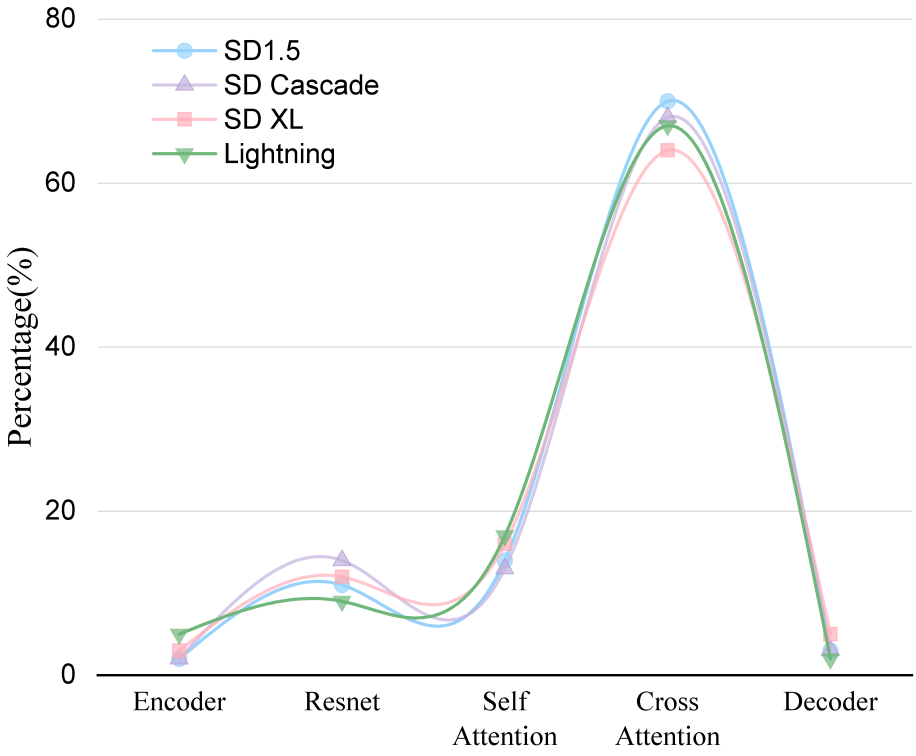
Response to reviewer 2bX6's questions:

- **Q3: (about the additional analysis regarding)**

  We sincerely appreciate your constructive suggestions. We conducted a statistical analysis of stereotype neuron distribution across five key layers: ResNet, Cross-Attention, Self-Attention, Encoder, and Decoder. The results are as follows:



Fiugre 1: Distribution ratio of stereotype neurons in different network layers

Analysis and Insights: Figure 1 compares the distribution of stereotype neurons across different diffusion models. We found that in all four diffusion models, over 65% of the stereotype neurons were located in the cross-attention layers. Additionally, approximately 10%–18% of the stereotype neurons were located in the ResNet and self-attention layers.

In diffusion models, the cross-attention layer enables the model to process its features while attending to and integrating features from conditional inputs, such as text, images, or other modalities. This mechanism allows the model to incorporate conditional information into the denoising process, thereby enhancing the quality and relevance of the generated outputs.

Based on these findings, the cross-attention layer, when integrating self-features and conditional input features, inadvertently activates sensitive attribute features associated with the conditional input due to fixed stereotype associations learned during training. This erroneous activation leads to the generation of stereotype-biased images. Our analysis further reveals that suppressing the activation of stereotype neurons within the cross-attention layer can significantly reduce their influence on model outputs. This suppression mechanism may underlie the effectiveness of SNS in mitigating stereotypes.

Response to reviewer tdR3's questions:

- **Q2: (neuronal characterization analysis)**

  To explore this, we conducted a statistical analysis of stereotype neurons identified across multiple sensitive attribute dimensions. We statistic their distribution across different layers of the diffusion model network. The results are shown in Figure 1.

  Analysis and Insights: Figure 1 compares the distribution of stereotype neurons across different diffusion models. We found that in all four diffusion models, over 65% of the stereotype neurons were located in the cross-attention layers. Additionally, approximately 10%–18% of the stereotype neurons were located in the ResNet and self-attention layers.

  In diffusion models, the cross-attention layer enables the model to process its features while attending to and integrating features from conditional inputs, such as text, images, or other modalities. This mechanism allows the model to incorporate conditional information into the denoising process, thereby enhancing the quality and relevance of the generated outputs.

  Based on these findings, the cross-attention layer, when integrating self-features and conditional input features, inadvertently activates sensitive attribute features associated with the conditional input due to fixed stereotype associations learned during training. This erroneous activation leads to the generation of stereotype-biased images. Our analysis further reveals that suppressing the activation of stereotype neurons within the cross-attention layer can significantly reduce their influence on model outputs. This suppression mechanism may underlie the effectiveness of SNS in mitigating stereotypes.

Response to reviewer KNZV's questions:

- **Q4: (about the prompt)**

  For each stereotype type across each sensitive attribute dimension, we randomly selected **10** prompts. For every prompt, we generated **100** images, resulting in a total of **1,000** images used in the evaluation for that stereotype.

- **Q5: (objects and attributes detection)**

  In order to avoid the uncertainty caused by the accuracy of the detector, we use manual detection.

- **Q6: (the sota T2I models)**

  We added the mitigation effect of our solution based on the transformer architecture. We tested Stable diffusion 3.5 and the results are as follows:

Model	Single Object ↓					Multi Object ↓				FID ↓
	Gender	Race	Region	Age	G×R	Gender	Race	Age	G×R	
SD-3.5	.76±.13	.67±.17	.66±.10	.84±.11	.80±.10	.78±.16	.82±.11	.87±.06	.74±.14	.65±.06
SNS(Ours)	.29±.11	.26±.16	.28±.13	.30±.14	.32±.11	.27±.09	.28±.16	.30±.11	.33±.09	.69±.04

  We evaluate the ViT-based Stable Diffusion 3.5 model, where the SD-3.5 row represents the original model's Stereotype Distribution Total Variation (SDTV) distance score. The results indicate that the original model exhibits severe stereotypes across multiple sensitive attributes, including gender, race, and age. After applying our proposed mitigation approach, a substantial reduction in these biases is observed.

- **Q7: (about expression)**

We sincerely appreciate your constructive suggestions. We will modify the original text as follows:

> **Line 16:** which may cause potential negative impacts or even severe consequences in T2I applications.
>
> **Line 92:** By utilizing a backpropagation algorithm, DDTD traces the influence of each neuron on stereotype encoding, starting from the output layer and moving forward, progressively decomposing each neuron's contribution to the stereotype content in the model's output.

Response to reviewer PDkP's questions:

- **Q2: (about obtaining sensitive attributes)**

  Considering the potential inaccuracy of the classifier, we adopt a statistical approach for manual evaluation. Additionally, we supplement FID experiments, as shown in the following table:

Model	CLIP-T2I ↑		CLIP-I2I ↑	FID ↓	
	Original	Ours		Original	Ours
SD-1.5	.40±.05	.38±.05	.79±.09	15.5±1.30	18.1±1.00
SD XL	.34±.04	.33±.06	.80±.11	16.1±1.00	20.3±0.80
Lightning	.33±.05	.33±.05	.85±.07	22.4±1.20	24.6±1.50
Turbo	.32±.03	.30±.04	.78±.14	20.6±1.90	23.3±1.60
Cascade	.42±.04	.41±.02	.90±.06	22.9±2.00	25.3±1.60

These results show that although our stereotype neuron suppression leads to a slight increase in FID values—indicating a minor decline in image quality—the overall degradation is minimal. We believe this trade-off is acceptable given the benefits in mitigating harmful stereotypes.