Assessing the Reliability of Image Quality Metrics and Mitigating Quality Bias in Generative Models

Supplementary Material

A. Details of Synthetic Dataset

A.1. Constructing Evaluation Dataset for DQA

We simulate realistic scenarios encountered in text-to-image generation of human images using Stable Diffusion Inpainting [60]. Our baseline follows the recommended settings from [49], where image quality degradation is achieved by adjusting specific hyperparameters. Each modification 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% of sampling, denoted by $\tau_{\text{refine}} = 0.2$. The degradation scenarios are defined as follows:

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 [38], 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 [16, 27, 49], including flight attendant, nurse, secretary, teacher, veterinarian, engineer, pilot, firefighter, surgeon, and builder. Additionally, we include four racial groups identified in [49]: Asian, Black, Indian, and White Caucasian. Example datasets illustrating the applied degradations are shown in Figure 8.

A.2. Validation of Controlled Dataset Degradation

A.2.1. Analysis of Individual Hyperparameters

This section details the process for generating images with systematic variations by modifying key hyperparameters. We vary the scale of Classifier-Free Guidance (CFG), s, and the number of diffusion sampling steps, T, both of which are known to directly impact image quality. As a preliminary check, we measure the DINO-MMD. The results confirm that quality degrades (higher MMD) as s is lowered and T is reduced, as shown in Table 2 and Table 3.

Table 2. Impact of varying CFG scale (s) on DINO-MMD. Lower s values lead to higher (worse) MMD scores, indicating poorer quality.

s **(default 7.5)**	**DINO-MMD** $\downarrow$
5	**17.94**
2	23.31
1	25.82

Table 3. Impact of varying diffusion steps (T) on DINO-MMD. Fewer steps lead to higher (worse) MMD scores.

T **(default 40)**	**DINO-MMD** $\downarrow$
30	**12.18**
20	12.21
10	13.42

A.2.2. Statistical Validation of Degradation Scenarios (T1-T6)

While human evaluation is the gold standard for perceptual quality, conducting large-scale annotation over our full dataset is infeasible. To rigorously validate our degradation scenarios, we conducted an additional statistical validation using No-Reference Image Quality Assessment (NR-IQA) metrics [43].

The goal of this analysis is to verify two properties:

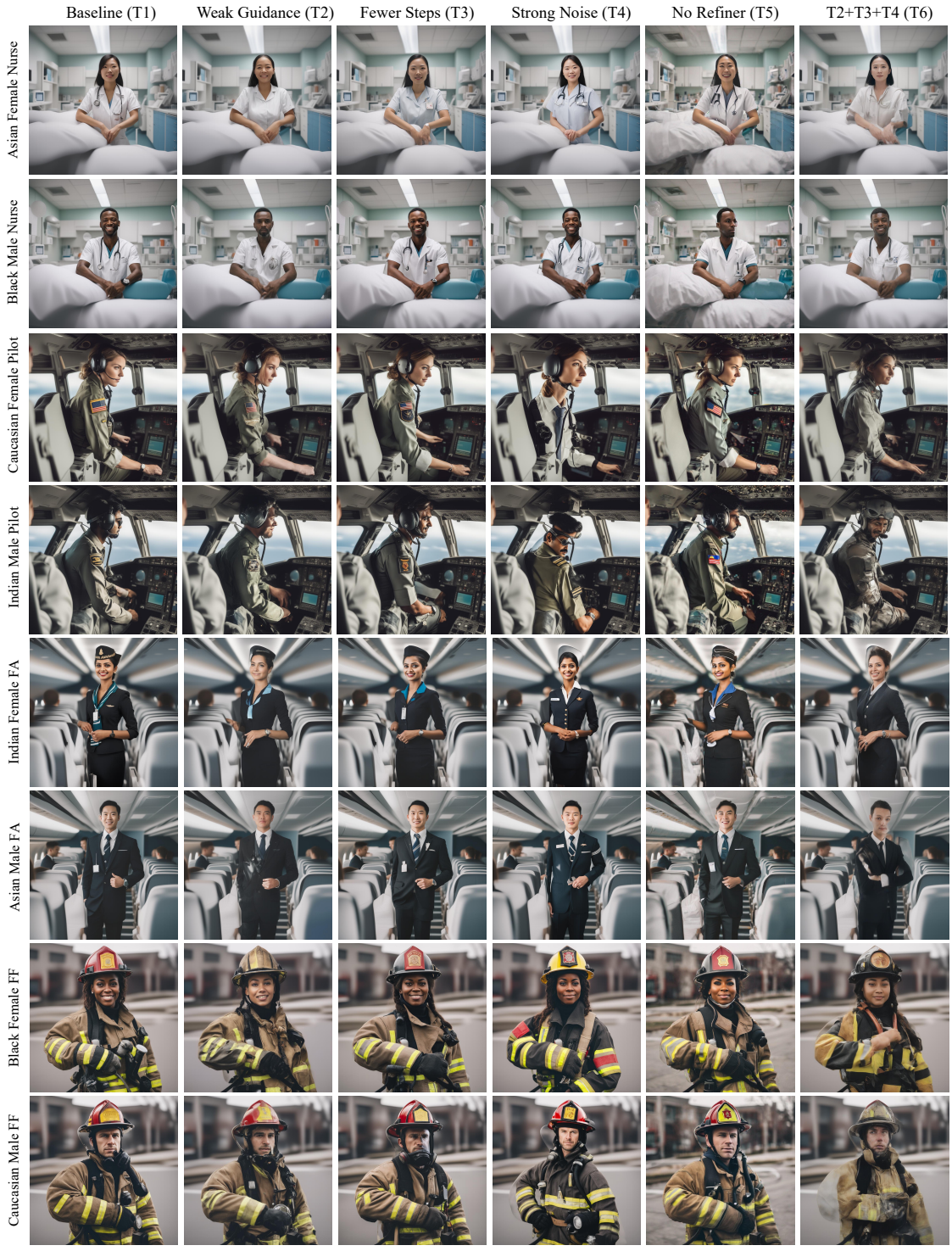


Figure 8. Examples of constructed evaluation datasets for DQA under various text-to-image generation scenarios to controlled degradation of generated image. The scenarios include Baseline, Weak Guidance (T2), Fewer Steps (T3), Strong Noise (T4), No Refiner, and a combination of T2, T3, and T4. Each setting adjusts specific hyperparameters of Stable Diffusion Inpainting [60] 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.

Table 4. H1 Results (Quality Drop): All p-values are < 0.05, confirming degradation is significant.

Degradation	TOPIQ-FLIVE↑	TOPIQ-FPAQ↑	WaDIQaM↑	MUSIQ↑
T1 (Base)	0.7536	0.5399	-0.1906	72.5817
T2	0.7502 (p=0.0002)	0.5304 (p=0.0035)	-0.2067 (p=0.0102)	72.2058 (p=0.0317)
T3	0.7434 (p=0.0000)	0.5089 (p=0.0000)	-0.2282 (p=0.0000)	70.8131 (p=0.0000)
T4	0.7395 (p=0.0000)	0.5017 (p=0.0000)	-0.2277 (p=0.0000)	70.0091 (p=0.0000)
T5	0.7400 (p=0.0000)	0.5013 (p=0.0000)	-0.2268 (p=0.0000)	70.2069 (p=0.0000)
T6	0.7400 (p=0.0000)	0.5012 (p=0.0000)	-0.2265 (p=0.0000)	70.1266 (p=0.0000)

Table 5. H2 Results (Group Parity): All p-values are > 0.05, showing degradation is consistent across gender.

Degradation	TOPIQ-FLIVE↑	TOPIQ-FPAQ↑	WaDIQaM↑	MUSIQ↑
T1	-	-	-	-
T2	0.0006 (p=0.7064)	0.0017 (p=0.7888)	0.0105 (p=0.3994)	0.1470 (p=0.6576)
T3	0.0015 (p=0.4056)	0.0001 (p=0.9858)	0.0183 (p=0.2180)	0.5022 (p=0.1898)
T4	0.0008 (p=0.6732)	0.0013 (p=0.8490)	0.0136 (p=0.3700)	0.4770 (p=0.2584)
T5	0.0001 (p=0.9646)	0.0041 (p=0.5424)	0.0149 (p=0.3352)	0.0982 (p=0.8120)
T6	0.0005 (p=0.8278)	0.0033 (p=0.6264)	0.0198 (p=0.2060)	0.5927 (p=0.1332)

- Each degradation level (T2–T6) significantly reduces perceptual quality compared to the clean baseline (T1).
- The level of degradation is statistically indistinguishable across demographic groups (e.g., gender).

We use four state-of-the-art NR-IQA models from the `pyiqa` library [12], all of which are trained neural networks designed to estimate perceptual image quality without reference images: TOPIQ-FLIVE, TOPIQ-FPAQ [13], WaDIQaM [8], and MUSIQ [37] (higher scores = better quality).

Hypothesis tests. We define two hypotheses and test them using 10,000-sample bootstrapping:

H1 (Quality Drop): The mean metric score at level **T1** (clean) is equal to the score at each degraded level **T2–T6**. (This hypothesis is expected to be rejected.)

H2 (Group Parity): For each degradation level x, the degradation gap between groups, $\Delta_x = (\bar{T}_1^{\text{male}} - \bar{T}_x^{\text{male}}) - (\bar{T}_1^{\text{female}} - \bar{T}_x^{\text{female}})$, is zero. (This hypothesis is expected to be retained.)

Results: H1 - Degradation is Significant. As shown in Table 4, all p-values are well below 0.05. This confirms that each degradation scenario (T2–T6) meaningfully reduces perceptual quality compared to the clean baseline. We therefore **reject H1**.

Results: H2 - Degradation is Consistent Across Gender. As shown in Table 5, all p-values are well above 0.05. This means we **do not reject H2**, as there is no statistical evidence that the degradation effect differs between the male and female groups.

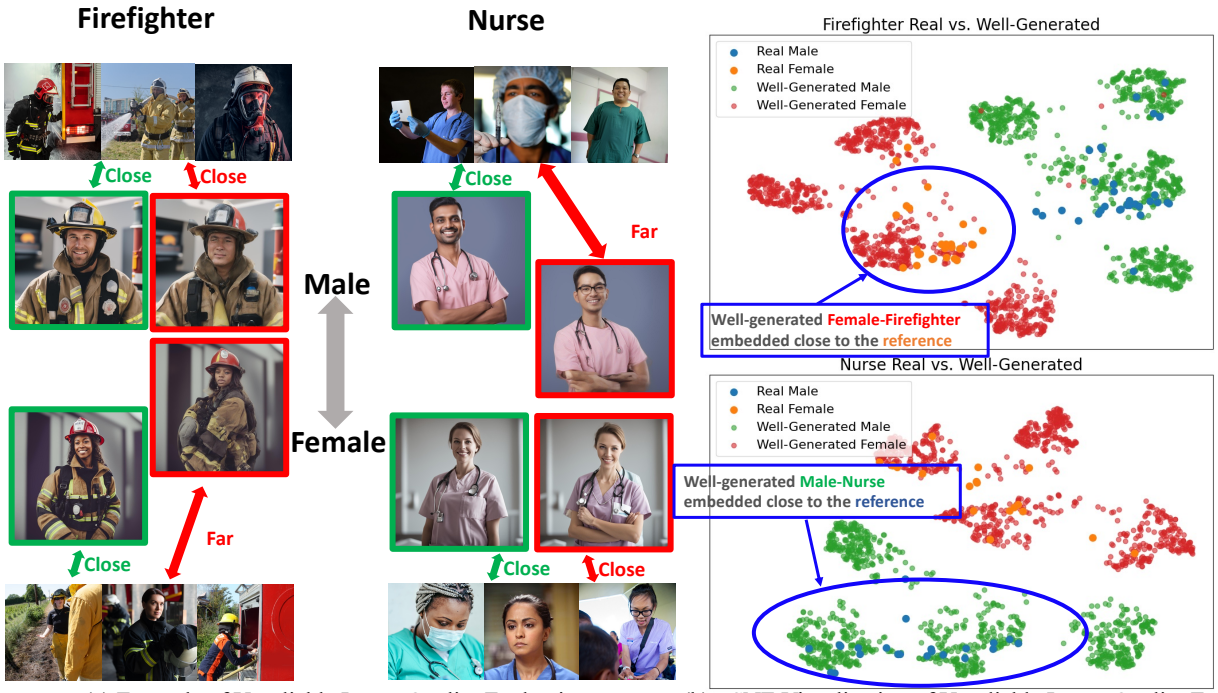
This statistical analysis supports our claim that the synthetic dataset is well-controlled: each degradation level reduces image quality, and the degradation effect is consistent across demographic groups.

B. Details of Synthetic Data in Figure 3

To construct the synthetic dataset to analyze the impact of reference set, 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) Example of Unreliable Image Quality Evaluation

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

Figure 9. (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 [57] image encoder shows clear separation between gender clusters and correct placement of most samples, highlighting the encoder's unreliable behavior under poor-quality conditions.

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

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.

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

Fig. 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. 9 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. 9 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.

Table 6. Ablation on the image encoder used for DQA-Guidance ($\lambda_1 = 20, \lambda_2 = 100$). A more reliable encoder (DINO-RN50) yields stronger improvements.

Model	Avg.MMD ↓	MMD Gap ↓	FID ↓	FID Gap ↓	TOPIQ ↑	TOPIQ Gap ↓	BLIP ↑	BLIP Gap ↓
No Guidance	109.93	12.57	137.19	12.85	0.7224	0.0167	0.9925	0.0150
DQA-Guidance (DINO-RN50)	**96.38**	**9.12**	138.97	11.22	0.7225	0.0164	0.9925	0.0150
DQA-Guidance (DINO-RN50, Numerator-Only)	104.79	9.52	137.00	11.56	0.7223	0.0166	0.9900	0.0200
DQA-Guidance (CLIP-VIT)	110.23	10.68	138.06	10.85	0.7224	0.0170	0.9925	0.0150

D. Additional Ablation Studies

This section provides further ablation studies on the components and generalizability of DQA-Guidance.

D.1. Impact of Guidance Encoder

To analyze the effect of the encoder choice within DQA-Guidance, we ablate our default DINO-RN50 with CLIP-ViT. This test confirms that DQA-Guidance is model-agnostic and compatible with any image encoder. However, as shown in Table 6, the effectiveness varies with the encoder's reliability. The more reliable DINO-RN50 (as identified in our main paper) yields stronger improvements in both overall quality (DINO-MMD) and fairness (DINO-MMD Gap). We also report FID-based evaluation and general IQA methods (Appendix D.3.1) results for reference.

D.2. Ablation on Guidance Formulation: Numerator-Only

One might question why our guidance (Eq. 4) uses the full DQA score (Eq. 1) instead of just its numerator: $\left|D\big(f(A_{\text{gen}}), f(A_{\text{ref}})\big) - D\big(f(B_{\text{gen}}), f(B_{\text{ref}})\big)\right|$. While using only the numerator as the fairness term (λ_1) simplifies the objective, it could inadvertently degrade global quality, as the term no longer penalizes a high denominator (poor overall quality).

Our full method prevents this by using the complete DQA score (Num/Den) for the λ_1 term and adding the denominator as a separate global quality regularizer, controlled by λ_2. As shown in Table 6, this "numerator-only" approach reduces the fairness gap, but the overall image quality (Avg.MMD) is significantly worse than our full method.

D.3. Generalization to Other Diffusion Models

To test the adaptability of DQA-Guidance, we apply it to the DeepFloyd [2] model, which consists of three diffusion stages. Since it follows a standard denoising process, our guidance can be integrated without architectural changes. As shown in Table 7, DQA-Guidance successfully improves both image quality (Avg. MMD) and fairness (MMD Gap), suggesting its adaptability across different diffusion frameworks.

Table 7. DQA-Guidance applied to the DeepFloyd model.

Model	Avg.MMD ↓	MMD Gap ↓
Baseline (No Guidance)	196.86	17.80
DQA-Guidance ($\lambda_1 = 1, \lambda_2 = 1$)	193.72	15.01
DQA-Guidance ($\lambda_1 = 5, \lambda_2 = 1$)	176.90	13.01

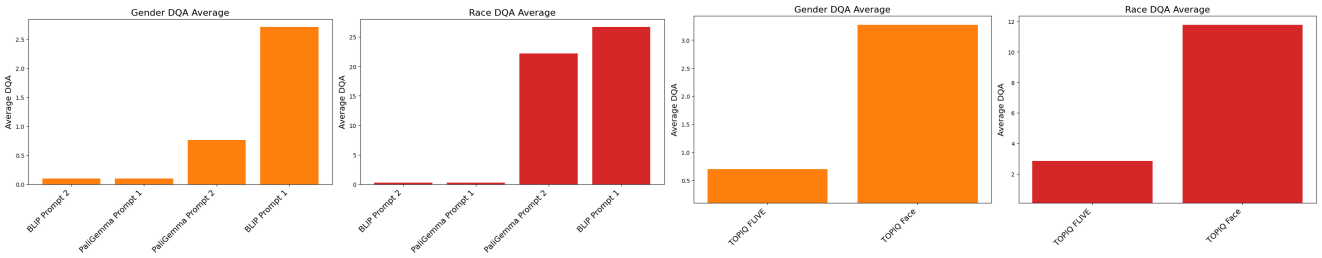
D.3.1. DQA on Different Types of Image Quality Assessment

In addition to our approach, other methods for assessing image quality include visual question answering (VQA) [49] and neural networks specifically trained for quality evaluation [13, 41, 68].

In [49], 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.

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

We also adapt DQA to image quality assessment (IQA) models that output indicators of general image quality. For example, TOPIQ [13] is a supervised network designed for image quality evaluation. It is trained on datasets such as FLIVE [76] for general images or CGFIQA [15] for facial images, using a regression task to predict quality scores.



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

Figure 10. 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.

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:

- **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 [44] and PaliGemma [5] 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 [76] for general images and another trained on the CGFIQA dataset [9] for facial images.

The experimental results for these different types of image quality assessments are visualized in Fig. 10. 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.

E. Experimental Results for Medical Image

E.1. Details of the Chest X-ray Dataset

We use the NIH ChestX-ray14 dataset [74], 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.

E.2. Details of Synthetic Chest X-ray Generation

To generate synthetic Chest X-ray images, we use a pretrained ImageGen model [61] trained on the ROCO dataset [53], which contains paired image and text data for medical purposes. The pretrained model is available on HuggingFace [75] 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}."

E.3. Negative Impact of Quality Bias in Generative Models

E.3.1. Quality Bias in Synthetic Data Augmentation Affects Classification Unfairness

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

Table 8. 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(ΔAUC) 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) ↓	max(ΔAUC) ↓	$\overline{\text{DQA}}$ ↓
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 ↓)	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 ↑)	85.54±0.12	77.95±0.32	77.81±0.33	**2.62±0.39**	**8.93±2.46**	0.5495

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 [28] using chest X-ray images from the Chest X-ray dataset [74], a dataset known to exhibit fairness issues, as evidenced by differing AUC scores across demographic groups [42]. To enhance classifier's performance, a user might employ text-to-medical-image generation models [61] trained on the ROCO dataset [53] as a data augmentation strategy. In our initial experiments, we generate 1,000 images per gender and class for augmentation.

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 8, even worsen (denoted as Full Augmentation). This is evidenced by higher values of Avg(ΔAUC) and max(ΔAUC), calculated as

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

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.

E.3.2. Positive Impact of Quality-Fair Data Augmentation

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 [19]. 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 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. E.3.1. Initial experiments in Sec. E.3.1 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 demo-

graphic 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 8, 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.

E.4. DQA analysis for Medical Image

E.4.1. Constructing Reference Dataset for Medical Image

In the medical image, we utilize the Chest X-ray diagnosis dataset in Sec. E.3.1 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 [7] by applying transformations to the test set, creating poor-quality images as shown in Fig. 11 (a).

E.4.2. 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) [14], 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.

E.5. DQA-Guidance for Medical Image

E.5.1. 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 3,000 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

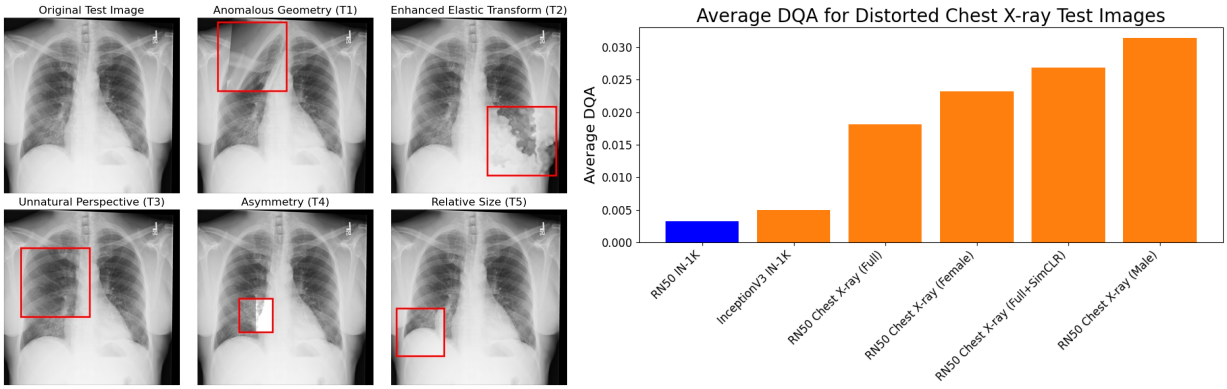
Algorithm 1 Finding Fair and Unfair Subsets Using Influence Scores for DQA

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 $\text{DQA}^{(m,\text{fair})} <$ best_DQA **then**
    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 $\text{DQA}^{(m,\text{unfair})} >$ worst_DQA **then**
    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: $(S_A^{\text{fair}}, S_B^{\text{fair}}), (S_A^{\text{unfair}}, S_B^{\text{unfair}})$

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.

E.5.2. Result Analysis for DQA-Guidance

Fig. 12 demonstrates the clear impact of DQA-Guidance on medical image generation. Compared to the baseline ($\lambda_1 = 0$), increasing λ_1 effectively reduces quality dis-



(a) Example of Transforms Mimicking Image Generation Failure

(b) DQA of Various Models for Distorted Images

Figure 11. (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.

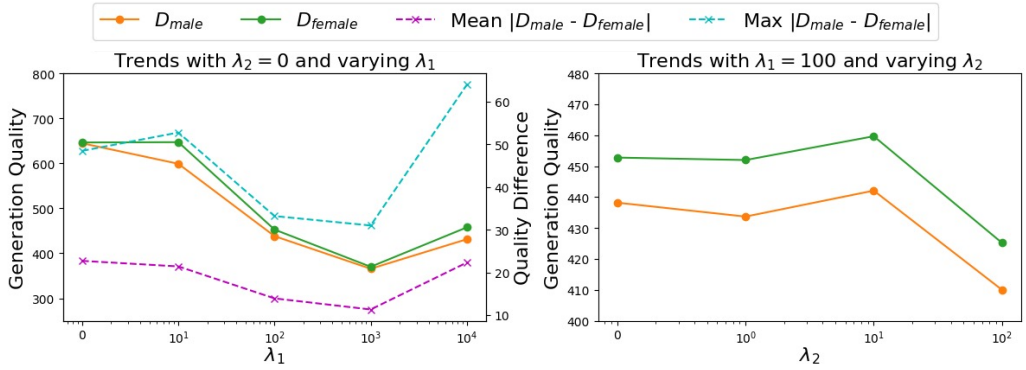


Figure 12. Experimental results for generation quality and quality disparities with DQA-Guidance. The left plot shows the impact of λ_1 on generation quality for each demographic group in Chest X-ray image generation (lower values indicate better quality) and displays the average and maximum quality 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.

parities 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. 13. 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.

E.6. Impact of DQA-Guidance on Downstream Tasks

In line with Appendix E.3.2, 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 8 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 9 shows the classification performance when generative samples created with DQA-Guidance are used for

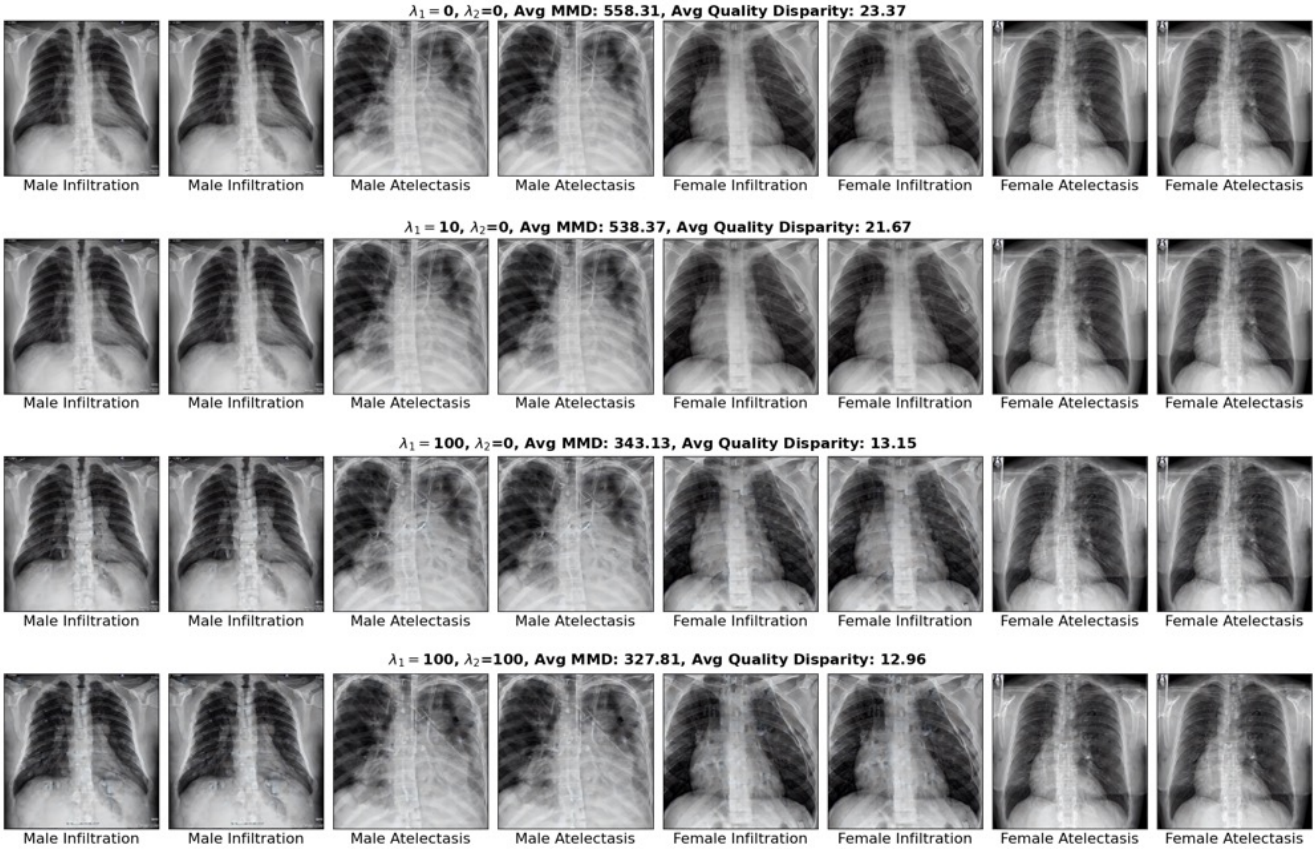


Figure 13. Qualitative results of DQA-Guidance for medical image generation. The examples highlight improvements primarily in texture refinement, demonstrating the method's ability to enhance overall image quality while addressing disparities across different conditions.

Table 9. 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} ↑	$\text{AUC}^{\text{female}}$ ↑	Avg(ΔAUC) ↓	max(ΔAUC) ↓
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

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.