Understanding Adversarial Weakness in Vision-Language Models

Multimodal learning has become a key aspect of modern artificial intelligence, with models such as CLIP combining visual and textual information to solve complex tasks. However, these models remain vulnerable to adversarial manipulation, and a fundamental question remains unresolved, which modality, image or text, is more fragile under attack. This work investigates this question using the Facebook Hateful Memes dataset, a benchmark that pairs images with captions for binary hate speech detection. We generate adversarial examples with FGSM and PGD on both modalities. By freezing one encoder at a time and retraining the other with adversarial inputs, we are able to isolate modality specific weaknesses and evaluate them with both task performance metrics and embedding level analyses.

Our results demonstrate that the vision model of CLIP is consistently the weaker side. Perturbations to images cause greater reductions in F1 score and produce larger shifts in the embedding space than perturbations to text. As shown in Table 1, PGD attacks reduce F1 from 75.0 to 62.3 when images are perturbed, compared with smaller declines for text perturbations. Adversarial training on a single modality partially improves robustness, but complete recovery is achieved only when both image and text encoders are updated together.

Table 1: CLIP F1 scores under FGSM and PGD. C=Clean, P=Perturbed. For rows with *C/P*, the F1 shows results for C then P.

Method	Vision	Text	F1 (FGSM)	F1 (PGD)
Baseline	C	C	75.00	75.00
Attack Vision	P	C	64.00	62.34
Attack Language	C	P	63.55	62.29
Attack Vision and Language	P	P	62.60	61.12
Retrain Vision Freeze Lang. model	P	C/P	70.10 / 62.12	71.65 / 71.76
Retrain Lang Freeze Vision model	C/P	P	68.94 / 64.13	72.76 / 70.80
Retrain Vision and Lang model	P	P	69.89	71.39

This study makes two main contributions. First, it introduces a clear and reusable protocol for measuring cross-modality robustness in multimodal models, combining task-level performance with embedding-space analysis to reveal how adversarial perturbations distort cross-modality representations. Second, the findings consistently show that images are the primary point of failure, adversarial perturbations to the vision modality cause greater performance degradation than perturbations to text, highlighting the need for defenses that place greater emphasis on the visual channel and maintain cross-modal consistency. Future research will extend this framework to stronger attacks, larger benchmarks, and more diverse architectures to validate and generalize the findings, ultimately aiming to build multimodal systems that remain reliable under adversarial stress.