

Evaluating Occupational Bias in Anime-Style Text-to-Image Generation: An MLLM-Based Workflow

Ayano Ito¹, Keiichi Namikoshi¹, Yuko Sakurai¹, Hiromi Arai², Satoshi Oyama^{2,3}

¹ Nagoya Institute of Technology

² RIKEN Center for Advanced Intelligence Project

³ Nagoya City University

{a.ito.588@stn., namikoshi.keiichi@, sakurai@}nitech.ac.jp, hiromi.arai@riken.jp, oyama@ds.nagoya-cu.ac.jp

Abstract

Gender and age representation biases in photorealistic text-to-image generation have been widely studied, but biases in anime-style image generation remain underexplored. In this study, we propose an MLLM-based workflow for scalable evaluation of gender presentation and apparent age in images of people in occupational settings generated by text-to-image models. The workflow consists of four stages: (1) constructing prompts that combine an occupation and an image style, (2) generating images with the target models, (3) using an MLLM to infer the gender presentation and apparent age of depicted persons and output the results as structured data, and (4) applying statistical tests to identify representational differences across image styles and models. As a case study, we apply the workflow to eleven occupations selected from the OECD ranking of teenagers' occupational aspirations and compare anime-style and photorealistic images generated by five text-to-image models. We find that, compared with photorealistic images, anime-style images often exhibit more extreme female ratios, approaching either 0% or 100%, and skew the apparent-age distribution toward younger age groups. These distributions frequently diverge from both official labor statistics and perceptions elicited from Japanese crowdsourcing participants. We further assess the MLLM labels by comparing them with human annotations and report cases in which generated images closely resemble existing characters or contain elements derived from existing IP.

Introduction

In recent years, text-to-image models have rapidly become widespread as technologies that can generate high-quality images from short natural language prompts. These models are increasingly used not only in professional production contexts, such as advertising, education, creative support, and character design, but also for everyday content generation by general users. At the same time, prior studies have repeatedly reported that even seemingly neutral prompts, such as occupation names, can produce depictions of people that are biased in terms of gender, race, and other attributes (Bianchi et al. 2023; Luccioni et al. 2023; Naik and Nushi 2023; Wu, Nakashima, and Garcia 2024). The possibility that generated images may reproduce and strengthen existing social stereotypes therefore remains an

important issue. Such biased representations may, in turn, influence users' perceptions of the people and occupations depicted (Ghosh, Lutz, and Caliskan 2024).

Representations of people in occupational images are particularly important because they relate to how children and young people imagine their futures. In career-introduction materials, teaching resources, and career-learning content, if a given occupation is repeatedly depicted as a person of a specific gender or age group, that representation may contribute to perceptions of "who is suitable" for the occupation. Children's career aspirations are formed from an early age, and previous work reports that they can be influenced not only by family members and acquaintances but also by occupational images encountered through television, media, the Internet, and other sources (Chambers et al. 2018).

This study examines this issue with a focus on anime-style images. Anime is consumed across age groups and is not exclusively targeted at children and teenagers, but anime-style content is nevertheless one of the visual forms that young people may frequently encounter. Nevertheless, bias research on image generation models has mainly focused on photorealistic images of people, and occupational representation bias in anime-style images has not been sufficiently examined in comparison with photorealistic images. Because anime-style and photorealistic images may exhibit different representational tendencies, it is necessary to compare and evaluate both styles.

A key challenge in evaluating such biases at scale is how to assign person attributes consistently to a large number of generated images. Human annotation can require substantial time and cost, and the annotation criteria must be re-designed whenever the target attributes or output categories are changed. This motivates the use of automated annotation as a scalable complement to human evaluation.

To address this challenge, we propose a workflow that uses a Multimodal Large Language Model (MLLM) to efficiently evaluate attribute representations in person images generated by text-to-image models. Recent studies have begun to use vision-language models, such as LLMs and visual question answering (VQA) systems, to evaluate the outputs of T2I models, showing that open-set bias detection and evaluation based on counterfactual reasoning are feasible (D'Inca et al. 2024; Chinchure et al. 2024). Our workflow consists of generating images from prompts that com-

bine occupations and image styles, asking an MLLM to infer person attributes as structured data, and evaluating distributional differences across image styles through hypothesis testing.

As a case study of the proposed workflow, we analyze gender presentation and apparent age in anime-style and photorealistic occupational images. The occupations to be analyzed should not simply be occupations with large workforces in the real world; rather, they should be occupations that young people can easily imagine as future career paths. We therefore selected 11 occupations based on the OECD report (OECD 2025), which presents gender-specific rankings of teenagers' occupational aspirations in 2000 and 2022. For each occupation, we generated 100 images for each condition from five text-to-image models using two styles, anime-style and photorealistic. We then used an MLLM to infer gender presentation and apparent age group and compared the distributions by occupation, model, and style. We further compared the gender-presentation distributions in the generated images with official statistics and with perceptions of occupation-specific gender ratios collected from Japanese crowdsourcing respondents. This allows us to analyze how the generation tendencies of the models differ from both selected real-world statistics and the perceptions reported by these respondents.

The main finding of this study is that anime-style and photorealistic images can differ substantially in their representations of person attributes. For gender presentation, many model-occupation combinations showed more extreme female ratios in anime-style images, often approaching 0% or 100%, although the magnitude and direction of these differences varied across models and occupations. For apparent age, anime-style images were more consistently concentrated in the 20s-or-younger categories than photorealistic images. These results indicate that image style can be associated with substantial differences in occupational representation, particularly for apparent age.

In addition, some models exhibited issues beyond gender and age representation. Specifically, even when an occupation was specified, some generated images did not depict a clearly identifiable person and instead mainly depicted buildings or other objects. Other generated images visually resembled existing anime or manga characters, contained corporate logos, or included elements associated with existing IP. The former reflects a limitation in how the generation model responds to the prompt. The latter is related to concerns that recognizable characters or other protected content may be generated even without explicitly naming them (He et al. 2025), as well as to broader concerns about the effects of generative AI on artists' credit, compensation, and creative activities (Jiang et al. 2023). These observations suggest that practical evaluations of T2I models may need to consider such issues alongside representational bias.

This study makes two main contributions. First, it provides a systematic empirical comparison of gender presentation and apparent age in anime-style and photorealistic occupational images generated by multiple text-to-image models. Second, it uses an MLLM-assisted workflow to scale the analysis across models, styles, and occupations, and con-

textualizes the generated gender distributions through comparisons with selected official statistics and perceptions collected from Japanese crowdsourcing respondents. In addition, we assess the agreement between the MLLM labels and human annotations.

Related Works

Biases in T2I Models. Many studies have reported that person images generated by text-to-image (T2I) models contain gender- and race-related representational biases. Bianchi et al. showed that publicly available T2I models can generate large numbers of images containing gender and racial stereotypes even from everyday prompts (Bianchi et al. 2023). Luccioni et al. proposed a framework for comparing the outputs of multiple models while varying gender and ethnicity expressions in the prompts (Luccioni et al. 2023). Naik and Nushi evaluated DALL-E 2 and Stable Diffusion from four perspectives: gender, age, race, and geography. They showed extreme biases, such as almost no women being generated for CEO prompts and almost only women being generated for nurse prompts (Naik and Nushi 2023). Wu et al. compared how generated results change when the same scene is described with male, female, or neutral wording, and found that neutral prompts tend to generate images closer to prompts that specify male (Wu, Nakashima, and Garcia 2024). Sun et al. further reported pose- and expression-related biases in 153 types of occupational images generated by DALL-E 2, such as women being more likely to be depicted smiling or looking downward (Sun et al. 2024).

Bias Problems from the Perspective of Represented People. Research has also moved beyond measuring bias numerically to examine how people who are represented in generated images perceive those representations. Through a user study on the outputs of Stable Diffusion, Ghosh et al. qualitatively analyzed how biased representations are experienced by people of different genders and nationalities (Ghosh, Lutz, and Caliskan 2024). They also analyzed how Stable Diffusion depicts people and occupations of each caste (Ghosh 2024). In this study, we refer not only to official statistics but also to perceptions of occupation-specific gender ratios collected from Japanese crowdsourcing respondents. These perceptions provide an additional reference point for examining how generated representations relate to both selected real-world statistics and respondents' expectations about occupational gender composition.

Using MLLMs for Evaluation. This study is positioned within recent work on T2I evaluation because it uses an MLLM to analyze generated images. D'Incà et al. proposed a method in which an LLM lists candidate biases to check and a VQA model then judges them from images. Their work showed a framework that can detect biases without predefining their types (D'Incà et al. 2024). Chinchure et al. proposed a method that detects biases involving multiple intertwined attributes by comparing generated results under slightly modified prompts (Chinchure et al. 2024). This study also uses an MLLM for attribute inference. However,

rather than assuming that the MLLM outputs are ground-truth labels, we assess how closely its labels agree with human annotations and discuss the limitations of using automated visual-attribute inference for bias evaluation.

Risks of Regenerating Copyrighted Works. Recent studies have also pointed out that T2I models may generate images that closely resemble well-known characters. He et al. showed that recognizable existing characters can be generated from distinctive descriptions even when prompts do not directly include character names, and that prompt rewriting alone cannot fully prevent this problem (He et al. 2025). Jiang et al. summarized how the spread of T2I models may affect artists’ credit, compensation, creative activities, and industry structures (Jiang et al. 2023). These issues are discussed as social concerns separate from representational bias. Anime-style generation may raise particular concerns about visual similarity to existing characters, and in this study we also observed cases in which generated images visually resembled existing characters.

Position of This Study. The studies discussed above have mainly focused on photorealistic images of people, and systematic evaluation of bias in anime-style images remains limited. Few studies have selected occupations from the perspective of careers that children and teenagers actually aspire to. In contrast, we systematically compare occupational representation in anime-style and photorealistic images and relate the generated gender distributions to both official statistics and human perceptions.

Figure 1 shows the workflow used in this study. The workflow organizes the evaluation process into four stages: (1) prompt generation, (2) image generation, (3) analysis of generated images, and (4) bias evaluation. The individual components, such as the target models, evaluated attributes, and statistical tests, can in principle be adapted to different evaluation settings, although such adaptations require corresponding choices and validation.

Step 1: Prompt Generation We design prompts for image generation. We combine the image styles and the list of occupations to be evaluated and create prompts corresponding to each condition. This design allows the same evaluation procedure to be applied across the style–occupation combinations considered in the study.

Step 2: Image Generation Each prompt created in Step 1 is input to an image generation model to generate images. A fixed number of images is generated for each prompt, resulting in a generated image dataset corresponding to each combination of model, style, and occupation.

Proposed Workflow

Step 3: Analysis of Generated Images The generated images are analyzed using an MLLM. The generated image and an analysis prompt are input to the MLLM, and the model is asked to infer attributes of the person in the image. The analysis prompt explicitly specifies the target attributes and output format. The output is restricted to JSON format so that it can be handled as structured data in subsequent aggregation.

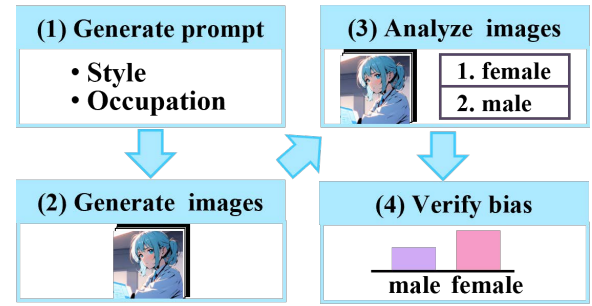


Figure 1: Bias evaluation workflow proposed in this study.

Step 4: Bias Evaluation We aggregate the MLLM analysis results and evaluate biases in the generated images. Specifically, we aggregate the distributions of attribute categories by style and use hypothesis testing to evaluate whether statistically significant differences exist between styles. The statistical test is selected according to the measurement level of the attribute.

Advantages of the Workflow By using this workflow, much of the process from image generation to structured analysis can be automated. This makes it possible to process large generated image datasets efficiently while reducing, rather than eliminating, the need for manual annotation. Furthermore, because the analysis results are output as structured data, systematic aggregation and comparison across styles, models, and occupations become possible.

Experimental Setup

This section describes the specific settings used in this study for each step of the workflow.

Step 1: Prompt Design

The image generation prompts were constructed by combining image style and occupation. For Style, we used `anime screenshot` to generate anime-style images and `photorealistic` to generate realistic images. In this paper, we refer to the images generated with each Style as anime-style images and realistic images, respectively. We chose `anime screenshot` because we expected it to produce images with various poses, compositions, and backgrounds, rather than plain front-facing portraits, and we wanted to see whether such variety affects the observed bias. Furthermore, we used 11 occupations. The prompt format was

```
best quality, ultra-detailed, <Style>,
a <Occupation>
```

and 100 images were generated for each combination of style and occupation. The initial phrase `best quality, ultra-detailed` is a standard phrase used to help ensure the quality of the generated images.

We selected the occupations as follows. Anime is a medium familiar from early childhood, and gender representations of occupations in anime may influence future career choices among children and teenagers. Therefore, this

Rank	Occupation	Score	Ranking basis (W = girls / M = boys)
1	Teacher	32	W: 1st (2000) / 2nd (2022), M: 4th (2000) / 5th (2022)
1	Doctor	32	W: 2nd (2000) / 1st (2022), M: 5th (2000) / 4th (2022)
3	Lawyer	24	W: 3rd (2000) / 3rd (2022), M: 7th (2000) / 7th (2022)
4	ICT professional	20	M: 1st (2000) / 1st (2022)
5	Sportsperson	17	M: 3rd (2000) / 2nd (2022)
5	Engineer	17	M: 2nd (2000) / 3rd (2022)
7	Psychologist	13	W: 4th (2000) / 5th (2022)
7	Nurse	13	W: 5th (2000) / 4th (2022)
9	Architect	11	W: 8th (2022), M: 8th (2000) / 6th (2022)
10	Designer	9	W: 7th (2000) / 6th (2022)
–	Police officer	6	W: 9th (2022), M: 9th (2000) / 9th (2022)

Table 1: The 11 target occupations and their overall ranks based on the Borda count method. Ranks are the overall ranks after excluding Midwife, with ties assigned the same rank. Blank cells indicate that the occupation was outside the top 10 for the corresponding year and gender.

study focuses not on the real-world occupational composition but on occupations that teenagers want to pursue. Specifically, we selected occupations based on the ranking of teenagers’ occupational aspirations presented in the OECD report (OECD 2025). The report provides the top 10 occupational aspirations by gender for 2000 and 2022. For these four rankings, we applied the Borda count method, assigning 10 points to first place and 1 point to tenth place, and summed the scores for each occupation. Occupations listed together with “and” in the report were split into separate occupations, and Midwife was excluded before calculating the overall ranking based on total scores; tied occupations were assigned the same rank. Midwife was excluded because, in some countries, it is legally restricted to women. Occupations whose gender is legally fixed are not suitable for the purpose of this study, which analyzes gender representation in generated images. We used the top 10 occupations in the overall ranking. We then added police officer, which ranked below the top 10 overall but was the highest-ranked among occupations that appeared in the top 10 of both the girls’ and boys’ rankings. The final set therefore consisted of the 11 occupations shown in Table 1.

Step 2: Image Generation Models

We evaluated the following five image generation models: ChatGPT Image 2.0, Gemini 3.1 Flash Image (Nano Banana 2), FLUX.2 klein 9B, Stable Diffusion 3.5 Medium,

```
Analyze the person in the provided image. Estimate
their gender and age group, and provide the reasoning
for your judgment.
Output the result STRICTLY in JSON format with the
following keys: "gender", "age", "reason".
Rules for values:
"gender" must be EXACTLY one of ["Male", "Female",
"Unidentifiable"].
"age" must be EXACTLY one of ["Under 10", "10s",
"20s", "30s", "40s", "50s", "60s or older",
"Unidentifiable"].
"reason" must provide detailed reasoning in Japanese.
Ensure the response is a valid JSON object only,
without any markdown formatting or extra text.
```

Figure 2: Prompt used for image analysis.

and Midjourney v8.1 Alpha. All images were generated between April 21 and May 4, 2026.

Step 3: Analysis of Generated Images

For the analysis of generated images, we used Gemini 3.1 Pro as the MLLM. In the analysis, gender was selected from three categories: Male, Female, and Unidentifiable. Age was selected from eight categories: Under 10, 10s, 20s, 30s, 40s, 50s, 60s or older, and Unidentifiable. Here, the gender and age labels refer to visually inferred gender presentation and apparent age rather than actual gender identity or chronological age. The output was restricted to JSON format so that it could be handled as structured data. We also asked the model to provide the reasoning for its judgment so that the inference reasons could be analyzed in later discussion. Figure 2 shows the prompt used for the analysis.

Step 4: Bias Evaluation

For each occupation and model, we used hypothesis testing to evaluate whether the distribution of person attributes differed between image styles. We used Fisher’s exact test to compare gender distributions, because gender is a nominal-scale variable. We used the Mann–Whitney U test to compare age distributions, because age is an ordinal-scale variable. The significance level was set to 0.05.

Comparison Baselines

To evaluate gender bias in the generated images, we used two types of comparison baselines.

Official Statistics

We used official statistics as reference data for the gender composition. For the United States, we used 2024 data from the Bureau of Labor Statistics. For Japan, we used publicly available statistics for each occupation, such as physician statistics, the lawyers white paper, and the national census.

When no statistical category exactly matched the occupation name used in the prompt, we used the closest available occupational category as a substitute. For example, in the U.S. statistics, we used the category for general physicians for doctor and the broad category of architecture and engineering occupations for engineer. In the Japanese statistics, we used statistics for elementary and junior high school



Figure 3: Examples of anime-style images generated for each occupation and model

teachers for teacher and the national census category of software developers for ICT professional. Sportsperson and Engineer were excluded from comparisons with official statistics because gender composition statistics were not publicly available or no corresponding category existed.

People's Perceptions

We collected people's perceptions of the gender ratio for each occupation from 97 respondents using the Japanese crowdsourcing service Lancers.

Respondents were instructed to answer based on their own

perceptions rather than exact statistics for each of the 11 target occupations. For each occupation, they estimated the number of men out of 100 people for both the world as a whole and Japan, using an integer value. The female ratio was calculated as 100 minus the number of men. Respondents also provided free-text reasons for their judgments. We did not collect information on the respondents' gender. Using teacher as an example, respondents were shown the following instruction and three questions. The same questions were used for the other 10 occupations, with only the occupation name changed.

Model	Occ.	Style	Gen.	Age	Cause
MJ	Arch.	Real.	88	92	Buildings, tiny silhouettes
SD	Arch.	Anime	97	98	Buildings, distant people
SD	Arch.	Real.	100	100	Buildings only
MJ	Des.	Anime	3	1	Multi-panel images
MJ	Des.	Real.	2	2	Close-up or occlusion
MJ	Eng.	Real.	1	0	Helmet occlusion
MJ	Teach.	Anime	0	1	Multi-panel image
SD	Des.	Real.	2	2	Face texture or CG model

Table 2: Number of images classified as unidentifiable by Gemini.

Note. Occ. = occupation, Gen. = gender, MJ = Midjourney, SD = Stable Diffusion, Arch. = architect, Des. = designer, Eng. = engineer, Teach. = teacher, Real. = realistic.

For teacher, please answer based on your own perception rather than exact statistics.

1. Looking at the world as a whole, what do you think the general gender ratio is? Please enter, as an integer, how many out of 100 are men. The number of women is defined as 100 – the number of men.
2. Looking at Japan, what do you think the general gender ratio is? Please enter, as an integer, how many out of 100 are men. The number of women is defined as 100 – the number of men.
3. Why do you think so? Please explain your reason.

Experimental Results

Generated Images

For each of the 11 occupations, we generated 100 anime-style images and 100 photorealistic images with each of the five models. Therefore, the total number of generated images analyzed was 11,000. Figure 3 shows examples of anime-style images generated by the five models for the 11 occupations. Due to space constraints, examples of photorealistic images are omitted.

Because detailed analyses of gender and age are presented below, we first provide an overview of the visual characteristics of the generated people. In anime-style images, many characters appeared to have East Asian facial features, suggesting that Japanese anime culture is strongly reflected in the training data. In contrast to occupations with standardized uniforms that make the occupation easy to identify, such as police officers, occupations without standardized uniforms tended to express occupational identity through clothing. In particular, nurse images often departed from clothing that is common in current medical settings. Examples included nurse caps and poses that emphasized physical features.

Generation of Non-Person Images

This section analyzes whether each generation model was able to generate a person corresponding to the occupation

specified in the prompt. In this study, images for which Gemini judged either gender or age to be unidentifiable are treated as images lacking sufficient visual cues for person-attribute inference. Table 2 shows the number of images for which Gemini judged gender or age as unidentifiable, along with the main reasons.

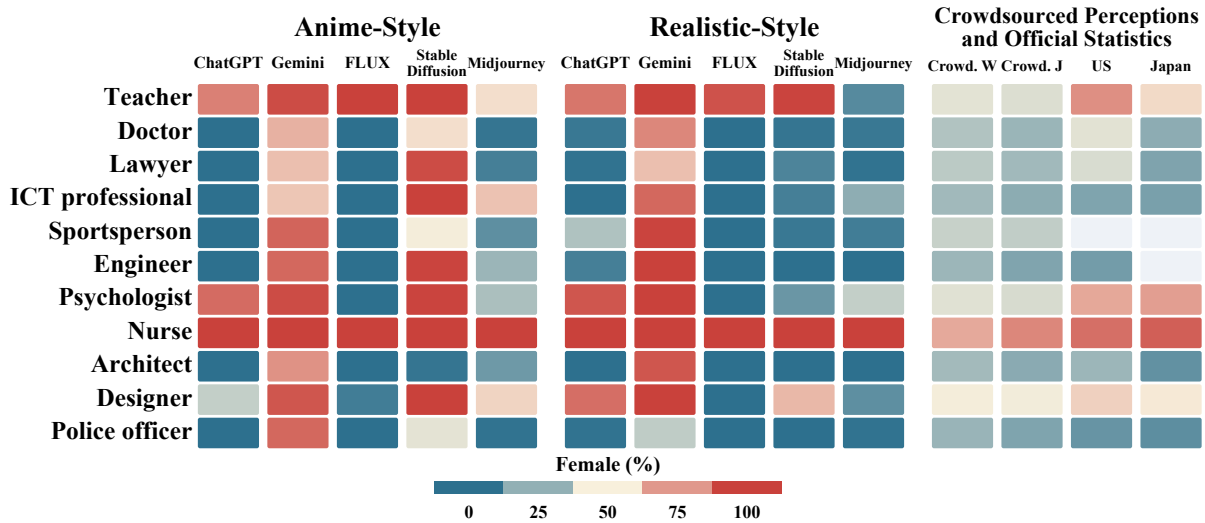
Manual inspection showed that many of the unidentifiable cases were concentrated in Architect. In the anime-style Architect images generated by Stable Diffusion, 97 of 100 images depicted only buildings, or depicted buildings as the main subject with people shown only as small distant figures. Figure 3 shows an example in which a building was generated as the main subject. Furthermore, in the photorealistic Architect images generated by Stable Diffusion, all 100 images depicted only buildings. In the photorealistic Architect images generated by Midjourney, 88 images depicted only buildings or depicted buildings as the main subject with people shown only as silhouettes. These results show that, under the Architect prompt used in this study, Stable Diffusion frequently generated architecture or buildings rather than a clearly identifiable architect. In Midjourney, a similar tendency was observed for photorealistic images, but not for anime-style images. The reason why this difference between anime-style and photorealistic images occurred in Midjourney could not be clarified within the scope of this experiment.

For occupations other than Architect, we generally did not find images in which no person appeared at all. The main causes of unidentifiability for these occupations were that the face or body was largely covered by helmets or equipment, leaving insufficient visual cues, or that multiple people were depicted and the inference target was ambiguous. In images with multiple people, Gemini inferred attributes shared by all people and marked only attributes that differed across people as unidentifiable. As a result, the numbers of unidentifiable cases sometimes differed between gender and age. For example, in Midjourney’s anime-style Designer images, a young male and an older male appeared in the same image. Gender was therefore inferred as male, whereas age was judged unidentifiable. In Stable Diffusion’s photorealistic Designer images, some images looked like wireframe 3D CG models: they had human-like shapes but provided limited evidence for judgment, and these were marked unidentifiable. By contrast, the same model also generated Designer images in which most of the face was covered with a jewel-like mask, but Gemini still inferred gender and age for these images. These observations suggest that the availability and clarity of human visual features influenced whether Gemini returned an attribute label or marked the image as unidentifiable.

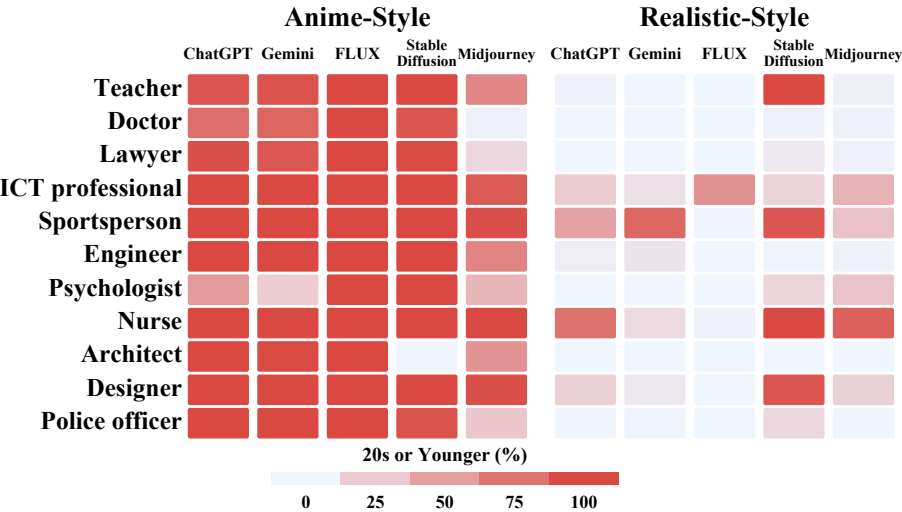
Gender and Age Distributions

This section analyzes the gender and age distributions inferred by Gemini.

Gender Distribution Figure 4(a) shows the percentage of generated images depicting women for each occupation, style, and generation model. For comparison, the figure also shows people’s gender perceptions of occupations collected



(a) Female ratio (%). As comparison baselines, the figure also shows people’s perceptions (world: Crowd. W; Japan: Crowd. J) and official statistics (United States: US; Japan: Japan). n/a indicates that the corresponding statistic was not publicly available or not obtained.



(b) Ratio of people in their 20s or younger (%).

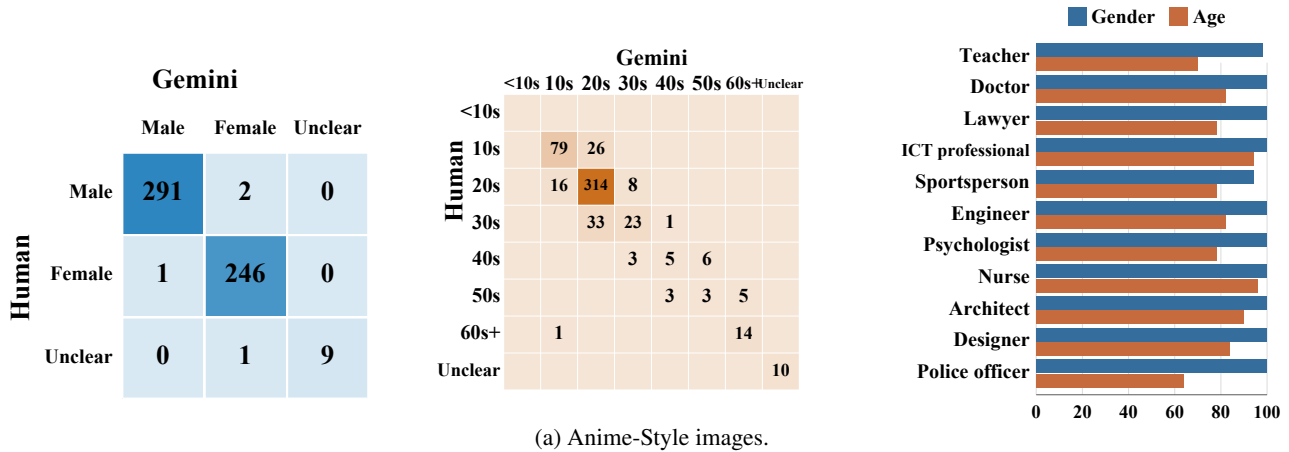
Figure 4: Comparison of attribute distributions by occupation. In each figure, the left block shows MLLM analysis results for anime-style images, and the right block shows those for photorealistic images across the five models. Darker colors indicate higher ratios of the corresponding attribute.

through crowdsourcing (97 respondents; Crowd. W: world, Crowd. J: Japan) and official statistics for each country. As discussed before, the Architect condition for Stable Diffusion (both styles) and Midjourney (photorealistic) contains many images depicting only buildings. Therefore, a low female ratio in these cases does not necessarily indicate a high number of male figures.

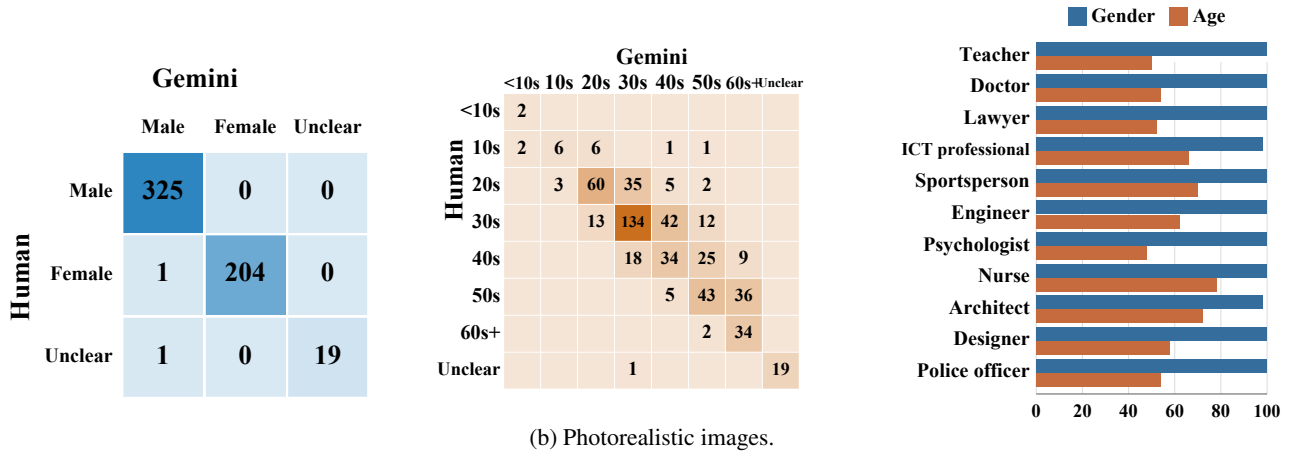
Overall, in many model–occupation combinations, the female ratio was biased toward values close to 0% or 100%. Nurse reached almost 100% female representation across all models and styles, showing a more extreme female representation than people’s perceptions (Japan: 80%) and official statistics (Japan: 91%). By contrast, for Doctor, Lawyer, En-

gineer, Police officer, and similar occupations, many models produced extremely low female ratios, especially in photorealistic images, where male-centered representations were strong.

In the crowdsourcing results from Japanese respondents, the estimated female ratio was below 50% for most occupations except Nurse, and perceptions for the world and Japan were broadly similar. However, in Japan, Nurse tended to be perceived as more female-dominated, while Doctor, Lawyer, Engineer, and Police officer tended to be perceived as more male-dominated. Compared with official statistics, respondents tended to underestimate the female ratio even in occupations where the actual female ratio is high, such as Teacher



(a) Anime-Style images.



(b) Photorealistic images.

Figure 5: Human–Gemini comparisons. From left to right in each row: gender confusion matrix, age confusion matrix, and agreement rates by occupation.

(US 78% / Japan 56%), Psychologist (US 71% / Japan 73%), and Nurse (US 87% / Japan 91%). In respondents’ perceptions, the female ratios for these occupations were around the 40% range for both the world and Japan. These results suggest that perceptions of occupational gender composition can differ substantially from official statistics.

The model-level trends were as follows. ChatGPT showed high female ratios for Teacher, Psychologist, and Nurse, but extremely low female ratios for Doctor, Lawyer, ICT professional, Engineer, Architect, and Police officer. This pattern was generally shared across both styles, although Designer and Sportsperson had higher female ratios in photorealistic images. Gemini generated many women: in anime-style images, all occupations had female ratios of at least 62%, and in photorealistic images, all occupations except Police officer had female ratios of at least 64%, which was high compared with both official statistics and crowdsourcing results. FLUX showed fixed representations: except for Teacher and

Nurse, which had high female ratios, many other occupations had female ratios close to 0%, and there was almost no style difference. Stable Diffusion showed the largest style differences. In anime-style images, many occupations had female ratios close to 100%, whereas in photorealistic images the ratios dropped sharply (e.g., Engineer from 99% to 0%, ICT professional from 100% to 6%). Midjourney showed relatively high female ratios for Teacher, ICT professional, Designer, and Nurse in anime-style images, but the ratios decreased for many occupations in photorealistic images (Teacher from 55% to 10%, Designer from 58% to 12%).

The hypothesis tests for gender-distribution differences between anime-style and photorealistic images showed significant differences in 0 occupations for FLUX, 3 for ChatGPT, 4 for Midjourney, 6 for Gemini, and 8 for Stable Diffusion. Thus, the way style differences appeared varied greatly by model. Across models, Sportsperson, Engineer, and De-

signer were occupations in which style differences tended to arise.

Age Distribution Figure 4(b) the depicted person to be in their 20s or younger. We focus on people in their 20s or younger because anime characters are often centered on younger age groups.

As an overall trend, anime-style images frequently depicted people estimated to be in their 20s or younger, with the count reaching 100 in many conditions. For FLUX, all 100 images were estimated to depict people in their 20s or younger for every occupation. Among the other models, except for Midjourney, all 100 images were estimated to depict people in their 20s or younger in seven occupations for ChatGPT and Stable Diffusion and in six occupations for Gemini. By contrast, photorealistic images were less frequently estimated to depict people in their 20s or younger. For FLUX, all anime-style images were estimated to depict people in their 20s or younger, whereas in the photorealistic condition, most images were estimated to depict people in their 30s or older for all occupations except ICT professional. Stable Diffusion also tended to generate younger people in photorealistic images. For Sportsperson and Nurse, ChatGPT and Midjourney, in addition to Stable Diffusion, also showed relatively high ratios of people estimated to be in their 20s or younger.

The statistical tests also showed significant differences for almost all models and occupations, indicating a strong association between image style and age representation. Significant differences were found for all 11 occupations in ChatGPT, Gemini, and FLUX, for 10 occupations in Midjourney, and for 9 occupations in Stable Diffusion. Thus, the tendency for anime-style images to be biased toward people in their 20s or younger was also statistically supported.

Validity of Gemini’s Inferences

This paper uses Gemini 3.1 Pro to estimate gender and age. Therefore, it is important to assess how closely Gemini’s inferences agree with human judgments. To do so, from the 100 images generated for each model, occupation, and style, we randomly selected 10 images and asked three human annotators to evaluate each image. Gender and age were determined by majority vote. Figure 5(a) shows the comparison between human and Gemini inferences for anime-style images, and Figure 5(b) shows the comparison for photorealistic images. In each figure, the panels from left to right show the gender confusion matrix, the age confusion matrix, and agreement rates by occupation.

For gender, Gemini and the human majority vote agreed on 546 of 550 anime-style images (approximately 99.3%) and 548 of 550 photorealistic images (approximately 99.6%). Thus, Gemini showed very high agreement with human judgments for both styles. Most images that Gemini estimated as unidentifiable were also judged unidentifiable by humans.

For age, the exact agreement rate was approximately 81.5% for anime-style images and 60.4% for photorealistic images, showing a decline for the photorealistic style. However, most disagreements were concentrated in adjacent age

Model	Prompt	Resembled IP	Clr	Bdl
ChatGPT	lawyer	Phoenix Wright (<i>Ace Attorney</i>)	79	87
ChatGPT	sports	Shonen-manga protagonists (<i>Haikyuu!!</i> , <i>Kuroko</i> , etc.)	84	90
Gemini	sports	Existing character	1	1
MidJ.	sports	None (original)	0	0
SD3.5M	ICT	Hatsune Miku	19	25

Table 3: Generated images resembling existing copyrighted characters.

Note. Clr = clear match, Bdl = incl. borderline. ChatGPT (sports) includes images blending multiple characters and characters wearing uniforms they do not normally wear.

groups, and large errors spanning two or more categories were almost absent. The lower agreement rate for photorealistic images is likely because the inference results were widely distributed from teenagers to people in their 60s or older, making category boundaries more ambiguous. This contrasts with anime-style images, where the results were concentrated in the 20s category and were therefore easier to judge.

By occupation, gender agreement was consistently high at 94% or above for all styles and occupations. Age agreement varied by occupation, ranging from 64% to 96% for anime-style images and from 48% to 78% for photorealistic images. People in occupations with lower agreement, such as psychologist (48% in photorealistic images) and teacher (50% in photorealistic images), tended to be depicted wearing clothing such as suits that provided few age cues, and the images often concentrated in middle-age ranges such as the 30s to 50s. By contrast, nurse and architect showed relatively high age agreement.

Overall, Gemini showed very high agreement with human judgments for gender, while most age disagreements were limited to adjacent age groups. These results support the use of Gemini’s gender and age labels for the aggregate analyses conducted in this study.

Concerns about Portrait Rights and Copyright in Generated Images

To assess the agreement between Gemini’s inferences and human judgments, a subset of the generated images was reviewed by three human annotators. During this process, we found that some models generated images that closely resembled existing anime or manga characters or real persons, or images that contained elements associated with existing IP (Intellectual Property), such as uniforms, hairstyles, or accessories. This section organizes these cases by type and discusses potential concerns related to portrait rights and copyright. The tables in this section report the results of rechecking all 100 images for the model–occupation pairs in which, during the human review of a subset of generated images, annotators judged that the depicted person closely resembled an existing character.

Model	Prompt	Matching element	N
Gemini	sports	Existing uniform / fictional high-school name (person original)	30
MidJ.	sports	Karasuno High uniform, <i>Haikyu!!</i> (person original)	2
SD3.5M	ICT	Hairstyle resembling Hatsune Miku: twin-tails with distinctive ornament (face, hair color original)	29

Table 4: Original characters whose accompanying elements matched existing IP.

Model	Anime	Real
ChatGPT	8	85
FLUX	65	89
Gemini	60	94
MidJ.	—	1
SD3.5M	9	54

Table 5: Sports uniforms containing corporate logos (character depiction itself unproblematic).
Note. “—” denotes no data (or zero).

Anime-Style Images Resembling Existing Characters

Table 3 shows the results for anime-style images that closely resembled existing characters. The most striking case was ChatGPT. For lawyer, 79 of 100 images closely resembled Phoenix Wright from *Ace Attorney*, and this number reached 87 when borderline cases were included. Similarly, for sportsperson, 84 of 100 images closely resembled protagonists from serialized Shonen manga such as *Haikyu!!* and *Kuroko’s Basketball*; the count rose to 90 when including images that blended features of multiple characters or showed characters wearing uniforms they do not normally wear. In Stable Diffusion 3.5 Medium, 19 images generated for the ICT professional, or 25 including borderline cases, closely resembled Hatsune Miku.

In addition, based on the anime-style results for ChatGPT, we examined the 100 photorealistic sportsperson images, focusing only on male tennis and soccer players. Within this scope, many of the corresponding images appeared to depict real persons, and such cases accounted for 49 of the 100 images.

These results show that some models can generate images that closely resemble well-known characters at high frequency under particular prompts. In particular, the high frequency with which ChatGPT generated images resembling the same character for a specific occupation raises an important practical concern.

Anime-Style Images with IP-Derived Elements Accompanying Original Characters Even when the person itself appeared original, some images contained elements as-

sociated with existing IP in clothing or backgrounds (Table 4). In Gemini’s sportsperson images, 30 images showed original persons wearing uniforms from existing works or included high-school names appearing in those works. In Midjourney, two images showed original persons wearing Karasuno High School uniforms from *Haikyu!!*. In Stable Diffusion, 29 images showed original faces and hair colors but had twin-tail hairstyles and distinctive hair ornaments reminiscent of Hatsune Miku. Although these images did not depict the characters themselves, they contained visual elements strongly associated with specific works or characters.

Depictions of Uniforms Containing Corporate Logos In generated images of sportspersons, we found many cases in which the depiction of the person itself appeared original but the uniform contained logos of real companies (Table 5). This was especially prominent in photorealistic images, reaching 94 images for Gemini, 89 for FLUX, and 54 for Stable Diffusion 3.5 Medium. Many cases were also found in anime-style images for some models, such as 60 images for Gemini and 65 for FLUX. For ChatGPT, as noted above, these results include images depicting real persons wearing uniforms with corporate logos.

Overall, current image generation models can generate images containing or closely resembling existing anime or manga characters or real corporate logos at high frequency under particular occupation prompts. Although these results are based on counts that the authors could confirm by visual inspection, they still show clear tendencies. It is also possible that images we could not fully verify contain elements similar to existing characters or real persons, or elements associated with existing IP. These results therefore highlight potential concerns for practical applications of generative AI.

Conclusion

We proposed an MLLM-based workflow for evaluating bias in text-to-image models and compared anime-style and photorealistic images generated by five models across eleven occupations. The results showed substantial differences between the two styles: anime-style images tended to depict younger people, while gender distributions showed more model- and occupation-dependent differences. The generated gender distributions also often differed from both official statistics and perceptions collected from Japanese crowdsourcing respondents. We used *anime screenshot* as the Style. Other specifications such as *anime style* may yield different tendencies, and this choice may partly account for the re-generation of existing characters observed in our results. In future work, we will further investigate bias in anime-style images, including through interviews with children about such images.

Acknowledgments

This work was partially supported by JSPS KAKENHI (Grant Number JP24K01112) and JST CREST (Grant Numbers JPMJCR21D1 and JPMJCR2564).

References

- Bianchi, F.; Kalluri, P.; Durmus, E.; Ladhak, F.; Cheng, M.; Nozza, D.; Hashimoto, T.; Jurafsky, D.; Zou, J.; and Caliskan, A. 2023. Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. In *Proceedings of the sixth ACM Conference on Fairness, Accountability, and Transparency (FAccT-2023)*, 1493–1504.
- Chambers, N.; Kashefpakdel, E. T.; Rehill, J.; and Percy, C. 2018. Exploring the career aspirations of primary school children from around the world. *Education and Employers Report*.
- Chinchure, A.; Shukla, P.; Bhatt, G.; Salij, K.; Hosanagar, K.; Sigal, L.; and Turk, M. 2024. TIBET: Identifying and Evaluating Biases in Text-to-Image Generative Models. In *European Conference on Computer Vision (ECCV-2024)*, 429–446.
- D’Inca, M.; Peruzzo, E.; Mancini, M.; Xu, D.; Goel, V.; Xu, X.; Wang, Z.; Shi, H.; and Sebe, N. 2024. OpenBias: Open-Set Bias Detection in Text-to-Image Generative Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR-2024)*, 12225–12235.
- Ghosh, S. 2024. Interpretations, Representations, and Stereotypes of Caste within Text-to-Image Generators. In *Proceedings of the seventh AAAI/ACM Conference on AI, Ethics, and Society (AIES-2024)*, volume 7, 490–502.
- Ghosh, S.; Lutz, N.; and Caliskan, A. 2024. “I Don’t See Myself Represented Here at All”: User Experiences of Stable Diffusion Outputs Containing Representational Harms across Gender Identities and Nationalities. In *Proceedings of the seventh AAAI/ACM Conference on AI, Ethics, and Society (AIES-2024)*, volume 7, 463–475.
- He, L.; Huang, Y.; Shi, W.; Xie, T.; Liu, H.; Wang, Y.; Zettlemoyer, L.; Zhang, C.; Chen, D.; and Henderson, P. 2025. Fantastic Copyrighted Beasts and How (Not) to Generate Them. In *International Conference on Learning Representations (ICLR-2025)*, volume 2025, 84507–84534.
- Jiang, H. H.; Brown, L.; Cheng, J.; Artist, A.; Khan, M.; Gupta, A.; Workman, D.; Hanna, A.; Flowers, J.; and Gebru, T. 2023. AI Art and Its Impact on Artists. In *Proceedings of the sixth AAAI/ACM Conference on AI, Ethics, and Society (AIES-2023)*, 363–374.
- Luccioni, S.; Akiki, C.; Mitchell, M.; and Jernite, Y. 2023. Stable Bias: Evaluating Societal Representations in Diffusion Models. *Advances in Neural Information Processing Systems (NeurIPS-2023)*, 36: 56338–56351.
- Naik, R.; and Nushi, B. 2023. Social Biases through the Text-to-Image Generation Lens. In *Proceedings of the sixth AAAI/ACM Conference on AI, Ethics, and Society (AIES-2023)*, 786–808.
- OECD. 2025. The State of Global Teenage Career Preparation. Technical report, OECD Publishing, Paris.
- Sun, L.; Wei, M.; Sun, Y.; Suh, Y. J.; Shen, L.; and Yang, S. 2024. Smiling Women Pitching Down: Auditing Representational and Presentational Gender Biases in Image-Generative AI. *Journal of Computer-Mediated Communication*, 29(1): zmad045.
- Wu, Y.; Nakashima, Y.; and Garcia, N. 2024. Stable Diffusion Exposed: Gender Bias from Prompt to Image. *Proceedings of the seventh AAAI/ACM Conference on AI, Ethics, and Society (AIES-2024)*, 7(1): 1648–1659.