How Bias Binds: Measuring Hidden Associations for Bias Control in Text-to-Image Compositions

Jeng-Lin Li[1], **Ming-Ching Chang**[2], **Wei-Chao Chen**[1]

[1]Inventec Corporation, No.66, Hougang St., Shilin Dist., Taipei City 111059, Taiwan
[2]University at Albany, State University at New York, Albany, NY, 12222, USA
li.johncl@inventec.com, mchang2@albany.edu, chen.wei-chao@inventec.com

Abstract

Text-to-image generative models often exhibit bias related to sensitive attributes. However, current research tends to focus narrowly on single-object prompts with limited contextual diversity. In reality, each object or attribute within a prompt can contribute to bias. For example, the prompt "an assistant wearing a pink hat" may reflect female-inclined biases associated with a pink hat. Neglecting joint semantic bindings in prompts leads to significant failures of current debiasing methods. We present a preliminary investigation into how bias manifests under semantic binding, where contextual associations between objects and attributes affect generative outcomes. We demonstrate that the underlying bias distribution can be amplified based on these associations. Therefore, we introduce a bias adherence score that quantifies how specific object-attribute bindings activate bias. To delve deeper, we develop a training-free context-bias control framework to explore how token decoupling can facilitate the debiasing of semantic bindings. This framework achieves over 10% debiasing improvement in compositional generation tasks. Our analysis of bias scores across various attribute-object bindings and token decorrelation highlights a fundamental challenge: reducing bias without disrupting essential semantic relationships. These findings expose critical limitations in current debiasing approaches when applied to semantically bound contexts, underscoring the need to reassess prevailing bias mitigation strategies.

Introduction

Diffusion models have unlocked various applications of text-to-image (T2I) generative models. However, these models often capture spurious correlations from the training datasets, which can introduce bias during deployment in testing scenarios (Wu, Nakashima, and Garcia 2024). Unrecognized biases embedded in models lead to skewed decision-making and societal impacts, including the reinforcement of stereotypes and concerns about fairness, which are inclined to amplify and perpetuate existing societal inequities (Luccioni et al. 2023). Despite the biased correlation being observed years ago (Grover et al. 2019), many large-scale benchmarks still report the persistent biases in state-of-the-art T2I generative models (Vice et al. 2025) and biases during further distillation (Luo et al. 2024). The phenomena signify the embedded biased correlations from diverse and complex data distributions.

Biases in T2I models stem from latent correlations in training data, entangled with related attributes (Udandarao et al. 2024), making debiasing difficult without harming semantic fidelity. Recent methods address this by avoiding retraining, reducing reliance on balanced data (Smith et al. 2023) and counterfactual examples (Jung et al. 2024). Common techniques for bias mitigation include manipulating prompts (Bansal et al. 2022; Ding et al. 2021; Chuang et al. 2023), inserting features (Li et al. 2024b), rescaling noise guidance, and learning inclusive tokens (Zhang et al. 2023; Shrestha et al. 2024; Teo et al. 2024). The common underlying notion of modifying the concept-bias relationship frequently compromises essential generative structures, leading to pronounced collapse. Therefore, we investigate the deeper relationships between biases, objects, and attributes.

Modern bias assessment studies focus on fairness and bias metrics independent of complex prompts. Nevertheless, most studies focus on generating images of a single object (concept), failing to account for the increasing diversity of generation targets. For example, gender biases in occupational representations can be mitigated in simple prompts like "a headshot of an assistant" using current debiasing algorithms. However, in context-rich prompts such as "a headshot of an assistant wearing a pink hat," these biases may persist, indicating limitations in existing debiasing techniques. The contexts "wearing a pink hat" vary the token probability and amplify the spurious correlation. We empirically find significant failure using state-of-the-art debiasing algorithms (Li et al. 2024b; Teo et al. 2024; Parihar et al. 2024) in Figure 1 (right), which results in an unreal visual style, missing professional properties, and even violation of the Stable Diffusion Safe Checker. This prompts the research question: *How do object and attribute bindings influence generative bias beyond the original single-object setting?* As the first to explore this question, we focus the discussion on gender bias in occupations involving human-associated objects within compositional bias.

In this work, we quantitatively assess bias using a bias adherence score (BA-Score) in the compositional T2I generation task and reveal the insights for designing a training-free context bias control (CBC) framework. Our focus is on

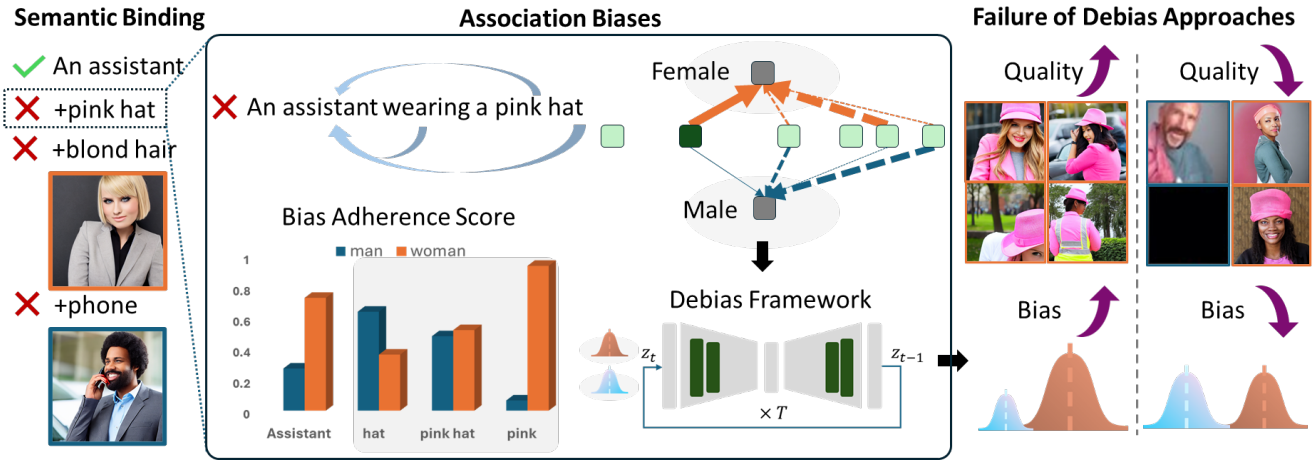

Figure 1: Conceptual illustration of the association bias in compositional text-to-image generation. Orange and blue boxes denote the gender bias. Current debiasing frameworks neglect the hidden bias stemming from semantic bindings, such as pink hat or blond hair, causing low-quality generation or persistent biases after debiasing.

exploring how contextual biases influence the main object and impact the quality of its generation. Figure 1 shows the prompt "an assistant wearing a pink hat" producing lower-quality generations for males than females due to the underlying biases toward females. This underscores a critical consideration: the pursuit of unbiased generation may compromise reliability, as it can conflict with the inherently learnt features in the model. Therefore, increasing importance emerges in quantifying biases in complex contexts alongside the inter-object relationship.

Our CBC framework decouples the embedding into the attribute-orthogonal embedding and the attribute-residual embedding. We use attribute-orthogonal embedding as the model input and adaptively inject the attribute residual to steer the bias tendency at each generation step. At each iteration, we measure the distance between the latent embedding and the attribute cluster centers. The BA-Score serves as an initialization of this bias measure, as the latent embedding is unavailable in the initial stage without performing any forward step. Our experiments demonstrate the promising bias mitigation using our training-free CBC framework and also reveal the importance of the BA-score initialization. We summarize our contribution as follows:

- The investigation of underexplored semantic binding biases in T2I generation, with embedding analysis uncovering key debiasing challenges.
- A training-free bias control framework yielding over 10% improvements of debiasing performance without quality degradation in compositional generation.
- In-depth experiments to reveal various compositions that illustrate underlying token correlation effects.

Related Work

Bias Measurements for T2I Models

Standard evaluation protocols for T2I models have progressively integrated more complex conditions, yet still include limited bias assessments (Li et al. 2024a). Given the implicit nature of bias, uncovering hidden correlations is a crucial step toward meaningful assessment. Counterfactual reasoning has proven valuable for improving the explainability of bias evaluation (Chinchure et al. 2024), while multi-aspect stereotype scoring extends traditional metrics, such as standard deviation across sensitive groups, by analyzing quantitative relationships within the latent space and across denoising steps (Dehdashtian, Sreekumar, and Boddeti 2025). Recent approaches even utilize large language models to identify open-set biases beyond predefined attributes (D'Incà et al. 2024). Moreover, bias patterns may intersect with other persistent challenges in generative models, including hallucinations (Huang et al. 2025) and object omission (Vice et al. 2025), further complicating reliable evaluation.

Model Debias

Debiasing through model retraining. A foundational approach in mitigating bias in T2I models relies heavily on strategies that explicitly define targeted biases through data resampling and tailored loss functions. Mainstream research casts fairness as the distribution alignment and optimization problem to ensure fair loss update and sampling of underrepresented classes (Shen et al. 2024; Khalafi, Ding, and Ribeiro 2024; Zhou et al. 2024). Limited access to unbiased data has driven efforts to train unbiased models from biased datasets (Kim et al. 2024b) using selective finetuning (Zhao et al. 2025) and inpainting (Hirota et al. 2024) techniques.
Debias without Retraining. Increasing training-free debiasing studies focus on identifying key intervention factors, including prompt enhancement, concept editing, and generating guidance. **Prompt enhancement** is to intuitively insert semantically fair expressions to balance the resulting attribute ratio (Bansal et al. 2022; Ding et al. 2021; Chuang et al. 2023). While prompt learning advances these techniques by inserting learnable inclusive tokens with either text or image references (Zhang et al. 2023; Shrestha et al.

2024)), FairQueue (Teo et al. 2024) proposes a prompt queuing mechanism to avoid the unexpected attention distortion in prompt learning. **Concept editing** studies (Gandikota et al. 2024) usually identify the latent-space concept vector, termed h-vector, to manipulate the generated result. Disentangling latent features purely for sensitive attributes (Shi et al. 2025) can be realized without building a classifier every time (Li et al. 2024b). **Generating guidance** techniques comprise latent feature imputation (Jung, Jang, and Wang 2024), attention map selection (Jiang et al. 2024), and minority class sampling (Kim et al. 2024a). Parihar et al. introduce the distribution guidance using an attribute distribution predictor to intervene in the latent space with the targeted distribution (Parihar et al. 2024). These studies only considered simple prompts without compositional cases. Therefore, our work extends bias mitigation into the compositional regime, which prior works did not address.

Compositional Text-to-Image Generation

Compositional generation is gaining traction for tackling object omission and attribute mixing (Bakr et al. 2023). FreeCustom (Ding et al. 2024) designs a multi-reference self-attention to refine the alignment between multiple provided concepts and the generated image. To reduce the inconvenience of using reference images, Hu et al. observed that text embeddings exhibit information coupling and additive properties, enabling the token merging within the same concept and disentangling multiple concepts within a prompt (Hu et al. 2024). Wang et al. introduce a self-consistency guidance to refine attention maps for multi-concept attribute binding (Wang et al. 2025). These algorithms overlook the bias metric for evaluation, which leaves a huge risk in real-world usage. Recent attempts to identify object-to-gender bias have been highlighted in language models (Sabir and Padró 2023) while its presence and implications in T2I generation remain overlooked.

Method

We first explore how each concept (token) embeds with gender biases toward female or male tokens using text embedding similarity comparison. Figure 3 shows the overall CBC framework containing token semantic bias decoupling, BA-Score, and token residual injection. Our idea is to decouple the sensitive attribute-related components and control these components to balance the input embeddings at each forwarding step. The bias indicator is initialized with BA-Score and then depends on the latent embedding distance towards the gender prototypes (cluster centers of males or females). When the bias indicator skews toward a specific group, we inject attribute-related embeddings from other groups and adjust the attention to regulate the bias. Detailed definitions and mathematical formulations are in the supplementary.

Motivation: Bias in Semantic Embeddings

We preliminarily examine the text similarity between each input token and the gender prototypes to reveal the potential bias. The CLIP tokenizer transforms a given prompt with L tokens into text embeddings $C = \{c_1, c_2, ..., c_L\}$. Given

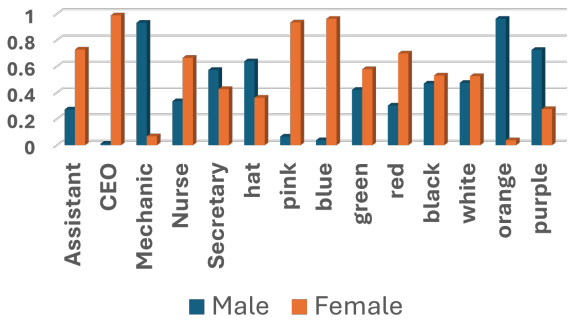

Figure 2: Text correlation to male or female prototype embeddings extracted via CLIP encoder.

sensitive attribute $\mathcal{S} = \{s_k\}_{k=1}^{K}$ with K groups, we represent the two-group gender attributes with $\{s_1, s_2\}$. Prototype embeddings ($\{p_1, p_2\}$) are the average embeddings over 1000 images generated by prompting "a photo of a female" or "a photo of a male". Considering a prompt "an assistant" containing c_2 = "assistant" that combines with "wearing a pink hat". The term "pink" (c_6) is endowed with high correlation to the "woman" prototype embedding, while "hat" (c_7) has over 0.6 correlation to the "man" prototype embedding in Figure 2. The combined effect of the pink hat likely shifts the bias toward more feminine features, further amplifying the inherently female-leaning latent representation of the assistant role. In contrast, the token embedding for 'orange' shows a stronger correlation with 'man', raising whether such a distinct bias direction could mitigate the original bias associated with 'assistant'. The results motivate us to design a decoupling approach for effective control of sensitive components in tokens.

Initialization: Bias Adherence Score (BA-Score)

Typical prompts include a unit template structure, "[main object] + [contexts]", where contexts are composed of context tokens. Given a main object embedding c_m, other identified $M-1$ nouns and adjectives result in context token embeddings $C = \{c_i\}_{i \in I}$ where I indicates a set of selected tokens. I is regarded as a hyperparameter tuned for generation without compromising the contexts. We define a Bias Adherence Score (BA-Score) to indicate the percentage of influence from context tokens and the main object contributing to the bias. We calculate the cosine similarity of context tokens to the prototype embedding p_k:

$$B_{m,k} = \frac{\sum_{i=1}^{M} \mathbf{I}_{i \neq m} \exp((\cos(c_m, c_i) + \cos(p_k, c_i))/\tau)}{\sum_{i=1}^{M} \exp((\cos(c_m, c_i) + \cos(p_k, c_i))/\tau)}, \tag{1}$$

where $\mathbf{I}_{i \neq m}$ is an indicative function, τ controls softmax normalization sharpness. The similarity from context tokens is weighted by $\cos(c_m, c_i)$, their relevance to the main object. The BA-Score measures the maximum deviation of sensitive attribute groups, defined as $B_m = \max_k |\pi - B_{m,k}|$ with $\pi = 0.5$ as the target balanced contribution. A larger deviation indicates stronger imbalance between the main ob-

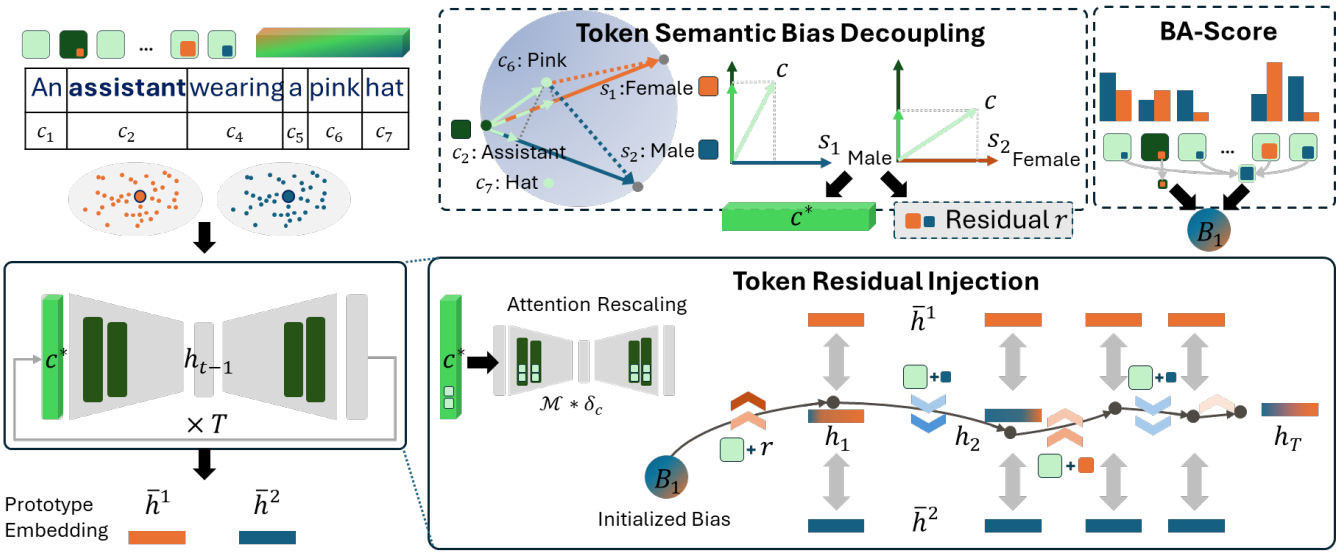

Figure 3: Our proposed CBC framework that decouples the token from sensitive attributes, leverages the BA-Score to measure bias tendency, and adjusts the token with sensitive attribute residuals for unbiased generation.

ject and its contextual attributes, suggesting a higher degree of spurious correlation and a greater risk of semantic distortion after bias denoising.

Context Bias Control (CBC)

We decouple the context token's embedding and remove the sensitive attribute components, yielding a *attribute-decoupled* token embedding and the corresponding attribute residual vector. During diffusion, we continually measure the bias using the latent vectors and inject the residual vector of other attributes into the prompt token embedding to balance the generation for the next time step.

Token Semantic Bias Decoupling We decouple the context tokens from sensitive attributes by orthogonizing text prompt embeddings, and derive attribute-orthogonal token embeddings C^*. The Schmidt orthogonalization projects a token embedding c to a new direction orthogonal to the sensitive attribute embedding s_k:

$$c^* = c - r_k = c - \frac{\langle c, s_k \rangle}{\langle s_k, s_k \rangle} s_k, \tag{2}$$

where $\langle c, s_k \rangle$ denotes inner product between c and s_k. We term r_k as the attribute-specific residual vector that entails attribute information adhering to c. The projection ensures that the new concept embedding c^* eliminates dependent information in s. Therefore, the following T2I generation regards pure semantics from the concept rather than spurious correlation from sensitive attributes.

In contrast to prior works that directly manipulate latent vectors or introduce loss-based guidance to steer the main concept semantics toward a less sensitive space (Gandikota et al. 2024; Shi et al. 2025), we propose decoupling the influences of individual objects and attributes before intervening in the T2I generation process. The model thus forms a pure relation in attention maps from the attribute-orthogonal input embeddings rather than conceptually mixed text embeddings. The decoupled residual vector for the sensitive attributes is preserved for the dynamic denoising steps.

Token Residual Injection We inject the residual embedding of the context tokens to adjust the bias tendency. The injection depends on the measure of current bias tendency via a bias deviation score. We inject the average residual embedding from the other attributes if we find the current generating step deviates from an attribute group s_k. Considering the residual embedding of other sensitive attributes written as $\bar{r} = \frac{1}{K-1} \sum_{j \neq k} r_j$, we can use a weighting factor δ_r to add control over the conditional text embedding $c_t^* = \delta_r * \bar{r} + (1 - \delta_r) * c_{t-1}^*$ for the next generative iteration. For the bias deviation score, we adapt the calculation of the BA-score to the latent space from the original semantic space. That is, measuring the BA-score with latent embedding h_{t-1} and the latent prototype embedding $\bar{h}^k$ for sensitive attribute k. Note that $\bar{h}^k$ is extracted by a contrastive network module trained on 1000 images from each attribute group (Li et al. 2024b). To generalize the $\bar{h}^k$ for all the time steps, we include the randomly sampled images from each time step during the contrastive training. We utilize text embedding-based BA-Score for initialization and latent-space BA-Score for adjusting the residual direction in the intermediate time steps. This residual not only controls the bias but also compensates for the corrupted spurious correlation due to token decoupling.

Token residual injection can introduce numerical instability because the injected vector also affects other compositional attributes and objects through cross-attention. To mitigate this, we apply an attention rescaling mechanism that adjusts the attention weight of the injected tokens. For the i^{th} injected token with the residual embedding, its attention vector $\mathcal{M}_i \in \mathbb{R}^{1 \times L}$ over the L input tokens is scaled

Baseline	Assistant			CEO			Mechanic			Nurse			Secretary			Average		
	FD	VQA	AFS	FD	VQA	AFS	FD	VQA	AFS	FD	VQA	AFS	FD	VQA	AFS	FD	VQA	AFS
SD-1.5	0.26	0.58	0.65	0.91	0.64	0.16	0.96	**0.60**	0.08	0.97	**0.67**	0.06	0.59	**0.73**	0.53	0.69	**0.74**	0.44
SelfDisc	0.32	0.52	0.59	0.17	0.59	0.69	0.09	0.43	0.58	0.98	0.54	0.04	0.38	0.66	0.64	0.62	0.61	0.47
DGDebias	0.35	**0.61**	0.63	0.04	**0.66**	**0.78**	0.86	0.51	0.22	0.89	0.58	0.18	0.52	0.66	0.56	0.68	0.64	0.43
FairQueue	0.01	0.32	0.48	0.03	0.52	0.68	0.05	0.19	0.32	0.02	0.12	0.21	0.01	0.43	0.60	**0.03**	0.33	0.49
CBC	**0.01**	0.60	**0.75**	**0.03**	0.63	0.76	**0.02**	0.59	**0.74**	**0.01**	0.66	**0.79**	**0.01**	0.70	**0.82**	0.04	0.68	**0.80**

Composition	Assistant			CEO			Mechanic			Nurse			Secretary			Average		
	FD	VQA	AFS	FD	VQA	AFS	FD	VQA	AFS	FD	VQA	AFS	FD	VQA	AFS	FD	VQA	AFS
SD-1.5	0.35	0.60	0.62	0.92	0.64	0.14	0.69	**0.62**	0.35	0.63	**0.67**	0.40	0.68	0.65	0.42	0.69	**0.63**	0.41
SelfDisc	0.37	0.54	0.58	0.37	0.55	0.59	0.42	0.56	0.56	0.94	0.60	0.14	0.76	0.63	0.37	0.61	0.58	0.42
DGDebias	0.55	0.43	0.43	0.61	0.55	0.43	0.93	0.55	0.12	0.90	0.51	0.17	0.78	0.62	0.31	0.73	0.62	0.35
FairQueue	0.03	0.52	0.67	0.04	0.58	0.72	0.11	0.44	0.59	0.04	0.40	0.54	0.04	0.38	0.53	0.04	0.45	0.60
CBC	**0.03**	**0.63**	**0.76**	**0.03**	**0.65**	**0.78**	**0.04**	**0.62**	**0.75**	**0.03**	0.60	**0.77**	**0.03**	**0.66**	**0.79**	**0.04**	0.62	**0.75**

Table 1: FD (↓), VQA (↑), AFS (↑) results of SD-1.5 and debiasing approaches. Full results are in the supplementary.

by a weight scalar δ_c, producing a calibrated mask $\mathcal{M}_i^*$ that strengthens attention to regions most relevant to the subject of the prompt: $\mathcal{M}_i^* = w(t)\delta_c\mathcal{M}_i$ where $w(t) = 1 - \frac{t}{T}$ is time-aware strength attenuation function, and t and T denote the current and total diffusion steps, respectively.

Experiments

Dataset: We evaluate the occupation bias on Winobias (Zhao et al. 2018) benchmark that includes 36 professions known to exhibit gender biases. Each prompt template, "a head of a [occupation] [semantic binding]", is used to generate 200 images for each occupation evaluation. Full composition settings are listed in the supplementary.

Evaluation metric: We consider the metrics: 1) Fairness Discrepancy (FD) used to assess the bias regarding a sensitive attribute (Teo et al. 2024; Li et al. 2024b; Orgad, Kawar, and Belinkov 2023), 2) VQAScore (Lin et al. 2024) (VQA) used for text alignment of the image generation, 3) Alignment-aware Fairness Score (AFS) to reflect the intrinsic debiasing results without compromising text-alignment quality. We define the score, AFS $= \frac{2*(1-\text{FD})*\text{VQA}}{(1-\text{FD})+\text{VQA}}$. The highest score of AFS $= 1$ can be achieved only if FD $= 0$ and VQA $= 1$. Higher harmonic mean in this metric suggests a better balancing of the bias-utility tradeoff (Zhao and Gordon 2022). VQAScore has shown superior alignment to human evaluation for image generation faithfulness (Li et al. 2024a). The implementation of FD depends on CLIP ViT-L-14, and the VQAScore is calculated with the outputs from LLaVA-1.5 (Liu et al. 2024).

Compared Models and Configurations: We use pretrained Stable Diffusion v1.5 as our base model to generate images with 512×512 resolution using a single Nvidia A40 GPU. The generation process is implemented with a guidance scale equal to 7.5 in 50 steps. For attention weighting control, $\delta_c = 2$ and $\delta_r = 0.2$ are empirically optimal. The compared debiasing approaches mainstream debiasing concepts include concept editing (SelfDisc (Li et al. 2024b)), generative guidance (DGDebias (Parihar et al. 2024)), and prompt enhancement (FairQueue (Teo et al. 2024)).

Association Bias Mitigation Results

Quantitative comparison: Semantic bindings are constructed using prompts such as "wearing a [object]" and "wearing a [color] [object]", with the verb adapted to the object, *e.g.*, "carrying a briefcase" instead of "wearing a briefcase". Table 1 reports averaged results across professions known to be biased in the SD model, augmented with these compositional bindings. The findings reveal clear limitations of SoTA debiasing methods: SelfDisc and DGDebias struggle under compositional prompts, producing high FD scores and low AFS values (0.41, 0.42, and 0.35 on average). FairQueue demonstrates the typical quality–debiasing tradeoff, achieving low FD (0.04) but considerably reduced VQAScore (0.45) and AFS of only 0.60. These quality drops stem from overly aggressive gender balancing, which conflicts with the model's learned distribution and harms visual fidelity. In contrast, our CBC framework attains 0.04 FD, 0.62 VQA, and 0.75 AFS, demonstrating stronger compositional debiasing with minimal degradation. These outcomes support our hypothesis that compositional bias arises from semantic correlations: context tokens are especially sensitive to uniform bias suppression, leading to severe distortion. By continually injecting calibrated token residuals, CBC refines contextual contributions, preserving semantic dependencies while effectively reducing bias.

Qualitative comparison: The generated images in Figure 4 show the superior text alignment quality from our proposed framework compared to other debiasing approaches. Looking into the images generated from DGDebias, the second image (orange hat assistant) shows a typical failure case that introduces redundant hats and irregular components. The DGDebias model erroneously assigns a dark hat and orange clothing to the woman in the first image, highlighting an issue of attribute mixing. FairQueue exhibits a different limitation where many of its generated images lack recognizable features associated with the intended professions. This issue is particularly evident in prompts for doctors,

Figure 4: Generated images using different debiasing approaches.

where the resulting images closely resemble those of assistants. Other debiasing methods typically rely on the presence of a stethoscope to fulfill the "doctor" prompt. These results suggest that current debiasing approaches struggle with properly performing composition generation. FairQueue often neglects the prompt's intended subject, while DGDebias overcorrects borderline cases, leading to visible image distortion. SelfDisc demonstrates relatively stronger text-to-image alignment, although its FD scores remain suboptimal, as shown in Table 1.

Ablation and Compositional Effect Analysis

Effects of BA-Score and hyper-parameters: Table 2 presents the ablation results for our CBC framework. Discarding BA-Score initialization leads to 7% AFS drops, and using simple semantic similarity for initialization brings 11% AFS degradation. We empirically derive the best hyper-parameters ($\delta_c = 2$ and $\delta_r = 0.2$), which show superior results over other values. The ablation study indicates the effectiveness of initialization with BA-Score, which highly impacts the control of generating steps.

Decoupling Effects: We examine the effects of token decoupling in terms of the underlying bias. Figure 6 illustrates the generated images with different selected tokens being decorrelated. Given the prompt "a photo of an assistant wearing a pink hat" with a high BA-Score due to the "assistant" and "pink hat", we project the selected tokens to be orthogonal to the "woman" token embedding. As a result, decoupling "an assistant" (tokens 4,5) and "a pink hat" (tokens 7,8,9) yields male characteristics, while several other combinations fail to change the bias. Performing orthogonization on "assistant", "pink", or "hat" cannot significantly modify the bias tendency. This phenomenon echoes an observed effect in composition T2I research called token information leakage. The information of the noun might leak to proximate tokens due to cross-attention calculation. Intriguingly, overdecorrelation degrades the semantics. For example, simultaneously decoupling five tokens for "an assistant" and "a pink hat" (4, 5, 7, 8, 9) results in an image without a human but with a hat. A similar result occurs when decorrelating "wearing a pink hat", which misleads the model to ignore the wearing action towards a human. This is because the semantic space of "woman" is always associated with humans. The decoupling removes the semantic relations to the woman, bringing about a side effect of lower confidence in generating humans. Therefore, our empirical finding suggests that keeping the main subject, such as an assistant, and debiasing the associated effects on the compositional attributes and objects can alleviate the semantic decoupling.

	FD	VQA	AFS
CBC (δ_c=2, δ_r=0.2)	0.04	0.62	0.75
without BA-Score Initialization	0.16	0.58	0.68
Semantic-Score Initialization	0.19	0.53	0.64
δ_c=1	0.08	0.61	0.73
δ_c=5	0.11	0.59	0.71
δ_r=0.3	0.04	0.60	0.74
δ_r=0.5	0.12	0.58	0.70

Table 2: Ablation results for BA initialization, δ_c and δ_r.

Object Binding Prompts: We regard binding objects and use the prompt "a photo of a [profession] wearing a [object]", where the objects are provided by prompting GPT-3.5 for common human accessories. The verb is adapted based on the object. For example, carrying a briefcase is preferred over wearing one. In Figure 5 (b), adding "carrying a briefcase" reduces bias in female-leaning roles like secretary. In contrast, "wearing a scarf" sharply increases bias, e.g., the assistant's bias ratio jumps from below 0.6 to over 0.9.

Attribute Binding Prompts: We include colors: blue, red, green, orange, black, white, pink, in the prompt "a photo of a [profession] wearing a [color] hat". Notably, changing hat

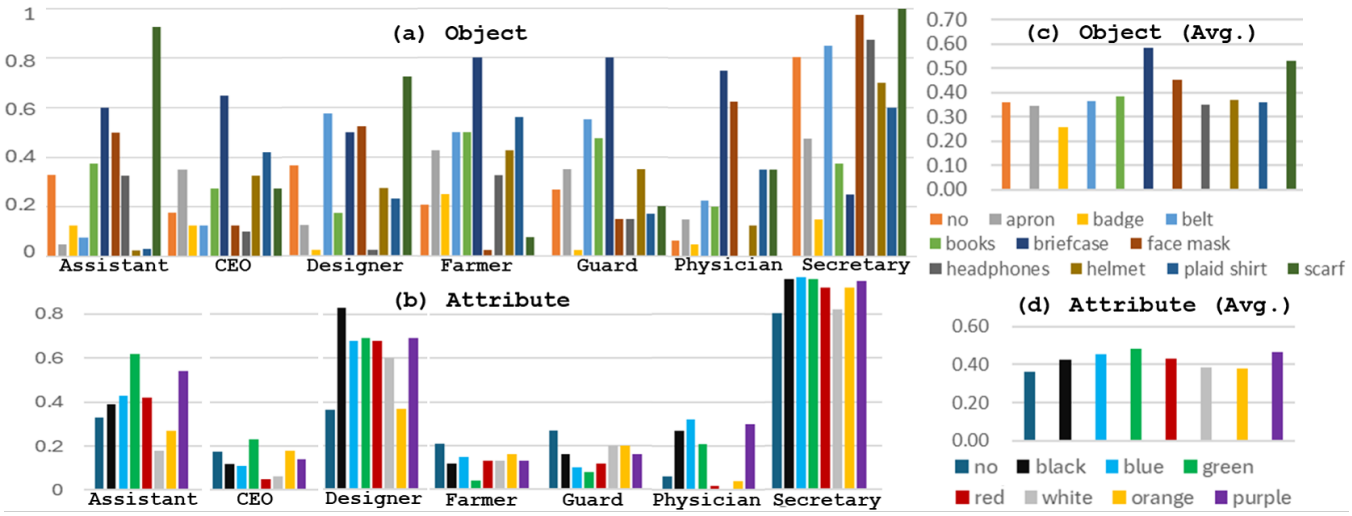

Figure 5: Bias results (FD) in binding conditions with (a) objects and (b) color attributes. The corresponding averaged results for objects and attributes are presented in (c) and (d). Results from other professions are reported in the supplementary.

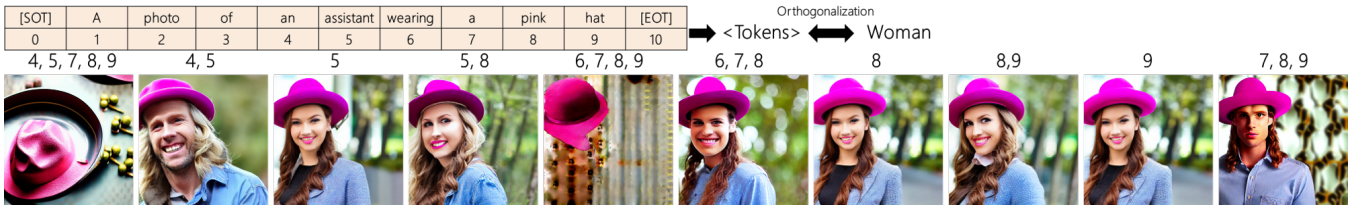

Figure 6: Visualization of the decoupling effects on different tokens. The annotated number of tokens is decoupled to be orthogonal to the "woman" token.

color alone can shift bias across professions. For example, designers show amplified bias with black hats, while physicians exhibit higher FD with blue hats, likely reflecting the common use of surgical caps. These trends align with social stereotypes and dataset distributions. Intriguingly, the green hat triggers substantial bias increases among multiple white-collar professions, e.g., assistants, typically not associated with this color. Conversely, professions like farmers and guards exhibit less bias when paired with green hats, suggesting less distortion from established associations. Color specifications can intensify latent biases, amplifying or attenuating them depending on the subject's occupational context. These results are compared with the pre-inference text embedding relationship in the supplementary.

Associations to Other Spurious Correlation: We observe that the successfully generated images shown in Figure 4 potentially imply other spurious correlations. For instance, the prompt, "An assistant wearing a pink hat" frequently produces outdoor scenes with green fields and trees, likely reflecting the rarity of pink hats in office settings. A similar trend is observed in SelfDisc's output for a doctor wearing a pink hat, highlighting persistent spurious associations across different professions and debiasing strategies. Moreover, the failed cases for "carrying a briefcase" in Figure 4 are accompanied by a suit, regardless of the debiasing results. When evaluating prompt alignment, the token correlation can suggest either a meaningful association or an unintended bias, depending on the sensitivity of the attribute involved. These observations highlight a research direction: stratify less sensitive attributes and leverage their underlying correlations to support realism, thereby alleviating the diffusion model's struggles with underrepresented bindings.

Conclusion

In this work, we introduce the notion of bias adherence in compositional T2I generation, a setting where existing debiasing methods often fail. By measuring bias scores across bound objects and attributes, our context-bias control framework decouples token embeddings and dynamically adjusts residuals to guide generation toward more balanced and less corrupted outputs. Experiments show that our method effectively mitigates bias while preserving strong text-image alignment. Additional analyses reveal how different objects and attributes contribute to bias, opening a path toward broader debiasing scenarios. Our results on complex compositions and decorrelation further suggest that leveraging token relationships can be an effective debiasing strategy. Future work includes building large-scale benchmarks for compositional T2I debiasing and exploring alternative composition mechanisms for more refined attention guidance.

References

Bakr, E. M.; Sun, P.; Shen, X.; Khan, F. F.; Li, L. E.; and Elhoseiny, M. 2023. Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 20041–20053.

Bansal, H.; Yin, D.; Monajatipoor, M.; and Chang, K.-W. 2022. How well can Text-to-Image Generative Models understand Ethical Natural Language Interventions? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 1358–1370.

Chinchure, A.; Shukla, P.; Bhatt, G.; Salij, K.; Hosanagar, K.; Sigal, L.; and Turk, M. 2024. Tibet: Identifying and evaluating biases in text-to-image generative models. In *European Conference on Computer Vision*, 429–446. Springer.

Chuang, C.-Y.; Jampani, V.; Li, Y.; Torralba, A.; and Jegelka, S. 2023. Debiasing vision-language models via biased prompts. *arXiv preprint arXiv:2302.00070*.

Dehdashtian, S.; Sreekumar, G.; and Boddeti, V. N. 2025. OASIS Uncovers: High-Quality T2I Models, Same Old Stereotypes. *arXiv preprint arXiv:2501.00962*.

D'Incà, M.; Peruzzo, E.; Mancini, M.; Xu, D.; Goel, V.; Xu, X.; Wang, Z.; Shi, H.; and Sebe, N. 2024. Openbias: Open-set bias detection in text-to-image generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12225–12235.

Ding, G.; Zhao, C.; Wang, W.; Yang, Z.; Liu, Z.; Chen, H.; and Shen, C. 2024. Freecustom: Tuning-free customized image generation for multi-concept composition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9089–9098.

Ding, M.; Yang, Z.; Hong, W.; Zheng, W.; Zhou, C.; Yin, D.; Lin, J.; Zou, X.; Shao, Z.; Yang, H.; et al. 2021. Cogview: Mastering text-to-image generation via transformers. *Advances in neural information processing systems*, 34: 19822–19835.

Gandikota, R.; Orgad, H.; Belinkov, Y.; Materzyńska, J.; and Bau, D. 2024. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5111–5120.

Grover, A.; Song, J.; Kapoor, A.; Tran, K.; Agarwal, A.; Horvitz, E. J.; and Ermon, S. 2019. Bias correction of learned generative models using likelihood-free importance weighting. *Advances in Neural Information Processing Systems*, 32.

Hirota, Y.; Andrews, J.; Zhao, D.; Papakyriakopoulos, O.; Modas, A.; Nakashima, Y.; and Xiang, A. 2024. Resampled Datasets Are Not Enough: Mitigating Societal Bias Beyond Single Attributes. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 8249–8267. Miami, Florida, USA: Association for Computational Linguistics.

Hu, T.; Li, L.; van de Weijer, J.; Gao, H.; Shahbaz Khan, F.; Yang, J.; Cheng, M.-M.; Wang, K.; and Wang, Y. 2024. Token Merging for Training-Free Semantic Binding in Text-to-Image Synthesis. *Advances in Neural Information Processing Systems*, 37: 137646–137672.

Huang, P.-H.; Li, J.-L.; Chen, C.-P.; Chang, M.-C.; and Chen, W.-C. 2025. Who Brings the Frisbee: Probing Hidden Hallucination Factors in Large Vision-Language Model via Causality Analysis. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 6125–6135.

Jiang, Y.; Lyu, Y.; He, Z.; Peng, B.; and Dong, J. 2024. Mitigating Social Biases in Text-to-Image Diffusion Models via Linguistic-Aligned Attention Guidance. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 3391–3400.

Jung, H.; Jang, T.; and Wang, X. 2024. A Unified Debiasing Approach for Vision-Language Models across Modalities and Tasks. *Advances in Neural Information Processing Systems*, 37: 21034–21058.

Jung, S.; Yu, S.; Chun, S.; and Moon, T. 2024. Do Counterfactually Fair Image Classifiers Satisfy Group Fairness?–A Theoretical and Empirical Study. *Advances in Neural Information Processing Systems*, 37: 56041–56053.

Khalafi, S.; Ding, D.; and Ribeiro, A. 2024. Constrained Diffusion Models via Dual Training. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Kim, E.; Kim, S.; Park, M.; Entezari, R.; and Yoon, S. 2024a. Rethinking Training for De-biasing Text-to-Image Generation: Unlocking the Potential of Stable Diffusion. *arXiv preprint arXiv:2408.12692*.

Kim, Y.; Na, B.; Park, M.; Jang, J.; Kim, D.; Kang, W.; and Moon, I.-C. 2024b. Training unbiased diffusion models from biased dataset. In *The Twelfth International Conference on Learning Representations*.

Li, B.; Lin, Z.; Pathak, D.; Li, J.; Fei, Y.; Wu, K.; Xia, X.; Zhang, P.; Neubig, G.; and Ramanan, D. 2024a. Evaluating and improving compositional text-to-visual generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5290–5301.

Li, H.; Shen, C.; Torr, P.; Tresp, V.; and Gu, J. 2024b. Self-discovering interpretable diffusion latent directions for responsible text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12006–12016.

Lin, Z.; Pathak, D.; Li, B.; Li, J.; Xia, X.; Neubig, G.; Zhang, P.; and Ramanan, D. 2024. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, 366–384. Springer.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26296–26306.

Luccioni, S.; Akiki, C.; Mitchell, M.; and Jernite, Y. 2023. Stable bias: Evaluating societal representations in diffusion models. *Advances in Neural Information Processing Systems*, 36: 56338–56351.

Luo, H.; Deng, Z.; Chen, R.; and Liu, Z. 2024. Faintbench: A holistic and precise benchmark for bias evaluation in text-to-image models. *arXiv preprint arXiv:2405.17814*.

Orgad, H.; Kawar, B.; and Belinkov, Y. 2023. Editing implicit assumptions in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7053–7061.

Parihar, R.; Bhat, A.; Basu, A.; Mallick, S.; Kundu, J. N.; and Babu, R. V. 2024. Balancing act: distribution-guided debiasing in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6668–6678.

Sabir, A.; and Padró, L. 2023. Women Wearing Lipstick: Measuring the Bias Between an Object and Its Related Gender. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Shen, X.; Du, C.; Pang, T.; Lin, M.; Wong, Y.; and Kankanhalli, M. 2024. Finetuning Text-to-Image Diffusion Models for Fairness. In *The Twelfth International Conference on Learning Representations*.

Shi, Y.; Li, C.; Wang, Y.; Zhao, Y.; Pang, A.; Yang, S.; Yu, J.; and Ren, K. 2025. Dissecting and Mitigating Diffusion Bias via Mechanistic Interpretability. *arXiv preprint arXiv:2503.20483*.

Shrestha, R.; Zou, Y.; Chen, Q.; Li, Z.; Xie, Y.; and Deng, S. 2024. Fairrag: Fair human generation via fair retrieval augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11996–12005.

Smith, B. A.; Farinha, M.; Hall, S. M.; Kirk, H. R.; Shtedritski, A.; and Bain, M. 2023. Balancing the Picture: Debiasing Vision-Language Datasets with Synthetic Contrast Sets. In *NeurIPS 2023 Workshop on Synthetic Data Generation with Generative AI*.

Teo, C. T.; Abdollahzadeh, M.; Ma, X.; and Cheung, N.-m. 2024. Fairqueue: Rethinking prompt learning for fair text-to-image generation. *Advances in Neural Information Processing Systems*, 37: 22878–22926.

Udandarao, V.; Prabhu, A.; Ghosh, A.; Sharma, Y.; Torr, P.; Bibi, A.; Albanie, S.; and Bethge, M. 2024. No" zero-shot" without exponential data: Pretraining concept frequency determines multimodal model performance. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Vice, J.; Akhtar, N.; Hartley, R.; and Mian, A. 2025. Exploring Bias in over 100 Text-to-Image Generative Models. *arXiv preprint arXiv:2503.08012*.

Wang, S.; Lin, W.; Huang, H.; Wang, H.; Cai, S.; Han, W.; Jin, T.; Chen, J.; Sun, J.; Zhu, J.; et al. 2025. Towards transformer-based aligned generation with self-coherence guidance. *arXiv preprint arXiv:2503.17675*.

Wu, Y.; Nakashima, Y.; and Garcia, N. 2024. Stable diffusion exposed: Gender bias from prompt to image. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 1648–1659.

Zhang, C.; Chen, X.; Chai, S.; Wu, C. H.; Lagun, D.; Beeler, T.; and De la Torre, F. 2023. Iti-gen: Inclusive text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3969–3980.

Zhao, H.; and Gordon, G. J. 2022. Inherent tradeoffs in learning fair representations. *Journal of Machine Learning Research*, 23(57): 1–26.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In Walker, M.; Ji, H.; and Stent, A., eds., *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 15–20. New Orleans, Louisiana: Association for Computational Linguistics.

Zhao, Z.; Liu, Z.; Cao, Y.; Gong, S.; and Patras, I. 2025. AIM-Fair: Advancing Algorithmic Fairness via Selectively Fine-Tuning Biased Models with Contextual Synthetic Data. *arXiv preprint arXiv:2503.05665*.

Zhou, J.; Gao, J.; Zhao, X.; Yao, X.; and Wei, X. 2024. Association of Objects May Engender Stereotypes: Mitigating Association-Engendered Stereotypes in Text-to-Image Generation. *Advances in Neural Information Processing Systems*, 37: 51754–51786.