One Word to Stereotype: Prompt-Induced Bias in Text-to-Image Generation

Magda Costa
Faculty of Sciences of the University of Porto & Faculty of Engineering of the University of Porto

(Dated: January 6, 2026)

We audit representational gender bias in a text-to-image model by testing whether neutral occupation prompts match real-world workforce gender distributions, and whether adding a single adjective changes the gender of generated subjects. Using Stable Diffusion v1.5, we generated images for six occupations under a baseline prompt and three adjective variants (*aggressive*, *friendly*, *successful*) and manually labeled outputs as woman, man, or ambiguous (excluded from the primary analysis). Baseline generations largely follow BLS workforce trends but often exaggerate gender skews, producing more polarized portrayals for several occupations. Adjective effects, measured as changes in female share relative to baseline, are not consistent across occupations and are concentrated in *loan officer*, where all tested adjectives shift portrayals toward men. These results show that even minimal prompt edits can measurably influence representational outcomes, motivating careful auditing and prompt design in applied use.

I. INTRODUCTION AND MOTIVATION

The field of text-to-image (T2I) models, such as Stable Diffusion and DALL-E, has gained significant attention for their ability to generate high-quality images from natural language descriptions[1]. Their increasing accessibility has enabled a wide range of creative and practical uses, like art generation and support in educational and professional contexts. However, as their use becomes more widespread, there is growing concern regarding the biases these models may perpetuate, particularly those related to gender and ethnicity[2,3].

This project is motivated by a broader interest in how prompt-to-image interaction shapes what these systems depict. Prior exposure to classroom use of image-generation tools highlighted how small changes in prompting can substantially alter outputs, suggesting the need for clearer guidance when using these tools. Further research and experimentation emphasized this need, particularly in addressing professional bias and gender representation. Building on this, the project audits gender bias in occupational portrayals by examining whether subtle prompt variations, such as the inclusion of different adjectives, shift the perceived gender representation of generated images. The presence of such biases is especially concerning when these models are employed in educational or professional settings, as they risk reinforcing harmful societal stereotypes. Moreover, the field of AI ethics emphasises the importance of fairness and diversity in model outputs, underscoring the relevance and necessity of this project in advancing the responsible use of AI tools in society[4,5].

To ground this evaluation, model outputs will be compared against external workforce statistics from the U.S. Bureau of Labor Statistics (BLS), which reports gender proportions across occupations and is widely used in bias benchmarking for occupational prompts[1,2,6]. This provides a concrete reference point for assessing whether T2I models reproduce or exaggerate real-world skews, and whether prompt wording can meaningfully change those distributions.

Ultimately, the goal of this project is to provide insight into the ethical considerations of AI tools and to evaluate how gender biases manifest in T2I models like Stable Diffusion. By comparing the model's outputs to real-world data, this study will contribute to the broader conversation on how these models influence societal representations. Additionally, by testing how adjectives affect gender representation, we will better understand the nuances in how these models respond to changes in prompt structure. As AI-generated images become increasingly integrated into various sectors, it is important to understand and critically assess these biases to ensure that these tools are used responsibly and do not perpetuate harmful stereotypes.

II. BACKGROUND AND RELATED WORK

A. Text-to-Image Models

T2I generation has advanced rapidly in recent years, enabled by powerful machine learning models that can generate high-quality images from natural language descriptions[7]. These models have become widely accessible and are used in various domains such as creative arts, advertising, and even training data for other machine learning models[7]. Notable models in this field include Stable Diffusion, DALL·E, and Imagen, which employ different underlying architectures, such as diffusion models and transformers, to produce images that closely match the given textual prompts[7,8].

One of the primary breakthroughs in T2I generation has been the advent of diffusion models. These models generate images iteratively, progressively reducing noise until the final image is produced. The model's ability to generate diverse outputs stems from its capacity to explore a large latent space, enabling the synthesis of a wide range of images from a single prompt[8].

Stable Diffusion, for instance, is a latent diffusion model that operates in a compressed latent space, significantly reducing computational complexity without compromising on image fidelity. This model, alongside others, has demonstrated state-of-the-art performance in creating realistic images, often with remarkable accuracy to the details specified in text prompts. The model's ability to generate high-resolution images from text has made it a valuable tool for both professional and recreational users.[8][7]

B. Prompt Engineering for T2I Generation

To produce a desired image with T2I tools, it might be necessary to refine the prompt iteratively, given the open-ended nature of text prompts and stochastic generation. Users frequently refine wording through repeated trial and error, which can feel "random and unprincipled" without guidance. This motivates treating prompt writing as a form of prompt engineering adapted to visual generation[7]. In this sense, prompt engineering acts as a practical "control layer" that can influence both output quality and content, including style, subject clarity, and consistency across variations[7].

It involves exploring prompt wording and generation settings, such as random initialisation and optimisation length, which can materially affect outputs even when the "idea" of the prompt remains the same[7]. Prompt engineering is also relevant for harder prompts: diffusion models can perform well on simple text but may become confused by complex descriptions, including prompts with multiple objects or relationships[9]. To address this, prompt-learning approaches aim to learn improved, input-specific prompts that better align generated images with the text, rather than relying only on manual rewording[9].

C. Ethical Concerns: Representational Bias and Stereotype Amplification

T2I models have made significant advancements in recent years, but they also raise ethical concerns due to the potential for these models to perpetuate harmful social biases[4]. Biases present in the training data can be unintentionally amplified when models generate images. Even with neutral prompts (e.g., specifying an occupation without mentioning gender or race), models can still produce heavily skewed portrayals. In the context of Stable Diffusion, occupation prompts have been shown to lead to "near-total stereotype amplification," exaggerating existing demographic skews rather than moderating them[1].

This amplification is especially visible in gendered occupational stereotypes. Models often depict women in nurturing or caregiving roles and men in technical or leadership positions[2]. As a result, jobs like nurses are more frequently generated as female, while engineers and CEOs are more often generated as male. This kind of occupational bias can reinforce traditional gender roles, influencing perceptions of who is "appropriate" for certain careers and contributing to unequal opportunities in the real world[3].

Bias also affects racial and ethnic representation: outputs can skew towards Western or 'default' norms unless diversity is explicitly prompted, marginalising non-Western identities and narrowing who is represented as typical or authoritative[12]. In applications like advertising, professional media, or public information, these defaults risk presenting whiteness as the "standard," reinforcing exclusionary ideas.

D. Bias Measurement and Mitigation

A common strategy for measuring occupational bias in T2I systems is to benchmark model outputs against real-world workforce statistics, particularly those reported by the U.S. Bureau of Labor Statistics (BLS). One approach is to generate images using occupation prompts (e.g., "A photo of the face of [OCCUPATION]") and then compare the demographic distribution of generated outputs to BLS workforce distributions, which can reveal cases where relatively modest real-world skews correspond to much more extreme imbalances in generated images[1]. In this sense, BLS statistics operate as an external reference point for quantifying occupational representational skew, making it possible to move beyond anecdotal examples and towards reproducible measurement.

In parallel, some evaluation frameworks aim to capture bias in a way that is less dependent on a fixed benchmark set and more sensitive to the prompt itself. Because biases can vary across prompts, TIBET proposes a counterfactual evaluation pipeline that identifies prompt-relevant bias axes, generates counterfactual prompts, and quantifies differences using metrics such as Concept Association Score (CAS) and Mean Absolute Deviation (MAD)[4]. This complements BLS-based benchmarking: while BLS enables grounded comparison for occupations with available labour data, prompt-aware tools such as TIBET help reveal bias patterns that may not be captured by occupational statistics alone.

On mitigation, downstream interventions like fairness-aware prompting are appealing because they do not require retraining or access to proprietary model weights. However, evidence suggests that results are inconsistent and model-dependent: controlled prompts can shift demographic portrayals substantially, but effects range from meaningful diversification to overcorrection into unrealistic uniformity, or even minimal responsiveness depending on the system[3]. This supports treating prompt-based mitigation as useful for probing and auditing model behaviour, but not as a complete solution by itself[3].

III. METHODOLOGY AND EXPERIMENTAL SETUP

A. Study Design and Goal

This project audits representational gender bias in a T2I system by generating synthetic images for occupation prompts under tightly controlled prompt edits. The audit addresses: Baseline vs. reality → ”Do gender proportions in neutral occupation generations align with real-world workforce distributions?”, and Linguistic variation → ”Does adding a single adjective systematically shift gender representation relative to the neutral baseline?”

This design follows prior audits that quantify demographic stereotype amplification in T2I outputs and benchmark occupation-related gender patterns against labor statistics.[1,2,6]

B. Model Selection

This audit focuses on Stable Diffusion v1.5, an open-weight text-to-image model that can be executed locally. This choice supports the project goals in two ways: it enables a fully reproducible workflow (full control over prompts, parameters, and seeds), and avoids paywalled or restricted APIs, keeping the experimental setup cost-free in practice for local inference. In this project, the checkpoint was obtained from Hugging Face[10] and run locally via AUTOMATIC1111[11] as the inference interface.

Architecturally, Stable Diffusion is based on Latent Diffusion Models (LDMs), which run the diffusion process in a compressed latent space learned by an autoencoder, substantially reducing computation compared to pixel-space diffusion while preserving image quality. This efficiency is a practical fit for local experimentation and supports generating hundreds of images under controlled prompt conditions.[8,10,12]

Stable Diffusion v1 models were trained on subsets of LAION–2B(en), i.e., predominantly English-captioned image–text pairs, and v1.5 is described as resumed from earlier v1 checkpoints and trained 595,000 steps at 512 × 512 on “laion-aesthetics v2 5+”. This is relevant to the audit because the model card explicitly warns that the system can reinforce social biases, noting that “white and western cultures are often set as the default”.[10] Finally, using Stable Diffusion 1.5 improves comparability with prior bias-audit work that explicitly evaluates occupational prompts using this exact version[4].

C. Prompt set Construction and Controlled Conditions

Prompting follows a minimal-edit design: the template is held constant and only one lexical element (the adjective) changes. This strengthens attribution of distribution shifts to the linguistic change rather than broader prompt restructuring.[1]

To reduce multi-person outputs and improve labelability, prompts explicitly request a single person, with a visible/centered face and a realistic “photo” framing, consistent with bias-evaluation prompt practices (template-based prompts such as “a photo of [DESCRIPTION]”, including professions/adjectives).[1,2,3,7]

The template prompts used in this project are the following:

- Baseline: “A photo of a person working as/a OCCUPATION. Depict a single person. Face visible, centered, realistic photo.”
- Variants: “A photo of a ADJECTIVE person working as/a OCCUPATION. Depict a single person. Face visible, centered, realistic photo.”

We keep wording stable (rather than many rephrasings) because prompt-engineering evidence emphasizes that outcomes depend mainly on core keywords (subject/style) and rephrasings with the same keywords do not reliably change results.[7]

Occupations were selected to cover male-dominated, female-dominated, and near-neutral distributions using BLS CPS Annual Averages, Table 11 (“percent women”). This table is used because its occupation labels map well onto natural-language occupation prompts.

Occupation	Nurses (Registered)	Electrician	Dental Hygienist	Software Developer	Retail Salesperson	Loan Officer
Women (%)	86.8%	1.2%	93.9%	20.3%	47.5%	52.8%

TABLE I: Share of women by occupation

Regarding the adjectives that will be used in the variants, we will test how simple social traits can affect the distribution of the baseline. Therefore, we will use: ”Successful,” ”Friendly,” and ”Aggressive.”

D. Image Generation Protocol and Experimental Configuration

For each occupation × condition (baseline plus each adjective variant), we generated 120 images (100 initial generations plus 20 additional generations), saving outputs together with the full generation metadata (prompt, seed, and parameters). This sample size balances feasibility with statistical stability: prior large-scale audits show that small samples can introduce substantial measurement error. To isolate adjective effects, generation settings are held constant and only the adjective token changes across paired prompts.[2,5]

Because the experiments were run locally on a machine with 4GB of VRAM, generation settings were chosen to remain computationally feasible while keeping all conditions comparable: Model→ v1-5-pruned-emaonly.safetensors (Stable Diffusion v1.5), Sampler → Euler a, Steps → 15, CFG → 6, Resolution → 512 × 512, Negative Prompt → "blurry, out of focus, lowres, bad anatomy, deformed, cartoon, illustration, anime, painting, 3d render, cropped, out of frame, multiple people"

E. Annotation and Dataset Cleaning

All generated images are manually labeled for perceived gender presentation using three categories: female / male / ambiguous. "Ambiguous" includes unclear gender perception, non-human depictions, or conflicting cues. Ambiguous outputs are not used in the primary gender-share estimates. Instead, we compute metrics on the confidently labeled subset only. Concretely, for each occupation × condition we define the effective sample size as $n_{\text{clear}} = \#W + \#M = 120 - \#A$, and we report $\#A$ (and thus n_{clear}) to document labelability and data quality variation across prompts.[26]

F. Metrics and Analysis

For each occupation × condition, we compute the female share (excluding ambigouos):

$$R_f = \frac{\#(\text{female-labeled})}{\#(\text{female-labeled}) + \#(\text{male-labeled})}$$

Baseline R_f is compared against BLS CPS Annual Averages, Table 11 ("percent women") as an external reference for occupational gender distributions.[16]

For each adjective variant, we compute the shift:

$$\Delta R_f = R_f(\text{adjective}) - R_f(\text{baseline})$$

Using paired seeds supports interpreting ΔR_f as an adjective-driven change rather than random variation.[45]

G. Reproducibility

To support reproducibility, we store generation metadata (prompt, seed, parameters) for every image and release the dataset plus analysis code on GitHub[13]. This follows the transparency norm in bias-audit work that publishes data/code to enable verification and reuse.

IV. RESULTS AND DISCUSSION

A. Dataset Labelability and Ambiguaty

120 images were generated per prompt condition (100 initial generations plus 20 additional generations) for each occupation–condition pair, covering 6 occupations and 4 conditions (baseline plus three adjective prompts), for a total of 2,880 images. Each image was annotated into one of three categories: man (M), woman (W), or ambiguous (A) when gender presentation could not be labeled confidently. In the primary analysis, ambiguous images were excluded, resulting in an effective sample size per cell of $n_{clear} = M + W = 120 - A$.

Across all cells, 2,333 images (81.0%) were confidently labeled (usable), while 547 images (19.0%) were labeled ambiguous and excluded. Effective sample sizes varied substantially across occupation–condition combinations, with n_{clear} ranging from 64 to 117 (median: 102). Ambiguity was not uniformly distributed: the highest ambiguity rates occurred for loan officer (39.2% ambiguous across conditions) and electrician (27.1%), whereas dental hygienist (6.3%) and nurse (8.8%) showed relatively low ambiguity. At the condition level, ambiguity was highest for the baseline (23.5%) and aggressive prompts (22.6%) and lowest for friendly (13.9%). Because effective sample sizes differ across cells, subsequent analyses report n_{clear} and quantify uncertainty (e.g., via confidence intervals), and we include a sensitivity check to verify that conclusions are not driven by the treatment of ambiguous cases.

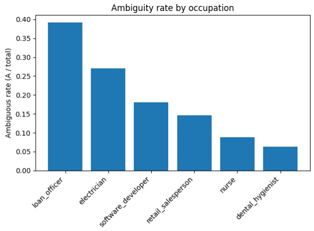

FIG. 1: Ambiguity rate per occupation

B. Baseline VS Real-world Workforce (BLS)

To assess whether neutral occupation prompts reproduce or amplify real-world gender skews, we compared the female share of generated baseline images, R_f^{base} (computed over confidently labeled images only, excluding ambiguous), against U.S. Bureau of Labor Statistics (BLS) workforce proportions ("percent women", CPS Annual Averages, Table 11). For each occupation, $R_f^{\text{base}} = W/(W + M)$, and we report 95% Wilson confidence intervals to reflect uncertainty given the effective sample size $n_{clear} = W + M$. This comparison is visualized in Fig. 2, where points above (below) the diagonal indicate that baseline generations depict women more (less) frequently than the BLS reference.

Across the six occupations, baseline generations strongly preserved the overall ordering of occupations by female share (Pearson correlation $r \approx 0.98$), but exhibited meaningful deviations from the BLS reference (mean absolute error ≈ 7.8 percentage points). As shown in Fig. 3, the largest discrepancies occurred in occupations that are already gender-skewed in the workforce: nurse and dental hygienist were generated as 100% women (vs.

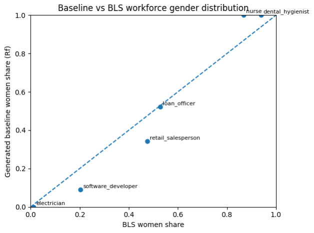

FIG. 2: Baseline vs BLS workforce female share by occupation. Each point shows the generated baseline female share R_f^{base} (excluding ambiguous labels) versus BLS "percent women"; the dashed line indicates perfect agreement.

BLS 86.8% and 93.9%, respectively), indicating a shift toward the extreme. In contrast, software developer and retail salesperson showed underrepresentation of women relative to BLS (≈ 9.0% vs. 20.3%, and 34.3% vs. 47.5%, respectively), while electrician remained near-zero female representation (0% vs. 1.2%). Loan officer was the closest match to BLS (≈ 52.3% vs. 52.8%). Overall, these results suggest that neutral occupation prompting can retain real-world trends while still pushing portrayals toward more polarized gender representations, consistent with stereotype amplification effects reported in prior text-to-image bias audits.

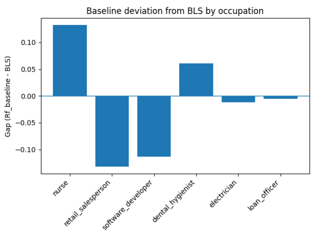

FIG. 3: Deviation from BLS by occupation, computed as $R_f^{\text{base}} - R_f^{\text{BLS}}$. Positive values indicate over-representation of women in baseline generations.

C. Adjectives Effect (Baseline VS Adjectives)

To quantify how descriptive adjectives shift gender representation, we compare each adjective prompt to the baseline prompt within the same occupation using the change in female share:

$$\Delta R_f = R_f^{\text{adj}} - R_f^{\text{base}}, \qquad R_f = \frac{\#W}{\#W + \#M}.$$

To preserve the paired experimental design, we compute ΔR_f using a complete-case paired subset: for each occupation–adjective pair, we match baseline and adjective generations by seed, and retain only seeds where both images are confidently labeled (W/M), excluding ambiguous images in either condition. We report 95% confidence intervals for ΔR_f via a paired bootstrap over seeds.

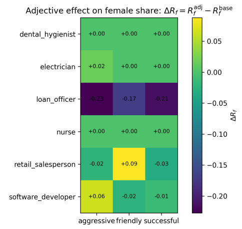

FIG. 4: Adjective effect on gender representation by occupation. Each cell shows $\Delta R_f = R_f^{\text{adj}} - R_f^{\text{base}}$ computed on the complete-case paired subset (ambiguous excluded). Positive values indicate an increase in the share of women; negative values indicate a decrease.

Figure 4 summarizes ΔR_f across occupations and adjectives. The dominant effect is concentrated in loan officer, where all three adjectives shift portrayals toward men: aggressive yields $\Delta R_f = -0.23$ (95% CI $[-0.36, -0.11]$), friendly yields $\Delta R_f = -0.17$ (95% CI $[-0.29, -0.05]$), and successful yields $\Delta R_f = -0.21$ (95% CI $[-0.34, -0.11]$). These are the only occupation–adjective effects whose confidence intervals exclude zero, indicating a consistent and statistically detectable shift in gender representation for this occupation. For several other occupations, adjective effects are smaller and typically overlap zero; for example, retail salesperson shows a positive shift for friendly ($\Delta R_f \approx +0.09$) but with wide uncertainty, while software developer exhibits a modest increase under aggressive ($\Delta R_f \approx +0.06$) that is not statistically distinguishable from zero given the available sample size. Finally, occupations already at extreme baseline representations (e.g., nurse and dental hygienist generated as all women in the clear-labeled subset) show $\Delta R_f \approx 0$ across adjectives due to a ceiling effect.

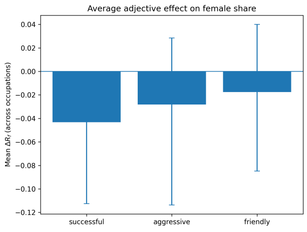

FIG. 5: Mean adjective effect across occupations. Bars show the average ΔR_f per adjective; error bars indicate 95% bootstrap confidence intervals.

Aggregating across occupations, Fig. 5 reports the mean ΔR_f per adjective. On average, adjective prompts tend to slightly reduce the female share relative to baseline, with the largest average decrease for successful (mean $\Delta R_f \approx -.0.4$, bootstrap CI excluding zero by a small margin), though this aggregate result is strongly influenced by the large negative shifts observed for loan officer. Overall, the results indicate that adjective modifiers can change gender portrayals, but the effect is not uniform across occupations; instead, it appears concentrated in specific occupation–adjective combinations.

D. Robustness Checks

Because a non-trivial fraction of generations were labeled ambiguous, our primary adjective-effect estimates use a complete-case paired analysis (baseline and adjective matched by seed, retaining only seeds with confident W/M labels in both conditions). To test whether conclusions depend on excluding ambiguous samples, we reran the paired analysis under a soft-label sensitivity scheme where ambiguous images contribute 0.5 to the female indicator (i.e., $A \mapsto 0.5$). Figure 6 compares effect sizes ΔR_f under the primary and soft-label analyses. The main qualitative conclusion remains unchanged: for *loan officer*, all three adjectives consistently reduce the share of women (primary: *aggressive* $\Delta R_f = -0.23$, *friendly* $\Delta R_f = -0.17$, *successful* $\Delta R_f = -0.21$; soft-label: -0.15, -0.15, -0.15), with confidence intervals excluding zero under both schemes. In addition, the soft-label analysis reveals several smaller negative shifts (e.g., *electrician* under *friendly*/*successful* and *software developer* under *friendly*) that are not detectable under the complete-case subset because the clear-labeled images in those cells are overwhelmingly male. Overall, robustness checks indicate that the strongest occupation–adjective effects (notably for *loan officer*) are stable to ambiguity handling, while some weaker effects are sensitive to how ambiguous samples are treated.

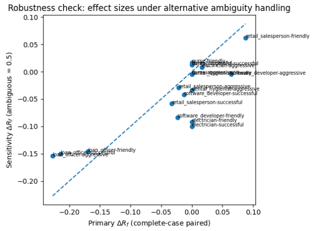

FIG. 6: Robustness check for adjective effects under alternative ambiguity handling. Each point is an occupation–adjective pair, comparing the primary complete-case paired estimate of ΔR_f (ambiguous excluded) to a sensitivity estimate where ambiguous labels are coded as 0.5. The dashed line indicates equality.

V. CONCLUSIONS AND FUTURE WORK

Neutral occupation prompts can reproduce real-world gender patterns while also amplifying them toward more polarized portrayals relative to BLS statistics. Moreover, one-word adjective edits can measurably shift gender representation, but the effect depends strongly on the occupation. In our data, the clearest and most stable shifts occur for *loan officer*. Future work should expand occupations/adjectives, compare across models/checkpoints, and strengthen annotation (multiple raters and clearer treatment of ambiguity). Finally, prompt-based mitigations (e.g., fairness-aware prompting or explicit diversity constraints) should be evaluated to reduce amplification without collapsing outputs into unrealistic uniformity.

[1] F. Bianchi, P. Kalluri, E. Durmus, F. Ladhak, M. Cheng, D. Nozza, T. Hashimoto, D. Jurafsky, J. Zou, and A. Caliskan, in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23 (Association for Computing Machinery, New York, NY, USA, 2023) p. 1493–1504.

[2] L. Girrbach, S. Alaniz, G. Smith, and Z. Akata, "A large scale analysis of gender biases in text-to-image generative models," (2025).

[3] S. Raza, M. Powers, P. P. Saha, M. Raza, and R. Qureshi, "Prompting away stereotypes? evaluating bias in text-to-image models for occupations," (2025), arXiv:2509.00849 [cs.CL].

[4] A. Chinchure, P. Shukla, G. Bhatt, K. Salij, K. Hosanagar, L. Sigal, and M. Turk, "Tibet: Identifying and evaluating biases in text-to-image generative models," (2024).

[5] Y. Wu, Y. Nakashima, and N. Garcia, Journal of Imaging **11** (2025), 10.3390/jimaging11020035.

[6] S. Luccioni, C. Akiki, M. Mitchell, and Y. Jernite, in *Advances in Neural Information Processing Systems*, Vol. 36, edited by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Curran Associates, Inc., 2023) pp. 56338–56351.

[7] V. Liu and L. B. Chilton, "Design guidelines for prompt engineering text-to-image generative models," (2023), arXiv:2109.06977 [cs.HC].

[8] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," (2022), arXiv:2112.10752 [cs.CV].

[9] C. Yu, J. Peng, X. Zhu, Z. Zhang, Q. Tian, and Z. Lei, "Seek for incantations: Towards accurate text-to-image diffusion synthesis through prompt engineering," (2024), arXiv:2401.06345 [cs.CV].

[10] "Stable diffusion v1-5," Hugging Face Hub.

[11] AUTOMATIC1111, "stable-diffusion-webui," GitHub repository.

[12] CompVis, "stable-diffusion: A latent text-to-image diffusion model," GitHub repository.

[13] Maguids, "Stable diffusion gender bias and adjective's impact on occupations," .