AutoGrader: Hybrid Evaluation of GenAI Product Images via Contrastive Embeddings and LLM-Generated Grading Notes

Ramesh Chadalavada*
Amazon, USA
Bellevue, WA, USA
hchadala@amazon.com

Sumanta Kashyapi*
Amazon, USA
Bellevue, WA, USA
sumakash@amazon.com

Ramkumar Subramanian
Amazon, USA
Bellevue, WA, USA
ramkumsu@amazon.com

Matt Witten
Amazon, USA
Bellevue, WA, USA
mwitten@amazon.com

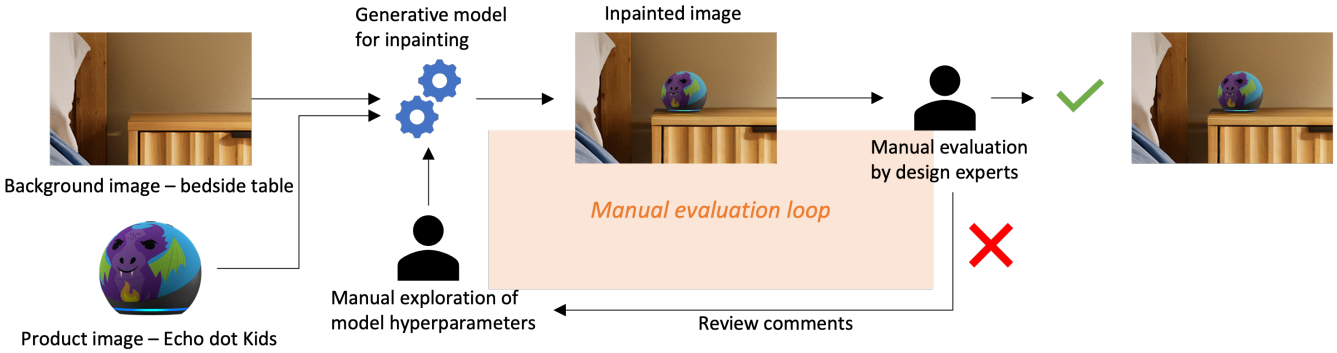

Figure 1: Typical implementation of GenAI-assisted asset generation workflows for product marketing using inpainting. Although generative models (e.g., Stable Diffusion) can produce realistic image assets, navigating the hyperparameter space to achieve marketing-grade quality remains a manual, iterative process requiring expertise from scientists and design specialists.

Abstract

High-quality product image generation is essential for scaling commercial content creation, yet evaluating AI-generated visuals remains a costly bottleneck requiring extensive human review. We propose a hybrid automated evaluation framework that removes the need for exhaustive manual inspection by integrating contrastive learning-based models with Vision-Language Model (VLM) reasoning. Our system filters low-quality images, generates scene-specific grading criteria via LLMs, and conducts weighted, criteria-driven evaluation per device-background pair. Trained on explicit human feedback and implicit image signals, the hybrid model achieves 70% agreement with design standards, flags only uncertain outputs for review, and improves high-quality image selection by 64%. This increases the volume of lifestyle imagery, enriches personalization assets, and enables scalable content generation. The method outperforms baselines and SoTA generative models, potentially saving 5,400+ hours of manual effort annually and unlocking 2M+ unit sales. Experiments across a public benchmark and an Amazon dataset show a 50% cut in image turnaround time, positioning adaptive evaluation as key to scalable, high-fidelity generation workflows.

1 Introduction

High-quality product visualization plays a pivotal role in e-commerce success, particularly for electronic devices. Eye-tracking studies [1, 2] consistently demonstrate that product images are the primary focal point for customers during the initial product discovery phase. This visual-first approach has led marketing teams to invest substantial resources in producing compelling imagery to drive engagement and sales. However, the increasing cadence of marketing campaigns, seasonal promotions, and new product launches has created unprecedented demand for visual content, straining traditional production pipelines that rely heavily on specialized artistic expertise.

To improve scalability, content creators frequently employ *image inpainting* techniques [3–6], which reuse background assets while compositing new product imagery. While these techniques improve

*Both authors contributed equally to the paper

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/XXXXXXX.XXXXXXX

asset reuse, applying them at scale still demands significant manual effort and creative oversight.

The recent emergence of Generative AI, particularly in realistic image synthesis [7–9], offers promising opportunities to scale content creation. These advances have the potential to enhance artist productivity [10] and democratize the production of professional-grade imagery [11]. In the product marketing domain, image inpainting via generative models [12] has become an increasingly popular approach to seamlessly blend products into diverse lifestyle settings. However, despite these improvements, deploying generative models in production workflows still introduces critical challenges:

- **Model hallucination and brand risk**: Generative models are susceptible to producing subtle distortions that can undermine product authenticity, leading to potential erosion of brand trust.
- **Quality assurance bottlenecks**: While generative models accelerate content creation, maintaining professional quality standards requires intensive human evaluation, creating a bottleneck that limits scalability (Figure 1).
- **Hyperparameter optimization complexity**: The generative output quality is highly sensitive to hyperparameter choices, necessitating manual tuning that remains costly and inefficient.

A promising direction to mitigate these issues is to integrate an automated image evaluation model into the asset generation workflow. Such a model would assess the quality of generated outputs in real time and provide actionable feedback to the generative process. Ideally, this system would not only filter low-quality outputs but also automate hyperparameter optimization, drastically reducing reliance on human oversight. However, realizing such a framework presents several key challenges:

- **Data scarcity**: Obtaining sufficient annotated examples of generated imagery for supervised training remains difficult, particularly in niche, product-specific use cases.
- **Limitations of existing evaluation models**: While no-reference evaluation methods [13–16] exist, they primarily focus on generic aesthetic metrics such as prompt alignment [17] or photographic quality [18], falling short of the fine-grained fidelity requirements for commercial product imagery.
- **Identifying subtle quality differences**: In iterative generation workflows, many inpainted outputs are near-identical, and distinguishing subtle defects across batches is beyond the capability of most existing evaluation approaches.
- **Varying evaluation granularity**: Effective marketing evaluation spans a wide range — from coarse scene coherence (e.g., lighting consistency) to fine-grained details (e.g., pixel-level distortions on the product surface) — a range existing methods are ill-equipped to handle.

To address these challenges, we propose a **hybrid automated evaluation framework** tailored for marketing-grade product imagery. Our framework tackles the limitations of prior work along four key dimensions:

- **Few-shot adaptability**: To mitigate data scarcity, our approach operates with minimal supervision, leveraging a small number of high-level annotations from expert designers.
- **Product-specific quality focus**: Beyond general aesthetic alignment, we explicitly model product fidelity and contextual integrity in lifestyle settings.
- **Fine-grained visual sensitivity**: Our model distinguishes subtle variations among near-identical inpainted batches, ensuring that minor defects do not go undetected.
- **Multi-granularity evaluation**: We introduce a dual-model strategy — combining a ResNet-based contrastive model and a Vision-Language Model (VLM) — to capture both global and local quality signals effectively.

By fusing contrastive learning with LLM-driven, scene-specific grading criteria, our framework delivers semantic sensitivity and fine-grained visual fidelity in tandem. It achieves 70% alignment with expert design feedback, improves marketing-grade image selection by 64%, and reduces review time by 50% through confident auto-approval. This significantly expands the pool of ready-to-use lifestyle imagery and supports personalization at scale. Our system consistently outperforms internal and state-of-the-art baselines and could save 5,400 hours of manual work per year while enabling an estimated 2 million incremental sales.

The rest of the paper is organized as follows: we present the relevant literature in Section 2. The detailed methodologies are discussed in Section 3, followed by the experimental results in Section 4. Finally, we conclude the paper in Section 5.

2 Related Work

The field of generative AI has witnessed rapid advancements in photorealistic image synthesis, enabling transformative applications across digital marketing, advertising, and e-commerce [7, 8, 19–22]. Diffusion models in particular have emerged as a dominant paradigm for generating high-fidelity, controllable images. However, deploying such models in commercial workflows—especially for product marketing—introduces unique challenges surrounding consistency, brand safety, and scalability. A critical bottleneck in production workflows is quality assurance (QA). Even when using high-performing diffusion models, issues such as subtle hallucinations, texture inconsistencies, and lighting mismatches often require extensive manual review, particularly in product-centric settings [23, 24]. Although model fine-tuning or prompt engineering can help mitigate these artifacts, they are computationally expensive and do not eliminate the need for robust downstream evaluation. Existing optimization strategies such as Bayesian optimization [25] and reinforcement learning [26] offer some automation for hyperparameter search, but they rely on handcrafted reward functions and struggle with feedback sparsity in high-dimensional generation tasks. To overcome this, recent work has explored Vision-Language Models (VLMs) as automated "judges" for image evaluation. State-of-the-art models such as Flamingo [27], Prometheus-Vision [28], and GPT-4V [29] exhibit impressive cross-modal reasoning abilities and are used in captioning [30], visual QA [31], and aesthetic analysis [32]. However, most existing applications apply VLMs with static evaluation rubrics and ignore product-specific or scene-aware nuances, which are essential in high-stakes visual marketing. In parallel, contrastive learning has proven highly effective for visual representation learning, particularly in retrieval

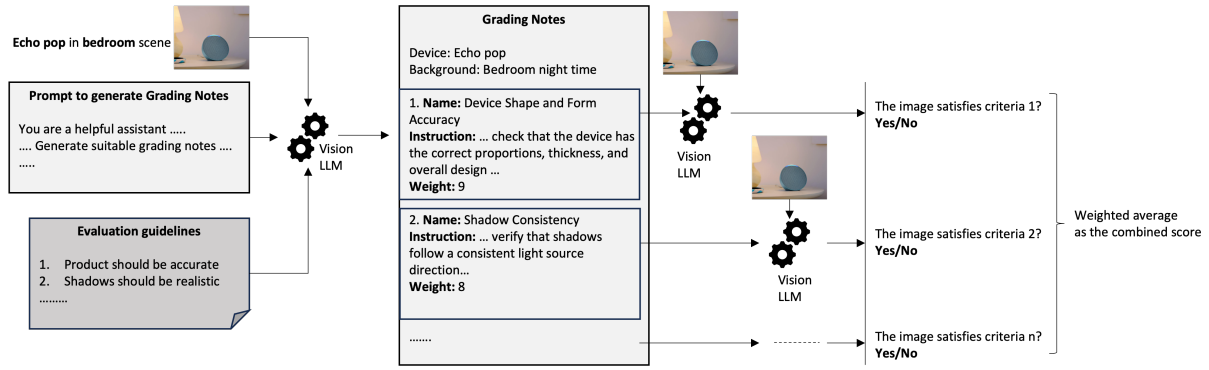

Figure 2: Architecture overview for the LLM-as-a-Judge approach with dynamic grading notes. A generic evaluation guideline is adapted to the specific device-background context via a Vision LLM, producing weighted, context-specific evaluation criteria. Each image is then scored against these criteria through a chain-of-prompts mechanism.

and ranking tasks [33]. In the image quality domain, recent approaches [13, 15] use contrastive objectives to separate pristine from distorted samples, focusing on global degradations such as blur or noise. Yet, these models fall short when evaluating near-identical generations that differ in fine details like edge integrity or contact shadows—details critical in professional marketing images. GLIPS [34] attempts to bridge this gap by combining local and global features for evaluating photorealism in AI-generated images. While promising, GLIPS was not designed for batch-wise comparative ranking and lacks sensitivity to contextual requirements like lighting directionality or compositional fit—key quality signals in e-commerce imagery. Broadly, current evaluators—whether handcrafted, CNN-based, or LLM-based—are not designed to detect subtle, context-dependent differences among high-resolution generative outputs. This limitation is particularly problematic for workflows involving product placement in lifestyle scenes, where the evaluation signal must balance photorealism, product fidelity, and scene coherence. To the best of our knowledge, our framework is the first to combine VLM-based semantic reasoning with contrastive fine-grained perceptual analysis in a no-reference setting, tailored for iterative selection of high-quality generative outputs. By introducing dynamic grading notes and contrastive image embeddings, our approach fills a critical gap in scalable, automated QA pipelines for commercial image generation, delivering both explainability and production-ready accuracy.

3 Approach

Our objective is to develop a robust, no-reference evaluation framework that can quantitatively assess the quality of high-resolution inpainted product images in lifestyle settings. These images are synthesized using generative models and must meet stringent visual fidelity standards, even in the absence of ground-truth references. Specifically, we aim to identify the highest-quality image from a batch of inpainted candidates corresponding to a given product-background composition. We frame this as an image ranking problem:

Problem Statement: Given a background image $b_i \in B$, a product image $d_j \in D$, and a set of K inpainted images $\{p_{ij}^k\}_{k=1}^K$ generated from these inputs, our task is to compute quality scores s_{ij}^k for each inpainted image using an evaluation model $\mathcal{M}_\theta$:

$$s_{ij}^k = \mathcal{M}_\theta(p_{ij}^k, b_i, d_j)$$

where θ represents the learnable parameters. We design $\mathcal{M}_\theta$ as a hybrid framework that integrates two complementary components: (i) a semantic evaluator based on Vision-Language Models (VLMs), and (ii) a fine-grained quality detector based on contrastive learning using convolutional embeddings.

3.1 LLM as a Judge with Dynamic Grading Notes

Recent Vision-Language Models (VLMs) such as Flamingo [27] and GPT-4V [29] exhibit strong capabilities in multimodal understanding and have been leveraged for tasks including visual question answering [31], captioning [30], and aesthetic scoring [32]. Inspired by these, we implement an LLM-as-a-Judge approach to assess image quality.

Rather than applying a static rubric across all product-background compositions, we dynamically generate evaluation criteria for each specific scenario. A VLM is prompted with a generic guideline authored by expert designers, along with a representative inpainted image. It returns a structured set of scene-specific grading notes, each representing a criterion with an associated weight indicating importance.

Each criterion includes a vision-specific grading instruction (e.g., "Check whether shadows cast by the product align with the background lighting."). During evaluation, the image is independently assessed for each criterion using a chain-of-prompt method, and scores are aggregated as a weighted average to compute the final semantic quality score s_{VLM}^k.

3.2 Contrastive Image Embedding

While the VLM-based evaluation captures global and semantic quality attributes, it may overlook subtle but important visual artifacts such as edge distortions, shadow mismatches, or texture inconsistencies. To complement this, we train a contrastive convolutional model that learns to distinguish fine-grained quality variations.

3.2.1 Selective Oversampling of Limited Examples. One major challenge in training such a model is the scarcity of annotated low-quality inpainted examples. We address this by selectively oversampling from the hyperparameter space of the generative model.
Starting from a known high-quality hyperparameter configuration $h_+ \in \mathcal{H}$, we define a neighborhood $\mathcal{N}(h_+)$ and sample negative hyperparameter sets $h_- \in \mathcal{H}^- = \mathcal{H} \setminus \mathcal{N}(h_+)$, ensuring a diverse distribution of failure modes. These suboptimal configurations are used to generate visually degraded images, which are incorporated into training to strengthen the model's ability to distinguish high-quality from flawed outputs.

3.2.2 Contrastive Learning of Inpainting Iterations. We model quality ranking as a dense retrieval task. Each product-background composite image is treated as a query q_i, and its corresponding inpainted outputs are treated as documents to be ranked. We train a dual-encoder architecture with ResNet18 [35] backbones (f_q for the query and f_d for the document) using a triplet loss formulation:

$$\mathcal{L} = \sum_i \max\left\{\text{dist}(\vec{q}_i, \vec{p}_i) - \text{dist}(\vec{q}_i, \vec{n}_i) + M, 0\right\}$$

where $\vec{q}_i = f_q(q_i)$, $\vec{p}_i$ and $\vec{n}_i$ are the positive and negative inpainted samples respectively, and M is a margin. Randomized augmentations are applied to reduce overfitting.
At inference time, we compute the quality score $s^k_{\text{Contrastive}}$ for an inpainted image e_i as the average cosine similarity between the query embedding $\vec{q}_i$ and a set of augmented versions of e_i:

$$s^k_{\text{Contrastive}} = \frac{1}{K}\sum_{k=1}^{K} \text{sim}(\vec{q}_i, \vec{e_{ik}})$$

3.3 Adaptive Combination

The final quality score s^k_{ij} for each inpainted image is computed as a convex combination of the semantic and perceptual scores:

$$s^k_{ij} = \lambda s^k_{\text{VLM}} + (1-\lambda) s^k_{\text{Contrastive}}$$

Here, $\lambda \in [0, 1]$ is a tunable hyperparameter chosen via cross-validation to optimize alignment with expert ratings. This adaptive fusion leverages the strengths of both components, enabling our model to perform robust, reference-free evaluation across a wide range of generative scenarios.

4 Experiments

To evaluate the effectiveness of our proposed hybrid evaluation framework, we conduct a series of experiments comparing our method against state-of-the-art baselines across multiple quality metrics. We assess both quantitative alignment with expert designer judgments and qualitative interpretability of our scores. Our experiments are structured into four parts: dataset creation, evaluation benchmarks, comparative results, and ablation studies.

4.1 Dataset

There are currently no standardized datasets that support benchmarking of inpainting quality for product marketing scenarios. To address this, we construct a proprietary benchmark dataset using images generated with Amazon's internal inpainting engine based on Stable Diffusion XL. The dataset contains over 150 high-resolution lifestyle images with inpainted Amazon devices blended into diverse real-world backgrounds. These images span a variety of device types, room settings, lighting conditions, and compositional layouts.
Each image is evaluated by a panel of expert designers using a standardized guideline covering key criteria including: lighting realism, shadow consistency, contextual blending, and product fidelity. Designers provided binary accept/reject judgments as well as textual comments. For this study, we use the binary approval labels for supervised evaluation. This dataset enables rigorous testing of model performance under real-world production constraints.

4.2 Baselines and Metrics

We compare our hybrid approach against three categories of baseline methods:

- **LLM-as-a-Judge (Static):** A state-of-the-art VLM is prompted to evaluate inpainted images using static, non-contextual rubrics.
- **ViT-based Classifier:** A supervised image classifier trained on the binary labels from our dataset using a ViT backbone.
- **BRISQUE:** A widely-used no-reference quality metric for natural image distortions [36].

Note that, all LLM-based models and BRISQUE are evaluated in zero-shot setting. For fair comparison with models involving supervised training such as ViT-based classifier and the contrastive module of our hybrid approach, we report the consolidated evaluations of five-fold cross validation by considering the predicted scores of only the test-folds.

4.3 Results

Table 1 summarizes the performance of each model. Our hybrid approach outperforms all baselines by a significant margin across all metrics. Notably, the Vision LLM baseline, when used with static rubrics, performs poorly—highlighting the importance of scene-specific grading notes. Specifically, we notice the static LLM baseline lacks the variance of predicted scores resulting in identical scores for more than 20% of the dataset, resulting in poor ranking performances in terms of MRR, average precision and agreement@40 even though obtaining high precision@1 score. The ViT classifier shows marginal improvement but still lacks generalization. Our method achieves 70.4% agreement with expert-labeled images in the top 40%, demonstrating high reliability. The MRR and Precision@1 scores confirm that the model is highly effective at surfacing the most marketing-ready image from a batch.

4.4 Ablation Study

To isolate the contribution of each component in our hybrid framework, we conduct an ablation study evaluating the Vision LLM module (with dynamic grading notes) and the contrastive ResNet independently, as well as their combined performance. Results indicate that both components contribute meaningfully to the final score. The contrastive model captures fine-grained visual artifacts, while the LLM module adds semantic, context-aware judgment. Their combination yields superior alignment with human evaluations.

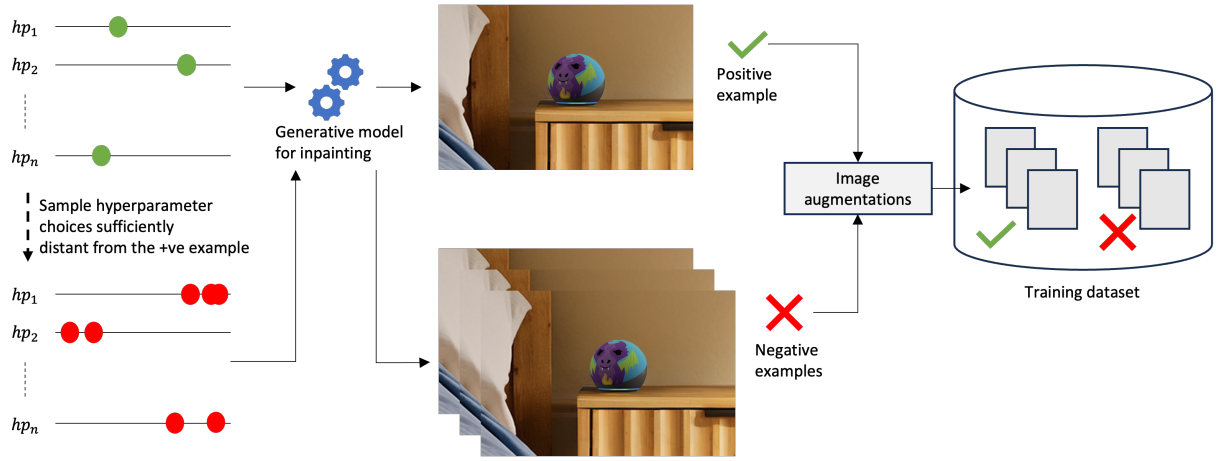

Figure 3: Selective oversampling workflow to address negative data scarcity by generating meaningful low-quality examples based on distance from known good hyperparameters.

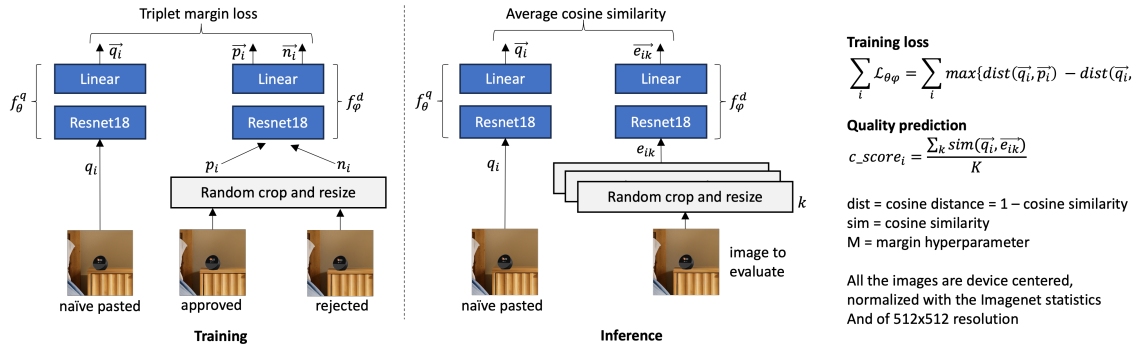

Figure 4: Architecture of the contrastive ResNet model for fine-grained, context-aware inpainting quality evaluation.

Table 1: Experimental Results on Inpainted Image Evaluation. Evaluation metrics used MRR: Mean Reciprocal Rank, Avg. Prec.: Average precision, Prec@1: Precision@1. Agreement@40: denotes percentage overlap between top-40% of model-ranked images and human-approved images.

Method	MRR	Avg. Prec.	Prec@1	Agreement@40
LLM-as-a-Judge (Static)	0.3871	0.3765	0.4815	0.1481
ViT Classifier	0.5339	0.5512	0.3333	0.4074
BRISQUE	0.5710	0.6000	0.2963	0.5556
Ours (Hybrid)	**0.6554**	**0.6876**	**0.4815**	**0.7037**

Table 2: Ablation Study of the Proposed Hybrid Approach

Component	MRR	Avg. Prec.	Prec@1	Agreement@40
LLM-as-a-Judge (Dynamic Grading Notes)	0.4882	0.4957	0.2963	0.3704
Contrastive ResNet	0.6462	0.6864	0.4444	0.6667
Hybrid (Ours)	**0.6554**	**0.6876**	**0.4815**	**0.7037**

4.5 Qualitative Study

In addition to the quantitative evaluation, we conducted an in-depth qualitative analysis to examine the interpretability, visual sensitivity, and real-world applicability of our hybrid evaluation framework. This study aimed to understand how effectively the model replicates expert design judgments and whether it can distinguish nuanced visual differences in inpainted outputs. Our analysis revealed three core behaviors:

- **Semantic reasoning by the LLM module:** The LLM-driven grading component reliably penalizes contextually inconsistent renderings. It detects lighting mismatches (e.g., harsh directional shadows in evenly lit environments), misaligned object placement, and spatial incoherence. These results affirm the effectiveness of dynamically generated, scene-specific grading notes tailored to each product-background pair.
- **Fine-grained detection by the contrastive module:** The contrastive model excels in capturing subtle artifacts often missed by high-level semantic models. It consistently downgrades images with issues such as edge bleeding, inconsistent shadow gradients, and over-smoothed textures—visual flaws that impact photorealism yet remain challenging to identify via traditional aesthetic metrics.
- **Complementary strengths in hybrid scoring:** The integrated hybrid score synthesizes both global semantic reasoning and local perceptual cues, producing rankings that closely mirror expert human evaluations. Top-ranked images frequently exhibit strong visual coherence and brand alignment, while lower-ranked outputs present minor—but perceptually important—defects.

To illustrate, Figure 5 presents a representative test case involving an Echo Dot Kids device placed on a bedside table. A batch of inpainted variants was generated under varying hyperparameter conditions and evaluated by our framework. The top-ranked image received the highest hybrid score and was independently selected by Amazon design experts for deployment in a live campaign. During the design review, experts noted the presence of realistic bilateral contact shadows and appropriate light falloff on the tabletop as key approval factors—elements our model scored highly. In contrast, lower-ranked images suffered from subtle but detrimental issues such as overly sharp reflections, improper object grounding, or lighting direction mismatches. These were appropriately penalized by the system, demonstrating its capacity to capture both visual and contextual quality. This alignment between model predictions and professional rationale highlights the framework's practical value. By enabling reference-free evaluation that accounts for both aesthetic and compositional fidelity, the system supports scalable content review while offering actionable, interpretable feedback. As personalized visual content becomes central to modern marketing, such hybrid evaluation systems are essential for bridging the gap between generative output and production-grade quality.

5 Conclusion

We present a hybrid automated evaluation framework designed to meet the high standards of product image generation in e-commerce and marketing contexts. By integrating the semantic reasoning capabilities of Vision-Language Models (VLMs) with the fine-grained perceptual sensitivity of contrastive learning-based convolutional models, our system provides a robust, reference-free solution for ranking and filtering inpainted lifestyle images.

Our framework dynamically adapts evaluation criteria to the product-background context using LLM-generated grading notes, ensuring contextual relevance and interpretability. Complementing this, the contrastive module captures subtle artifacts that are often missed by semantic models alone. Together, the system delivers high agreement with professional designer judgments (70%), boosts high-quality image selection by 64%, and reduces image review turnaround time by 50%. Importantly, it enables reviewers to focus only on low-confidence edge cases, drastically reducing manual effort while maintaining visual quality standards.

This capability has direct implications for scaling creative production workflows. By expanding the pool of available marketing-grade assets, the system supports more personalized, timely, and visually coherent customer experiences. With an estimated annual savings of over 5,400 hours of manual review and a potential uplift of 2 million unit sales, our approach demonstrates tangible business impact.

Looking ahead, this framework offers a strong foundation for next-generation generative systems where quality assurance is integrated into the generation loop. Future work includes incorporating reinforcement learning-based feedback mechanisms, extending to multi-modal content (e.g., video, 3D renders), and generalizing the approach to other product verticals beyond consumer electronics. Our results highlight the importance of automated, adaptive evaluation in unlocking the full potential of generative AI. The framework not only cuts operational burden but also unlocks strategic value—fueling personalized customer journeys, faster launches, and brand-safe content at scale. Future directions include integrating reinforcement-driven feedback loops, expanding to 3D/AR content, and adapting across product domains, ensuring the system evolves alongside generative capabilities.

References

[1] Jiazhang Wang, Tianfu Wang, Bingjie Xu, Oliver Cossairt, and Florian Willomitzer. Accurate eye tracking from dense 3d surface reconstructions using single-shot deflectometry. *Nature Communications*, 16(1):2902, 2025.

[2] Patrick Mikalef, Kshitij Sharma, Sheshadri Chatterjee, Ranjan Chaudhuri, Vinit Parida, and Shivam Gupta. All eyes on me: Predicting consumer intentions on social commerce platforms using eye-tracking data and ensemble learning. *Decision Support Systems*, 175:114039, 2023.

[3] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 28(3):24, 2009.

[4] Alexandru Telea. An image inpainting technique based on the fast marching method. *Journal of graphics tools*, 9(1):23–34, 2004.

[5] Marcelo Bertalmio, Andrea L Bertozzi, and Guillermo Sapiro. Navier-stokes, fluid dynamics, and image and video inpainting. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I. IEEE, 2001.

[6] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514, 2018.

[7] 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.

[8] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

[9] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

[10] Eric Zhou and Dokyun Lee. Generative artificial intelligence, human creativity, and art. *PNAS nexus*, 3(3):pgae052, 2024.

[11] Tojin Eapen, Daniel J Finkenstadt, Josh Folk, and Lokesh Venkataswamy. How generative ai can augment human creativity. *Harvard Business Review*, 101(4), 2023.

[12] Jochen Hartmann, Yannick Exner, and Samuel Domdey. The power of generative marketing: Can generative ai create superhuman visual marketing content? *International Journal of Research in Marketing*, 42(1):13–31, 2025.

Figure 5: Ranked inpainting results by our proposed hybrid approach featuring an Echo Dot Kids device on top of a bedside table. The top-ranked image by our approach is also approved by design experts for use in live Amazon.com campaigns.

[13] Yuxuan Yang, Zhichun Lei, and Changlu Li. No-reference image quality assessment combining swin-transformer and natural scene statistics. *Sensors*, 24(16):5221, 2024.

[14] Jinsong Shi, Pan Gao, and Jie Qin. Transformer-based no-reference image quality assessment via supervised contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4829–4837, 2024.

[15] Ibrahim Kajo, Abderrazak Chahi, Mohamed Kas, and Yassine Ruichek. No-reference quality evaluation of realistic hazy images via singular value decomposition. *Neurocomputing*, 610:128574, 2024.

[16] Prianka Ramachandran Radhabai, Kavitha Kvn, Ashok Shanmugam, and Agbotiname Lucky Imoize. An effective no-reference image quality index prediction with a hybrid artificial intelligence approach for denoised mri images. *BMC Medical Imaging*, 24(1):208, 2024.

[17] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

[18] Junyu Chen, Jie An, Hanjia Lyu, Christopher Kanan, and Jiebo Luo. Learning to evaluate the artness of ai-generated images. *IEEE Transactions on Multimedia*, 2024.

[19] Xuan Ju, Junhao Zhuang, Zhaoyang Zhang, Yuxuan Bian, Qiang Xu, and Ying Shan. Image inpainting models are effective tools for instruction-guided image editing. *arXiv preprint arXiv:2407.13139*, 2024.

[20] Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In *European Conference on Computer Vision*, pages 150–168. Springer, 2024.

[21] Gabriele Valvano, Antonino Agostino, Giovanni De Magistris, Antonino Graziano, and Giacomo Veneri. Controllable image synthesis of industrial data using stable diffusion. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 5354–5363, 2024.

[22] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, pages 423–439. Springer, 2022.

[23] Dinesh Kumar and Nidhi Suthar. Ethical and legal challenges of ai in marketing: an exploration of solutions. *Journal of Information, Communication and Ethics in Society*, 22(1):124–144, 2024.

[24] Rita Ngoc To, Yi-Chia Wu, Parichehr Kianian, and Zhe Zhang. When ai doesn't sell prada: Why using ai-generated advertisements backfires for luxury brands. *Journal of Advertising Research*, pages 1–35, 2025.

[25] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, 25, 2012.

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

[27] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[28] Seongyun Lee, Seungone Kim, Sue Park, Geewook Kim, and Minjoon Seo. Prometheus-vision: Vision-language model as a judge for fine-grained evaluation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11286–11315, 2024.

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

[30] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[31] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19175–19186, 2023.

[32] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20144–20154, 2023.

[33] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

[34] Memoona Aziz, Umair Rehman, Muhammad Umair Danish, and Katarina Grolinger. Global-local image perceptual score (glips): Evaluating photorealistic quality of ai-generated images. *IEEE Transactions on Human-Machine Systems*, 2025.

[35] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[36] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012.