Reliable Image Quality Evaluation and Mitigating Quality Bias in Generative Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

Discrepancies in generation quality across demographic groups pose a substantial and critical challenge in image generative models. However, the Fréchet Inception Distance (FID) score, which is widely used as an image quality evaluation metric for generative models, introduces unintended bias when assessing quality across sensitive attributes. This undermines the reliability of the evaluation procedure. This paper addresses this limitation by introducing the **Difference in Quality Assessment (DQA)** score, a novel approach that quantifies the reliability of existing evaluation metrics, e.g. FID. DQA assesses discrepancies in evaluated quality across demographic groups under strictly controlled conditions to effectively gauge metric reliability. Our findings reveal that traditional quality evaluation metrics can yield biased assessments across groups due to inappropriate reference set selection and inherent biases in image encoder in FID. Furthermore, we propose DQA-Guidance within diffusion model sampling to reduce quality disparities across groups. Experimental results demonstrate the utility of the DQA score in identifying biased evaluation metrics and present effective strategies to mitigate these biases. This work contributes to the development of reliable and fair evaluation metrics for generative models and provides actionable methods to address quality disparities in image generation across groups.

1 Introduction

In recent years, image generative models such as Generative Adversarial Networks (GANs) [21], Denoising Diffusion Probabilistic Models (DDPMs) [27], and text-to-image generation [47, 49] systems have brought bias concerns to the forefront of generative modeling. While substantial research has focused on distributional fairness to ensure balanced sample generation across sensitive attributes [13, 54, 37, 42, 30], the **fairness in generation quality** across demographic groups remains an equally critical yet underexplored issue. For example, Fig. 1 demonstrates the existing bias in generation quality by producing better quality of image for certain demographic group.

Furthermore, in the classification task, text-to-image generative models can be used as data augmentation tools to improve classifier performance [32]. However, if the quality of generated images is inconsistent across demographic groups, it can negatively impact classification performance for certain groups, exacerbating fairness issues in prediction and introducing biases in decision-making. We empirically demonstrate in Appendix B that discrepancies in image generation quality can adversely affect real-world applications, e.g. medical imaging [20], particularly in classification performance and fairness [35]. We also show that achieving fair quality in generated images can lead to improved outcomes, underscoring the necessity of addressing this issue.

In response, recent studies [44, 41] have highlighted quality discrepancies in generative models related to gender-profession biases, relying on the Fréchet Inception Distance (FID) [26] to assess

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

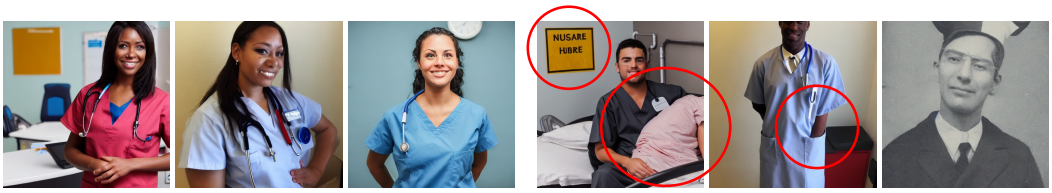


Figure 1: Using the same prompt template and seed, a generative model may produce varying image quality across different demographic groups, e.g., generating higher-quality nurse images for females while producing obscured objects, distorted limbs, or grayscale images for male nurses.

the quality of generated images. However, our analysis reveals that FID is unreliable for evaluating fairness in image quality for two reasons.

First, FID is sensitive to the selection of reference dataset due to distinct group distributions. As demonstrated in our synthetic data analysis in Sec. 3 and Fig. 3, the reference should be chosen group-specific manner. Choosing combined dataset as reference for FID not only leads to inaccurate quality evaluations for each group but also misidentifies the direction of bias, making FID an unreliable metric for detecting fairness issues in generative models observed in [44, 41].

Secondly, even with group-specific evaluation, traditional encoders can remain unreliable due to inherent biases in image encoders, which may produce inconsistent representations for images of similar quality across demographic groups. For example, as shown in Fig. 2, biased encoders such as InceptionV3 [56] and CLIP [46] yield unreliable evaluation results, misassessing certain demographic groups as having better image quality. We identify that this inconsistency arises from the biased representations produced by the encoder. To validate this issue, we use a t-SNE [59] plot of embeddings from a biased encoder, shown in Fig. 4 (b). The plot reveals a clear gender-based separation despite similar image quality, highlighting the encoder's failure to reliably evaluate quality discrepancies across demographic groups. Further details are provided in Sec. 3.2.

In summary, although quality bias exists in generative models, the commonly used evaluation metric, FID, and potential alternatives leveraging different backbone networks [29] are not reliable for assessing this bias. This raises the following key questions:

Q1: Which image encoder for evaluation metric can reliably assess quality bias, and how can it be quantified?

Q2: What strategies can effectively mitigate quality bias in generative models?

To address the first question, we introduce a novel score, the *Difference in Quality Assessment* (DQA), which serves as a **reliability score** for assessing the reliability of evaluation metrics' fairness across demographic groups. DQA quantifies whether an encoder introduces bias, by measuring discrepancies in evaluation results across demographic groups based on strictly controlled test dataset. An encoder with a lower DQA value is interpreted as more reliable and suitable for group-specific quality assessments to be used as an evaluation metric for image quality. DQA can identify the most reliable pre-trained foundational models in quality evaluation in Sec. 4, supporting fairness and reliability in future generative model applications for downstream tasks.

Furthermore, to address the second question, we propose a DQA-based regularization method, **DQA-Guidance** for diffusion models' sampling stage, which enhances both quality fairness and overall generation quality without re-training the diffusion model, as discussed in Sec. 5.

Overall, unlike prior work that has predominantly focused on distributional fairness, this study is the first to systematically address fairness in generation quality by introducing:

- **a reliable diagnostic tool (DQA)** to evaluate quality bias across demographic groups, and
- **a practical mitigation strategy (DQA-Guidance)** to reduce quality disparities during image generation.

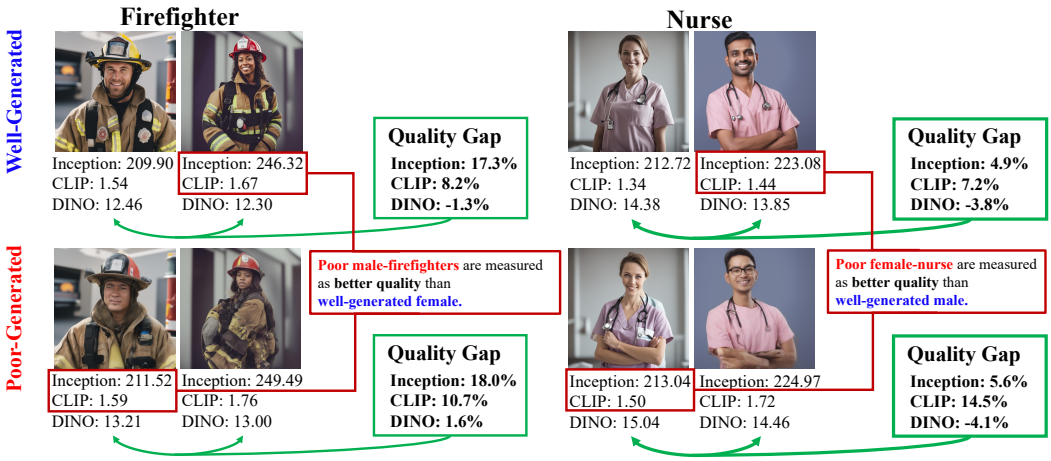


Figure 2: Using the same distance metric (Fréchet Distance, smaller is better), we compare image quality across varying professions and genders, with each set consisting of 1,000 images. Each image set is carefully controlled to include both well-generated and poorly-generated images. We evaluate image quality with three image encoders: InceptionV3 (FID), CLIP, and DINO. A biased encoder in quality evaluation leads to two forms of unreliable measurement. First, as shown in **green box**, InceptionV3 and CLIP exhibit significant measurement gaps across demographic groups for images of the same quality, whereas DINO shows relatively smaller discrepancies. Second, as shown in **red box**, InceptionV3 and CLIP misleadingly assess poor-quality images as having better quality for certain gender-profession subset, while DINO more accurately reflects true quality assessments.

2 Related Work

2.1 Generated Image Quality Assessment

FID is a widely used metric for assessing the quality of generated images by measuring the Wasserstein-2 distance [60] between embeddings of synthetic and real images extracted by the InceptionV3 [56]. This embedding-based distance measurement has thus become standard in generative model research [52, 33, 62, 3]. To enhance representational richness and relax distributional assumptions, MMD with the CLIP encoder [46] has been proposed [29]. While prior studies [5, 14, 28] have highlighted the unreliability of evaluation metrics under finite or imbalanced sample conditions, the reliability of these metrics from a fairness perspective remains largely unexplored.

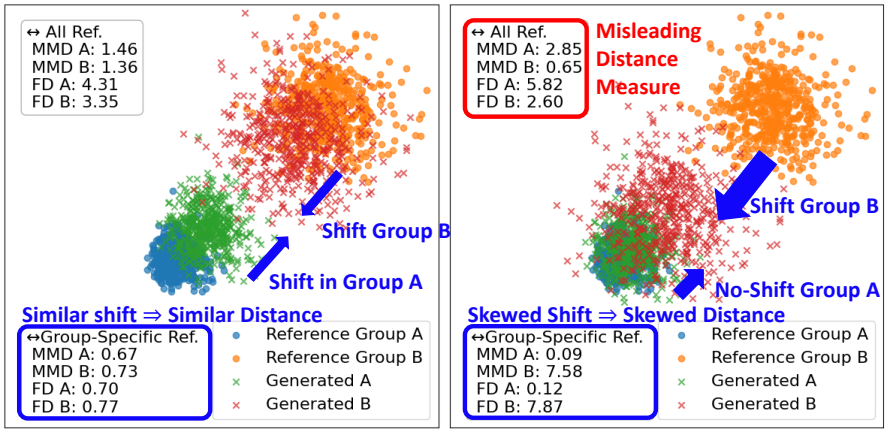
2.2 Fairness in Generative Models

Many studies have explored fairness in generative models but have primarily focused on addressing distributional bias, aiming to achieve an equal number of generated samples across demographic groups from a neutral prompt such as fine-tuning the entire model [13, 54], utilizing a pretrained classifier [37, 42], and manipulating intermediate embeddings [30]. Some works concentrated on new metric evaluating such biases [12, 51].

In contrast, beyond distributional bias, Perera and Patel [44] and Naik and Nushi [41] highlighted that quality bias in generated images across demographic groups, particularly in associating certain careers with specific genders. However, methods for mitigating quality bias have not been presented in the literature. We are the first to propose guiding the diffusion model's sampling stage to ensure fairness in image quality.

3 Bias in Image Quality Assessment for Generative Models

Recent studies have highlighted concerns about quality bias in generated images [44, 41]. To evaluate the quality of generated images and quantify this bias, the Fréchet Inception Distance (FID) [26] is



(a) Example of Fair Image Encoder (b) Example of Unfair Image Encoder

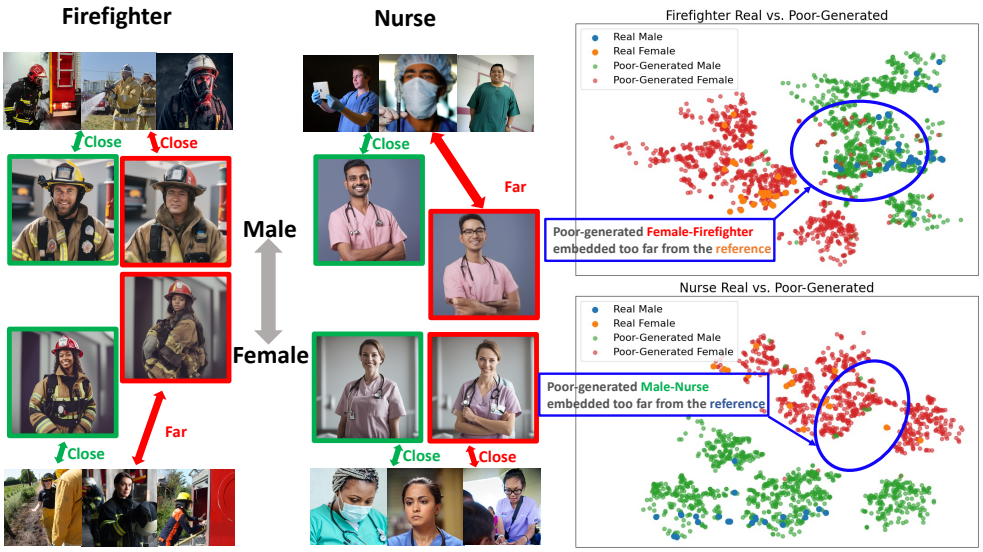
Figure 3: Illustration of quality bias in evaluation metrics using distance measures such as Maximum Mean Discrepancy (MMD) and Fréchet Distance (FD). The left figure depicts a fair scenario where generated data embeddings for both groups exhibit the same distribution shift, while the right figure shows an unfair scenario, with embeddings for one generated group skewed towards the other. Using group-specific references (e.g., $A_{gen} \leftrightarrow A_{ref}$) more accurately captures distribution shifts compared to an all-reference approach (e.g., $A_{gen} \leftrightarrow A_{ref} \cup B_{ref}$), which can produce misleading values in cases of biased image encoders. Thus, group-specific distance measures more accurately evaluate the quality under the biased representation.

widely used as a metric for assessing the similarity between the distributions of real and generated images. FID calculates the statistical distance between embeddings extracted from the InceptionV3 model [56] for both generated images and a reference dataset [7, 19, 50, 45]. However, relying on FID for quality evaluation has significant limitations, as discussed below.

3.1 Selection of Reference Dataset

Firstly, the measurement method should be group-specific to accurately capture differences across demographic groups. To formalize, let $D(\cdot, \cdot)$ denote a distance measurement such as Maximum Mean Discrepancy (MMD) [46] or Fréchet Distance (FD), and let f represent an image encoder. Define two demographic groups A and B, with corresponding reference datasets, A_{ref} and B_{ref}, and generated datasets, A_{gen} and B_{gen}. The combined reference and generated datasets are given by $\mathcal{I}_{\text{ref}} = A_{\text{ref}} \cup B_{\text{ref}}$ and $\mathcal{I}_{\text{gen}} = A_{\text{gen}} \cup B_{\text{gen}}$. In FID, D represents FD while f is typically the InceptionV3 model [56]. In the quality bias literature [44, 41], the generation quality of each group is calculated by $D(f(A_{\text{gen}}), f(\mathcal{I}_{\text{ref}}))$ and $D(f(B_{\text{gen}}), f(\mathcal{I}_{\text{ref}}))$ for groups A and B, respectively, while the bias measurement is given by $D(f(A_{\text{gen}}), f(\mathcal{I}_{\text{ref}})) - D(f(B_{\text{gen}}), f(\mathcal{I}_{\text{ref}}))$. Here, the magnitude represents the degree of bias, while the sign indicates its direction.

However, as demonstrated in our synthetic data analysis in Fig. 3, using a unified reference dataset can mask or amplify biases, potentially leading to unfair assessments of image quality across different groups. In this figure, blue and orange points represent reference embeddings for two demographic groups, while green and red points denote generated embeddings for each group. Fig. 3 (a) depicts a scenario where the embeddings of the generated data are similarly out-of-distribution from their respective reference datasets, suggesting a fair assessment. In contrast, Fig. 3 (b) shows a scenario where the generated data embeddings for one group are skewed toward the other group's reference data, indicating potential quality bias. According to Fig. 3 (b), the quality evaluation results for group B should be worse (higher) than for group A. However, when using the combined reference set, as denoted as "All Ref.", the measured distances indicate $D(f(A_{\text{gen}}), f(\mathcal{I}_{\text{ref}})) \gg D(f(B_{\text{gen}}), f(\mathcal{I}_{\text{ref}}))$, which is misleading. In contrast, in Fig. 3 (b), using group-specific references yields $D(f(A_{\text{gen}}), f(A_{\text{ref}})) \ll D(f(B_{\text{gen}}), f(B_{\text{ref}}))$, providing an accurate evaluation. Thus, the quality bias evaluation should be $D(f(A_{\text{gen}}), f(A_{\text{ref}})) - D(f(B_{\text{gen}}), f(B_{\text{ref}}))$, in a group-specific manner, rather than $D(f(A_{\text{gen}}), f(\mathcal{I}_{\text{ref}})) - D(f(B_{\text{gen}}), f(\mathcal{I}_{\text{ref}}))$.



(a) Example of Unreliable Image Quality Evaluation (b) t-SNE Visualization of Unreliable Image Quality Evaluation

Figure 4: (a) Images in **green** boxes represent "good" quality generated images, while **red** boxes denote "poor" quality images. A biased encoder embeds poor-quality images by associating specific genders with certain professions, leading to skewed evaluation results as these images are unfairly placed far from their respective reference groups. (b) The t-SNE visualization using a CLIP [46] image encoder illustrates this issue. Poor-quality images for certain gender and profession (e.g., red boxes in (a)) demonstrate a tendency for being embedded within the wrong gender cluster, resulting in biased evaluation outcomes despite similar quality levels.

3.2 Bias in Image Encoder Used in Evaluation

Secondly, when discrepancies in group-specific quality evaluations are observed, it remains unclear whether these differences stem from actual variations in image quality or from biases inherent in the image encoder. A biased encoder can distort embeddings, impacting the interpretation of image quality across groups and leading to skewed evaluation results, as observed in Fig. 2. We illustrate this issue in Fig. 4 (a), and verify this in Fig. 4 (b) using t-SNE plot. In Fig. 4 (b), although well-generated images are correctly located closer to each reference (See Appendix C), a poorly generated image of a "male nurse" may be embedded closer to the "female nurse" reference due to encoder bias, rather than reflecting its true quality. Conversely, a similarly poor-quality image of a "female nurse" remains within the in-distribution region of the "female nurse" reference, indicating inconsistency in quality evaluation across demographic groups. This leads to inaccuracies in both quality assessment and quality bias evaluation, such that $|D(f(A_{\text{gen}}), f(A_{\text{ref}})) - D(f(B_{\text{gen}}), f(B_{\text{ref}}))| \gg 0$, even though $TrueQuality(A_{\text{gen}}) \approx TrueQuality(B_{\text{gen}})$. The t-SNE plot for well-generated images is shown in Appendix C, further highlighting the unreliability of the image encoder with respect to image quality.

Given these limitations, it is crucial to identify evaluation metrics that can reliably distinguish between distribution shifts caused by actual quality discrepancies and those stemming from biases in the image encoder. By employing group-specific measurement and introducing a reliability score for evaluation metrics using a dataset with controlled quality, we gain a clearer understanding of the sources of quality bias and can improve the fairness and accuracy of image quality assessments across different demographic groups.

4 Reliability of Evaluation Metric for Generated Image Quality

In this section, we introduce a novel method to assess the reliability of evaluation metrics for generated image quality, focusing primarily on metrics that measure the distributional distance between generated and reference datasets. This emphasis arises from concerns that biased image encoders might handle poor-quality images inconsistently across sensitive groups, even when distances are calculated in a group-specific manner, as discussed in Sec. 3.1 and Sec. 3.2.

4.1 Difference in Quality Assessment

We consider two generated datasets, A_{gen} and B_{gen}, each containing images of comparable quality and equal quantity. In our experiments, we use MMD as a distance metric $D(\cdot, \cdot)$ instead of FD due to its efficiency and freedom from distributional assumptions [29]. *Difference in Quality Assessment* (DQA) aims to identify bias in the evaluation metric $D(f(\cdot), f(\cdot))$. Recalling the combined reference and generated datasets as $\mathcal{I}_{\text{ref}} = A_{\text{ref}} \cup B_{\text{ref}}$ and $\mathcal{I}_{\text{gen}} = A_{\text{gen}} \cup B_{\text{gen}}$, DQA is formulated as:

$$\text{DQA} = \frac{\left| D\big(f(A_{\text{gen}}), f(A_{\text{ref}})\big) - D\big(f(B_{\text{gen}}), f(B_{\text{ref}})\big) \right|}{D\big(f(\mathcal{I}_{\text{gen}}), f(\mathcal{I}_{\text{ref}})\big)} \tag{1}$$

By employing group-specific distance measurements, Eq. (1) isolates the bias inherent in the encoder by comparing the embeddings of generated images with consistent quality across different demographic groups. The numerator captures the difference in quality between generated data for groups A and B relative to their respective reference sets. A large numerator implies significant quality disparity between groups, whereas a small or zero value suggests that the encoder treats both groups equally. The denominator captures the global generation quality by measuring the distance between the combined reference and generated datasets. A smaller denominator value indicates generated data closely matches the reference set, while a larger value signifies deviation. Hence, DQA quantifies the relative quality discrepancy between groups compared to the overall distribution shift in generation. A low DQA suggests fair treatment of both groups by the encoder, while a high DQA indicates significant bias. Therefore, DQA serves as a **reliability score** for quantifying bias in image encoders.

4.2 Constructing the Evaluation Dataset for DQA

To effectively apply the DQA score for finding reliable image encoders in practice, it is essential to construct controlled reference and generated datasets. To assess the reliability of image encoders, we construct a dataset with six different versions, ranging from well-generated to poorly generated sets, capturing realistic scenarios encountered in text-to-image generation of human images using Stable Diffusion XL (SDXL) [45]. Following the recommended settings from [40] as our baseline, we degrade image quality in various ways by adjusting hyperparameters. The scenarios include the baseline, weak guidance, reduced sampling steps in diffusion, increased noise influence on the initial image, and the absence of refinement methods. The baseline serves as the reference dataset, while the other scenarios represent controlled generated datasets. For each image seed, we prepare datasets under all six scenarios. We generate 250 images for each combination of profession, gender, and race, resulting in 20,000 images per scenario (10 professions, 2 genders, and 4 races). This ensures that each attribute has the same number of reference images, avoiding inaccuracies caused by imbalanced attribute distributions [28]. Detailed descriptions of each degradation, along with the professions and races used, are provided in Appendix D, and visualizations are presented in Fig. 5 (a).

4.3 DQA for Multiple Attributes (e.g., Race)

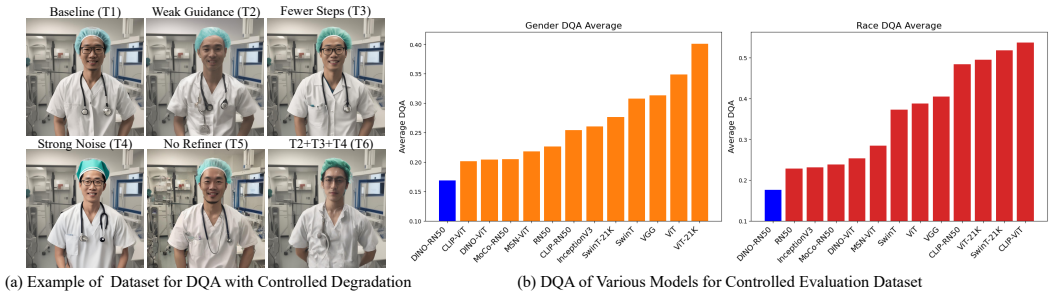
Let Eq.(1) be denoted as $\text{DQA}(A_{gen}, B_{gen}; f)$ for groups A and B given encoder f. Let $\mathcal{G} = \{G_1, \cdots, G_n\}$ represent the set of n groups. We aggregate pairwise DQA across all combinations to provide a comprehensive measure of fairness in image quality assessment across multiple attributes.

$$\text{AvgDQA}(\mathcal{G}) = \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} \text{DQA}(G_i, G_j; f), \tag{2}$$

4.4 Reliability Analysis for Pre-trained Image Encoders

To assess the reliability of image encoders in evaluating generated image quality fairly across demographic groups, we apply the DQA score to various pre-trained models, considering differences in architecture, training dataset, and training scheme. In this analysis, we calculate the average DQA score across all degradation types.

We evaluate models including InceptionV3, VGG [55], ResNet-50 (RN50) [24], ViT-B/16 [17], and Swin Transformer [39], all trained on the ImageNet-1K (IN-1K) [16] dataset using supervised learning. We also compare models trained on IN-1K and IN-21K [48] for ViT-B/16 and Swin Transformer architectures to examine the effect of training dataset size. Additionally, we explore different training



(a) Example of Dataset for DQA with Controlled Degradation

(b) DQA of Various Models for Controlled Evaluation Dataset

Figure 5: (a) Examples of generated images under controlled degradation scenarios. The figure illustrates samples from both the well-generated baseline (reference, T1) and the intentionally degraded cases (T2 - T6), where image quality is systematically reduced by adjusting specific hyperparameters. This controlled degradation enables effective measurement of the DQA score to assess the reliability of an image encoder. (b) Across all pre-trained encoders and various degradation in generated images, DINO-RN50 achieves the lowest DQA in average for both gender and race bias, indicating it is the most reliable encoder for evaluating the quality of generated images.

schemes by evaluating models trained with self-supervised methods like MoCo-RN50 [25], MSN-ViT [1], and DINO [8] and CLIP using both RN50 and ViT-B/16 architectures.

Impact of Training Scheme on DQA. Our results, summarized in Fig. 5 (b) indicates that self-supervised models using the RN50 architecture, particularly DINO-RN50 and MoCo-RN50, achieve the lower DQA scores in general compared to supervised models. This suggests that the combination of self-supervised learning and the RN50 architecture effectively reduces bias, leading to fairer embeddings across demographic groups. We analyze this as self-supervised models learn representations without explicit labels, which helps them avoid inheriting biases tied to label information.

Impact of Backbone Network on DQA. In contrast, self-supervised models using the ViT architecture, such as DINO-ViT and MSN-ViT, exhibit slightly higher DQA scores, implying that RN50 may be better suited for learning unbiased representations in self-supervised settings. We analyze the architectural differences between convolutional neural networks (CNNs) [53] and Transformers [61]. RN50, as a CNN, incorporates locality and spatial patterns through its convolutional layers. This structure allows CNNs to capture both local and global image features, making them more robust to distortions in the image [58]. In contrast, Transformer-based models rely on self-attention mechanisms that process images as sequences of tokens, without the same spatial locality constraints [58]. The token-based approach enables the model to capture complex global dependencies, but it may also make it more sensitive to specific variations in distorted images [22], resulting in larger discrepancies between reference and generated datasets.

Impact of Training Dataset on DQA. We also examine the effect of training dataset size by comparing models trained on IN-1K and IN-21K for both ViT-B/16 and Swin Transformer. The results show that models trained on the larger dataset, IN-21K, actually exhibit higher DQA scores compared to their IN-1K counterparts. This suggests that increasing the dataset size alone does not necessarily improve fairness in the encoder's representations. Similarly, models like CLIP, despite being trained on large-scale image-text datasets, show higher DQA scores especially in racial bias, indicating that large-scale multimodal training does not necessarily guarantee fairness in embeddings.

4.5 Validity of DQA

To validate the effectiveness of DQA for quality assessment, we apply it to data augmentation in a medical image classification task. As detailed in Appendix B, datasets generated by text-to-image models for medical images can be used for data augmentation but often exacerbate fairness issues due to quality bias in the generative model, resulting in significant performance gaps across demographic groups in classification. Leveraging a reliable image encoder, we construct both fair and unfair generated datasets based on their DQA scores as detailed in Algorithm 1. Fair dataset enhances classification fairness when used for augmentation, whereas unfair dataset exacerbates disparities. This demonstrates DQA's ability to identify reliable image encoders and its practical utility in enabling

Table 1: Experimental results for generation quality and quality disparities with DQA-Guidance.

Method	Avg.MMD	Mean $\lvert D_{male} - D_{female} \rvert$	Max $\lvert D_{male} - D_{female} \rvert$
Baseline (Stable Diffusion)	109.93	12.57	17.77
+ DQA-Guidance ($\lambda_1 = 20, \lambda_2 = 100$)	103.89	**6.21**	**6.94**
+ DQA-Guidance ($\lambda_1 = 20, \lambda_2 = 1000$)	**85.72**	10.16	11.87

DQA-based data augmentation. These findings underscore the benefit of DQA in generative models for classification applications, as further elaborated in Appendix B.

5 Mitigating Quality Bias in Diffusion Models

DQA serves not only as a reliability indicator for the evaluation metric but can also act as an energy function in generative models to regularize equal image quality across demographic groups. Specifically, we employ guided diffusion [66, 38, 18, 2] during sampling in diffusion models rather than training a model from scratch. By interpreting DQA as an energy function, its gradient can be integrated into the diffusion sampling process following energy-based guidance principles, steering the generation process toward desired outcomes without modifying the pre-trained model parameters.

5.1 DQA-Guidance for Diffusion

In our context, the DQA score quantifies relative discrepancies in image quality assessments across demographic groups. By computing the gradient of DQA with respect to latent variables z_t at each diffusion timestep, we obtain the latent direction that reduces this discrepancy. Incorporating this gradient into noise prediction adjusts the sampling trajectory to favor samples that minimize quality differences across groups.

Assume we identify a reliable image encoder f^* for evaluating generated image quality. Let g be the base generative model that samples from latent variable z_t^A and z_t^B for each group. We apply DQA-Guidance in diffusion modeling by taking the gradient of DQA with respect to $z_t = [z_t^A; z_t^B]$:

$$\tilde{\epsilon}_\theta(z_t) = \epsilon_\theta(z_t) + \sigma_t \lambda_1 \nabla_{z_t} \text{DQA}(g(z_t^A), g(z_t^B); f^*), \tag{3}$$

where $\epsilon_\theta(z_t)$ is the estimated noise, θ represents the pre-trained weights of the diffusion model, σ_t scales the gradient term according to the noise level at timestep t, and λ_1 is a hyperparameter controlling the strength of the DQA-Guidance in diffusion process.

Since reducing DQA (Eq.(1)) could unintentionally increase the denominator (representing the overall quality), we introduce a regularizer to ensure that both the numerator and denominator are minimized. Specifically, we add the gradient of the denominator of DQA, the overall distributional distance between generated and reference datasets $D\big(f^*(\mathcal{I}_{\text{gen}}), f^*(\mathcal{I}_{\text{ref}})\big)$, to improve overall quality,

$$\tilde{\epsilon}_\theta(z_t) = \epsilon_\theta(z_t) + \sigma_t \nabla_{z_t} \Big(\lambda_1 \text{DQA}(g(z_t^A), g(z_t^B); f^*) + \lambda_2 D\big(f^*(\mathcal{I}_{\text{gen}}), f^*(\mathcal{I}_{\text{ref}})\big) \Big), \tag{4}$$

where λ_2 is a hyperparameter balancing the influence of the quality regularizer. Consequently, by treating DQA as an energy function and integrating the gradients of both DQA and the overall quality term into the sampling process, the model is guided to reduce quality disparities across demographic groups while preserving high image fidelity.

5.2 Experimental Details for DQA-Guidance

To verify the effectiveness of DQA-Guidance in mitigating quality bias, we conduct human image generation experiments using Stable Diffusion [49]. We utilize the well-generated (Baseline) dataset introduced in Appendix D, which contains images generated by the state-of-the-art SDXL model [45], as a reference set to maintain consistent quality and context across demographic groups during the diffusion process. DQA-Guidance is applied to Stable Diffusion to mitigate quality disparities while enhancing overall image quality. An extension of DQA-Guidance for medical image generation with ImageGen [50] is presented in Appendix H.

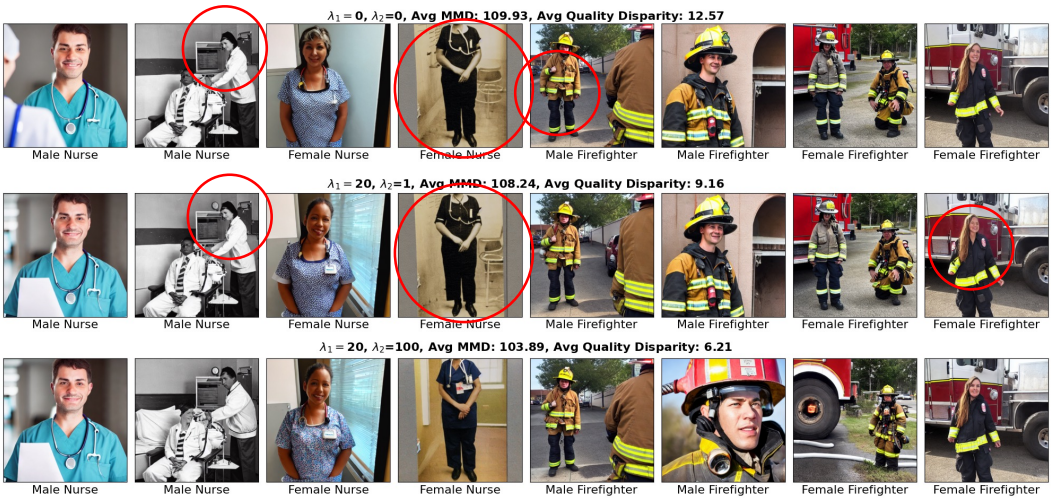


Figure 6: Qualitative results of DQA-Guidance. The examples demonstrate improvements in artifact reduction, color correction, and texture and background refinement (**red circles**). These enhancements illustrate the impact of DQA-Guidance in balancing quality across demographic groups.

5.3 Result Analysis for DQA-Guidance

Table 1 demonstrates the impact of DQA-Guidance on image generation. DQA-Guidance not only reduces quality disparities, but also substantially enhances overall image quality. These findings indicate that DQA serves not only as a reliable fairness evaluation metric but also as an effective regularizer when applied during the sampling stage of diffusion models. Moreover, increasing values of λ_2 are intuitively associated with improved generation quality. Qualitative results are presented in Fig. 6, highlighting reductions in artifacts and more balanced image quality across groups. The effects and sensitivity of the hyperparameters λ_1 and λ_2 are further discussed in the Appendix E.

5.4 Potential Limitation

DQA-Guidance is the first approach to demonstrate that quality bias-aware guidance can effectively steer diffusion models toward fairer outputs. As our work represents the first attempt to mitigate fairness issues in generative models from a quality perspective, Table 1 lacks comparative methods. We will continue seeking appropriate baselines for future evaluation.
In terms of practicality, DQA-Guidance introduces additional computational cost, requiring an auxiliary image encoder and gradient computation, which may increase memory usage. However, it opens a promising direction for fairness-aware generation, and future work may explore more efficient variants of this strategy.
Additionally, although our controlled dataset enables systematic evaluation, it may not fully reflect human-perceived image quality. Developing human-validated reference datasets, such as those based on perceptual surveys, would further enhance the validity of DQA and provide a more robust benchmark for auditing image evaluation metrics.

6 Conclusion and Societal Impact

This paper addresses the underexplored issue of quality disparities in image generation and introduces the Difference in Quality Assessment (DQA) score, a novel metric for evaluating the reliability of quality assessment methods. Through extensive analysis, we reveal that commonly used metrics, such as FID, can introduce unintended biases, resulting in misinterpretation of quality discrepancies. DQA mitigates these issues by identifying reliable image encoders, enabling fairer and more dependable quality evaluations. We further extend its utility through DQA-Guidance, which steers diffusion models toward reducing quality disparities while preserving image fidelity. Overall, our work offers an equitable and reliable generative AI, fostering responsible innovation in technologies that promote societal fairness and support decision-making.

References

[1] M. Assran, M. Caron, I. Misra, P. Bojanowski, F. Bordes, P. Vincent, A. Joulin, M. Rabbat, and N. Ballas. Masked siamese networks for label-efficient learning. In *European Conference on Computer Vision*, pages 456–473. Springer, 2022.

[2] A. Bansal, H.-M. Chu, A. Schwarzschild, S. Sengupta, M. Goldblum, J. Geiping, and T. Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 843–852, 2023.

[3] A. Bansal, E. Borgnia, H.-M. Chu, J. Li, H. Kazemi, F. Huang, M. Goldblum, J. Geiping, and T. Goldstein. Cold diffusion: Inverting arbitrary image transforms without noise. *Advances in Neural Information Processing Systems*, 36, 2024.

[4] L. Beyer, A. Steiner, A. S. Pinto, A. Kolesnikov, X. Wang, D. Salz, M. Neumann, I. Alabdulmohsin, M. Tschannen, E. Bugliarello, T. Unterthiner, D. Keysers, S. Koppula, F. Liu, A. Grycner, A. Gritsenko, N. Houlsby, M. Kumar, K. Rong, J. Eisenschlos, R. Kabra, M. Bauer, M. Bošnjak, X. Chen, M. Minderer, P. Voigtlaender, I. Bica, I. Balazevic, J. Puigcerver, P. Papalampidi, O. Henaff, X. Xiong, R. Soricut, J. Harmsen, and X. Zhai. PaliGemma: A versatile 3B VLM for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

[5] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

[6] A. Borji. Qualitative failures of image generation models and their application in detecting deepfakes. *Image and Vision Computing*, 137:104771, 2023.

[7] M. Brack, F. Friedrich, D. Hintersdorf, L. Struppek, P. Schramowski, and K. Kersting. Sega: Instructing text-to-image models using semantic guidance. *Advances in Neural Information Processing Systems*, 36:25365–25389, 2023.

[8] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[9] C. Chen, J. Mo, J. Hou, H. Wu, L. Liao, W. Sun, Q. Yan, and W. Lin. Topiq: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing*, 2024.

[10] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[11] W.-T. Chen, G. Krishnan, Q. Gao, S.-Y. Kuo, S. Ma, and J. Wang. Dsl-fiqa: Assessing facial image quality via dual-set degradation learning and landmark-guided transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2931–2941, 2024.

[12] J. Cho, A. Zala, and M. Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3043–3054, 2023.

[13] Y. Choi, J. Park, H. Kim, J. Lee, and S. Park. Fair sampling in diffusion models through switching mechanism. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21995–22003, 2024.

[14] M. J. Chong and D. Forsyth. Effectively unbiased fid and inception score and where to find them. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6070–6079, 2020.

[15] R. D. Cook and S. Weisberg. Characterizations of an empirical influence function for detecting influential cases in regression. *Technometrics*, 22(4):495–508, 1980.

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[17] A. Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[18] D. Epstein, A. Jabri, B. Poole, A. Efros, and A. Holynski. Diffusion self-guidance for controllable image generation. *Advances in Neural Information Processing Systems*, 36:16222–16239, 2023.

[19] W. Feng, X. He, T.-J. Fu, V. Jampani, A. Akula, P. Narayana, S. Basu, X. E. Wang, and W. Y. Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*, 2022.

[20] F. Garcea, A. Serra, F. Lamberti, and L. Morra. Data augmentation for medical imaging: A systematic literature review. *Computers in Biology and Medicine*, 152:106391, 2023.

[21] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[22] Y. Guo, D. Stutz, and B. Schiele. Improving robustness of vision transformers by reducing sensitivity to patch corruptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4108–4118, 2023.

[23] L. Gustafson, C. Rolland, N. Ravi, Q. Duval, A. Adcock, C.-Y. Fu, M. Hall, and C. Ross. Facet: Fairness in computer vision evaluation benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20370–20382, 2023.

[24] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[25] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[26] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[27] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[28] A. Jain, N. Memon, and J. Togelius. Fair gans through model rebalancing with synthetic data. *arXiv preprint arXiv:2308.08638*, 2023.

[29] S. Jayasumana, S. Ramalingam, A. Veit, D. Glasner, A. Chakrabarti, and S. Kumar. Rethinking fid: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9307–9315, 2024.

[30] H. Jung, T. Jang, and X. Wang. A unified debiasing approach for vision-language models across modalities and tasks. *arXiv preprint arXiv:2410.07593*, 2024.

[31] J. Kim, Z. Wang, and Q. Qiu. Model-agnostic human preference inversion in diffusion models. *arXiv preprint arXiv:2404.00879*, 2024.

[32] K. Kim, Y. Na, S.-J. Ye, J. Lee, S. S. Ahn, J. E. Park, and H. Kim. Controllable text-to-image synthesis for multi-modality mr images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7936–7945, 2024.

[33] J. Y. Koh, D. Fried, and R. R. Salakhutdinov. Generating images with multimodal language models. *Advances in Neural Information Processing Systems*, 36, 2024.

[34] Y. A. Kolchinski, S. Zhou, S. Zhao, M. Gordon, and S. Ermon. Approximating human judgment of generated image quality. *arXiv preprint arXiv:1912.12121*, 2019.

[35] A. J. Larrazabal, N. Nieto, V. Peterson, D. H. Milone, and E. Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23):12592–12594, 2020.

[36] J. Li, D. Li, C. Xiong, and S. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

[37] J. Li, L. Hu, J. Zhang, T. Zheng, H. Zhang, and D. Wang. Fair text-to-image diffusion via fair mapping. *arXiv preprint arXiv:2311.17695*, 2023.

[38] N. Liu, S. Li, Y. Du, A. Torralba, and J. B. Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, pages 423–439. Springer, 2022.

[39] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[40] N. Lui, B. Chia, W. Berrios, C. Ross, and D. Kiela. Leveraging diffusion perturbations for measuring fairness in computer vision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 14220–14228, 2024.

[41] R. Naik and B. Nushi. Social biases through the text-to-image generation lens. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 786–808, 2023.

[42] R. Parihar, A. Bhat, A. Basu, S. Mallick, J. N. Kundu, and R. V. Babu. Balancing act: Distribution-guided debiasing in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6668–6678, 2024.

[43] O. Pelka, S. Koitka, J. Rückert, F. Nensa, and C. M. Friedrich. Radiology objects in context (roco): a multimodal image dataset. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3*, pages 180–189. Springer, 2018.

[44] M. V. Perera and V. M. Patel. Analyzing bias in diffusion-based face generation models. In *2023 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2023.

[45] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[46] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[47] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

[48] T. Ridnik, E. Ben-Baruch, A. Noy, and L. Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021.

[49] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. 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.

[50] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494, 2022.

[51] A. Sathe, P. Jain, and S. Sitaram. A unified framework and dataset for assessing gender bias in vision-language models. *arXiv preprint arXiv:2402.13636*, 2024.

[52] A. Sauer, D. Lorenz, A. Blattmann, and R. Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pages 87–103. Springer, 2025.

[53] J. Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

[54] X. Shen, C. Du, T. Pang, M. Lin, Y. Wong, and M. Kankanhalli. Finetuning text-to-image diffusion models for fairness. In *The Twelfth International Conference on Learning Representations*.

[55] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[56] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[57] Y. Tian, Z. Ni, B. Chen, S. Wang, H. Wang, and S. Kwong. Generalized visual quality assessment of gan-generated face images. *arXiv preprint arXiv:2201.11975*, 2022.

[58] S. Tuli, I. Dasgupta, E. Grant, and T. L. Griffiths. Are convolutional neural networks or transformers more like human vision? *arXiv preprint arXiv:2105.07197*, 2021.

[59] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[60] L. N. Vaserstein. Markov processes over denumerable products of spaces, describing large systems of automata. *Problemy Peredachi Informatsii*, 5(3):64–72, 1969.

[61] A. Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[62] J. Wang, Z. Yue, S. Zhou, K. C. Chan, and C. C. Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, pages 1–21, 2024.

[63] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

[64] T. Wolf. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

[65] Z. Ying, H. Niu, P. Gupta, D. Mahajan, D. Ghadiyaram, and A. Bovik. From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3575–3585, 2020.

[66] M. Zhao, F. Bao, C. Li, and J. Zhu. Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations. *Advances in Neural Information Processing Systems*, 35:3609–3623, 2022.

A Details of Synthetic Data in Figure 3

To construct the synthetic dataset, we generated non-Gaussian data for groups A and B by combining multivariate normal and exponential distributions. Each group has distinct means, covariances, and exponential scaling factors to ensure variability and non-Gaussian characteristics in the data. For group A, we define the mean as $\boldsymbol{\mu}_A$ and covariance as Σ_A. Samples for group A were drawn from a multivariate normal distribution, $\mathcal{N}(\boldsymbol{\mu}_A, \Sigma_A)$, and combined with exponential noise with a scale parameter λ_A. Similarly, for group B, we define the mean as $\boldsymbol{\mu}_B$ and covariance as Σ_B. Samples are drawn from $\mathcal{N}(\boldsymbol{\mu}_B, \Sigma_B)$ and combined with exponential noise with a scale parameter λ_B.

$$A_{\text{ref}} = \mathcal{N}(\boldsymbol{\mu}_A, \Sigma_A) + \text{Exp}(\lambda_A)$$
$$B_{\text{ref}} = \mathcal{N}(\boldsymbol{\mu}_B, \Sigma_B) + \text{Exp}(\lambda_B)$$

To introduce distribution shift as examples for fair and unfair case, translations are applied to each group. Let $\mathbf{t}_A$ and $\mathbf{t}_B$ represent the translations for groups A and B respectively. The test data for each group is generated as:

$$A_{\text{gen}} = \mathcal{N}(\boldsymbol{\mu}_A, \Sigma_A) + \mathbf{t}_A + \text{Exp}(\lambda_A)$$
$$B_{\text{gen}} = \mathcal{N}(\boldsymbol{\mu}_B, \Sigma_B) + \mathbf{t}_B + \text{Exp}(\lambda_B)$$

where $\boldsymbol{\mu}_A = [\mu_{A1}, \mu_{A2}]$ and $\Sigma_A = \begin{bmatrix} \sigma_{A1}^2 & 0 \\ 0 & \sigma_{A2}^2 \end{bmatrix}$ denote the mean and covariance of group A, $\boldsymbol{\mu}_B = [\mu_{B1}, \mu_{B2}]$ and $\Sigma_B = \begin{bmatrix} \sigma_{B1}^2 & 0 \\ 0 & \sigma_{B2}^2 \end{bmatrix}$ denote the mean and covariance of group B, λ_A and λ_B represent the exponential scaling factors for groups A and B, and $\mathbf{t}_A$ and $\mathbf{t}_B$ are translations applied to groups A and B, respectively.

Using this structure, we introduce non-Gaussianity through the combination of multivariate normal and exponential distributions with group-specific parameters $\boldsymbol{\mu}_A, \Sigma_A, \lambda_A$, and $\boldsymbol{\mu}_B, \Sigma_B, \lambda_B$. Test (generated) datasets maintain only the mean parameters for each group, but covariance and scaling factors are shifted as well as translations to mimic the distribution shift in generative models.

For the reference set, we choose $\mu_{A1} = \mu_{A2} = 0$, $\sigma_{A1}^2 = \sigma_{A2}^2 = 1$, $\lambda_A = 1$, $\mu_{B1} = \mu_{B2} = 15$, $\sigma_{B1}^2 = \sigma_{B2}^2 = 8$, and $\lambda_B = 2$. For the generated set, we change the covariance as $\sigma_{A1}^2 = \sigma_{A2}^2 = 3$ and $\sigma_{B1}^2 = \sigma_{B2}^2 = 12$, and shift the scaling $\lambda_A \leftarrow \lambda_A + 0.2$, and $\lambda_B \leftarrow \lambda_B + 0.2$. Moreover, we apply different scaling and translations for fair and unfair synthetic dataset. Specifically, we choose $\mathbf{t}_A = [3, 3]$ and $\mathbf{t}_B = [-3, -3]$, to depict a fair scenario, while $\mathbf{t}_A = [1, 1]$ and $\mathbf{t}_B = [-11, -11]$ are chosen to simulate unfairly skewed distribution for group B.

B Impact of Quality Bias in Generative Models in Downstream Task and Validity of DQA

B.1 Negative Impact of Quality Bias in Generative Models

Unfairness in generated image quality across demographic groups poses a critical issue in generative modeling. Generative models, especially those trained on uncurated datasets, often produce images of systematically lower quality for specific demographic groups, such as those defined by gender, race, or age. This quality discrepancy not only undermines visual representation fairness but also risks reinforcing biases when these generated images are used for data augmentation in training pipelines, potentially transferring such biases into downstream models. Addressing this issue requires robust strategies to ensure consistent image quality across all demographic attributes.

To highlight the practical implications of quality bias, we conduct a classification task with a ResNet-50 model [24] using chest X-ray images from the Chest X-ray dataset [63], a dataset known to exhibit fairness issues, as evidenced by differing AUC scores across demographic groups [35]. To enhance classifier's performance, a user might employ text-to-medical-image generation models [50] trained on the ROCO dataset [43] as a data augmentation strategy. In our initial experiments, we generate 1,000 images per gender and class for augmentation. The details of Chest X-ray dataset and the generation details are introduced in Appendix F.

However, despite using an equal quantity of generated images for each demographic group, fairness issues in the classification model not only persist but, as shown in Table 2, even worsen. This is

evidenced by higher values of Avg(ΔAUC) and $\max(\Delta\text{AUC})$, calculated as

$$\text{Avg}(\Delta\text{AUC}) = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} |\text{AUC}_c^{\text{male}} - \text{AUC}_c^{\text{female}}|, \quad \max(\Delta\text{AUC}) = \max_{c\in\mathcal{C}} |\text{AUC}_c^{\text{male}} - \text{AUC}_c^{\text{female}}|,$$

where $\mathcal{C}$ denotes the set of classes. These results imply that generated images may exacerbate fairness issues, likely due to quality discrepancies across demographic groups.

Table 2: Comparison of classification performance and fairness metrics using different data augmentation strategies on the Chest X-ray dataset. **Blue** indicates an improvement in fairness, while **Red** denotes a deterioration compared to the baseline. All augmented data are generated by a text-to-medical-image model, with Fair and Unfair subsets selected from the entire generated dataset using Algorithm 1. Full augmentation worsens fairness, suggesting quality bias issues in the generated images. Data augmentation with the Fair Subset uses generated data of equal quality across genders, identified by lower DQA scores, yields lower Avg(ΔAUC) and $\max(\Delta\text{AUC})$ values without applying any fairness-specific technique. This outcome suggests that DQA effectively identifies reliable evaluation metrics for assessing fairness in generated image quality.

	Overall AUC	AUC^{male}	$\text{AUC}^{\text{female}}$	Avg(ΔAUC) $\downarrow$	$\max(\Delta\text{AUC})\downarrow$	$\overline{\text{DQA}}$
Baseline	83.10±0.13	72.78±0.33	71.96±0.35	2.40±0.36	7.08±1.82	-
Full Augmentation	85.35±0.12	78.12±0.32	77.71±0.33	**2.45±0.35**	**8.13±2.04**	-
Fair Subset (DQA $\downarrow$)	85.27±0.12	77.35±0.35	77.24±0.35	**2.16±0.36**	**6.98±2.54**	0.0868
Unfair Subset (DQA $\uparrow$)	85.54±0.12	77.95±0.32	77.81±0.33	**2.62±0.39**	**8.93±2.46**	0.5495

B.2 Validity of DQA

To validate the effectiveness of DQA in identifying reliable image encoders for quality assessment, we construct both fair and unfair generated datasets in terms of quality as identified by their DQA scores. The fair generated dataset is expected to enhance fairness in classification when used for data augmentation, while the unfair generated dataset is anticipated to exacerbate fairness issues.

These datasets are characterized by lower (fair) and higher (unfair) DQA scores, evaluated using a reliable image encoder f^*. Specifically, let A_{gen} and B_{gen} represent two groups of generated data, with subsets $S_A \subset A_{\text{gen}}$ and $S_B \subset B_{\text{gen}}$, each of size $k = 0.2 \times |A_{\text{gen}}|$. We define the fair and unfair subsets as $(S_A^{\text{fair}}, S_B^{\text{fair}}) = \arg\min_m \text{DQA}(S_A^{(m)}, S_B^{(m)}; f^*)$ and $(S_A^{\text{unfair}}, S_B^{\text{unfair}}) = \arg\max_m \text{DQA}(S_A^{(m)}, S_B^{(m)}; f^*)$, selected from M candidate subsets $\{(S_A^{(m)}, S_B^{(m)})\}_{m=1}^M$.

To construct meaningful candidate pairs, we employ influence scores as a probabilistic measure of each image's impact on the DQA score, calculated via influence functions [15]. These scores are normalized and used in a multinomial sampling scheme, allowing us to prioritize high-impact images in both fair and unfair selection processes. Algorithm 1 in Appendix B.3 details the steps for sampling fair and unfair subsets, using influence-based probabilities to guide the selection.

For the classification task, we train a ResNet-50 model on the Chest X-ray diagnosis dataset, as outlined in Sec. B.1. Initial experiments in Sec. B.1 used an augmentation set containing 1000 images per gender and class. For DQA-guided augmentation, we add either the fair subset $(S_A^{\text{fair}}, S_B^{\text{fair}})$ or the unfair subset $(S_A^{\text{unfair}}, S_B^{\text{unfair}})$, each consisting of 200 images per gender and class, to assess how these augmentations impact model performance and demographic fairness. This setup enables a comparative evaluation of overall accuracy and fairness across demographic groups, thereby justifying the validity of DQA as an indicator of reliability.

The experimental results, shown in Table 2, demonstrate the effectiveness of the DQA score: the fair subset identified by low DQA improves fairness in classification AUC scores across demographic groups, even though DQA is not specifically designed for classification fairness, whereas the unfair subset (high DQA) worsens fairness outcomes.

B.3 Fair/Unfair Subset Sampling Algorithm with DQA

Algorithm 1 Finding Fair and Unfair Subsets Using Influence Scores for DQA

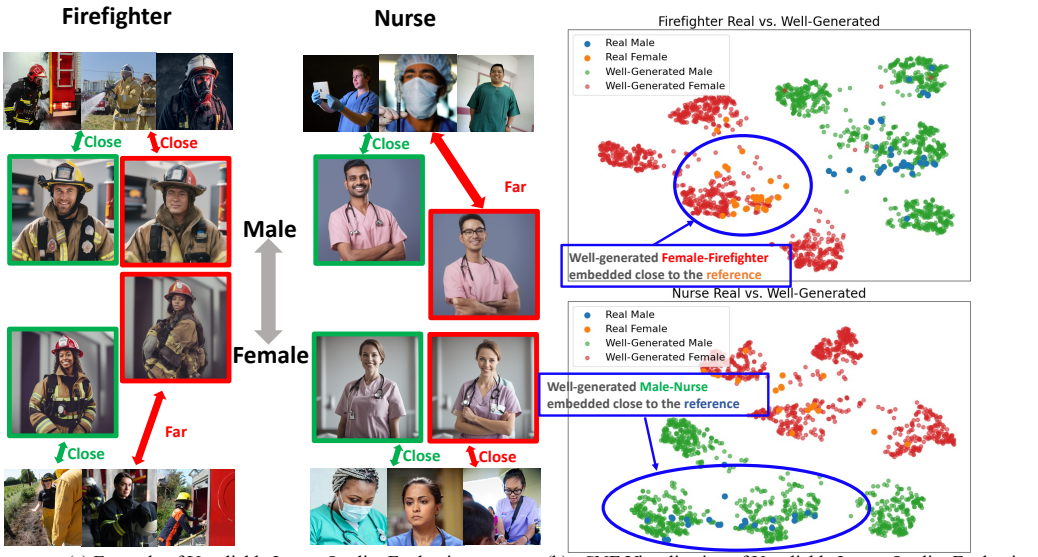
Input: Generated datasets A_{gen} and B_{gen}; reference datasets A_{ref} and B_{ref}; reliable encoder f^*; subset size k; number of samples M; small constant ϵ
Output: Fair/Unfair subsets $(S_A^{\text{fair}}, S_B^{\text{fair}})$, $(S_A^{\text{unfair}}, S_B^{\text{unfair}})$
$F_A, F_B, F_{A_{\text{ref}}}, F_{B_{\text{ref}}} \leftarrow \{f^*(x_i) \mid x_i \in A_{\text{gen}}, B_{\text{gen}}, A_{\text{ref}}, B_{\text{ref}}\}$
$\text{DQA}_{\text{original}} \leftarrow \text{DQA}(F_A, F_B, F_{A_{\text{ref}}}, F_{B_{\text{ref}}})$
for each $x_i \in A_{\text{gen}}$ and $x_j \in B_{\text{gen}}$ **do**
  $F_A^{-i}, F_B^{-j} \leftarrow F_A \setminus \{f^*(x_i)\}, F_B \setminus \{f^*(x_j)\}$
  $\delta_i^A \leftarrow \text{DQA}_{\text{original}} - \text{DQA}(F_A^{-i}, F_B, F_{A_{\text{ref}}}, F_{B_{\text{ref}}})$
  $\delta_j^B \leftarrow \text{DQA}_{\text{original}} - \text{DQA}(F_A, F_B^{-j}, F_{A_{\text{ref}}}, F_{B_{\text{ref}}})$
end for
Adjust influence scores for sampling:
For fair subsets, invert influence scores:
$p_i^{A,\text{fair}}, p_j^{B,\text{fair}} \leftarrow \frac{-\delta_i^A - \min\{-\delta_i^A\} + \epsilon}{\sum_i (-\delta_i^A - \min\{-\delta_i^A\}) + \epsilon}, \frac{-\delta_j^B - \min\{-\delta_j^B\} + \epsilon}{\sum_j (-\delta_j^B - \min\{-\delta_j^B\}) + \epsilon}$
For unfair subsets, use original influence scores:
$p_i^{A,\text{unfair}}, p_j^{B,\text{unfair}} \leftarrow \frac{\delta_i^A - \min\{\delta_i^A\} + \epsilon}{\sum_i (\delta_i^A - \min\{\delta_i^A\}) + \epsilon}, \frac{\delta_j^B - \min\{\delta_j^B\} + \epsilon}{\sum_j (\delta_j^B - \min\{\delta_j^B\}) + \epsilon}$
Initialize: best_DQA $\leftarrow \infty$, worst_DQA $\leftarrow -\infty$
for $m = 1$ to M **do**
  Sample fair/unfair candidate subsets:
  $S_A^{(m,\text{fair})}, S_B^{(m,\text{fair})} \leftarrow \text{Sample}(A_{\text{gen}}, k, p_i^{A,\text{fair}}), \text{Sample}(B_{\text{gen}}, k, p_j^{B,\text{fair}})$
  $\text{DQA}^{(m,\text{fair})} \leftarrow \text{DQA}(S_A^{(m,\text{fair})}, S_B^{(m,\text{fair})}, F_{A_{\text{ref}}}, F_{B_{\text{ref}}})$
  Compute DQA for fair/unfair candidate:
  if $\text{DQA}^{(m,\text{fair})} < \text{best\_DQA}$ **then**
    $\text{best\_DQA} \leftarrow \text{DQA}^{(m,\text{fair})}$
    $(S_A^{\text{fair}}, S_B^{\text{fair}}) \leftarrow (S_A^{(m,\text{fair})}, S_B^{(m,\text{fair})})$
  end if
  $S_A^{(m,\text{unfair})}, S_B^{(m,\text{unfair})} \leftarrow \text{Sample}(A_{\text{gen}}, k, p_i^{A,\text{unfair}}), \text{Sample}(B_{\text{gen}}, k, p_j^{B,\text{unfair}})$
  $\text{DQA}^{(m,\text{unfair})} \leftarrow \text{DQA}(S_A^{(m,\text{unfair})}, S_B^{(m,\text{unfair})}, F_{A_{\text{ref}}}, F_{B_{\text{ref}}})$
  if $\text{DQA}^{(m,\text{unfair})} > \text{worst\_DQA}$ **then**
    $\text{worst\_DQA} \leftarrow \text{DQA}^{(m,\text{unfair})}$
    $(S_A^{\text{unfair}}, S_B^{\text{unfair}}) \leftarrow (S_A^{(m,\text{unfair})}, S_B^{(m,\text{unfair})})$
  end if
end for
Return: $(S_A^{\text{fair}}, S_B^{\text{fair}}), (S_A^{\text{unfair}}, S_B^{\text{unfair}})$

C Supplementary for Fig. 4: Illustration for Well-Generated Sample

Figu. 4 shows how poor-quality images are frequently misembedded into the wrong gender cluster due to encoder bias and sensitivity to image degradation. To complement this, Fig. 7 presents the case of well-generated images. While the encoder fails to reliably embed poor-quality images in Fig. 4, well-generated samples in Fig. 7 demonstrate clearer separation between gender groups and are mostly placed correctly within their demographic clusters. This comparison underscores the unreliability of the encoder, which performs inconsistently depending on the quality of the input images.

D Constructing Evaluation Dataset for DQA

We consider realistic scenarios encountered in text-to-image generation for human image datasets using Stable Diffusion Inpainting [49]. Our baseline follows the recommended settings from [40], where image quality degradation is achieved by adjusting specific hyperparameters. Each modification



(a) Example of Unreliable Image Quality Evaluation (b) t-SNE Visualization of Unreliable Image Quality Evaluation

Figure 7: (a) Images in **green** boxes represent "good" quality generated images, while **red** boxes indicate "poor" quality images. Poor-quality images are prone to misembedding by the encoder, as shown in Fig. 4. (b) t-SNE visualization of well-generated images using a CLIP [46] image encoder shows clear separation between gender clusters and correct placement of most samples, highlighting the encoder's unreliable behavior under poor-quality conditions.

is grounded in prior literature, ensuring that the degradations reflect practical and interpretable variations in generation quality. Specifically, the baseline parameters include a sampling step size of $T = 40$, noise strength $s_n = 0.7$, guidance scale $s_g = 7.5$, and a refinement phase during the last 20

1. **Baseline**: Uses sufficient diffusion steps with a balanced influence between the initial image and noise. This represents high-quality generation with the standard configuration $(T, s_n, s_g, \tau_{\text{refine}}) = (40, 0.7, 7.5, 0.2)$.
2. **Weak Guidance**: In classifier-free guidance (CFG), a higher guidance scale enforces stronger adherence to the text prompt, while lower values weaken this connection. We reduce s_g to simulate a scenario where the model struggles to align the image with the intended prompt, leading to reduced coherence or incomplete rendering of attributes $(40, 0.7, \mathbf{1.0}, 0.2)$.
3. **Fewer Steps**: As established in [31], reducing the number of diffusion steps often results in poorer visual quality due to incomplete denoising. We halve T to 20 to intentionally increase residual noise and visible artifacts, thereby decreasing the model's capacity to refine image details $(\mathbf{20}, 0.7, 7.5, 0.2)$.
4. **Strong Noise**: For inpainting, increased noise strength s_n preserves more of the original image, which can hinder the model's ability to apply the target attribute modifications. By increasing s_n to 0.9, we introduce more randomness, degrading coherence and making the attribute editing task more difficult $(40, \mathbf{0.9}, 7.5, 0.2)$.
5. **No Refiner**: According to the SDXL paper, a dedicated refiner network improves visual fidelity and detail. Removing the refiner by setting $\tau_{\text{refine}} = 0.0$ allows us to directly test the quality drop, particularly in terms of fine-grained details and overall realism $(40, 0.7, 7.5, \mathbf{0.0})$.
6. **Combination**: We combine the weak guidance, fewer steps, and strong noise conditions to create an extremely degraded setting. This tests the model's robustness under simultaneous quality impairments (**20, 0.9, 1.0, 0.0**).

We select 10 professions commonly referenced in the literature [40, 23, 12], including flight attendant, nurse, secretary, teacher, veterinarian, engineer, pilot, firefighter, surgeon, and builder. Additionally,

we include four racial groups identified in [40]: Asian, Black, Indian, and White Caucasian. Example datasets illustrating the applied degradations are shown in Figure 8.

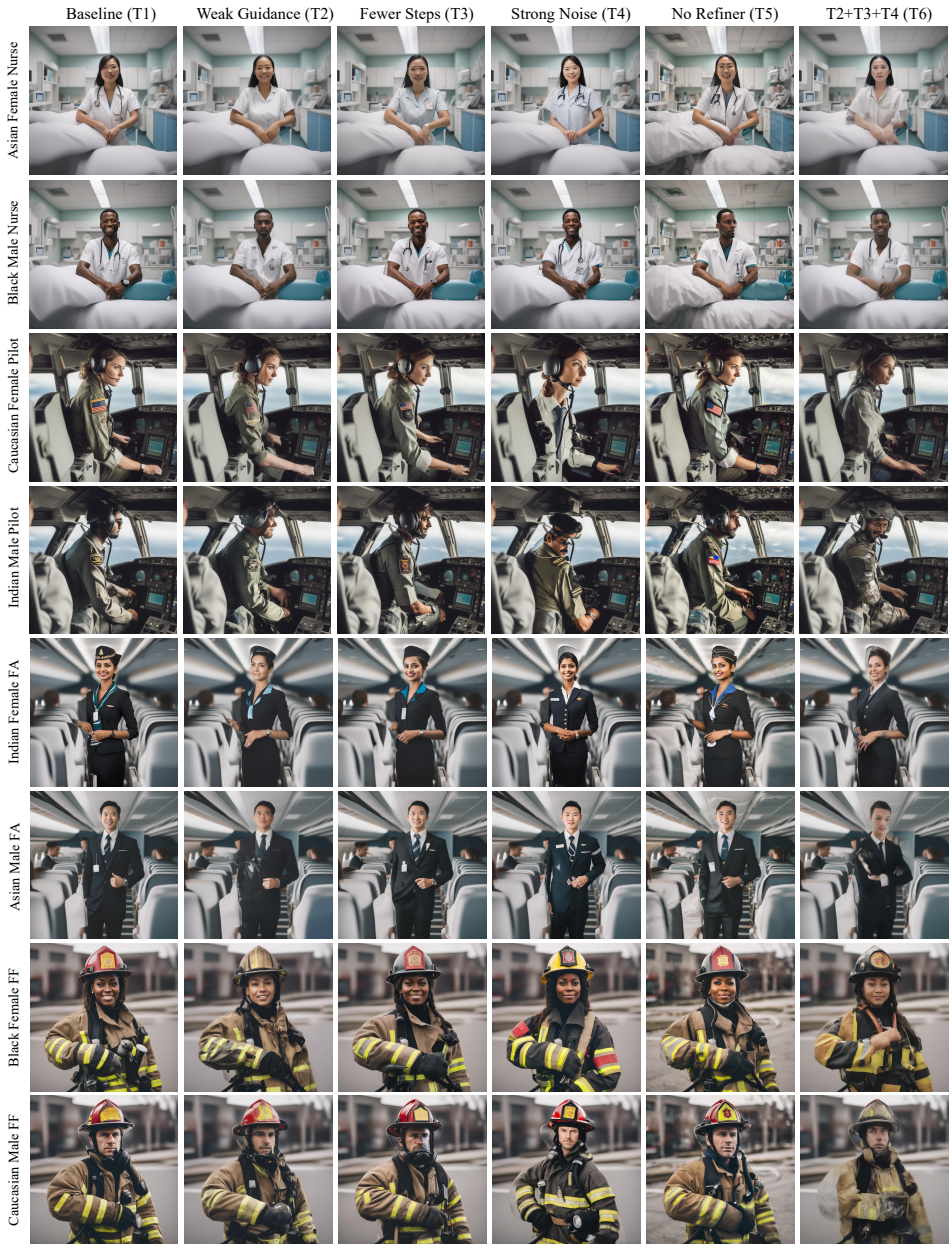


Figure 8: Examples of constructed evaluation datasets for DQA under various text-to-image generation scenarios to controlled degradation of generated image. The scenarios include Baseline, Weak Guidance (T2), Fewer Steps (T3), Strong Noise (T4), No Refiner, and a combination of T2, T3, and T4. Each setting adjusts specific hyperparameters of Stable Diffusion Inpainting [49] to simulate realistic degradations in image quality. The datasets represent 10 professions and 4 racial groups, illustrating the diversity and quality variations used for evaluation while four professions (Nurse, Pilot, Flight Attendant (FA), and fire fighter (FF)) are presented in the example.

E Ablation Study: Impact of Hyperparameter

Fig. 9 demonstrates the clear impact of DQA-Guidance on image generation. Compared to the baseline ($\lambda_1 = 0$), increasing λ_1 effectively reduces quality disparities in generated images while substantially improving overall image quality, especially $\lambda_1 = 20$ and $\lambda_1 = 30$. However, setting λ_1 too high introduces excessive noise, leading to a decline in image quality. These findings suggest that DQA not only provides a reliable measure for evaluating fairness but also serves as an effective regularizer, enhancing fairness in image generation when applied as guidance in diffusion models. Additionally, larger values of λ_2 intuitively contribute to improved generation quality, as demonstrated in Fig. 9 (b).

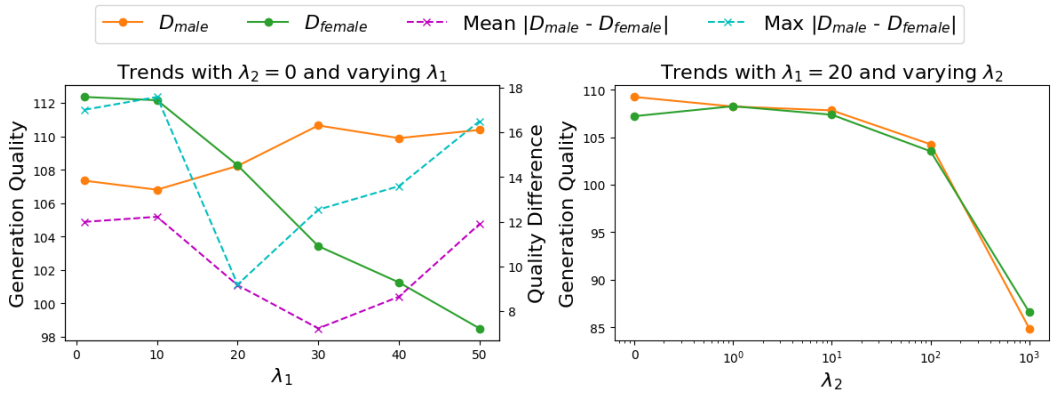


Figure 9: Experimental results for generation quality and quality disparities with DQA-Guidance with Stable DIffusion. The left plot shows the impact of λ_1 on generation quality for each demographic group (lower values indicate better quality) and displays the average and maximum quality gap across all disease classes (lower values indicate reduced disparity). The right plot illustrates the effect of λ_2 on overall generation quality. Here, $\lambda_1 = 0$ denotes no DQA-Guidance, while higher λ_1 values reflect a stronger influence of DQA-Guidance. DQA-Guidance effectively enhances generation quality and reduces quality disparities across demographic groups.

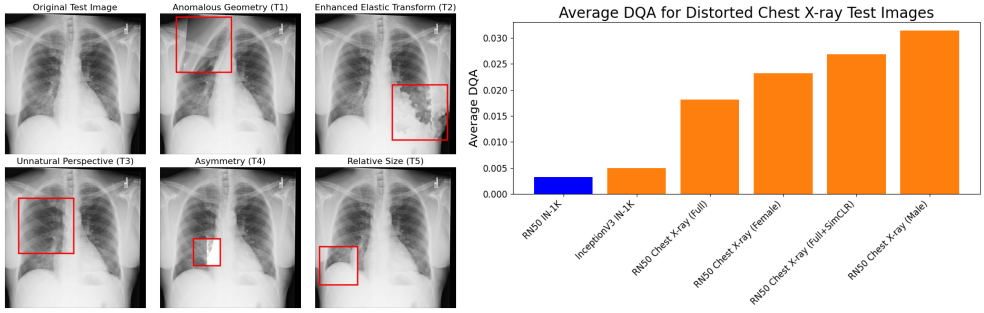
F Details in Chest X-ray Dataset and Generation

F.1 Details of the Chest X-ray Dataset

We use the NIH ChestX-ray14 dataset [63], a large repository containing 112,120 chest X-ray images from 30,805 patients, annotated with 14 common thoracic disease categories, including Hernia, Pneumonia, Fibrosis, Emphysema, Edema, Cardiomegaly, Pleural Thickening, Consolidation, Mass, Pneumothorax, Nodule, Atelectasis, Effusion, and Infiltration. By including 'No Findings' as a benign case, the dataset expands to 15 classes. It also includes demographic information, with approximately 56.5% male and 43.5% female patients.

F.2 Details of Synthetic Chest X-ray Generation

To generate synthetic Chest X-ray images, we use a pre-trained ImageGen model [50] trained on the ROCO dataset [43], which contains paired image and text data for medical purposes. The pretrained model is available on HuggingFace [64] under the model ID Nihirc/Prompt2MedImage. We generate 1,000 images per gender and class, resulting in a total of 30,000 images across 2 genders and 15 classes. The input prompt format for generation is "Chest X-ray image of a {GENDER} patient showing a/an {DISEASE}."



(a) Example of Transforms Mimicking Image Generation Failure

(b) DQA of Various Models for Distorted Images

Figure 10: (a) To assess the DQA across varying qualities of generated medical images, we simulate generative model failures by applying transformations to test images that reflect common failure patterns in generative models. (b) By incrementally applying these transformations and evaluating the reliability of various pretrained encoders, we find that a ResNet-50 model pretrained on ImageNet-1K demonstrates greater reliability in quality assessment, consistently handling poor-quality images across demographic groups by showing lowest DQA in average. In contrast, the same model trained on reference data shows higher DQA scores, indicating unreliable image quality assessment.

G DQA analysis for Medical Image

G.1 Constructing Reference Dataset for Medical Image

In the medical image, we utilize the Chest X-ray diagnosis dataset in Sec. B.1 as the reference, given its consistent image quality across genders, controlled through human annotations. This consistency makes it an effective benchmark for quality assessment. Specifically, we designate the training set of Chest X-ray images as the reference dataset, while the test set and its transformations are used as a mimic of the generated dataset to help identify a reliable image encoder. In more detail, the real test data remains in-distribution relative to the training dataset, while we simulate generative model failures [6] by applying transformations to the test set, creating poor-quality images as shown in Fig. 10 (a).

G.2 Reliability Analysis for Image Encoders for Medical Image

For medical images, we assess encoders such as InceptionV3 and RN50 pretrained on IN-1K, alongside RN50 models trained directly on the Chest X-ray dataset using supervised learning, self-supervised learning (SimCLR) [10], and supervised learning on a single-gender subset. The RN50 pretrained on IN-1K achieves the lowest DQA score, suggesting that pretraining on a diverse dataset helps mitigate biases inherent in domain-specific data. In contrast, models trained directly on medical images exhibit higher DQA scores, potentially due to the amplification of existing biases within the specialized dataset.

H DQA-Guidance for Medical Image

H.1 Experimental Details

To verify the effectiveness of DQA-Guidance in mitigating quality bias, we utilize a medical dataset and a generative model for medical images, consistent with the setup in previous sections. Specifically, we apply Eq. (4) to the text-to-medical-image model during the sampling stage, generating 100 images per gender and class, resulting in a total of 3000 images (2 genders and 15 classes). For each gender, the prompt "Chest X-ray image of a {GENDER} patient showing a {DISEASE_NAME}." is used, with the Chest X-ray training data for each gender serving as a reference to compute empirical DQA during the sampling stage. In the experiments, we vary λ_1 while fixing $\lambda_2 = 0$ to examine the impact of DQA-Guidance on both generation quality and the quality gap between groups.

H.2 Result Analysis for DQA-Guidance

Fig. 11 demonstrates the clear impact of DQA-Guidance on medical image generation. Compared to the baseline ($\lambda_1 = 0$), increasing λ_1 effectively reduces quality disparities in generated images while substantially improving overall image quality. However, setting λ_1 too high introduces excessive noise, leading to a decline in image quality. These findings suggest that DQA not only provides a reliable measure for evaluating fairness but also serves as an effective regularizer, enhancing fairness in image generation when applied as guidance in diffusion models. Additionally, larger values of λ_2 intuitively contribute to improved generation quality. Qualitative results of DQA-Guidance is shown in Fig. 12. Similar to DQA-Guidance for human images, the improvements primarily focus on refining texture. While these improvements may appear subtle from a user's perspective, the measured quality confirms that the hyperparameters λ_1 and λ_2 play a significant role in enhancing overall quality and reducing quality disparities.

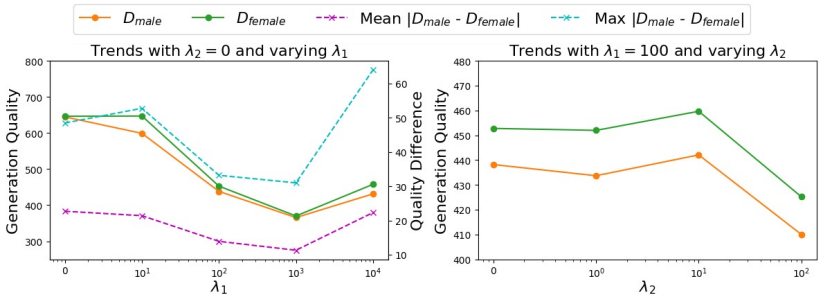


Figure 11: Experimental results for generation quality and quality disparities with DQA-Guidance. The left plot shows the impact of λ_1 on generation quality for each demographic group in Chest X-ray image generation (lower values indicate better quality) and displays the average and maximum quality gap across all disease classes (lower values indicate reduced disparity). The right plot illustrates the effect of λ_2 on overall generation quality. Here, $\lambda_1 = 0$ denotes no DQA guidance, while higher λ_1 values reflect a stronger influence of DQA-Guidance. DQA-Guidance effectively enhances generation quality and reduces quality disparities across demographic groups.

I DQA on Different Types of Image Quality Assessment

In addition to our approach, other methods for assessing image quality include visual question answering (VQA) [40] and neural networks specifically trained for quality evaluation [34, 57, 9].

In [40], VQA models are asked questions such as Prompt 1: "Is this image real or fake?" or Prompt 2: "Are this person's limbs distorted?" to detect unreal aspects of a given image. However, as the image encoder used in VQA models may exhibit bias, the distribution of VQA answers could also be biased. To quantify this bias, we adapt DQA in Eq. (1) by replacing $D(f(\cdot), f(\cdot))$ with $p(h(\cdot), \mathcal{T})$, where h denotes the VQA model and p represents the probability of detecting abnormalities based on the text prompt $\mathcal{T}$. This approach utilizes the probability of realism detected by the VQA model as the image quality assessment metric.

$$\text{DQA}^{\text{VQA}} = \frac{|p(h(A_{\text{gen}})) - p(h(B_{\text{gen}}))|}{p(h(\mathcal{I}_{\text{gen}}))}$$

We also adapt DQA to image quality assessment (IQA) models that output indicators of general image quality. For example, TOPIQ [9] is a supervised network designed for image quality evaluation. It is trained on datasets such as FLIVE [65] for general images or CGFIQA [11] for facial images, using a regression task to predict quality scores. Let $s(\cdot)$ an IQA model's outcome, then we adapt DQA in Eq. (1) by replacing $D(f(\cdot), f(\cdot))$ with $\bar{s}(\cdot)$, the mean of quality score over each group.

$$\text{DQA}^{\text{IQA}} = \frac{|\bar{s}(A_{\text{gen}}) - \bar{s}(B_{\text{gen}})|}{\bar{s}(\mathcal{I}_{\text{gen}})}$$

To summarize the quality assessment methods utilized throughout the paper:

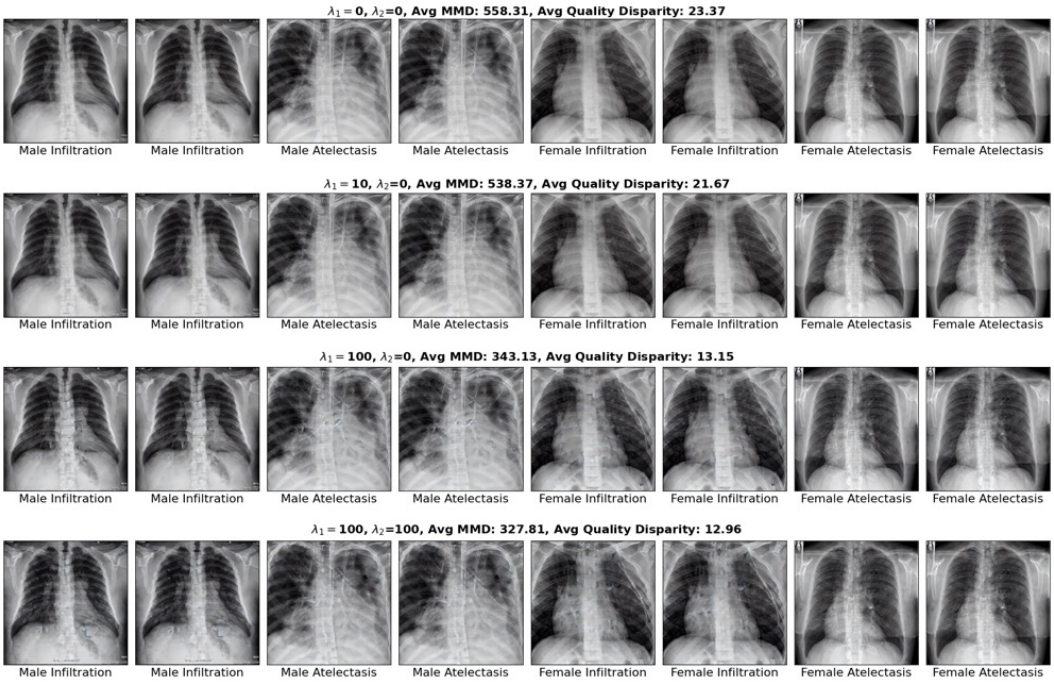
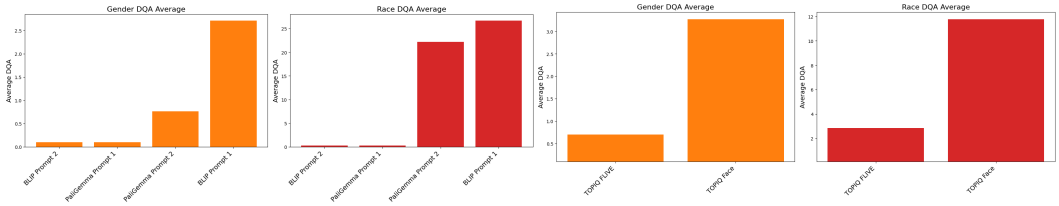


Figure 12: Qualitative results of DQA-Guidance for medical image generation. The examples highlight improvements primarily in texture refinement, demonstrating the method's ability to enhance overall image quality while addressing disparities across different conditions.



(a) DQA for visual question answering (VQA) models (b) DQA for image quality assessment (IQA) models

Figure 13: DQA on different types of image quality assessments. We compare DQA scores for gender and racial fairness across VQA models (BLIP and PaliGemma) under two prompts, as well as IQA models trained on general and facial datasets. Results highlight varying tendencies in DQA across models and prompts, with racial fairness remaining a significant challenge and facial dataset-trained IQA models showing higher DQA scores.

- **Distance-based methods**: Measure the similarity between the feature distributions of generated images and real images to determine image quality (e.g., FID).
- **VQA-based methods**: Assess visual realism and detect whether images are free from noticeable distortions or errors.
- **General IQA methods**: Evaluate objective image quality metrics such as blur, noise, sharpness, and color saturation.

We use BLIP [36] and PaliGemma [4] as representative VQA models with two different prompts. Additionally, we utilize two pre-trained versions of TOPIQ for general IQA: one trained on the FLIVE dataset for general images and another trained on the CGFIQA dataset for facial images.

The experimental results for these different types of image quality assessments are visualized in Fig. 13. Interestingly, VQA models exhibit varying tendencies. For gender-based DQA, PaliGemma

demonstrates reliability with low DQA for Prompt 1 but shows relatively high DQA for Prompt 2. Conversely, BLIP achieves reliable results with Prompt 2 but exhibits high DQA for Prompt 1. For racial DQA, both models exhibit similar tendencies with gender-based DQA; however, the overall DQA values are significantly higher, indicating that racial bias remains a pressing concern in fair evaluation.

In the case of IQA models, the version trained on a general dataset exhibits greater reliability with low DQA, whereas the version trained on facial datasets demonstrates significantly higher DQA. This result highlights potential challenges in achieving fairness when applying models trained on specific datasets.

J Impact of DQA-Guidance on Downstream Tasks

In line with Appendix B.2, we further investigate the impact of DQA-Guidance on fairness in AUC across gender in medical image classification. We compare the classification performance using different versions of generated samples. For this analysis, we use 100 images per gender and class as augmentation, while Table 2 reports results based on 1,000 images per gender and class for full augmentation and 200 images per gender and class for fair and unfair subsets.

Table 3 shows the classification performance when generative samples created with DQA-Guidance are used for data augmentation. To isolate the impact of λ_1, we eliminate the influence of λ_2 by setting $\lambda_2 = 0$.

Compared to baseline augmentation (No Guidance), DQA-Guidance improves the overall AUC and significantly reduces both the mean and maximum AUC gaps between demographic groups. This enhancement is achieved without explicit fairness constraints, relying solely on improved quality parity between groups.

Table 3: Classification performance and fairness metrics on the Chest X-ray dataset using DQA-Guidance for data augmentation. The table compares results across augmentation strategies using 100 images per gender and class. λ_1 is varied while λ_2 is set to 0 to isolate its effect. Compared to No Guidance, DQA-Guidance improves overall AUC and significantly reduces both the mean and maximum AUC gaps between demographic groups, demonstrating its effectiveness in enhancing quality parity without applying explicit fairness constraints.

	Overall AUC	AUC^{male}	$\text{AUC}^{\text{female}}$	$\text{Avg}(\Delta\text{AUC}) \downarrow$	$\max(\Delta\text{AUC}) \downarrow$
Baseline (No Augmentation)	83.10±0.13	72.78±0.33	71.96±0.35	2.40±0.36	7.08±1.82
No Guidance	85.21±0.12	77.46±0.30	77.00±0.33	2.52±0.33	8.96±2.04
DQA-Guidance ($\lambda_1 = 10$)	85.26±0.12	76.28±0.33	76.40±0.37	2.17±0.35	8.07±2.43
DQA-Guidance ($\lambda_1 = 20$)	85.74±0.12	77.90±0.34	78.04±0.32	2.22±0.38	7.82±2.86
DQA-Guidance ($\lambda_1 = 100$)	85.55±0.12	77.65±0.35	77.22±0.35	2.31±0.36	7.81±2.42
DQA-Guidance ($\lambda_1 = \lambda_2 = 100$)	85.70±0.11	78.06±0.35	77.62±0.34	2.28±0.38	8.06±2.66

K Experimental Result with Fréchet distance

The effectiveness of DQA-Guidance is demonstrated in Table 1, using the MMD metric with the DINO-RN50 encoder. In addition, we report the Fréchet Distance for generated images with and without DQA-Guidance to further evaluate generation quality and disparities across demographic groups in Table 4.

Table 4: Experimental results for generation quality and quality disparities with DQA-Guidance.

Method	Avg.MMD	Mean $\lvert D_{male} - D_{female} \rvert$	Max $\lvert D_{male} - D_{female} \rvert$
Baseline (Stable Diffusion)	29.09	1.26	1.77
+ DQA-Guidance ($\lambda_1 = 20, \lambda_2 = 100$)	28.53	**0.09**	**0.12**
+ DQA-Guidance ($\lambda_1 = 20, \lambda_2 = 1000$)	**26.27**	0.29	0.44

L Computational Resource

Table 5: Compute Resources Used for Experiments

Component	**Details**
CPU	AMD EPYC 7313 16-Core Processor
GPU	NVIDIA RTX A5000

M Licenses for existing assets

Table 6: Licenses for each asset

Dataset	**License**
ROCO Dataset	CC BY-NC-SA 4.0
Stable Diffusion	creativeml-openrail-m
SDXL	openrail++
Prompt2MedImage	wtfpl

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Abstract and introduction accurately reflect the paper's contributions, scope, and all necessary claims.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: The limitation of work is dicussed in Section 5.4.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: No theoretical result is included in the paper.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The details of experimental setting is presented, while code and data are available via supplementary materials.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The data is publicly available, or reproducible by image generation with open access code. The code for DQA-Guidance is available in the supplementary material, and will be published on GitHub after the acceptance of the paper.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The details of experimental setting are provided.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: For classification task, Table 2, the confidence interval is presented.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Computational resource is mentioned in Appendix L.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed our code according to the NeurIPS Code of Ethics, and no deviation or issue is detected.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Mentioned in the Conclusion section

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Licenses are mentioned in Appendix M, while each paper are correctly cited in the main contents.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [No]

Justification: LLM is used only for refining authors' original writing.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.