Think Before You Act: A Two-Stage Framework for Mitigating Gender Bias Towards Vision-Language Tasks

Yunqi Zhang[1], Songda Li[1], Chunyuan Deng[2], Luyi Wang[1], Hui Zhao[1,3,*]
[1]Software Engineering Institute, East China Normal University
[2]Georgia Institute of Technology
[3]Shanghai Key Laboratory of Trustworthy Computing, Shanghai, China
{yunqi.zhang,songda.li,luyi.wang}@stu.ecnu.edu.cn
cdeng73@gatech.edu
hzhao@sei.ecnu.edu.cn

Abstract

Gender bias in vision-language models (VLMs) can reinforce harmful stereotypes and discrimination. In this paper, we focus on mitigating gender bias towards vision-language tasks. We identify object hallucination as the essence of gender bias in VLMs. Existing VLMs tend to focus on salient or familiar attributes in images but ignore contextualized nuances. Moreover, most VLMs rely on the co-occurrence between specific objects and gender attributes to infer the ignored features, ultimately resulting in gender bias. We propose GAMA, a task-agnostic generation framework to mitigate gender bias. GAMA consists of two stages: narrative generation and answer inference. During narrative generation, GAMA yields all-sided but gender-obfuscated narratives, which prevents premature concentration on localized image features, especially gender attributes. During answer inference, GAMA integrates the image, generated narrative, and a task-specific question prompt to infer answers for different vision-language tasks. This approach allows the model to rethink gender attributes and answers. We conduct extensive experiments on GAMA, demonstrating its debiasing and generalization ability.[1]

1 Introduction

Vision-language models (VLMs) have attracted significant attention in recent years due to their widespread applications in image captioning (Li et al., 2020; Nguyen et al., 2022), image-text retrieval (Wang et al., 2020; Qu et al., 2021), and visual question answering (Antol et al., 2015; Jiang et al., 2020). Remarkable advancements have been achieved in these tasks, primarily measured by task performance metrics (Radford et al., 2021; Li et al., 2023a). However, there is a growing concern about the undesirable social bias (e.g., gender, race) in VLMs (Hendricks et al., 2018; Ross et al., 2021; Zhang et al., 2022). An example is presented in Figure 1 (bottom), implying stereotypes associating sports with masculinity. More alarmingly, biased VLMs have the potential to propagate and even exacerbate existing stereotypes and inequalities.

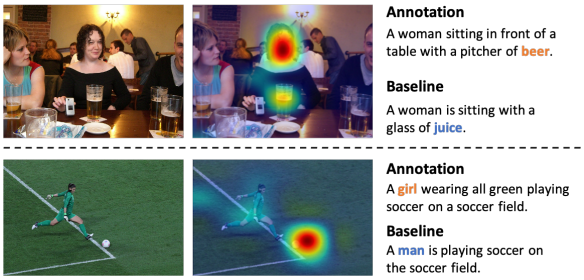

Figure 1: **Examples in image captioning with annotations and captions generated by a baseline model, SAT (Xu et al., 2015).** We overlay images with attention heatmaps of SAT on the right. In the top example, SAT focuses on the woman and predicts "juice", a word highly co-occurring with females. In the bottom example, the gender is incorrectly predicted, as "soccer" highly co-occurs with males in the training set.

Previous methods for mitigating social bias towards vision-language tasks can be grouped into two categories: 1) *task-specific methods* are designed for task-specific datasets and models, which mitigate bias by re-sampling datasets (Zhao et al., 2017; Wang et al., 2021), synthesizing negative samples for training (Hirota et al., 2023), or introducing debiasing modules or training objectives to task-specific models (Hendricks et al., 2018; Tang et al., 2021; Seth et al., 2023); and 2) *task-agnostic methods* aim to pre-train a debiased encoder for downstream tasks, most commonly by adopting an adversarial approach to remove unwanted features (Wang et al., 2019), or leveraging counterfactual samples to minimize biased representations (Zhang et al., 2022). Despite these efforts, task-specific

*Corresponding author.
[1]Our code is available at https://github.com/zyq0000/GAMA.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 773–791
June 16-21, 2024

methods often lack generalization ability, while existing task-agnostic methods primarily address bias at the feature level and fall short of probing the essence of social bias.

In this work, we focus on gender bias as it is a crucial aspect of social bias. To address potential limitations in existing research, our initial step is to explore the essence of gender bias. We posit that gender bias is a manifestation of object hallucination in VLMs (Rohrbach et al., 2018). Specifically, models understand the form rather than the meaning.[2] This leads to a tendency for VLMs to focus on the most salient or familiar objects or attributes but ignore the rest of the image. Moreover, as depicted in Figure 1, VLMs are inclined to hallucinate objects that co-occur with gender words, and vice versa for gender attributes closely associated with specific objects in the training set. As a result, VLMs may generate answers containing objects or gender attributes inconsistent with the given image.

We propose GAMA, a novel task-agnostic generation framework to mitigate gender bias towards vision-language tasks. Unlike previous methods, GAMA is a multi-level method, addressing bias both in external object co-occurrences and in internal bias features. The framework comprises two stages: *narrative generation* and *answer inference*. During *narrative generation*, GAMA creates an all-sided narrative for a given image, which prevents premature focus on localized details. Besides, we disentangle features through contrastive learning (He et al., 2020; Chen et al., 2020) to obfuscate gender information in the generated narratives. This approach helps mitigate the influence of gender attributes on context generation. The trained model can seamlessly adapt to different vision-language tasks for narrative generation without retraining. During *answer inference*, the image, generated narrative, and task-specific question prompt are utilized to derive answers for different tasks. As gender information is obfuscated in generated narratives, the model is encouraged to rethink gender attributes in this stage, which contributes to more appropriate and unbiased answers.

We conduct extensive experiments to demonstrate the effectiveness of GAMA. First, GAMA is evaluated across two vision-language tasks: image captioning and image search. Additionally, zero-shot experiments are performed on two benchmarks for measuring gender bias, namely VisoGender (Hall et al., 2023) and VL-Bias (Zhang et al., 2022). Experimental results show that GAMA performs well against previous debiasing methods on task performance and gender bias mitigation. Notably, GAMA exhibits remarkable generalization ability on both benchmarks. Moreover, we showcase the effectiveness of our proposed modules in reducing object hallucination and gender bias. For further explanation, we probe the connection between object hallucination and gender bias.

[2]Following Bender and Koller (2020), we define *form* as the visible realization of vision and language, such as pixels or bytes in digital representations of text or image, and *meaning* as the relationship between form and external elements to vision and language, like communicative intent.

2 Related Work

2.1 Sources of gender bias

The growing popularity of multi-modality has prompted research into sources of gender bias in vision-language tasks. First, *datasets* are a prominent factor, as object labeling is closely linked to our conceptualization (Brown, 1958). Thus real-life stereotypes subtly infiltrate datasets. Hirota et al. (2022a) and Harrison et al. (2023) highlighted the underrepresentation of women in datasets. Second, *pre-trained VLMs* (PVLMs) have enhanced vision-language tasks by leveraging extensive knowledge from pre-training data. Consequently, PVLMs inherit such bias from both language and vision sources (Wang et al., 2019; Ross et al., 2021; Srinivasan and Bisk, 2022). Third, *model structures* may amplify gender bias in datasets (Zhao et al., 2017). Researchers have observed stereotype exaggeration in various task-specific VLMs (Kay et al., 2015; Bhargava and Forsyth, 2019; Wang et al., 2021).

Given these diverse bias sources, gender bias may not be effectively addressed by merely pre-processing data or debiasing pre-training features.

2.2 Gender bias mitigation

Research interest in gender bias mitigation for vision-language tasks has notably increased. Previous studies addressed gender bias in unimodal models, such as language (He et al., 2022; Shaikh et al., 2023) or vision (Wang et al., 2019; Steed and Caliskan, 2021). Zhou et al. (2022) uncovered both intra-modal and inter-modal gender bias. Previous work on gender bias mitigation for VLMs can be divided into three classes: 1) *adding debiasing modules to existing task-specific VLMs*, which aims to pre-process imbalanced training sets (Zhao et al.,

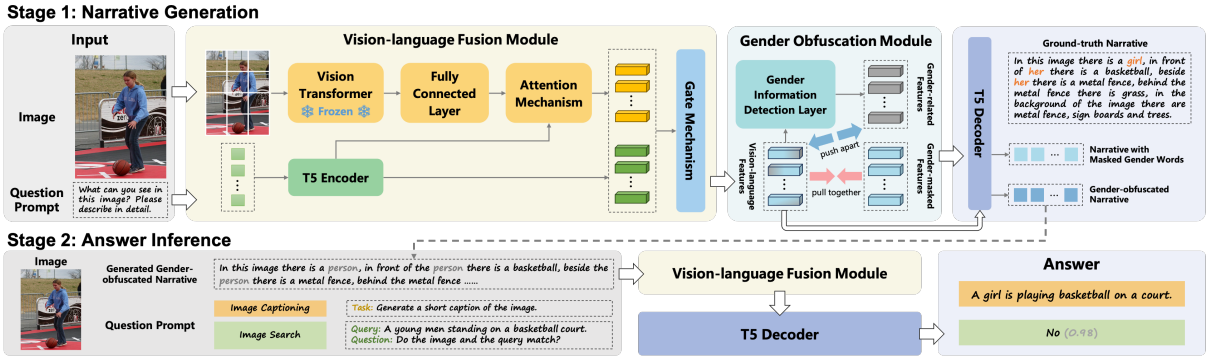

Figure 2: **The overall framework of GAMA.** We briefly provide task-specific question prompts and answers, which are detailed in Appendix C. We take the token probability of the decoder as the match score in image search.

2017), process intermediate features (Hendricks et al., 2018; Wang et al., 2021), or post-process biased model outputs (Hirota et al., 2023; Janghorbani and De Melo, 2023); 2) *re-designing training objectives of PVLMs* to produce debiased features through negative (Wang et al., 2019) or counterfactual (Zhang et al., 2022) samples; and 3) *proposing new model structures* designed to learn fair representations by capturing gender visual evidence (Tang et al., 2021), or adding a negative residual (Seth et al., 2023). However, previous studies superficially address gender bias by removing biased data or features and ignore the correlation between gender bias and object hallucination.

2.3 Object hallucination

Despite the great success of VLMs in vision-language tasks, they still suffer from object hallucination. Object hallucination refers to the fact that the content generated by the model contains objects inconsistent with or absent from the given image (Rohrbach et al., 2018). Rohrbach et al. (2018) proposed two metrics, CHAIRs and CHAIRi, to measure object hallucination. A recent approach, POPE (Li et al., 2023b), is a polling-based query method with enhanced flexibility. Efforts to mitigate object hallucination in VLMs fall into two categories. Some methods aim to disentangle object co-occurrence patterns. For instance, Biten et al. (2022) introduced object labels and altered objects in the captions, while Zhou et al. (2023) replaced error-prone objects with placeholder tags. Others focus on minimizing logical errors, such as leveraging chain-of-thought (CoT) prompting to generate intermediate reasoning chains as rationales (Zhang et al., 2023).

3 Method

In this section, we will first overview the framework of GAMA and then introduce the training data. Finally, we detail our model architecture.

3.1 Overview

GAMA is composed of two stages: narrative generation and answer inference, as illustrated in Figure 2. During narrative generation, GAMA takes an image and a question prompt as input to yield an all-sided but gender-obfuscated narrative. During answer inference, different vision-language tasks are formulated as generation tasks. GAMA utilizes the image, generated narrative, and task-specific question prompt to generate a task-specific answer. We formulate both stages as taking an image I and a text sequence $X = [x_1, x_2, \cdots, x_n]$ with n tokens as input and outputting a target sequence $Y = [y_1, y_2, \cdots, y_m]$ with m tokens.

The models in the two stages are trained independently. The trained narrative generation model can be applied to different tasks without retraining.

GAMA comprises three key modules: a vision-language fusion module to extract features in both stages, a gender obfuscation module to mask gender-related information during narrative generation, and a decoder for answer generation.

3.2 Training data

During narrative generation, GAMA strives to generate an all-sided narrative for the given image to avoid premature focus on localized details. To this end, we train the model with the Localized Narratives dataset (Voigtlaender et al., 2023) for Open Images (Kuznetsova et al., 2020). The dataset is crafted to depict image regions covered by a mouse trace, associating specific image regions with spe-

cific words in the narrative. Consequently, the narratives are spread throughout the entire image, rather than confined to specific localized regions.

3.3 Vision-language fusion module

The vision-language fusion module is designed to learn vision-language representations in two stages.

Encoder We use the T5 (Raffel et al., 2020) encoder to learn language representations, and a Vision Transformer (ViT) (Dosovitskiy et al., 2020) to extract vision features:

$$\mathbf{H}_l = \text{T5Encoder}([x_1, x_2, \cdots, x_n]), \quad (1)$$

$$\mathbf{H}_v = \mathbf{W}_1 \text{ViT}(I), \quad (2)$$

where $\mathbf{H}_l, \mathbf{H}_v \in \mathbb{R}^{n \times d}$, and d is the hidden dimension. $\mathbf{W}_1$ is a projection matrix to convert the shape of features extracted by the ViT. Notably, the parameters of the ViT are frozen.

Feature fusion We utilize an attention mechanism (Vaswani et al., 2017) to capture the interactions between language and vision features:

$$\hat{\mathbf{H}}_v = \text{Attention}(\mathbf{H}_l, \mathbf{H}_v, \mathbf{H}_v). \quad (3)$$

We aggregate $\mathbf{H}_l$ and $\hat{\mathbf{H}}_v$ with a gate mechanism (Li et al., 2022a; Fang and Feng, 2022). The gate λ and the fused features $\mathbf{H}$ are calculated as:

$$\lambda = \text{Sigmoid}(\mathbf{W}_2\mathbf{H}_l + \mathbf{W}_3\hat{\mathbf{H}}_v), \quad (4)$$

$$\mathbf{H} = (1 - \lambda) \cdot \mathbf{H}_l + \lambda \cdot \hat{\mathbf{H}}_v, \quad (5)$$

where $\mathbf{W}_2$ and $\mathbf{W}_3$ are trainable parameters.

3.4 Gender obfuscation module

The gender obfuscation module is employed during narrative generation to obfuscate gender information in narratives. This module prevents the model from biasing the context due to gender attributes.

Pre-processing First, we replace gender words with a special token `[GENDER]` and leave other words unchanged to obtain a narrative with masked gender words $\bar{Y} = [\bar{y}_1, \bar{y}_2, \cdots, \bar{y}_m]$.[3]

Gender information detection Then we construct the gender-masked features $\bar{\mathbf{H}}$. We detect gender-related features $\mathbf{H}_g$ and remove them from the vision-language features $\mathbf{H}$ as follows:

$$\mu = \text{Sigmoid}(\mathbf{W}_4\mathbf{H}), \quad (6)$$

$$\mathbf{H}_g = \mu \cdot \mathbf{H}, \quad (7)$$

$$\bar{\mathbf{H}} = \mathbf{H} - \mathbf{H}_g, \quad (8)$$

where $\mathbf{W}_4$ is a trainable parameter.

Contrastive loss Motivated by contrastive learning (He et al., 2020; Chen et al., 2020), we employ a contrastive loss to keep the vision-language features $\mathbf{H}$ close to the gender-masked features $\bar{\mathbf{H}}$ and away from the gender-related features $\mathbf{H}_g$. This approach helps obscure gender details as well as preserve context information. We define our contrastive loss as:

$$\mathcal{L}_{con} = -\log \frac{e^{\text{s}(\mathbf{H}, \bar{\mathbf{H}})/\tau}}{e^{\text{s}(\mathbf{H}, \bar{\mathbf{H}})/\tau} + e^{\text{s}(\mathbf{H}, \mathbf{H}_g)/\tau}}, \quad (9)$$

where $\text{s}(\mathbf{U}, \mathbf{V}) = \mathbf{U}^\top \mathbf{V} / \|\mathbf{U}\| \|\mathbf{V}\|$ denotes the cosine similarity between $\mathbf{U}$ and $\mathbf{V}$, and τ is a temperature hyper-parameter.

3.5 Decoder

The T5 decoder takes the encoder output $\tilde{\mathbf{H}}$, text sequence X, and previously generated tokens $\tilde{Y}_{<t}$ as inputs to get the t-th token probability distribution:

$$\tilde{\mathbf{h}}_t = \text{T5Decoder}(\tilde{\mathbf{H}}, X, \tilde{Y}_{<t}), \quad (10)$$

$$P(\tilde{y}_t | \tilde{\mathbf{H}}, X, \tilde{Y}_{<t}) = \text{Softmax}(\mathbf{W}_5 \tilde{\mathbf{h}}_t), \quad (11)$$

where $\mathbf{W}_5$ is a trainable parameter to map the hidden dimension to the vocabulary size. For each symbol $\tilde{a} \in \{\tilde{\mathbf{H}}, \tilde{\mathbf{h}}_t, \tilde{Y}_{<t}, \tilde{y}_t\}$, $\tilde{a}$ is either a or $\bar{a}$.

3.6 Training objective

During the training phase, we use the teacher forcing to train the models in the two stages. The models in two stages are trained independently.

Narrative generation In the narrative generation stage, we train the model with the cross-entropy loss on ground-truth narratives and narratives with masked gender words, which can be defined as:

$$\mathcal{L}_{ce} = -\sum_{t=1}^{m} \log P(y_t | \mathbf{H}, X, Y_{<t}), \quad (12)$$

$$\bar{\mathcal{L}}_{ce} = -\sum_{t=1}^{m} \log P(\bar{y}_t | \bar{\mathbf{H}}, X, \bar{Y}_{<t}). \quad (13)$$

In total, the narrative generation stage contains three losses, which can be formulated as:

$$\mathcal{L}_1 = \mathcal{L}_{con} + \mathcal{L}_{ce} + \bar{\mathcal{L}}_{ce}. \quad (14)$$

Answer inference The answer inference model is trained with a cross-entropy loss:

$$\mathcal{L}_2 = -\sum_{t=1}^{m} \log P(y_t | \mathbf{H}, X, Y_{<t}). \quad (15)$$

[3]The gender word list comes from Hirota et al. (2023), which can be found in Appendix A.

4 Experiment Setup

To demonstrate the effectiveness and generalization ability of GAMA, we conduct extensive experiments across two vision-language tasks (image captioning and image search) and two benchmarks for measuring gender bias (VisoGender (Hall et al., 2023) and VL-Bias (Zhang et al., 2022)).[4]

4.1 Datasets

The statistics of the datasets are listed in Table 1.

Gender label Following previous work (Hendricks et al., 2018; Wang et al., 2021; Hirota et al., 2023), we utilize ground-truth captions to label the gender attributes of images. Specifically, an image will be labeled as "male (female)" if at least one of its captions contains male (female) gender words and no captions contain female (male) words. Otherwise, it will be labeled as "neutral".

Narrative generation The Localized Narratives dataset (Voigtlaender et al., 2023) for Open Images (Kuznetsova et al., 2020) is utilized to train the narrative generation model.[5]

Image captioning We experiment on MSCOCO captions (Chen et al., 2015). Following Hirota et al. (2023), we use the original MSCOCO training set for training, a subset of the MSCOCO validation set from Zhao et al. (2021) for test,[6] and the remaining images for validation. Each image is associated with five human-annotated captions. Notably, no images in the test set are labeled as "neutral".

Image search Our evaluation for image search involves the MSCOCO (Chen et al., 2015) and Flickr30K (Young et al., 2014) datasets. Following Wang et al. (2021), we use the Karpathy split (Karpathy and Fei-Fei, 2015) for training and validation. As test sets in the Karpathy split are gender-unbalanced, we randomly select the same number of images with their gender-neutral queries [7] under "male", "female" and "neutral" labels for test.

[4]Due to space constraints, experiments on VL-Bias can be found in Appendix D.1.

[5]We ascertain the absence of any overlap between Localized Narratives and other test sets through the cosine similarity among image tensors with a threshold of 0.9. See Appendix D.2 for further studies on the training set size of Localized Narratives.

[6]Access the test split at https://princetonvisualai.github.io/imagecaptioning-bias/.

[7]The queries are available at https://github.com/eric-ai-lab/Mitigate-Gender-Bias-in-Image-Search.

Stage	Task	Dataset	Train	Dev	Test
1	Narrative Generation	Localized Narratives	507,444	41,691	126,020
2	Image Captioning	MSCOCO	82,783	29,724	10,780
	Image Search	MSCOCO	113,287	5,000	1,500
		Flickr30K	29,000	1,000	300
	VisoGender		/	/	690

Table 1: **The statistics of datasets.** We show the counts of images within each split of the datasets.

VisoGender VisoGender (Hall et al., 2023) benchmarks occupation-related gender bias in VLMs. VLMs are required to align images with correct gender pronouns in a resolution task and retrieve the top-K images for a gender-neutral caption of a given occupation in a retrieval task.

4.2 Baselines

Image captioning GAMA is compared with the following methods for image captioning: 1) **Equalizer** (Hendricks et al., 2018), which focuses on the "person" segmentation to make gender-specific predictions; 2) **GAIC**$_{es}$ (Tang et al., 2021), which encourages capturing gender visual evidence through self-guided visual attention; and 3) **LIBRA** (Hirota et al., 2023), which leverages state-of-the-art captioning models to generate high-quality captions and debiases through an additional editing model.

Image search We evaluate GAMA against the following methods for image search: 1) **SCAN-FS** (Wang et al., 2021), which applies a fair sampling method to the representative image search baseline SCAN (Lee et al., 2018); 2) **CLIP-clip** (Wang et al., 2021), which introduces a feature pruning algorithm to the features generated by the PVLM CLIP (Radford et al., 2021); and 3) **FairVLP** (Zhang et al., 2022), which trains the PVLM ALBEF (Li et al., 2021) with counterfactual samples to obtain debiased representations.

VisoGender Following Hall et al. (2023), we evaluate GAMA against state-of-the-art pre-trained vision-language encoders (**CLIP** (Radford et al., 2021), **OpenCLIP** (Cherti et al., 2023) trained on LAION 2B and 400M, **SLIP** (Mu et al., 2022), **DeCLIP** (Li et al., 2022b) and **FILIP** (Yao et al., 2022)) and pre-trained captioning models (**BLIP-2** (Li et al., 2023a) and **GIT** (Wang et al., 2022)).

Model		Gender Bias Metrics ↓			Image Captioning Metrics ↑				
		LIC	Error	BiasAmp	BLEU-4	CIDEr	METEOR	SPICE	CLIPScore
Equalizer †		0.7	8.1	-0.50	27.2	79.8	24.1	16.8	69.9
GAIC$_{es}$ †		1.4	5.9	-0.70	32.6	94.5	27.4	18.3	72.7
NIC (Vinyals et al., 2015)		-0.3	5.7	-1.47	24.6	72.0	24.2	16.5	71.7
SAT (Xu et al., 2015)		-1.4	3.9	-0.48	34.6	95.9	27.8	20.0	73.6
FC (Rennie et al., 2017)		-0.2	4.3	-1.11	32.8	95.9	27.3	19.7	72.9
Att2in (Rennie et al., 2017)		-0.3	4.6	-3.39	35.9	101.7	28.5	20.6	73.8
UpDn (Anderson et al., 2018)	+LIBRA ♭	1.5	4.5	-2.23	37.7	110.1	29.6	22.0	74.6
Transformer (Vaswani et al., 2017)		2.3	5.0	-0.26	33.9	98.7	28.6	20.9	75.7
OSCAR (Li et al., 2020)		0.3	4.6	-1.95	37.2	113.1	31.1	23.2	75.7
ClipCap (Mokady et al., 2021)		-1.5	4.5	-0.57	33.8	100.6	29.3	21.4	76.0
GRIT (Nguyen et al., 2022)		0.7	4.1	1.57	40.5	116.8	30.6	22.6	75.9
GAMA		-1.1	3.4	-3.40	38.2	115.1	31.0	22.7	75.4

Table 2: **Results of image captioning.** The best results are highlighted in green, and the second-best are in blue. For gender bias metrics, lower is better. For image captioning metrics, higher is better. Gender bias metrics are scaled by 100. †: the results are reproduced with official implementation; ♭: the results are retrieved from Hirota et al. (2023).

Dataset	Model	Gender Bias Metrics			Image Search Metrics ↑		
		Bias@1	Bias@5	Bias@10	Recall@1	Recall@5	Recall@10
MSCOCO	SCAN-FS	-0.1043	-0.1716	-0.2392	25.4	54.8	66.1
	CLIP-clip	-0.1173	-0.1940	-0.2528	28.9	57.2	68.5
	FairVLP	0.0334	0.1293	0.1965	58.7	80.2	90.4
	GAMA	0.0273	0.1281	0.1995	63.7	83.6	93.5
Flickr30K	SCAN-FS	-0.1281	-0.1857	-0.2469	36.4	67.6	78.3
	CLIP-clip	-0.1050	-0.1603	-0.2307	64.0	86.5	91.9
	FairVLP	0.0514	0.1012	0.1731	77.4	95.2	97.1
	GAMA	0.0449	0.0941	0.1676	83.1	95.8	97.9

Table 3: **Results of image search.** For gender bias metrics, closer to 0 is better. Baselines are reproduced with official implementation. We report the average across 3 runs. Note that we evaluate models with gender-balanced test sets instead of the Karpathy test sets utilized by Wang et al. (2021) and Zhang et al. (2022).

4.3 Metrics

We present metrics for evaluation, covering task performance, gender bias, and object hallucination. Calculation details are presented in Appendix B.

Task performance metrics are used to evaluate model performance on specific tasks. For *image captioning*, we use established referenced-based metrics BLEU-4 (Papineni et al., 2002), CIDEr (Vedantam et al., 2015), METEOR (Denkowski and Lavie, 2014) and SPICE (Anderson et al., 2016) as well as a reference-free metric CLIPScore (Hessel et al., 2021). For *image search*, we employ Recall@K to calculate the ratio of correct images in the top-K retrieved images.

Gender bias metrics are employed to measure the model performance on gender bias mitigation. For *image captioning*, we use LIC (Hirota et al., 2022b) to measure gender bias in the context of gender words, Error (Hendricks et al., 2018) to assess the gender misclassification ratio of generated captions, and BiasAmp (Zhao et al., 2017) to quantify bias amplification based on word-gender co-occurrence. For *image search*, we follow Wang et al. (2021) and adopt Bias@K to measure gender bias among the top-K images. For *VisoGender*, we follow the setting in Hall et al. (2023). The resolution accuracy gap (Δ_{RA}) is used to measure resolution bias, while Bias@K (Wang et al., 2021), Skew@K (Geyik et al., 2019) and NDKL (Geyik et al., 2019) are used to measure retrieval bias.

Object hallucination metrics are utilized to probe the connection between object hallucination and gender bias. CHAIRs and CHAIRi (Rohrbach et al., 2018) are used to evaluate incorrect object generation at the sentence and object levels, respectively. Rohrbach et al. (2018) used a synonym list (Lu et al., 2018) to map words to MSCOCO objects. We refine it with hierarchical object relationships.[8]

4.4 Implementation details

We use `flan-t5-base` (Chung et al., 2022) as the backbone. Our image encoder is `vit-base-patch16-384` (Dosovitskiy et al., 2020), and its parameters are frozen during training. For narrative generation, we set the temperature $\tau = 0.1$ and conduct further studies in Appendix D.2. Experiments on VisoGender are conducted on GAMA search models under a zero-shot setting. More details are presented in Appendix C.

5 Results and Analysis

In this section, we seek to answer the following research questions: **RQ1:** Does GAMA perform well across different vision-language tasks in task performance metrics and gender bias metrics? (Section 5.1) **RQ2:** Do our gender obfuscation module and two-stage framework help GAMA think before acting? (Section 5.2) **RQ3:** Are object hallucination and gender bias closely related? (Section

[8]Details can be found in Appendix B.2.

5.2) **RQ4:** Can GAMA be generalized to specialized datasets for gender bias measurement under a zero-shot setting? (Section 5.3)

5.1 Overall results

We conduct experiments on GAMA and baselines. Results on image captioning and image search are shown in Table 2 and Table 3, respectively.

To begin with, GAMA outperforms or closely aligns with the baselines in both two tasks. It is noteworthy that most of these baselines are tailored to a specific task. In contrast, GAMA is a task-agnostic method and can be easily applied to different vision-language tasks.

Then we compare GAMA with each baseline separately. First, GAMA comfortably outperforms Equalizer and GAIC$_{es}$. As Equalizer forces the model to focus on persons in images, it loses necessary information to generate the correct context of gender words. Similarly, GAIC$_{es}$ is designed to improve gender classification accuracy with additional gender evide nce. Therefore, these two baselines prove less effective.

Second, LIBRA achieves noticeable results with different captioning models. LIBRA considers gender bias manifested as gender misclassification as well as biased context. However, there are some problems. 1) LIBRA is limited by the performance of captioning models, as it is designed to revise their outputs. 2) LIBRA is designed to edit biased captions after a captioning model, but it is trained with synthesized data instead of actual model outputs, potentially causing error propagation. Consequently, the performance of LIBRA depends on the specific captioning model in use, rendering its results on gender bias metrics less stable.

Third, GAMA also performs better than SCAN-FS, CLIP-clip and FairVLP in image search. We observe that SCAN-FS and CLIP-clip exhibit overcorrection, leading to the underrepresentation of males. These baselines tackle gender bias mainly by removing biased data or features. Given the diverse sources of bias, it cannot be effectively mitigated by merely processing data or debiasing pre-training features.

Finally, we compare models on task performance and gender bias mitigation ability in Figure 3. Results show that GAMA strikes a good balance between task performance and gender bias mitigation.

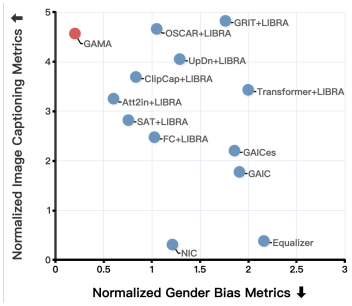

Figure 3: **Comparison on task performance and gender bias mitigation ability.** We normalize the metrics separately and sum the normalized gender bias metrics and image captioning metrics, respectively.

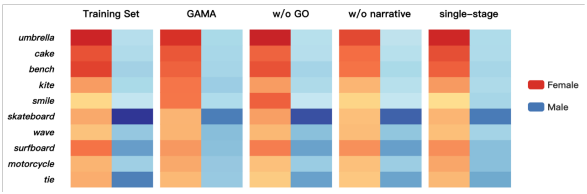

Figure 4: **Heatmap visualization of the co-occurrence frequency between gender attributes and certain words.** We respectively select five words highly co-occurring with females and males in the training set. We show the frequency of co-occurrence between gender attributes and words in the model predictions. Darker colors indicate higher frequencies.

5.2 Ablation study

To further explain the results, we carry out ablation studies on GAMA as well as explore the connection between object hallucination and gender bias in image captioning. Table 4 shows the results. We summarize the main observations as follows.

The gender obfuscation module mitigates gender bias in the context. First, this module effectively obfuscates gender-related information during narrative generation, leading to a reduction in the number of narratives with gender words. Next, the absence of this module is observed to cause an increase in the LIC score, indicating amplified gender bias in the generated context. Additionally, the BiasAmp score shows an increase. We speculate it is because GAMA without this module is exposed to direct gender information, leading to the generation of more words closely related to gender attributes. The rise in HR^g_C@10 and Figure 4 manifest that GAMA without this module hallucinates more objects that highly co-occur with gender words in the training set, thus affirming our hypothesis. Lastly, GAMA without generated narratives also achieves competitive results against previous

Model	Gender Bias Metrics ↓			Gender-related Statistics ↓	Object Hallucination Metrics ↓		Co-occurrence Statistics ↓	
	LIC	Error	BiasAmp	#Gender	CHAIRs	CHAIRi	HR^g_C@10	HR^o_C@10
GAMA	-1.1	3.4	-3.40	55.61	10.94	6.02	38.15	50.30
w/o GO ♦	0.5	3.1	-1.26	62.49	12.81	7.36	41.52	50.09
w/o narrative ★	-0.6	4.1	-2.87	/	13.02	7.84	41.37	51.94
single-stage ♣	-0.9	3.6	-2.13	/	12.40	7.02	39.91	51.76

Table 4: **Results of ablation studies.** #Gender denotes the proportion of generated narratives with gender words in the test set. Note that no images are labeled as "neutral" in the test set. CHAIRs and CHAIRi are calculated based on our refined synonym list. HR^o_C@10 (Li et al., 2023b) and HR^g_C@10 are used to quantify whether the model is prone to hallucinate objects that frequently co-occur with ground-truth objects and gender attributes, respectively, which are detailed in Appendix B.3. ♦: the gender obfuscation module (GO) is removed from narrative generation; ★: the model is trained to generate gender-obfuscated captions instead of gender-obfuscated narratives in the first stage; ♣: the narrative generation stage is removed, and GO is adopted in the answer inference stage. Note that the single-stage framework does not incorporate narratives as input.

Model		Resolution			Retrieval									
		Accuracy ↑	Δ_{RA}		Bias@5		Bias@10		MaxSkew@5 ↓		MaxSkew@10 ↓		NDKL ↓	
			OO	OP	Mean	σ	Mean	σ	Mean	σ	Mean	σ	Mean	σ
CLIP		0.75	-0.14	-0.27	0.11	0.38	0.16	0.22	0.27	0.15	0.18	0.13	0.19	0.07
$OpenCLIP_{2B}$		0.78	-0.07	-0.37	0.10	0.44	0.08	0.23	0.29	0.17	0.18	0.11	0.18	0.07
$OpenCLIP_{400M}$		0.74	-0.27	-0.29	0.17	0.47	0.11	0.22	0.33	0.18	0.16	0.13	0.19	0.07
SLIP		0.60	0.14	0.14	0.06	0.52	0.00	0.24	0.32	0.21	0.17	0.12	0.19	0.09
DeCLIP		0.70	0.06	-0.17	0.11	0.40	0.15	0.26	0.28	0.16	0.20	0.14	0.17	0.07
FILIP		0.45	0.06	0.36	0.01	0.43	0.03	0.26	0.29	0.16	0.17	0.13	0.18	0.07
BLIP-2		0.84	-0.09	0.07	/	/	/	/	/	/	/	/	/	/
GIT		0.84	-0.07	-0.27	/	/	/	/	/	/	/	/	/	/
GAMA	**Flickr30K**	0.81	0.06	0.09	0.04	0.42	0.01	0.26	0.26	0.15	0.17	0.12	0.18	0.07
	MSCOCO	0.82	0.04	-0.09	0.01	0.40	-0.04	0.24	0.26	0.14	0.18	0.11	0.17	0.08

Table 5: **Results of zero-shot experiments on VisoGender.** The best results are highlighted in green, and the second-best are in blue. The results of GAMA are obtained by the search models trained on MSCOCO and Flickr30K, respectively. Other results are retrieved from Hall et al. (2023). Occupation-object (OO) and occupation-participant (OP) cases denote the single-person and two-person resolution tasks, respectively. For Δ_{RA} and Bias@K, closer to 0 is better. We report mean and standard deviation (σ) for the retrieval task across all occupations.

debiasing methods. The results further demonstrate the effectiveness of the gender obfuscation module.

The two-stage framework can facilitate debiased answers. We note an increase in the Error score in the single-stage framework. The narrative generation stage is designed to prompt GAMA to grasp the overall image before giving ultimate answers. The rise in BiasAmp and HR^o_C@10 indicates that the single-stage framework relies more on word co-occurrence for answer generation. Hence, the two-stage framework serves as an effective method to prevent GAMA from prematurely focusing on localized features and hallucinating ignored objects.

Object hallucination and gender bias exhibit a close correlation. As shown in Table 4, mitigating gender bias leads to a decrease in object hallucination. Object hallucination mainly results from the frequent occurrence of certain objects (including persons) and the co-occurrence between objects in the training set (Li et al., 2023b). Similarly, gender bias is manifested as the overrepresentation of a certain gender or the high-frequency co-occurrence between gender attributes and objects in datasets. In essence, gender bias can be thought of as a form of object hallucination in VLMs. Consequently, efforts to mitigate gender bias in VLMs result in a simultaneous reduction of object hallucination.

5.3 Experiments on generalization ability

The datasets employed in the above experiments are not specifically constructed to detect gender bias in VLMs, which may contain gender bias in annotations, e.g., underrepresentation of women (Zhao et al., 2017; Harrison et al., 2023). Consequently, we turn to VisoGender for further experiments. As shown in Table 5, GAMA exhibits good overall performance across bias metrics for both tasks, further demonstrating its generalization ability and effectiveness in gender bias mitigation.

6 Conclusion

In this paper, we present GAMA, a two-stage task-agnostic generation framework to mitigate gender

bias towards vision-language tasks. GAMA is encouraged to gain a comprehensive understanding of images during narrative generation and to rethink gender attributes and answers during answer inference. Experimental results demonstrate GAMA's superiority in both task performance and gender bias metrics over previous methods. Furthermore, we conduct ablation studies and analyze the close connection between object hallucination and gender bias. Finally, we evaluate GAMA under a zero-shot setting to showcase its generalization ability. We hope that GAMA can contribute to the future exploration of fairness in VLMs.

Acknowledgements

We would like to thank the anonymous reviewers for their insightful feedback. This work is supported by National Key Research and Development Program of China (No. 2022YFC3302600).

Limitations

Following previous research (Wang et al., 2021; Hirota et al., 2023), we utilize a gender word list for pre-processing. Although prior studies have covered most gender words, some may still be omitted. One potential solution is to train a model to obfuscate gender information in a sentence via synthesized data.

Another consideration is the narrative generation model. Despite its generalization ability to various vision-language tasks and datasets, the model requires additional computing resources and datasets for training. Due to the boom in large VLMs (LVLMs), we will investigate the feasibility of generating gender-obfuscated narratives using these models.

Theoretically, we can replace the model for answer inference with any state-of-the-art task-specific generative model, which will explored in our future work.

Ethics Statement

In our paper, we focus on mitigating gender bias towards vision-language tasks. Due to the inherent challenges associated with human analysis, including substantial manual effort and time investment, we leverage existing datasets and benchmarks for gender bias measurement. While quantitative metrics provide valuable insights, we acknowledge their potential limitations in capturing nuanced gender bias. Additionally, current datasets and benchmarks only consider binary gender, which oversimplifies the intricate and diverse nature of gender identity. Therefore, GAMA remains to be improved. We hope that the advancements in GAMA, as presented in this paper, will serve as a catalyst for inspiring further valuable research in gender bias mitigation towards vision-language tasks.

References

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. NoCaps: Novel Object Captioning at Scale. In *2019 IEEE/CVF International Conference on Computer Vision*, pages 8947–8956.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: Semantic Propositional Image Caption Evaluation. In *European Conference on Computer Vision 2016*, pages 382–398.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *2015 IEEE International Conference on Computer Vision*, pages 2425–2433.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Shruti Bhargava and David Forsyth. 2019. Exposing and Correcting the Gender Bias in Image Captioning Datasets and Models. *arXiv preprint arXiv:1912.00578*.

Ali Furkan Biten, Lluís Gómez, and Dimosthenis Karatzas. 2022. Let there be a clock on the beach: Reducing Object Hallucination in Image Captioning. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2473–2482.

Roger Brown. 1958. How Shall a Thing Be Called? *Psychological review*, 65(1):14.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1597–1607.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft COCO Captions: Data Collection and Evaluation Server. *arXiv preprint arXiv:1504.00325*.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible Scaling Laws for Contrastive Language-Image Learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, and others. 2022. Scaling Instruction-Finetuned Language Models. *arXiv preprint arXiv:2210.11416*.

Michael Denkowski and Alon Lavie. 2014. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and others. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv preprint arXiv:2010.11929*.

Qingkai Fang and Yang Feng. 2022. Neural Machine Translation with Phrase-Level Universal Visual Representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5687–5698, Dublin, Ireland. Association for Computational Linguistics.

Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. 2019. Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2221–2231.

Siobhan Mackenzie Hall, Fernanda Gonçalves Abrantes, Hanwen Zhu, Grace Sodunke, Aleksandar Shtedritski, and Hannah Rose Kirk. 2023. VisoGender: A Dataset for Benchmarking Gender Bias in Image-Text Pronoun Resolution. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Sophia Harrison, Eleonora Gualdoni, and Gemma Boleda. 2023. Run Like a Girl! Sport-Related Gender Bias in Language and Vision. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14093–14103, Toronto, Canada. Association for Computational Linguistics.

Jacqueline He, Mengzhou Xia, Christiane Fellbaum, and Danqi Chen. 2022. MABEL: Attenuating Gender Bias using Textual Entailment Data. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9681–9702, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9726–9735.

Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women Also Snowboard: Overcoming Bias in Captioning Models. In *European Conference on Computer Vision 2018*, pages 793–811.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yusuke Hirota, Yuta Nakashima, and Noa Garcia. 2022a. Gender and Racial Bias in Visual Question Answering Datasets. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1280–1292.

Yusuke Hirota, Yuta Nakashima, and Noa Garcia. 2022b. Quantifying Societal Bias Amplification in Image Captioning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13440–13449.

Yusuke Hirota, Yuta Nakashima, and Noa Garcia. 2023. Model-Agnostic Gender Debiased Image Captioning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15191–15200.

Sepehr Janghorbani and Gerard De Melo. 2023. Multi-Modal Bias: Introducing a Framework for Stereotypical Bias Assessment beyond Gender and Race in Vision–Language Models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1725–1735, Dubrovnik, Croatia. Association for Computational Linguistics.

Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. 2020. In Defense of Grid Features for Visual Question Answering. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10264–10273.

Andrej Karpathy and Li Fei-Fei. 2015. Deep Visual-Semantic Alignments for Generating Image Descriptions. *2015 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137.

Matthew Kay, Cynthia Matuszek, and Sean A. Munson. 2015. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3819–3828.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, and others. 2020. The Open Images Dataset V4: Unified Image Classification, Object Detection, and Visual Relationship Detection at Scale. *International Journal of Computer Vision*, 128(7):1956–1981.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked Cross Attention for Image-Text Matching. In *European Conference on Computer Vision 2018*, pages 212–228.

Bei Li, Chuanhao Lv, Zefan Zhou, Tao Zhou, Tong Xiao, Anxiang Ma, and JingBo Zhu. 2022a. On Vision Features in Multimodal Machine Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6327–6337, Dublin, Ireland. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *arXiv preprint arXiv:2301.12597*.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. In *Advances in Neural Information Processing Systems*, volume 34, pages 9694–9705.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. OSCAR: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In *European Conference on Computer Vision 2020*.

Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. 2022b. Supervision Exists Everywhere: A Data Efficient Contrastive Language-Image Pre-training Paradigm. In *International Conference on Learning Representations*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating Object Hallucination in Large Vision-Language Models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2018. Neural Baby Talk. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7219–7228, Los Alamitos, CA, USA. IEEE Computer Society.

Ron Mokady, Amir Hertz, and Amit H. Bermano. 2021. ClipCap: CLIP Prefix for Image Captioning. *arXiv preprint arXiv:2111.09734*.

Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. 2022. SLIP: Self-Supervision Meets Language-Image Pre-Training. In *European Conference on Computer Vision 2022*, pages 529–544.

Van-Quang Nguyen, Masanori Suganuma, and Takayuki Okatani. 2022. GRIT: Faster And Better Image Captioning Transformer Using Dual Visual Features. In *European Conference on Computer Vision 2022*, pages 167–184.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, volume 32.

Leigang Qu, Meng Liu, Jianlong Wu, Zan Gao, and Liqiang Nie. 2021. Dynamic Modality Interaction Modeling for Image-Text Retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1104–1113.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and others. 2021. Learning Transferable Visual Models from Natural Language Supervision. In *International Conference on Machine Learning*, volume 139, pages 8748–8763.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-Critical Sequence Training for Image Captioning. In *2017*

IEEE Conference on Computer Vision and Pattern Recognition, pages 1179–1195.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object Hallucination in Image Captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium. Association for Computational Linguistics.

Candace Ross, Boris Katz, and Andrei Barbu. 2021. Measuring Social Biases in Grounded Vision and Language Embeddings. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 998–1008, Online. Association for Computational Linguistics.

Ashish Seth, Mayur Hemani, and Chirag Agarwal. 2023. DeAR: Debiasing Vision-Language Models with Additive Residuals. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6820–6829.

Omar Shaikh, Hongxin Zhang, William Held, Michael Bernstein, and Diyi Yang. 2023. On Second Thought, Let's Not Think Step by Step! Bias and Toxicity in Zero-Shot Reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4454–4470, Toronto, Canada. Association for Computational Linguistics.

Tejas Srinivasan and Yonatan Bisk. 2022. Worst of Both Worlds: Biases Compound in Pre-trained Vision-and-Language Models. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing*, pages 77–85, Seattle, Washington. Association for Computational Linguistics.

Ryan Steed and Aylin Caliskan. 2021. Image Representations Learned With Unsupervised Pre-Training Contain Human-like Biases. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 701–713.

Ruixiang Tang, Mengnan Du, Yuening Li, Zirui Liu, Na Zou, and Xia Hu. 2021. Mitigating Gender Bias in Captioning Systems. In *Proceedings of the Web Conference 2021*, pages 633–645.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-Based Image Description Evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and Tell: A Neural Image Caption Generator. In *2015 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164.

Paul Voigtlaender, Soravit Changpinyo, Jordi Pont-Tuset, Radu Soricut, and Vittorio Ferrari. 2023. Connecting Vision and Language with Video Localized Narratives. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2461–2471.

Feng Wang and Huaping Liu. 2021. Understanding the Behaviour of Contrastive Loss. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2495–2504.

Haoran Wang, Ying Zhang, Zhong Ji, Yanwei Pang, and Lin Ma. 2020. Consensus-Aware Visual-Semantic Embedding for Image-Text Matching. In *European Conference on Computer Vision 2020*, pages 18–34, Cham. Springer International Publishing.

Jialu Wang, Yang Liu, and Xin Wang. 2021. Are Gender-Neutral Queries Really Gender-Neutral? Mitigating Gender Bias in Image Search. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1995–2008, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. GIT: A Generative Image-to-text Transformer for Vision and Language. *arXiv preprint arXiv:2205.14100*.

Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. 2019. Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations. In *2019 IEEE/CVF International Conference on Computer Vision*, pages 5309–5318.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. 2020. Measuring and Reducing Gendered Correlations in Pre-Trained Models. *arXiv preprint arXiv:2010.06032*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *Proceedings of the 32nd International*

Conference on International Conference on Machine Learning, pages 2048–2057.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2022. FILIP: Fine-grained Interactive Language-Image Pre-Training. In *International Conference on Learning Representations*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference Over Event Descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Yi Zhang, Junyang Wang, and Jitao Sang. 2022. Counterfactually Measuring and Eliminating Social Bias in Vision-Language Pre-Training Models. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4996–5004. Event-place: Lisboa, Portugal.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal Chain-of-Thought Reasoning in Language Models. *arXiv preprint arXiv:2302.00923*.

Dora Zhao, Angelina Wang, and Olga Russakovsky. 2021. Understanding and Evaluating Racial Biases in Image Captioning. In *2021 IEEE/CVF International Conference on Computer Vision*, pages 14810–14820.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Kankan Zhou, Eason Lai, and Jing Jiang. 2022. VLStereoSet: A Study of Stereotypical Bias in Pretrained Vision-Language Models. In *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 (Volume 1: Long Papers)*, pages 527–538, Online only. Association for Computational Linguistics.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2023. Analyzing and Mitigating Object Hallucination in Large Vision-Language Models. *arXiv preprint arXiv:2310.00754*.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *2017 IEEE International Conference on Computer Vision*, pages 2242–2251.

A List of Gender Words

We use the gender words from Hirota et al. (2023), which are listed in Table 6.

B Metrics

B.1 Gender bias metrics

LIC As for LIC, we follow the method proposed by Hirota et al. (2022b).

First, we pre-process captions by masking gender words. To measure gender bias amplification in image captioning models, we need to quantify the difference between bias in the generated captions set $\hat{\mathcal{D}}$ and bias in the ground-truth captions in the training data $\mathcal{D}$. Then we train two gender classifiers f and $\hat{f}$ on the two masked caption sets $\mathcal{D}$ and $\hat{\mathcal{D}}$, respectively. Finally, we compare the accuracy of two gender classifiers as follows:

$$\text{LIC}_D = \frac{1}{|\mathcal{D}|} \sum_{(y,a)\in\mathcal{D}} s_a(y)\mathbb{1}[f(y) = a], \quad (16)$$

$$\text{LIC}_M = \frac{1}{|\hat{\mathcal{D}}|} \sum_{(\hat{y},a)\in\hat{\mathcal{D}}} \hat{s}_a(\hat{y})\mathbb{1}[\hat{f}(\hat{y}) = a], \quad (17)$$

$$\text{LIC} = \text{LIC}_M - \text{LIC}_D, \quad (18)$$

where y and $\hat{y}$ are the ground-truth caption and generated caption, respectively. We denote a as the gender label of the caption. $s_a(\cdot)$ and $\hat{s}_a(\cdot)$ are the probabilities that f and $\hat{f}$ classify the gender of the caption as a, respectively.

The higher accuracy of the classifier trained on the masked captions indicates that the context contains more information related to gender. LIC > 0 indicates that the model amplifies gender bias with respect to the training data, and mitigates it otherwise.

Following Hirota et al. (2023), we construct the classifiers with `bert-base-uncased` (Devlin et al., 2019) as the backbone and two fully connected layers with Leaky ReLU activation on top. We finetune the model for 5 epochs with a learning rate of 1×10^{-5}.

Error Error represents the ratio of gender misclassification in the generated captions. Following Hendricks et al. (2018) and Hirota et al. (2023), we utilize the gender word list to identify the gender label of the generated captions as detailed in Section 4.1. And "neutral" labels are not considered as errors.

Gender	Gender Words
Female	woman, female, lady, mother, girl, aunt, wife, actress, princess, waitress, sister, queen, chairwoman, policewoman, girlfriend, pregnant, daughter, she, her, hers, herself
Male	man, male, father, gentleman, boy, uncle, husband, actor, prince, waiter, son, brother, guy, emperor, dude, cowboy, boyfriend, chairman, policeman, he, his, him, himself

Table 6: Gender word list.

BiasAmp BiasAmp is proposed to quantify the bias amplification of the model. Following Zhao et al. (2017) and Hirota et al. (2023), we use the top $1,000$ common words in captions and filter the words that are not strongly associated with humans, leaving a set $\mathcal{L}$ of high-frequency words.[9]

We calculate the bias of the word $l \in \mathcal{L}$ on the gender $a \in \mathcal{A} = \{m, f\}$ as follows:

$$b_{a,l} = \frac{c_{a,l}}{\sum_{a\in\mathcal{A}} c_{a,l}}, \quad (19)$$

$$\hat{b}_{a,l} = \frac{\hat{c}_{a,l}}{\sum_{a\in\mathcal{A}} \hat{c}_{a,l}}, \quad (20)$$

where $c_{a,l}$ and $\hat{c}_{a,l}$ are the number of co-occurrences of a and l in the training data and in the model predictions, respectively. Then bias amplification is defined as:

$$\text{BiasAmp} = \frac{1}{\mathcal{L}} \sum_{a\in\mathcal{A},l\in\mathcal{L}} (\hat{b}_{a,l} - b_{a,l})\mathbb{1}[b_{a,l} > \frac{1}{|\mathcal{A}|}]. \quad (21)$$

BiasAmp > 0 indicates that gender bias is amplified by the model, and otherwise mitigated.

Bias@K Wang et al. (2021) proposed Bias@K to measure gender bias in image search by comparing the proportions of masculine and feminine images in search results. Let q denote the search query and $\mathcal{R}_q^K$ denote the top-K retrieved image set. The gender bias of $\mathcal{R}_q^K$ is defined as:

$$\text{Bias@K}(q) = \begin{cases} 0, & \text{if } N_m + N_f = 0 \\ \frac{N_m - N_f}{N_m + N_f}, & \text{otherwise} \end{cases}, \quad (22)$$

where N_m and N_f denote the number of images labeled "male" and "female" in $\mathcal{R}_q^K$, respectively. Then Bias@K is calculated as the average of Bias@K(q) over all queries.

[9]The word list in available at https://github.com/uclanlp/reducingbias.

A positive Bias@K indicates a higher frequency of retrieving images featuring males compared to females. This metric is most suitable when the candidate images are gender-balanced, as it does not consider the distribution of candidates.

MaxSkew@K Hall et al. (2023) measured the difference between the desired proportion of gender attributes in $\mathcal{R}_q^K$ and the actual proportion. Let $\mathcal{A} = \{m, f\}$ denote the gender attribute set. The skew of $\mathcal{R}_q^K$ for the gender attribute $a \in \mathcal{A}$ is defined as:

$$\text{Skew}_a@\text{K}(q) = \ln \frac{P_{\mathcal{R}_q^K,q,a}}{P_{d,q,a}}, \tag{23}$$

where $P_{d,q,a}$ and $P_{\mathcal{R}_q^K,q,a}$ are desired proportion of gender attributes in the test set and the actual proportion in $\mathcal{R}_q^K$, respectively.

Then we calculate the maximum value of Skew@K among all gender attributes for the retrieved images:

$$\text{MaxSkew}@\text{K}(q) = \max_{a \in \mathcal{A}} \text{Skew}_a@\text{K}(q), \tag{24}$$

Finally, MaxSkew@K is calculated as the average of MaxSkew@K(q) over all queries.

NDKL Geyik et al. (2019) and Hall et al. (2023) measured the distance of the prediction distribution from a fair distribution over all candidate results. Let $\mathcal{R}_q$ denote the candidate set of the query q. The NDKL of $\mathcal{R}_q$ is defined as:

$$\text{NDKL}(q) = \frac{1}{Z} \sum_{K=1}^{|\mathcal{R}_q|} \frac{1}{\log_2(K+1)} d_{KL}(D_{\mathcal{R}_q^K} \| D), \tag{25}$$

where $Z = \sum_{K=1}^{|\mathcal{R}_q|} \frac{1}{\log_2(K+1)}$ is a normalizing factor, and $d_{KL}(\cdot)$ denotes the KL-divergence. $D_{\mathcal{R}_q^K}$ and D denote the actual distribution of gender attributes over the top-K retrieved images and the desired distribution, respectively.

Resolution bias Hall et al. (2023) defined resolution accuracy (RA) as the proportion of correctly resolved pronouns. Let $\mathcal{O}$ denote the occupation set. We calculate the resolution accuracy of the gender attribute a on the occupation $o \in \mathcal{O}$ as:

$$RA_a(o) = \frac{\hat{n}_{a,o}}{n_{a,o}}, \tag{26}$$

where $n_{a,o}$ denotes the total number of pronouns of the gender attribute a in occupation o, and $\hat{n}_{a,o}$ denotes the number of correctly resolved pronouns of a in o.

Then the resolution bias on the occupation o is defined as the gender resolution accuracy gap:

$$\Delta_{RA}(o) = RA_m(o) - RA_f(o). \tag{27}$$

Finally, we calculate the resolution bias Δ_{RA} as the average of $\Delta_{RA}(o)$ over all occupations. $\Delta_{RA} > 0$ indicates that the model performs better in resolving males within occupations, and vice versa.

B.2 Object Hallucination Metrics

Definition CHAIR (Rohrbach et al., 2018) is a popular metric for object hallucination measurement with two variants, CHAIRi and CHAIRs. CHAIRi and CHAIRs evaluate object hallucination at the object and sentence levels, respectively. Let N_o^H denote the number of hallucinated objects, N_o denote the total number of mentioned objects, N_c^H denote the number of captions with hallucinated objects, and N_c denote the total number of captions. CHAIRi and CHAIRs are defined as:

$$\text{CHAIRi} = \frac{N_o^H}{N_o}, \tag{28}$$

$$\text{CHAIRs} = \frac{N_c^H}{N_c}. \tag{29}$$

CHAIRi describes the proportion of hallucinated objects among all generated objects. CHAIRs describes the proportion of generated captions with hallucinated objects.

Details Rohrbach et al. (2018) utilized a synonym list (Lu et al., 2018) to map words to MSCOCO objects (e.g., "player" to "person"). However, it is notable that the list is coarse-grained. For instance, it considers terms like "woman" and "man" as well as "purse" and "briefcase" as synonyms. Although "woman" and "man" both fall under the"person" class, and "purse" and "briefcase" belong to the category of bags, it is crucial to recognize that they represent distinct objects. Therefore, we refine the list with hierarchical object relationships among objects based on the fine-grained classes defined in NoCaps (Agrawal et al., 2019). If a word in a sub-category (e.g., "woman") is predicted in its super-category (e.g., "person"), we do not consider it as a hallucinated object. Conversely, if a word in a category (e.g., "woman") is predicted in its sibling category (e.g., "man"), we consider it as a hallucinated object.

B.3 Hit Ratio

Li et al. (2023b) utilized the hit ratio (HR_C^o) to measure the object co-occurrence in object hallucination. They demonstrated that VLMs mostly hallucinate objects that frequently co-occur with ground-truth objects in the image.

Let $\mathcal{H}_i$ denote the set of hallucinated objects in the i-th image, and $\mathcal{C}_o$ denote the set of the top-K frequently co-occurring objects with $\hat{o}$ in the training set. The top-K hit ratio of the probing object $\hat{o}$ is defined as:

$$\text{HR}_C^o@\text{K}(\hat{o}) = \frac{1}{M_{\hat{o}}} \sum_{i=1}^{M_{\hat{o}}} \frac{|\mathcal{H}_i \cap \mathcal{C}_{\hat{o}}|}{|\mathcal{H}_i|}, \tag{30}$$

where $M_{\hat{o}}$ is the total number of images containing $\hat{o}$.

Similarly, we define the top-K hit ratio of the probing gender a as:

$$\text{HR}_C^g@\text{K}(a) = \frac{1}{M_a} \sum_{i=1}^{M_a} \frac{|\mathcal{H}_i \cap \mathcal{C}_a|}{|\mathcal{H}_i|}, \tag{31}$$

where M_a is the total number of images with the gender label a, and $\mathcal{C}_a$ denotes the set of the top-K frequently co-occurring objects with the gender a.

C Implementation Details

We select the best models based on the loss on the validation set for all tasks. Following Hendricks et al. (2018) and Tang et al. (2021), we set the beam size as 5 during inference. Table 7 shows our hyperparameters for training in different tasks. Table 8 lists the input and output formats of GAMA.

We implement GAMA with PyTorch (Paszke et al., 2019) and Huggingface Transformers (Wolf et al., 2020). We train the models with a single NVIDIA Ge-Force RTX 4090 GPU. More implementation details can be found in our code.

Image search With a probability of 50%, we randomly sample the caption of another training image to create a negative pair for training. The model predicts answers with "yes" or "no" as shown in Table 8. We take the token probability of the decoder as the match score.

VisoGender In VisoGender, each occupation appears with two caption templates: a person with a possessive pronoun to an object (e.g., "the doctor and his/her stethoscope") and a person with a possessive pronoun to a participant (e.g., "the doctor and his/her patient"). We evaluate the performance of the trained GAMA search models on MSCOCO and Flickr30K without additional training, respectively. Our experiments on VisoGender follow the original setting proposed by Hall et al. (2023). In the resolution task, we provide the model with queries containing "his" and "her", respectively. We determine the answers based on the match scores. In the retrieval task, we use neutral queries incorporating the pronoun "their" and provide the model with candidate images featuring correct occupations with balanced gender distributions.

D Additional Experiments

D.1 Experiments on VL-Bias

We conduct experiments on VL-Bias (Zhang et al., 2022) under a zero-shot setting.

Dataset VL-Bias serves as a benchmark for measuring gender bias in VLMs, encompassing 52 activities and 13 occupations related to humans.[10] The dataset includes images sourced from the Internet, as well as existing image datasets such as MSCOCO (Chen et al., 2015) and Flickr30K (Young et al., 2014). The captions are in the format of "The {gender} is {target}", where "target" represents an activity (e.g., "shopping") or an occupation (e.g., "engineer"). In total, VL-Bias comprises 24K image-text pairs, including 13K pairs for the 52 activities and 11K for the 13 occupations.

Baselines Following Zhang et al. (2022), we evaluate GAMA against three common and effective debiasing methods: 1) Gender Swapping (**GS**) (Zhao et al., 2018), which swaps gender words in the input text to mitigate the effect of gender attributes, and is adapted to vision-language tasks based on CycleGAN (Zhu et al., 2017); 2) Dropout Regularization (**DR**) (Webster et al., 2020), which is designed to prevent model from overfitting to gender attributes by increasing the dropout rate; and 3) **FairVLP** (Zhang et al., 2022), which trains the PVLM ALBEF (Li et al., 2021) with counterfactual samples to obtain debiased representations.

Metrics Zhang et al. (2022) proposed the vision-language bias for VL-Bias. Let t denote a target word, and a denote the gender attribute. The vision-language bias on t towards a on an image-text pair

[10]The dataset is available at https://github.com/VL-Bias/VL-Bias.

Hyper-parameters	Narrative Generation	Image Captioning	Image Search	
			MSCOCO	Flickr30K
Hidden Dimension d	768	768	768	768
Maximum Epochs	15	10	5	5
Learning Rate	4×10^{-5}	1×10^{-4}	2×10^{-5}	3×10^{-5}
Weight Decay	0.01	0.01	0.01	0.01
Batch Size	20	24	32	32
Maximum Input Sequence Length	128	256	256	256
Maximum Output Sequence Length	128	64	3	3

Table 7: Hyper-parameter settings of GAMA.

Stage	Task	Input Sequence	Target Sequence
1	Narrative Generation	Task: What can you see in this image? Please describe it in detail. Answer:	[NARRATIVE]
2	Image Captioning	Context: [NARRATIVE]. Task: Generate a short caption of the image. Answer:	[CAPTION]
	Image Search	Context: [NARRATIVE]. Query: [CAPTION]. Question: Do the image and the query match? Answer:	Yes/No

Table 8: **The input and output formats of GAMA.** For image search, we take the token probability of the decoder as the match score. For the second stage, model-generated narratives are utilized instead of relying on ground-truth narratives from datasets.

s can be formulated as:

$$B_s(t,a) = \frac{P_T(t|s_c) - P_T(t|s)}{P_A(a|s_c) - P_A(a|s)}, \quad (32)$$

where s_c is the counterfactual image-text pair of s. $P_T(t|s)$ denotes the probability that the model predicts the masked target word as t, and $P_A(a|s)$ denotes the probability that the model predicts the masked gender word as a.

Let $\mathcal{A} = \{m, f\}$ denote the gender attribute set, and S_t denote all image-text pairs with the target t. The gender bias over the target t is defined as:

$$B_{VL}(t) = \frac{1}{|S_t|} \sum_{s \in S_t} B_s(t,m) - B_s(t,f). \quad (33)$$

A positive $B_{VL}(t)$ indicates that the target t is biased towards males, and vice versa. The gender bias of the dataset is calculated as the average of $B_{VL}(t)$ over all targets.

Implementation details As the baseline models are pre-trained on MSCOCO (Chen et al., 2015), we utilize the GAMA search model trained on MSCOCO for evaluation. We take the token probability of the decoder as the probability for a target or a gender attribute. Counterfactual texts are constructed by reversing the gender words, while counterfactual images are generated with an adversarial attack based on the official implementation of Zhang et al. (2022).

Experimental results As shown in Table 9, GAMA obtains remarkable debiasing performance against baselines. The results underscore its generalization ability as well as its effectiveness in gender bias mitigation.

Model	Activity 13K	Occupation 11K
GS	11.21	12.47
DR	11.17	13.52
FairVLP	6.97	7.74
GAMA	5.96	6.83

Table 9: **Results on VL-Bias.** The best results are highlighted in green. For the results, closer to 0 is better. The baseline results are retrieved from Zhang et al. (2022).

D.2 Further study

In this section, we analyze the impact of the temperature hyper-parameter, the data size of Localized Narratives (Voigtlaender et al., 2023), and the frozen parameters.

Temperature We show the results in Table 10. The contrastive loss with a larger temperature is less sensitive to the hard negative samples as discussed in prior work (Wang and Liu, 2021). Therefore, it is hard for GAMA with a large temperature to distinguish gender-related features from gender-masked features, leading to increased LIC and BiasAmp scores.

Data size Although we have demonstrated the effectiveness of the model without narrative generation in Section 5.2, we wonder about the influence of the data size on the results. Therefore, we conduct an ablation study on the training set size of

Temperature	Gender Bias Metrics↓			Image Captioning Metrics ↑				
	LIC	Error	BiasAmp	BLEU-4	CIDEr	METEOR	SPICE	CLIPScore
0.01	-1.4	3.5	-3.42	37.9	115.2	30.9	22.5	75.5
0.1	-1.1	3.4	-3.40	38.2	115.1	31.0	22.7	75.4
1	-0.7	3.3	-3.28	37.7	114.6	30.6	22.4	75.6

Table 10: Results of the ablation study on the temperature hyper-parameter.

Localized Narratives (Voigtlaender et al., 2023) for narrative generation.

As our objective is to alleviate gender bias, we randomly select the images categorized under the "Person" class for experiments. Results are illustrated in Table 11 and Table 12.

We observe that the data size seems to have little impact on GAMA's task performance and its gender bias mitigation ability in image captioning. However, the data size affects the generalization ability of GAMA on VisoGender. We consider that it is because a large training set enhances the zero-shot generalization ability of GAMA in narrative generation, thereby ensuring robust model performance during answer inference.

Parameters To minimize the cost of training the model, we investigate the feasibility of reducing additional training. We experiment with freezing a portion of the parameters in GAMA, and the results are summarized in Table 13 below.

Freezing the T5 encoder in GAMA leads to results that are only slightly inferior to those of unfrozen GAMA, which is an encouraging finding. However, the outcomes of freezing both the flan-T5 encoder and decoder are less promising. We think it is because the fusion of language features and vision features creates a new feature space, which is distinct from the original features learned by T5. Consequently, the frozen decoder cannot be effectively adapted to this new feature space.

Data Size	Gender Bias Metrics ↓			Image Captioning Metrics ↑				
	LIC	Error	BiasAmp	BLEU-4	CIDEr	METEOR	SPICE	CLIPScore
100%	-1.1	3.4	-3.40	38.2	115.1	31.0	22.7	75.4
100% Person	-1.0	3.2	-3.35	38.0	114.5	30.9	22.5	75.3
30% Person	-0.8	3.5	-3.02	37.9	113.6	30.6	22.2	75.2

Table 11: **Results of the ablation study on the data size in image captioning.** "100% Person" and "30% Person" denote that we randomly construct the training set for narrative generation with 100% and 30% of the images under the "Person" class, respectively. We report the average across 3 runs.

Data Size	Resolution			Retrieval									
	Accuracy ↑	Δ_{RA}		Bias@5		Bias@10		MaxSkew@5 ↓		MaxSkew@10 ↓		NDKL ↓	
		OO	OP	Mean	σ	Mean	σ	Mean	σ	Mean	σ	Mean	σ
100%	0.82	0.04	-0.09	0.01	0.40	-0.04	0.24	0.26	0.14	0.18	0.11	0.17	0.08
100% Person	0.81	0.06	0.10	0.07	0.39	0.10	0.23	0.26	0.16	0.19	0.12	0.18	0.08
30% Person	0.76	0.14	0.19	0.11	0.39	0.15	0.26	0.31	0.16	0.21	0.13	0.19	0.09

Table 12: Results of the ablation study on the data size in VisoGender.

Model	Gender Bias Metrics ↓			Image Captioning Metrics ↑				
	LIC	Error	BiasAmp	BLEU-4	CIDEr	METEOR	SPICE	CLIPScore
GAMA	-1.1	3.4	-3.40	38.2	115.1	31.0	22.7	75.4
Frozen encoder	-1.2	3.4	-3.35	38.0	114.6	30.7	22.2	75.1
Frozen encoder & decoder	-0.6	3.7	-2.79	37.1	112.5	29.8	21.6	74.4
w/o GO Frozen encoder	0.4	3.2	-1.12	37.8	114.2	30.3	22.1	74.9

Table 13: **Results of partial parameters frozen.** We freeze the T5 encoder in GAMA, both the T5 encoder and decoder in GAMA, and the T5 encoder in GAMA w/o GO, respectively.