

# RefAdGen: High-Fidelity Advertising Image Generation

Yiyun Chen<sup>1</sup>, Weikai Yang<sup>1\*</sup>

<sup>1</sup>The Hong Kong University of Science and Technology (Guangzhou)  
ychen780@connect.hkust-gz.edu.cn, weikaiyang@hkust-gz.edu.cn

## Abstract

The rapid advancement of Artificial Intelligence Generated Content (AIGC) techniques has unlocked opportunities in generating diverse and compelling advertising images based on referenced product images and textual scene descriptions. This capability substantially reduces human labor and production costs in traditional marketing workflows. However, existing AIGC techniques either demand extensive fine-tuning for each referenced image to achieve high fidelity, or they struggle to maintain fidelity across diverse products, making them impractical for e-commerce and marketing industries. To tackle this limitation, we first construct AdProd-100K, a large-scale advertising image generation dataset. A key innovation in its construction is our dual data augmentation strategy, which fosters robust, 3D-aware representations crucial for realistic and high-fidelity image synthesis. Leveraging this dataset, we propose RefAdGen, a generation framework that achieves high fidelity through a decoupled design. The framework enforces precise spatial control by injecting a product mask at the U-Net input, and employs an efficient Attention Fusion Module (AFM) to integrate product features. This design effectively resolves the fidelity-efficiency dilemma present in existing methods. Extensive experiments demonstrate that RefAdGen achieves state-of-the-art performance, showcasing robust generalization by maintaining high fidelity and remarkable visual results for both unseen products and challenging real-world, in-the-wild images. This offers a scalable and cost-effective alternative to traditional workflows.

**Code** — <https://github.com/yunsoft2019/RefAdGen>

## Introduction

In the fast-paced digital marketing and e-commerce landscape, there is a growing demand for quickly generating visually engaging advertising images (Marwan, Harkim, and Sugiharto 2024). Traditional content creation relies on costly photoshoots and manual design, creating bottlenecks that hinder both speed and scale (Cui, Liu, and Yuan 2025; Adepoju et al. 2024). While Artificial Intelligence Generated Content (AIGC), particularly diffusion models (Ho, Jain, and Abbeel 2020; Rombach et al. 2022), promises to



Figure 1: Examples of generated advertisement images.

revolutionize this landscape, a core challenge prevents its widespread commercial deployment: how to generate high-fidelity advertising images that faithfully preserve the visual characteristics of given product images while achieving compelling visual results according to user specifications?

Existing subject-driven generation methods reveal a critical dilemma between visual fidelity and computational efficiency. On the one hand, tuning-based methods like DreamBooth (Ruiz et al. 2023) and Textual Inversion (Gal et al. 2022) achieve remarkable fidelity in preserving product details. However, their “one-model-per-subject” paradigm incurs prohibitive training and storage costs, making them impractical for e-commerce platforms managing a large number of products. On the other hand, tuning-free methods like IP-Adapter (Ye et al. 2023) and PhotoMaker (Li et al. 2024) offer the required efficiency and scalability, but they often fail to preserve product details such as the unique textures, shapes, and logos, which are essential for maintaining brand identity in advertising contexts. This fidelity-efficiency dilemma represents the primary barrier to applying AIGC techniques in advertising image generation.

To bridge this fidelity-efficiency gap, we propose RefAdGen, a novel generation method that achieves high fidelity within a tuning-free paradigm (i.e., no per-SKU test-time fine-tuning). Its core innovation is a decoupled generation

\*Corresponding author.

architecture that disentangles the complex synthesis process into two distinct tasks: spatial layout control and identity feature fusion. Spatial control is achieved by injecting a product mask at the U-Net’s input, which implicitly guides the model’s spatial awareness throughout the generation process, as opposed to applying an explicit mask during feature fusion. By handling these objectives separately, our design philosophy prevents feature entanglement, enabling the model to achieve state-of-the-art product fidelity while maintaining high efficiency (Figure 1).

However, developing and validating our method requires high-quality training and evaluation data, which is still lacking in this domain. Therefore, we introduce AdProd-100K, the first large-scale benchmark specifically designed for reference-based advertising image generation. We first collect 100,000 high-quality advertising images from various e-commerce platforms and brand websites, and then generate corresponding product images and textual scene descriptions. We further enhance the diversity of dataset with a novel dual data augmentation strategy that combines multi-view synthesis (Kerbl et al. 2023) and image degradation, which encourages models to learn intrinsic 3D product properties rather than learning a superficial 2D “copy-paste” shortcut. This dataset not only supports the training and evaluation of our proposed method but also provides researchers the basic foundation they need to tackle this important commercial challenge.

The main contributions of this work are as follows:

- We introduce AdProd-100K, the first large-scale benchmark for reference-based advertising generation, featuring a novel dual augmentation strategy to foster robust, 3D-aware learning.
- We propose RefAdGen, a framework with a novel decoupled design that combines mask-guidance at the U-Net input with an efficient Attention Fusion Module (AFM) to enable high-fidelity product preservation without per-product training.
- We conduct extensive evaluation on both AdProd-100K and in-the-wild images to demonstrate the effectiveness of RefAdGen in maintaining fidelity and generating compelling advertising images based on textual description.

## Related Work

We review controllable image generation, focusing on fine-tuning-based and tuning-free subject-driven methods, and spatial control approaches.

### Subject-Driven Generation via Fine-Tuning

Fine-tuning-based methods adapt pretrained diffusion models by updating network parameters. Textual Inversion (Gal et al. 2022) uses learned pseudo-tokens for personalized synthesis. DreamBooth (Ruiz et al. 2023) tunes the U-Net with subject-specific images. Extensions include optimization acceleration (Tewel et al. 2023), multi-concept blending (Kumari et al. 2023), and flexible tuning paradigms (Jiang et al. 2023). Domain-specific works (Shen et al. 2025d,c) achieve high realism but require subject-specific training. All rely on “one-model-per-subject”, demanding separate training per

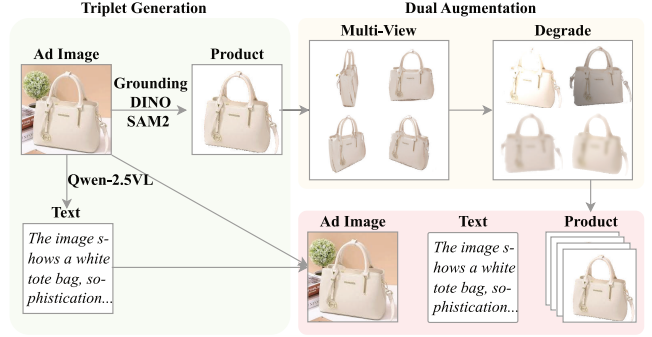


Figure 2: The construction pipeline of AdProd-100K. Starting from the advertising images, we first construct triplets by extracting the product images and textual scene description, and then enhance them with dual data augmentation.

product with prohibitive costs, unsuitable for commercial platforms.

## Tuning-Free Controllable Generation

Tuning-free methods inject subject and structure information via external conditions. IP-Adapter (Ye et al. 2023) uses a visual encoder and adapter to inject reference features. PhotoMaker (Li et al. 2024) extends this for faces, while others explore multi-subject injection (Ma et al. 2024). Although efficient, these often compromise identity preservation, failing to maintain critical attributes (material, shape, branding) for advertising. Spatial control methods inject structured priors through auxiliary encoders. ControlNet (Zhang, Rao, and Agrawala 2023) and T2I-Adapter (Mou et al. 2023) use edge maps or pose skeletons. Hierarchical approaches (Cheng et al. 2024) address layout-to-image through multi-level modeling. Fashion-domain methods (Shen and Tang 2024; Shen et al. 2025b) leverage pose or mask guidance. While excelling at structural fidelity, these do not address instance-level identity preservation. Instruction-based editing (Brooks, Holynski, and Efros 2022) guides synthesis via natural language but often overwrites subject appearance. Inspired by IMAGHarmony (Shen et al. 2025a), we adopt a decoupled design: masks injected into the U-Net for layout control, and identity features fused through lightweight attention, enabling high fidelity without fine-tuning. Our task is product-to-scene with mask layout control and identity re-injection.

## Building AdProd-100K with Dual Augmentation

The key to advertising image generation is a large collection of high-quality triplets containing textual scene descriptions, product images, and advertising images. To achieve this, we constructed AdProd-100K, a large-scale dataset designed for both training and evaluating models. Figure 2 shows our dataset construction pipeline, which consists of two main stages: triplet generation and dual augmentation.

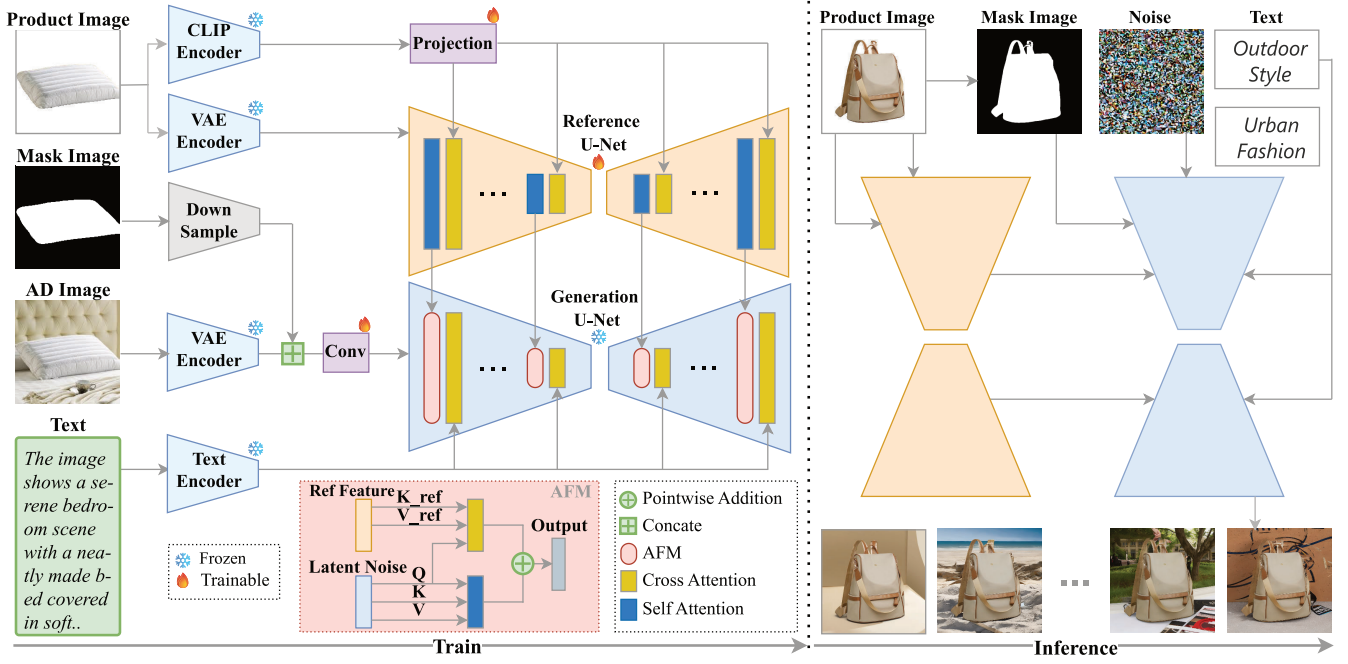


Figure 3: The overall model architecture of RefAdGen, featuring a decoupled dual U-Net design. The Generation U-Net receives the noisy latent and the product mask  $M'$  at its input for spatial control. At each level of the network, the Attention Fusion Module (AFM) fuses identity features from the Reference U-Net with the scene features of the Generation U-Net.

### Triplet Generation

Our process begins by collecting over 400,000 images from various e-commerce platforms and brand websites, spanning 30 product categories. We then used Qwen-2.5VL (Bai et al. 2025) to generate the textual scene description, and combined Grounding DINO (Liu et al. 2023) and SAM2 (Ravi et al. 2024) to obtain referenced product images, which are effective and widely used in corresponding tasks. Throughout this process, we performed a rigorous filtering process to ensure data quality, discarding images with significant watermarks, embedded text, occlusions, or low resolution. This stage results in 100,000 high-quality triplets, each containing a textual scene description, a product image, and an advertising image.

### Dual Augmentation

Directly training models on these triplets, however, leads to significant limitations and poor generalizability. First, since the product images are directly segmented from the advertising images, models tend to learn a superficial “copy-paste” shortcut. This results in unrealistic composite images and a lack of diverse viewing angles, as the model merely replicates rather than understands the products. Second, user-provided product images are inherently of varying quality due to factors like inconsistent lighting, camera shake, and sensor noise. Without training on such diverse and imperfect data, models struggle to generalize effectively to varied real-world scenarios.

To address these challenges and enhance model robustness, we designed a dual data augmentation strategy: **Multi-**

**View Image Generation.** This strategy addresses the limitations of a single, fixed camera angle. Instead of single-view 3D Gaussian Splatting, we use minor viewpoint perturbations to foster 3D-aware features. We leverage 3D Gaussian Splatting (Kerbl et al. 2023) to render novel views from a single product image. Crucially, we recognize that in practical applications, a user’s reference image is typically close to their desired output angle. Therefore, our strategy focuses on rendering novel yet similar views, rather than drastically different perspectives. Training with these subtle variations enhances the model’s 3D awareness, incentivizing it to learn the product’s underlying geometry rather than being restricted to a single, static viewpoint.

**Image Degradation.** This strategy simulates real-world imaging variations. By applying degradations like Gaussian noise, shadow variations, and minor geometric warps to the product images, models are forced to distinguish between a product’s intrinsic, invariant features (e.g., logo, material) and incidental extrinsic, photographic artifacts. This shifts the model’s focus from pixel-level replication to essential feature extraction, enabling it to better handle imperfect input images.

For each original triplet, we generated five augmented variants by modifying the product images. The mask of each product image is also extracted using SAM2 for model training. Table 1 provides a detailed variant distribution across different categories.

Category	Num	Category	Num	Category	Num
Backpack	25000	Eyeshadow	13960	Pens	14650
Bench	21855	Fork	18115	Pillows	14595
Body wash	25000	Foundation	10715	Rugs	8760
Bottle	25000	Handbag	25000	Shampoo	22990
Car	25000	Hats	36740	Snacks	24990
Cell phone	25000	Headphones	22645	Sneakers	24995
Chargers	21095	Kite	13030	Sports ball	7045
Clothing	24995	Lipstick	10585	Toothbrush	7105
Coffee	24275	Motorcycle	24995	Umbrella	15675
Cup	16860	Notebooks	24995	Wine glass	24990

Table 1: 30 categories and their sample counts.

## RefAdGen

In addition to constructing AdProd-100K, we proposed RefAdGen to generate high-fidelity advertising images without fine-tuning on each input product. The core idea is to separate the task of scene generation from product injection and then precisely merge them. Figure 3 illustrates the design of RefAdGen, which consists of two core components: 1) a dual U-Net backbone, where the Generation U-Net is spatially guided by an input product mask, and 2) an Attention Fusion Module for injecting identity features from the Reference U-Net into the Generation U-Net.

### Dual U-Net Architecture

To address the inherent conflict between diverse scene generation and high-fidelity identity preservation, we adopted a dual U-Net architecture that decouples these competing objectives into specialized processing streams. This architecture leverages two networks derived from the same pre-trained model, ensuring natural feature space compatibility that enables seamless identity-scene fusion while allowing each U-Net to excel in its dedicated task.

**Reference U-Net.** This network functions as a dedicated identity feature extractor. Its weights are fully fine-tuned during training to learn and extract the most critical features for identity preservation from the input product image. This design enables it to supply highly relevant and condensed identity information to the generation process.

**Generation U-Net.** This network serves as the primary denoising backbone for synthesizing the final advertising images based on the extracted product feature and the user-provided scene description. To incorporate precise spatial guidance, we modified its input to accept a 5-channel tensor, formed by concatenating the standard 4-channel noisy latent  $\mathbf{z}_t$  with the 1-channel product mask  $\mathbf{M}'$ . The first convolutional layer is modified accordingly to accept this extra channel. To preserve the powerful generative prior and achieve parameter-efficient adaptation, most parameters are copied from the pre-trained model and frozen. Only the modified input layer and the projection matrices within our AFM modules are made trainable.

### Attention Fusion Module

To effectively inject identity features into the scene, we employ an Attention Fusion Module (AFM) that computes self-

attention and cross-attention in parallel and then sums their outputs. Specifically, for a given query feature  $\mathbf{Q}$  from the Generation U-Net, the module’s final output  $\mathbf{O}_{\text{AFM}}$  combines 1) a self-attention term for scene structure, calculated using the Generation U-Net’s own key ( $\mathbf{K}$ ) and value ( $\mathbf{V}$ ); and a cross-attention term for identity injection, calculated using the key ( $\mathbf{K}_{\text{ref}}$ ) and value ( $\mathbf{V}_{\text{ref}}$ ) from the Reference U-Net. The entire operation is captured by:

$$\mathbf{O}_{\text{AFM}} = \text{softmax} \left( \frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}} \right) \mathbf{V} + \text{softmax} \left( \frac{\mathbf{Q}\mathbf{K}_{\text{ref}}^T}{\sqrt{d_k}} \right) \mathbf{V}_{\text{ref}}, \quad (1)$$

where  $d_k$  is the dimension of the key vectors.

**Implicit Spatial Control Mechanism.** To enable 3D-aware generation with novel viewpoints, our method injects masks at the input of Generation U-Net rather than applying rigid feature-level masking at each level. By providing the mask as an early spatial prior, we allow it to naturally propagate through the network, creating spatially-aware query vectors ( $\mathbf{Q}$ ) that adaptively attend to relevant identity features based on spatial context. This implicit mechanism delivers two benefits: it maintains a precise spatial layout while encouraging the model to learn intrinsic 3D object representations rather than memorizing static 2D silhouettes. Consequently, RefAdGen achieves geometrically coherent synthesis across novel viewpoints.

### Training Objective

We train RefAdGen end-to-end using the standard noise prediction objective from latent diffusion models. The model  $\epsilon_\theta$ , represented by our AFM-integrated U-Net, is trained to predict the sampled noise  $\epsilon$  from the noisy latent  $\mathbf{z}_t$ . The loss function  $\mathcal{L}$  is the mean squared error between the predicted and sampled noise, conditioned on the text prompt  $\mathbf{c}_{\text{text}}$ , the spatial product mask  $\mathbf{M}$ , and the feature  $\mathbf{z}_{\text{ref}}$ , which is encoded from the product image by our Reference U-Net.

$$\mathcal{L} = \mathbb{E}_{\mathbf{z}_0, \mathbf{c}_{\text{text}}, \mathbf{z}_{\text{ref}}, \mathbf{M}, \epsilon, t} \left[ \|\epsilon - \epsilon_\theta(\text{concat}(\mathbf{z}_t, \mathbf{M}'), t, \mathbf{c}_{\text{text}}, \mathbf{z}_{\text{ref}})\|_2^2 \right]. \quad (2)$$

## Experiments

This section systematically validates our proposed framework, RefAdGen, through a series of comprehensive experiments conducted on our AdProd-100K. In addition to numerical evaluation, we further tested our method on a diverse set of in-the-wild photographs captured with mobile phones, confirming its robustness and broad applicability. A user study is also conducted to assess human perception of our generated results.

### 5.1 Experimental Setup

**Dataset and Metrics.** Our experiments were conducted on the AdProd-100K dataset, using a 9:1 training-to-testing split. We evaluated performance on five metrics across several key dimensions: CLIP-Score (Hessel et al. 2021) for text adherence, FID (Seitzer 2020) for realism, ImageReward (Xu et al. 2023) for human preference, and MP-LPIPS (Chen et al. 2024) and LPIPS (Zhang et al. 2018)



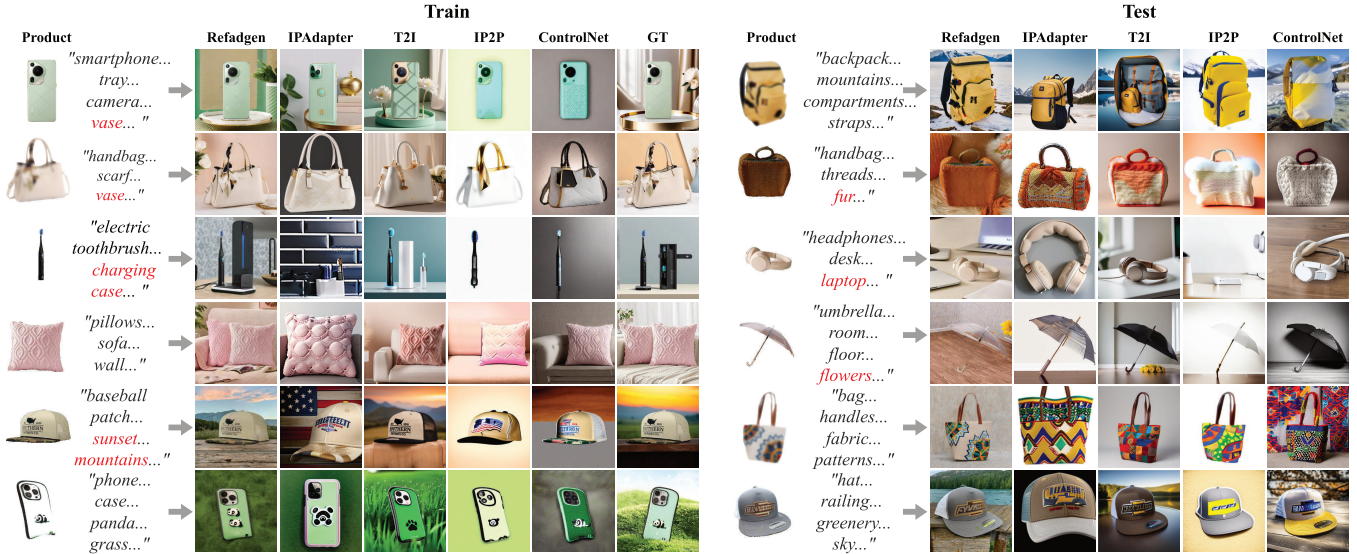


Figure 4: Qualitative comparisons on AdProd-100K. Prompts are simplified for clarity. Both the training samples on the left and the test samples on the right showcase the consistent advantages of RefAdGen in identity consistency, scene realism, and overall aesthetic quality.



Figure 5: Qualitative comparison on in-the-wild images. The inputs are real-world photos captured using mobile phones, from which our pipeline automatically segments the product image and generates advertising images using RefAdGen.

Hyperparameter	Value	Hyperparameter	Value
Optimizer	AdamW	Weight Decay	0.01
Batch Size	3	Learning Rate	$1 \times 10^{-5}$
Noise Offset	0.05	LR Scheduler	Linear
Training Epochs	8	Warmup Steps	500

Table 2: Hyperparameter configuration for the model.

for perceptual similarity and identity preservation. We also evaluated on in-the-wild images to assess generalization.

**Baselines.** We compared RefAdGen against state-of-the-art methods representing diverse control strategies: IP-

Adapter (Ye et al. 2023) that uses identity as reference, ControlNet (Zhang, Rao, and Agrawala 2023) and T2I-Adapter (Mou et al. 2023) that uses structure information as reference, and Instruct-Pix2Pix (IP2P) (Brooks, Holynski, and Efros 2022) that performs instruction-guided image editing. To ensure a fair comparison, all baselines and our model are built upon the Stable Diffusion v1.5 backbone.

**Implementation Details.** Our model was trained on two NVIDIA 5090D GPUs. All input images were resized to  $512 \times 640$ . Detailed hyperparameter settings are provided in Table 2. At inference time, our method generates a single image in approximately 1 second on an NVIDIA 5090D GPU.

## 5.2 Main Results and Analysis

**Quantitative Comparison.** The quantitative results in Table 3 demonstrate the effectiveness of our RefAdGen. A central challenge is maintaining perceptual similarity while ensuring faithful text adherence. Methods like T2I perform well in text adherence (34.27 CLIP-Score) but exhibit significant weaknesses at preserving the fine-grained details (0.3382 MP-LPIPS). Conversely, identity-reference methods like IP-Adapter compromise on text adherence and overall realism in favor of perceptual similarity. In contrast, our method achieves state-of-the-art performance in both dimensions, scoring highest on CLIP-Score (34.5106), MP-LPIPS (0.2612), and LPIPS (0.5487). This demonstrates its robust ability to integrate the reference product seamlessly into a novel scene described by the text prompt.

In addition, this high-fidelity generation does not compromise overall image quality or aesthetic appeal. RefAdGen achieves the best realism with an FID score of 50.58, which indicates that our generated scenes are not only accurate but

Model	CLIP Score $\uparrow$	FID $\downarrow$	ImageReward $\uparrow$	MP-LPIPS $\downarrow$	LPIPS $\downarrow$
IP-Adapter	32.6308	62.6666	-0.2572	0.3974	0.7159
T2I	34.2737	59.1770	0.1777	0.3382	0.6063
IP2P	32.4433	64.1141	-0.5842	0.3517	0.6336
ControlNet	33.0226	57.1835	-0.3668	0.3748	0.6578
RefAdGen (Ours)	<b>34.5106</b>	<b>50.5843</b>	<b>0.2391</b>	<b>0.2612</b>	<b>0.5487</b>

Table 3: Performance comparison of our model (RefAdGen) against several baselines on five key metrics. The arrow indicates whether a higher value ( $\uparrow$ ) or a lower value ( $\downarrow$ ) is better. The **best** result in each column is highlighted in bold.



Figure 6: Qualitative ablation analysis of architectural components. The Full Model better preserves details, while the ablation variant without U-Net input masks exhibits degradation in object placement precision and scene coherence.

also visually harmonious. Our model also scores highest on ImageReward (0.2391), which suggests that our results possess strong commercial appeal for advertising applications. **Qualitative Comparison.** Qualitative analysis on AdProd-100K (Figure 4) shows that RefAdGen generates novel images without overfitting (training set) and maintains high fidelity for unseen items (test set), demonstrating strong generalization capabilities.

To further assess robustness and practical utility, we evaluated model performance on challenging in-the-wild images (Figure 5). For these mobile phone-captured images, we first extracted product images using the same method employed in our data generation pipeline, and then generated advertising images using RefAdGen. The results show that RefAdGen not only handles this domain shift successfully but also exhibits remarkable lighting adaptability, rendering objects with physically plausible illumination that seamlessly integrates with the target scene context.

### Ablation Studies

To systematically dissect the contributions of our design choices, we conducted a comprehensive ablation study covering two key aspects: model architectural components and data augmentation strategies.

**Analysis of Model Architectural Components.** We first analyze the contributions of our product mask component. As shown in Table 4, removing mask guidance from the



Figure 7: Qualitative ablation analysis of the Dual Augmentation strategy. With augmentation, the generation results exhibit substantially enhanced visual clarity and improved robustness compared to the non-augmented baseline.

generation U-Net input results in a significant drop in performance, particularly for realism (FID) and identity preservation (MP-LPIPS). This provides empirical justification for our model design: injecting the mask at the input layer provides essential spatial bias, enabling the network to learn selective application of identity features. Without this spatial guidance, the model fails to perform implicit feature gating, resulting in feature bleeding and consequent degradation in generation quality.

**The Foundational Role of Dual Augmentation.** Next, we ablate our data generation method. As shown in Table 4, removing dual data augmentation results in a catastrophic drop in performance, with the FID score increasing dramatically from 50.58 to 68.72. This suggests that without our augmentation strategy, the model cannot effectively generalize to create realistic, well-integrated scenes and instead resorts to superficial “copy-paste” behavior from training data.

Figure 7 provides the evidence supporting these quantitative findings. Across all test cases, the model trained with our augmentation produces clearer, more realistic, and more robust results. It also successfully handles varied lighting, generates novel viewpoints (e.g., the cup and sneakers), and reconstructs details from imperfect inputs. In contrast, the model trained without augmentation learns a superficial “copy-paste” of the input.





Figure 8: Image generation fidelity improves markedly and remains consistently high for all  $\lambda$  values above 0.7.

Configuration	CLIP Score $\uparrow$	FID $\downarrow$	ImageReward $\uparrow$	MP-LPIPS $\downarrow$	LPIPS $\downarrow$
<b>Full Model (Ours)</b>	<b>34.5106</b>	<b>50.5843</b>	<b>0.2391</b>	<b>0.2612</b>	<b>0.5487</b>
w/o Masks	33.2415	55.0244	-0.2293	0.3638	0.6602
w/o Dual Augmentation	32.7952	68.7243	-0.3554	0.3014	0.6064

Table 4: Ablation study of our framework’s core components and data strategy. “Full Model” represents our complete design, while each subsequent row ablates one key element. The results highlight the critical contributions of each component. The **best** score in each column is highlighted in bold.

	RefAdGen	ControlNet	IP2P	IP-Adapter	T2I
J2b	<b>38.70</b>	23.60	8.40	0.90	28.40
G2R	<b>90.70</b>	77.30	47.30	6.70	78.00

Table 5: Results of our user study comparing RefAdGen with baseline methods. The **best** score is highlighted in bold.

## User Study

To assess human perception, we also conducted a subjective study with 30 volunteers using two protocols. For Realism Judgment (G2R), participants were shown individual images and asked to distinguish whether they were real photographs or AI-generated, thereby measuring perceived authenticity. For Preference Selection (J2b), we presented the image from our method alongside the images from all baselines simultaneously, and asked users to select the single best result based on realism and aesthetic quality. As shown in Table 5, our method significantly outperforms all baselines on both metrics. This indicates that images generated by RefAdGen are not only highly realistic but are also more aesthetically favored by users.

## 5.5 Further Analysis and Discussion

**Robustness to Fusion Strength.** Balancing multiple guidance signals often requires careful hyperparameter calibration. We thus tested our identity fusion’s sensitivity to a trade-off parameter,  $\lambda$ , in the fusion mechanism ( $\mathbf{O}_{AFM} = \mathbf{O}_{self} + \lambda \cdot \mathbf{O}_{cross}$ ). As shown in Figure 8, our method is remarkably stable, maintaining consistently high-quality results where  $\lambda \geq 0.7$ . This robustness stems from our decoupled architecture. The spatial guidance provided by the input mask enables the model to apply identity features only within the intended foreground, mitigating feature conflicts that would otherwise necessitate a trade-off. This spatial disentanglement eliminates the criticality of  $\lambda$ , allowing us to set a default value of 1.0.

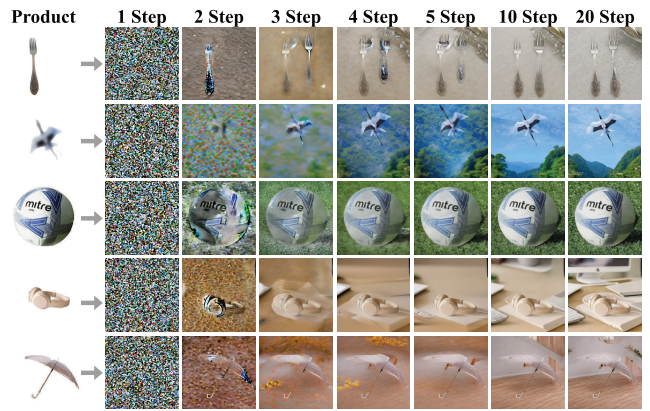


Figure 9: Visualization of the denoising process.

**Analysis of Generation Process.** The progressive denoising process, illustrated in Figure 9, demonstrates the computational efficiency of our method. The overall structure emerges within the first 3 steps, with fine-grained details being refined from step 10. High-quality results are achieved by approximately step 20, demonstrating a good balance between computational efficiency and output quality.

## Conclusion and Future Work

This paper addresses the fidelity-efficiency trade-off in reference-based generation. We introduce AdProd-100K, a large-scale benchmark for this task, and RefAdGen, a tuning-free framework with a decoupled design. It achieves state-of-the-art identity preservation through the synergy of its mask-guidance at the U-Net input and an efficient Attention Fusion Module (AFM).

**Future Work.** We aim to extend our framework to video ad generation and to build a closed-loop system that optimizes content directly for business metrics, such as conversion rates.

## Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 6250070266.

## References

- Adepoju, A. H.; Eweje, A.; Collins, A.; and Austin-Gabriel, B. 2024. Automated offer creation pipelines: An innovative approach to improving publishing timelines in digital media platforms. *International Journal of Multidisciplinary Research and Growth Evaluation*, 5(6): 1475–1489.
- Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.
- Brooks, T.; Holynski, A.; and Efros, A. A. 2022. Instruct-Pix2Pix: Learning to Follow Image Editing Instructions. *arXiv preprint arXiv:2211.09800*.
- Chen, W.; Gu, T.; Xu, Y.; and Chen, C. 2024. Magic Clothing: Controllable Garment-Driven Image Synthesis. *arXiv preprint arXiv:2404.09512*.
- Cheng, B.; Ma, Y.; Wu, L.; Liu, S.; Ma, A.; Wu, X.; Leng, D.; and Yin, Y. 2024. HiCo: hierarchical controllable diffusion model for layout-to-image generation. In *Advances in Neural Information Processing Systems*.
- Cui, W.; Liu, M. J.; and Yuan, R. 2025. Exploring the Integration of Generative AI in Advertising Agencies: A Co-Creative Process Model for Human-AI Collaboration. *Journal of Advertising Research*, 1–23.
- Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.
- Hessel, J.; Holtzman, A.; Forbes, M.; Bras, R. L.; and Choi, Y. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *EMNLP*.
- Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, 6840–6851.
- Jiang, Z.; Mao, C.; Huang, Z.; Ma, A.; Lv, Y.; Shen, Y.; Zhao, D.; and Zhou, J. 2023. Res-Tuning: A Flexible and Efficient Tuning Paradigm via Unbinding Tuner from Backbone. In *Advances in Neural Information Processing Systems*.
- Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (TOG)*, 42(4): 1–14.
- Kumari, N.; Zhang, B.; Zhang, R.; Shechtman, E.; and Zhu, J.-Y. 2023. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1931–1941.
- Li, Z.; Cao, M.; Wang, X.; Qi, Z.; Cheng, M.-M.; and Shan, Y. 2024. PhotoMaker: Customizing Realistic Human Photos via Stacked ID Embedding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; et al. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.
- Ma, J.; Liang, J.; Chen, C.; and Lu, H. 2024. Subject-Diffusion: Open Domain Personalized Text-to-Image Generation without Test-time Fine-tuning. In *ACM SIGGRAPH 2024 Conference Papers*, SIGGRAPH '24. Association for Computing Machinery. ISBN 9798400705250.
- Marwan, A.; Harkim, H.; and Sugiharto, B. 2024. The Impact of Visual Marketing on Purchasing Behavior in E-Commerce: A Case Study in The Fashion Industry. *Golden Ratio of Data in Summary*, 4(2): 1022–1031.
- Mou, C.; Wang, X.; Xie, L.; Wu, Y.; Zhang, J.; Qi, Z.; Shan, Y.; and Qie, X. 2023. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*.
- Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; Mintun, E.; Pan, J.; Alwala, K. V.; Carion, N.; Wu, C.-Y.; Girshick, R.; Dollár, P.; and Feichtenhofer, C. 2024. SAM 2: Segment Anything in Images and Videos. *arXiv preprint arXiv:2408.00714*.
- Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.
- Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 22500–22510.
- Seitzer, M. 2020. pytorch-fid: FID Score for PyTorch. <https://github.com/mseitzer/pytorch-fid>. Version 0.3.0.
- Shen, F.; Du, X.; Gao, Y.; Yu, J.; Cao, Y.; Lei, X.; and Tang, J. 2025a. IMAGHarmony: Controllable Image Editing with Consistent Object Quantity and Layout. *arXiv preprint arXiv:2506.01949*.
- Shen, F.; Jiang, X.; He, X.; Ye, H.; Wang, C.; Du, X.; Li, Z.; and Tang, J. 2025b. Imagdressing-v1: Customizable virtual dressing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 6795–6804.
- Shen, F.; and Tang, J. 2024. Imapose: A unified conditional framework for pose-guided person generation. *Advances in neural information processing systems*, 37: 6246–6266.
- Shen, F.; Wang, C.; Gao, J.; Guo, Q.; Dang, J.; Tang, J.; and Chua, T.-S. 2025c. Long-Term TalkingFace Generation via Motion-Prior Conditional Diffusion Model. *arXiv preprint arXiv:2502.09533*.
- Shen, F.; Yu, J.; Wang, C.; Jiang, X.; Du, X.; and Tang, J. 2025d. IMAGGarment-1: Fine-Grained Garment Generation for Controllable Fashion Design. *arXiv preprint arXiv:2504.13176*.
- Tewel, Y.; Gal, R.; Chechik, G.; and Atzmon, Y. 2023. Key-Locked Rank One Editing for Text-to-Image Personalization. *ACM SIGGRAPH 2023 Conference Proceedings*.



Xu, J.; Liu, X.; Wu, Y.; Tong, Y.; Li, Q.; Ding, M.; Tang, J.; and Dong, Y. 2023. ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation. *arXiv:2304.05977*.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3836–3847.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*.