# Stay Fair!

## Ensuring Group Fairness in Diffusion Models Across Guidance Scales

**Myeongsoo Kim[1],** Eunji Kim[2*], Minwoo Chae[1], Sangwoo Mo[1]

[1]POSTECH [2]Amazon (*Work done prior to joining Amazon)

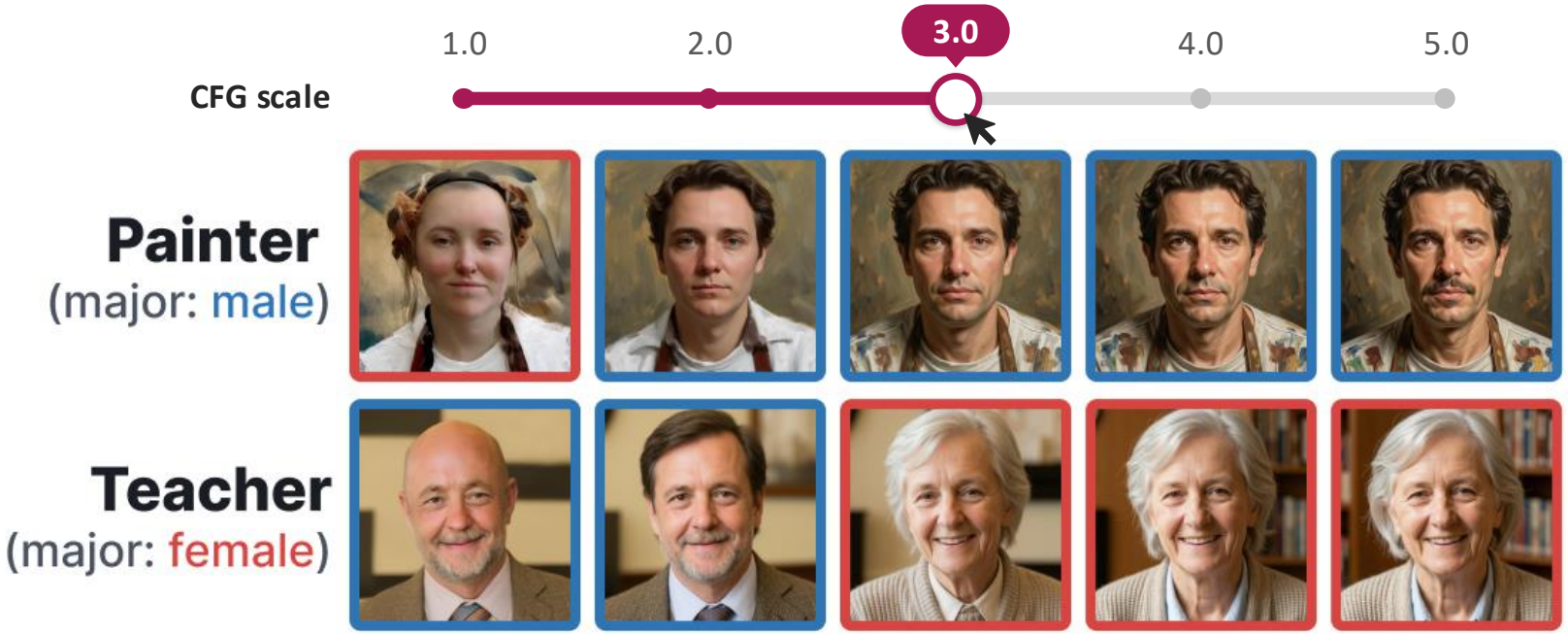
NeurIPS 2026

**Paper:** https://openreview.net/pdf?id=tIDWhFyqX7

**Project Page:** https://kim-myeong-soo.github.io/stayfair/

# Background

- Diffusion models generate **imbalanced outputs** across sensitive attributes (e.g., gender, race).
- **Guidance, a widely used technique** for prompt alignment, also **amplifies the prompt's social bias**.
- In practice, **users adjust the guidance scale** across a wide range to match their preferences.




**Users pick the scale** by preference

**Occupational details added**

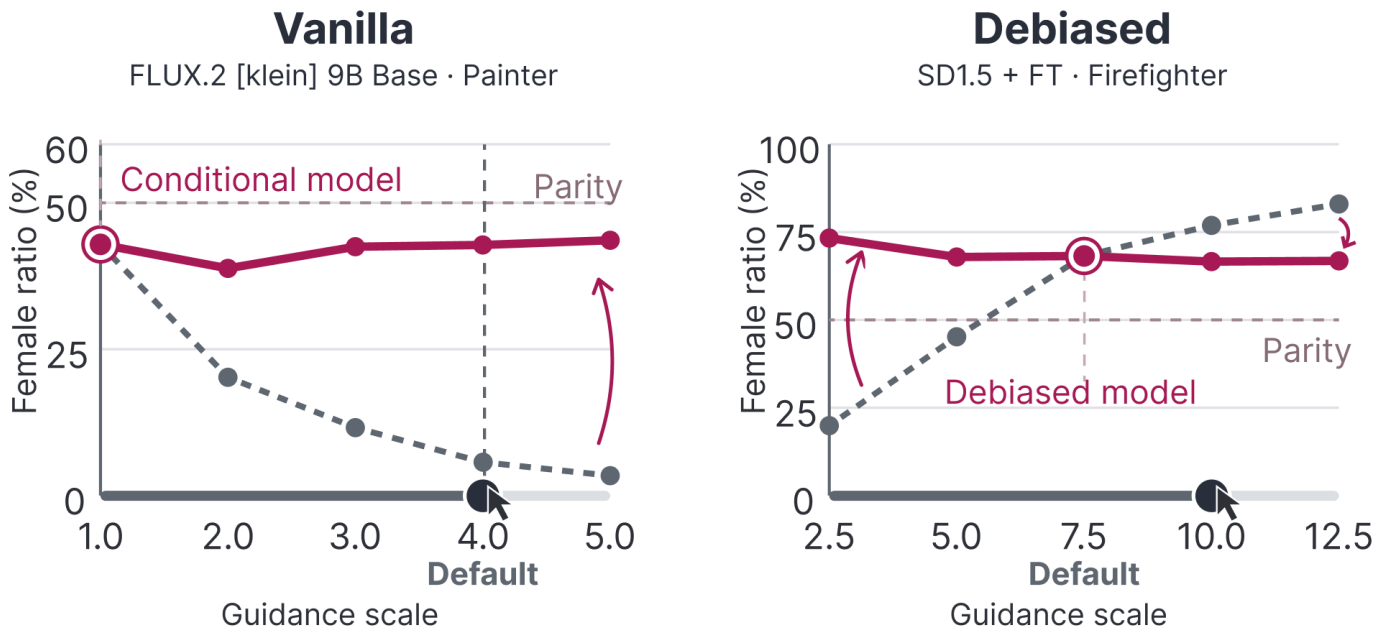
e.g., paint-stained apron, bookshelves

**Gender shifts to the majority**

Painter → male, Teacher → female

# Motivation

- Prior work evaluates and debiases **only at the default guidance scale**.
- However, fairness **collapses at other scales**, for both vanilla and debiased models.
- **Our goal:** ensure group fairness **across all user-selected guidance scales**.




**CFG**

Fairness collapses at unconsidered scales

**StayFair (Ours)**

Fairness holds across user-selected scales

# Problem Formulation

- We decompose the **total bias** into static **model bias** and scale-dependent **guidance bias**.

**Total bias = Guidance bias + Model bias**

$$Q^w - T = Q^w - Q^{w_{\text{ref}}} + Q^{w_{\text{ref}}} - T$$

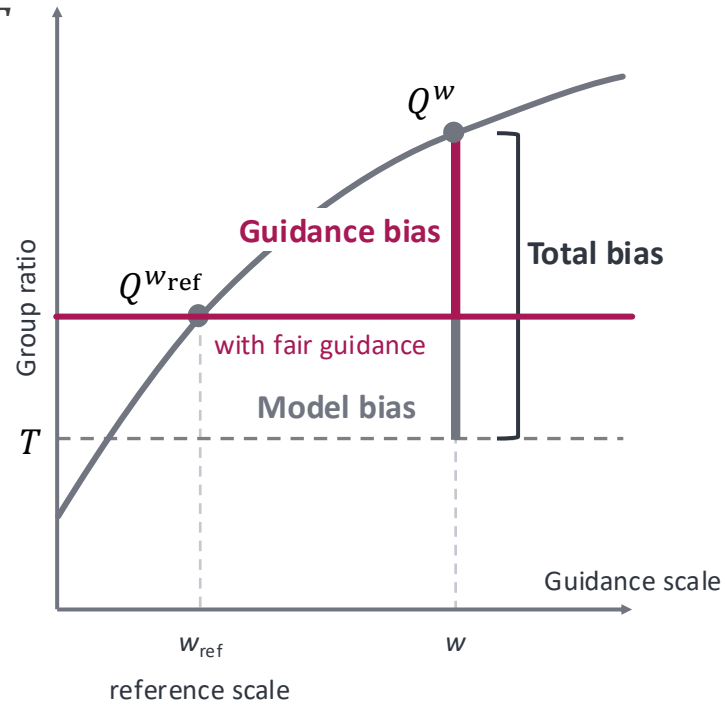
where $Q^w$ is the group distribution generated at scale $w$, $T$ is the target distribution.

- **Guidance is fair** if its bias is identically zero at any scale a user may select:

$$Q^w - Q^{w_{\text{ref}}} = 0 \text{ for any } w$$

- **StayFair** debiases the **guidance function *f*** across group $a$.

$$\nabla \log p_\theta^w(x_t \mid y) = \nabla \log p_\theta(x_t) + w \nabla \log f(x_t, t)$$



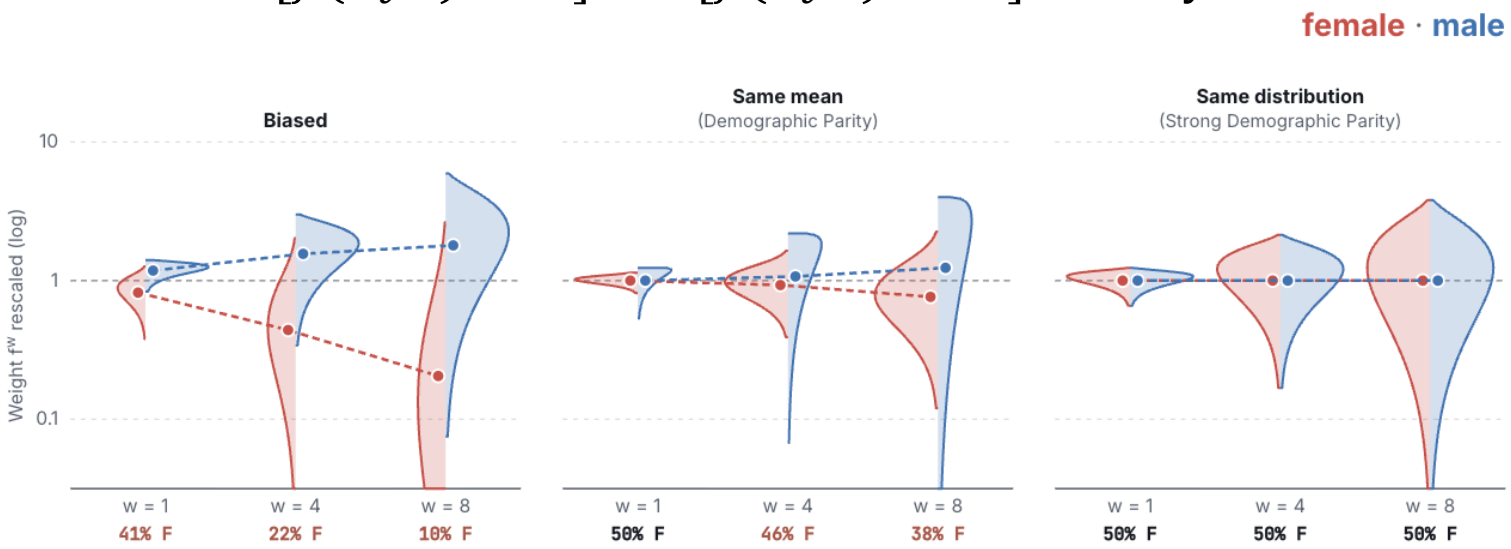
# Proposed Method: StayFair in CG

- **Classifier Guidance (CG):** Train classifier ***f*** toward **Strong Demographic Parity (SDP)**.

$$f(X_t, t) \mid a \overset{d}{=} f(X_t, t) \mid a', \qquad \forall a, a'$$

- Then $p_\theta^w(x_t \mid y) \propto p_\theta(x_t) f(x_t, t)^w$ **stays fair at every guidance scale**.

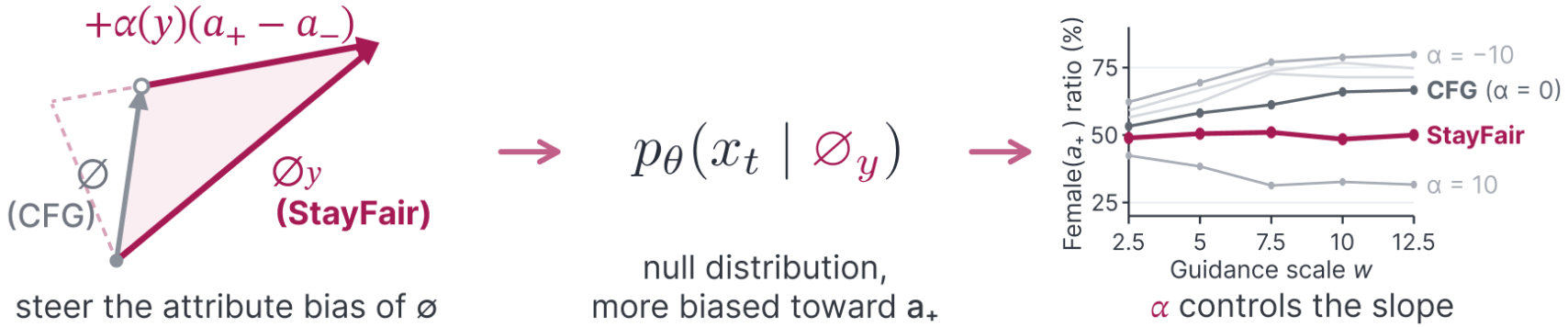
$$\mathbb{E}[f(X_t, t)^w \mid a] = \mathbb{E}[f(X_t, t)^w \mid a'] \text{ for any } w$$



# Proposed Method: StayFair in CFG

- In CFG, $f$ is implicit, so SDP condition cannot be enforced directly.
- StayFair uses an **adaptive null prompt $\emptyset_y$** in place of the fixed null prompt $\emptyset$.

**StayFair guidance:** $\nabla \log f(x_t, t) = \nabla \log p_\theta(x_t \mid y) - \nabla \log p_\theta(x_t \mid \emptyset_y)$




- All we have to do is find $\alpha(y)$
  - **StayFair (Search):** search $\alpha(y)$ for the flattest curve
  - **StayFair (Estimate):** look up $\alpha(y)$ in a search-built table, indexed by female counts in 5 images at min/max scales

# Experimental Results: Classifier Guidance (CG)

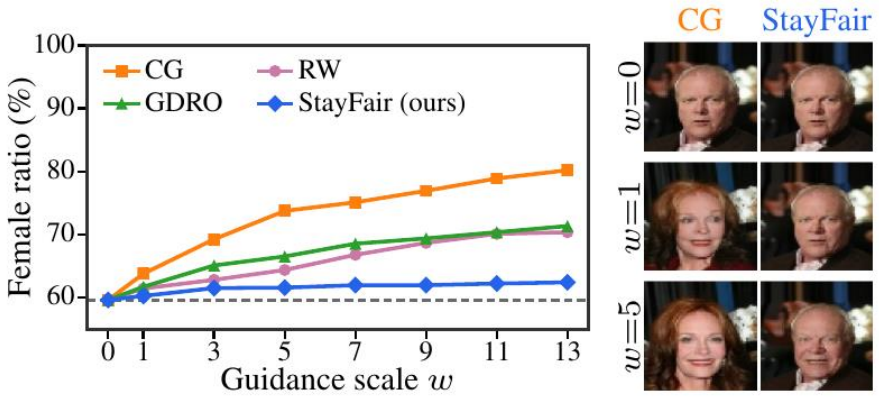
**Setup:** ADM-G on CelebA (64×64) with gender as the sensitive attribute, across w ∈ {0, 1, 3, …, 13}.

| (c) CG on CelebA | Smiling | | | Blond Hair | | |
|---|---|---|---|---|---|---|
| Classifier | Range (%) ↓ | FID ↓ | $FD_{DINOv2}$ ↓ | Range (%) ↓ | FID ↓ | $FD_{DINOv2}$ ↓ |
| CG | 20.6 ±2.2 | 17.99 | 87.51 | 39.2 ±3.5 | 22.70 | 151.04 |
| RW | 11.0 ±3.4 | 16.96 | 83.39 | 32.9 ±3.1 | 20.05 | **136.55** |
| GDRO | 11.9 ±3.5 | 17.50 | 85.28 | 29.2 ±2.8 | 22.29 | 152.70 |
| StayFair (ours) | **3.5** ±1.2 | **16.66** | **82.56** | **12.4** ±2.6 | **19.68** | 139.65 |

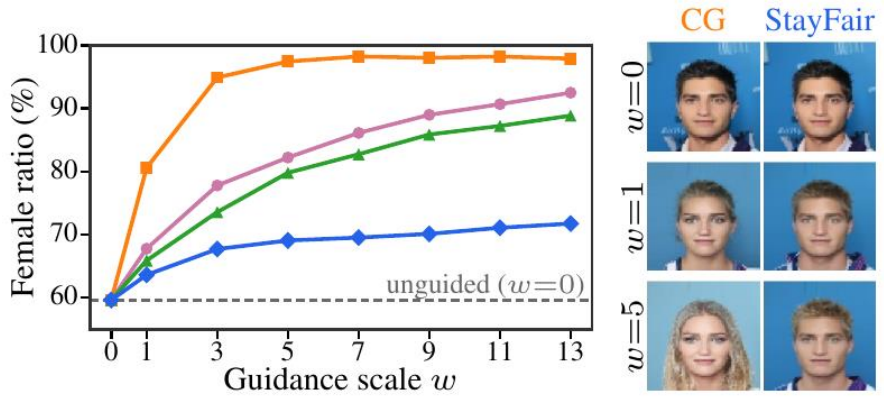
**Smiling / Blond Hair:** weak / strong gender imbalance. **FID, $FD_{DINOv2}$:** fair-intra versions (reference set matched to target ratios).
**Range:** max − min female ratio across w (guidance bias). **±:** std over six seeds.

1. **Bias range ↓**
   20.6 → 3.5, 39.2 → 12.4

2. **Best fair-intra FID**
   in both conditions




(a) Smiling




(b) Blond Hair

# Experimental Results: Vanilla Model

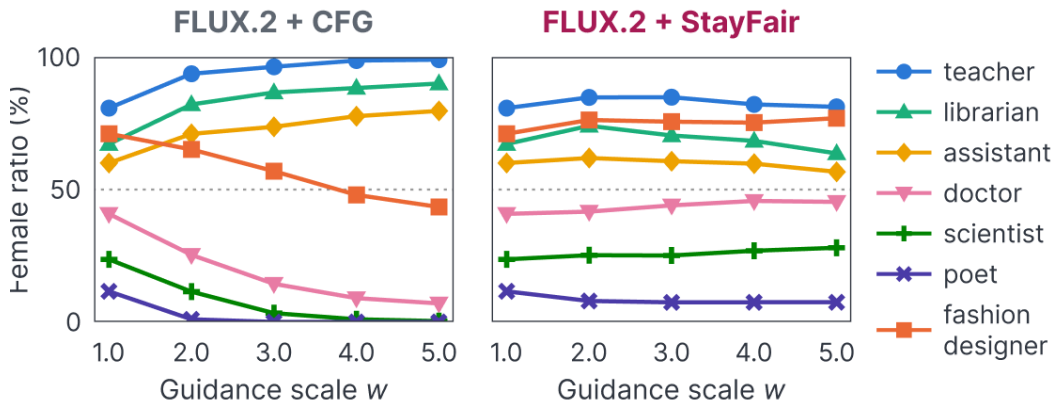
**Setup:** We measure gender–occupation bias for 33 occupations across 5 scales on 3 T2I diffusion models.

| (a) Vanilla generation | | Bias (%) ↓ | | | Quality ↑ | | |
|---|---|---|---|---|---|---|---|
| Model | Guidance | Avg. | Worst | Range | CLIP | Aesth. | Pick |
| SD1.5 | CFG | 28.0 | 32.5 | 10.7 | 27.61 | **6.70** | **19.99** |
| | **StayFair (ours)** | **27.6** (−0.4) | **30.8** (−1.7) | **7.2** (−3.5) | **27.62** (+0.01) | **6.70** (0.00) | 19.96 (−0.03) |
| SD3 | CFG | 39.7 | 43.0 | 9.1 | 27.36 | **6.62** | **20.29** |
| | **StayFair (ours)** | **33.9** (−5.8) | **36.6** (−6.4) | **6.3** (−2.8) | **27.38** (+0.02) | 6.61 (−0.01) | 20.23 (−0.06) |
| FLUX.2 | CFG | 39.4 | 44.3 | 15.9 | **28.18** | **6.93** | 21.15 |
| | **StayFair (ours)** | **32.7** (−6.7) | **36.0** (−8.3) | **8.4** (−7.5) | 28.17 (−0.01) | **6.93** (0.00) | **21.18** (+0.03) |

**SD1.5:** Stable Diffusion 1.5 **SD3:** Stable Diffusion 3 Medium **FLUX.2:** FLUX.2 [klein] 9B Base
**Bias** = |female ratio − 50%| (pp). **Avg. / Worst:** mean / max bias. **Range:** max − min female ratio (guidance bias).

1 **Bias range ↓**
**image quality ≈ CFG**

2 **Avg. & Worst bias ↓**

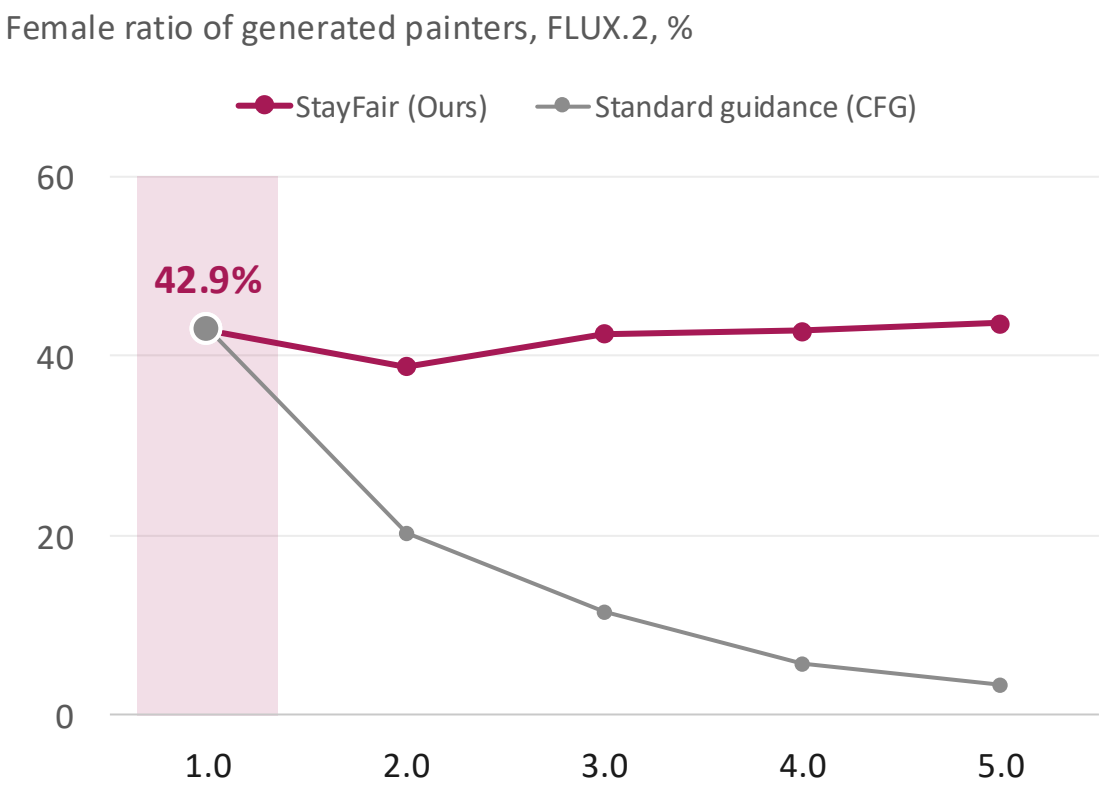



**CFG** drifts to the extremes.

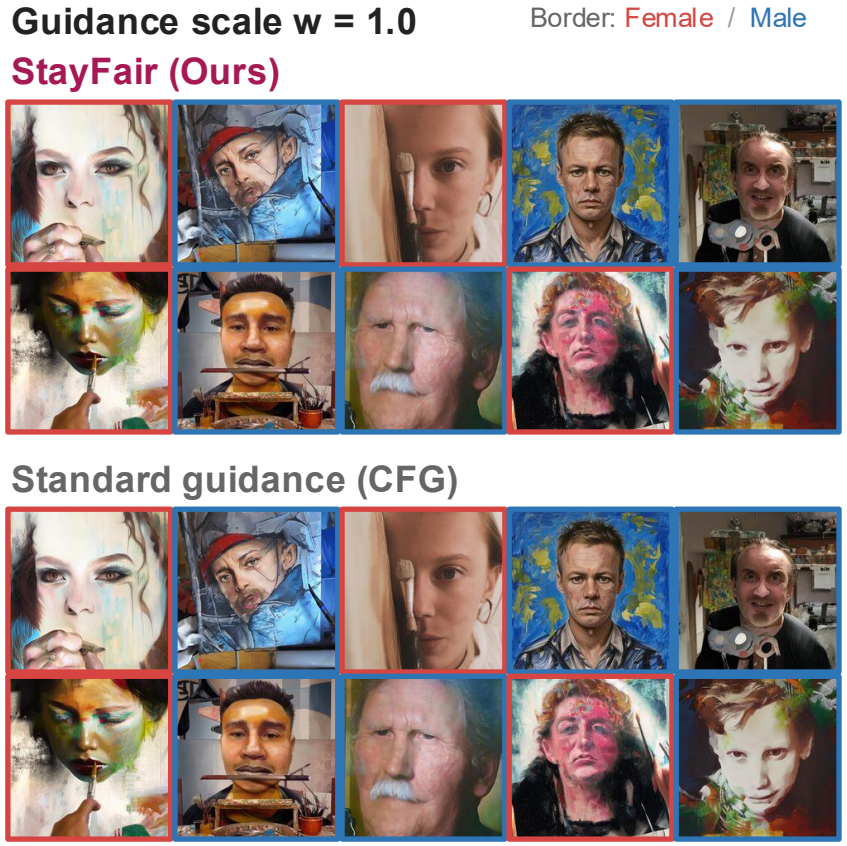
**StayFair** holds them steady.

# Experimental Results: Qualitative Examples

- StayFair preserves minority gender across scales, with similar visual context as CFG.

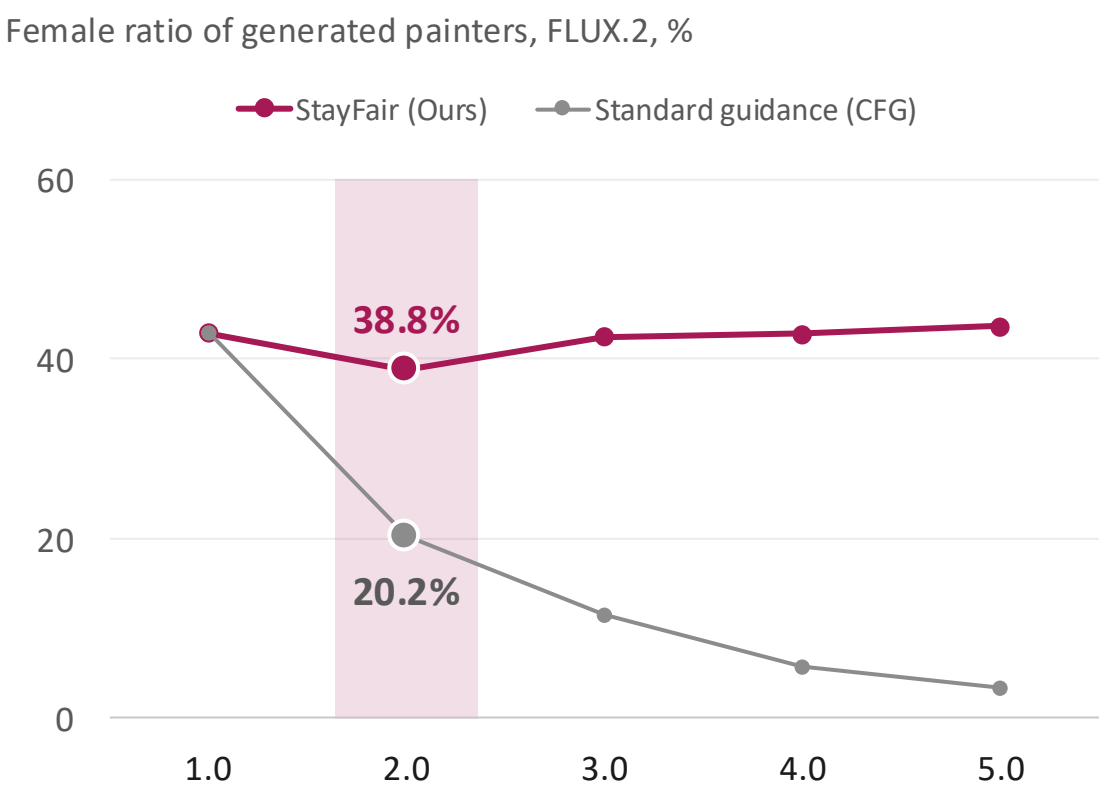



Source: StayFair project page, Motivation demo (prompt "A photo of the face of a painter")

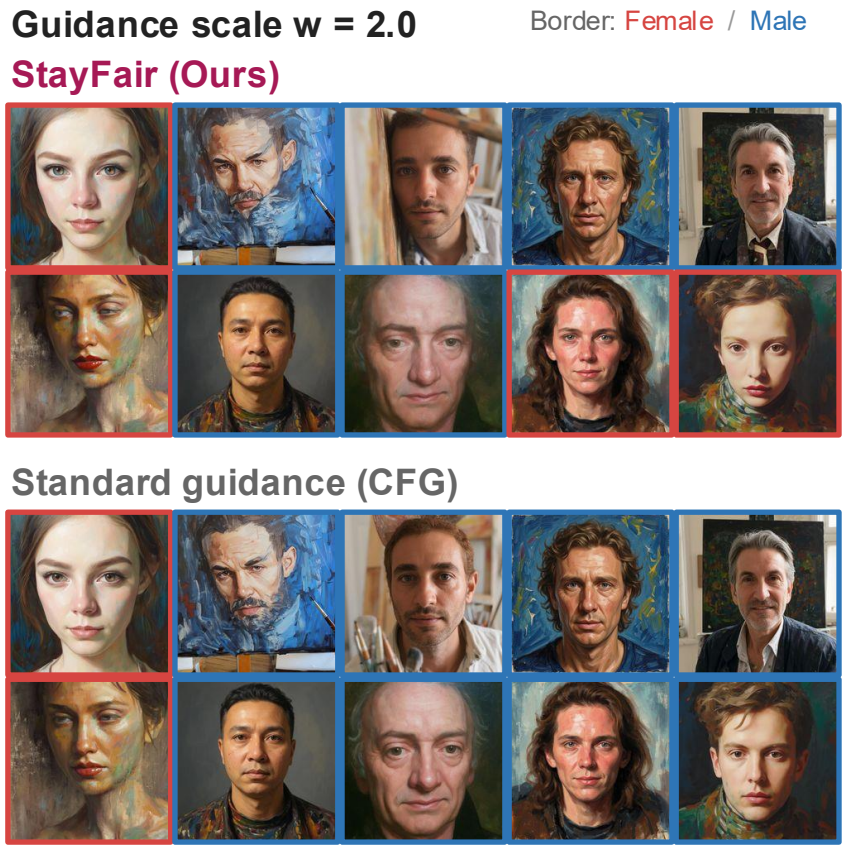


# Experimental Results: Qualitative Examples

- StayFair preserves minority gender across scales, with similar visual context as CFG.

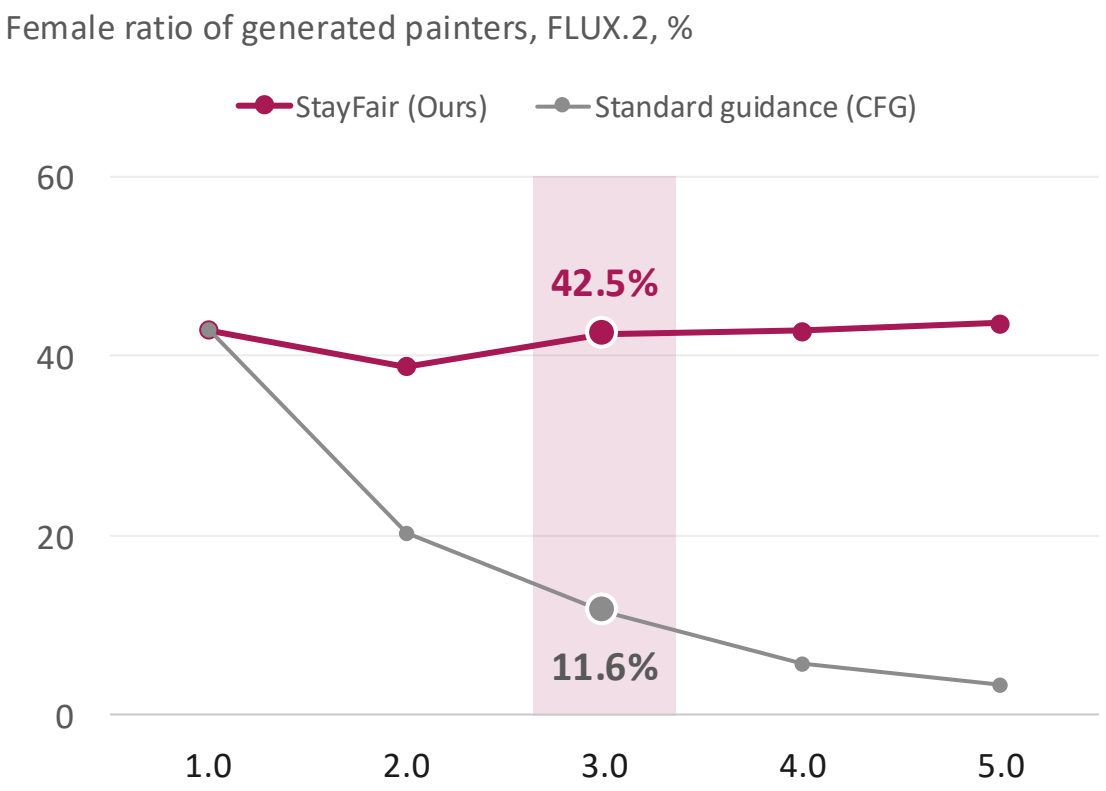



Source: StayFair project page, Motivation demo (prompt "A photo of the face of a painter")

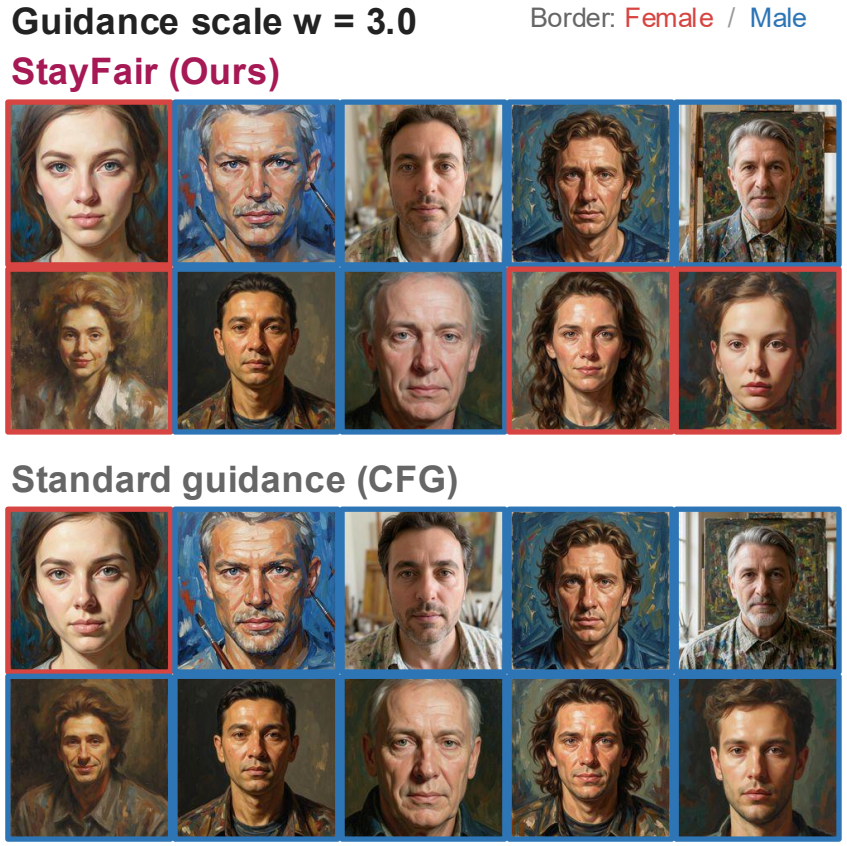


# Experimental Results: Qualitative Examples

- StayFair preserves minority gender across scales, with similar visual context as CFG.

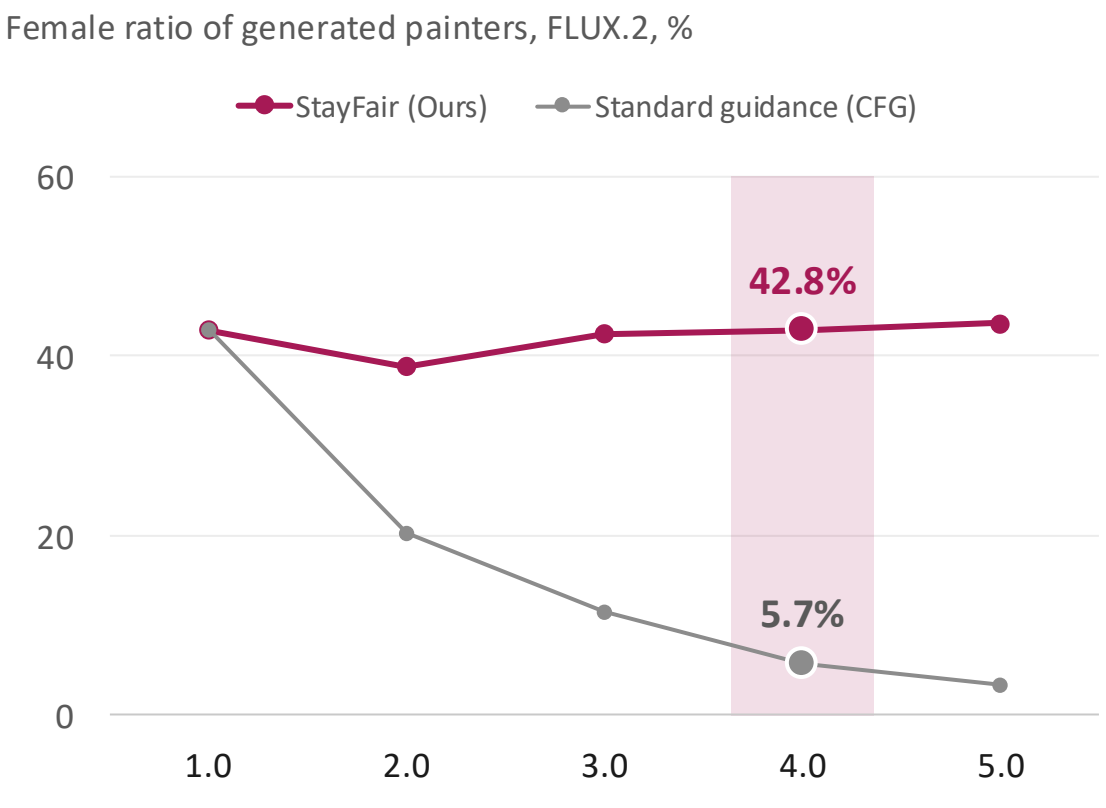



Source: StayFair project page, Motivation demo (prompt "A photo of the face of a painter")

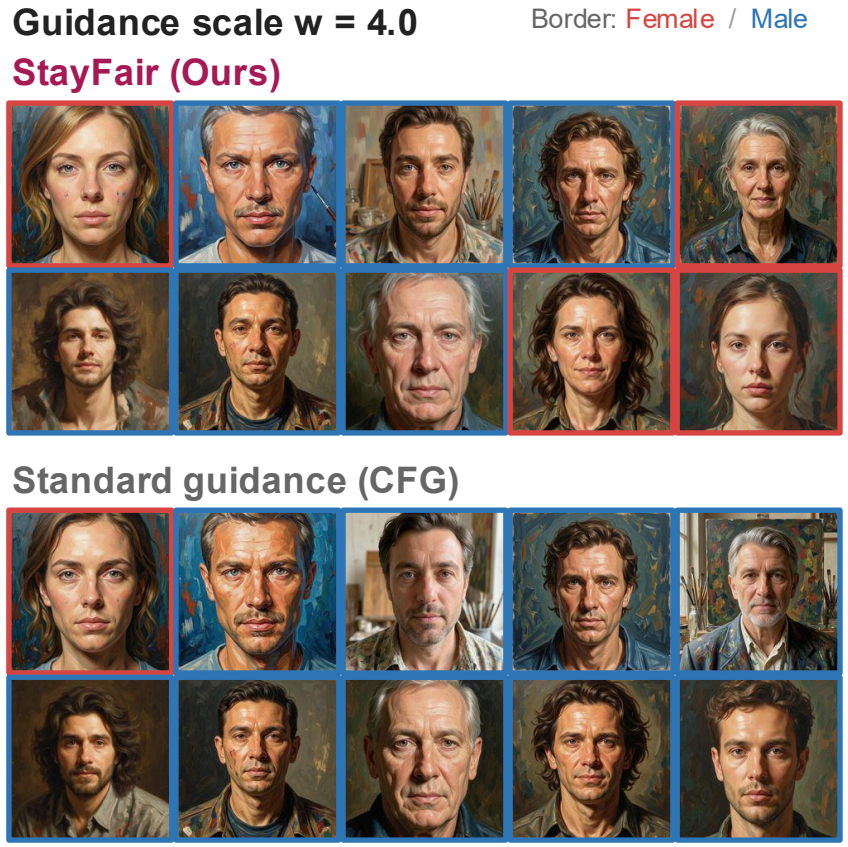


# Experimental Results: Qualitative Examples

- StayFair preserves minority gender across scales, with similar visual context as CFG.

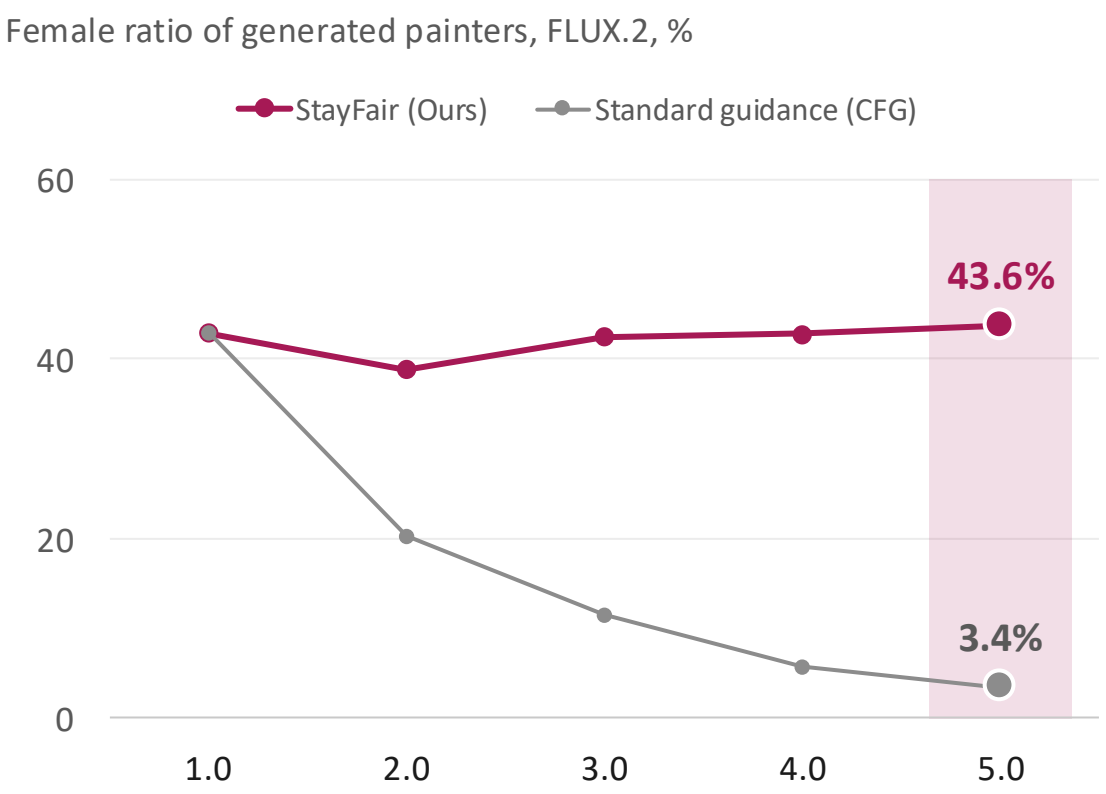



Source: StayFair project page, Motivation demo (prompt "A photo of the face of a painter")

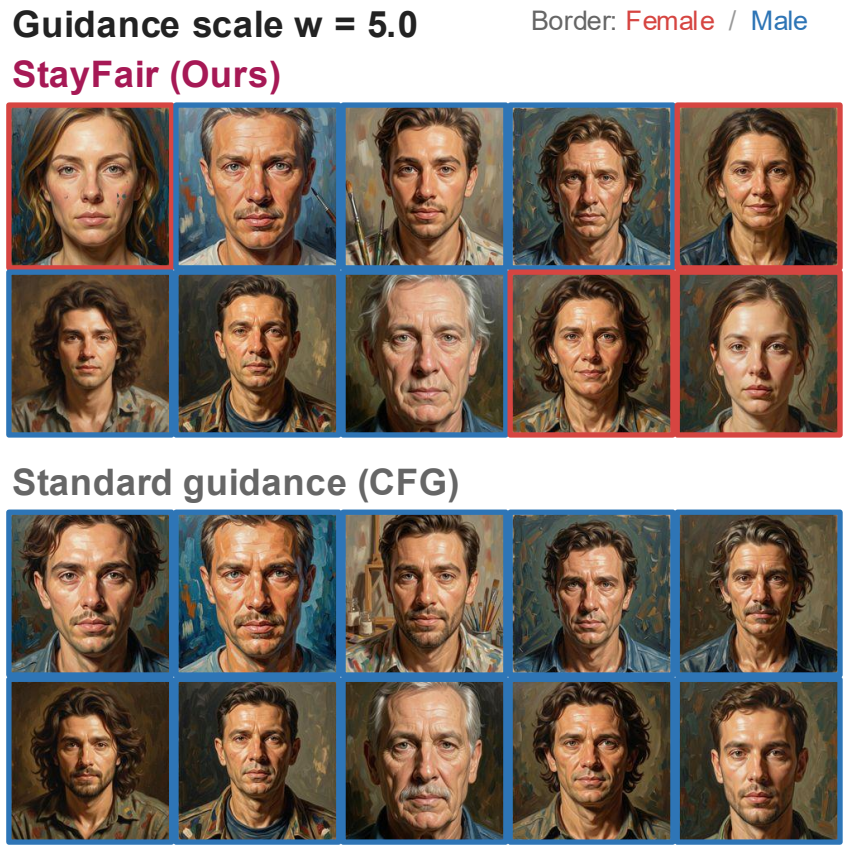


# Experimental Results: Qualitative Examples

- StayFair preserves minority gender across scales, with similar visual context as CFG.



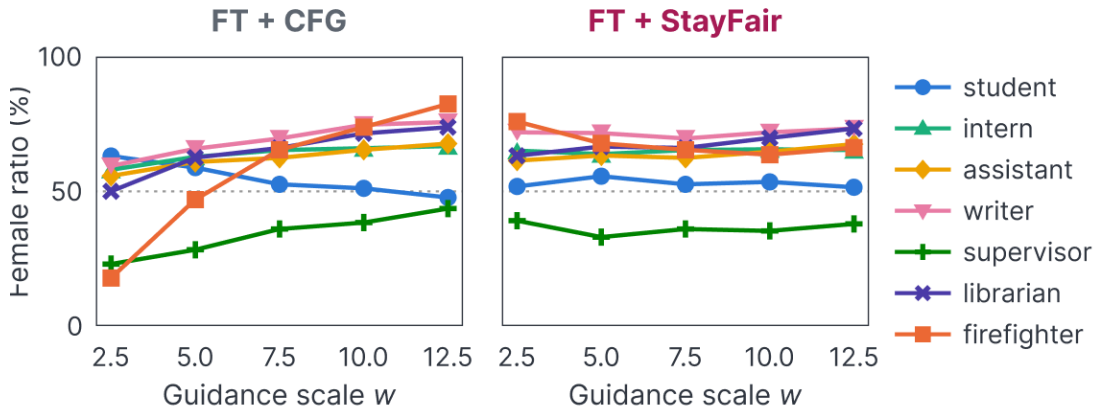

Source: StayFair project page, Motivation demo (prompt "A photo of the face of a painter")



# Experimental Results: Debiased Model

| (b) Debiased generation | | | Bias (%) ↓ | | | Quality ↑ | | |
|---|---|---|---|---|---|---|---|---|
| Method | Scales | Guidance | Avg. | Worst | Range | CLIP | Aesth. | Pick |
| UCE | Single | CFG | 18.0 | 23.4 | 13.2 | 27.28 | 6.70 | **19.96** |
| | Multi (random) | CFG | 19.1 | 24.2 | 13.7 | **27.38** | 6.71 | 19.89 |
| | Multi (per-scale) | CFG | **17.7** | 29.4 | 30.6 | 27.32 | **6.71** | 19.92 |
| | Single | StayFair (ours) | 18.5 (+0.5) | **22.2** (−1.2) | **7.6** (−5.6) | 27.28 (0.00) | 6.70 (0.00) | **19.96** (0.00) |
| FT | Single | CFG | 14.1 | 21.3 | 19.3 | 27.54 | 6.77 | 20.03 |
| | Multi (random) | CFG | 16.5 | 21.3 | 12.0 | **27.63** | **6.79** | **20.07** |
| | Multi (per-scale) | CFG | **13.6** | 22.0 | 17.9 | 27.62 | 6.77 | 20.03 |
| | Single | StayFair (ours) | 13.8 (−0.3) | **18.2** (−3.1) | **9.4** (−9.9) | 27.53 (−0.01) | 6.77 (0.00) | 20.03 (0.00) |

**Bias** = |female ratio − 50%| (pp). **Avg. / Worst:** mean / max bias. **Range:** max − min female ratio (guidance bias).



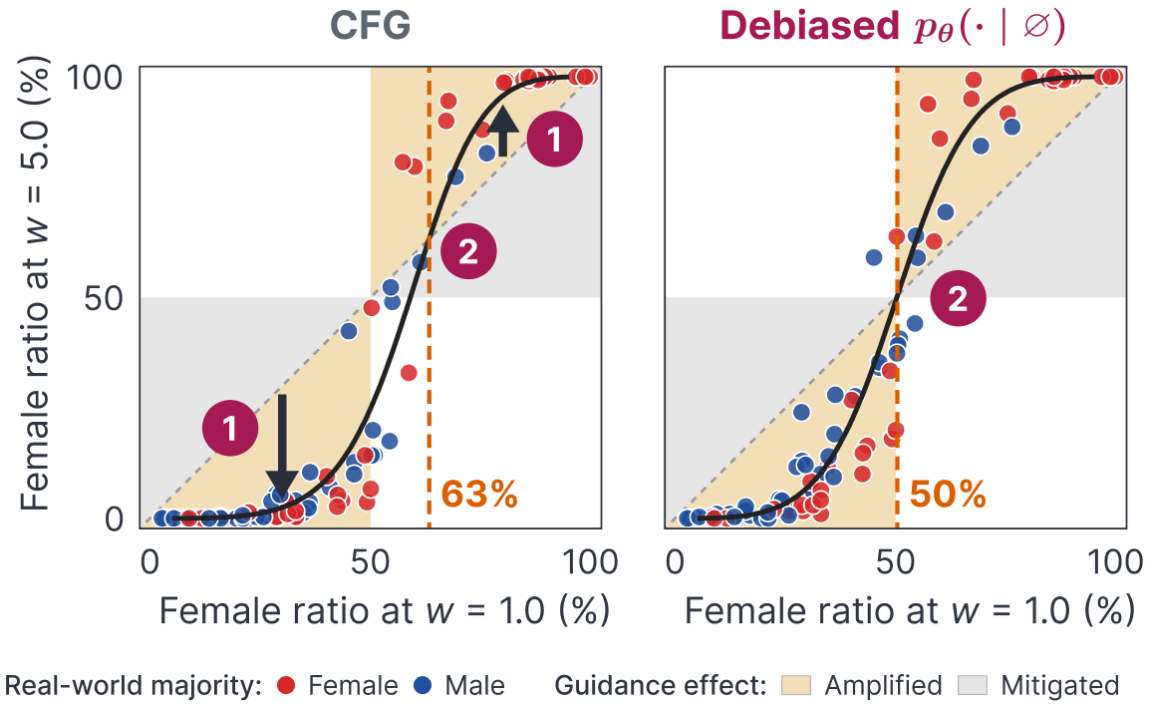

1. **Bias range ↓ image quality ≈ CFG**
2. **Worst bias ↓**
3. **Lower bias range than Multi** (extended debiasing)

**CFG** drifts to the extremes.

**StayFair** holds them steady.

# Experimental Results: Analysis

**Phenomenon** Relative to workforce statistics, T2I models (SD3, FLUX.2) **overrepresent male majorities** more often than female ones. **Why?**




1. Guidance amplifies **male majorities** more than female ones:
   - **more often** (**96.2%** vs. 39.6%)
   - **more strongly** (**18.2** vs. 13.1 pp)

2. Surprisingly, the **marginal** $p_\theta(\cdot \mid \varnothing)$ is biased toward **female**
   - amplification **flips direction** at **63%** (>50%)
   - **debiasing** $p_\theta(\cdot \mid \varnothing)$ restores **50%**

**Guidance bias offers a lens to analyze bias in T2I diffusion models.**

# Takeaways

What does a realistic fairness setting require?

**Fairness must hold across user-selected scales**

How should we evaluate bias in T2I diffusion models?

**Guidance bias gives a deeper look beyond total bias**

What should future debiasing target?

**Jointly debias the model and the guidance**

# Thank you!

**Paper:** https://openreview.net/pdf?id=tIDWhFyqX7

**Project Page:** https://kim-myeong-soo.github.io/stayfair/