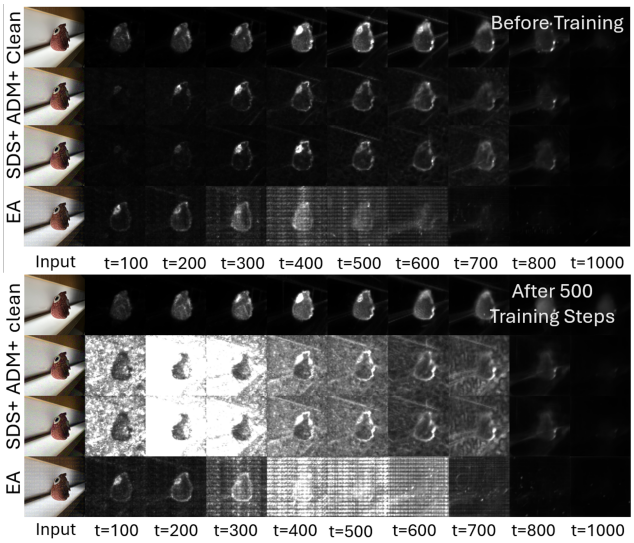

Figure 9. Comparison between SSMs of clean and AE samples at different noise timesteps (a) before TI training and (b) after 500 steps of TI training.

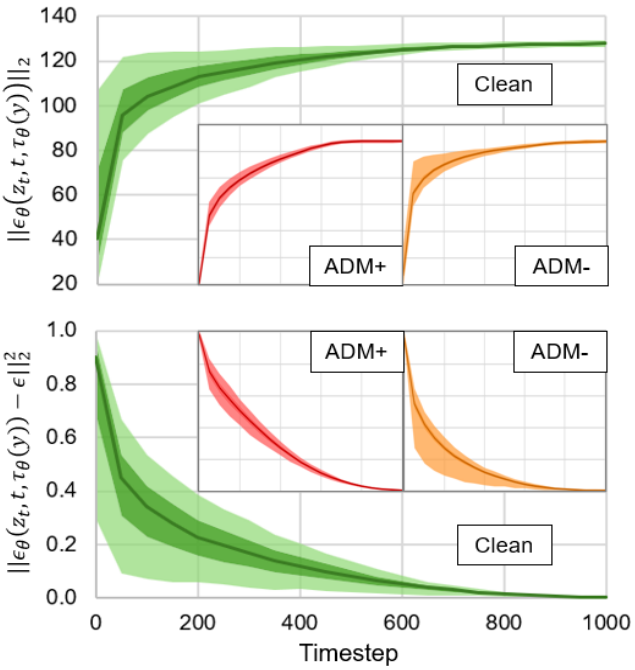

Figure 10. Noise prediction magnitude (top) and loss (bottom) distributions for Stable Diffusion 1.5 on 1000 LAION Aesthetic images at each of 21 interspersed timesteps.

A. Additional Semantic Sensitivity Maps

We include an additional study of SSMs for the FishDoll concept from NovelConcepts10 in Fig. 9. Trends here generally reflect those noted in Sec. 4.1.

B. Noise-Prediction Distribution

We assume that the distribution of LDM latents z_0 follows some Gaussian $\mathcal{N}(\mu_z, \Sigma_z)$.[2] The ground truth noise ϵ is sampled from a standard Gaussian $\mathcal{N}(\mathbf{0}, I)$. For some timestep t, we calculate z_t following Eq. 1. The distribution of z_t can be described as the Gaussian achieved by scaling and combining the z_0 and ϵ distributions.

$$z_t \sim \mathcal{N}(\sqrt{\overline{\alpha}_t}\mu_z, \Sigma_{zt}), \tag{8}$$

where $\Sigma_{zt} = \overline{\alpha}_t\Sigma_z + (1-\overline{\alpha}_t)I$. We assume that Σ_{zt} is invertible. The joint distribution of ϵ and z_t is given by

$$p(\epsilon, z_t|t) = \mathcal{N}\left(\begin{bmatrix} 0 \\ \sqrt{\overline{\alpha}_t}\mu_z \end{bmatrix}, \begin{bmatrix} I & \sqrt{1-\overline{\alpha}_t}I \\ \sqrt{1-\overline{\alpha}_t}I & \Sigma_{zt} \end{bmatrix}\right), \tag{9}$$

[2]We note a distinction between distributions $p(z_0|x)$ and $p(z_0)$. In LDMs where the encoder is from a VAE [31, 58], $p(z_0|x)$ is Gaussian by definition, as z_0 is randomly sampled from a multivariate Gaussian that describes the encoding space (i.e., $\mathcal{E}(x) = [\mu_z, \Sigma_z]$). However, across many samples x, $p(z_0)$ is not necessarily Gaussian. Our analysis in this section is exactly correct in the single-image case ($p(z_0|x)$), but its correctness in the multi-image case ($p(z_0)$) hinges on the correctness of the Gaussian assumption for $p(z_0)$. In a brief experiment (not shown), we sample z_0 for 1000 LAION Aesthetic images and find that the distributions of z_0 dimensions are generally unimodal and akin to bell-curves, but they are decidedly not Gaussian. Even so, the empirical results for the multi-image case in Fig 10 conform to the behavior expected by our derivation in this section, suggesting that the Gaussian assumption for $p(z_0)$ is permissible.

where we apply the linearity property of covariance ($cov(A, A+B) = cov(A, A) + cov(A, B)$) and the independence between z_0 and ϵ to derive the off-diagonal covariance matrices. We seek to predict the distribution of noise when given the same inputs as a noise-prediction DM. We use the definition of a conditional distribution, $p(x|y) = p(x, y)/p(y)$ to derive the distribution of ϵ conditioned upon inputs z_t and t. The conditional distribution can be derived in closed form from the joint distribution [75].

$$\begin{aligned} p(\epsilon|z_t, t) = \mathcal{N}(&\sqrt{1-\overline{\alpha}_t}(\Sigma_{zt})^{-1}(z_t - \sqrt{\overline{\alpha}_t}\mu_x), \\ &I - (1-\overline{\alpha}_t)(\Sigma_{zt})^{-1}), \end{aligned} \tag{10}$$

Though initially complicated, this conditional distribution simplifies beautifully in the limits as $t \to 0$ and $t \to T$. As $t \to 0$, $\overline{\alpha}_t \to 1$, so the conditional distribution approaches $\mathcal{N}(\mathbf{0}, I)$. This is sensible because $z_t \to z_0$, which contains no information about the ground-truth noise, so the noise distribution $\epsilon = \mathcal{N}(\mathbf{0}, I)$ is independent of z_0. As $t \to T$, $\overline{\alpha}_t \to 0$, so the distribution approaches $\mathcal{N}(z_T, \mathbf{0})$. This is sensible as $z_t \to z_T = \epsilon$.

This derivation can be extended to the distribution of the noise-prediction error (Eq. 2). As both components are Gaussian, the noise-prediction error $p(\epsilon|z_t, t) - \epsilon$ is also Gaussian and the covariance of Eq. 10 is shifted by I. Notably, the noise prediction error has maximum variance at

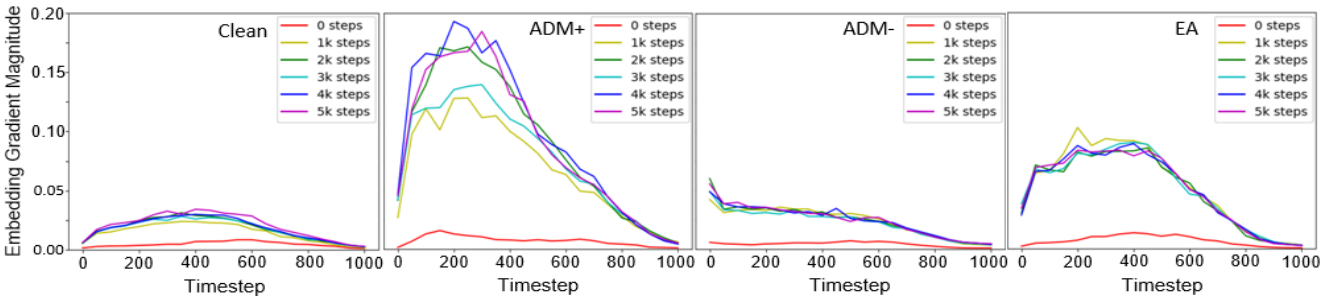

Figure 11. Plots of text embedding loss gradient magnitude as a function of timestep measured throughout training. Values at each timestep are averaged across concepts in the NovelConcepts10 dataset.

$t = 0$ which monotonically decreases as t increases.

$$p(\epsilon|z_t, t) - \mathcal{N}(\mathbf{0}, I) = \mathcal{N}(\sqrt{1-\overline{\alpha}_t}(\Sigma_{zt})^{-1}(z_t - \sqrt{\overline{\alpha}_t}\mu_x), \quad 2I - (1-\overline{\alpha}_t)(\Sigma_{zt})^{-1}), \tag{11}$$

Fig. 10 is an expanded version of Fig. 4 that includes plots for the magnitudes of the predicted noise. As discussed in Sec. 4.2, the model predicts noise near 0 at $t = 0$ and perfectly predicts noise at $t = T$. The magnitude of the predicted noise grows from near 0 at $t = 0$ to a maximum of 128 at $t = T$ (which aligns with $E[||\epsilon||_2] = \sqrt{tr(I)} = 128$ when ϵ is a 4x64x64 vector as in Stable Diffusion). The median loss monotonically decreases from a maximum at $t = 0$ to 0 at $t = T$. Additionally, the width of the loss distribution at each timestep indicates that the variance of the noise-prediction error is indeed a maximum at $t = 0$ and diminishes to 0 at $t = T$.

The noise-prediction DM always seeks to minimize loss by maximizing the conditional probability of the noise (Eq. 11). At $t = 0$, it will return the mean of $\epsilon \sim \mathcal{N}(\mathbf{0}, I)$. At high timesteps, it will increasingly predict its own input. At middle timesteps, it will actually learn to distinguish noise from image. Therefore, all useful learning will occur at lower-middle timesteps.

Lastly, we note that this behavior is reversed (but still present) for input-prediction models (i.e., $z_{0,\theta}(\cdot)$). An input-prediction model will return its input at $t = 0$ and the average image (i.e., a plain gray image) at $t = T$.

C. Textual Inversion Gradients

In Fig. 11 we display the l^2 magnitude of the loss gradient on the trainable text embedding as a function of timestep t throughout TI training. For each concept in the NovelConcepts10 dataset, we extract checkpoints every 1000 steps during TI training and calculate the expected gradient magnitude for each timestep then average the magnitudes across concepts. Aligning with our findings from SSMs (Sec. 4.1) and timestep learning bias (Sec. 4.2), all gradient magnitudes are biased towards lower-middle timesteps, with near-0 gradient magnitudes at high timesteps for all datasets. The ADM- poison, which seeks to minimize DM loss, mitigates the low-timestep bias. The EA poison, which optimizes poisons on the LDM encoder and avoids the DM entirely during optimization, still permits a bias towards lower timesteps during TI training. This is likely because EA focuses on perturbing z_0, thereby adversarially affecting high signal-to-noise z_t latents (i.e., those at lower-middle timesteps).

We note that the dropoff in gradient at very low timesteps for Clean, ADM+, and EA samples is due to the tendency of the noise-prediction model to predict $\mathbf{0}$ at $t = 0$. Although the loss at this point is maximal (as shown in Fig. 10 (bottom)), the conditional distribution of ϵ (from Eq. 10) at $t = 0$ is independent of z_t, as noted in Sec. B. Therefore, there is no loss gradient with respect to the inputs at $t = 0$. In summary, learning is generally minimal at high timesteps and at $t = 0$, regardless of whether the model is learning a true concept or an adversarial signal.

D. JPEG Compression Analysis

In Fig. 12, we extend the analysis performed for Fig. 7 in Sec. 4.4 to all poisons. Across all poisons, we find that JPEG compression has the same effect: the poison perturbation distribution is converted from a bimodal distribution with modes that skew towards the perturbation limits ($\pm\kappa$) to a bell-curve distribution centered at or near 0.

To verify claim (2) from Sec 4.4, we also investigate the impact of JPEG compression on the LDM latent space. Fig. 13 displays the radially averaged power spectrum density curves for the mean latent encodings of Clean, ADM+, ADM-, and EA images with and without JPEG compression, averaged across all 50 images in the NovelConcepts10 dataset. Before JPEG compression, the power curve for ADM+ latents is consistently higher than that of clean latents at high frequencies, whereas the powers for ADM- and EA latents are consistently lower. However, after JPEG compression (in pixel space), the power spectra of all images are centralized, lying closer to the power spectra of

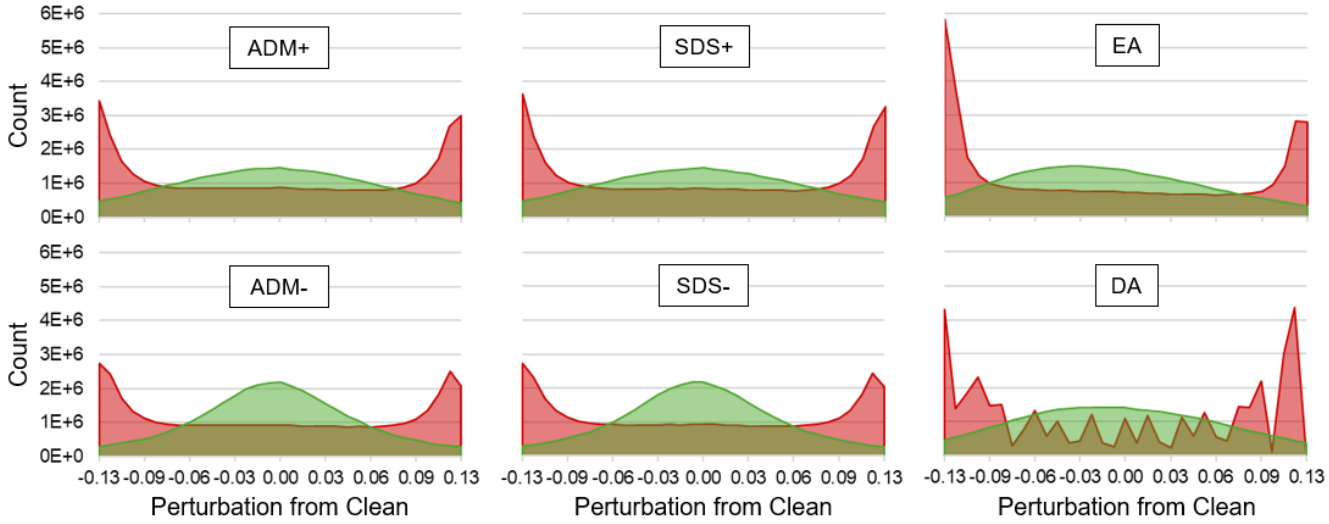

Figure 12. Histograms of pixel-space perturbations (relative to clean images) for poisoned NovelConcept10 images without (red) and with (green) JPEG compression. Input images are shifted and scaled to $[-1, 1]$

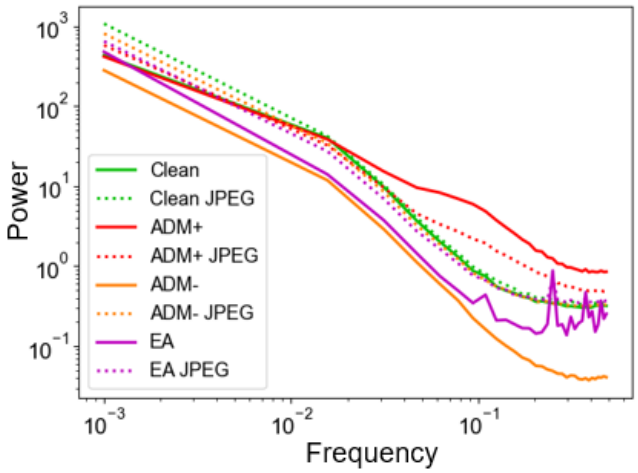

Figure 13. Radially averaged power spectrum density curves for mean latents with and without JPEG compression. Curves are averaged across all images in NovelConcepts10.

clean images. This result suggests that JPEG compression tends to "standardize" poisoned images, forcing their latents to conform to the power spectra expected by LDMs.

E. NovelConcepts10 Dataset

The NovelConcepts10 dataset consists of five images for each of 10 distinct concepts, for a total of 50 images. Each object is located roughly at the center of each image, as is common in most LAION [62] and ImageNet [10] images. Angle (or pose) and background are different for each image of a single object. The images are resized to 512x512 and stored in PNG format to preserve quality. When choosing concepts to capture, we included both non-unique objects (e.g., CokeCan) and unique objects (e.g., FishDoll). Non-unique objects are easily recognizable and are likely included in large training datasets whereas as unique objects are rare and unlikely to be captured in most training datasets. Fig. 14 shows an example of each concept in the NovelConcepts10 dataset.

Figure 14. One example image for each of the 10 concepts contained within the NovelConcepts10 dataset.

F. Poison Descriptions

We analyze multiple adversarial methods. In particular, we use AdvDM (ADM+) [39], SDS (SDS+) [78], as well as EncoderAttack (EA) and DiffusionAttack (DA) [61]. We implement ADM+ and EA via attack modes 0 and 1 from the MIST library [38]. We also analyze the gradient descent versions of ADM+ and SDS+, ADM- and SDS- respectively, since gradient descent poisons are also effective [78]. Table 3 describes the attack direction, targeting type,

Table 2. NovelConcepts10 DINOv2 Similarity for various TI hyperparameter settings. Bold values indicate best settings per poison type.

	LR Sched.	Constant			Linear			Cosine		
	LR	$5e^{-5}$	$5e^{-4}$	$5e^{-3}$	$5e^{-5}$	$5e^{-4}$	$5e^{-3}$	$5e^{-5}$	$5e^{-4}$	$5e^{-3}$
Clean	1k	0.14	0.36	0.43	0.11	0.31	0.41	0.11	0.36	0.41
	5k	0.33	0.44	0.47	0.28	0.42	0.47	0.26	0.40	0.48
	10k	0.37	0.44	0.49	0.33	0.44	0.50	0.36	0.43	**0.51**
ADM+	1k	0.11	0.09	0.08	0.11	0.10	0.10	0.11	**0.12**	0.11
	5k	0.10	0.10	0.08	0.10	0.09	0.09	0.10	0.09	0.08
	10k	0.10	0.10	0.09	0.11	0.11	0.09	0.11	0.11	0.09
ADM-	1k	0.11	0.29	0.30	0.11	0.24	0.29	0.11	0.25	0.28
	5k	0.22	0.29	0.33	0.21	0.30	0.30	0.20	0.29	0.34
	10k	0.22	0.26	0.37	0.24	0.29	0.35	0.24	0.28	**0.40**
EA	1k	**0.11**	0.09	0.07	**0.11**	**0.11**	0.07	**0.11**	0.10	0.06
	5k	**0.11**	0.08	0.08	0.10	0.09	0.07	0.10	0.07	0.07
	10k	0.09	0.09	0.07	0.10	0.08	0.09	0.10	0.09	0.09

Table 3. Characteristics of examined poisons.

Poison	Direction	Targeted	Objective
ADM+ [39]	↑	No	DM Loss
ADM-	↓	No	DM Loss
SDS+ [78]	↑	No	DM Loss
SDS- [78]	↓	No	DM Loss
EA [61]	↓	Yes	Encoding Dist
DA [61]	↑	No	DM Sampling

and objective of each poison. We restrict each poison perturbation to a l^∞ bound of size $\kappa = 16/256$. For ADM+, ADM-, SDS+, SDS-, and EA, we apply 100 projected gradient steps of strength $\eta = 1/256$. For DA, we follow the default sampling and optimization settings. We use the high-contrast MIST image from the MIST library as the target image for EA, which was cited by [38] as a better target than gray images.

G. Hyperparameters Ablation

We ablate the hyperparameter settings (learning rate, learning rate schedule, training steps) for TI training on clean and poisoned NovelConcepts10 datasets. Results are shown in Tables 2, 4, and 5 (we apologize for the nonconsecutive table order - the formatting for this section was difficult).

In general, we find that high learning rates and more training steps benefit performance for Clean and ADM- datasets, but reduce performance on ADM+ and EA. Low learning rates and training steps can inhibit learning (of both clean and adversarial signals). The default learning rate of $5e^{-4}$ yields balanced performance across all datasets and thus we focus on this setting. At a learning rate of $5e^{-4}$, a constant learning rate schedule generally gives best performance. Finally, 5000 training steps outperforms 1000 training steps across most poisoned datasets, and the marginal improvements (if any) seen with 10000 training steps are not worth the extended training time.

We note additional interesting behaviors from the hyperparameter ablation. In particular, we find that training for 1000 is almost equally affected by poisoning as training for 10000 steps, even as clean performance increases. This suggests that early stopping cannot avoid adversarial signals. Likewise, a higher learning rate boosts performance for clean images but reduces performance on concepts poisoned by ADM+ or EA. Using linear or cosine decay reduces performance when training for 1000 steps but does not significantly impact results for 5000 or 10000 steps. Finally, it appears that ADM- is the weakest poison across a wide range of settings.

Table 4. NovelConcepts10 FID for various TI hyperparameter settings. Best settings per poison in bold.

LR Sched.		Constant			Linear			Cosine		
	LR	$5e^{-5}$	$5e^{-4}$	$5e^{-3}$	$5e^{-5}$	$5e^{-4}$	$5e^{-3}$	$5e^{-5}$	$5e^{-4}$	$5e^{-3}$
Clean	1k	397	307	285	415	330	289	414	301	296
	5k	317	285	266	335	287	273	341	297	270
	10k	301	285	269	315	285	262	305	286	**258**
ADM+	1k	417	449	457	412	446	446	**411**	434	447
	5k	440	453	459	439	452	453	436	447	459
	10k	431	449	457	433	438	454	438	447	454
ADM-	1k	414	344	337	416	359	341	414	348	343
	5k	370	339	333	368	341	337	373	348	323
	10k	370	358	318	358	347	324	365	352	**300**
EA	1k	**414**	435	440	**414**	418	444	**414**	424	448
	5k	419	443	457	416	437	450	416	446	467
	10k	425	440	465	424	442	448	417	439	457

Table 5. NovelConcepts10 CLIP Score for various TI hyperparameter settings. Best settings per poison in bold.

LR Sched.		Constant			Linear			Cosine		
	LR	$5e^{-5}$	$5e^{-4}$	$5e^{-3}$	$5e^{-5}$	$5e^{-4}$	$5e^{-3}$	$5e^{-5}$	$5e^{-4}$	$5e^{-3}$
Clean	1k	0.37	0.48	0.50	0.34	0.47	0.46	0.34	0.50	**0.51**
	5k	0.47	**0.51**	0.46	0.45	**0.51**	0.48	0.44	**0.51**	**0.51**
	10k	0.50	0.49	0.43	0.46	**0.51**	0.48	0.50	0.47	0.48
ADM+	1k	0.40	0.50	0.46	0.35	**0.52**	0.48	0.35	0.49	0.51
	5k	0.46	0.46	0.42	0.48	0.46	0.44	0.46	0.48	0.45
	10k	0.45	0.45	0.40	0.48	0.46	0.44	0.49	0.45	0.44
ADM-	1k	0.38	0.50	0.47	0.37	0.46	**0.51**	0.37	0.45	0.50
	5k	0.48	0.45	0.45	0.44	0.48	0.45	0.43	0.45	0.47
	10k	0.45	0.46	0.41	0.46	0.43	0.45	0.46	0.47	0.43
EA	1k	0.37	**0.49**	0.46	0.35	0.44	0.46	0.36	0.45	0.46
	5k	0.45	0.45	0.43	0.39	0.45	0.43	0.39	0.46	0.45
	10k	0.44	0.44	0.38	0.43	0.44	0.43	0.42	0.44	0.40

H. Timestep Range Ablation

We ablate methods of restricting training to higher timesteps. We investigate high thresholding ($t \sim \mathcal{U}(\rho, 1)$), power distributions ($p(t) \propto t^{\rho}$), and tanh distributions ($p(t) \propto tanh(\rho(t-0.5))/2+0.5$). For comparison, we also evaluate low thresholding ($t \sim \mathcal{U}(0, \rho)$). We abuse notation and use ρ for various function parameters that control the shape of each sampling distribution; increasing ρ increases sampling probability for higher timesteps. Here, t is sampled in domain $[0, 1]$ and then the sampled output is rescaled to $[0, 1000]$ during training. Fig. 15 displays the tanh and power probability distributions for various ρ values.

It can be seen from ablation results in Tables 6, 7, and 8 that performance on datasets poisoned by ADM+ improves significantly as timestep sampling shifts towards higher timesteps. Performance on EA-poisoned concepts also increases slightly while performance on ADM- is relatively stable across methods. Simple high-thresholding is the often the best timestep restriction method. In all cases, we note a performance decrease when t is concentrated at extremely high timesteps (e.g., $t \geq 900$) since learning true features in this high-noise range is challenging. Lastly, we empirically validate the hypothesis that adversarial signals are concentrated at lower-middle timesteps by demonstrating that performance for low thresholding ($t \sim \mathcal{U}(0, \rho)$) is consistently worse than nominal sampling.

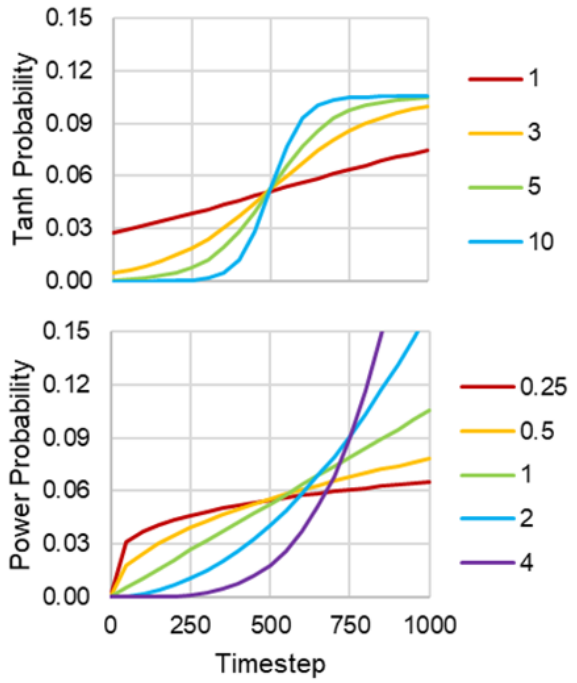

Figure 15. Probability curves for timestep sampling, displaying $tanh$ distributions (top) and power distributions (bottom). Distribution parameters for each curve are given in the legends.

Table 6. NovelConcepts10 DINOv2 Similarity for various timestep range restrictions at various ρ values. Best settings per poison in bold.

Curve	Nominal	Tanh				Power				
ρ	-	1	3	5	10	0.25	0.5	1	2	4
Clean	0.44	0.44	0.47	**0.49**	0.45	0.45	0.46	0.45	0.47	0.47
ADM+	0.10	0.13	0.11	0.20	0.18	0.10	0.10	0.12	0.24	0.27
ADM-	0.29	**0.31**	0.30	0.28	0.27	0.28	0.29	0.30	0.26	0.25
EA	0.08	0.10	0.09	0.10	0.11	0.08	0.08	0.10	0.10	0.11
Curve		Threshold Low				Threshold High				
ρ	0.1	0.2	0.3	0.4	0.5	0.5	0.6	0.7	0.8	0.9
Clean	0.17	0.24	0.29	0.31	0.37	0.46	0.44	0.39	0.37	0.31
ADM+	0.08	0.08	0.08	0.10	0.09	0.22	0.21	0.28	**0.33**	0.26
ADM-	0.13	0.16	0.20	0.27	0.26	0.26	0.25	0.23	0.19	0.18
EA	0.07	0.07	0.08	0.07	0.07	0.11	0.13	**0.14**	0.13	0.10

Table 7. NovelConcepts10 FID for various timestep range restrictions at various ρ values. Best settings per poison in bold.

Curve	Nominal	Tanh				Power				
ρ	-	1	3	5	10	0.25	0.5	1	2	4
Clean	285	288	271	**265**	283	281	274	279	270	**265**
ADM+	453	440	443	401	399	448	448	430	376	360
ADM-	339	**335**	341	344	351	351	344	**335**	349	362
EA	443	445	440	439	432	442	442	433	425	424
Curve		Threshold Low				Threshold High				
ρ	0.1	0.2	0.3	0.4	0.5	0.5	0.6	0.7	0.8	0.9
Clean	387	365	342	334	323	279	285	304	312	336
ADM+	442	448	449	451	447	388	389	352	**336**	364
ADM-	426	407	387	360	363	351	361	369	393	400
EA	437	441	445	443	451	424	414	**413**	**413**	439

Table 8. NovelConcepts10 CLIP Score for various timestep range restrictions at various ρ values. Best settings per poison in bold.

Curve	Nominal	Tanh				Power				
ρ	-	1	3	5	10	0.25	0.5	1	2	4
Clean	0.51	**0.53**	0.51	0.50	0.52	0.51	0.51	**0.53**	0.47	0.50
ADM+	0.46	0.49	0.46	**0.53**	0.50	0.47	0.47	0.47	0.49	0.52
ADM-	0.45	0.47	**0.49**	0.47	0.44	**0.49**	0.48	**0.49**	0.47	0.44
EA	0.45	0.47	0.46	0.45	0.44	0.47	0.46	0.46	0.48	0.45
Curve		Threshold Low				Threshold High				
ρ	0.1	0.2	0.3	0.4	0.5	0.5	0.6	0.7	0.8	0.9
Clean	0.44	0.45	0.43	0.49	0.49	0.51	0.49	0.50	0.49	0.45
ADM+	0.38	0.41	0.43	0.44	0.44	0.49	0.50	0.48	0.48	0.47
ADM-	0.44	0.45	0.45	0.48	0.46	0.44	0.45	0.48	**0.49**	0.48
EA	0.41	0.43	0.40	0.43	0.45	0.44	0.46	0.48	**0.49**	**0.49**

I. Masking Ablation

Table 9. NovelConcepts10 DINOv2 Similarity for different masking methods.

Poison	Nominal	LM	IM	LIM	ZM
Clean	0.44	**0.45**	0.38	0.42	0.20
ADM+	0.10	**0.36**	0.28	0.33	0.11
ADM-	0.29	**0.41**	0.31	0.31	0.22
SDS+	0.11	**0.33**	0.30	0.26	0.14
SDS-	0.25	**0.39**	0.29	0.29	0.16
EA	0.08	**0.30**	0.29	0.26	0.15
DA	0.27	**0.38**	0.34	0.31	0.16
Psn Avg	0.18	**0.36**	0.30	0.29	0.16

Table 10. NovelConcepts10 FID for different masking methods.

Poison	Nominal	LM	IM	LIM	ZM
Clean	**285**	**285**	299	293	406
ADM+	453	**309**	344	342	439
ADM-	339	**299**	330	341	378
SDS+	439	**341**	**341**	365	428
SDS-	352	**304**	346	342	405
EA	443	359	**348**	365	421
DA	352	**309**	328	333	434
Psn Avg	396	**320**	340	348	417

Table 11. NovelConcepts10 CLIP Score for different masking methods.

Poison	Nominal	LM	IM	LIM	ZM
Clean	0.51	**0.52**	0.48	**0.52**	0.46
ADM+	0.46	0.47	0.48	**0.52**	0.42
ADM-	0.45	0.54	0.50	**0.51**	0.44
SDS+	0.44	**0.48**	**0.48**	**0.48**	0.45
SDS-	0.45	0.49	0.49	**0.50**	0.41
EA	0.45	**0.50**	0.49	0.48	0.42
DA	0.49	0.48	0.52	**0.53**	0.46
Psn Avg	0.46	0.49	0.49	**0.50**	0.43

We define the various objectives used by loss masking (LM), input masking (IM), loss-input masking (LIM), and latent masking (ZM). The notation here follows that of Sections 3 and 4.5. The LM objective is given by

$$L_{LM}(x,t,c,M_x) = ||(\epsilon_\theta(z_t,t,c) - \epsilon) \odot M_z||_2^2, \quad (12)$$

which is similar to the SZT objective from Eq. 7. IM applies masking only to the input x and is given by

$$L_{IM}(x,t,c,M_x) = ||\epsilon_\theta(z_{tM},t,c) - \epsilon||_2^2, \quad (13)$$

where z_{tM} is the noised latent of a masked input image, $z_{tM} = \sqrt{\overline{\alpha}_t}\mathcal{E}(x \odot M_x) + \sqrt{1-\overline{\alpha}_t}\epsilon$. LIM combines the L_{LM} and L_{IM} objectives as

$$L_{LIM}(x,t,c,M_x) = ||(\epsilon_\theta(z_{tM},t,c) - \epsilon) \odot M_z||_2^2. \quad (14)$$

ZM applies masking to the latent vector z_t and is given by

$$L_{ZM}(x,t,c,M_x) = ||\epsilon_\theta(z_t \odot M_z,t,c) - \epsilon||_2^2. \quad (15)$$

We evaluate masking types on NovelConcepts10 across all poisons. Tables 9, 10, and 11 demonstrate that LM outperforms all other forms of masking for poison defense, as it is the only method that fully preserves background information in the forward process. IM performs underwhelmingly and LIM is apparently limited by the input mask applied by IM. ZM consistently gives the lowest performance.

To further evaluate the impact of LM, we measure the proportion of SSM values within novel concept regions compared to the sum of all SSM values throughout the image. This metric can be captured using the ratio defined below, with notation mirroring that of Sec. 4.1:

$$R_{n,M_z} = \frac{\sum_{i,j}(SSM(x,t,\widehat{e},n) \odot M_z)_{i,j}}{\sum_{i,j} SSM(x,t,\widehat{e},n)_{i,j}}. \quad (16)$$

We measure R_{n,M_z} for clean and poisoned versions of NovelConcepts10 at steps 100, 500, and 900 during TI training and average the values across all concepts. The results in Fig. 16 show that the ratio of SSM within novel concept regions naturally increases throughout training for clean data. For poisoned datasets without LM, the proportion of SSM in the novel concept regions never increases, indicating that learning is distracted away from the novel concept regions by adversarial signals. Only with LM does TI focus on the novel concept regions for poisoned datasets.

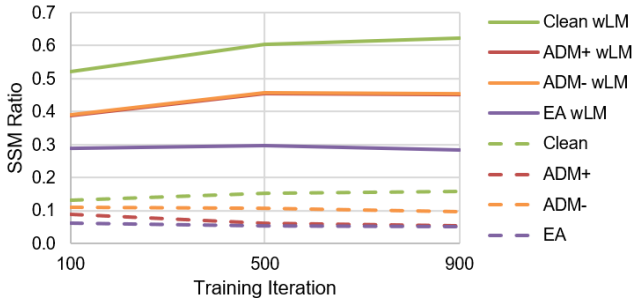

Figure 16. Average R_{n,M_z} for clean and poisoned versions of NovelConcepts10 throughout TI training. "wLM" denotes TI with loss masking.

J. Mask Dilation Ablation

In initial experiments with SZT, we found that combining T600 and LM tended to perform below expectations. Although the performance of T600+LM was typically better than established defenses like Regen and AdvClean, it underperformed other SZT ablations like JPEG+T600 or JPEG+LM. We hypothesize that strong restrictions in both time (i.e., T600) and space (i.e., LM) may be too restrictive and may hinder concept learning during TI training. Therefore, we evaluated additional configurations of SZT that ease temporal and spatial restrictions when used in combination. In particular, we investigate lower timestep threshold (e.g., $t \geq 500$, denoted as "T500") combined with dilated concept masks. To implement mask dilations for LM, we apply the `ImageFilter.MaxFilter` method from PIL with 8, 16, or 24 pixels of dilation to the 512x512 binary masks, then rescale to 64x64 for latent space. Intuitively, applying mask dilation includes extra background information outside of the novel concept region during loss backpropagation.

Tables 12, 13, and 14 display the DINOv2 similarity, FID, and CLIP Score for various combinations of T500, T600, and dilated masks (denoted "LM-D08", "LM-D16", and "LM-D24"). Of the various LM configurations, 16 pixels of dilation (LM-D16) demonstrates the highest robustness to poisons. Furthermore, when combining T500 or T600 with LM, using T500+LM-D16 gives the best performance, supporting our "too restrictive" hypothesis above. Our final implementation of SZT uses JPEG preprocessing with T500+LM-D16 and further improves poison defense beyond all other ablations.

K. Defense Comparison for CustomConcept101

We additionally include FID and CLIP Score metrics for CustomConcept101 in Tables 15 and 16, complementing the DINOv2 Similarity results in Table 1. Trends in defense methods are generally similar as those observed in Sec. 5.2.4 As observed for the DINOv2 Similarity results, SZT is the best method for poison defense.

For emphasis, we plot the DINOv2 Similarity versus CLIP Score values for "Psn Avg" on CustomConcept101 for all defenses in Fig. 17. As discussed in Sec. 5.2.4, Regen, PDMPure, and AdvClean all offer minor improvements in poison performance. We note the that despite a drastic improvement in DINOv2 Similarity, JPEG is limited in its prompt fidelity (measured by CLIP Score). This aligns with qualitative observations of its limited concept learning from Fig. 8. SZT improves DINOv2 Similarity and CLIP Score beyond all existing defenses, and most ablations of SZT also perform well. We note that all ablations of SZT can beat existing defenses in at least one metric.

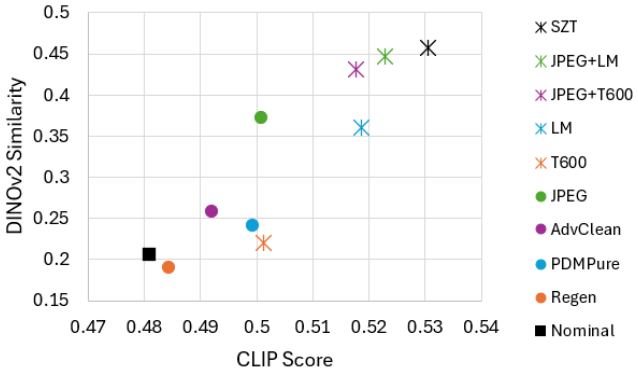

Figure 17. DINOv2 Similarity versus CLIP Score for all defenses on CustomConcept101. Values are from the "Psn Avg" rows in Tables 1 and 16.

L. Defense Comparison for NovelConcepts10

We repeat the same study as Sections 5.2.4 and K for NovelConcepts10 and report results in Tables 17, 18, and 19. All trends for NovelConcepts10 generally mirror those seen for CustomConcept101, validating our prior observations.

M. Generated Images after TI

We display generated images after TI for various concepts, poisons, and defenses in figures 18, 19, 20, and 21. For each concept/poison/defense, we utilized DINOv2 similarity between generated images and training images to identify the most faithful image for display. Fig. 18 compares five concepts from CustomConcept101, each poisoned by ADM+. Fig. 19 focuses on the things-bottle1 concept from CustomConcept101 across multiple poisons. Figures 20 and 21 are analogous, but for NovelConcepts10. For convenience, we also include samples of the original training images in Figures 18 and 20 (left-most column). Observations for all figures here reflect those made for Fig. 8 in Sec. 5.2.4.

Table 12. NovelConcepts10 DINOv2 Similarity for various timestep-restriction and masking ablations.

Defense → Poison ↓	Nominal	LM	LM D08	LM D16	LM D24	JPEG +LM	JPEG+ LM-D16	T500 +LM	T500+ LM-D16	T600 +LM	T600+ LM-D16
Clean	0.46	0.45	0.45	0.45	**0.47**	0.43	0.46	0.44	0.46	0.44	0.46
ADM+[39]	0.10	0.36	0.39	**0.45**	**0.45**	0.43	0.44	0.39	0.40	0.33	0.36
ADM-	0.32	0.41	0.43	**0.44**	0.42	0.43	0.42	0.25	0.31	0.22	0.24
SDS+[78]	0.09	0.30	**0.44**	**0.44**	0.41	0.43	0.42	0.24	0.25	0.21	0.24
SDS-[78]	0.11	0.33	0.40	**0.47**	0.44	0.42	0.44	0.36	0.42	0.33	0.37
EA[61]	0.25	0.39	0.41	0.41	**0.42**	0.41	**0.42**	0.23	0.28	0.19	0.20
DA[61]	0.27	0.38	**0.43**	**0.43**	0.42	0.42	0.42	0.29	0.34	0.33	0.34
Psn Avg	0.19	0.36	0.42	**0.44**	0.43	0.42	0.43	0.29	0.33	0.27	0.29

Table 13. NovelConcepts10 FID for various timestep-restriction and masking ablations.

Defense → Poison ↓	Nominal	LM	LM D08	LM D16	LM D24	JPEG +LM	JPEG+ LM-D16	T500 +LM	T500+ LM-D16	T600 +LM	T600+ LM-D16
Clean	281	285	280	281	281	291	281	284	275	289	**272**
ADM+[39]	451	309	296	279	**275**	291	290	297	284	312	309
ADM-	323	299	291	**287**	296	293	302	375	341	374	378
SDS+[78]	434	359	**288**	291	302	296	300	372	358	390	381
SDS-[78]	439	341	297	**273**	281	299	288	311	283	318	303
EA[61]	352	304	298	295	**285**	296	297	378	354	381	389
DA[61]	352	309	295	**286**	295	299	298	350	321	343	328
Psn Avg	392	320	294	**285**	289	296	296	347	324	353	348

Table 14. NovelConcepts10 CLIP Score for various timestep-restriction and masking ablations.

Defense → Poison ↓	Nominal	LM	LM D08	LM D16	LM D24	JPEG +LM	JPEG+ LM-D16	T500 +LM	T500+ LM-D16	T600 +LM	T600+ LM-D16
Clean	**0.54**	0.52	0.48	0.52	0.52	0.50	0.52	0.50	0.45	0.50	0.48
ADM+[39]	0.46	0.47	0.48	0.50	0.48	0.51	0.52	0.52	**0.53**	0.50	0.48
ADM-	0.49	**0.54**	0.48	0.51	0.46	0.50	0.49	0.45	0.48	0.48	0.46
SDS+[78]	0.45	0.50	0.50	0.48	0.48	0.47	0.47	0.49	0.51	**0.52**	0.49
SDS-[78]	0.44	0.48	0.46	0.51	0.50	0.50	**0.54**	0.45	0.53	0.50	0.49
EA[61]	0.45	0.49	0.47	0.46	0.46	0.51	**0.53**	0.46	0.49	0.49	0.48
DA[61]	0.49	0.48	0.47	0.46	0.48	**0.53**	**0.53**	0.49	0.48	0.51	**0.53**
Psn Avg	0.46	0.49	0.48	0.49	0.48	0.50	**0.51**	0.48	0.50	0.50	0.49

Table 15. CustomConcept101 FID for various poison defenses.

Defense → Poison ↓	Nominal	Regen [84]	PDMPure [77]	AdvClean [65]	JPEG	T600	LM	JPEG +T600	JPEG +LM	SZT
Clean	251	255	356	256	257	**235**	250	246	251	245
ADM+[39]	422	386	363	399	292	344	332	269	259	**257**
ADM-	302	349	367	297	271	333	267	256	256	**252**
SDS+[78]	420	389	357	393	301	348	322	275	261	**259**
SDS-[78]	311	357	367	299	273	355	272	259	261	**248**
EA[61]	413	421	365	367	302	386	329	260	262	**248**
DA[61]	323	341	347	298	278	358	273	260	255	**249**
Psn Avg	365	374	361	342	286	354	299	263	259	**252**

Table 16. CustomConcept101 CLIP Score for various poison defenses.

Defense → Poison ↓	Nominal	Regen [84]	PDMPure [77]	AdvClean [65]	JPEG	T600	LM	JPEG +T600	JPEG +LM	SZT
Clean	**0.53**	**0.53**	0.48	0.52	0.52	**0.53**	**0.53**	0.51	**0.53**	**0.53**
ADM+[39]	0.47	0.47	0.50	0.47	0.49	0.53	0.50	0.52	0.51	**0.54**
ADM-	0.50	0.50	0.48	0.51	0.50	0.47	**0.52**	0.51	**0.52**	**0.52**
SDS+[78]	0.46	0.49	0.50	0.48	0.50	**0.53**	0.51	**0.53**	**0.53**	0.52
SDS-[78]	0.48	0.48	0.50	0.51	0.50	0.46	**0.53**	0.50	0.52	**0.53**
EA[61]	0.47	0.48	0.50	0.47	0.49	0.48	0.51	0.52	0.52	**0.53**
DA[61]	0.50	0.49	0.51	0.51	0.52	**0.54**	**0.54**	0.52	0.53	**0.54**
Psn Avg	0.48	0.48	0.50	0.49	0.50	0.50	0.52	0.52	0.52	**0.53**

Table 17. NovelConcepts10 DINOv2 Similarity for various poison defenses.

Defense → Poison ↓	Nominal	Regen [84]	PDMPure [77]	AdvClean [65]	JPEG	T600	LM	JPEG +T600	JPEG +LM	SZT
Clean	0.44	0.43	0.24	0.44	0.41	0.44	0.45	0.45	0.43	**0.47**
ADM+[39]	0.10	0.14	0.24	0.13	0.28	0.21	0.36	0.36	**0.43**	**0.43**
ADM-	0.29	0.27	0.21	0.31	0.34	0.25	0.41	0.42	0.43	**0.44**
SDS+[78]	0.11	0.16	0.22	0.12	0.23	0.18	0.33	0.36	0.42	**0.44**
SDS-[78]	0.25	0.25	0.20	0.30	0.30	0.21	0.39	0.40	0.41	**0.45**
EA[61]	0.08	0.09	0.19	0.16	0.28	0.13	0.30	0.40	0.43	**0.44**
DA[61]	0.27	0.26	0.23	0.28	0.33	0.25	0.38	0.40	0.42	**0.47**
Psn Avg	0.18	0.19	0.22	0.21	0.29	0.21	0.36	0.39	0.42	**0.45**

Table 18. NovelConcepts10 FID for various poison defenses.

Defense → Poison ↓	Nominal	Regen [84]	PDMPure [77]	AdvClean [65]	JPEG	T600	LM	JPEG +T600	JPEG +LM	SZT
Clean	285	285	380	281	294	285	285	280	291	**276**
ADM+[39]	453	416	378	443	349	389	309	311	291	**280**
ADM-	339	353	390	336	324	361	299	294	293	**279**
SDS+[78]	439	410	389	433	374	399	341	317	299	**283**
SDS-[78]	352	366	392	338	335	379	304	300	296	**277**
EA[61]	443	450	394	410	349	414	359	308	296	**282**
DA[61]	352	350	373	354	323	367	309	305	299	**269**
Psn Avg	396	391	386	385	342	385	320	306	296	**278**

Table 19. NovelConcepts10 CLIP Score for various poison defenses.

Defense → Poison ↓	Nominal	Regen [84]	PDMPure [77]	AdvClean [65]	JPEG	T600	LM	JPEG +T600	JPEG +LM	SZT
Clean	0.51	0.47	0.49	**0.53**	0.48	0.49	0.52	0.51	0.50	0.52
ADM+[39]	0.46	0.47	0.49	0.46	0.47	0.50	0.47	0.50	0.51	**0.52**
ADM-	0.45	0.50	0.46	0.50	0.46	0.45	**0.54**	0.51	0.50	0.49
SDS+[78]	0.44	0.49	0.47	0.48	0.47	0.48	0.48	0.48	**0.50**	0.49
SDS-[78]	0.45	0.48	0.48	0.48	0.48	0.44	0.49	0.51	0.51	**0.52**
EA[61]	0.45	0.45	0.47	0.48	0.48	0.46	**0.50**	0.48	0.47	0.48
DA[61]	0.49	0.49	0.48	0.49	0.48	0.52	0.48	0.52	**0.53**	0.46
Psn Avg	0.46	0.48	0.48	0.48	0.47	0.47	0.49	**0.50**	**0.50**	0.49

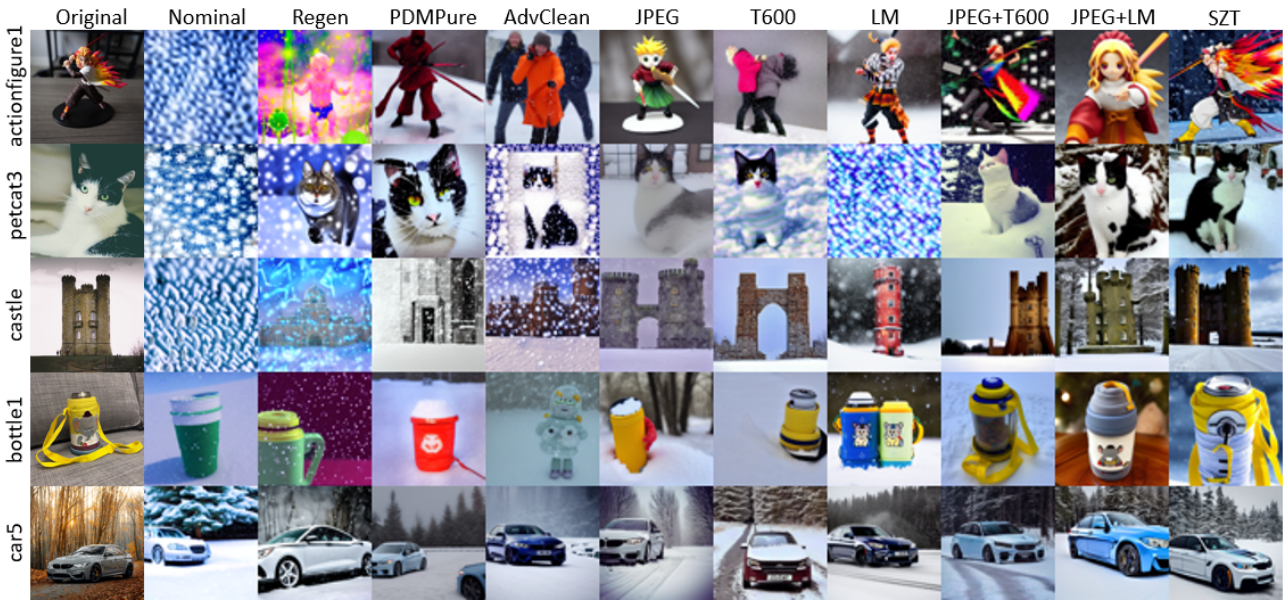

Figure 18. Images generated after TI on various concepts from CustomConcept101 poisoned by ADM+. Prompt: "a $R*$ in the snow"

Figure 19. Images generated after TI on things-bottle1 (CustomConcept101) for various poisons. Prompt: "a $R*$ in the snow"

Figure 20. Images generated after TI on various concepts from NovelConcepts10 poisoned by ADM+. Prompt: "a $R*$ on the beach"

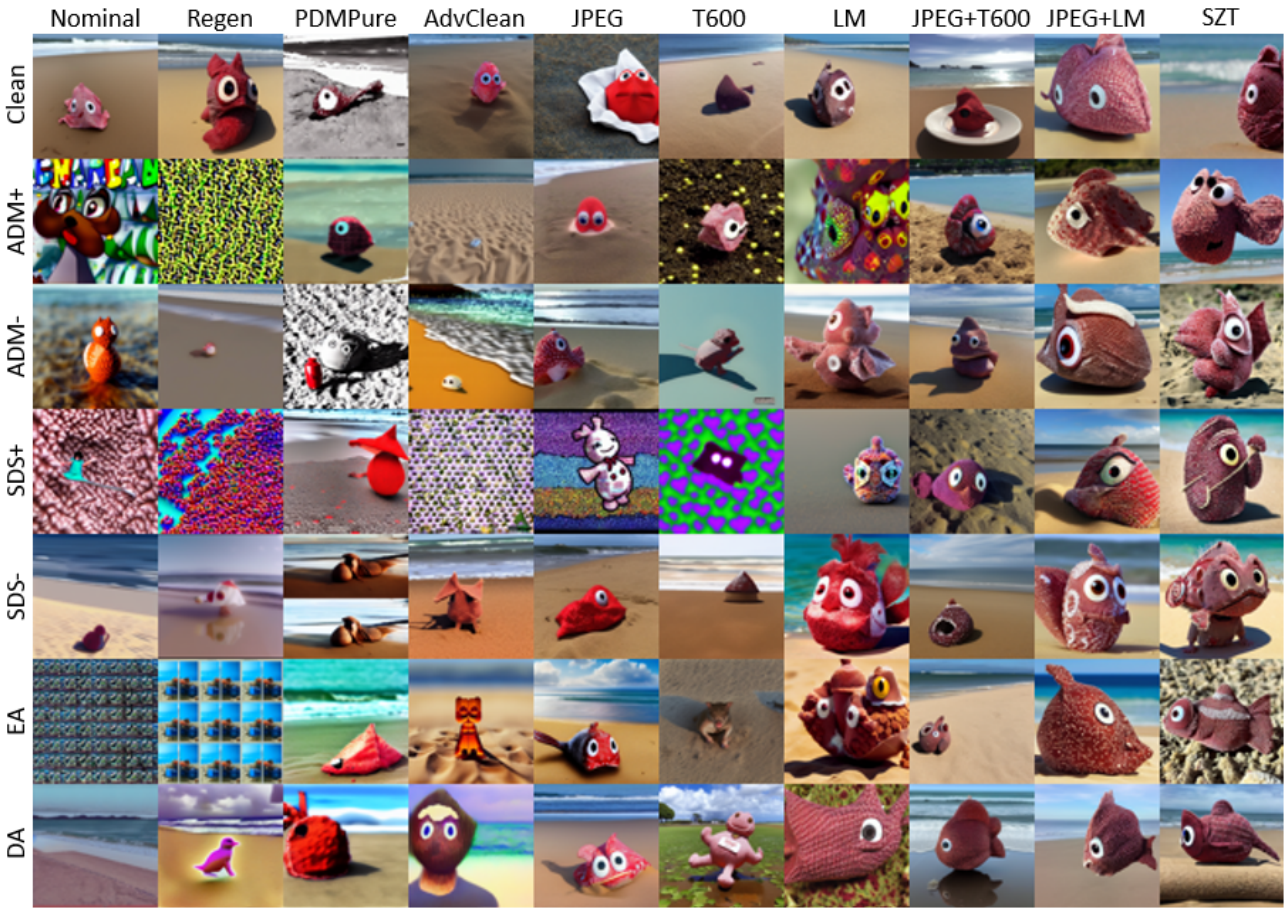

Figure 21. Images generated after TI on FishDoll (NovelConcepts10) for various poisons. Prompt: "a $R*$ on the beach"

N. JPEG Preprocessing for Other Defenses

Given the success of JPEG with SZT in Sec. 5.2.4, we further analyze JPEG preprocessing for existing defenses in Tables 20, 21, and 22. In general, adding JPEG preprocessing improves performance for Regeneration and AdverseCleaner, but reduces performance for PDMPure. SZT still outperforms all baseline defenses with JPEG processing.

O. Additional Diffusion Models

We further validate the effectiveness of SZT and existing defenses for TI applied to multiple LDMs. We analyze Stable Diffusion 1.5 (rectified flow version) [43], Stable Diffusion 2.1 (noise-prediction), Stable Diffusion 2.1 (velocity-prediction) [40], and SDXL [55]. For convenience, we abbreviate the models as SD15, SD15rf, SD21, SD21v, and SDXL. We evaluate all models on NovelConcepts10 for all poisons. Training settings are generally similar to those described in Sec. 5.1.2 except that we use 2500 training steps. Due to the large size of SDXL, we instead use 500 training steps, each with 5 steps of gradient accumulation. We were unable to track CLIP Score for SD21, SD21v, or SDXL due to their text encoders' incompatibilities with the OpenAI CLIP image encoder.

This experiment approaches a "black-box" poisoning scenario, as we craft all poisons using SD15 and then test defenses on different Stable Diffusion versions. This is not a major concern, as multiple prior works have demonstrated the strong transferability of poisons across Stable Diffusion models [37, 78]. In multiple tables for SD15rf, SD21, and SD21v results below, the "Nominal" defense column indeed shows that all poisons are effective. Network-based defenses (Regen and PDMPure) shown here still use the same models for purification.

We evaluate SD15rf in Tables 23, 24, and 25. We suspect that the poor performance of SD15rf is due to the fact that it derives from a reflow procedure [42]. We hypothesize that reflow, which relies on deterministic sampling for data/noise pairs to finetune models, acts as a type of model distillation and forces models to unlearn natural ODE paths between probability distributions. The rigid ODE paths enforced by distilled models may lack the flexibility to insert new concepts into the model via personalization. Existing defenses like Regen and AdvClean barely improve the poison performance. Even so, SZT improves generative quality for TI on poisoned data to match that of clean data.

Poison and defense performance on SD21 in Tables 26 and 27 is generally similar to that of SD15, which is sensible given their similar architectures, training data, and training objectives. The trends for SD21v in Tables 28 and 29 are generally similar to those of SD15 and SD21, though the effect of poisoning (i.e., the difference between "Clean" and "Psn Avg" values) is diminished relative to SD15 and SD21. Even so, SZT and its ablations are still effective for improving generation quality on poisoned data. In Tables 30 and 31, SDXL demonstrates the largest rift in performance trends relative to SD15, being unaffected by most poisons. Though SZT and its ablations can improve performance on poison data they are generally not necessary. We are uncertain of the exact cause of the low degree of poison transferability from SD15 to SDXL.

P. Additional Personalization Methods

In addition to TI, we investigate additional personalization methods for Stable Diffusion 1.5, namely LoRA [25] and CustomDiffusion [35]. We use NovelConcepts10 for these experiments. LoRA personalizes LDMs by finetuning low-rank adapters for U-Net weights (and optionally for the text encoder). LoRA does not finetune any text embeddings, though it does associate finetuned weights with a target prompt. CustomDiffusion is a compact version of DreamBooth [59] that only finetunes the cross-attention weights of the U-Net as well as a new text token (as in TI). Notably, the CustomDiffusion paper also introduced a crop/rescale augmentation that applies as a mask during loss calculation, similar to our LM strategy.

In our implementation of LoRA, we use rank 4 adapters with a target prompt that includes "*" as a dummy token to signify LoRA usage and trained for 1000 steps. We did not find any improvement in generative quality when using the prior preservation loss or when training text encoder weights, and thus our implementation does not include these methods. We report LoRA results in Tables 32, 33, and 34. We find that poisons are moderately effective against LoRA, with EA being the most severe. Existing defenses like PDMPure and AdvClean slightly improve defense against poisons but Regen generally reduces generative quality. SZT and its ablations outperform existing defenses to achieve clean-level generative quality for poisoned images. We note that as LoRA does not provide any finetuned text tokens, the CLIP Score becomes a metric for background fidelity; the CLIP text encoder will ignore the dummy token ("*") and instead focus on context words.

In our implementation of CustomDiffusion, we train for 1000 steps and do not use the prior preservation loss. We report results both without their novel crop/rescale augmentation (Tables 35, 36, and 37) and with it (Tables 38, 39, and 40). Without crop/rescale, performance is generally similar to that of LoRA, with SZT outperforming existing defenses to improve poison performance. With crop/rescale, all poisons are almost ineffective and most defenses (with the exception of JPEG+T600) only reduce generative performance. This suggests that their novel crop/rescale augmentation, which is applied during loss calculation, acts similarly to our LM method, effectively blocking out adversarial signals outside the novel concept region.

Table 20. NovelConcepts10 DINOv2 Similarity for existing defenses with JPEG preprocessing.

Defense → Poison ↓	Nominal	Regen [84]	PDMPure [77]	AdvClean [65]	JPEG	JPEG + Regen	JPEG + PDMPure	JPEG + AdvClean	SZT
Clean	0.44	0.43	0.24	0.45	0.41	0.37	0.19	0.38	**0.47**
ADM+[39]	0.10	0.14	0.24	0.13	0.28	0.26	0.20	0.30	**0.43**
ADM-	0.29	0.27	0.21	0.31	0.34	0.35	0.19	0.33	**0.44**
SDS+[78]	0.11	0.16	0.22	0.12	0.23	0.23	0.19	0.29	**0.44**
SDS-[78]	0.25	0.25	0.20	0.30	0.30	0.35	0.19	0.35	**0.45**
EA[61]	0.08	0.09	0.19	0.16	0.28	0.29	0.22	0.32	**0.44**
DA[61]	0.27	0.26	0.23	0.28	0.33	0.33	0.17	0.35	**0.47**
Psn Avg	0.18	0.19	0.22	0.21	0.29	0.30	0.19	0.32	**0.45**

Table 21. NovelConcepts10 FID for existing defenses with JPEG preprocessing.

Defense → Poison ↓	Nominal	Regen [84]	PDMPure [77]	AdvClean [65]	JPEG	JPEG + Regen	JPEG + PDMPure	JPEG + AdvClean	SZT
Clean	285	285	380	281	294	303	389	303	**276**
ADM+[39]	453	416	378	443	349	355	389	335	**280**
ADM-	339	353	390	336	324	303	387	315	**279**
SDS+[78]	439	410	389	433	374	364	388	335	**283**
SDS-[78]	352	366	392	338	335	310	398	309	**277**
EA[61]	443	450	394	410	349	335	370	321	**282**
DA[61]	352	350	373	354	323	317	396	311	**269**
Psn Avg	396	391	386	385	342	331	388	321	**278**

Table 22. NovelConcepts10 CLIP Score for existing defenses with JPEG preprocessing.

Defense → Poison ↓	Nominal	Regen [84]	PDMPure [77]	AdvClean [65]	JPEG	JPEG + Regen	JPEG + PDMPure	JPEG + AdvClean	SZT
Clean	0.51	0.47	0.49	**0.53**	0.48	0.49	0.46	0.52	0.52
ADM+[39]	0.46	0.47	0.49	0.46	0.47	0.49	0.51	0.47	**0.52**
ADM-	0.45	**0.50**	0.46	**0.50**	0.46	0.48	0.49	**0.50**	0.49
SDS+[78]	0.44	0.49	0.47	0.48	0.47	0.50	**0.51**	0.47	0.49
SDS-[78]	0.45	0.48	0.48	0.48	0.48	0.50	0.49	0.51	**0.52**
EA[61]	0.45	0.45	0.47	0.48	0.48	**0.49**	**0.49**	0.48	0.48
DA[61]	**0.49**	**0.49**	0.48	**0.49**	0.48	0.48	0.44	**0.49**	0.46
Psn Avg	0.46	0.48	0.48	0.48	0.47	**0.49**	**0.49**	**0.49**	**0.49**

Table 23. NovelConcepts10 DINOv2 Similarity for various poison defenses with Stable Diffusion 1.5 (rectified flow).

Defense → Poison ↓	Nominal	Regen [84]	PDMPure [77]	AdvClean [65]	JPEG	T600	LM	JPEG +T600	JPEG +LM	SZT
Clean	0.21	0.23	0.20	0.20	0.18	0.17	**0.30**	0.19	0.28	0.28
ADM+[39]	0.03	0.04	0.17	0.02	0.08	0.18	0.09	0.16	0.21	**0.24**
ADM-	0.12	0.13	0.20	0.13	0.15	0.12	0.18	0.14	0.20	**0.23**
SDS+[78]	0.04	0.03	0.15	0.04	0.09	0.17	0.10	0.15	0.21	**0.24**
SDS-[78]	0.13	0.12	0.18	0.14	0.12	0.10	0.17	0.15	**0.22**	0.18
EA[61]	0.07	0.09	0.21	0.10	0.14	0.10	0.14	0.13	**0.22**	0.19
DA[61]	0.10	0.08	0.17	0.10	0.15	0.11	0.19	0.16	0.21	**0.23**
Psn Avg	0.08	0.08	0.18	0.09	0.12	0.13	0.14	0.15	0.21	**0.22**

Table 24. NovelConcepts10 FID for various poison defenses with Stable Diffusion 1.5 (rectified flow).

Defense → Poison ↓	Nominal	Regen [84]	PDMPure [77]	AdvClean [65]	JPEG	T600	LM	JPEG +T600	JPEG +LM	SZT
Clean	385	364	380	386	391	389	**343**	385	356	352
ADM+[39]	440	443	394	455	419	388	416	396	379	**367**
ADM-	415	408	381	408	407	403	393	402	383	**370**
SDS+[78]	440	442	401	438	426	398	417	400	367	**364**
SDS-[78]	407	416	397	409	409	410	395	398	**371**	399
EA[61]	430	429	384	421	399	419	409	408	**375**	387
DA[61]	422	422	400	425	400	408	395	389	374	**368**
Psn Avg	426	427	393	426	410	404	404	399	**375**	376

Table 25. NovelConcepts10 CLIP Score for various poison defenses with Stable Diffusion 1.5 (rectified flow).

Defense → Poison ↓	Nominal	Regen [84]	PDMPure [77]	AdvClean [65]	JPEG	T600	LM	JPEG +T600	JPEG +LM	SZT
Clean	0.44	0.41	0.46	0.43	0.42	0.46	**0.47**	0.43	**0.47**	0.44
ADM+[39]	0.40	0.41	0.45	0.38	0.43	0.42	0.40	0.42	0.46	**0.48**
ADM-	0.41	0.44	**0.46**	0.43	0.44	0.45	**0.46**	0.43	0.45	0.45
SDS+[78]	0.37	0.37	0.44	0.39	0.42	0.45	0.40	0.44	**0.47**	0.46
SDS-[78]	0.39	0.45	0.45	0.45	0.42	0.44	**0.47**	0.44	**0.47**	0.46
EA[61]	0.40	0.42	0.45	0.44	0.42	0.43	0.43	0.44	0.45	**0.46**
DA[61]	0.43	0.45	0.45	0.43	0.44	0.43	0.46	0.44	0.45	**0.47**
Psn Avg	0.40	0.42	0.45	0.42	0.43	0.44	0.43	0.43	**0.46**	**0.46**

Table 26. NovelConcepts10 DINOv2 Similarity for various poison defenses with Stable Diffusion 2.1 (noise-prediction).

Defense → Poison ↓	Nominal	Regen [84]	PDMPure [77]	AdvClean [65]	JPEG	T600	LM	JPEG +T600	JPEG +LM	SZT
Clean	0.41	0.39	0.31	0.41	0.38	0.43	**0.44**	0.41	**0.44**	0.43
ADM+[39]	0.10	0.14	0.29	0.10	0.30	0.20	0.34	0.38	0.40	**0.44**
ADM-	0.21	0.29	0.26	0.33	0.34	0.28	0.42	0.37	**0.44**	0.40
SDS+[78]	0.08	0.13	0.24	0.10	0.32	0.18	0.36	0.32	**0.44**	0.43
SDS-[78]	0.23	0.24	0.24	0.26	0.31	0.22	0.41	0.37	0.42	**0.43**
EA[61]	0.09	0.07	0.26	0.10	0.26	0.12	0.33	0.38	**0.43**	0.41
DA[61]	0.27	0.23	0.26	0.31	0.35	0.27	0.42	0.38	**0.45**	0.44
Psn Avg	0.16	0.18	0.26	0.20	0.31	0.21	0.38	0.36	**0.43**	0.42

Table 27. NovelConcepts10 FID for various poison defenses with Stable Diffusion 2.1 (noise-prediction).

Defense → Poison ↓	Nominal	Regen [84]	PDMPure [77]	AdvClean [65]	JPEG	T600	LM	JPEG +T600	JPEG +LM	SZT
Clean	290	305	356	298	301	295	**279**	294	288	286
ADM+[39]	440	421	355	429	339	393	319	306	295	**276**
ADM-	377	342	365	320	320	341	291	313	**283**	303
SDS+[78]	445	439	374	433	322	399	317	323	289	**286**
SDS-[78]	362	368	383	356	318	371	299	313	**290**	295
EA[61]	421	448	372	406	355	417	335	310	**281**	303
DA[61]	350	370	374	340	310	361	299	313	**280**	291
Psn Avg	399	398	370	380	327	380	310	313	**286**	292

Table 28. NovelConcepts10 DINOv2 Similarity for various poison defenses with Stable Diffusion 2.1 (velocity-prediction).

Defense → Poison ↓	Nominal	Regen [84]	PDMPure [77]	AdvClean [65]	JPEG	T600	LM	JPEG +T600	JPEG +LM	SZT
Clean	0.28	0.31	0.26	0.34	0.30	0.29	0.39	0.31	**0.41**	0.39
ADM+[39]	0.25	0.23	0.29	0.30	0.29	0.24	0.37	0.24	**0.42**	0.37
ADM-	0.20	0.16	0.25	0.16	0.22	0.17	0.31	0.28	0.36	**0.38**
SDS+[78]	0.21	0.21	0.31	0.27	0.25	0.23	0.33	0.28	**0.40**	0.38
SDS-[78]	0.16	0.14	0.28	0.17	0.35	0.17	0.27	0.26	**0.38**	**0.38**
EA[61]	0.09	0.09	0.28	0.21	0.26	0.13	0.27	0.26	0.38	**0.40**
DA[61]	0.21	0.17	0.26	0.28	0.29	0.16	0.28	0.27	**0.37**	**0.37**
Psn Avg	0.19	0.17	0.28	0.23	0.28	0.18	0.30	0.27	**0.38**	**0.38**

Table 29. NovelConcepts10 FID for various poison defenses with Stable Diffusion 2.1 (velocity-prediction).

Defense → Poison ↓	Nominal	Regen [84]	PDMPure [77]	AdvClean [65]	JPEG	T600	LM	JPEG +T600	JPEG +LM	SZT
Clean	336	328	343	321	326	334	308	330	**295**	314
ADM+[39]	355	362	332	332	335	370	309	360	**292**	309
ADM-	376	389	353	381	368	381	329	347	**313**	319
SDS+[78]	393	385	327	332	341	367	323	350	**294**	309
SDS-[78]	388	401	336	383	318	386	341	345	**306**	317
EA[61]	431	436	336	370	348	420	371	352	**300**	316
DA[61]	369	378	348	331	341	382	361	357	**310**	315
Psn Avg	385	392	339	355	342	384	339	352	**303**	314

Table 30. NovelConcepts10 DINOv2 Similarity for various poison defenses with SDXL.

Defense → Poison ↓	Nominal	Regen [84]	PDMPure [77]	AdvClean [65]	JPEG	T600	LM	JPEG +T600	JPEG +LM	SZT
Clean	0.35	0.33	0.25	0.33	0.22	0.41	**0.42**	0.34	0.39	0.39
ADM+[39]	0.36	0.29	0.24	0.31	0.25	0.39	**0.42**	0.32	0.34	0.32
ADM-	0.30	0.26	0.23	0.34	0.20	**0.40**	0.38	0.32	0.38	0.39
SDS+[78]	0.34	0.37	0.27	0.32	0.29	0.39	0.41	0.39	0.38	**0.42**
SDS-[78]	0.32	0.30	0.26	0.35	0.17	0.36	**0.43**	0.35	0.37	0.36
EA[61]	0.30	0.16	0.22	0.30	0.20	0.35	**0.42**	0.30	0.31	0.33
DA[61]	0.34	0.31	0.29	0.33	0.21	0.38	**0.41**	0.36	0.35	0.38
Psn Avg	0.33	0.28	0.25	0.32	0.22	0.38	**0.41**	0.34	0.35	0.37

Table 31. NovelConcepts10 FID for various poison defenses with SDXL.

Defense → Poison ↓	Nominal	Regen [84]	PDMPure [77]	AdvClean [65]	JPEG	T600	LM	JPEG +T600	JPEG +LM	SZT
Clean	325	335	374	333	374	303	**293**	332	319	322
ADM+[39]	314	340	383	342	358	319	**311**	326	327	329
ADM-	341	351	382	330	370	**307**	316	347	314	310
SDS+[78]	317	313	370	332	342	321	312	317	314	**298**
SDS-[78]	336	322	371	**321**	386	325	297	323	322	332
EA[61]	337	421	387	334	376	328	**303**	337	344	321
DA[61]	325	334	356	331	378	325	319	318	324	**313**
Psn Avg	328	347	375	332	368	321	**310**	328	324	317

Table 32. NovelConcepts10 DINOv2 Similarity for various poison defenses with LoRA.

Defense → Poison ↓	Nominal	Regen [84]	PDMPure [77]	AdvClean [65]	JPEG	T600	LM	JPEG +T600	JPEG +LM	SZT
Clean	0.35	0.35	0.25	0.34	0.35	**0.47**	0.39	0.42	0.38	0.44
ADM+[39]	0.26	0.26	0.28	0.28	**0.35**	0.28	0.31	**0.35**	0.31	**0.35**
ADM-	0.33	0.29	0.29	0.35	0.32	0.28	0.32	0.40	0.36	**0.42**
SDS+[78]	0.32	0.26	0.31	0.29	0.30	0.28	0.28	0.31	0.32	**0.34**
SDS-[78]	0.31	0.27	0.28	0.35	0.32	0.26	0.31	**0.41**	0.35	0.38
EA[61]	0.11	0.13	0.30	0.18	0.29	0.09	0.13	0.35	0.31	**0.36**
DA[61]	0.29	0.27	0.27	0.32	0.32	0.24	0.28	**0.40**	0.34	0.38
Psn Avg	0.27	0.25	0.29	0.30	0.32	0.24	0.27	**0.37**	0.33	**0.37**

Table 33. NovelConcepts10 FID for various poison defenses with LoRA.

Defense → Poison ↓	Nominal	Regen [84]	PDMPure [77]	AdvClean [65]	JPEG	T600	LM	JPEG +T600	JPEG +LM	SZT
Clean	322	311	365	316	315	**261**	301	284	303	280
ADM+[39]	364	363	357	367	312	383	360	**310**	337	325
ADM-	328	347	360	318	319	345	350	**284**	316	287
SDS+[78]	335	355	348	346	339	375	363	340	330	**323**
SDS-[78]	338	350	357	318	320	355	345	**282**	317	298
EA[61]	435	433	347	414	351	432	425	328	339	**318**
DA[61]	348	356	365	342	323	388	352	**292**	323	311
Psn Avg	358	367	356	351	327	380	366	**306**	327	310

Table 34. NovelConcepts10 CLIP Score for various poison defenses with LoRA.

Defense → Poison ↓	Nominal	Regen [84]	PDMPure [77]	AdvClean [65]	JPEG	T600	LM	JPEG +T600	JPEG +LM	SZT
Clean	0.48	0.48	**0.51**	0.48	0.48	0.41	0.44	0.44	0.45	0.42
ADM+[39]	0.49	0.49	**0.50**	**0.50**	0.48	0.46	0.45	0.45	0.48	0.45
ADM-	0.47	0.48	0.48	0.47	**0.49**	0.42	0.46	0.43	0.46	0.43
SDS+[78]	0.48	0.49	0.47	**0.50**	0.49	0.45	0.47	0.46	0.48	0.45
SDS-[78]	0.48	0.46	0.47	0.48	**0.49**	0.44	0.45	0.44	0.46	0.44
EA[61]	0.42	0.42	0.46	0.43	**0.48**	0.41	0.43	0.44	0.47	0.45
DA[61]	0.47	0.48	0.48	0.48	**0.49**	0.44	0.47	0.44	0.47	0.45
Psn Avg	0.47	0.47	**0.48**	**0.48**	**0.48**	0.44	0.45	0.44	0.47	0.44

Table 35. NovelConcepts10 DINOv2 Similarity for various poison defenses with CustomDiffusion (no crop/rescale augmentation).

Defense → Poison ↓	Nominal	Regen [84]	PDMPure [77]	AdvClean [65]	JPEG	T600	LM	JPEG +T600	JPEG +LM	SZT
Clean	0.30	0.31	0.27	0.30	0.30	**0.37**	0.30	0.34	0.32	**0.37**
ADM+[39]	0.26	0.28	0.29	0.28	0.30	0.27	0.31	0.33	0.34	**0.35**
ADM-	0.29	0.25	0.27	0.30	0.31	0.20	0.29	0.32	0.32	**0.33**
SDS+[78]	0.25	0.27	0.28	0.26	0.29	0.26	0.29	0.33	0.33	**0.35**
SDS-[78]	0.27	0.25	0.28	0.30	0.32	0.19	0.29	0.32	0.34	**0.38**
EA[61]	0.14	0.15	0.24	0.19	0.29	0.13	0.19	0.30	**0.33**	**0.33**
DA[61]	0.28	0.25	0.27	0.30	0.31	0.23	0.31	0.29	**0.34**	**0.34**
Psn Avg	0.25	0.24	0.27	0.27	0.30	0.21	0.28	0.31	0.33	**0.35**

Table 36. NovelConcepts10 FID for various poison defenses with CustomDiffusion (no crop/rescale augmentation).

Defense → Poison ↓	Nominal	Regen [84]	PDMPure [77]	AdvClean [65]	JPEG	T600	LM	JPEG +T600	JPEG +LM	SZT
Clean	336	333	346	331	327	**305**	335	316	324	307
ADM+[39]	347	346	342	339	331	347	330	322	311	**306**
ADM-	335	352	349	338	334	374	338	332	324	**320**
SDS+[78]	357	345	341	354	334	350	341	319	325	**309**
SDS-[78]	344	348	344	333	324	383	343	330	314	**302**
EA[61]	404	393	360	382	339	408	389	336	**318**	321
DA[61]	342	351	347	332	333	368	331	339	315	**310**
Psn Avg	355	356	347	346	333	372	345	330	318	**312**

Table 37. NovelConcepts10 CLIP Score for various poison defenses with CustomDiffusion (no crop/rescale augmentation).

Defense → Poison ↓	Nominal	Regen [84]	PDMPure [77]	AdvClean [65]	JPEG	T600	LM	JPEG +T600	JPEG +LM	SZT
Clean	0.37	0.38	0.38	0.38	0.38	0.39	0.38	0.39	**0.40**	0.39
ADM+[39]	**0.40**	0.39	**0.40**	0.39	0.39	0.39	**0.40**	**0.40**	**0.40**	0.39
ADM-	0.38	0.39	0.39	0.38	0.38	**0.41**	0.39	0.38	0.40	0.39
SDS+[78]	**0.40**	0.38	0.38	**0.40**	0.39	**0.40**	0.39	0.39	0.39	**0.40**
SDS-[78]	0.37	0.39	0.38	0.38	0.39	**0.41**	0.40	0.39	0.39	0.40
EA[61]	0.38	0.38	0.38	0.37	0.38	0.39	**0.40**	0.39	0.39	0.39
DA[61]	**0.39**	0.38	0.38	0.38	0.38	**0.39**	**0.39**	**0.39**	0.38	**0.39**
Psn Avg	0.39	0.39	0.38	0.38	0.39	**0.40**	0.39	0.39	0.39	0.39

Table 38. NovelConcepts10 DINOv2 Similarity for various poison defenses with CustomDiffusion (crop/rescale augmentation).

Defense → Poison ↓	Nominal	Regen [84]	PDMPure [77]	AdvClean [65]	JPEG	T600	LM	JPEG +T600	JPEG +LM	SZT
Clean	0.35	0.35	0.33	**0.36**	0.33	**0.36**	0.34	**0.36**	0.34	0.32
ADM+[39]	0.32	0.30	0.33	**0.34**	0.31	**0.34**	0.33	0.33	0.33	0.30
ADM-	0.33	0.25	0.33	0.33	0.33	0.33	0.31	**0.34**	**0.34**	0.30
SDS+[78]	0.33	0.31	**0.34**	0.31	0.32	0.32	0.32	**0.34**	**0.34**	0.30
SDS-[78]	0.32	0.26	0.32	**0.34**	0.33	0.31	0.31	**0.34**	0.32	0.30
EA[61]	0.31	0.21	0.29	0.31	0.31	0.29	0.32	0.32	**0.33**	0.30
DA[61]	**0.36**	0.33	0.32	0.32	0.32	0.34	0.33	0.35	0.33	0.30
Psn Avg	0.33	0.28	0.32	0.32	0.32	0.32	0.32	**0.34**	0.33	0.30

Table 39. NovelConcepts10 FID for various poison defenses with CustomDiffusion (crop/rescale augmentation).

Defense → Poison ↓	Nominal	Regen [84]	PDMPure [77]	AdvClean [65]	JPEG	T600	LM	JPEG +T600	JPEG +LM	SZT
Clean	316	**313**	321	315	323	314	320	**313**	320	325
ADM+[39]	329	330	325	**319**	326	321	323	324	322	333
ADM-	324	353	320	327	**320**	327	327	321	321	334
SDS+[78]	323	331	319	329	325	328	326	320	**316**	328
SDS-[78]	324	357	326	320	**317**	338	329	324	327	337
EA[61]	332	375	336	335	331	349	324	331	**323**	337
DA[61]	**311**	326	329	327	327	320	325	321	324	332
Psn Avg	324	345	326	326	325	330	326	324	**322**	333

Table 40. NovelConcepts10 CLIP Score for various poison defenses with CustomDiffusion (crop/rescale augmentation).

Defense → Poison ↓	Nominal	Regen [84]	PDMPure [77]	AdvClean [65]	JPEG	T600	LM	JPEG +T600	JPEG +LM	SZT
Clean	0.41	0.40	**0.42**	0.41	0.40	0.40	**0.42**	0.41	**0.42**	0.40
ADM+[39]	0.41	0.41	0.41	0.40	0.41	0.41	0.42	0.41	**0.43**	0.41
ADM-	0.42	0.42	0.41	0.40	0.41	0.40	0.42	0.39	**0.43**	0.42
SDS+[78]	0.42	**0.43**	0.42	0.40	**0.43**	0.41	0.42	0.40	0.42	0.42
SDS-[78]	**0.42**	0.40	0.41	0.41	0.41	**0.42**	0.41	0.41	**0.42**	0.41
EA[61]	0.41	0.40	**0.42**	0.41	0.40	0.41	0.40	0.40	**0.42**	0.41
DA[61]	0.41	**0.42**	**0.42**	0.40	**0.42**	0.41	**0.42**	0.40	**0.42**	0.40
Psn Avg	0.41	0.41	**0.42**	0.40	0.41	0.41	**0.42**	0.40	**0.42**	0.41

Q. Ethical Statement

As our research primarily concerns the subfield of data poisoning, we are keenly aware of our work's ethical proximity to copyright theft and artistic style copying. Practical applications of SZT (and related methods) may realize as improved attacks against attempts by copyright holders and artists to protect their works. Even so, we believe that our research is necessary. The individual components of SZT are not complex, simply relying on JPEG compression, biased timestep sampling, and loss masking. Rather, we believe that the effectiveness of SZT, despite its simplicity, exposes the vulnerabilities of existing poisons and demands further research on robust poisons.

R. Contributions

Name order generally denotes share of contribution.
Dataset Collection: Styborski, Kapur, Lu
Poison Curation: Styborski, Kapur
Defense Baselines: Lu, Kapur, Styborski
Semantic Sensitivity Map Studies: Lyu
Timestep Learning Bias Studies: Styborski
Spatial Learning Bias Studies: Lyu, Styborski
JPEG Compression Studies: Styborski
Hyperparameter Ablation: Styborski
Model Ablation: Lyu, Styborski, Lu
Additional Personalization Methods: Styborski
Writing: Styborski, Lyu, Kong
Editing: Kong, Styborski, Lyu
Advising and Guidance: Kong