aesthetic target used for reward modeling. This design is aligned with the designer reward model described in Section **??**, where aesthetic quality is learned from appearance, visual layout, textual layout, and style coherence.

B.3 User Preference Annotation

In addition to designer annotations, our dataset contains real-world CTR preference pairs for user-side supervision. Figure 10(b) shows representative real-world CTR-based user preference pairs. For the same product, two advertising images are associated with their observed click-through rates (CTR), forming pairwise user preference supervision for functional reward modeling. Following the notation in the main paper, each sample is represented as $(x^{(j)}, I_+^{(j)}, I_-^{(j)})$, where $I_+^{(j)}$ and $I_-^{(j)}$ denote the higher-CTR and lower-CTR images, respectively.

We emphasize that this pairwise construction is important for two reasons. First, it controls the product variable, making CTR differences more attributable to visual communication and image design rather than to intrinsic product popularity. Second, pairwise supervision is naturally compatible with preference learning objectives, which are more stable than directly regressing absolute CTR values across heterogeneous product categories.

B.4 Dataset Comparison

Table 4 compares our dataset with representative resources for advertisement layout, graphic design, advertisement generation, and advertising preference modeling. Existing datasets focus on layout, graphic elements, or structured posters, whereas our data provide the end-to-end multimodal signals required by AdsPrompter: product images, taglines, image-oriented prompts, target advertising images, designer scores, and CTR preference pairs. The full internal corpus contains over 200K images and more than 10K real-world CTR preference pairs; we will release a 5K-sample test set built with the same annotation protocol.

B.5 Data Quality Control

All prompt annotations are generated with an open-source Qwen VLM using the structured protocol described above and are screened through cross-modal reconstruction checks. For designer supervision, ten professional designers manually review sampled VLM ratings; examples judged unreasonable by more than three reviewers are removed. We further filter raw advertisements below 1024×1024, blurry, or heavily watermarked, manually spot-check foreground matting boundaries, and use OCR cross-verification to discard advertisements with garbled or severely corrupted text. For CTR preference data, we remove visually near-identical pairs with anomalously large CTR gaps to reduce residual non-visual delivery noise.

B.6 Online A/B Testing Protocol

We sample 100K active products across categories and generate identical-sized creatives for each method. Live traffic is randomly assigned to three disjoint buckets, each receiving 5% of the traffic; only the generated creatives for the sampled products are displayed in these buckets. The test runs



(a) Background Synthesis



(b) Full Image Generation

Figure 11: More qualitative results of AdsPrompter.

continuously for one week, and each bucket accumulates more than 1.2M impressions and 60K clicks. We test CTR differences with a two-sample proportion z-test on the collected logs. This protocol controls the delivery setting at the experiment level and evaluates whether the generated creative changes engagement under the same randomized deployment procedure.

C Independent Evaluation and Reward Validation

To complement the in-family VLM judge used for the main ADE metrics, we cross-evaluate representative methods using GPT-5.5 and Gemini-3-Pro. Table 5 shows that the relative ordering remains consistent across independent VLM judges and online CTR.

Table 4: Comparison between our dataset and representative related datasets.

Dataset	Application	Instances / Split	Label Types	Avail.
PKU Poster-Layout (Hsu et al., 2023)	Content-aware layout	9,974 / 905 (train / test)	BBox, category, canvas, saliency	Public
Crello (Yamaguchi, 2021)	Vector graphic design	18,714 / 2,316 / 2,331 (train / val / test)	Group, layer, font, color, BBox, canvas size, pre-rendered elements	Public
CGL-Dataset (Xu et al., 2023)	E-commerce poster generation	60,548 / 1,035 (train / test)	BBox, category, canvas size	Public
Creative-Ranking (Wang et al., 2021)	Ad creative ranking / CTR modeling	Large-scale industrial logs	Multiple creatives for the same product, user feedback, CTR-related signals	Internal
Ours	End-to-end multimodal advertising design generation	200K+ image samples; 10K+ CTR preference pairs; 5K public test set	Product image, taglines, image-oriented prompt, advertising image, designer scores, CTR preference pairs	To be released

Table 5: Independent cross-evaluation with GPT-5.5, Gemini-3-Pro, and online CTR. Higher is better.

Method	GPT-5.5				Gemini-3-Pro				CTR (%)
	VLS	TLS	APS	STS	VLS	TLS	APS	STS	
CAIG (Chen et al., 2025c)	7.51	-	7.68	7.22	7.48	-	7.62	7.18	4.950
FLUX-Fill (Black Forest Labs, 2024)	8.10	-	7.70	7.80	8.05	-	7.65	7.74	4.210
USO (Wu et al., 2025b)	8.22	-	7.75	8.02	8.18	-	7.71	7.96	4.380
AutoPoster (Lin et al., 2023)	8.05	6.98	7.50	7.92	8.01	6.92	7.45	7.88	5.020
Qwen-Image-Edit (Wu et al., 2025a)	7.98	6.85	7.69	7.95	7.92	6.80	7.63	7.90	5.080
Wan 2.7 (Team Wan et al., 2025)	7.80	6.78	7.41	7.65	7.75	6.72	7.35	7.59	4.820
Nano Banana Pro (Zuo et al., 2025)	8.24	7.06	7.74	8.09	8.19	7.01	7.69	8.03	5.110
GPT-Image-2 (OpenAI, 2026)	8.34	6.94	7.86	8.21	8.29	6.89	7.80	8.16	5.160
Ours (Full)	**8.40**	**7.20**	**7.81**	**8.32**	**8.35**	**7.15**	**7.76**	**8.27**	**5.439**

Independent validation is performed on 10K held-out real-world CTR preference pairs for URM. For DRM, five professional designers score 500 randomly selected advertisements on the four design dimensions using a 1–10 scale; we compare model predictions with the mean human ratings. Table 6 supports the use of both models as proxy objectives.

D Additional Experimental Results

D.1 Qualitative Evaluation

In addition to the representative examples shown in the main paper, we provide more qualitative results of AdsPrompter in Figure 11. These examples are intended to further demonstrate the robustness and versatility of our framework across different advertising generation settings and canvas sizes. Specifically, we include both *background synthesis* results and *full image generation* results. The former mainly evaluates whether AdsPrompter can synthesize visually coherent and product-relevant background scenes while preserving the identity and appearance of the input product. The latter is more challenging, since it requires the model to jointly generate full advertising images with coordinated product presentation, promotional text, layout design, and overall visual style.

Moreover, the presented examples cover images of different aspect ratios and design styles, further showing that AdsPrompter is not restricted to a single template or fixed canvas format. Across these diverse cases, our framework consistently produces visually appealing results that align well with both commercial communication goals and design quality requirements. These additional exam-

Table 6: Independent validation of the user and designer reward models.

URM Validation		DRM Validation	
Metric	Value	Metric	Value
Pairwise Accuracy	0.824	Spearman ρ (VLS)	0.712
AUC	0.845	Spearman ρ (TLS)	0.745
Pearson r with online CTR	0.691	Spearman ρ (APS)	0.701
Spearman ρ with online CTR	0.682	Spearman ρ (STS)	0.728
Kendall τ with online CTR	0.514	Mean Kendall τ	0.564



Figure 12: Qualitative comparison of different VLMs under the same instruction template and the same image renderer.

ples complement the qualitative comparisons in Section 5 of the main paper and provide a more comprehensive view of the generation capability of AdsPrompter.

D.2 Prompt Generator Comparison

Figure 12 compares different prompt generation models, including GPT-4o (von Werra et al., 2020), Gemini-3 (Gemini Team, 2025), Qwen3-VL-235B-A22B-Instruct (Bai et al., 2025), and our AdsPrompter, under the same instruction template and the same image renderer, Qwen-Image-Edit (Wu et al., 2025a). Since the prompt format is aligned across methods, the quality gap mainly comes from the prompt content and its adaptation to advertising design generation.

In the first row, GPT-4o and Gemini-3 show poor text readability, while Gemini-3 and Qwen3-VL-235B-A22B-Instruct further produce obvious text rendering errors, such as incorrectly merging two unrelated price-related promotional snippets into a single text region. In the second row, GPT-4o generates text that is both hard to read and partially occludes the product, while Gemini-3 and Qwen3-VL-235B-A22B-Instruct produce weaker layouts. In the third row, AdsPrompter achieves the best layout, with more appropriate whitespace and a less crowded composition. These examples show that the prompt generation model plays a crucial role in determining the final quality of advertising images.

D.3 Prompt Format Comparison

We further compare two prompt formulation strategies for advertising image generation: a JSON-structured prompt and our organized natural-language prompt. The JSON-structured



Figure 13: Qualitative comparison between the JSON-structured prompt and our organized natural-language prompt.

prompt explicitly decomposes the image into hierarchical fields, including *scene_composition* and *global_attributes*, where the former is further divided into background, object, and text-graphic layers, and the latter summarizes global properties such as material finish, depth of field, and overall art style. In contrast, our prompt expresses the same information in a compact natural-language form, describing scene layout, object relations, typography, decorative elements, visual attributes, and overall style in a coherent paragraph. Compared with the JSON format, our prompt is more concise and typically contains only 500–700 characters.

For fair comparison, we separately fine-tune Qwen-Image-Edit-2511 (Wu et al., 2025a) using JSON-structured prompts and our prompts, respectively, and then use the corresponding prompt format at inference time for image generation. The qualitative results are shown in Figure 13. Although the two prompt types encode largely similar semantic content, they lead to noticeably different generation behaviors.

First, our prompt yields more stable text rendering and overall composition. In the first row, the model trained with the JSON-structured prompt produces redundant text, where one promotional phrase is rendered twice. In the second row, the JSON-based result shows poor background generation and also truncates part of the main object, leading to an incomplete composition. In the third row, the JSON-based result contains obvious Chinese text errors. By contrast, the model trained with our prompt generates cleaner layouts, more coherent backgrounds, and more reliable Chinese text rendering across all examples.