From Prompt to Production: Automating Brand-Safe Marketing Imagery with Text-to-Image Models

Parmida Atighehchian Henry Wang Andrei Kapustin Boris Lerner Tiancheng Jiang
Taylor Jensen Negin Sokhandan
Amazon Web Services
{patigheh, yuanhenw, kapustan, blerner, tjiangg, tdjensen, ngnsl}@amazon.com

Abstract

Text-to-image models have made significant strides, producing impressive results in generating images from textual descriptions. However, creating a scalable pipeline for deploying these models in production remains a challenge. Achieving the right balance between automation and human feedback is critical to maintain both scale and quality. While automation can handle large volumes, human oversight is still an essential component to ensure that the generated images meet the desired standards and are aligned with the creative vision. This paper presents a new pipeline that offers a fully automated, scalable solution for generating marketing images of commercial products using text-to-image models. The proposed system maintains the quality and fidelity of images, while also introducing sufficient creative variation to adhere to marketing guidelines. By streamlining this process, we ensure a seamless blend of efficiency and human oversight, achieving a 30.77% *increase in marketing object fidelity using DINOV2 and a* 52.00% *increase in human preference over the generated outcome.*

1. Introduction

The rapid advancement of generative artificial intelligence has transformed content creation workflows across industries, with marketing and advertising emerging as particularly active domains [10, 12]. Recent breakthroughs in text-to-image generation models have enabled high-quality visual creation from natural language descriptions, leading to widespread experimentation in commercial applications [4, 8, 10, 12, 13]. Despite these advances, integrating such models into production marketing workflows remains challenging, limiting their practical deployment at scale.

The transition from experimental AI-generated content to production-ready marketing materials reveals critical gaps in current approaches. While text-to-image models excel at producing visually appealing content, they struggle to consistently meet the stringent requirements of professional marketing applications of a commercial product (like a shoe), where adherence to brand guidelines, product fidelity, and scalable workflows is paramount. Here, brand guidelines mean accurate representation of product-specific features (e.g., colorways, logos, stitching). Efforts to address these limitations through architectural innovations such as ControlNet [21], adapters [5], and IP-Adapters [19] provide improved control but fall short of delivering the comprehensive solution required for enterprise use.

We identify three fundamental challenges preventing current generative AI systems from achieving production-grade marketing content generation:

Alignment with brand guidelines: Marketing content must adhere to brand specifications (e.g., colors, background features) while preserving product characteristics without alteration. The inherent variability of generative models can compromise brand consistency and product accuracy. Control mechanisms such as ControlNet [21] and fine-tuning methods like AnyDoor [2] offer structural guidance. However, they cannot guarantee the pixel-level fidelity required for marketing campaigns, where even small deviations can compromise brand integrity. Outpainting models aim to maintain reference object fidelity, but common failure cases include duplicating objects, hallucinations (i.e. generating spurious elements), or ignoring the object entirely.

Predictable outcome and quality consistency: A common solution for campaign creation is to incorporate outpainting. However, seamlessly integrating products into scenes remains challenging and varies significantly based on product and scene characteristics. Skilled creators can iteratively refine generations using visual heuristics, prompt tuning, and trial and error, but creators rarely achieve quality outcomes on the first attempt.

Scalability and workflow integration: Production workflows demand consistent, high-throughput content

generation with minimal human oversight. Current approaches require repeated manual iteration, prompt engineering, and quality checks, creating bottlenecks that hinder scalability. Reliance on constant human input makes these systems impractical for enterprise campaigns that require hundreds of variations.

To address these challenges, we present a novel end-to-end automated system for generating composite marketing images that respect brand guidelines while enabling scalable and robust workflows. Our approach creates a production-ready workflow by deferring human validation to the final stage and automating all intermediate generations and quality assessments through an agentic pipeline.

Our system introduces several key innovations: (1) a structured prompt decomposition mechanism that converts complex marketing requirements into machine-readable specifications, (2) an intelligent asset retrieval system with specialized guardrails for content types, (3) a rich caption generator and a multi-modal composition planer that determines optimal product placement and scaling, (4) a grid search approach for marketing image variants, and (5) an intelligent quality control pipeline that evaluates content across multiple dimensions without human intervention. By treating the core generation module as plug-and-play, the system adapts to diverse industry needs at scale.

Section 3 details each pipeline module. Our experiments show that the proposed approach achieves significant gains in generation quality compared to existing methods, while maintaining strict adherence to brand guidelines and composition standards. The system handles diverse marketing scenarios and provides the scalability required for enterprise deployment.

2. Related Work

While various research efforts have explored improving control over text-to-image generation, to our knowledge, no prior work has proposed a scalable workflow with comprehensive testing for production-ready deployment. We therefore review the most relevant approaches to object compositing (the core component of our system), particularly in the context of marketing applications.

2.1. Fine-tuned Object Insertion and Harmonization

ICLight [22] is a diffusion-based framework for illumination harmonization and editing that operates on uncurated, in-the-wild data. ICLight enforces light transport consistency, enabling physically plausible relighting and compositional edits. However, ICLight can alter product textures and colors, creating a challenge for brand-sensitive marketing content. AnyDoor [2] enables zero-shot object-level customization by teleporting a target object into a new scene at a user-specified location without per-object fine-tuning. It balances identity preservation with local variations in lighting and pose. ObjectStitch [16] similarly performs object insertion using conditional diffusion, with adjustments such as geometry correction, color harmonization, and shadow synthesis. Despite these capabilities, AnyDoor and ObjectStitch struggle to maintain strict object fidelity, limiting their reliability in marketing scenarios.

2.2. Training-free Insertion

Insert Diffusion [6] proposes a training-free approach for inserting objects into scenes. However, the method requires manual positioning and scaling, and a second diffusion pass, which can cause unwanted pixel-level manipulation.

2.3. Outpainting

Inpaint Anything [20] introduces a modular inpainting framework built around a "clicking and filling" paradigm: objects can be removed, replaced, or preserved while modifying the background based on user prompts. While flexible, the method relies on manual object selection, preventing full automation. More broadly, outpainting models often hallucinate artifacts near objects or replicate them unnaturally within the scene, further complicating their use in marketing image generation.

Overall, while prior methods advance object compositing and harmonization, they fall short of generating brand-consistent, production-ready marketing content without extensive manual refinement. Section 3 details how our pipeline overcomes these limitations through systematic generation and artifact filtering.

3. Method

We present an automated system for generating high-quality composite marketing images from natural language prompts. Our approach translates marketing requirements into visually compelling product compositions via a four-stage pipeline: structured prompt analysis, intelligent asset retrieval, composition planning and variation generation, and multi-modal quality assessment.

3.1. System Overview

Given a marketing prompt such as *"Shoe on the floor on an urban street at sunset"*, the system automatically: (1) decomposes the prompt into structured components, (2) retrieves relevant visual assets, (3) determines optimal composition parameters using a caption generator and a multi-modal advisor, generates multiple image variations, and (4) performs comprehensive quality filtering to select the top-k final images. This automation minimizes manual intervention while maintaining professional-grade quality (Figure 1.)

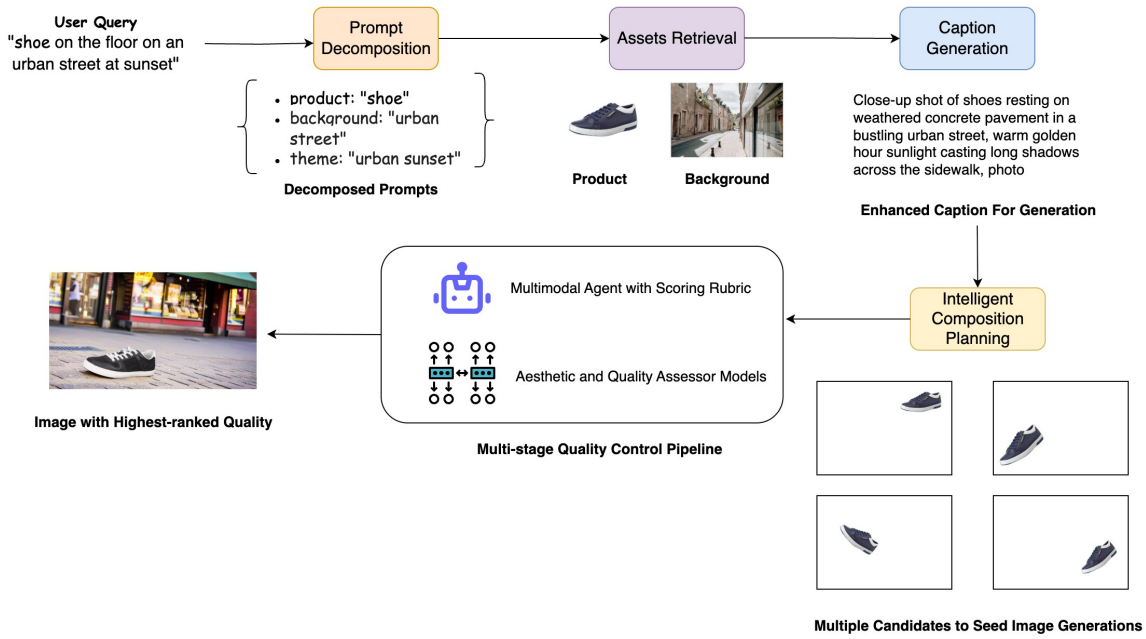

Figure 1. Overview of our pipeline. The pipeline starts with decomposing the user query. It then retrieves relevant visual assets, and rewrites the initial prompts for optimized image generation quality. The visual assets are scaled and positioned at various positions of the canvas to maximize generation diversity. After generation, candidate images are filtered and ranked based on rubric scores and aesthetic scores.

3.2. Structured Prompt Decomposition

Complex marketing prompts often contain multiple semantic elements and implicit requirements. We address this by converting natural language prompts into machine-readable specifications using a large language model. Our decomposition module extracts three key components: (1) **primary product** - the main item being promoted, (2) **background elements** - environmental context and scene descriptors, (3) **theme** - the promotional context or occasion. This structured representation ensures all specified elements are considered during asset retrieval and composition planning. The decomposition process uses few-shot prompting with predefined output schemas to maintain consistency across diverse input formats.

3.3. Multi-modal Asset Retrieval with Guardrails

Precise asset selection is critical for composite image generation. We employ a two-tier retrieval system with specialized guardrails for different asset types.

3.3.1. Product Asset Retrieval

Product assets are retrieved using similarity-based dense vector representations. Candidates are filtered by a similarity threshold $\tau_p = 0.39$, empirically optimized to balance precision and recall, ensuring visually and semantically appropriate products.

3.3.2. Background Asset Retrieval with LLM Validation

Background selection is more complex due to subjective scene appropriateness. While similarity scores provide an initial ranking, they often miss nuanced contextual requirements. After retrieving the top-k most relevant matches, we employ an LLM-based validation step that evaluates semantic alignment between each candidate background B_i and prompt P:

$$R(B_i, P) = \text{LLM}_{\text{validator}}(B_i, P) \in 0, 1 \tag{1}$$

This step ensures robust selection by considering context, theme, and visual compatibility.

3.4. Caption Generator

The caption generator is an LLM agent prompted to produce descriptive captions that guide image generation and composition planning. It considers the background, primary product, and theme to emphasize the product while positioning other elements appropriately. If a relevant background is found, the caption generator uses the image to enrich the text with additional visual features. Otherwise, it relies on the user query. This caption not only is used for image generation but also plays an important role in the composition planning stage.

3.5. Intelligent Composition Planning

Our composition module determines optimal object placement, scaling, and spatial relationships using multi-modal analysis and grid search.

3.5.1. Multi-modal Composition Analysis

A critical design decision is the ability to continue with the image generation even if a background match is not found. Using either an empty canvas or a resized background, the multi-modal advisor leverages the generated caption to recommend scaling factors for the product in the scene:

Given a background-filled or an empty canvas C, the product object O, and the generated caption Ca our composition advisor analyzes both visual and textual elements to determine optimal scaling factors:

$$S = \{s_w, s_h\} = f_{\text{advisor}}(C, O, Ca) \tag{2}$$

where s_w and s_h represent the recommended width and height scaling factors relative to the canvas dimensions.

3.5.2. Variation Generation Strategy

To maximize creative diversity, products are systematically placed across canvas regions (left, center, right thirds) with rotation angles $\theta \in \{0°, 15°, 345°\}$. This ensures broad exploration of the composition space while maintaining visual coherence.

3.5.3. Scene Generation

For marketing campaign generation, maintaining product fidelity is crucial. Our scene generation module is designed as a plug-and-play component that works with any image generation model accepting reference images and captions. For marketing applications, the generative model must preserve product fidelity while properly integrating objects into scenes. Given these base performance requirements, any outpainting, insertion and harmonization, or object composition model is suitable.

This module takes a product-on-canvas composition and extend the background while preserving visual integrity. Although the original background cannot be maintained, our feature-rich caption generation injects key background characteristics back into the scene, enabling better product composition. This approach ensures seamless product integration with realistic lighting and perspective relationships.

The image variants encourage the model to sample diversely from its learned latent space. This systematic exploration of the latent space increases the probability of generating high-quality marketing images by leveraging the model's full generative capacity.

3.6. Multi-Stage Quality Control Pipeline

Quality assessment evaluates semantic accuracy, visual realism, and aesthetic appeal. Our pipeline combines rule-based validation with learned metrics and is fully customizable for different use cases.

3.6.1. Multi-modal Quality Controller

Our primary quality filter employs a multi-modal Large Language Model that evaluates generated images across four primary critical dimensions:

1. **Caption Alignment**: Verification that all elements specified in the prompt are accurately depicted in the generated image.
2. **Product Uniqueness**: Confirmation that the scene contains exactly one instance of the product—a safeguard against common model hallucinations (e.g., unintended duplicates)
3. **Physical Realism**: Assessment of object placement consistency with real-world physics and spatial relationships.
4. **Lighting Consistency**: Evaluation of shadow and reflection accuracy relative to the scene's lighting conditions.

Each criterion receives a binary score $c_i \in \{0, 1\}$. In most production setups, the overall score is computed as below:

$$G = \prod_{i=1}^{4} c_i \tag{3}$$

This multiplicative formulation ensures that images failing any critical criterion are filtered out, maintaining high quality standards. Where data is sufficient, we recommend using a weighted average, where the weights are optimized to the use case in hand. In Section 4, we use a hierarchical approach to relax the less critical guardrails for our setup. Since the generative model is probabilistic, this component is a way to trigger a retry if necessary in a production setup and have a safe fall back.

3.6.2. Aesthetic and Semantic Quality Assessment

For critical applications such as marketing image creation, we recommend maintaining human oversight while minimizing intervention to enable scalable deployment. To facilitate scaling, the number of final images presented for human review is configurable through a hyperparameter k. When more than k candidates pass the guardrail evaluation, they undergo additional quality assessment using complementary metrics:

Aesthetic Scoring: We employ a pre-trained aesthetic quality model [14] that processes CLIP [9] visual features to predict aesthetic scores. Images below the set aesthetic threshold are filtered out.

CLIP-based Semantic Alignment: We optionally measure text-image alignment using CLIP similarity between generated images and prompt descriptions as an additional filter. The CLIP [9] score is computed as:

$$\text{CLIP}(I, T) = w \cdot \max(0, \cos(\phi(I), \psi(T))) \quad (4)$$

where $\phi(I)$ and $\psi(T)$ are the CLIP encodings of image I and text T, respectively, and $w = 2.5$ is a scaling factor.

To reduce the cost overhead and latency, we chose to incorporate metrics that need minimal computation power. However, each of the modules are fully customizable and can be adapted to different sets of requirements.

3.6.3. Combined Quality Ranking

Final image ranking combines aesthetic and semantic quality scores:

$$Q_{\text{final}} = \alpha \cdot Q_{\text{aesthetic}} + \beta \cdot Q_{\text{CLIP}} \quad (5)$$

where α and β are weighting parameters that can be adjusted based on application requirements. This multi-faceted approach ensures that selected images excel across both perceptual quality and semantic accuracy dimensions.

The complete pipeline processes multiple image variations and returns only those meeting all quality criteria, ranked by their combined quality scores.

4. Experiments

4.1. Baseline

We select baseline candidates based on their ability to perform outpainting and harmonization using reference images and prompts. Below is the list of our selected candidates:

- **ICLight [22]**: ICLight represents state-of-the-art object-scene compositing with specialized fine-tuning for lighting and shadow adjustment. Unlike standard outpainting models that focus solely on semantic alignment, ICLight's illumination capabilities enable more natural product integration.
- **Inpaint Anything [20]**: Replace Anything module from the Inpaint Anything framework segments the object of interest from the background. It then replaces the background with a text-guided inpainting model. This model requires an input to select the object for segmentation (a "clicking and filling" paradigm). In our image variants, the object location is determined. We use this prior placement knowledge to automate the object selection.
- **Amazon Nova Canvas [1]**: Nova Canvas receives the product on an empty canvas. Additionally, it has capabilities to accept a mask image or a mask prompt to mask out the product and perform text guided outpainting on the scene. For simplicity, we use the mask-prompt in our experiments.

We compare our approach to these baselines along two verticals critical to our application domain: (1) Product Fidelity and (2) Scene Integration.

4.2. Data

We use the Amazon-Berkeley Objects (ABO) dataset [3] as the primary source for product images. The ABO dataset contains high-quality, multi-view images of real-world products across diverse categories with annotations. These product images are clean and have a white background which eliminates the need for any preprocessing. We select 21 distinct categories, ranging from clothing and accessories to household items to cover a diverse range of items. We sample a minimum of 10 images per category if there are a sufficient number of images to sample from. If the number of images in a category is less than 10, we include all of them in our test dataset. Our final sampled dataset consists of 206 images representing 21 different categories.

To further increase our test sample size, we send each product type to a Visual Language Model (VLM) and ask the model to generate 5 rich, diverse and creative captions for scenes that the product can fit in. With 5 captions per product, we have a total of 1030 product-caption pairs for our experiments.

4.3. Experiment setup

We evaluate all three baselines on the same set of 1,030 product–caption pairs. Each method requires a mask image specifying product placement. In real-world workflows, designers typically adjust product position and rotation through repeated trials until a visually coherent result is achieved. Our pipeline automates this process by placing the product at multiple positions and rotations on an empty canvas, enabling the model to generate diverse variations in a single batch. This increases the likelihood of obtaining a high-quality composition without manual intervention.

To ensure fair comparison, we restrict each baseline to a single fixed placement–rotation configuration, which represents a non-automated workflow where no iterative adjustment is performed. By contrasting one-shot baseline outputs with our multi-variation pipeline, we directly test the contribution of automated variation generation.

For pipeline evaluation, we initialize from ground-truth product–caption pairs, bypassing retrieval and captioning, and begin at the composition stage. The setup follows Section 3.5, using an empty canvas. The final pipeline outputs are selected through our quality assessment module. In this experiment, we follow algorithm 1 to progressively relax the lower-priority rules if needed, while keeping rules 2 and 3 strict (see Section 3.6 for the list of rules).

Figure 2 shows a comparison of our pipeline generated images against the baseline models' generations (additional results can be found in Supplementary Material section 1.1, Figures 1, 2, and 3).

Algorithm 1 Find images aligned with guidelines.

Input: A list of lists containing 4 binary $scores$
Output: A list of image indices matching the one of the best pattern, or $\emptyset$
$patterns \leftarrow [[1,1,1,1],[0,1,1,1],[0,1,1,0]]$
for each $pattern$ in $patterns$ **do**
    $indices \leftarrow \{i \mid scores[i] = pattern\}$
    if $indices \neq \emptyset$ **then**
        return $indices$
    end if
end for
return $\emptyset$

4.4. Quantitative Results

We use the following metrics to evaluate performance across the three key axes:

- **Product Fidelity**: We use DINOv2 [7] features and MS-SSIM [17] to evaluate how well the original product is preserved in the generated image. To do so, one needs to isolate the generated product from the scene. We accomplish this by using Amazon Nova Pro [1] as an image analyzer agent to predict bounding box coordinates around the product, which are then passed with the image to SAM2 [11] for segmentation. We then compare the extracted object with the ground-truth product using the cosine similarity of DINOv2 features and calculated MS-SSIM values. DINOv2 similarity score captures both semantic and fine-grained visual similarities, while MS-SSIM provides a strong signal of structural fidelity, detecting distortions or visual degradation. Combined, these two metrics offer a robust measure of product fidelity in the generated scenes.
- **Overall text to image alignment**: We compute the HPSv2 [18] to measure the overall quality of the image and alignment with text. While not a direct measure of quality for images, HPSv2 correlates strongly with human visual preference.

Tables 1 report results for each baseline and their integration into our pipeline. Across all models, our pipeline consistently improves both structural similarity and DINOv2 scores, supporting our claim that object size and placement play a critical role in outpainting quality. Although we observe a slight drop in HPSv2 for ICLight, our human evaluation indicates a strong preference for pipeline-generated images. This drop is partly explained by our pipeline's design: when no images pass the quality check, the output defaults to a score of 0. This effect is more pronounced in models like ICLight, where the baseline often alters an object's texture or color. In such cases, the baseline still receives a score for its output, even if it deviates from fidelity requirements, while the pipeline registers no result (see Limitations 4.6). Consequently, more test cases yield zero outputs on the pipeline side, lowering the average HPSv2. In contrast, we observe a slight improvement in HPSv2 for Replace Anything. This improvement stems from selecting more effective product placement and rotation, which enables better integration. The baseline Replace Anything frequently hallucinates around products, whereas our pipeline mitigates this by providing stronger initialization.

4.5. Qualitative Evaluation

Evaluating image quality remains a challenging task. While numerous metrics exist, each captures only limited aspects, and none adequately assess whether an object is convincingly integrated into a scene—critical for inpainting and outpainting. Consequently, we rely on qualitative evaluation through a user study. We employ Amazon Ground Truth Labeling Service [15] to compare images generated by the baseline model and our pipeline.

We randomly sample 50 generation pairs per model, shuffle them, and collect responses from five annotators per pair. For each task, the annotators are shown a caption, the original product image, and two generated images depicting the product within the described scene. They are instructed to choose the best generated outcome according to the following criteria (in order of importance):

Carefully read the provided caption and look at the provided reference image for the product. Compare the two generated images and select the one that matches the below criteria best. The requirements are listed in the order of importance. Make sure to move down to the next requirement only if both of the images have equivalent performance on the current requirement.

1. *Integration of product in the scene*
2. *Maintaining product look and fidelity*
3. *Alignment with caption: The generated image clearly demonstrates the elements described in the caption and their relative positions in the scene.*

Since our pipeline has a strict quality assessment module, there exist cases where none of the generated images passed our agentic quality assessment. During sampling image pairs, if we sample a rare case like this, we take a note and consider an extra win for the baseline model when the preference rate based on annotations is calculated.

As shown in Figure 3, the results for three of our test models show a clear preference of the pipeline over the baseline run of the models.

In summary, our numerical metrics show a consistent improvement of the pipeline compared to the base model in maintaining product's integrity and our human evaluation results further show that when choosing a proper base model for the task in hand, pipeline can significantly help the model to generate more quality images.

Table 1. Comparison of metrics across three models with and without the proposed pipeline. Values are mean ± std.

	IClight		Nova Canvas		Replace Anything	
Metric	Baseline	Pipeline (ours)	Baseline	Pipeline (ours)	Baseline	Pipeline (ours)
hpsv2	0.240 ± 0.024	0.231 ± 0.063	0.232 ± 0.030	0.233 ± 0.024	0.213 ± 0.036	$\mathbf{0.221 \pm 0.027}$
ms_ssim	0.239 ± 0.210	$\mathbf{0.282 \pm 0.179}^*$	0.382 ± 0.269	$\mathbf{0.422 \pm 0.219}^*$	0.289 ± 0.228	$\mathbf{0.334 \pm 0.103}^*$
dinov2	0.401 ± 0.282	$\mathbf{0.589 \pm 0.274}^*$	0.645 ± 0.213	$\mathbf{0.781 \pm 0.185}^*$	0.538 ± 0.235	$\mathbf{0.669 \pm 0.196}^*$

Note. * Statistically significant improvement over baseline ($p < 0.05$, paired t-test).

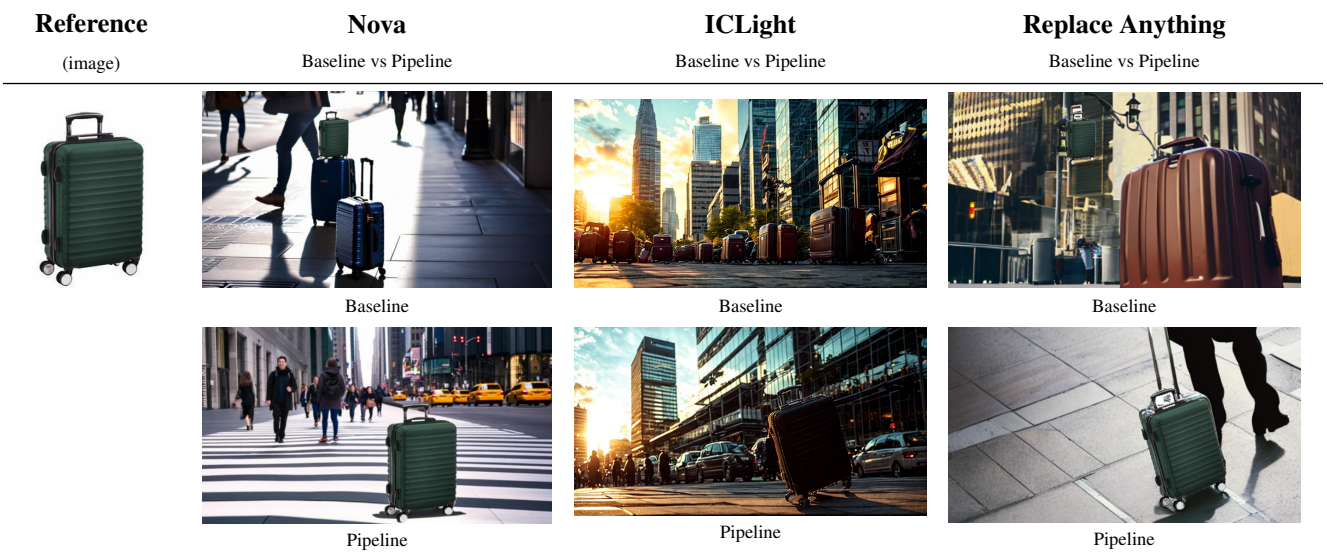

Figure 2. Baseline vs. pipeline outputs across three models for multiple product references. Prompts span the full width below each image row.

4.6. Limitations

We identify two primary sources of failure in our pipeline: dataset limitations and base model limitations.

4.6.1. Dataset Limitations

The ABO dataset contains a small number of mis-labeled product images. Although infrequent, these errors can confuse the generation process and lead to nondeterministic

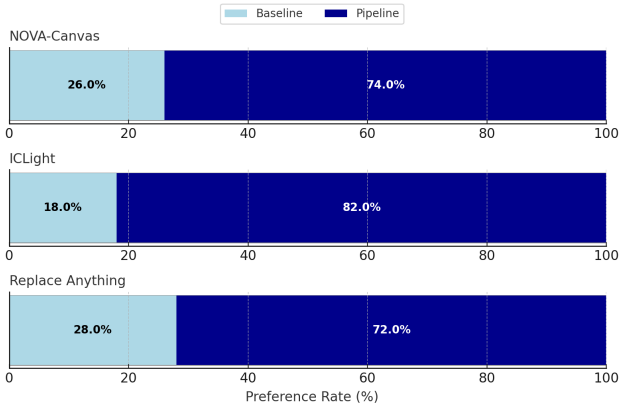

Figure 3. Preference rates for Baseline (light blue) vs. Pipeline (dark blue) across different models.

outputs (see Supplementary Material, Section 1.2, Figures 7, 8, and 9).

4.6.2. Base Model Limitations

The majority of failure cases stem from the underlying generative models. First, such models often hallucinate by generating spurious element around the masked region. In our set of baseline models we observe this failure case mostly in the outcomes of Replace Anything (see Supplementary Material, Section 1.2, Figure 4). Our placement strategy reduces - but does not eliminate - this effect, and when the base model fails, the pipeline inherits these errors.

Second, when the reference product appears relatively too small in the scene, models tend to ignore the object entirely. The model in this case might generate a different instance of the product, or omit it altogether. Our current quality module detects the latter case (rule number 2 in Section 3.6) but does not explicitly check for product similarity to balance cost and latency. Repeated cases of this error type would require the integrating of an object fidelity test into the Quality Control Module (see Supplementary Material, Section 1.2, Figures 5, and 6).

Finally, individual base models exhibit their own quirks. For example, despite being trained to preserve object features, our experiments reveal pixel-level changes to products when using ICLight as the image generator. Particularly, ICLight often alters product texture or color when adjusting illumination. Since such behaviors are model-specific, we recommend testing the base model on a small subset of data to surface recurring issues and adjusting the quality checks accordingly (see Supplementary Material, Section 1.2, Figure 10).

5. Conclusion and Future Work

In this paper, we introduce a modular pipeline for scalable generation of marketing images. Although tailored to marketing, the design is highly customizable and can be adapted to a wide range of content-creation tasks that follow visual guidelines. By automating the iterative generation process and enforcing quality through guardrail checks, our approach reduces the need for constant human oversight. Instead, human effort is focused where it is most valuable—making final selections from images that already satisfy predefined criteria. While our current system focuses primarily on maintaining product integrity and visual composition, brands often require adherence to strict and diverse guidelines such as color palettes, copy style, and layout constraints. Future work includes extending the pipeline to automatically incorporate these brand-specific rules, enabling more comprehensive compliance and further streamlining large-scale production workflows. We view this work as a step toward the productionization of image generation, enabling deployment at scale and addressing real industry needs.

References

[1] Amazon. Amazon nova: Foundation models for general intelligence. https://aws.amazon.com/nova/, 2025. 5, 6

[2] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization, 2024. 1, 2

[3] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, et al. Abo: Dataset and benchmarks for real-world 3d object understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21126–21136, 2022. 5

[4] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 1

[5] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models, 2023. 1

[6] Phillip Mueller, Jannik Wiese, Ioan Craciun, and Lars Mikelsons. Insertdiffusion: Identity preserving visualization of objects through a training-free diffusion architecture, 2024. 2

[7] 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. 6

[8] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1

[9] Alec Radford, Jong Wook Kim, Chris Xu, Greg Brockman, Patrick McLeavey, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021. 4

[10] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. 1

[11] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 6

[12] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[13] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. 1

[14] christoph schuhmann. improved-aesthetic-predictor public. https://github.com/christophschuhmann/improved-aesthetic-predictor, 2023. 4

[15] Amazon Web Services. Amazon sagemaker ground truth. https://aws.amazon.com/sagemaker/groundtruth/, 2024. 6

[16] Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, and Daniel Aliaga. Objectstitch: Generative object compositing, 2022. 2

[17] Z. Wang, E.P. Simoncelli, and A.C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, pages 1398–1402 Vol.2, 2003. 6

[18] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023. 6

[19] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models, 2023. 1

[20] Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. Inpaint anything: Segment anything meets image inpainting, 2023. 2, 5

[21] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 1

[22] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Scaling in-the-wild training for diffusion-based illumination harmonization and editing by imposing consistent light transport. In *The Thirteenth International Conference on Learning Representations*, 2025. 2, 5