

Supplementary Materials: Transferable Diffusion-based Unrestricted Adversarial Attack on Pre-trained Vision-Language Models

Anonymous Authors

1 DENOISING DIFFUSION IMPLICIT MODELS

Denoising Diffusion Probabilistic Models (DDPMs) [1] are the state-of-the-art generative models employed for image generation, where an iterative denoising of an initial Gaussian noise is conducted. DDPMs contain both an inversion process and a denoising process. In the inversion process, Gaussian noise is gradually added to an image x_0 for a total timestep T to get latents x_1, x_2, \dots, x_T in the following equation:

$$q(x_t|x_{t-1}) = \mathcal{N}(\sqrt{1 - \beta_t}x_{t-1}, \beta_t\mathbf{I}), \quad (1)$$

where $t \in [1, T]$ and $\beta \in (0, 1)$. When T is large enough, the last latent x_T will be approximately a pure Gaussian noise. According to the additivity of Gaussian distributions, x_t can be directly calculated using x_0 :

$$\begin{aligned} q(x_t|x_0) &= \mathcal{N}(\sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t)\mathbf{I}), \\ x_t &= \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \end{aligned} \quad (2)$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{s=0}^t \alpha_s$.

In the denoising process, a new sample is drawn from x_T by iteratively sampling from $q(x_{t-1}|x_t)$. Since $q(x_{t-1}|x_t)$ is hard to calculate for the distribution of x_0 is unknown, a neural network p_θ with parameters denoted as θ is trained to predict the mean and covariance of $q(x_{t-1}|x_t)$:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(\mu_\theta(x_t, t), \Sigma_\theta(x_t, t)). \quad (3)$$

Ho *et al.* [1] proposed to predict the noise $\epsilon_\theta(x_t, t)$ to simplify the objective:

$$\min_{\theta} \mathcal{L}(\theta) = \mathbb{E}_{x_0, \epsilon, t} \|\epsilon - \epsilon_\theta(x_t, t)\|_2^2. \quad (4)$$

After p_θ is fully trained, we can sample x_{t-1} as follows:

$$x_{t-1} = \mu_\theta(x_t, t) + \sigma_t z, z \sim \mathcal{N}(0, \mathbf{I}). \quad (5)$$

Song *et al.* [3] proposed Denoising Diffusion Implicit Models (DDIMs) to accelerate sampling in DDPMs. In DDIMs, the denoising process is no more a Markovian process:

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \left(\frac{x_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(x_t)}{\sqrt{\bar{\alpha}_t}} \right) + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2} \cdot \epsilon_\theta(x_t) + \sigma_t z. \quad (6)$$

where $z \sim \mathcal{N}(0, \mathbf{I})$. In DDIMs, σ_t is set to 0 to make the denoising process determined. The reversion of the denoising process is called DDIM inversion.

2 EXPERIMENTAL RESULTS ON MSCOCO DATASET

Besides Flickr30k Dataset, we also conduct experiments on image-text and text-image retrieval tasks on MSCOCO dataset. According to the conclusions in the ablations studies (Section 4.2), increasing either relative weight control factor μ or guidance scale ω could effectively improve the attack performance under both white-box and black-box settings. With relative weight control factor $\mu = 0.5$ and guidance scale $\omega = 2.5$, our method MDA have surpassed all other compared attack methods except SGA under white-box settings, as shown in Section 4.2. Therefore, different from the experiments conducted on Flickr30k dataset, we here choose a larger guidance scale $\omega = 5$ to test whether our method MDA could surpass state-of-the-art attack method SGA [2] under white-box settings. The experimental results are listed in Table 1. From Table 1 we could see that our method MDA achieves the best results both under white-box and black-box settings, successfully surpassing the attack performance of SGA. (in image-text retrieval task, MDA outperforms SGA by over 5% under white-box settings and over 30% under black-box settings.) This is because a larger guidance scale indicates a stronger guidance by adversarial text, which introduces more distortions in the denoising process, thus enhancing attack performance.

REFERENCES

- [1] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.
- [2] Dong Lu, Zhiqiang Wang, Teng Wang, Weili Guan, Hongchang Gao, and Feng Zheng. 2023. Set-level guidance attack: Boosting adversarial transferability of vision-language pre-training models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 102–111.
- [3] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.

Table 1: ASR on image-text and text-image retrieval tasks on MSCOCO dataset. ALBEF is adopted as the surrogate model. * indicates the performance under white-box attack. The best results are highlighted in bold.

Task	Attack	ALBEF			TCL			CLIP _{ViT}			CLIP _{CNN}		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Image-Text	PGD	76.70*	67.49*	62.47*	12.46	5.00	3.14	13.96	7.33	5.21	17.45	9.08	6.45
	BERT-Attack	24.39*	10.67*	6.75*	24.34	9.92	6.25	44.94	27.97	22.55	47.73	29.56	23.10
	Sep-Attack	82.60*	73.20*	67.58*	32.83	15.52	10.10	44.03	27.60	21.84	46.96	29.83	23.15
	Co-Attack	79.87*	68.62*	62.88*	32.62	15.36	9.67	44.89	28.33	21.89	47.30	29.89	23.29
	SGA	96.75*	92.83*	90.37*	58.56	39.00	30.68	57.06	39.38	31.55	58.95	42.49	34.84
	MDA	99.51*	98.64*	97.51*	93.35	80.61	73.55	93.11	80.62	75.40	94.01	81.50	76.15
Text-Image	PGD	86.30*	78.49*	73.94*	17.77	8.36	5.32	23.10	12.74	9.43	23.54	13.26	9.61
	BERT-Attack	36.13*	23.71*	18.94*	33.39	20.21	15.56	52.28	38.06	32.04	54.75	41.39	35.11
	Sep-Attack	89.88*	82.60*	78.80*	42.92	27.04	20.65	54.46	40.12	33.46	55.88	41.30	35.18
	Co-Attack	87.83*	80.16*	75.98*	43.09	27.32	21.35	54.75	40.00	33.81	55.64	41.48	35.28
	SGA	96.95*	93.44*	91.00*	65.38	47.61	38.96	65.25	50.42	43.47	66.52	52.44	45.05
	MDA	99.63*	98.57*	96.91*	95.67	84.66	76.52	94.80	83.98	77.22	94.91	84.05	78.84