# **Cost Savings from Automatic Quality Assessment of Generated Images**

Xavier Giro-i-Nieto Nefeli Andreou Anqi Liang Manel Baradad Francesc Moreno-Noguer Aleix Martinez Amazon

## **Abstract**

Deep generative models have shown impressive progress in recent years, making it possible to produce high quality images with a simple text prompt or a reference image. However, state of the art technology does not yet meet the quality standards offered by traditional photographic methods. For this reason, production pipelines that use generated images often include a manual stage of image quality assessment (IQA). This process is slow and expensive, especially because of the low yield of automatically generated images that pass the quality bar. The IQA workload can be reduced by introducing an automatic pre-filtering stage, that will increase the overall quality of the images sent to review and, therefore, reduce the average cost required to obtain a high quality image. We present a formula that estimates the cost savings depending on the precision and pass yield of a generic IQA engine. This formula is applied in a use case of background inpainting, showcasing a significant cost saving of 51.61% obtained with a simple AutoML solution.

# 1. Motivation

The automatic production of images has experienced a rapid progress thanks to the broad adoption of deep neural networks in generative models learning. The current solutions based on diffusion [5, 21] and auto-regressive [30, 44] models can produce high quality images, sometimes indistinguishable to the naked eye from their photographic counterparts. This represents an opportunity for industry because automatic image generation is orders of magnitude less expensive than a manual capture, which typically requires a photography studio and trained professionals. Moreover, the time necessary to generate an image is virtually zero compared to a manual production, and the scale is basically limited by the available computational resources. These advantages make generative deep learning an attractive technology for reducing costs and increasing the scale.

However, current technology is not perfect, and may introduce noticeable artifacts in a significant portion of the generated images. The nature of these defects may be di-

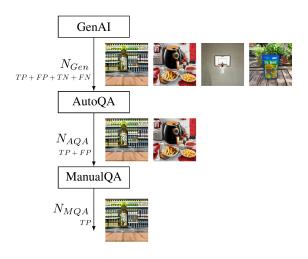


Figure 1. Vertical pipeline for image generation and quality assessment with sample images.  $N_{Gen}$ ,  $N_{AQA}$  and  $N_{MQA}$  represent the number of images output after each block. TP/FP/TN/FN refer to true/false positive/negatives.

verse, whether directly visible on the generated images, or failing to follow the provided guidelines, which typically take the form of a textual prompt or a visual reference. For example, in the *inpainting* study case that we explore in this work, the goal is generating a realistic visual context for an object depicted over a white background. This is a specially challenging problem, because not only the background must be perceptually realistic, but it must also be semantically coherent with the provided reference object. Figure 2 shows six different types of defects that are common in background inpainting. The details of these defects are described in Table 1. According to our studies with marketing professionals, existing technology offers low QA pass yield for this task: only 1.87% with the model considered in this work, based on Stability AI SDXL [22].

Commercial applications require high quality images, so an additional filtering is currently needed before delivering automatically generated images. A first solution is relying only on human annotators trained to detect the most common defects, an approach we refer to as *ManualOA*. This



Figure 2. Types of visual defects in background inpainting. Top: Reference object. Bottom: Generated image.

approach is simple in terms of technology, because it only requires a user interface to collect the annotations. However, ManualQA also presents challenges in terms of calibrating the annotators with the product requirements and, more importantly, scales very poorly in terms of time and costs. Another approach is introducing a computer vision system capable of detecting the anomalies in the generated images, which we will refer as *AutoQA*. This is a scalable solution whose speed is only limited by the allocated computational resources. On the other hand, existing computer vision systems offer a lower precision than human annotators.

Type of defect	Description			
Main Object Distortion	Black dots on the pan cover.			
Main Object Extension	Straps added to the backpack.			
Misplaced Object	The firepit is an outdoor object but is placed in an indoor environment.			
Scale Mismatch	Chair is much smaller than table.			
Bg. Objects Distortion	Unrealistic chair backs.			
Bg. Structural Distortion	Misaligned wall behind the table.			

Table 1. Detailed descriptions of the defects in Figure 2.

In this work, we consider the hybrid pipeline depicted in Figure 1, where ManualQA will only process those images that pass AutoQA. This scheme aims at increasing the pass yield of ManualQA by pre-filtering the generated images with AutoQA. As a result, the final cost of generating a given amount of images is lower thanks to the reduced amount of reviews needed.

Our contributions are three-fold: (a) a novel IQA task

for images generated by background inpainting, focused on the popular application of placing objects in generated contexts (b) a closed formula that determines the cost savings of introducing an AutoQA block in an image generation system, and a (c) a case study based on a zero-shot VLM and AutoML which, in their best set up, obtain financial cost savings of 51.61%.

#### 2. Related work

Image evaluation methodologies in the literature can be broadly categorized by (1) the rating of individual images [32] or comparative ranking across image sets [28] (2) by the availability of reference images that are used for comparison (full-reference methods [45], reduced-reference methods [24], or no-reference methods [20], or (3) by the main features used including traditional image features [4, 32] or deep features (such as representations extracted from pre-trained models) [13, 15, 39, 47]. Our work concentrates on individual image quality assessment in a reference-free manner. Image Quality Evaluation can span multiple dimensions, including image realism [3, 14, 26, 47], text-image alignment [13, 39, 41], image aesthetics [27, 40, 42] and human preferences [17, 37, 38, 48]. Our work considers the evaluation of realism in generated imagery.

Automated Image Quality Assessment (IQA). Automatic Image Quality Assessment has become a critical block when generating images, and multiple benchmarks and metrics have been proposed [11, 32]. When comparing an distorted image with its reference, PSNR and MSE perform pixel-level comparisons, while SSIM [32] captures perceptual changes through structural information by comparing luminance, contrast, and structure. [6] proposed an Image Quality Transformer (IQT) to measure the perceptual quality between an input image pair of distorted and reference images. Given the fact that reference images are not always available, [8, 43]

explore the performance of transformer-based no-Reference image quality assessment (NR-IQA). Along the NR-IQA research line, [31] proposed an algorithm based on the Swin Transformer [19] with fused features from multiple stages to better predict image quality.

Image Quality Assessment with VLMs The surprising understanding capabilities of large Vision-Language Models (VLMs) have been explored for image quality assessment. CLIPScore [13] uses CLIP embeddings for evaluating text-image alignment. Building upon this foundation, ImageReward [39] introduced a reward model finetuned on human preferences to better capture image quality aspects. PickScore and HPS [17, 38] align better with subjective human judgement in the evaluation. Similar to GenomeBench [9] which proposes a benchmark to evaluate the quality and text alignment of generated images through a series of questions and human annotations, recent works have leveraged large Vision-Language Models (VLMs) for nuanced assessment of image realism through visual question answering (VQA) [2]. Foundational models can engage in detailed dialogues about image artifacts and realism. InstructBLIP [10] demonstrates how instruction-tuning of VLMs can enable more targeted assessment of specific quality aspects. TIFA [15] leverages VQA to provide interpretable assessment of text-image alignment. Cho [7] built the set of questions using Davidsonian graphs. Gecko [34] proposed a fine-grained classification of defects in prompt alignment, named skills. VQAScore [18] showed better performance of a VLM than the CLIPScore for IQA.

### 3. Cost Estimation

We present a formula that estimates the cost savings from introducing AutoQA for automatic image generation, following the pipeline in Figure 1. The QA classification systems operate with binary classes: Clean (denoted with  $\checkmark$ ) and Defect (denoted with  $\checkmark$ ). An image generation engine (GenAI) produces  $N_{Gen}$  images that are first evaluated by the AutoQA block. Only the  $N_{AQA}$  samples that pass AutoQA are reviewed in ManualQA. The result are  $N_{MQA}$  high quality images that passed both QA controls.

We relate the volume of images after each block  $(N_{Gen}, N_{AQA}, N_{MQA})$  through the two metrics that characterize the blocks in Figure 1: the *yield* of images that pass AutoQA  $(y_{AQA})$  and ManualQA  $(y_{MQA})$ , and the *precision* of the image generator  $(P_{Gen})$  and AutoQA engine  $(P_{AQA})$ . The two metrics are interpretable and facilitate the estimation of financial costs in industrial scenarios.

#### 3.1. Volume of Images

As introduced in Figure 1, we develop a formulation that considers *Clean* (✓) as the positive class of our task, and uses the standard notation of TP/FP/TN/FP for referring to

true/false positives/negatives. Based on these definitions, we can relate  $N_{MQA} {\rm with}\ N_{AQA} {\rm as}$ 

$$N_{MQA} = TP = \frac{TP + FP}{TP + FP} \cdot \frac{TP}{TP} \cdot TP =$$

$$= (TP + FP) \cdot \frac{TP}{TP + FP} \cdot \frac{TP}{TP} =$$

$$= N_{AQA} \cdot P_{AQA}(\checkmark) = N_{AQA} \cdot y_{MQA},$$
(1)

where  $P_{AQA}(\checkmark)$  is the precision of the AutoQA module for the *Clean* class. Notice that  $P_{AQA}(\checkmark)$  can also be interpreted as the ManualQA yield,  $y_{MQA}$ , because we consider ManualQA as the source of true predictions.

Similarly, we can relate  $N_{MQA}$  with  $N_{Gen}$  as

$$\begin{split} N_{MQA} &= TP = \frac{TP + FP + TN + FN}{TP + FP + TN + FN} \cdot \frac{TP + FP}{TP + FP} \cdot TP = \\ &= (TP + FP + TN + FN) \frac{TP + FP}{TP + FP + TN + FN} \frac{TP}{TP + FP} = \\ &= N_{Gen} \cdot y_{AQA} \cdot P_{AQA}(\checkmark) = N_{Gen} \cdot y, \end{split} \tag{2}$$

where  $y_{\text{AQA}}$  is the AutoQA yield. We also define the overall yield  $y = y_{AQA} \cdot P_{AQA}(\checkmark) = \frac{TP}{TP + FP + TN + FN}$  to simplify the formulation in future equations.

#### 3.2. Total Cost

The total cost of a pipeline with AutoQA can be derived by weighting the unitary cost  $(c_i)$  of processing the images in each of the three stages  $i \in \{Gen, AQA, MQA\}$ , and summing their results as

$$C_{Total} = N_{Gen} \cdot c_{Gen} + N_{Gen} \cdot c_{AQA} + N_{AQA} \cdot c_{MQA} =$$

$$= \frac{N_{MQA}}{y} c_{Gen} + \frac{N_{MQA}}{y} c_{AQA} + \frac{N_{MQA}}{P_{AQA}(\checkmark)} c_{MQA} =$$

$$= N_{MQA} \cdot \left(\frac{c_{Gen}}{y} + \frac{c_{AQA}}{y} + \frac{c_{MQA}}{P_{AQA}(\checkmark)}\right). \tag{3}$$

#### 3.3. Cost Savings

The cost savings are estimated by comparing  $C_{TOTAL}$  with the costs of obtaining the same amount of images without AutoQA, which is

$$C_{Baseline} = N_{Gen} \cdot c_{Gen} + N_{Gen} \cdot c_{MQA} =$$

$$= N_{Gen} (c_{Gen} + c_{MQA}) =$$

$$= \frac{N_{MQA}}{P_{Gen}(\checkmark)} \cdot (c_{Gen} + c_{MQA}) =$$

$$= N_{MQA} \cdot \left(\frac{c_{Gen}}{P_{Gen}(\checkmark)} + \frac{c_{MQA}}{P_{Gen}(\checkmark)}\right),$$
(4)

where  $P_{Gen}(\checkmark)$  is the precision of the GenAI technology that produces the images, that is, the ManualQA yield.

The final cost savings  $\Delta C$  can be calculated by substracting the total costs defined in Equation 3 from the baseline costs in Equation 4, as

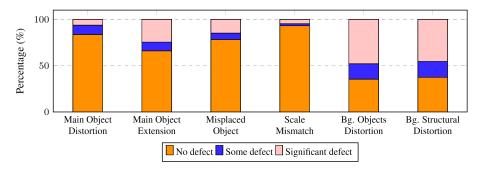


Figure 3. Distribution of defect severity across the six types of coarse defects. Each of the 8,304 generated images was manually labeled by a human annotator according to a scale of three levels: *No defect, Some defect,* or *Significant defect.* This figure does not consider agreement among annotators

$$\Delta C = C_{Baseline} - C_{Total} =$$

$$= N_{MQA} \cdot \left(\frac{c_{Gen}}{P_{Gen}(\checkmark)} + \frac{c_{MQA}}{P_{Gen}(\checkmark)} - \frac{c_{Gen}}{rP_{AQA}(\checkmark)} - \frac{c_{AQA}}{P_{AQA}(\checkmark)}\right) =$$

$$= N_{MQA} \left(\frac{rP_{AQA}(\checkmark) - P_{Gen}(\checkmark)}{y \cdot P_{Gen}(\checkmark)}c_{Gen} - \frac{1}{y}c_{AQA} + \frac{P_{AQA}(\checkmark) - P_{Gen}(\checkmark)}{P_{AQA}(\checkmark) \cdot P_{Gen}(\checkmark)}c_{MQA}\right)$$
(5)

# 4. Experiments

We apply the cost savings formula from Equation 5 in a realistic case for e-commerce: the generation of a virtual context from images that depict products over a white background. This problem is technically referred as background *inpainting*, and the resulting imagery *lifestyle* images. We adopt state of the art technical solutions for both generating and evaluating the quality of the generated images.

#### 4.1. Dataset

Our experiments were conducted with a dataset of 8,304 images of 195 different product types. The images were generated with a latent diffusion-based model open-sourced by Stability AI (SDXL) [22], which was adapted with ControlNet [46] to solve the inpainting task. We worked with a team of photography experts to define six categories of the possible defects, depicted in Figure 2.

Professional annotators from an external vendor were trained to provide a numerical rating on a scale from 1 (*No defect*) to 3 (*Significant defect*). Each image was labeled by 3 different annotators from a pool of 21. The average amount of annotations per worker was 1256.38, with a maximum of 1782 and a minimum of 815. The label distribution of the collected labels is presented in Figure 3. The plots show how the low-level distortions on the background are the most common defects (*Bg. Objects Distortion* and *Bg. Structural Distortion*), while the mismatch of scale between

the provided reference object and the rest of the scene is the most rare defect (*Scale Mismatch*).

The following sections report experiments based on different subsets of these annotated data. In each experiment, we only used samples where all annotators agreed in the labels under study. This way, metrics are more reliable and avoid the problem of low inter-annotator agreement, which is depicted in Table 3. Notice how obtaining alignment among humans for IQA task is very challenging, despite providing detailed instructions to annotators. In addition, the very few images where all annotators agreed on a rating of 2 (some issue), were also discarded.

### 4.2. AutoQA

Two technologies were considered to serve as AutoQA: a classic AutoML solution, and an off-the-shelf large visual language model (VLM).

**AutoML** solutions consider different machine learning techniques to solve a task defined by an annotated dataset. AutoML allows a quick development and provides solid baselines. In our work, we adopt AutoGluon [29], which leverages the *timm* model zoo of deep learning architectures for image analysis [33]. Under the hood, AutoGluon trains a variety of different models, uses bagging (bootstrap aggregation) to train them, and considers different architecture through stack-ensembling these models. In our use case, AutoGluon automatically opted for a solution based on an ensemble of fine-tuned vision transformers.

Large Visual Language Models (VLMs) are generative models pretrained with a very large dataset, typically, of Internet scale. They have shown promising results in a diversity of multimodal tasks, among them, visual-question answering (VQA). We initially considered two VLMs for image quality assessment formulated as VQA, Amazon Lite and Pro 1.0 [1]. After an initial comparison between them, we opted for Nova Pro only, given the poorer performance of Amazon Lite for IQA. The VLMs were conditioned with a prompt specific for each defect type, requesting an answer

Defect type	# img	Oracle	Oracle Autogluon 1.2		Nova Pro 1.0		Random 0.5	
2 circle type	8	$y_{AQA}$	$y_{AQA}$	$P_{\text{AQA}}(\mathbf{X})\uparrow$	$y_{AQA}$	$P_{\text{AQA}}(\mathbf{X})\uparrow$	$y_{AQA}$	$P_{\text{AQA}}(\mathbf{X}) \uparrow$
Main Object Distortion	199	0.96	1.00	0.000	0.98	0.000	0.46	0.037
Main Object extension	231	0.73	0.82	0.561	0.97	0.500	0.52	0.312
Product Placement	212	0.83	0.97	0.571	0.87	0.382	0.44	0.144
Scale Mismatch	199	0.97	0.98	0.667	0.80	0.022	0.52	0.021
Bg. Objects Distortion	254	0.68	0.77	0.508	0.82	0.370	0.56	0.312
Bg. Structural distortion	225	0.74	0.80	0.444	0.98	0.105	0.48	0.284

Table 2. Comparison of yield,  $y_{AQA}$ , and precision for the *Defect* class,  $P_{AQA}(X)$ , for the two considered AutoQA technologies. Column # img indicates the amount of test images considered when assessing each defect. Column *Oracle* represents the defect rate of images that a perfect AutoQA system would detect. Values in **bold** highlight which of the two technologies perform better to detect each defect type, by considering: (a) a  $y_{AQA}$  similar to the Oracle's, and (b) a  $P_{AQA} \approx 0.5$ . Values in red indicate failure cases, due to  $P_{AQA}$  similar to the *Random* case, and/or  $y_{AQA}$  very different from the Oracle's.

Defect Type	Agreement Rate ↑			
Main Object Distortion	0.6525			
Main Object Extension	0.5425			
Product Placement	0.6653			
Scale Mismatch	0.8582			
Bg. Objects Distortion	0.3868			
Bg. Structural Distortion	0.2982			
Any Defect	0.4579			

Table 3. Annotator Agreement Rates by type of defect, and also in the *Any Defect* case. This later case reflects the final IQA task, where a binary decision must be made between  $Clean(\checkmark)$  or  $Defect(\checkmark)$ .

in the same scale as the human annotators: 1 for *No defect*, 2 for *Some defect*, and 3 for *Significant defect*.

The VLMs were queried with the question prompts described in Table 4. These prompts were manually designed to detect each type of defect, and some of the defects were further divided in more fine-grained categories. The VLM was modulated with the role prompt presented in Table 5.

#### 4.2.1. Independent defects

Our first experiment explores which machine learning models can detect the considered defect types, so we focus our metrics on the Defect(X) class. For this experiment, we built a dataset of images that present none or a single a single defect only, that is, we discarded images with multiple types of defects.

We compared AutoGluon 1.2 (AG) only with Amazon Nova Pro 1.0 (NP). Similarly how Nova Pro used a specific prompt for each type of defect, an independent AutoGluon binary classifier was trained for each defect type. A different data subset was created for each defect, providing 90% of the available images to AutoGluon for finetuning, and keeping the remaining 10% for our tests with AutoGluon and Nova Pro. By default, AutoGluon uses the a partition 90% of the provided data for training, and the remaining 10% for

internal validation. The AutoGluon models were left to train as much time as needed, which was in the order of minutes for each binary classifier. Models were trained with a p3.2xlarge cloud desktop in Amazon Web Services (AWS) equipped with a single Tesla V100-SXM2 GPU with 16 GB of memory.

Table 2 shows that, in four out of the six considered defect types, AutoGluon performs better than Nova Pro. Only for the *product placement* defect Nova Pro competes with Autogluon with a better  $y_{AQA}(X)$ . The *main object distortion* defects are not captured by any of the two configurations.

#### 4.2.2. All defects

Based on the results of detecting defects independently, we explore the performance of a full AutoQA system that would detect all defects but *main object distortion*.

The dataset in this section no longer satisfies the restriction of presenting a single defect. In this case, the full dataset contains 3,802 images, and its test partition 380 samples with unbalanced binary labels. Only 18.7 % of the test images are Clean, and the rest contain one or more defects. The test partition covers 192 different object categories, so most categories only have two samples. On the other hand, the training partition used for AutoGluon only covers 140 of these 192 categories, and it is highly unbalanced, where 12 categories account for 50% of the training samples. We adopted this design to have a high coverage of object categories and avoid any bias towards any class. With this set up, most of the test samples can be considered as out-of-domain from the perspective of the object category. Figure 4 details how the train partition is unbalanced, but the test partition is very balanced.

By default, AutoGluon uses the a partition 90% of the provided data for training, and the remaining 10% for internal validation. The AutoGluon models were left to train as much time as needed, which was in the order of minutes for each binary classifier. Models were trained with a p3.2xlarge cloud desktop in Amazon Web Services (AWS)

Coarse defect	Detailed defect	Prompt			
Main Object Distortion	Surface texture  Color blending Structural distortion	Focus on the surface of the {object_class}. Is there any distortion on its texture?  Can you see weird color blending at its contours? Is there any structural distortion in the {object_class}?			
Main Object Extension	Product extension  Product attached	Does the {object_class} present a realistic shape? Compare the shape of the {object_class} in the first generated image to the reference image and its segmentation mask. Make sure that the {object_class} did not grow in extension when the background was generated Is there any other object attached to the {object_class}? If so, is this attachment common and natural?			
	Objects layout	What objects appear in the scene? Are their relative positions natural?			
Misplaced Object	Floating objects	Look at the {object_class}. It must be standing on a surface. Otherwise, consider that it is floating, which is a severe issue.			
	Matching location	In which locations is the normally found? Does the context in the image represent one of these probable locations?			
	Functional location	Where is the {object_class} located? Does it appear in a proper functional location?			
	Rich background	How is the background around the {object_class}? The background must contain rich semantic and be aesthetically appealing. A solid or uniform background is not acceptable.			
Scale Mismatch	Scale mismatch	There is an anomaly in the size of the {object_class} compared to the rest of objects in the scene. True or false?			
Background Objects Distortion	Objects distortion	What objects appear in the image? Is there any distortion in any of them?			
Background Structural Distortion	General Because occlusion	Is there any structural distortion in the scene? Is the background behind the {object_class} realistic? Make sure that there are no discontinuities in the generated background because of the occlusion of the {product_type}			

Table 4. Question prompts organised in a hierarchy of coarse and detailed defects.

equipped with a single Tesla V100-SXM2 GPU with 16 GB of memory.

Based on the results reported in Section 4.2.1, we consider three possible configurations: 1) a cascade of Autogluon binary classifiers (*Cascade AG*), 2) a similar cascade where AutoGluon is replaced by Nova Pro for the *product placement* defect (*Cascade (AG & NP)*), and 3) a single Autogluon binary classifier that does not distinguish between

defect types (*Single (AG only)*). The performance of each configuration is reported in Table 6, together with the metrics of two set ups that facilitate the interpretation of results: (1) *Random (0.5)* represents a lower bound baseline where the binary classifier would simply flip a coin, and (2) the *Oracle* is the upper bound set by perfect predictions.

The AutoQA results presented in Table 6 must be referred to the *Random* baseline. A *Random* AutoQA would have no

### Knowledge

You are a vision-language assistant responsible for assessing the quality of synthetically generated images. You have expertise in professional photography for e-commerce and design. You will receive a question and your task is to answer with the most appropriate score.

#### **Objective**

You are assessing the quality of a synthetically generated image depicting a {product\_type}. This image is generated by adding a background to an image of a {product\_type}. The main {product\_type} is the primary object of the image. The background is generated by a text-to-image model.

Table 5. Text prompts for System knowledge and objective.

	Defect(✗)			Clean(✓)			
Configuration	$y_{AQA}$	$P_{AQA}\uparrow$	$R_{AQA}\uparrow$	$y_{AQA}$	$P_{AQA}\uparrow$	$R_{AQA}\uparrow$	
Random (0.5)	0.521	0.809	0.521	0.500	0.184	0.465	
Cascade (AG)	0.761	0.848	0.793	0.239	0.297	0.380	
Cascade (AG NP)	0.847	0.823	0.858	0.153	0.241	0.197	
Single (AG)	0.882	0.835	0.916	0.118	0.400	0.211	
Oracle	0.813	1.000	1.000	0.187	1.000	1.000	

Table 6. Performance metrics on 380 test images for three AutoQA configurations.

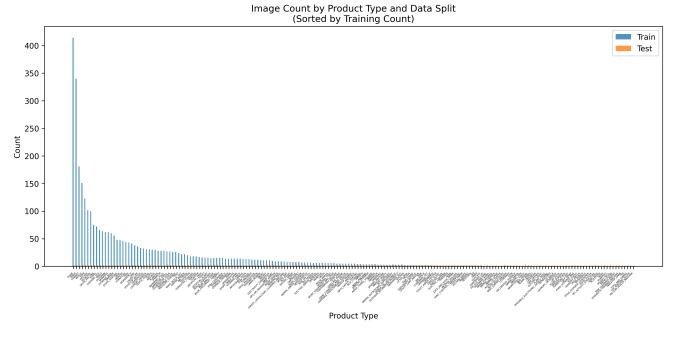


Figure 4. Histogram by object category of the train and test partition used to in the All Defects experiment.

effect on the quality of the generated images of the pipeline in Figure 1. Mathematically, this is equivalent to achieving a precision  $P_{\rm AQA}$  identical to the generator's one,  $P_{\rm Gen} = P_{\rm AQA}$ . In our use case, the quality of the generator is depicted in Table 4 as the yield of a perfect Oracle AutoQA is  $y_{\rm AQA} = 0.187$ , . This value matches the precision  $P_{\rm AQA} = 0.184$  of the Random AutoQA. The small difference is due to  $P_{\rm Gen}$  being computed over the full 380 test images, but  $P_{\rm AQA}$  over

half ( $y_{AOA} = 0.5$ ) the samples passed the *Random* AutoQA.

Table 6 shows that the three AutoQA configurations significantly improve over the  $P_{\rm AQA}(\checkmark)$  of the random baseline. Similarly, their  $P_{\rm AQA}(X)$  may look high, but they are actually close to the random baseline because of the 81.3% of *Defect* cases in the data.

As proved in Equation 3, the total cost depends from both  $y_{AQA}(\checkmark)$  and  $P_{AQA}(\checkmark)$ , so we will focus on these two met-

rics. When we do, we observe that using Nova Pro in the cascade does not offer gains over the AutoGluon only cascade. Similarly to [12, 36] we conclude that the considered off-the-shelf VLM does not provide reliable judgments for IQA, and that some additional adaptations would be needed, as proposed in [35, 50]. Between the two *AutoGluon only* configurations, there is no clear winner when looking at Table 6. For this reason, we compare these two configurations in the next section.

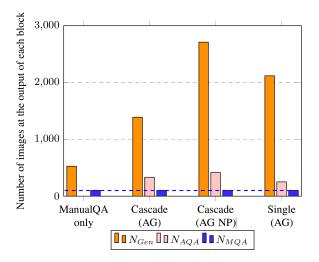


Figure 5. Volume of images needed after each block to obtain 100 high quality images.

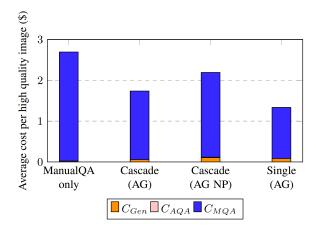


Figure 6. Cost composition over image generation ( $C_{Gen}$ ), AutoQA ( $C_{AQA}$ ), and ManualQA ( $C_{MQA}$ ).

#### 4.3. Cost savings

The characterization of AutoQA configurations in terms of their  $y_{\text{AQA}}(\checkmark)$  and  $P_{\text{AQA}}(\checkmark)$  allows estimating the cost savings, as presented in Equation 5. We use the measurements in Table 6 to compare the three AutoQA configurations in our use case.

As a first step, we calculate the number of images needed at the output of each block in the pipeline, as developed in Section 3.1. For simplicity, we consider a hypothetical case were the requirement is obtaining  $N_{\rm MQA}=100$  high quality images. We first leverage Equation 1 and  $P_{\rm AQA}(\checkmark)$  to obtain  $N_{AQA}$ , and afterwards Equation 2 and  $y_{\rm AQA}(\checkmark)$  to estimate  $N_{Gen}$ .

Figure 5 depicts the amount of  $N_{Gen}$  images that must be generated, and the amount of  $N_{AQA}$  images that must pass AutoQA, respectively.

The second step is weighting and aggregating the unitary cost of each stage with the volume of generated images, as presented in Equation 5. This requires establishing the unitary costs  $c_i$ , for  $i \in \{Gen, AQA, MQA\}$ :

**Image generation cost** ( $c_{Gen}$ ): The unitary cost per generated lifestyle image is of \$0.00400 per image approximately, empirically estimated based on our experience of producing images on a cloud server.

**AutoQA cost**  $(c_{AQA})$ : The cost of running inference with AutoGluon is negligible compared to the rest of the costs. However, the cost is not negligible if we consider one API call to Nova Pro on AWS Bedrock, which we approximate by \$0.00041 / image, assuming 3,000 input tokens and 300 output tokens.

**ManualQA cost** ( $c_{MQA}$ ): We follow the current suggestion in *Amazon SageMaker Ground Truth pricing*  $^1$  of 0.012 for image classification tasks. We consider each of the 13 detailed defects defined as an image classification task, an we also multiply by 3, as we deal with three annotations per image. This makes \$0.468, which we round up to  $c_{MQA} = \$0.5$ . The reader is referred to [23] for discussion over the challenges of manual annotations.

The individual costs allow estimating the aggregated costs for the four considered set ups, as plotted in Figure 6. The height of the bars show how the simple Single (AG) is the cheapest configuration, offering a significant reduction of 51.61% with respect to the ManualQA only baseline. Keep in mind that these high savings are estimated on a test set whose label distribution does not follow the training one. If both train and test partitions followed the same label distribution, we would expect to obtain a higher AutoQA precision, that would translate in even higher cost savings. The composition of the plot bars show how the aggregated cost is clearly dominated by ManualQA, which represents 99.5% of the total.

Finally, we focus on the impact of  $P_{AQA}(\checkmark)$ , plotting how cost savings would evolve for other values beyond the 0.4 of our use case. The plot in Figure 7 shows how a very low  $P_{AQA}(\checkmark)$  would make the AutoQA system harmful to a baseline without AutoQA, mostly because more images would need to be manually reviewed. This effect is mostly represented by the last term of Equation 5, where the cost

<sup>1</sup>https://aws.amazon.com/sagemaker-ai/groundtruth/pricing/

savings coming from ManualQA will become negative when  $P_{AQA}(\checkmark) < P_{Gen}(\checkmark)$ . Additionally, higher generation and AutoQA costs would also result from a low  $P_{AQA}(\checkmark)$ .

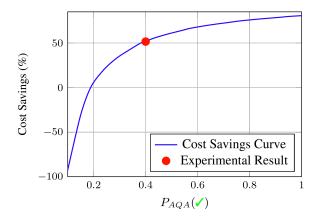


Figure 7. Cost savings as a function of AutoQA precision  $P_{AQA}(\checkmark)$  with  $P_{Gen}(\checkmark)$  set at 0.187, and  $r_{AQA}(\checkmark)$  at 0.118. The red dot shows the experimental result from our Single (AG only) solution.

### 4.4. Qualitative Results

includes The full predictions in the test set are provided in Figures 8-11. A visual inspection of the data does not show evidences of biases towards certain objects or backgrounds.

In the early stages of our research, we did observe that the VLM-based AutoQA was biased to only allow images with very simple backgrounds, which were less valuable for our target e-commerce application. For this reason, we extended the prompt to include the detailed defect *Rich background* shown in Table 4. When we experimented with AutoGluon we no longer observed this issue.

#### 5. Conclusions

We have shown how introducing AutoQA to an image generation pipeline can bring significant cost savings. The formula derived estimates these savings based on the AutoQA yield and precision. While we have applied this formula to the use case of image generation, it is valid for any GenAI task.

Our study case for background inpainting has shown significant cost savings of 51.61% even with a modest AutoQA precision of 0.4. This is because the cost of ManualQA clearly dominates over the costs of the automatic blocks of the pipeline. AutoQA comes almost for free, and it increases the quality of the images sent for AutoQA. In our best configuration, AutoQA only approved 11.8% of the generated images but, at the same time, the ManualQA yield increased from 18.7% to 40.0%. As a consequence, annotators need to review less images to reach a certain goal, and the total cost decreases.

The proposed technical solution is simple and based in



Figure 8. Full set of True Defect (X) predictions.

AutoML, which allows grounds for improvement based on the existing literature on IQA. In our set up, the zero-shot



Figure 9. Full set of False Defect (X) predictions.



Figure 10. Full set of True Clean (✓) predictions.



Figure 11. Full set of False Clean (✓) predictions.

VLMs mostly could not detect the defects in the images, a limitation aligned with existing works that question the effectiveness of this set up for VQA problems [25] and image quality assessment [12, 16, 49]. A fine-tuning of these models should be explored to improve their performance.

The significant cost saving achieved in our work motivates further scientific research. This research can be oriented in two directions: improving the image generator or the AutoQA engine. In any case, the ultimate goal is completely removing the need of a manual review in the pipeline.

### References

- [1] Amazon. The amazon nova family of models: Technical report and model card. *Amazon Technical Reports*, 2024. 4
- [2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE* international conference on computer vision, pages 2425– 2433, 2015. 3
- [3] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *International Conference on Learning Representations*, 2018. 2
- [4] Damon M. Chandler and Sheila S. Hemami. Vsnr: A wavelet-based visual signal-to-noise ratio for natural images. *IEEE Transactions on Image Processing*, 16(9):2284–2298, 2007.

- [5] Junsong Chen, Jincheng YU, Chongjian GE, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *The Twelfth International Conference on Learning Representa*tions, 2024. 1
- [6] Manri Cheon, Sung-Jun Yoon, Byungyeon Kang, and Junwoo Lee. Perceptual image quality assessment with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 433–442, 2021. 2
- [7] Jaemin Cho, Yushi Hu, Jason Michael Baldridge, Roopal Garg, Peter Anderson, Ranjay Krishna, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. In *ICLR*, 2024. 3
- [8] Marcos V Conde, Maxime Burchi, and Radu Timofte. Conformer and blind noisy students for improved image quality assessment. In *CVPR*, pages 940–950, 2022. 2
- [9] Ciprian Corneanu, Raghudeep Gadde, and Aleix M. Martinez. Latentpaint: Image inpainting in latent space with diffusion models. In WACV, 2024. 3
- [10] Wenliang Dai, Junnan Li, Dongxu LI, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose visionlanguage models with instruction tuning. *NeurIPS*, 36:49250– 49267, 2023. 3
- [11] Sebastian Hartwig, Dominik Engel, Leon Sick, Hannah Kniesel, Tristan Payer, Poonam Poonam, Michael Glöckler, Alex Bäuerle, and Timo Ropinski. A survey on quality metrics for text-to-image generation. *arXiv preprint arXiv:2403.11821*, 2024. 2
- [12] Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, et al. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. In EMNLP, 2024. 8, 10
- [13] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *EMNLP* (1), 2021. 2, 3
- [14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. NeurIPS, 30, 2017. 2
- [15] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20406–20417, 2023. 2, 3
- [16] Yipo Huang, Quan Yuan, Xiangfei Sheng, Zhichao Yang, Haoning Wu, Pengfei Chen, Yuzhe Yang, Leida Li, and Weisi Lin. Aesbench: An expert benchmark for multimodal large language models on image aesthetics perception. arXiv preprint arXiv:2401.08276, 2024. 10
- [17] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. Advances

- in Neural Information Processing Systems, 36:36652–36663, 2023. 2, 3
- [18] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In European Conference on Computer Vision, pages 366–384. Springer, 2024. 3
- [19] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 3
- [20] 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.
- [21] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF inter*national conference on computer vision, pages 4195–4205, 2023. 1
- [22] 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, 4
- [23] Tim Rädsch, Annika Reinke, Vivienn Weru, Minu D Tizabi, Nicholas Heller, Fabian Isensee, Annette Kopp-Schneider, and Lena Maier-Hein. Quality assured: Rethinking annotation strategies in imaging ai. In ECCV, pages 52–69. Springer, 2024. 8
- [24] Abdul Rehman and Zhou Wang. Reduced-reference image quality assessment by structural similarity estimation. *IEEE Transactions on Image Processing*, 21(8):3378–3389, 2012.
- [25] Candace Ross, Melissa Hall, Adriana Romero-Soriano, and Adina Williams. What makes a good metric? evaluating automatic metrics for text-to-image consistency. In *First Conference on Language Modeling*, 2024. 10
- [26] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. Advances in Neural Information Processing Systems, 29, 2016.
- [27] Xiangfei Sheng, Leida Li, Pengfei Chen, Jinjian Wu, Weisheng Dong, Yuzhe Yang, Liwu Xu, Yaqian Li, and Guangming Shi. Aesclip: Multi-attribute contrastive learning for image aesthetics assessment. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1117–1126, 2023. 2
- [28] Yiqing Shi, Yuzhen Niu, Wenzhong Guo, Yize Huang, and Jiamei Zhan. Pairwise learning to rank for image quality assessment. *IEEE Access*, 8:192352–192367, 2020. 2
- [29] Zhiqiang Tang, Haoyang Fang, Su Zhou, Taojiannan Yang, Zihan Zhong, Tony Hu, Katrin Kirchhoff, and George Karypis. Autogluon-multimodal (automm): Supercharging multimodal automl with foundation models. 2024. 4
- [30] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image

- generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2025. 1
- [31] Jing Wang, Haotian Fan, Xiaoxia Hou, Yitian Xu, Tao Li, Xuechao Lu, and Lean Fu. Mstriq: No reference image quality assessment based on swin transformer with multistage fusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pages 1269–1278, 2022. 3
- [32] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. 2
- [33] Ross Wightman. Pytorch image models. https: //github.com/rwightman/pytorch-image-models, 2019. 4
- [34] Olivia Wiles, Chuhan Zhang, Isabela Albuquerque, Ivana Kajić, Su Wang, Emanuele Bugliarello, Yasumasa Onoe, Chris Knutsen, Cyrus Rashtchian, Jordi Pont-Tuset, et al. Revisiting text-to-image evaluation with gecko: On metrics, prompts, and human ratings. arXiv preprint arXiv:2404.16820, 2024.
- [35] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Kaixin Xu, Chunyi Li, Jingwen Hou, Guangtao Zhai, et al. Q-instruct: Improving low-level visual abilities for multi-modality foundation models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 25490–25500, 2024. 8
- [36] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. In *International Conference* on Machine Learning, pages 54015–54029. PMLR, 2024. 8
- [37] 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. *CoRR*, 2023. 2
- [38] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-toimage models with human preference. In *Proceedings of* the IEEE/CVF International Conference on Computer Vision, pages 2096–2105, 2023. 2, 3
- [39] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: learning and evaluating human preferences for text-to-image generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 15903–15935, 2023. 2, 3
- [40] Yuzhe Yang, Liwu Xu, Leida Li, Nan Qie, Yaqian Li, Peng Zhang, and Yandong Guo. Personalized image aesthetics assessment with rich attributes. In *Proceedings of the IEEE/CVF* Conference on Computer Vision and Pattern Recognition (CVPR), pages 19861–19869, 2022. 2
- [41] Michal Yarom, Yonatan Bitton, Soravit Changpinyo, Roee Aharoni, Jonathan Herzig, Oran Lang, Eran Ofek, and Idan Szpektor. What you see is what you read? improving textimage alignment evaluation. *Advances in Neural Information Processing Systems*, 36:1601–1619, 2023. 2

- [42] Ran Yi, Haoyuan Tian, Zhihao Gu, Yu-Kun Lai, and Paul L Rosin. Towards artistic image aesthetics assessment: a large-scale dataset and a new method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22388–22397, 2023. 2
- [43] Junyong You and Jari Korhonen. Transformer for image quality assessment. In 2021 IEEE international conference on image processing (ICIP), pages 1389–1393. IEEE, 2021.
- [44] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*, 2022. 1
- [45] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386, 2011.
- [46] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 4
- [47] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In CVPR, 2018. 2
- [48] Sixian Zhang, Bohan Wang, Junqiang Wu, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. Learning multi-dimensional human preference for text-to-image generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8018–8027, 2024. 2
- [49] Zicheng Zhang, Haoning Wu, Erli Zhang, Guangtao Zhai, and Weisi Lin. Q-bench: A benchmark for multi-modal foundation models on low-level vision from single images to pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 10
- [50] Zicheng Zhang, Yingjie Zhou, Chunyi Li, Baixuan Zhao, Xiaohong Liu, and Guangtao Zhai. Quality assessment in the era of large models: A survey. ACM Transactions on Multimedia Computing, Communications and Applications, 2025. 8