Gender Bias in Vision-Language In-Context Learning

Tong Xiang, Yuta Nakashima, and Noa Garcia

The University of Osaka, Japan
tongxiang@is.ids.osaka-u.ac.jp, n-yuta@im.sanken.osaka-u.ac.jp,
noagarcia@ids.osaka-u.ac.jp

Abstract. In-context learning (ICL) enables large vision-language models (LVLMs) to perform tasks by following patterns in context examples, yet its potential to amplify societal biases remains underexplored. We systematically investigate how ICL influences gender bias in LVLMs through VL-BICLE, an evaluation framework comprising six ICL settings, three tasks, and four datasets. Our experiments on six LVLMs reveal that gendered ICL demonstrations act as a directional force, shifting model bias toward the demonstrated gender through a cross-gender mechanism that disproportionately degrades performance on the opposite gender. This effect appears in image captioning and pronoun prediction but not in visual question answering, indicating that gendered ICL influences bias only when the task output involves gendered language. Similarity-based retrieval methods inherit the training pool's gender imbalance and offer no debiasing advantage, while standard quality metrics remain blind to these bias shifts. To mitigate this bias, we replace real in-context images with synthetic ones from stable diffusion models while keeping captions unchanged. This simple intervention reduces gender bias without degrading caption quality. The code can be found at: https://github.com/mathfather/Gender-Bias-in-VL-ICL.

Keywords: Gender Bias · In-context Learning · Large Vision-Language Model

1 Introduction

Large Vision-Language Models (LVLMs) are multimodal extensions of Large Language Models (LLMs) [33]. Recent LVLMs [36,55] can take interleaved image-text data as input [1], enabling in-context learning (ICL) which adapts model outputs based on a few demonstration examples without fine-tuning [9, 12]. A substantial body of work [9,50,60,64] has explored ICL for vision-language tasks such as image captioning [46] and visual question answering (VQA) [3]. However, prior work has highlighted gender bias in LVLMs, raising concerns about their real-world impact [21]. Furthermore, biased contexts are found to exacerbate these issues by amplifying underlying model biases during inference [11]. These findings lead to the question: *How does ICL influence gender bias in LVLMs?*

To address this question, we propose **V**ision-**L**anguage Gender **B**ias in **ICL** **E**valuation (VL-BICLE), a systematic evaluation framework comprising four gender-composition and two similarity-based ICL settings, evaluated across three vision-language tasks (image captioning, pronoun prediction, and VQA) on four datasets. Experiments on six LVLMs reveal several key findings. Gendered ICL demonstrations act as a directional force: male-only context pushes models toward male bias[1] while female-only context pushes toward female bias, regardless of the model's baseline bias direction, sometimes flipping it entirely. This shift stems from a cross-gender mechanism where gendered demonstrations disproportionately degrade performance on the opposite gender. Similarity-based retrieval methods inherit demographic imbalances from their source pool and offer no debiasing advantage over random selection. Standard caption quality metrics remain stable even as gender bias shifts substantially. These gendered ICL effects, observed in captioning and pronoun prediction, are absent in VQA, indicating that ICL influences gender bias only when the task output involves gendered language. Finally, by replacing real demonstration images with synthetic ones from stable diffusion models (SDMs), we explore a simple yet effective gender bias mitigation method, which works across most models without degrading caption quality and is consistent across two different SDMs.

2 Related Work

Bias Mitigation in LVLMs Societal bias has been observed across vision-and-language tasks including image captioning [2, 15, 19, 44, 61], text-to-image search [47], and VQA [18], with studies showing that multimodal models can perpetuate stereotypes [24] and that LVLMs exacerbate gender bias in downstream tasks [17]. Several mitigation approaches have been proposed [7, 22, 47, 57], yet most target Vision-Language Models (VLMs) such as CLIP [40] rather than LVLMs. The few LVLM-focused methods rely on post-hoc logit calibration [58] or inference-time steering [41], but their effectiveness remains inconsistent across tasks [35, 56].

ICL for LVLMs ICL [49], originally used in natural language tasks [9], is defined as *a paradigm that allows language models to perform tasks given only a few examples in the form of demonstrations* [12]. Inspired by the advances of LLMs in ICL [50, 51, 60], researchers have begun to extend the idea to other modalities. Some initial studies explored ICL in purely visual settings, where vision models perform visual tasks (*e.g.*, segmentation, detection) via image inpainting, conditioned on a few image demonstrations without task-specific fine-tuning [6, 59]. More recently, with the proposal of multimodal models enabled with vision-language ICL [1, 32, 65] and trained on interleaved image-text datasets [5, 27, 28, 62], increasing efforts have been devoted to enhancing the ICL performance in these LVLMs [13, 64]. Still, how ICL impacts LVLMs during generation is not yet well understood [5, 31]. While a few efforts have been devoted to

[1] We refer to the case where an LVLM favors male samples over female ones as *male bias*, and the reverse as *female bias*.

optimizing in-context sequences to enhance ICL performance, such as retrieving representative examples [29, 54] or training a small assistant model for sample selection and ranking [53], these approaches do not aim to mitigate the societal biases present in LVLMs. To the best of our knowledge, this is the first study to analyze how ICL affects gender bias in LVLMs.

3 VL-BICLE

We introduce VL-BICLE (**V**ision-**L**anguage Gender **B**ias in **ICL E**valuation), a systematic pipeline for analyzing how ICL shapes gender bias in LVLMs.

3.1 Task Formalization

We conduct experiments on tasks that share a common paradigm: the input comprises both image and text, and the prediction is obtained from the model's generated output, either as a token sequence or as a next-token probability distribution.

Formally, given a dataset $\mathcal{D}_t = \{(I_i, y_i)\}_{i=1}^{N}$ for task t, containing N image-text pairs where I_i denotes the i-th image and y_i denotes the ground-truth label, a pretrained LVLM $\mathcal{M}$ can directly take a query image I_i as input conditioned on a task-specific instruction p_t and generate an output $\hat{y}_i$, without requiring any task-specific training. This is the zero-shot setting, formalized as:

$$\hat{y}_i \leftarrow P_{\mathcal{M}}(\hat{y}_i \mid I_i, p_t). \tag{1}$$

Here $\leftarrow$ denotes a decoding strategy, *e.g.*, greedy search. In the ICL (few-shot) setting, a few examples, usually sampled from $\mathcal{D}_t$, are prepended to the input to improve performance on downstream tasks. We formalize a k-shot ICL as:

$$\hat{y}_i \leftarrow P_{\mathcal{M}}(\hat{y}_i \mid \mathcal{S}_t^k, I_i, p_t), \tag{2}$$

where $\mathcal{S}_t^k = \{(I_1, y_1), \ldots, (I_k, y_k)\}$ denotes a sequence of in-context demonstrations sampled from $\mathcal{D}_t$. By systematically varying $\mathcal{S}_t^k$, we isolate how context impacts gender bias.

3.2 Evaluation Framework

To evaluate gender bias, we employ datasets of natural images paired with gender labels,[2] formally represented as $\mathcal{D}_t = \{(I_i, y_i, g_i)\}_{i=1}^{N}$ where $g_i \in \{\text{male}, \text{female}\}$, and design four selection strategies for constructing $\mathcal{S}_t^k$:

1. **Random Sample (RS)**: RS constructs $\mathcal{S}_t^k$ by uniformly sampling image-text pairs from $\mathcal{D}_t$. This selection method ensures that the selected examples follow a similar distribution to $\mathcal{D}_t$, and has been commonly used as a baseline in prior work [53, 59].

[2] We use binary gender categories following previous work [15, 20, 63]; we acknowledge that this does not capture the full spectrum of social identities.

2. **Male-only Sample (MS)**: MS constructs $\mathcal{S}_t^k$ by sampling only from the male-presenting subset $\{(I_i, y_i, g_i), g_i = \text{male}\}$ of $\mathcal{D}_t$.
3. **Female-only Sample (FS)**: FS constructs $\mathcal{S}_t^k$ by sampling only from the female-presenting subset $\{(I_i, y_i, g_i), g_i = \text{female}\}$ of $\mathcal{D}_t$.
4. **Balanced Sample (BS)**: BS randomly selects an equal number $(k/2)$ of male-presenting and female-presenting examples. The selected examples are then interleaved in an alternating gender order to construct $\mathcal{S}_t^k$. Here $\mathcal{S}_t^k$ always starts with a male example and ends with a female one.

This design tests whether attribute-consistent contexts amplify pre-existing gender bias by reinforcing group-specific priors, and whether balanced contexts can mitigate such bias. We additionally compare two similarity-based retrieval (SBR) methods introduced in previous work [29, 53] as our baselines:

5. **Similarity-based Image-Image Retrieval (SIIR)**: SIIR selects k image-text pairs from $\mathcal{D}_t$ with the highest image-to-image cosine similarity to the query image, computed between image features from a frozen CLIP model [40].
6. **Similarity-based Image-Text Retrieval (SITR)**: SITR selects the top-k image-text pairs whose *text* is most similar to the query image, ranked by cosine similarity between the query image features and textual features, both obtained from a frozen CLIP model.

For both SIIR and SITR, the more similar samples are placed closer to the query. Although SBRs have not been previously applied to gender bias analysis in LVLMs, these methods are among the few approaches explored for in-context example selection in vision-language tasks [53], and have demonstrated effectiveness in enhancing ICL performance. The ICL settings are illustrated in Fig. 1.

3.3 Datasets

We present the four datasets used in VL-BICLE, summarized in Tab. 1. All datasets are filtered to single-person images with explicit gender labels (either manually annotated or deterministically inferred from the available text). Each dataset is split into a training set (used as the ICL demonstration pool) and a validation set (used for evaluation): the training set preserves the original gender distribution, while the validation set is down-sampled for gender balance. Notably, all four datasets elicit natural outputs (free-form captions or unconstrained next-token predictions), better reflecting real-world model behavior and avoiding artifacts introduced by constrained formats (*e.g.*, multiple-choice [30]). See Sec. A of the supplementary material for details on data processing.

VisoGender VisoGender [14] contains images of people in 23 occupational settings, each annotated with a perceived binary gender label. Two sub-tasks are defined: Occupation-Object (OO), a single-subject image paired with a possessive pronoun referring to a professional object (*e.g.*, *the doctor and his/her stethoscope*); and Occupation-Participant (OP), a two-subject image paired with

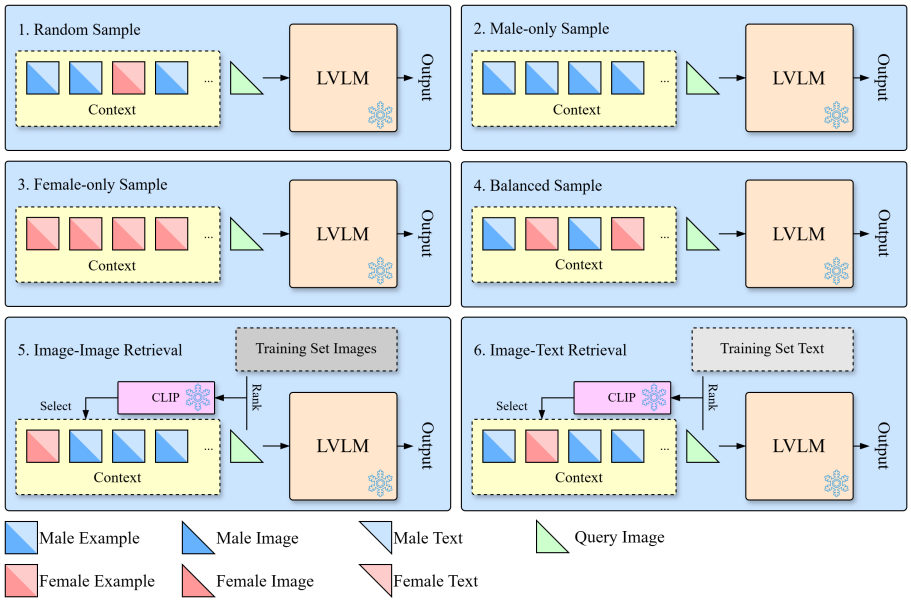

Fig. 1: The evaluation framework of VL-BICLE. All examples are image-text pairs; task-specific instructions are not shown in this illustration for simplicity and clarity. LVLM and CLIP models are frozen during inference.

a possessive pronoun referring to a secondary participant (*e.g.*, *the doctor and his/her patient*). During inference, given a query image and a template prefix (*e.g.*, *The doctor and* [`next-token`]), we obtain the next-token distribution and compare the probabilities of *his* and *her* to determine the predicted gender.

COCOBias We use the gender-annotated subset of MSCOCO [10] released by Zhao *et al.* [61], which we refer to as COCOBias. We filter the dataset to retain only single-person samples with unambiguous gender labels. During evaluation, models generate a free-form caption for each query image, and the predicted gender is inferred from the generated text using predefined gender word lists.

DCI Densely Captioned Images (DCI) [45] contains images from SA-1B with dense, human-annotated captions. Similar to COCOBias, we filter the dataset to contain only single-person images, yielding a subset of 371 images that is split into a training pool and a gender-balanced validation set. During evaluation, models generate a free-form caption for each query image, and the predicted gender is inferred using the same word lists as COCOBias.

VisualCoT VisualCoT [42] provides chain-of-thought (CoT) reasoning annotations for VQA; here we use its GQA [23] subset, which contains human images. Each ICL demonstration comprises an image, a question, an optional multi-step reasoning path, and a short answer (see Sec. C.3 of the supplementary material for the prompts). We conduct two types of inference: with CoT and without CoT (direct); inference without CoT directly outputs an answer given the input,

Table 1: Dataset overview. Training sets serve as ICL demonstration pools; validation sets are used for evaluation. All validation sets are gender-balanced.

Dataset	Task	Train			Valid		
		M	F	All	M	F	All
VisoGender-OO [14]	Pronoun prediction	46	46	92	69	69	138
VisoGender-OP [14]	Pronoun prediction	92	92	184	126	130	256
COCOBias [61]	Image captioning	538	267	805	1,065	1,065	2,130
DCI [45]	Image captioning	45	30	75	119	119	238
VisualCoT [42]	VQA	804	264	1,068	1,056	1,056	2,112

while inference with CoT produces intermediate reasoning steps before the final answer. To mitigate textual gender leakage during evaluation, test queries are neutralized (replacing explicit gender words with neutral ones, *e.g.*, *man* → *person*) while ICL demonstrations retain their original gendered phrasing, isolating the effect of ICL context from explicit gender cues.

3.4 Evaluation Metrics

Bias Metrics Given a protected attribute of interest, a natural strategy to quantify bias is to compute the performance disparity between its subgroups with respect to a chosen evaluation metric. We define gender-grouped error rates in Eq. (3) [25], where $\mathcal{D}_t^m$ and $\mathcal{D}_t^f$ denote the male-presenting and female-presenting subsets of $\mathcal{D}_t$, and $e_i \in [0, 1]$ is the per-sample error (1 indicating an error and 0 a correct prediction; continuous for VQA). ER_o denotes the overall error rate and $\mathrm{ER}_m - \mathrm{ER}_f$ denotes the error rate gap between gender subgroups. We use ER_{m-f} as the primary gender bias metric: $\mathrm{ER}_{m-f} > 0$ suggests a model makes more errors on male-presenting samples (female bias), and $\mathrm{ER}_{m-f} < 0$ suggests the opposite case (male bias); ER_{m-f} closer to zero indicates more equitable performance across gender subgroups.

$$\mathrm{ER}_m = \frac{1}{|\mathcal{D}_t^m|} \sum_{i \in \mathcal{D}_t^m} e_i, \quad \mathrm{ER}_f = \frac{1}{|\mathcal{D}_t^f|} \sum_{i \in \mathcal{D}_t^f} e_i$$
$$\mathrm{ER}_o = \frac{1}{|\mathcal{D}_t|} \sum_{i=1}^{|\mathcal{D}_t|} e_i, \quad \mathrm{ER}_{m-f} = \mathrm{ER}_m - \mathrm{ER}_f \ . \tag{3}$$

Performance Metrics For image captioning performance evaluation, we utilize the reference-based metric BLEU-4 [38, 39], which requires human-written captions as ground-truth references. We also apply a reference-free metric, CLIPScore [16], which relies on the image-text matching of the pre-trained CLIP [40].
Statistical Metrics We additionally report two statistics: (1) for image captioning, the average caption length (AvgL) provides a coarse measure of stylistic

Table 2: Zero-shot ER_{m-f} across all datasets and tasks. Negative numbers indicate male bias; positive numbers indicate female bias.

Dataset	QwenVL	MiniCPM	Idefics3	Phi35V	InternVL35	Qwen3VL
VisoGender-OO [14]	−2.90	−27.54	−1.45	13.04	−4.35	2.90
VisoGender-OP [14]	−8.89	−45.36	−8.10	35.91	−3.36	3.64
COCOBias [61]	−0.94	−0.09	−1.69	−0.09	0.00	0.19
DCI [45]	0.84	0.84	−1.68	0.00	−0.84	−0.84
VisualCoT [42]	–	−0.31	–	6.42	–	1.61

similarity to the reference captions; and (2) for image captioning and pronoun prediction, the reveal rate (RR) measures the proportion of samples in which the LVLM explicitly mentions gender. We define a per-sample binary indicator $r_i \in \{0, 1\}$, where $r_i = 1$ if the model output contains an explicit gender mention and $r_i = 0$ otherwise; gender-grouped reveal rates are then defined in Eq. (4):

$$\mathrm{RR}_m = \frac{1}{|\mathcal{D}_t^m|} \sum_{i \in \mathcal{D}_t^m} r_i, \quad \mathrm{RR}_f = \frac{1}{|\mathcal{D}_t^f|} \sum_{i \in \mathcal{D}_t^f} r_i, \quad \mathrm{RR}_o = \frac{1}{|\mathcal{D}_t|} \sum_{i=1}^{|\mathcal{D}_t|} r_i \tag{4}$$

A lower RR indicates a tendency toward gender-neutral captions (*e.g.*, generates *person* instead of *man*), which reduces the observable gender signals.

4 Experiments

LVLMs We examine how ICL impacts gender bias during generation across six widely used LVLMs: Qwen-VL (QwenVL) [4], Idefics3-8B-Llama3 (Idefics3) [26], Phi-3.5-vision-instruct (Phi35V) [34], MiniCPM-o 2.6 (MiniCPM) [55], InternVL3.5-8B (InternVL35) [48], and Qwen3-VL-8B-Instruct (Qwen3VL) [52]. These models support interleaved image-text inputs, which is necessary for evaluation under multimodal ICL settings. All LVLMs used have publicly available checkpoints. We evaluate all six models on all datasets, except for VisualCoT: here we evaluate only MiniCPM, Phi35V, and Qwen3VL, as only these three models are observed to have both consistent pattern-following ability and reasonable inference speed. More details regarding models can be found in Sec. D of the supplementary material.

Inference Details Our experiments are conducted under k-shot settings for $k \in \{0, 2, 4, 6, 8\}$. All experiments use greedy search during inference. The first four ICL settings (RS, MS, FS, BS) are applied to all datasets. SIIR and SITR are applied only to COCOBias: VisoGender and VisualCoT lack the descriptive captions that SITR requires, and thus we discard both datasets from retrieval methods; the demonstration pool of DCI is too small for meaningful similarity-based selection. For RS, MS, FS, and BS, we run each experiment five times using independently sampled in-context sequences $\mathcal{S}_t^k$ from the training set, where

$\mathcal{S}_t^k$ is obtained by extending $\mathcal{S}_t^{k-2}$ with two additional examples. For SIIR and SITR, since the in-context examples are retrieved deterministically, each setting is conducted only once. See Sec. E of the supplementary material for details.

Evaluation Details For COCOBias, the predicted gender is inferred from each generated caption via keyword matching against pre-defined gender word lists. A sample is counted as an error ($e_i = 1$) if the caption mentions the wrong gender, or if it mentions both genders at once (since we only use single-person images for image captioning); captions containing no gendered term are treated as unrevealed and correct ($r_i = 0$, $e_i = 0$). For VisoGender, the predicted gender is derived from the next-token probabilities of *his* versus *her*; a mismatch with the ground-truth counts as an error, while ties are assigned $e_i = 0$ with $r_i = 0$. For VisualCoT, no gender is inferred from the model's output; gender labels serve only to define evaluation subgroups. Following the evaluation procedure of Shao *et al.* [42], each answer is assessed by GPT-OSS-20B [37] in an LLM-as-judge manner (see Sec. F of the supplementary material for more details) that scores the semantic similarity between the predicted and ground-truth answers, resulting in a score $s_i \in [0, 1]$, and the per-sample error is defined as $e_i = 1 - s_i$. All ER and RR values are reported as percentages throughout the following sections.

4.1 Analysis with VL-BICLE

Full results are in Secs. I and K–M of the supplementary material. We summarize the main observations below.

Baseline gender bias varies across datasets and tasks The sign and magnitude of ER_{m-f} at zero-shot (k=0) depend on the model, dataset, and task. Among all models, Idefics3 and InternVL35 shows consistent male bias, and InternVL35 is never female-biased. In image captioning, only Idefics3 shows a consistent gender preference across COCOBias and DCI; three of the six models (QwenVL, MiniCPM, and Qwen3VL) evaluated on both datasets reverse direction between them. Across tasks, Phi35V is nearly unbiased in captioning ($\mathrm{ER}_{m-f} = -0.09$ on COCOBias) yet strongly female-biased in pronoun prediction ($\mathrm{ER}_{m-f} = 13.04$ on VisoGender-OO); conversely, InternVL35 is neutral in captioning but male-biased in pronoun prediction. These inconsistencies suggest that diagnosing a model's gender bias from a single dataset or task [15, 61] risks overgeneralizing findings that are specific to the dataset and task at hand. Table 2 reports the full zero-shot baselines across all settings.

Gendered ICL shifts bias toward the demonstrated gender Across image captioning and pronoun prediction tasks, MS consistently shifts $\Delta\mathrm{ER}_{m-f}^k := \mathrm{ER}_{m-f}^k - \mathrm{ER}_{m-f}^0$ negative (toward male bias), while FS shifts it positive (toward female bias); see Fig. 2. The effect is especially pronounced in pronoun prediction: on VisoGender-OP, all models show positive $\Delta\mathrm{ER}_{m-f}^{k=8}$ under FS, while five of six models show negative $\Delta\mathrm{ER}_{m-f}^k$ under MS; a similar pattern holds for OO. For image captioning, the same directional pattern holds but with smaller effect sizes: under FS, all six models shift positive on DCI and four of six on COCOBias; under MS, five of six models show negative $\Delta\mathrm{ER}_{m-f}^k$ on COCOBias at

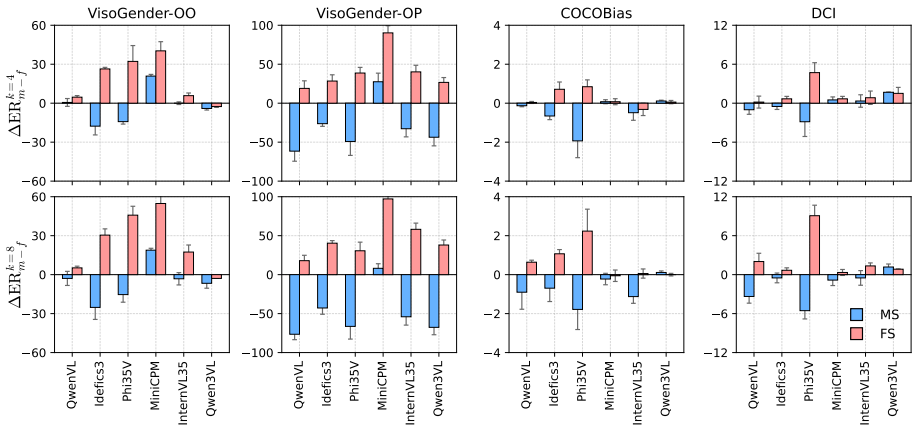

Fig. 2: Gendered ICL shifts bias toward the demonstrated gender.

8-shot (only Qwen3VL shifts positive). This directional shift is generally reliable: it holds regardless of the model's baseline bias direction. Even models that are female-biased at zero-shot (*e.g.*, Phi35V and Qwen3VL in pronoun prediction) are pushed toward male bias by MS (see Sec. I of the supplementary material). This indicates that single-gendered ICL is a directional force, rather than a debiasing tool; it can flip the bias to the opposite gender (*e.g.*, ER_{m-f} for MiniCPM flips from -27.5 at zero-shot to 27.2 under 8-shot FS on VisoGender-OO).

Gendered ICL hurts performance on the opposite gender The directional bias shift described above arises from a consistent cross-gender mechanism: gendered ICL demonstrations hurt performance on the opposite gender more than on the demonstrated gender. For example, under MS where only male examples are used in the context, ER_f rises while ER_m stays relatively flat; similarly under FS, ER_m rises while ER_f is preserved (Fig. 3). The same effect is present in both captioning and pronoun prediction yet differs in magnitude. In image captioning, the effect is modest: at 8-shot on COCOBias, the largest ER_f difference between MS and FS is 1.46 on Phi35V (2.25 under MS *vs*. 0.79 under FS); in pronoun prediction, the same mechanism produces gaps an order of magnitude larger: on VisoGender-OP at 8-shot, ER_f under MS reaches 85.5 for QwenVL (compared to 9.7 under FS), and ER_m under FS reaches 67.5 for Phi35V (compared to 9.4 under MS). Overall, the cross-gender effect is broadly robust; at 8-shot, five of six models exhibit it across both COCOBias and DCI in captioning, and all six models show it on both VisoGender subtasks. Qwen3VL is the only exception in captioning: it shows no cross-gender effect at any shot count on either dataset, because of its overall low error rates for both genders.

SBR methods inherit the training set's gender imbalance SIIR and SITR select ICL examples by semantic similarity rather than controlling for gender composition. Since the original pool of COCOBias is male-skewed, both SBR methods over-select male examples and produce bias levels comparable to

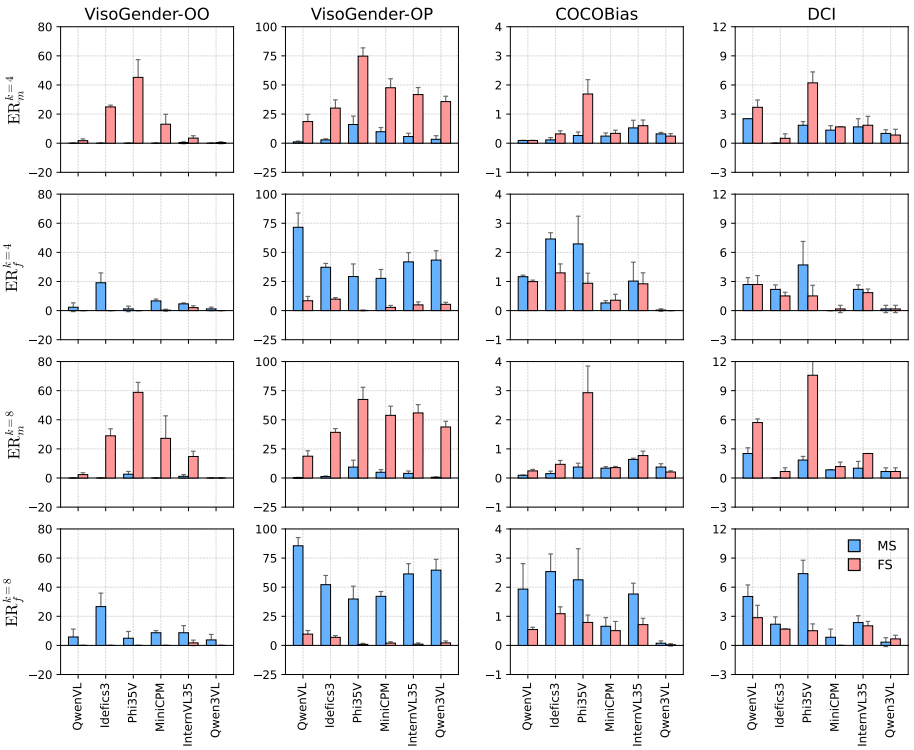

Fig. 3: Gendered ICL context raises opposite-gender error rates more than same-gender error rates.

(or exceeding) those of MS (as shown in Fig. 4). At 8-shot, Phi35V under SIIR and SITR yields ER_{m-f} of -2.25 and -2.35, respectively, showing more male bias than male-only ICL (-1.88 ± 1.03). For InternVL35 and MiniCPM, ER_{m-f} under SIIR and SITR remains within one standard deviation of RS, suggesting that retrieval-based methods, though male-skewed, have little additional effect on these models. Qwen3VL, which has shown little sensitivity to gendered ICL across all settings, behaves similarly under SIIR and SITR: ER_{m-f} remains near its baseline across all VL-BICLE settings. Overall, SIIR and SITR produce similar gender bias levels, and neither method is consistently more biased than the other. Furthermore, both SBR methods cannot consistently outperform MS regarding ER_{m-f}, suggesting that SBR methods offer no debiasing advantage. To separate the effect of the pool's composition from that of the SBR mechanism itself, we additionally run both methods over a down-sampled, gender-balanced pool (balanced pool in Fig. 4; full results in Sec. J of the supplementary material). Using a balanced pool recovers part of the gap opened by retrieval on the original pool, confirming that pool composition is one of the main reasons for gender bias; however, the recovery is only partial: QwenVL, Idefics3, and

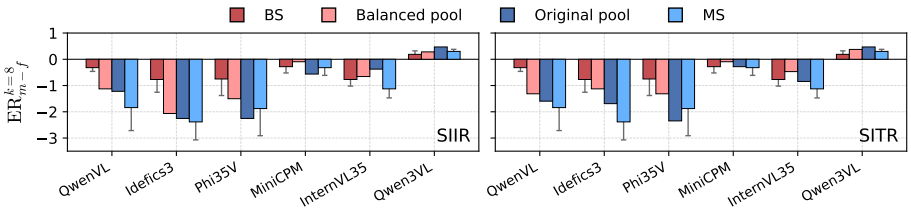

Fig. 4: ER_{m-f} at 8-shot for the two SBR methods across six LVLMs on COCOBias; negative values indicate male bias, positive values for female bias. In addition to the original pool, we further test the retrieval methods under a down-sampled balanced pool. Results on MS and BS settings are shown as references.

Phi35V underperform BS consistently on both SBR methods, indicating that these methods can still reintroduce bias even from a balanced pool, since the top-k retrieval does not force the retrieved context itself to be gender unbiased.

Caption quality metrics do not track gender bias As shown in Tab. 3 and Sec. I of the supplementary material, standard caption quality metrics remain largely stable across ICL settings even as gender bias shifts substantially. CLIPScore is the most insensitive: across all six models, two image captioning datasets, and all shot counts, the largest cross-experiment range of CLIPScore is 0.67, negligible relative to metric means (which lie between 25 and 35). BLEU-4 and AvgL show some variation across ICL settings for certain models, but this variation does not correlate with the direction or magnitude of gender bias shifts. BLEU-4 and CLIPScore gaps between male and female subsets are similarly uninformative: they fluctuate near zero without systematic patterns, even when ER_{m-f} shifts markedly. These results indicate that quality metrics cannot serve as proxies for gender bias.

Gendered ICL effects are absent in VQA The gendered ICL effects observed in image captioning and pronoun prediction do not directly transfer to VQA. As shown in Fig. 5, all four ICL settings (RS, MS, FS, BS) produce nearly identical ER_m and ER_f across all shot counts and all three models tested, regardless of the prompting style. In particular, ER_m and ER_f move in the same direction rather than in opposite directions across settings, so the underlying bias is essentially unchanged (ER_{m-f} differs by at most 1.82 across settings),

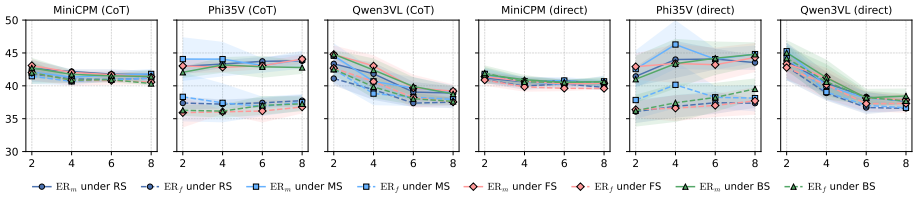

Fig. 5: ICL gender composition does not affect gender bias performance on QA.

Table 3: Caption quality and ER_{m-f} on COCOBias for Phi35V. ER_{m-f} closer to zero indicates lower bias. CLIP denotes CLIPScore.

ICL	k	BLEU↑	CLIP↑	AvgL	ER_{m-f}	ICL	BLEU↑	CLIP↑	AvgL	ER_{m-f}
RS	0	18.86	33.13	20.71	−0.09	MS	18.86	33.13	20.71	−0.09
	2	35.84	32.08	10.70	−1.30		34.15	32.19	11.19	−0.92
	4	36.19	31.77	10.09	−1.67		35.00	32.10	10.93	−2.03
	6	36.54	31.90	10.12	−1.13		36.06	31.99	10.57	−2.18
	8	35.89	31.84	10.20	−1.01		35.77	32.00	10.50	−1.88
FS	2	34.60	32.21	11.30	0.28	BS	33.21	32.21	11.21	−0.71
	4	36.60	31.90	10.41	0.75		35.64	31.94	10.14	−1.28
	6	36.44	31.85	10.19	1.43		36.03	32.01	10.26	−1.20
	8	36.41	31.73	10.05	2.14		36.50	32.02	10.59	−0.75
SIIR	2	34.22	32.35	11.58	−1.31	SITR	34.76	32.63	11.85	−1.41
	4	36.43	32.07	10.81	−1.97		36.34	32.40	11.07	−2.07
	6	36.04	32.15	10.99	−1.60		36.27	32.47	11.17	−2.35
	8	36.65	32.07	10.81	−2.25		37.16	32.39	10.96	−2.35

indicating that gendered ICL neither widens nor narrows the gender bias gap. Within VL-BICLE, where the models, ICL settings, and gender composition are held constant across tasks, we can reasonably attribute such contrast to the task's output space: gendered ICL influences gender bias level only when the task output involves gendered language.

5 Bias Mitigation

The preceding experiments show that the gender composition of ICL examples systematically shifts model behavior. To mitigate the effect of these contextual details, we replace real images with synthetic ones generated using SDMs, as gendered image generation allows for more control over what the image contains via the prompt. Specifically, given a dataset $\mathcal{D}_t$ consisting of N image-caption pairs, we utilize the captions as input to an SDM to generate corresponding synthetic images, thus forming a synthetic dataset $\mathcal{D}_s$ of synthetic image-caption pairs; we ensure that the captions used for image generation contain explicit gender words. Subsequently, we replicate the male-only (MS) and female-only (FS) settings of VL-BICLE, except that the in-context examples are now sampled from $\mathcal{D}_s$ rather than $\mathcal{D}_t$. For synthetic image generation, we employ two SDMs, FLUX [8] and Stable Diffusion-3.5-Large (SD35L) [43]. Since generation runs once offline over the sampling pool, no SDM is involved at inference time and inference latency is unchanged from using real images. More details regarding the experiment settings are shown in Sec. G of the supplementary material.

Caption quality stays stable while reveal rate drops Fig. 6 compares CLIPScore and reveal rate RR_o across two ICL settings with three different image distributions. CLIPScore remains stable across all models and settings,

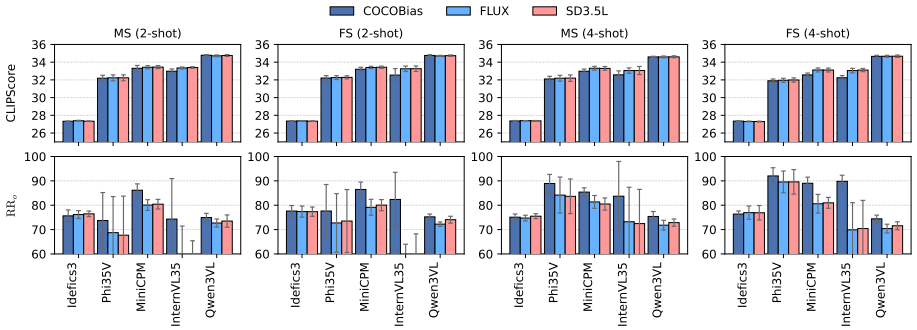

Fig. 6: CLIPScore and reveal rate RR_o across three visual distributions (COCOBias, FLUX, SD35L) at 2-shot and 4-shot under MS and FS settings. CLIPScore remains stable while reveal rate drops with synthetic images.

Table 4: $|\mathrm{ER}_{m-f}|$ on revealed samples (r_i=1) across $k \in \{4, 8\}$, for four LVLMs under male-only (MS) and female-only (FS) settings. We report the gap to the COCOBias baseline, where negative is less biased (better) and positive is more biased (worse).

Model	ICL	$k=4$						$k=8$					
		COCOBias	FLUX	SD35L	Black	S-NP	F-NP	COCOBias	FLUX	SD35L	Black	S-NP	F-NP
Idefics3	MS	2.82	−0.42	−0.27	−0.32	−0.12	−0.05	2.87	−0.12	−0.30	−0.33	−0.02	−0.25
Phi35V	MS	2.24	−0.36	−0.23	+2.48	+4.92	+3.70	2.04	−0.10	−0.20	+3.84	+9.41	+5.79
MiniCPM	MS	0.00	+0.05	+0.07	+0.43	+0.93	+0.93	0.32	−0.24	−0.08	+0.68	+1.63	+1.10
Qwen3VL	MS	0.44	−0.15	−0.07	−0.27	+2.35	+1.72	0.42	+0.00	+0.04	−0.31	+2.23	+1.74
Idefics3	FS	1.00	−0.02	−0.16	+0.95	+0.74	+1.01	0.51	+0.22	+0.00	+1.37	+1.27	+1.29
Phi35V	FS	0.91	−0.25	−0.12	+0.76	+3.16	+3.65	2.53	−0.90	−0.59	+3.36	+27.91	+19.52
MiniCPM	FS	0.00	+0.02	+0.05	+0.58	+0.24	+0.05	0.14	−0.12	−0.05	+1.47	+1.45	+0.94
Qwen3VL	FS	0.35	−0.12	−0.07	+0.18	+1.48	+1.33	0.27	−0.08	−0.02	+0.47	+6.15	+5.45

with differences typically within 0.5, confirming that replacing real context images with synthetic ones does not degrade caption quality. Reveal rate, however, decreases with synthetic context images for most models.

Gender bias is reduced on revealed samples when using synthetic images Evaluation results on ER_{m-f} show that using synthetic images (both FLUX and SD35L) yields better gender bias performance. Since RR_o drops with synthetic images, it is natural to question: does the superiority of using SDM-generated images come from the lower reveal rate (where LVLMs become more cautious about revealing gender, leading to loss of information density), or is it genuinely better? To control for this, we evaluate bias on revealed samples only. Tab. 4 reports $|\mathrm{ER}_{m-f}|$ conditioned on samples where the model does reveal gender ($r_i = 1$). With in-context synthetic images, $|\mathrm{ER}_{m-f}|$ decreases relative to the baseline for three of the four LVLMs under both MS and FS settings. The only exception is MiniCPM with $k = 4$, whose $|\mathrm{ER}_{m-f}|$ on COCOBias is already near zero for both MS and FS. FLUX and SD35L yield similar reductions, suggesting the effect stems from using synthetic images rather than a

specific generator. These results suggest that the visual content of ICL context images influences gender bias in model outputs, even though it has little effect on caption quality. Replacing real photographs with synthetic images offers a simple intervention: it reduces gender bias on revealed samples, without requiring changes to the example selection procedure or the caption text. We analyze why this intervention works in Sec. H of the supplementary material.

6 Conclusion

We presented VL-BICLE, a systematic evaluation framework for analyzing how ICL influences gender bias in LVLMs across three tasks and four datasets. Across six LVLMs, gendered ICL demonstrations act as a directional force, pushing bias toward the demonstrated gender through a cross-gender mechanism that degrades performance on the opposite gender. This effect holds in image captioning and pronoun prediction but is absent in VQA. SBR methods inherit their source pool's imbalance and offer no debiasing advantage, while quality metrics remain blind to these shifts. Replacing real in-context images with SDM-generated ones reduces gender bias without degrading caption quality. We hope that VL-BICLE and our findings encourage more careful consideration of context composition in ICL-based applications of LVLMs.

Acknowledgement

This work was partly supported by JSPS KAKENHI No. 23H00497, 22K12091, and 24K20795, JST ASPIRE Grant No. JPMJAP2502, and JST FOREST Grant No. JPMJFR216O. The first author was supported by JST BOOST, Japan Grant No. JPMJBS2402.

References

1. Alayrac, J., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J.L., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K.: Flamingo: a Visual Language Model for Few-Shot Learning. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022 (2022), http://papers.nips.cc/paper_files/paper/2022/hash/960a172bc7fbf0177ccccbb411a7d800-Abstract-Conference.html
2. Amend, J.J., Wazzan, A., Souvenir, R.: Evaluating Gender-Neutral Training Data for Automated Image Captioning. In: Chen, Y., Ludwig, H., Tu, Y., Fayyad, U.M., Zhu, X., Hu, X., Byna, S., Liu, X., Zhang, J., Pan, S., Papalexakis, V., Wang, J., Cuzzocrea, A., Ordonez, C. (eds.) 2021 IEEE International Conference on Big

Data (Big Data), Orlando, FL, USA, December 15-18, 2021. pp. 1226–1235. IEEE (2021). https://doi.org/10.1109/BIGDATA52589.2021.9671774, https://doi.org/10.1109/BigData52589.2021.9671774

3. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: Visual Question Answering. In: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015. pp. 2425–2433. IEEE Computer Society (2015). https://doi.org/10.1109/ICCV.2015.279, https://doi.org/10.1109/ICCV.2015.279
4. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-VL: A Frontier Large Vision-Language Model with Versatile Abilities. CoRR **abs/2308.12966** (2023). https://doi.org/10.48550/ARXIV.2308.12966, https://doi.org/10.48550/arXiv.2308.12966
5. Baldassini, F.B., Shukor, M., Cord, M., Soulier, L., Piwowarski, B.: What Makes Multimodal In-Context Learning Work? In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024 - Workshops, Seattle, WA, USA, June 17-18, 2024. pp. 1539–1550. IEEE (2024). https://doi.org/10.1109/CVPRW63382.2024.00161, https://doi.org/10.1109/CVPRW63382.2024.00161
6. Bar, A., Gandelsman, Y., Darrell, T., Globerson, A., Efros, A.A.: Visual Prompting via Image Inpainting. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022 (2022), http://papers.nips.cc/paper_files/paper/2022/hash/9f09f316a3eaf59d9ced5ffaefe97e0f-Abstract-Conference.html
7. Berg, H., Hall, S.M., Bhalgat, Y., Kirk, H., Shtedritski, A., Bain, M.: A Prompt Array Keeps the Bias Away: Debiasing Vision-Language Models with Adversarial Learning. In: He, Y., Ji, H., Liu, Y., Li, S., Chang, C., Poria, S., Lin, C., Buntine, W.L., Liakata, M., Yan, H., Yan, Z., Ruder, S., Wan, X., Arana-Catania, M., Wei, Z., Huang, H., Wu, J., Day, M., Liu, P., Xu, R. (eds.) Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, AACL/IJCNLP 2022 - Volume 1: Long Papers, Online Only, November 20-23, 2022. pp. 806–822. Association for Computational Linguistics (2022), https://aclanthology.org/2022.aacl-main.61
8. Black-Forest-Labs: Flux. https://github.com/black-forest-labs/flux (2024), last accessed 18 Jul 2026
9. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language Models are Few-Shot Learners. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual (2020), https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html
10. Chen, X., Fang, H., Lin, T., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft COCO Captions: Data Collection and Evaluation Server. CoRR **abs/1504.00325** (2015), http://arxiv.org/abs/1504.00325

11. Chuang, C., Jampani, V., Li, Y., Torralba, A., Jegelka, S.: Debiasing Vision-Language Models via Biased Prompts. CoRR **abs/2302.00070** (2023). https://doi.org/10.48550/ARXIV.2302.00070, https://doi.org/10.48550/arXiv.2302.00070
12. Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., Xia, H., Xu, J., Wu, Z., Chang, B., Sun, X., Li, L., Sui, Z.: A Survey on In-context Learning. In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 1107–1128. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024). https://doi.org/10.18653/v1/2024.emnlp-main.64, https://aclanthology.org/2024.emnlp-main.64/
13. Doveh, S., Perek, S., Mirza, M.J., Alfassy, A., Arbelle, A., Ullman, S., Karlinsky, L.: Towards Multimodal In-Context Learning for Vision & Language Models. CoRR **abs/2403.12736** (2024). https://doi.org/10.48550/ARXIV.2403.12736, https://doi.org/10.48550/arXiv.2403.12736
14. Hall, S.M., Abrantes, F.G., Zhu, H., Sodunke, G., Shtedritski, A., Kirk, H.R.: VisoGender: A dataset for benchmarking gender bias in image-text pronoun resolution. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023 (2023), http://papers.nips.cc/paper_files/paper/2023/hash/c93f26b1381b17693055a611a513f1e9-Abstract-Datasets_and_Benchmarks.html
15. Hendricks, L.A., Burns, K., Saenko, K., Darrell, T., Rohrbach, A.: Women Also Snowboard: Overcoming Bias in Captioning Models. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part III. Lecture Notes in Computer Science, vol. 11207, pp. 793–811. Springer (2018). https://doi.org/10.1007/978-3-030-01219-9_47, https://doi.org/10.1007/978-3-030-01219-9_47
16. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In: Moens, M., Huang, X., Specia, L., Yih, S.W. (eds.) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021. pp. 7514–7528. Association for Computational Linguistics (2021). https://doi.org/10.18653/V1/2021.EMNLP-MAIN.595, https://doi.org/10.18653/v1/2021.emnlp-main.595
17. Hirota, Y., Hachiuma, R., Yang, C.H., Nakashima, Y.: From Descriptive Richness to Bias: Unveiling the Dark Side of Generative Image Caption Enrichment. In: Al-Onaizan, Y., Bansal, M., Chen, Y. (eds.) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024. pp. 17807–17816. Association for Computational Linguistics (2024), https://aclanthology.org/2024.emnlp-main.986
18. Hirota, Y., Nakashima, Y., Garcia, N.: Gender and Racial Bias in Visual Question Answering Datasets. In: FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022. pp. 1280–1292. ACM (2022). https://doi.org/10.1145/3531146.3533184, https://doi.org/10.1145/3531146.3533184
19. Hirota, Y., Nakashima, Y., Garcia, N.: Quantifying Societal Bias Amplification in Image Captioning. In: IEEE/CVF Conference on Computer Vision and Pattern

Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. pp. 13440–13449. IEEE (2022). https://doi.org/10.1109/CVPR52688.2022.01309, https://doi.org/10.1109/CVPR52688.2022.01309
20. Hirota, Y., Nakashima, Y., Garcia, N.: Model-Agnostic Gender Debiased Image Captioning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 15191–15200. IEEE (2023). https://doi.org/10.1109/CVPR52729.2023.01458, https://doi.org/10.1109/CVPR52729.2023.01458
21. Howard, P., Fraser, K.C., Bhiwandiwalla, A., Kiritchenko, S.: Uncovering Bias in Large Vision-Language Models at Scale with Counterfactuals. CoRR **abs/2405.20152** (2024). https://doi.org/10.48550/ARXIV.2405.20152, https://doi.org/10.48550/arXiv.2405.20152
22. Howard, P., Madasu, A., Le, T., Lujan-Moreno, G.A., Bhiwandiwalla, A., Lal, V.: SocialCounterfactuals: Probing and Mitigating Intersectional Social Biases in Vision-Language Models with Counterfactual Examples. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024. pp. 11975–11985. IEEE (2024). https://doi.org/10.1109/CVPR52733.2024.01138, https://doi.org/10.1109/CVPR52733.2024.01138
23. Hudson, D.A., Manning, C.D.: GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 6700–6709. Computer Vision Foundation / IEEE (2019). https://doi.org/10.1109/CVPR.2019.00686, http://openaccess.thecvf.com/content_CVPR_2019/html/Hudson_GQA_A_New_Dataset_for_Real-World_Visual_Reasoning_and_Compositional_CVPR_2019_paper.html
24. Janghorbani, S., de Melo, G.: Multi-Modal Bias: Introducing a Framework for Stereotypical Bias Assessment beyond Gender and Race in Vision-Language Models. In: Vlachos, A., Augenstein, I. (eds.) Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023. pp. 1717–1727. Association for Computational Linguistics (2023). https://doi.org/10.18653/V1/2023.EACL-MAIN.126, https://doi.org/10.18653/v1/2023.eacl-main.126
25. Jung, H., Jang, T., Wang, X.: A Unified Debiasing Approach for Vision-Language Models across Modalities and Tasks. In: Globersons, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J.M., Zhang, C. (eds.) Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024 (2024), http://papers.nips.cc/paper_files/paper/2024/hash/254404d551f6ce17bb7407b4d6b3c87b-Abstract-Conference.html
26. Laurençon, H., Marafioti, A., Sanh, V., Tronchon, L.: Building and better understanding vision-language models: insights and future directions. CoRR **abs/2408.12637** (2024). https://doi.org/10.48550/ARXIV.2408.12637, https://doi.org/10.48550/arXiv.2408.12637
27. Laurençon, H., Saulnier, L., Tronchon, L., Bekman, S., Singh, A., Lozhkov, A., Wang, T., Karamcheti, S., Rush, A.M., Kiela, D., Cord, M., Sanh, V.: OBELICS: An Open Web-Scale Filtered Dataset of Interleaved Image-Text Documents. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023 (2023), http://papers.nips.cc/paper_files/

paper/2023/hash/e2cfb719f58585f779d0a4f9f07bd618-Abstract-Datasets_and_Benchmarks.html
28. Li, B., Zhang, Y., Chen, L., Wang, J., Pu, F., Yang, J., Li, C., Liu, Z.: MIMIC-IT: Multi-Modal In-Context Instruction Tuning. CoRR **abs/2306.05425** (2023). https://doi.org/10.48550/ARXIV.2306.05425, https://doi.org/10.48550/arXiv.2306.05425
29. Li, L., Peng, J., Chen, H., Gao, C., Yang, X.: How to Configure Good In-Context Sequence for Visual Question Answering. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024. pp. 26700–26710. IEEE (2024). https://doi.org/10.1109/CVPR52733.2024.02522, https://doi.org/10.1109/CVPR52733.2024.02522
30. Li, W., Li, L., Xiang, T., Liu, X., Deng, W., Garcia, N.: Can multiple-choice questions really be useful in detecting the abilities of LLMs? In: Calzolari, N., Kan, M.Y., Hoste, V., Lenci, A., Sakti, S., Xue, N. (eds.) Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). pp. 2819–2834. ELRA and ICCL, Torino, Italia (May 2024), https://aclanthology.org/2024.lrec-main.251/
31. Li, Y.: Advancing Multimodal In-Context Learning in Large Vision-Language Models with Task-aware Demonstrations. CoRR **abs/2503.04839** (2025). https://doi.org/10.48550/ARXIV.2503.04839, https://doi.org/10.48550/arXiv.2503.04839
32. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual Instruction Tuning. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023 (2023), http://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html
33. Manevich, A., Tsarfaty, R.: Mitigating Hallucinations in Large Vision-Language Models (LVLMs) via Language-Contrastive Decoding (LCD). In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Findings of the Association for Computational Linguistics: ACL 2024. pp. 6008–6022. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024). https://doi.org/10.18653/v1/2024.findings-acl.359, https://aclanthology.org/2024.findings-acl.359/
34. Microsoft: Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. CoRR **abs/2404.14219** (2024). https://doi.org/10.48550/ARXIV.2404.14219, https://doi.org/10.48550/arXiv.2404.14219
35. Niranjan, C., Jaidka, K., Yeo, G.C.: On the Limitations of Steering in Language Model Alignment. CoRR **abs/2505.01162** (2025)
36. OpenAI: GPT-4 Technical Report. CoRR **abs/2303.08774** (2023). https://doi.org/10.48550/ARXIV.2303.08774, https://doi.org/10.48550/arXiv.2303.08774
37. OpenAI: Introducing gpt-oss (2025), https://openai.com/index/introducing-gpt-oss/, last accessed 18 Jul 2026
38. Papineni, K., Roukos, S., Ward, T., Zhu, W.: Bleu: a Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA. pp. 311–318. ACL (2002). https://doi.org/10.3115/1073083.1073135, https://aclanthology.org/P02-1040/
39. Post, M.: A Call for Clarity in Reporting BLEU Scores. In: Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Jimeno-Yepes,

A., Koehn, P., Monz, C., Negri, M., Névéol, A., Neves, M.L., Post, M., Specia, L., Turchi, M., Verspoor, K. (eds.) Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018. pp. 186–191. Association for Computational Linguistics (2018). https://doi.org/10.18653/V1/W18-6319, https://doi.org/10.18653/v1/w18-6319

40. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (2021), http://proceedings.mlr.press/v139/radford21a.html
41. Ratzlaff, N., Olson, M.L., Hinck, M., Tseng, S., Lal, V., Howard, P.: Debiasing Large Vision-Language Models by Ablating Protected Attribute Representations. CoRR **abs/2410.13976** (2024). https://doi.org/10.48550/ARXIV.2410.13976, https://doi.org/10.48550/arXiv.2410.13976
42. Shao, H., Qian, S., Xiao, H., Song, G., Zong, Z., Wang, L., Liu, Y., Li, H.: Visual CoT: Advancing Multi-Modal Language Models with a Comprehensive Dataset and Benchmark for Chain-of-Thought Reasoning. In: Globersons, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J.M., Zhang, C. (eds.) Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024 (2024), http://papers.nips.cc/paper_files/paper/2024/hash/0ff38d72a2e0aa6dbe42de83a17b2223-Abstract-Datasets_and_Benchmarks_Track.html
43. StabilityAI: Stable Diffusion 3.5 (2024), https://github.com/Stability-AI/sd3.5, last accessed 18 Jul 2026
44. Tang, R., Du, M., Li, Y., Liu, Z., Zou, N., Hu, X.: Mitigating Gender Bias in Captioning Systems. In: Leskovec, J., Grobelnik, M., Najork, M., Tang, J., Zia, L. (eds.) WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021. pp. 633–645. ACM / IW3C2 (2021). https://doi.org/10.1145/3442381.3449950, https://doi.org/10.1145/3442381.3449950
45. Urbanek, J., Bordes, F., Astolfi, P., Williamson, M., Sharma, V., Romero-Soriano, A.: A Picture is Worth More Than 77 Text Tokens: Evaluating CLIP-Style Models on Dense Captions. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024. pp. 26690–26699. IEEE (2024). https://doi.org/10.1109/CVPR52733.2024.02521, https://doi.org/10.1109/CVPR52733.2024.02521
46. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015. pp. 3156–3164. IEEE Computer Society (2015). https://doi.org/10.1109/CVPR.2015.7298935, https://doi.org/10.1109/CVPR.2015.7298935
47. Wang, J., Liu, Y., Wang, X.E.: Are Gender-Neutral Queries Really Gender-Neutral? Mitigating Gender Bias in Image Search. In: Moens, M., Huang, X., Specia, L., Yih, S.W. (eds.) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021. pp. 1995–2008. Association for Computa-

tional Linguistics (2021). https://doi.org/10.18653/V1/2021.EMNLP-MAIN.151, https://doi.org/10.18653/v1/2021.emnlp-main.151

48. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., Wang, Z., Chen, Z., Zhang, H., Yang, G., Wang, H., Wei, Q., Yin, J., Li, W., Cui, E., Chen, G., Ding, Z., Tian, C., Wu, Z., Xie, J., Li, Z., Yang, B., Duan, Y., Wang, X., Hou, Z., Hao, H., Zhang, T., Li, S., Zhao, X., Duan, H., Deng, N., Fu, B., He, Y., Wang, Y., He, C., Shi, B., He, J., Xiong, Y., Lv, H., Wu, L., Shao, W., Zhang, K., Deng, H., Qi, B., Ge, J., Guo, Q., Zhang, W., Zhang, S., Cao, M., Lin, J., Tang, K., Gao, J., Huang, H., Gu, Y., Lyu, C., Tang, H., Wang, R., Lv, H., Ouyang, W., Wang, L., Dou, M., Zhu, X., Lu, T., Lin, D., Dai, J., Su, W., Zhou, B., Chen, K., Qiao, Y., Wang, W., Luo, G.: InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. CoRR **abs/2508.18265** (2025). https://doi.org/10.48550/ARXIV.2508.18265, https://doi.org/10.48550/arXiv.2508.18265
49. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E.H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., Fedus, W.: Emergent Abilities of Large Language Models. Trans. Mach. Learn. Res. **2022** (2022), https://openreview.net/forum?id=yzkSU5zdwD
50. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E.H., Le, Q.V., Zhou, D.: Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022 (2022), http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html
51. Wu, Z., Wang, Y., Ye, J., Kong, L.: Self-Adaptive In-Context Learning: An Information Compression Perspective for In-Context Example Selection and Ordering. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1423–1436. Association for Computational Linguistics, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.acl-long.79, https://aclanthology.org/2023.acl-long.79/
52. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., Qiu, Z.: Qwen3 technical report. CoRR **abs/2505.09388** (2025). https://doi.org/10.48550/ARXIV.2505.09388, https://doi.org/10.48550/arXiv.2505.09388
53. Yang, X., Peng, Y., Ma, H., Xu, S., Zhang, C., Han, Y., Zhang, H.: Lever LM: Configuring In-Context Sequence to Lever Large Vision Language Models. In: Globersons, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J.M., Zhang, C. (eds.) Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024 (2024), http://papers.nips.cc/paper_files/paper/2024/hash/b619cd6dcc986856b8a8da2b08d89396-Abstract-Conference.html

54. Yang, X., Wu, Y., Yang, M., Chen, H., Geng, X.: Exploring Diverse In-Context Configurations for Image Captioning. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023 (2023), http://papers.nips.cc/paper_files/paper/2023/hash/804b5e300c9ed4e3ea3b073f186f4adc-Abstract-Conference.html
55. Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., Cai, T., Li, H., Zhao, W., He, Z., Chen, Q., Zhou, H., Zou, Z., Zhang, H., Hu, S., Zheng, Z., Zhou, J., Cai, J., Han, X., Zeng, G., Li, D., Liu, Z., Sun, M.: MiniCPM-V: A GPT-4V Level MLLM on Your Phone. CoRR **abs/2408.01800** (2024). https://doi.org/10.48550/ARXIV.2408.01800, https://doi.org/10.48550/arXiv.2408.01800
56. Yin, H., Si, G., Wang, Z.: The Mirage of Performance Gains: Why Contrastive Decoding Fails to Address Multimodal Hallucination. CoRR **abs/2504.10020** (2025)
57. Zhang, M., Ré, C.: Contrastive Adapters for Foundation Model Group Robustness. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022 (2022), http://papers.nips.cc/paper_files/paper/2022/hash/8829f586a1ac0e6c41143f5d57b63c4b-Abstract-Conference.html
58. Zhang, Y., Yu, W., Wen, Q., Wang, X., Zhang, Z., Wang, L., Jin, R., Tan, T.: Debiasing Multimodal Large Language Models. CoRR **abs/2403.05262** (2024). https://doi.org/10.48550/ARXIV.2403.05262, https://doi.org/10.48550/arXiv.2403.05262
59. Zhang, Y., Zhou, K., Liu, Z.: What Makes Good Examples for Visual In-Context Learning? In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023 (2023), http://papers.nips.cc/paper_files/paper/2023/hash/398ae57ed4fda79d0781c65c926d667b-Abstract-Conference.html
60. Zhang, Z., Zhang, A., Li, M., Smola, A.: Automatic Chain of Thought Prompting in Large Language Models. In: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net (2023), https://openreview.net/forum?id=5NTt8GFjUHkr
61. Zhao, D., Wang, A., Russakovsky, O.: Understanding and Evaluating Racial Biases in Image Captioning. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. pp. 14810–14820. IEEE (2021). https://doi.org/10.1109/ICCV48922.2021.01456, https://doi.org/10.1109/ICCV48922.2021.01456
62. Zhao, H., Cai, Z., Si, S., Ma, X., An, K., Chen, L., Liu, Z., Wang, S., Han, W., Chang, B.: MMICL: Empowering Vision-language Model with Multi-Modal In-Context Learning. In: The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net (2024), https://openreview.net/forum?id=5KojubHBr8
63. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.: Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In: Palmer, M., Hwa, R., Riedel, S. (eds.) Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark,

September 9-11, 2017. pp. 2979–2989. Association for Computational Linguistics (2017). https://doi.org/10.18653/V1/D17-1323, https://doi.org/10.18653/v1/d17-1323

64. Zhou, Y., Li, X., Wang, Q., Shen, J.: Visual In-Context Learning for Large Vision-Language Models. In: Ku, L., Martins, A., Srikumar, V. (eds.) Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024. pp. 15890–15902. Association for Computational Linguistics (2024). https://doi.org/10.18653/V1/2024.FINDINGS-ACL.940, https://doi.org/10.18653/v1/2024.findings-acl.940
65. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. In: The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net (2024), https://openreview.net/forum?id=1tZbq88f27