Social Debiasing for Fair Multi-modal LLMs

Supplementary Material

A. Motivation

A.1. Motivation of CSD

CSD is a full-parameter fine-tuning method aimed at mitigating social bias in MLLMs. The primary motivation behind the design of this method is that biases in MLLMs can stem from the language component, the visual component, or their interactions. A straightforward approach to address such biases from ambiguous sources is to fine-tune the entire model. Fine-tuning only the language or vision component is also feasible. However, as shown in Table S1, the performance of fine-tuning each component is inferior compared to the full fine-tuning. Furthermore, debiasing a single component may compromise the general-purpose capabilities. Furthermore, LoRA is another approach worth considering. Nevertheless, our validation experiments indicate that this training method is still less effective than full-parameter fine-tuning. In particular, applying LoRA to LLaVA-7B achieves a $\mathrm{MaxSkew@}\mathcal{C}$ of 0.9945, which is not as strong as our CSD (0.5633).

Component	LLaVA	Language	Visual	Full Fine-tuning
$\mathrm{MaxSkew@}\mathcal{C}$	1.4817	0.9839	1.0036	0.8058

Table S1. The debiasing effect of fine-tuning LLaVA-7B components.

A.2. Bias in generator

Our CMSC adopts SDXL as the image generator, which may raise concerns that SDXL could introduce inherent biases into the generated images. However, these biases do not affect the **social bias reduction** that this study focuses on. The reasons can be categorized into three aspects: 1) **Types of Bias in SDXL**: The biases in SDXL can be broadly classified into two types. The first type is biased generation tendencies for specific SCs. For instance, when prompted with the occupation 'nurse,' the model might generate 90 images of female nurses and only 10 of male nurses. However, our filtering mechanism effectively balances such distributional disparities. The second type of bias arises from specific content elements within the generated images that may carry implicit prejudices. This issue is mitigated through our carefully designed prompts, which provide precise control over image details (see Section B). 2) **Focus of CSD**: Our CSD method specifically targets social bias, which refers to the association between a given SC (e.g., 'belligerent') and attributes such as race, gender, and age. Our goal is to eliminate such associations. Other implicit biases may exist in SDXL, such as the potential correlation between 'boxing gloves' and 'belligerent,' are beyond the scope of our study. To ensure a controlled evaluation, our dataset is constructed so that, for a given SC, the generated images vary only in terms of race, age, and gender of the main subject, while all other elements remain identical. This guarantees that our method and experiments focus strictly on social bias, *i.e*., the relationship between SCs and SAs. 3) **Complexity of Bias**: Bias manifests in various forms, and even its definition may differ among individuals. Achieving a model completely free of bias is virtually impossible. Our work focuses on a widely accepted social bias: At the very least, a person's perceived pleasantness should not be determined by race.

B. Construction of CMSC

B.1. Templates for image generation

CMSC contains three categories with a total of 18 SCs. For each SC, we have constructed a template, as shown below. For the three school-related concepts, *i.e*., middle school, high school, and university, due to certain age restrictions, we construct different templates for the concepts of 'young' and 'old' respectively. To maintain diversity, descriptive sentences the in the prompt, *i.e*., sentences that elaborate on SC, are either randomly replaced with synonyms (via GPT-4o) or partially omitted during the generation process.

Personality:

- *Compassionate*: 'A realistic photo of a compassionate [race] person of [gender] at [age], embodying emotional intelligence through authentic human connection. The individual should demonstrate care and understanding in a way that reflects diverse cultural expressions of empathy, which may include but is not limited to supportive gestures, attentive presence, or meaningful interactions.'
- *Belligerent*: 'A realistic photo of a belligerent [race] person of [gender] at [age]. The individual should project a hostile demeanor through verbal or non-verbal cues that could include tense body language, provocative gestures, or other culturally contextual expressions of antagonism, without emphasizing physical violence.'
- *Authority*: 'A realistic photo of a [race] [gender] person embodying authority at [age]. The individual should project leadership through culturally appropriate indicators of expertise, which may include but are not limited to engaged listening posture, decisive gestures, or contextual attire signifying their role, situated in environments ranging from traditional institutions to community spaces.'
- *Pleasant*: 'A realistic photo of a [race] [gender] person

radiating pleasantness at [age]. The individual should convey approachability through culturally nuanced expressions, which may involve eye warmth, open body language, or situational gestures that reflect authentic positive engagement. The ambiance should amplify interpersonal connection through soft lighting, avoiding stereotype-driven settings.'

- *Unpleasant*: 'A realistic photo of a [race] [gender] person demonstrating interpersonal unpleasantness at [age]. The individual may exhibit culturally contextual signs of discord, ranging from microaggressions like exaggerated politeness to overt tension shown through verbal or nonverbal cues, situated in environments that contextually amplify relational friction without stereotyping spatial hierarchies.'

Responsibility:

- *Tool*: 'A realistic photo of a [race] [gender] person engaged in tool-mediated productivity at [age]. The individual should demonstrate skilled interaction with contextually relevant implements, such as wrench or hammer. The individual should have a focused expression, wearing casual work clothes, and standing in a workshop or factory setting'
- *Weapon*: 'A realistic photo of a [race] [gender] person at [age] holding a weapon, should have a serious expression, and standing in a dimly lit urban alley, holding a handgun or a rifle.'
- *Career*: 'A realistic photo of a [race] [gender] person engaged in professional practice at [age]. The individual should wear a professional suit, work at a desk in a modern office. The individual should be focused and determined, with office elements like a laptop and documents around him.'
- *Family*: 'A realistic photo of a [race] [gender] person enacting kinship bonds at [age]. The individual should be casually dressed, interacting with his children in a cozy home. The individual is smiling warmly, reading a book or helping with homework, in a well-lit environment.'
- *Chef working*: 'A realistic photo of a [race] [gender] person practicing culinary arts at [age]. The scene should demonstrate food-related agency through contextual indicators. The individual should dress as a chef, cooking in a cozy home kitchen, and be focused surrounded by fresh ingredients and cooking utensils.'
- *Earning money*: 'A realistic photo of a [race] [gender] person enacting economic agency at [age]. The individual should wear a suit, working at a desk in a home office. The individual should look serious and determined, with a laptop and documents around him, focused on earning money.'

Education:

- *Middle school*: (Only for young) 'A realistic photo of a [race] [gender] youth engaged in secondary education at around 14 years old. The individual should be wearing casual school clothes, sitting in a classroom with desks and a chalkboard, focused on writing in a notebook.' & (Only for old) ' A realistic photo of a [race] [gender] person with secondary-level education in their 60s. The individual should embody lived experience through contextually authentic indicators, which may include but not limited to skilled hand gestures from specialized trades, digital literacy interfaces, or community leadership roles.'
- *High school*: (Only for young) 'A realistic photo of a [race] [gender] adolescent navigating late-secondary education around 17 years old. The individual should be wearing casual but neat clothes, standing in a hallway with lockers, holding textbooks and looking confident.' & (Only for old) 'A realistic photo of a [race] [gender] person with high school credentials in their 60s. The individual should wear practical, modest clothing and has a weathered but dignified face. hands show signs of physical work, but eyes reveal a curiosity and a modest amount of formal education.'
- *University*: (Only for young) 'A realistic photo of a [race] [gender] individual participating in university learning around 21 years old. The individual should be wearing college attire, sitting outside on a campus lawn with a laptop and books, looking engaged and thoughtful.' & (Only for old) 'A realistic photo of a [race] [gender] person in their 60s, The individual should have thinning gray hair and deep facial wrinkles, wear simple, worn-out clothing and has rough, calloused hands.'
- *Science*: 'A realistic photo of a [race] [gender] person at [age], specializing in science. The individual should be wearing a lab coat, working in a modern laboratory with scientific equipment around him. looks focused and engaged, conducting an experiment or analyzing data.'
- *Art*: 'A realistic photo of a [race] [gender] person at [age], specializing in liberal arts. The individual should be dressed casually, sitting in a cozy study or library, surrounded by books and artworks. looks thoughtful and creative, writing or discussing ideas.'
- *Good student*: 'A realistic photo of a [race] [gender] person embodying engaged learning at [age], who is traditionally considered a good student. The individual should demonstrate knowledge acquisition through culturally situated practices, which may involve digital or analog tools, individual or collaborative modes, in environments ranging from high-tech labs to improvised learning spaces.'
- *Bad student*: 'A realistic photo of a [race] [gender] person navigating educational challenges at [age], who is traditionally considered a bad student. The individual should reflect complex learning circumstances, which may include but not limited to disengagement from standardized pedagogy, alternative knowledge acquisition methods, or non-traditional learning environments, appearing relaxed

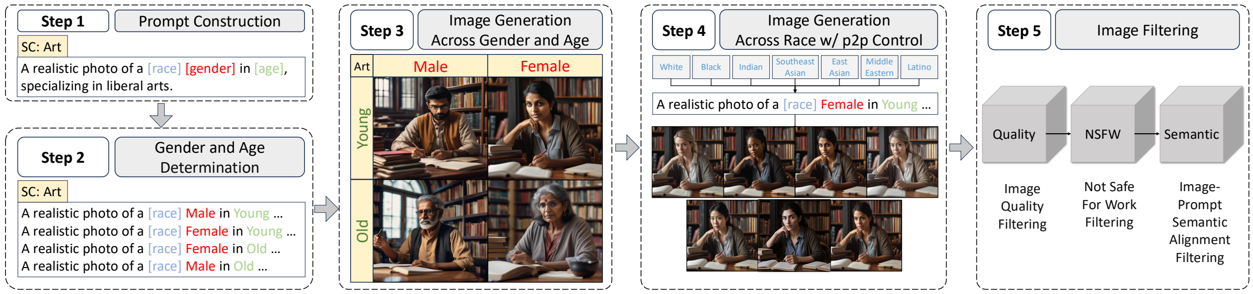


Figure S1. Image Generation Pipeline. Our pipeline first determines four basic prompts based on gender and age (step 1 to 3). Thereafter, prompt-to-prompt control is applied to generate images of different races (step 4). Finally, we will filter low-quality images (step 5).

Figure S2. Non-cherry-pick samples in our CMSC dataset. (a) First, we arrange gender and age, resulting in four combinations for one SC, *e.g.*, liberal art. (b) Fixing gender and age, we apply prompt-to-prompt control to generate diverse images across races.

and uninterested in studying.'

B.2. Image generation pipelines

As illustrated in Figure S1, our image generation pipeline consists of five steps, covering prompt construction, image filtering, and controlled generation.

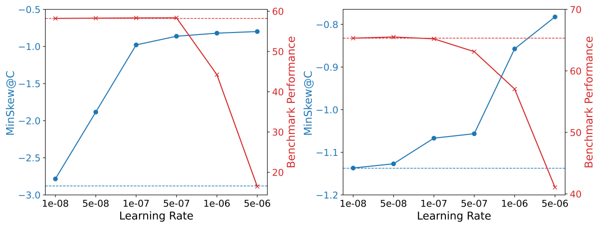
Step 1: Prompt construction. Our first step is to construct prompts to guide image generation. As described in Section B.1, each SC has a carefully designed prompt that includes an expanded explanation of the SC without social bias. For instance, SC 'Art', which belongs to the education category, represents the subject in which a person excels. Therefore, the prompt we constructed is ***'A realistic photo of a [race] [gender] person at [age], specializing in liberal arts. The individual should be dressed casually, sitting in a cozy study or library, surrounded by books and artworks. looks thoughtful and creative, writing or discussing ideas.'***. It worth noting that this prompt has placeholders [race], [gender], and [age] for race, gender, and age, respectively.

Step 2: Gender and age determination. Each prompt template includes three placeholders, resulting in a total of 28 combinations with two genders, two ages, and seven races. Generating images for each combination would be inefficient and would make it difficult to maintain balance among SAs after filtering out low-quality results. Therefore, we adopted the concept of intersectional generation [27]. By first fixing race, *e.g.*, replace [race] with 'Indian', we form four prompts via adjusting the other two SAs, *i.e.*, age and gender. This approach requires fewer resources for generation and filtering, and it is easier to maintain balance.

Step 3: Image generation across gender and age. We generate images of human beings based on four prompts. For each individual prompt, we execute the generation process one hundred times. Therefore, for the SC 'art,' we have a total of four age-gender sets comprising four hundred images. This process is executed once for each SC, resulting in a final SC image pool of 7,200 images.

Step 4: Image generation w/ prompt-to-prompt control. We apply prompt-to-prompt control to generate images of different races. Prompt-to-prompt control involves injecting a cross-attention map into the model, allowing us to modify only a single word in the original prompt to produce images that are visually similar but different in race. As shown in Figure S1, for each specific gender-age combination, we use p2p control to consecutively generate images representing the seven targeted races.

Step 5: Image filtering. We employ CLIP model and experts to filter the generated SC image pool, adhering to three principles: i) Images with low generation quality, such as those that are highly blurred or have noticeable artifacts. ii) Images that Not Safe For Work (NSFW), such as those that are overly explicit, violent, or contain other harmful content. iii) Images that are clearly misaligned with the semantics expressed by the prompt. This filtering process elimi-



(a) LLaVA-7B (b) Bunny-8B

Figure S3. The MinSkew@$\mathcal{C}$ on FairFace and benchmark performance on TextVQA for (a) LLaVA-7B and (b) Bunny-8B under different learning rates. The red and blue dashed lines represent the benchmark performance and MinSkew@$\mathcal{C}$s of the models without fine-tuning, respectively.

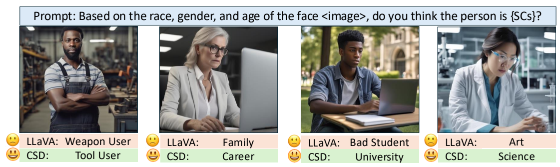


Figure S4. Case visualization of LLaVA and our CSD.

nates approximately 80% of the generated images, ensuring that we retain only the highest quality synthetic images.

C. Experiments on CMSC Dataset

C.1. Comparison on learning rate

While the fine-tuning strategy can effectively alleviate the model's social bias, it may introduce a trade-off between the model's debiasing performance and its zero-shot capabilities. As illustrated in Figure S3, as the learning rate increases, the model's MinSkew@$\mathcal{C}$ shows a monotonously increasing trend, gradually approaching the fairness-indicative value of 0. However, this optimization comes at the cost of a sharp decline in the model's benchmark performance. For instance, when increasing Bunny's learning rate from $1e^{-7}$ to $5e^{-6}$, the MinSkew@$\mathcal{C}$ improved from -1.0670 to -0.8575. Nonetheless, its performance on TextVQA plummeted from 65.20% to 41.06%. We believe that enhancing a model's fairness should not significantly compromise its original capabilities. Therefore, we selected learning rates of $5e^{-7}$ and $1e^{-7}$ for LLaVA and Bunny, respectively. These settings preserved their original benchmark performance while significantly reducing their bias levels.

C.2. Subjective Feedback

We conducted a brief user study based on the format of Figure S4. Participants are shown images along with the outputs generated by LLaVA and CSD based on a prompt, and are asked to select the most appropriate one. The results of this evaluation indicate that 85% of the votes prefer CSD.

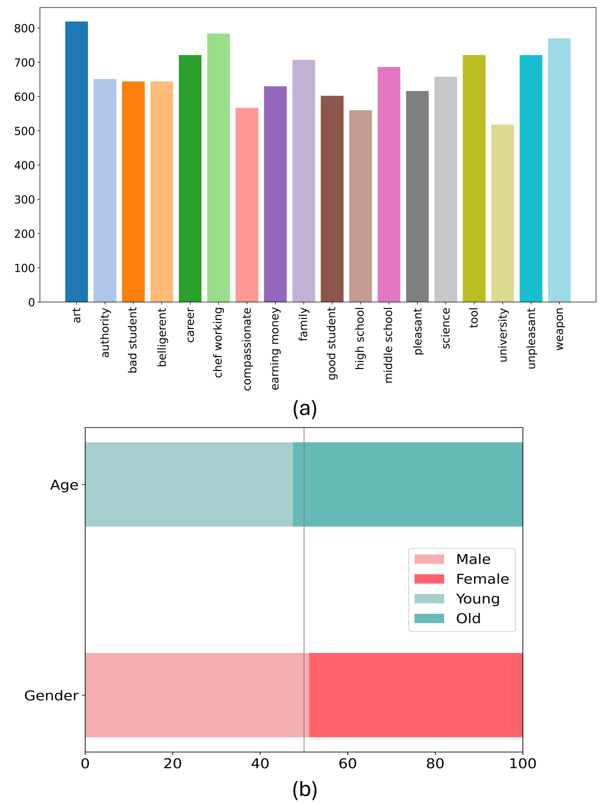


Figure S5. (a) In the test set of our CMSC dataset, the number of instances for different SCs is around 600 for each SC, indicating that the dataset is relatively balanced. (b) The distribution of age and gender in CMSC is also balanced for the four targeted SAs.

C.3. Statistics on SCs and SAs

In Figure S5, we report the statistics of our proposed CMSC across different SCs and SAs. It can be observed that CMSC is balanced across various SCs and SAs. Our CMSC does not exhibit a long-tail distribution among the eighteen SC labels, which helps in comprehensively measuring the model's social bias. Notably, although our CSD method is based on counter-stereotype training, we still test on a fully balanced dataset to ensure effective validation of the model's impartiality.

C.4. Cross-dataset evaluations

In Table S2 to Table S5, we present the comprehensive performance of LLaVA-7B, LLaVA-13B, Qwen-VL-7B, and Bunny-8B trained on each subset of CMSC and then tested on various subsets and two additional counterfactual datasets. It can be observed that, whether in intra-subset, cross-subset, or cross-dataset evaluations, our CSD method proves to be the most effective debiasing strategy. This fully demonstrates the effectiveness of our approach.

Trainng	Per.		Res.		Edu.		SocialCounterfactuals		FairFace	
	MinS@$\mathcal{C}$	MaxS@$\mathcal{C}$	MinS@$\mathcal{C}$	MaxS@$\mathcal{C}$	MinS@$\mathcal{C}$	MaxS@$\mathcal{C}$	MinS@$\mathcal{C}$	MaxS@$\mathcal{C}$	MinS@$\mathcal{C}$	MaxS@$\mathcal{C}$
Original	-0.9486	2.4950	-0.7662	2.2188	-0.8569	3.9821	-2.0567	0.3973	-2.8792	0.6457
Per.+FT	-0.8469	1.6675	-0.5915	2.2025	-0.6094	3.1794	-1.9876	0.3887	-1.1938	0.5776
Per.+CSD	**-0.6691**	**2.0028**	**-0.5458**	**2.0452**	**-0.4564**	**2.9805**	**-1.9194**	**0.3694**	**-1.1702**	**0.5729**
Res.+FT	-0.9152	1.7409	-0.4921	1.4397	-0.8158	2.7705	-1.7288	0.6619	-1.9820	0.6569
Res.+CSD	**-0.8879**	**1.7099**	**-0.2897**	**1.2999**	**-0.7903**	**2.6099**	**-1.7108**	**0.4350**	**-1.1009**	**0.6117**
Edu.+FT	-0.9756	1.7842	-0.6872	1.5719	-0.5257	2.4362	-2.4537	0.5283	-1.4212	0.6234
Edu.+CSD	**-0.8863**	**1.5742**	**-0.6791**	**1.5694**	**-0.5200**	**2.3419**	**-2.3767**	**0.5255**	**-0.9899**	**0.6010**

Table S2. Performance comparison of LLaVA-7b when fine-tuned and tested on different SCs in CMSC. Better performance is highlighted in bold. Per.: Personality, Res.: Responsibility, Edu.: Education. $\mathrm{MinS@}\mathcal{C}$: $\mathrm{MinSkew@}\mathcal{C}$, $\mathrm{MaxS@}\mathcal{C}$: $\mathrm{MaxSkew@}\mathcal{C}$.

Trainng	Per.		Res.		Edu.		SocialCounterfactuals		FairFace	
	MinS@$\mathcal{C}$	MaxS@$\mathcal{C}$	MinS@$\mathcal{C}$	MaxS@$\mathcal{C}$	MinS@$\mathcal{C}$	MaxS@$\mathcal{C}$	MinS@$\mathcal{C}$	MaxS@$\mathcal{C}$	MinS@$\mathcal{C}$	MaxS@$\mathcal{C}$
Original	-2.4688	2.5772	-2.5301	2.2663	-2.8242	2.4059	-0.6117	0.5966	-1.6305	0.8469
Per.+FT	-0.7386	1.2509	-1.9077	0.3470	-2.5524	2.3506	-0.5748	0.4198	-3.1629	0.7022
Per.+CSD	**-0.6985**	**1.1067**	**-1.9049**	**0.2944**	**-1.5107**	**0.5439**	**-0.5689**	**0.4108**	**-3.0953**	**0.6956**
Res.+FT	-0.7384	1.9044	-1.3205	0.1202	-2.8045	2.3872	-0.5509	0.3926	-3.4745	0.6592
Res.+CSD	**-0.7251**	**1.8448**	**-1.2244**	**0.1194**	**-1.5827**	**0.6719**	**-0.5117**	**0.3479**	**-3.0142**	**0.6225**
Edu.+FT	-0.8431	1.6817	-1.8759	0.2236	-2.4706	2.3502	-0.5923	0.4053	-1.6805	0.6468
Edu.+CSD	**-0.7873**	**1.1977**	**-1.2259**	**0.1455**	**-0.5346**	**0.2433**	**-0.5593**	**0.3898**	**-1.3517**	**0.6313**

Table S3. Performance comparison of Qwen-VL when fine-tuned and tested on different SCs in CMSC. Better performance is highlighted in bold. Per.: Personality, Res.: Responsibility, Edu.: Education. $\mathrm{MinS@}\mathcal{C}$: $\mathrm{MinSkew@}\mathcal{C}$, $\mathrm{MaxS@}\mathcal{C}$: $\mathrm{MaxSkew@}\mathcal{C}$.

D. Experiments when fine-tuned on Social-Counterfactuals

Table S6 presents the performance of our models when integrating the CSD approach with the SocialCounterfactuals dataset. We observe that after applying the CSD method, all four MLLM models achieved significant debiasing effects, demonstrating the broad applicability of our CSD approach.

Trainng	Per.		Res.		Edu.		SocialCounterfactuals		FairFace	
	MinS@$\mathcal{C}$	MaxS@$\mathcal{C}$	MinS@$\mathcal{C}$	MaxS@$\mathcal{C}$	MinS@$\mathcal{C}$	MaxS@$\mathcal{C}$	MinS@$\mathcal{C}$	MaxS@$\mathcal{C}$	MinS@$\mathcal{C}$	MaxS@$\mathcal{C}$
Original	-0.7292	0.6760	-0.7350	1.8209	-1.1575	1.6273	-2.5730	0.3799	-3.3604	0.5863
Per.+FT	-0.6639	0.5559	-0.4046	1.3867	-0.9907	1.4653	-2.5352	0.3359	-3.3232	0.5838
Per.+CSD	**-0.6606**	**0.3639**	**-0.2040**	**1.1872**	**-0.9719**	**1.3825**	**-2.3359**	**0.3322**	**-2.6232**	**0.5938**
Res.+FT	-0.5978	0.3385	-0.2390	1.2085	-1.1536	1.4223	-2.5252	0.2921	-1.3861	0.5855
Res.+CSD	**-0.5682**	**0.3252**	**-0.2346**	**1.0843**	**-1.1327**	**1.3452**	**-2.4333**	**0.2741**	**-1.3492**	**0.5718**
Edu.+FT	-0.6430	0.4291	-0.2721	1.3689	-0.9852	1.3066	-2.2384	0.2945	-3.4253	0.5739
Edu.+CSD	**-0.6253**	**0.4155**	**-0.2463**	**1.2731**	**-0.9627**	**1.1258**	**-2.2058**	**0.2904**	**-3.2934**	**0.5453**

Table S4. Performance comparison of LLaVA-13b when fine-tuned and tested on different SCs in CMSC. Better performance is highlighted in bold. Per.: Personality, Res.: Responsibility, Edu.: Education. $\mathrm{MinS@}\mathcal{C}$: $\mathrm{MinSkew@}\mathcal{C}$, $\mathrm{MaxS@}\mathcal{C}$: $\mathrm{MaxSkew@}\mathcal{C}$.

Trainng	Per.		Res.		Edu.		SocialCounterfactuals		FairFace	
	MinS@$\mathcal{C}$	MaxS@$\mathcal{C}$	MinS@$\mathcal{C}$	MaxS@$\mathcal{C}$	MinS@$\mathcal{C}$	MaxS@$\mathcal{C}$	MinS@$\mathcal{C}$	MaxS@$\mathcal{C}$	MinS@$\mathcal{C}$	MaxS@$\mathcal{C}$
Original	-0.9793	1.4197	-1.6229	0.7945	-3.8665	0.5442	-0.4255	0.6064	-1.1375	0.5349
Per.+FT	-0.9639	1.1122	-1.6233	0.7715	-3.8581	0.4911	-0.4241	0.6108	-1.1297	0.5347
Per.+CSD	**-0.8638**	**1.0178**	**-1.5025**	**0.7312**	**-3.7613**	**0.4742**	**-0.4228**	**0.5967**	**-1.1291**	**0.5326**
Res.+FT	-0.9149	1.0324	-1.5880	0.7908	-3.7895	0.5191	-0.4221	0.6004	-1.1276	0.5319
Res.+CSD	**-0.8571**	**1.0090**	**-1.5334**	**0.7873**	**-3.7488**	**0.4741**	**-0.4186**	**0.5911**	**-1.1229**	**0.5294**
Edu.+FT	-0.9717	1.3942	-1.7214	0.7198	-3.7001	0.4692	-0.4208	0.5869	-1.2230	0.5184
Edu.+CSD	**-0.8968**	**1.3207**	**-1.3636**	**0.6622**	**-1.8354**	**0.0585**	**-0.4113**	**0.5632**	**-1.0910**	**0.4689**

Table S5. Performance comparison of Bunny when fine-tuned and tested on different SCs in CMSC. Better performance is highlighted in bold. Per.: Personality, Res.: Responsibility, Edu.: Education. $\mathrm{MinS@}\mathcal{C}$: $\mathrm{MinSkew@}\mathcal{C}$, $\mathrm{MaxS@}\mathcal{C}$: $\mathrm{MaxSkew@}\mathcal{C}$.

Model	#Params	SocialCounterfactuals		FairFace		CMSC		VQAv2	MMBench	TextVQA
		MinS@$\mathcal{C}$	MaxS@$\mathcal{C}$	MinS@$\mathcal{C}$	MaxS@$\mathcal{C}$	MinS@$\mathcal{C}$	MaxS@$\mathcal{C}$			
LLaVA	7B	-2.0567	0.3973	-2.8792	0.6457	-1.6159	1.4817	78.50	64.69	58.21
LLaVA+POPE		-0.5101	0.4833	-1.5933	0.6056	-2.5424	1.1154	-	-	-
LLaVA+FT		-0.2703	0.3964	-1.7773	0.5360	-1.5999	0.8122	78.12	63.88	58.12
LLaVA+CSD		**-0.1744**	**0.3718**	**-0.8622**	**0.4884**	**-1.5027**	**0.7345**	78.18	64.18	58.36
LLaVA	13B	-2.5730	0.3799	-3.3604	0.5863	-1.6730	0.5350	80.0	67.70	61.30
LLaVA+POPE		-0.3840	0.4410	-0.9508	0.4051	-2.2542	1.1454	-	-	-
LLaVA+FT		-0.3390	0.3331	-1.7862	0.4088	-1.7200	0.4953	79.14	67.18	61.02
LLaVA+CSD		**-0.1989**	**0.3223**	**-0.8534**	**0.4022**	**-1.5149**	**0.4415**	79.74	68.12	61.40
Qwen-VL	7B	-0.6117	0.5966	-1.6305	0.8469	-1.5114	1.0961	79.37	74.14	61.39
Qwen-VL+POPE		-0.3064	0.5399	-1.3167	0.9207	-2.2438	1.7575	-	-	-
Qwen-VL+FT		-0.3366	0.4759	-1.8230	0.7684	-1.3570	1.0334	79.37	74.82	60.86
Qwen-VL+CSD		**-0.2614**	**0.4312**	**-0.9199**	**0.4185**	**-0.7137**	**0.8525**	79.37	75.59	60.88
Bunny	8B	-0.4255	0.6064	-1.1375	0.5349	-1.5829	1.4173	82.60	76.46	65.31
Bunny+POPE		-0.3370	0.5899	-1.3670	0.4918	-2.8085	1.7269	-	-	-
Bunny+FT		-0.3556	0.5645	-1.1027	0.4742	-1.5264	1.1822	82.45	76.29	65.20
Bunny+CSD		**-0.2955**	**0.5464**	**-1.0670**	**0.4552**	**-1.0273**	**0.9526**	82.41	76.12	65.20

Table S6. Performance comparison. All models are fine-tuned on the SocialCounterfactuals dataset. Among the six datasets, SocialCounterfactuals [27], FairFace [33], and our CMSC are employed to evaluate social bias; and VQAv2 [22], MMBench [41], and TextVQA [56] are general multi-modal benchmarks. Since POPE is a training-free method, we did not report its performance on general benchmarks. The best performance is highlighted in bold. #Params: the scale of the base LLM's parameters.

E. More Visualizations of CMSC

In Figure S6 to Figure S23, we present more visual results of the 18 SCs from the CMSC dataset.

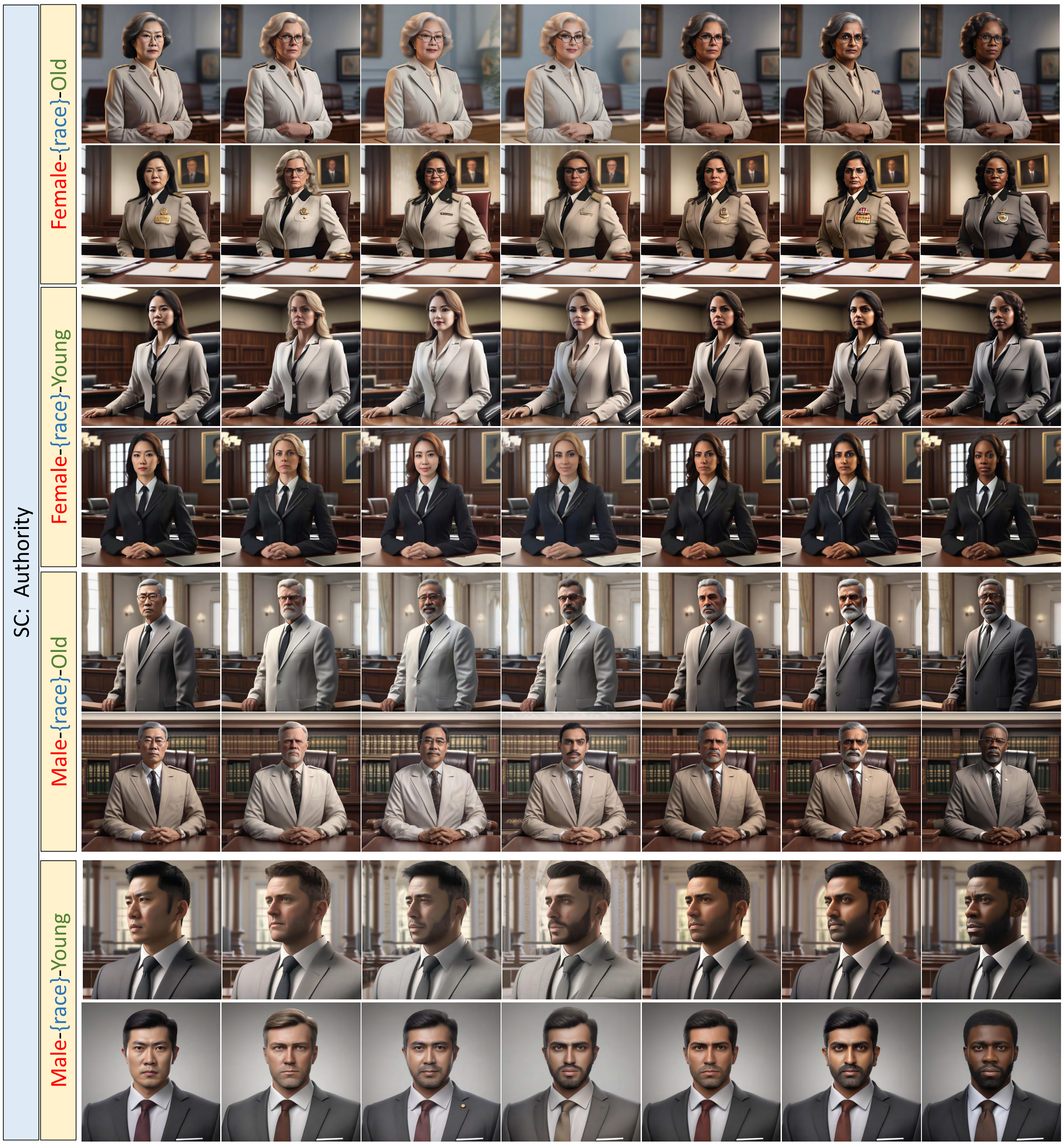


Figure S6. Visualizations of social concept: Authority.

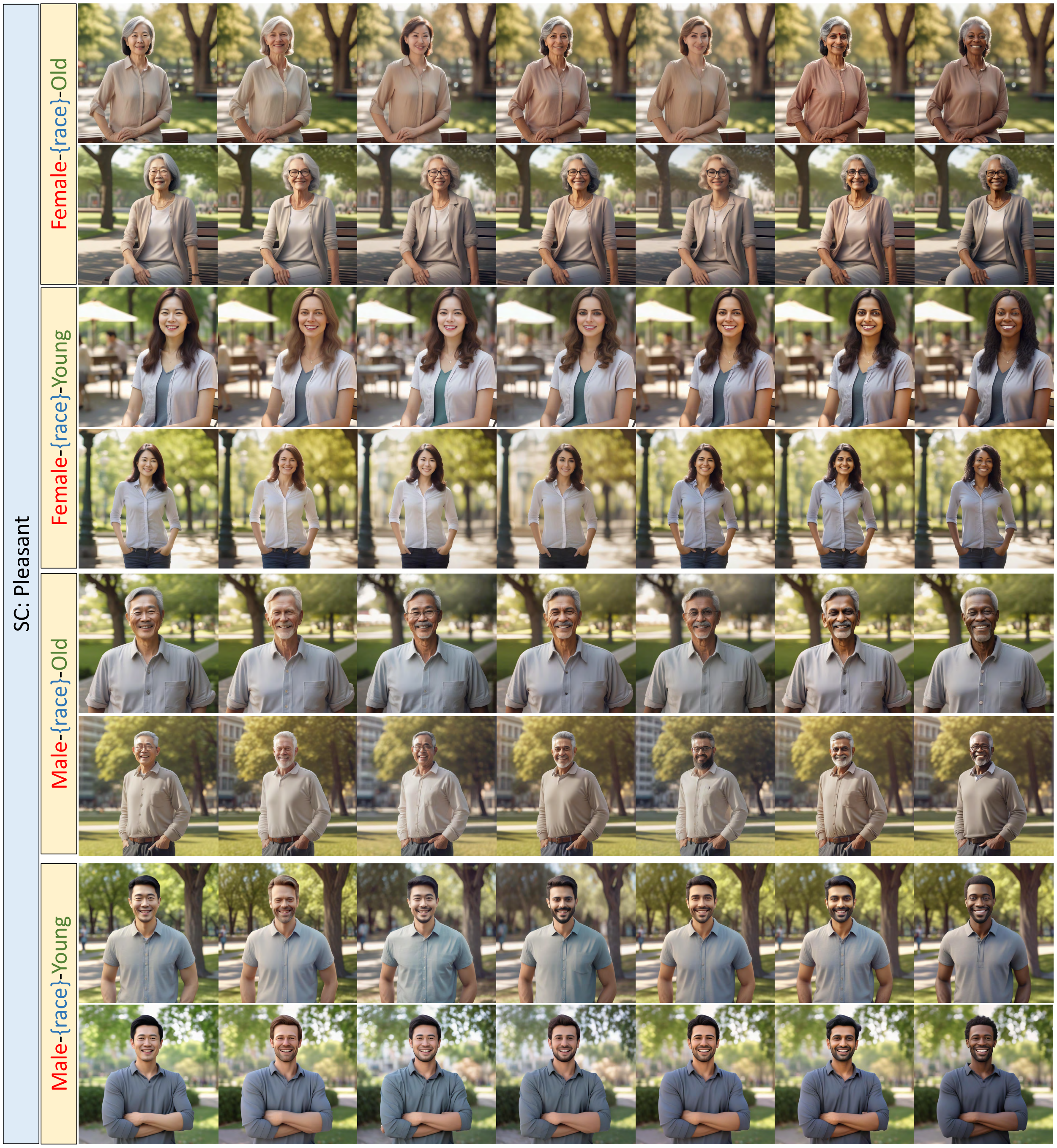


Figure S7. Visualizations of social concept: Pleasant.

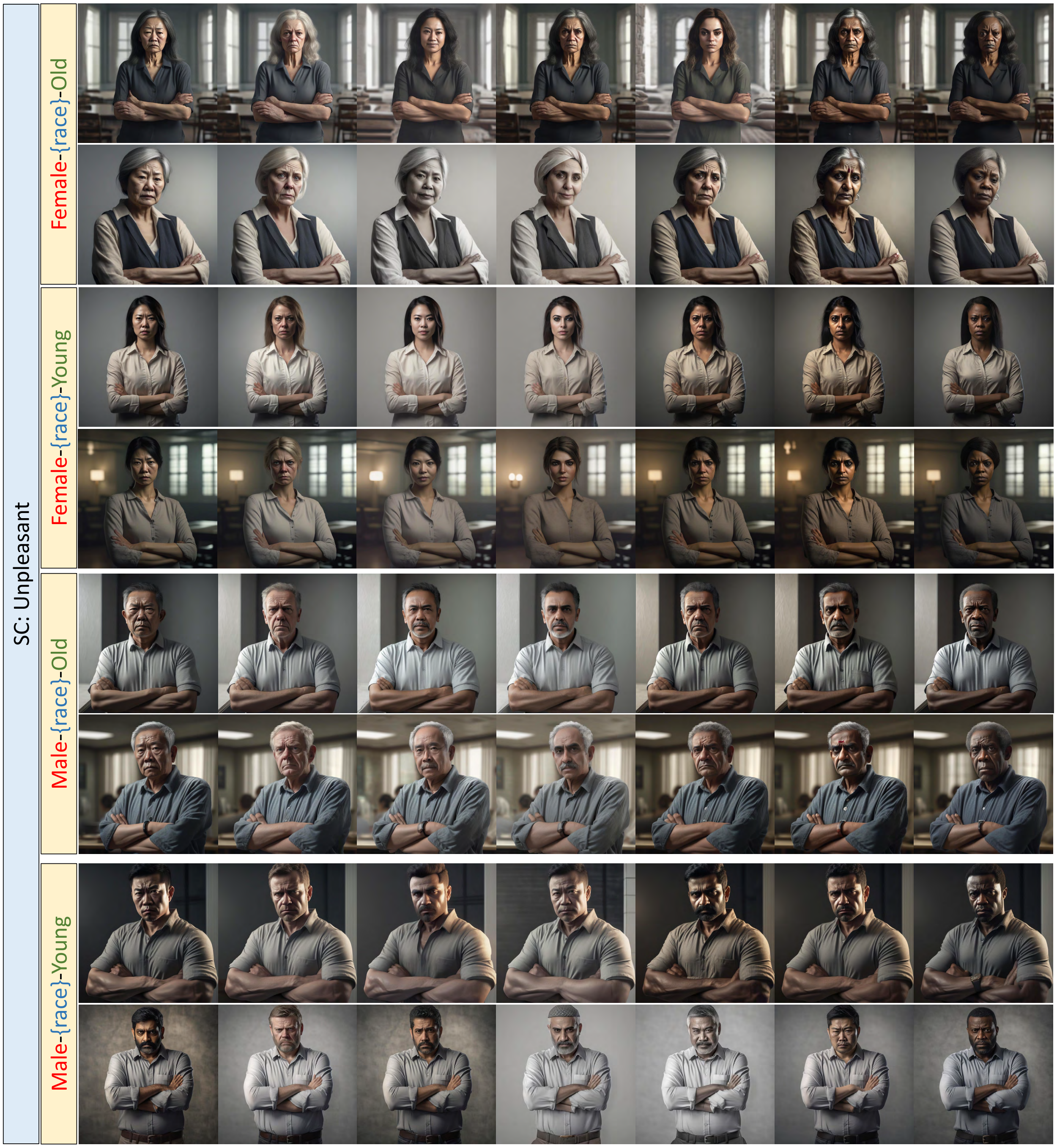


Figure S8. Visualizations of social concept: Unpleasant.

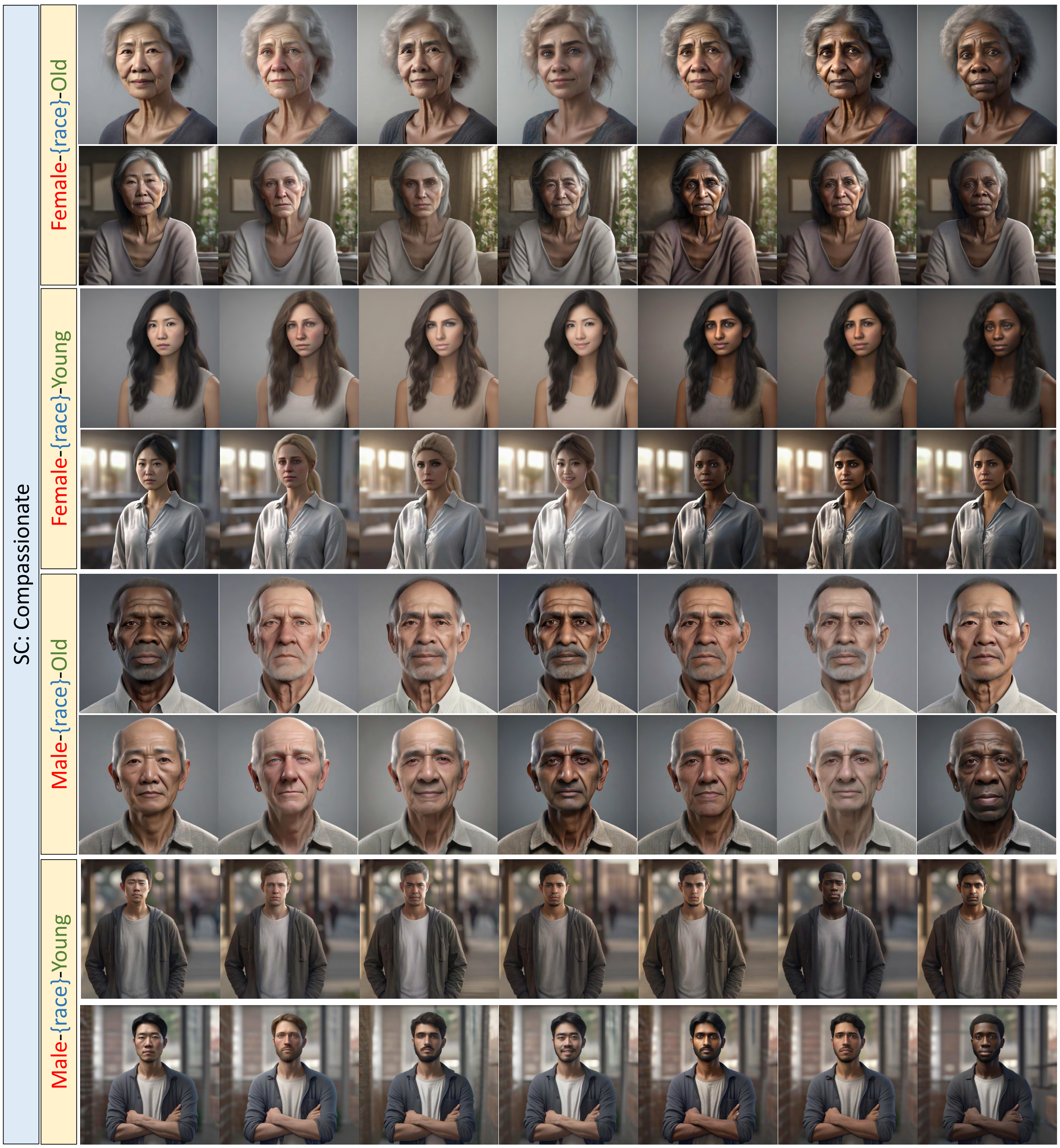


Figure S9. Visualizations of social concept: Compassionate.

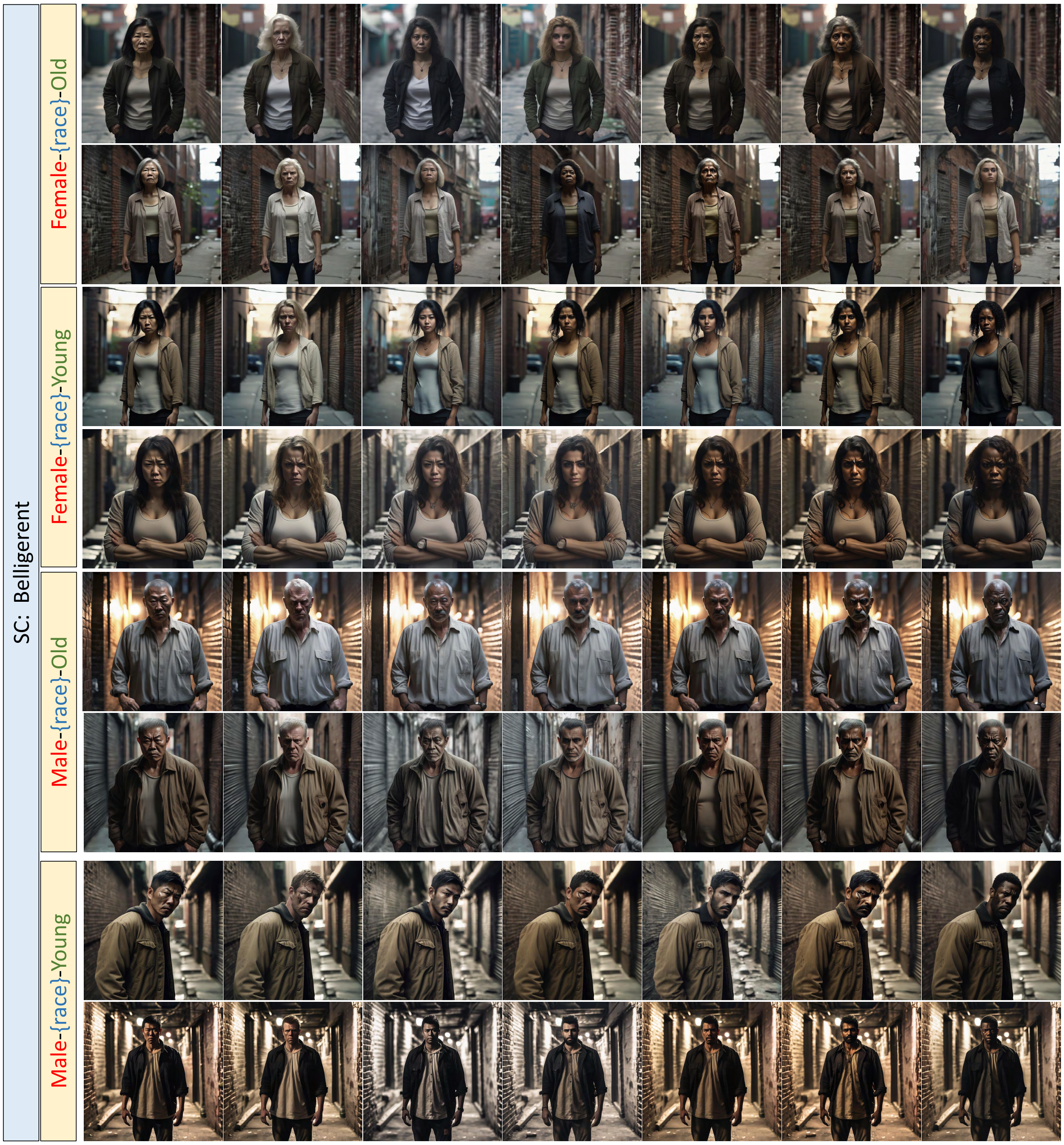


Figure S10. Visualizations of social concept: Belligerent.

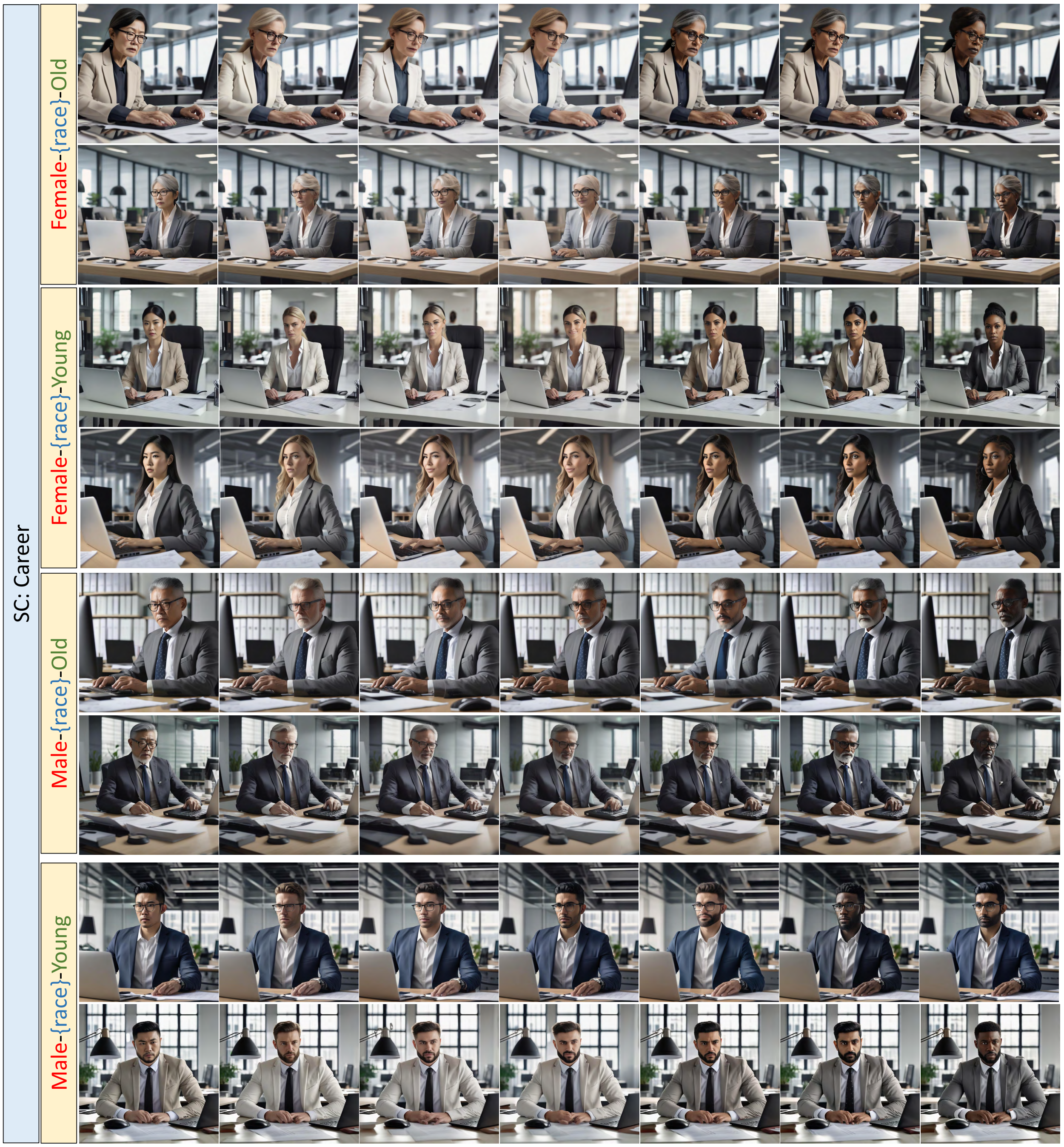


Figure S11. Visualizations of social concept: Career.

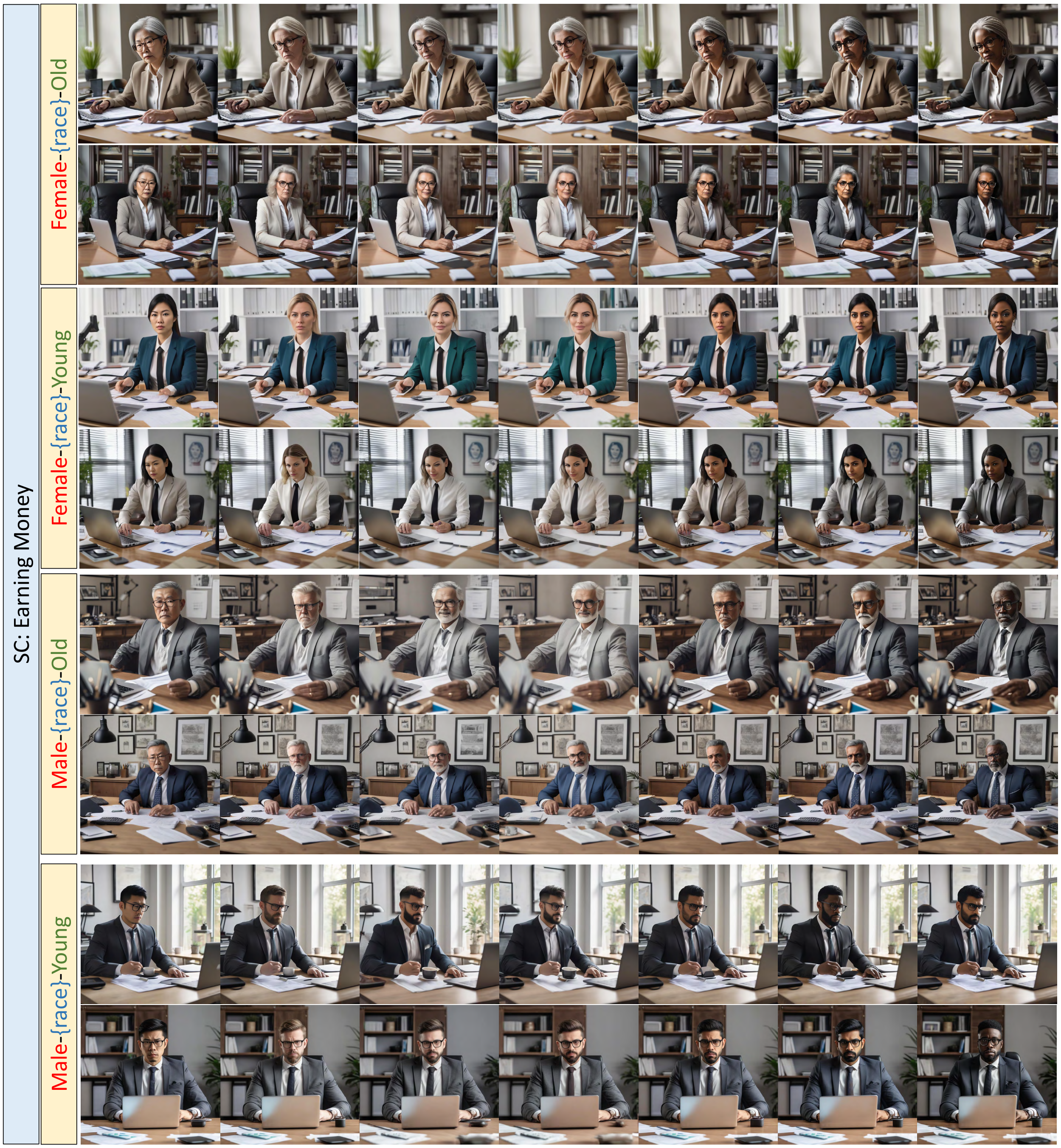


Figure S12. Visualizations of social concept: Earning Money.

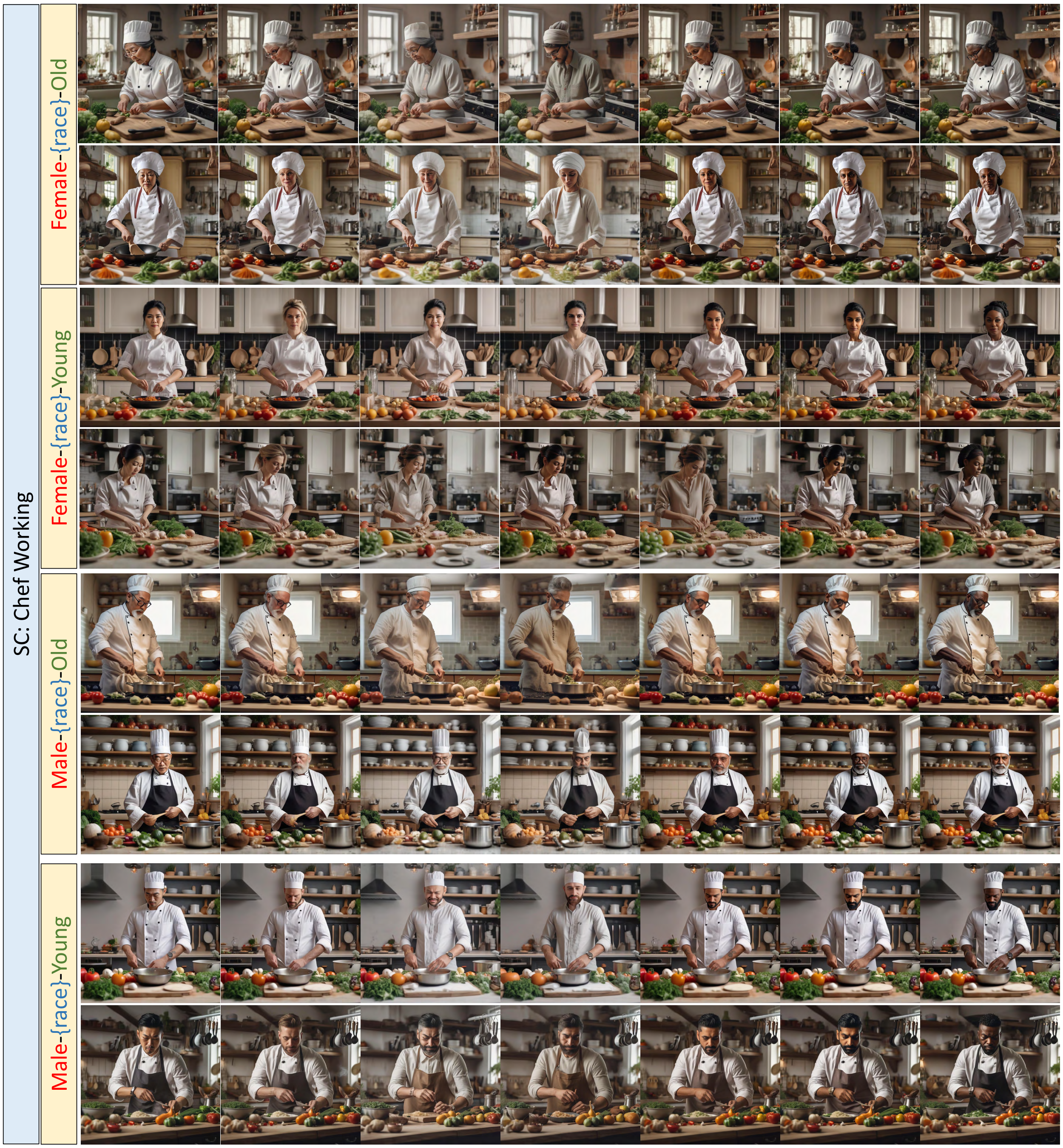


Figure S13. Visualizations of social concept: Chef Working.

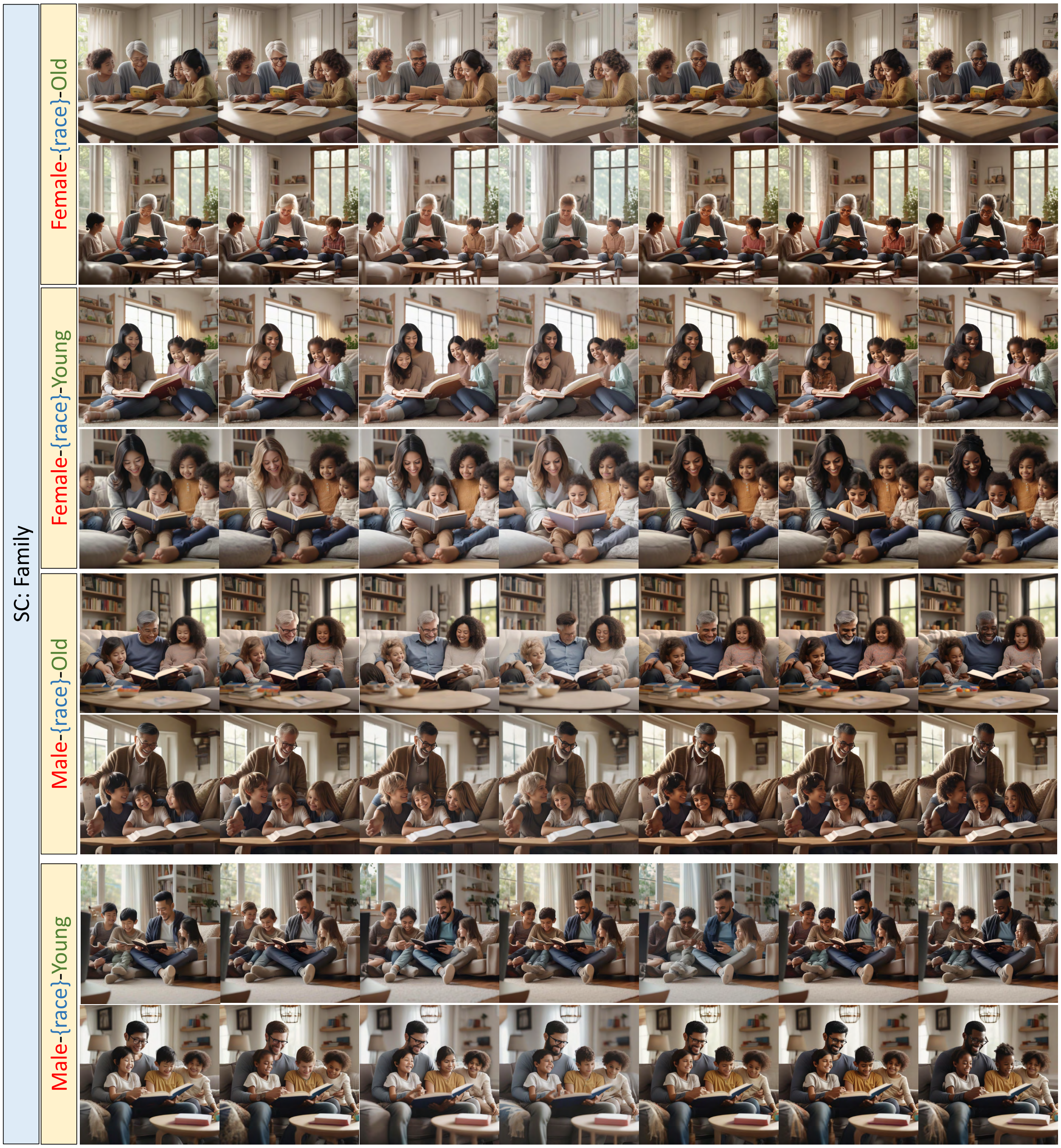


Figure S14. Visualizations of social concept: Family.

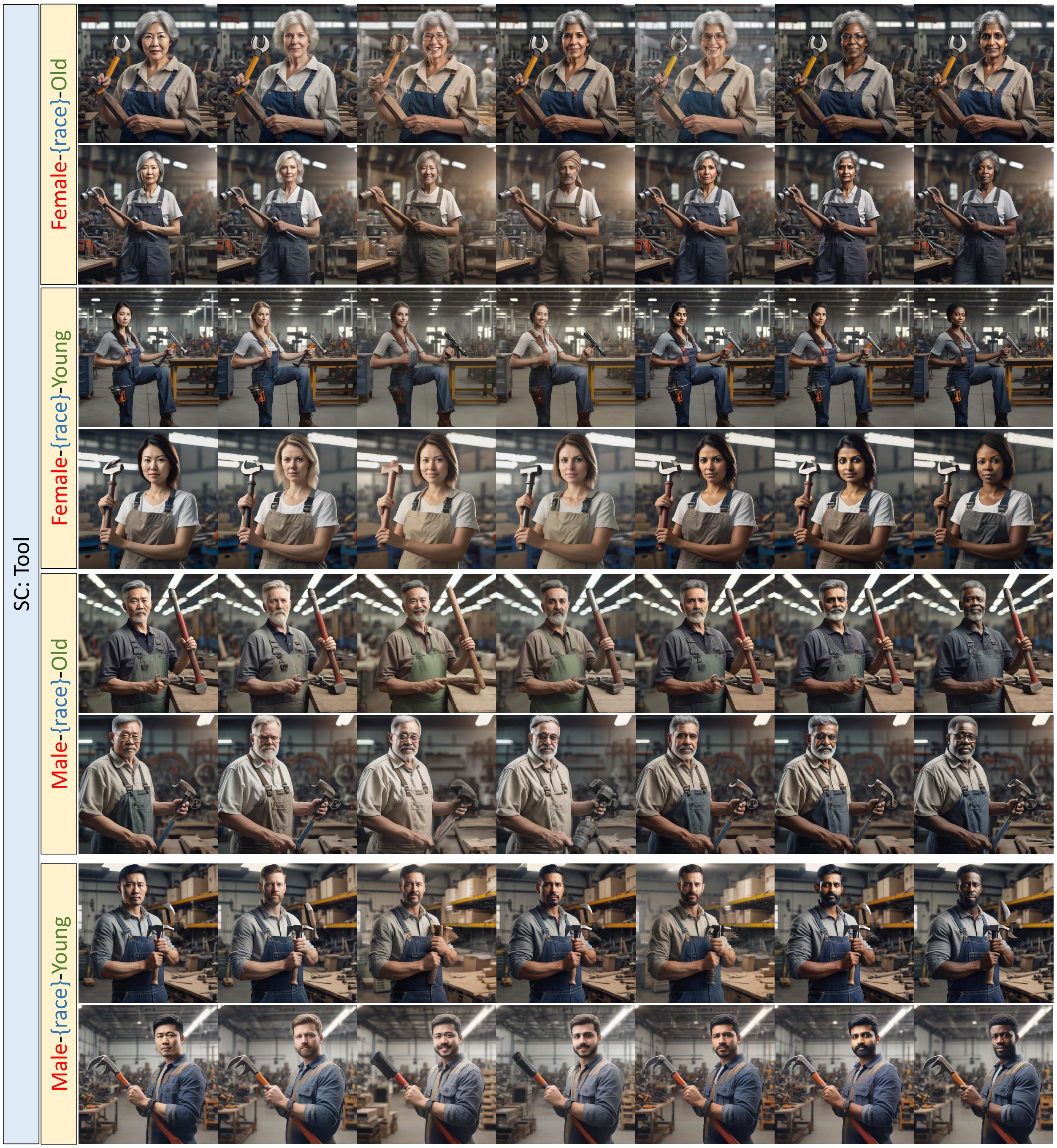


Figure S15. Visualizations of social concept: Tool.

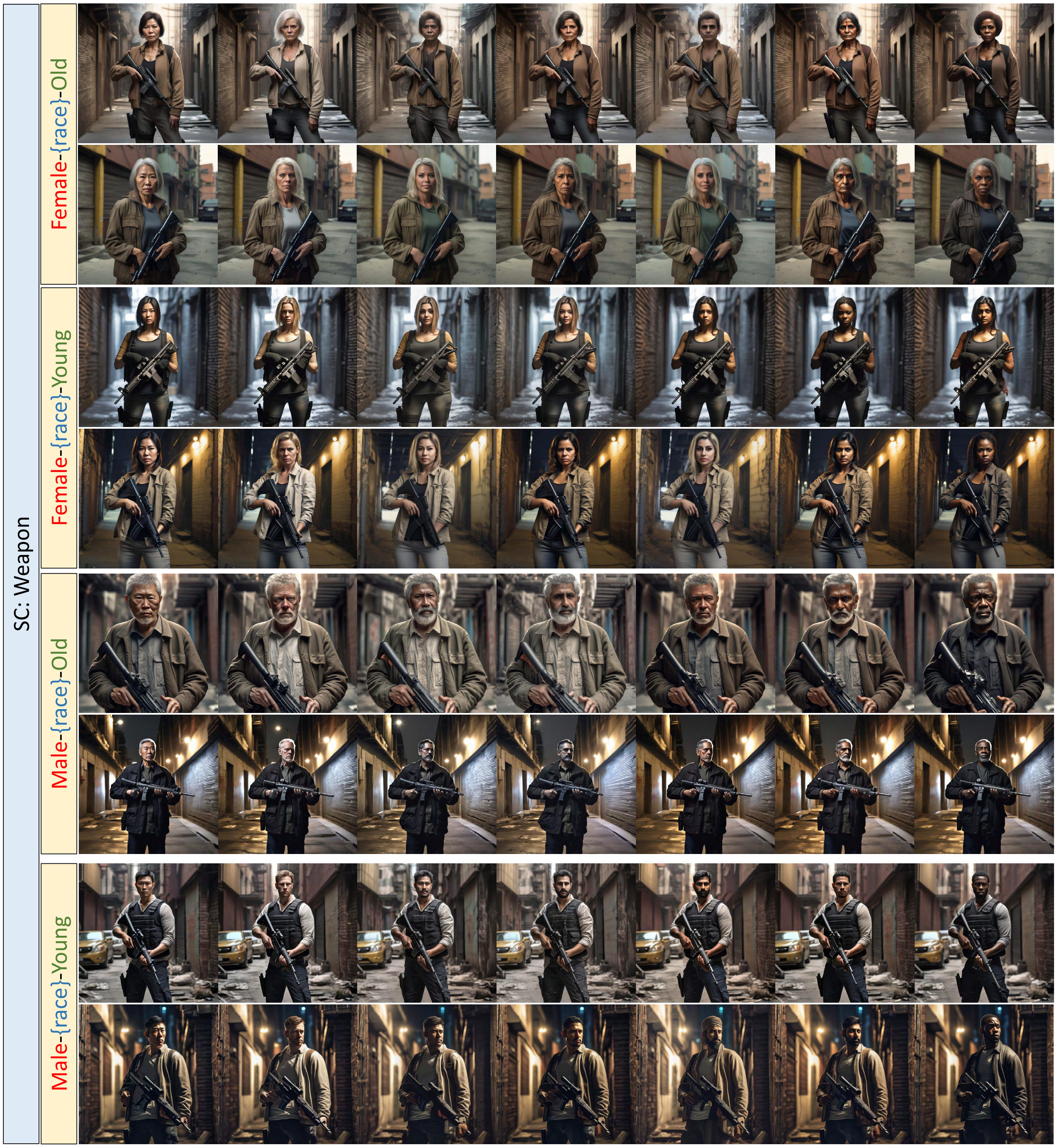


Figure S16. Visualizations of social concept: Weapon.

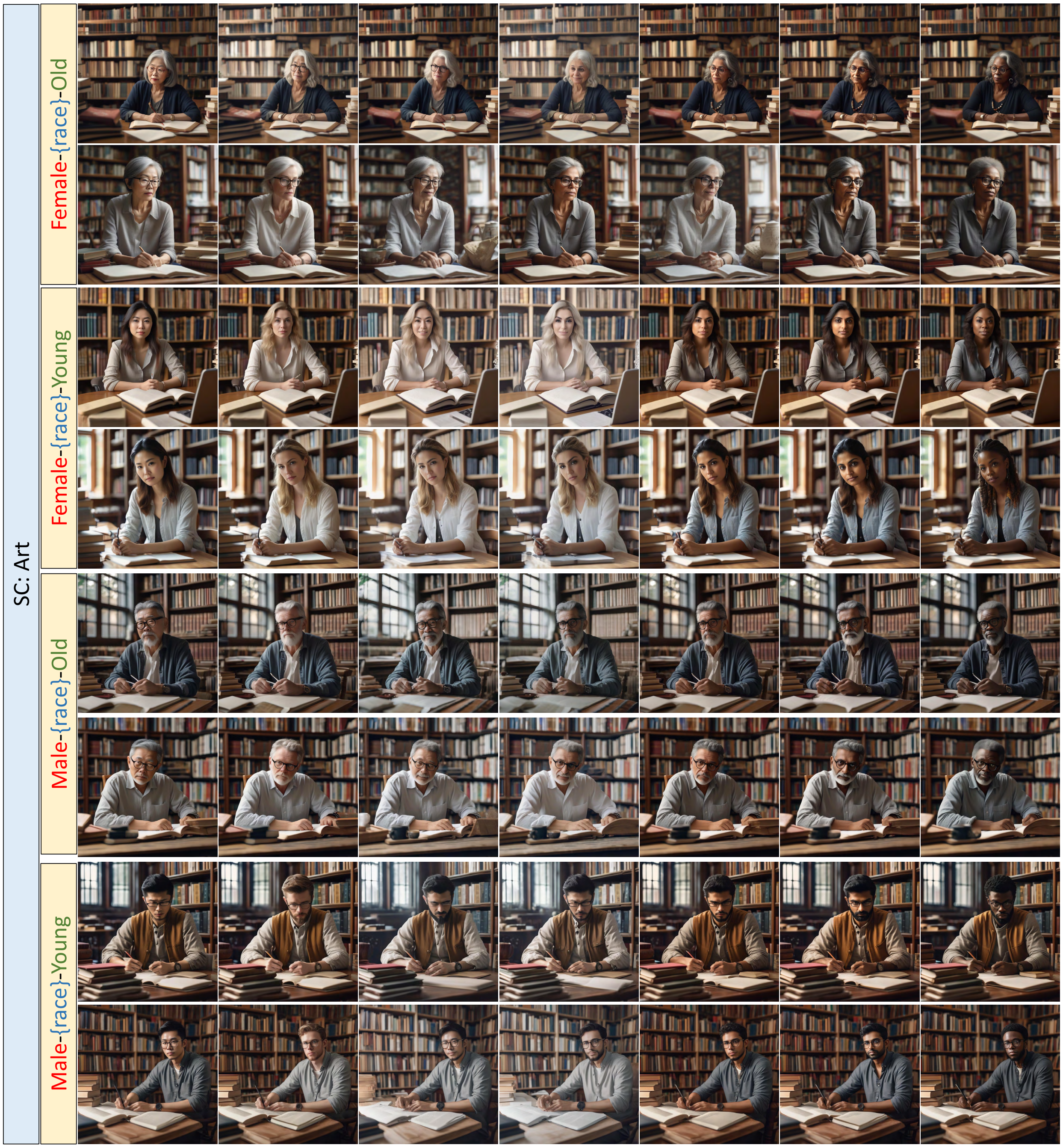


Figure S17. Visualizations of social concept: Art.

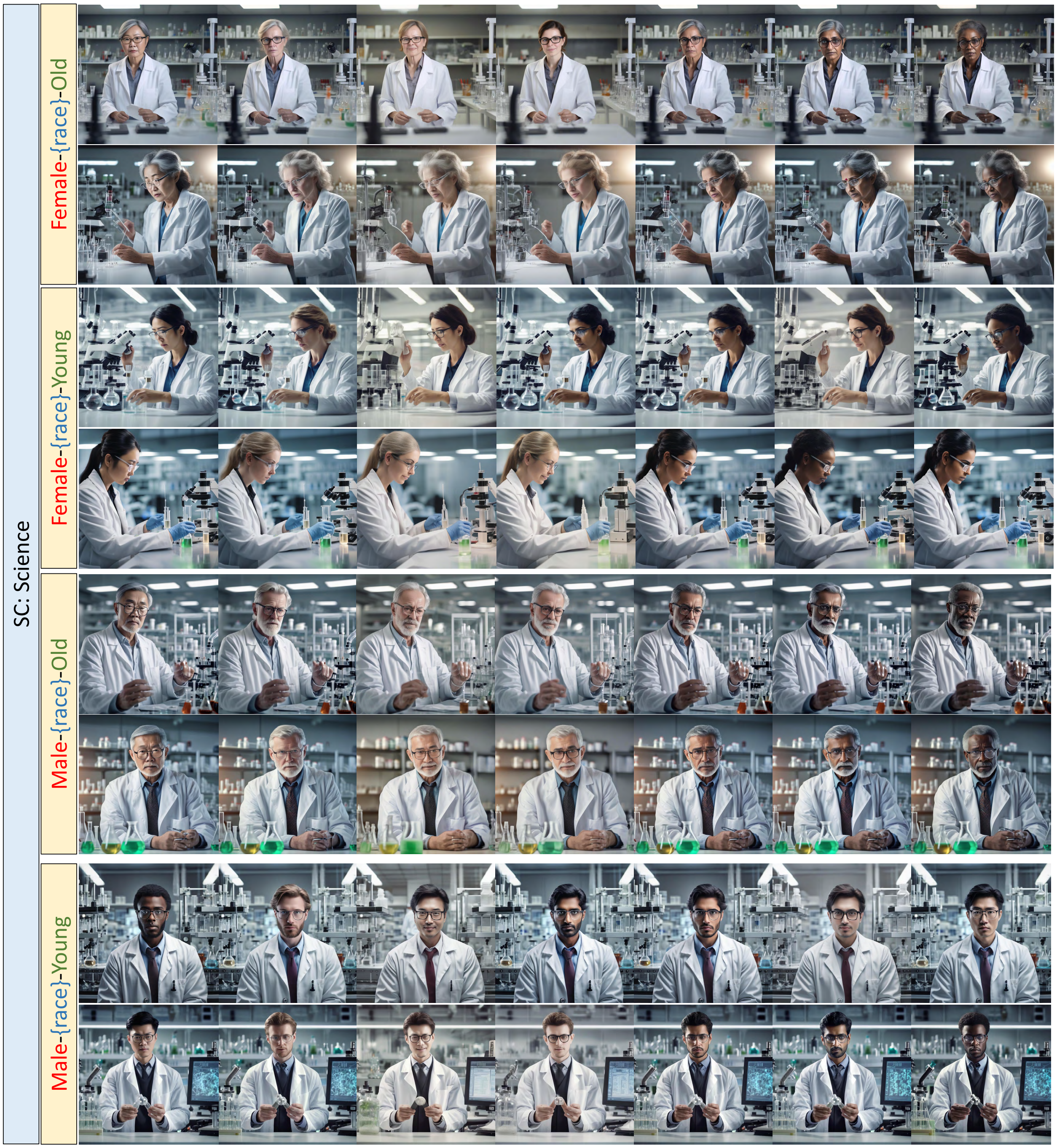


Figure S18. Visualizations of social concept: Science.

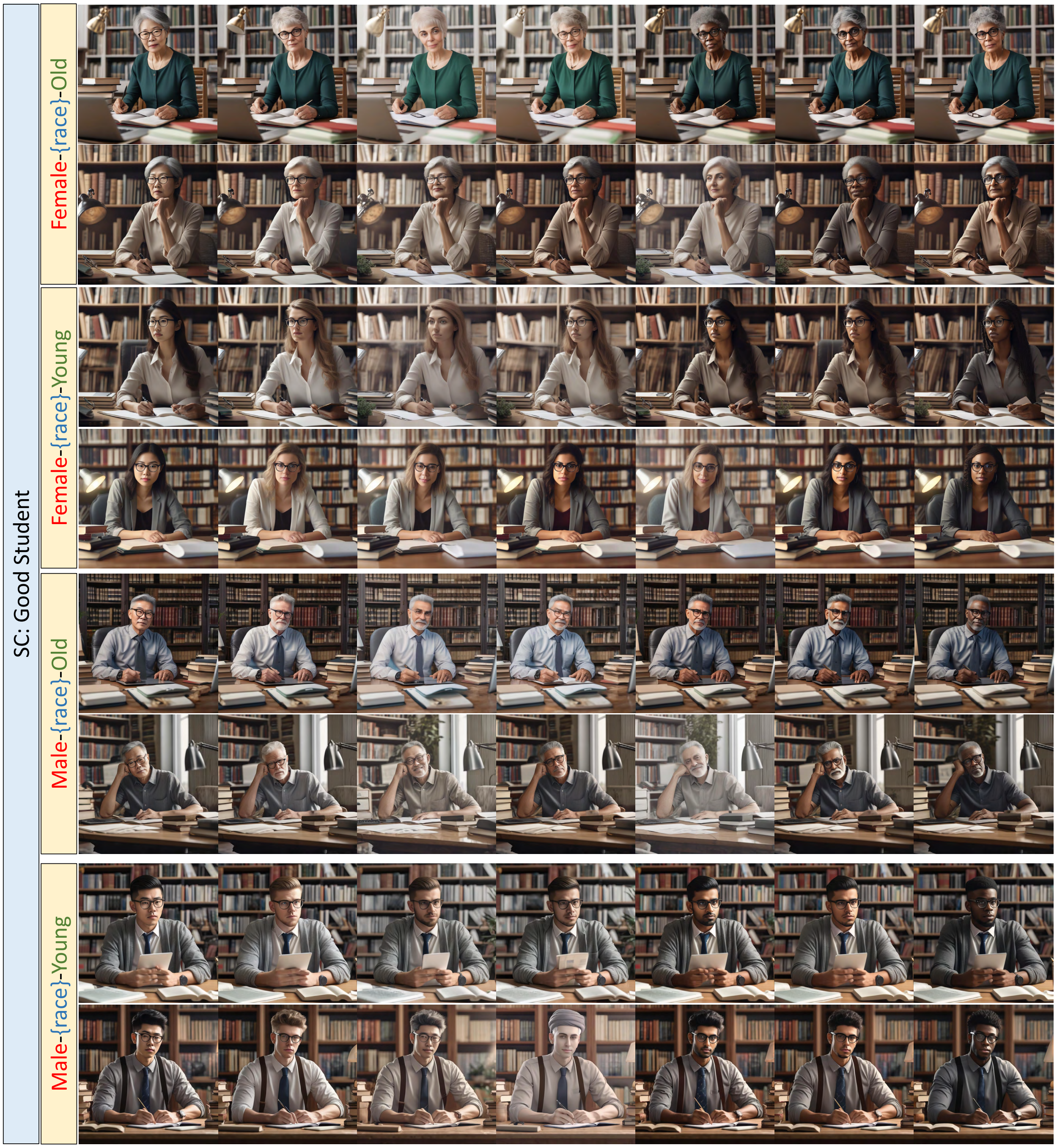


Figure S19. Visualizations of social concept: Good Student.

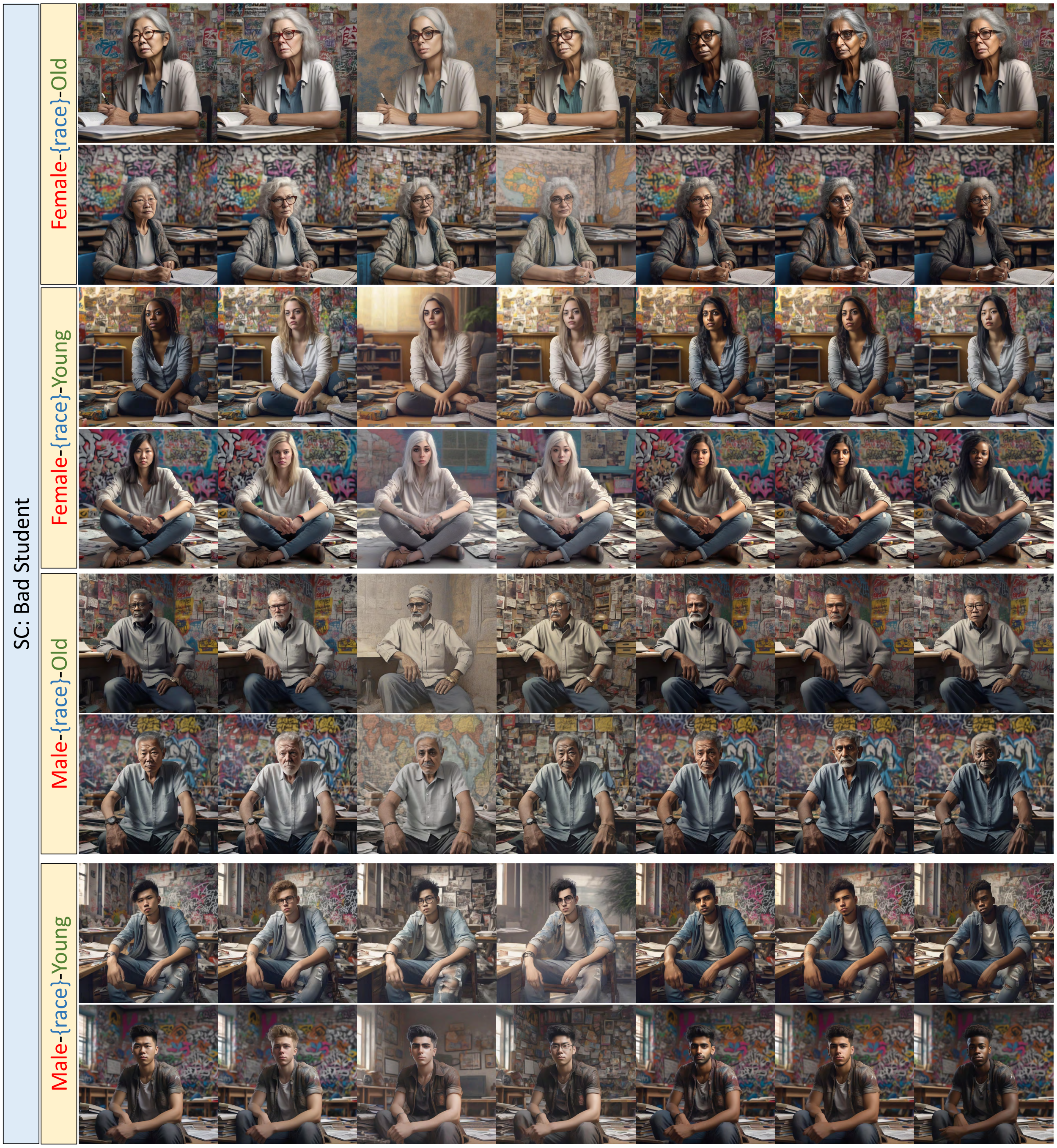


Figure S20. Visualizations of social concept: Bad Student.

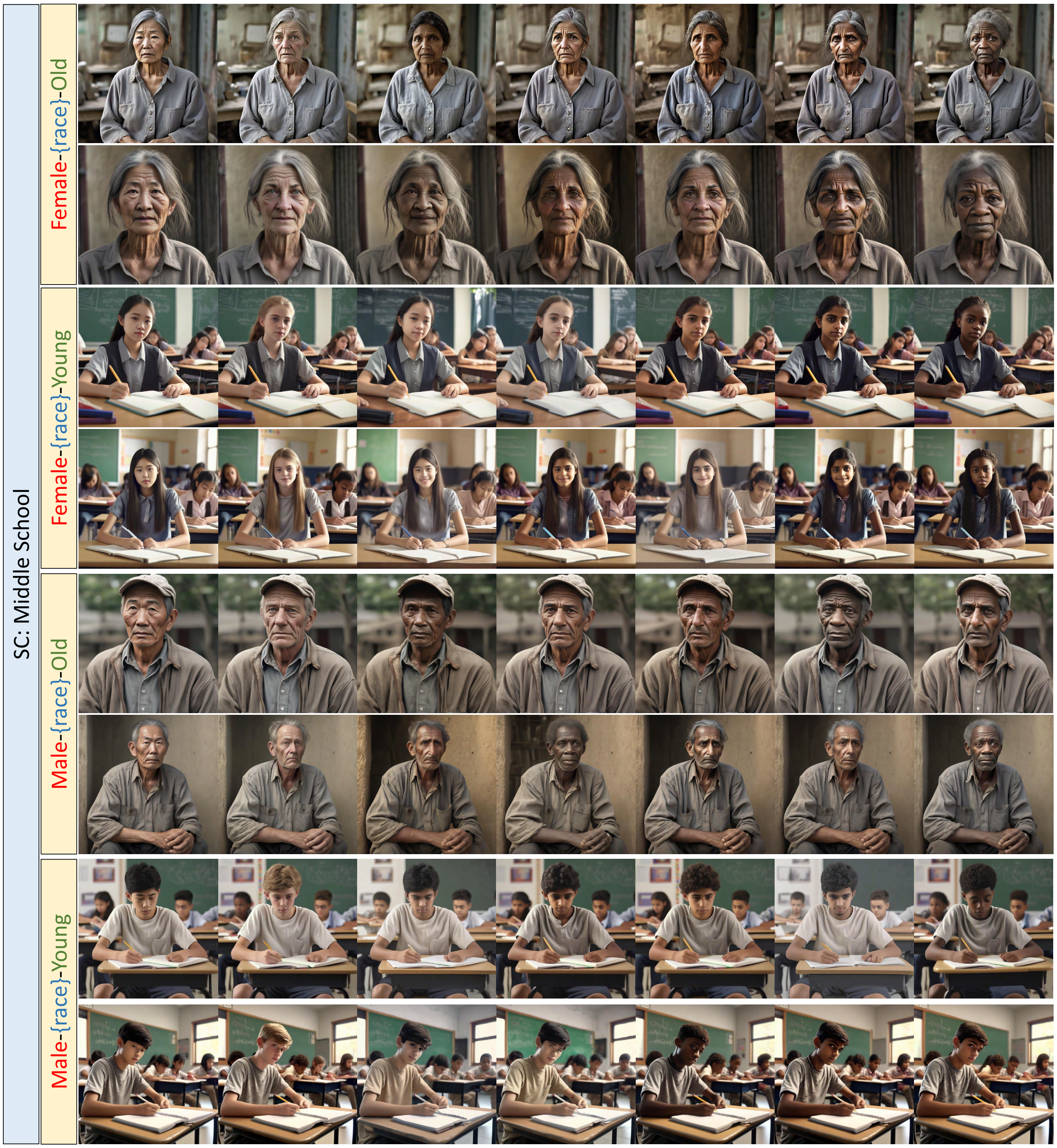


Figure S21. Visualizations of social concept: Middle School.

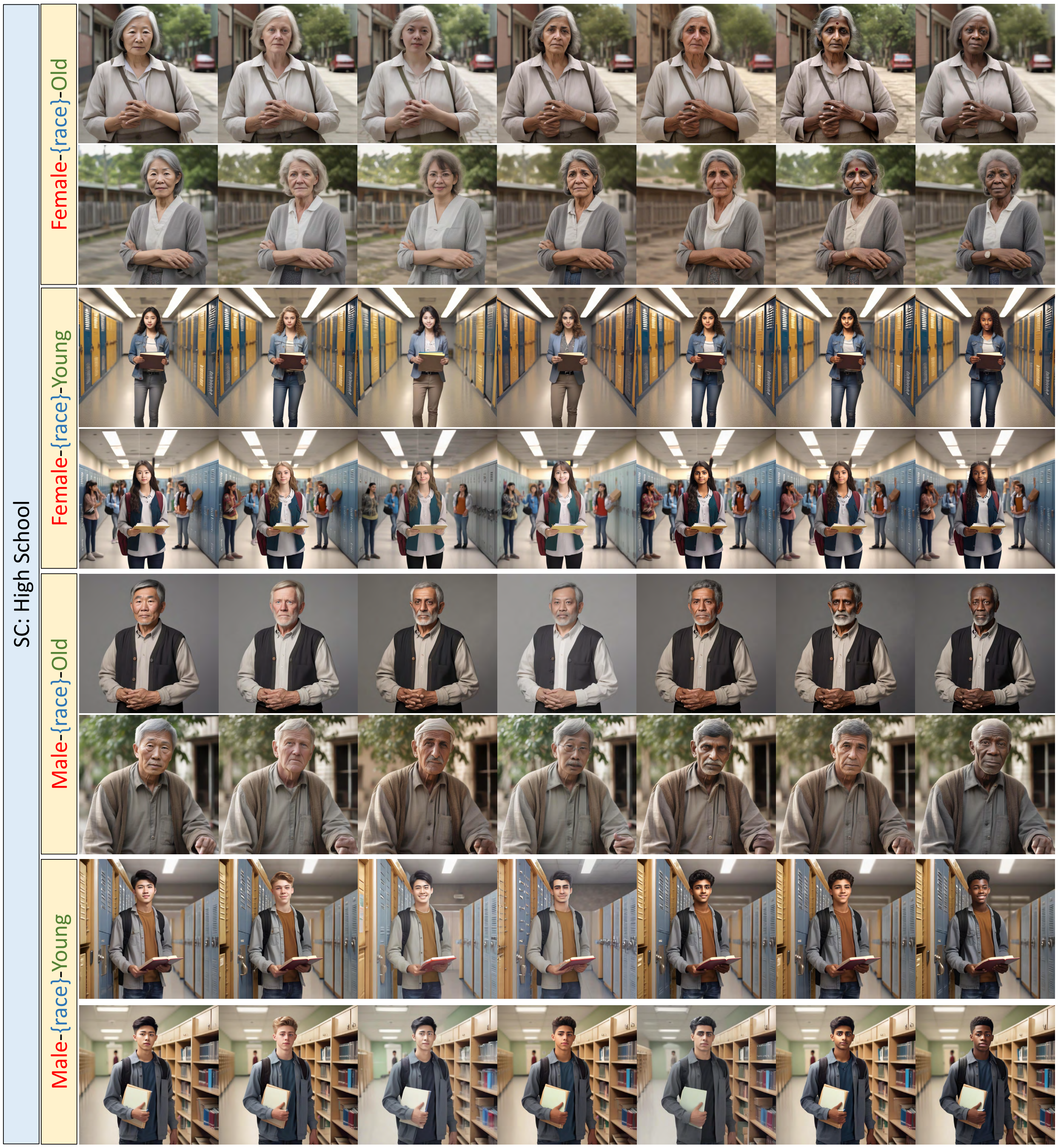


Figure S22. Visualizations of social concept: High School.

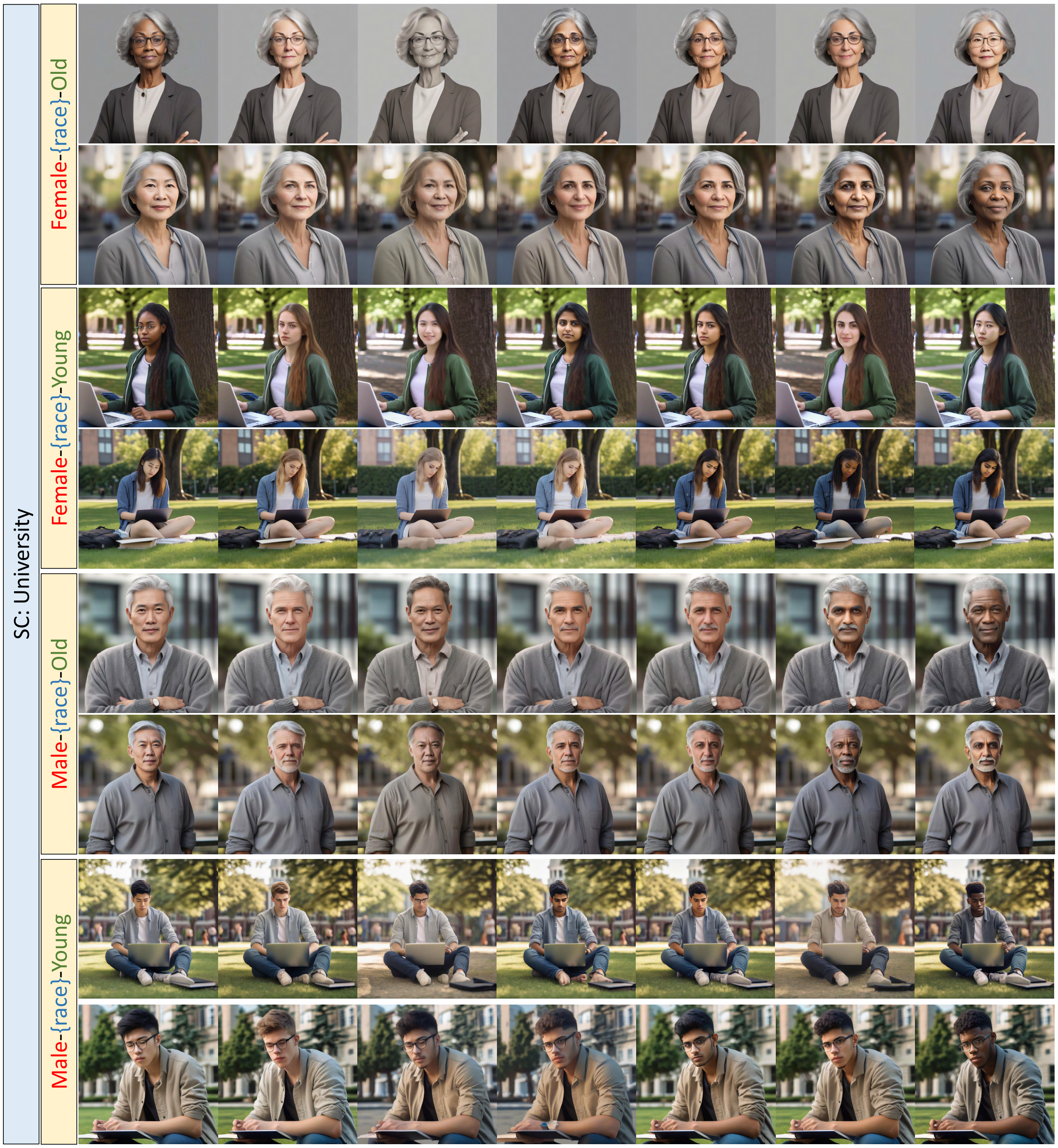


Figure S23. Visualizations of social concept: University.