Breaking the Illusion: Consensus-Based Generative Mitigation of Adversarial Illusions in Multi-Modal Embeddings

Supplementary Material

A.1. Cosine Similarity Analysis of Mitigation Methods

Table 3 reports the cosine similarity between the embedding vectors of original/perturbed images and the corresponding outputs produced by each mitigation method. A high cosine similarity for the original input indicates low distortion or side effects, whereas a low cosine similarity for the perturbed input suggests that the mitigation method effectively disrupts the illusion attack by removing or weakening the adversarial signal.

Among all evaluated approaches, VAE + Sampling exhibits the most favorable trade-off. It achieves the highest cosine similarity on original images (0.93), indicating that it preserves the clean content more faithfully than the other methods. At the same time, its similarity on perturbed images drops to 0.15, meaning that the filtered output is largely unaligned with the adversarially manipulated representation and thus effectively disrupts the attacker's intended semantics.

Other mitigation methods achieve comparable cosine similarity values on the perturbed inputs, indicating that they can also diminish the adversarial manipulation to some extent. However, these methods generally introduce stronger distortions to clean images. For example, Gaussian Blur yields the lowest cosine similarity on perturbed images (0.13), suggesting effective removal of the illusion signal, but it also reduces the similarity for clean images to 0.77—substantially lower than the preservation achieved by VAE + Sampling. This highlights that while several methods can weaken the illusion attack, VAE + Sampling offers the best balance between attack mitigation and fidelity to the original image.

Table 3. Cosine similarity between original/perturbed images and their augmented versions. Here, x denotes the original image and $\tilde{x}$ its perturbed counterpart. Method$(\cdot)$ represents the transformation produced by a given mitigation technique. CS$(\cdot,\cdot)$ computes the cosine similarity between the embedding vectors of two inputs.

Method	**CS(x, Method(x))**	**CS($\tilde{x}$, Method($\tilde{x}$))**
VAE + Sampling (Ours)	0.93 ± 0.02	0.15 ± 0.05
DM + Sampling (Ours)	0.92 ± 0.04	0.15 ± 0.05
JPEG	0.85 ± 0.04	0.15 ± 0.05
Gaussian Blur	0.77 ± 0.05	0.13 ± 0.04
Random Affine	0.84 ± 0.05	0.15 ± 0.04
Color Jitter	0.88 ± 0.04	0.28 ± 0.07
Horizontal Flip	0.90 ± 0.03	0.26 ± 0.10
Random Perspective	0.80 ± 0.06	0.16 ± 0.04

Table 4. Computational cost of VAE, Diffusion and Original Models in terms of parameters and FLOPs.

Method	**Parameters (M)**	**TFLOPs**
VAE	83.65	3.57
DM	859.52	33.9
Original	2157.52	155.6

A.2. Computational Cost Analysis of Mitigation Methods

Table 4 presents the parameters and FLOPs of the original adversarial illusion model and the VAE and diffusion mitigation modules. The originsl model contains 2,157.52M parameters and requires 155.6 TFLOPs for one image. The VAE mitigation is highly lightweight, adding only 83.65M parameters and 3.57 TFLOPs, which corresponds to merely a 3.9% increase in parameters and a 2.3% increase in computation over the original model. The diffusion model is heavier with a 39.8% parameter and 21.8% FLOP overhead compared to original model. Both mitigation modules introduce only a relatively small extra computational cost.