Generating Potent Poisons and Backdoors from Scratch with Guided Diffusion

Hossein Souri [1]   Arpit Bansal [2]   Hamid Kazemi [2]   Liam Fowl [3]   Aniruddha Saha [2]   Jonas Geiping [4]
Andrew Gordon Wilson [5]   Rama Chellappa [1]   Tom Goldstein [2]   Micah Goldblum [5]

Abstract

Modern neural networks are often trained on massive datasets that are web scraped with minimal human inspection. As a result of this insecure curation pipeline, an adversary can poison or backdoor the resulting model by uploading malicious data to the internet and waiting for a victim to scrape and train on it. Existing approaches for creating poisons and backdoors start with randomly sampled clean data, called base samples, and then modify those samples to craft poisons. However, some base samples may be significantly more amenable to poisoning than others. As a result, we may be able to craft more potent poisons by carefully choosing the base samples. In this work, we use guided diffusion to synthesize base samples from scratch that lead to significantly more potent poisons and backdoors than previous state-of-the-art attacks. Our Guided Diffusion Poisoning (GDP) base samples can be combined with any downstream poisoning or backdoor attack to boost its effectiveness. Our implementation code is publicly available at: https://github.com/hsouri/GDP.

1. Introduction

Large-scale neural networks have seen rapid improvement in many domains over recent years, enabled by web-scale training sets. These massive datasets are collected using automated curation pipelines with little to no human oversight. Such automated pipelines are vulnerable to data tampering attacks in which malicious actors upload harmful samples to the internet that implant security vulnerabilities in models trained on them, in the hopes that a victim scrapes the harmful samples and incorporates them in their training set. For example, *targeted data poisoning attacks* cause the victim model to misclassify specific test samples (Shafahi et al., 2018; Geiping et al., 2020), while *backdoor attacks* manipulate victim models so that they only misclassify test samples when the samples contain a specific backdoor trigger (Gu et al., 2017; Turner et al., 2018; Saha et al., 2020; Souri et al., 2022).

Typical data poisoning and backdoor attacks begin with randomly selected *base samples* from clean data, which the attacker perturbs to minimize a poisoning objective. The attacker often constrains the perturbations to be small so that the resulting poisons look similar to the original base samples and still appear correctly labeled (Geiping et al., 2020; Souri et al., 2022). However, a real world attacker may not be constrained around a particular set of randomly chosen base samples. Indeed, Souri et al. (2022) show that some base samples are far more effective for crafting backdoor poisons than others. Namely, Souri et al. (2022) select base samples which achieve a high gradient norm, $\|\nabla_\theta \ell(\theta)\|_2$, where $\ell(\theta)$ denotes the training loss of a classifier with parameters θ. Such a selection strategy enables more potent poisons, but the base samples are still limited to clean samples chosen from a limited dataset.

In this work, we use diffusion models to synthesize base samples from scratch that enable especially potent attacks. Rather than filtering out existing natural data, synthesizing the samples from scratch allows us to optimize them specifically for the poisoning objective. By weakly guiding the generative diffusion process using a poisoning objective, we craft images that are simultaneously near potent poisons while also looking precisely like natural images of the base class. Then, we can use existing poisoning and backdoor attack algorithms on top of our diffusion-generated base poisons to amplify the effectiveness of the downstream attacks, surpassing previous state-of-the-art for both targeted data poisoning and backdoor attacks. See Figure 1 for a schematic of our method and Figures 2, 3, 5, 6 (see Appendix) for corresponding example base images.

We use our Guided Diffusion Poisoning (GDP) base samples in combination with both targeted data poisoning and backdoor attacks (Geiping et al., 2020; Souri et al., 2022). In all cases, we boost the performance of existing state-of-
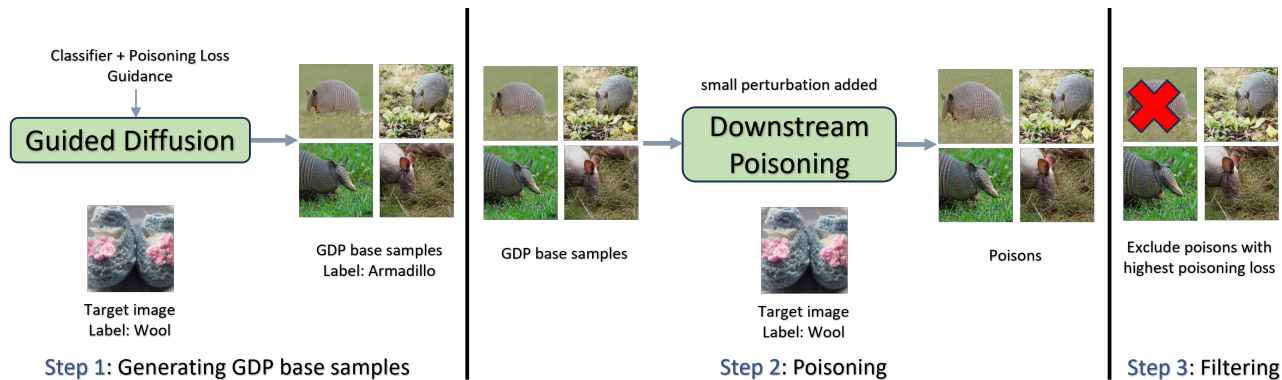
Figure 1: **Schematic of Guided Diffusion Poisoning (GDP).** GDP contains three stages: **(1)** generate base samples with a diffusion model weakly guided using a poisoning loss; **(2)** use the base samples as initialization for a downstream poisoning algorithm; **(3)** select poisons with the lowest poisoning loss and include them in the poisoned training set.

the-art attacks by a wide margin while preserving image quality and crucially without producing mislabeled poisons, measured through a human study. These resulting attacks bypass all 7 defenses we test, and they are successful in the hard black-box setting. Our approach also enables successful attacks with smaller perturbation budgets on the downstream poisoning algorithm and inserting far fewer poison images into the victim's training set than previous state-of-the-art attacks.

Our contributions are summarized as follows:

- We devise a method, GDP, for synthesizing clean-label poisoned training data from scratch with guided diffusion.

- Our approach achieves far higher success rates than previous state-of-the-art targeted poisoning and backdoor attacks, including in scenarios where the attacker is only allowed to poison a very small proportion of training samples.

- We show how GDP can be used to significantly boost the performance of previous data poisoning and backdoor attack algorithms.

2. Background

2.1. Poisoning Setup

Poisoning deep networks trained from scratch is a problem that cannot be solved exactly. This is due to the bi-level nature of the optimization problem, where the attacker tries to find perturbations $\delta = \{\delta_i\}_{i=1}^N$ constrained by conditions $\mathcal{C}$ (usually an ℓ_∞ bound) to minimize an adversarial objective $\mathcal{A}$. Formally stated, the attacker solves the following optimization problem:

$$\min_{\delta \in \mathcal{C}} \mathcal{A}(f_{\theta_*}) \quad \text{s.t.}$$

$$\theta_* \in \text{argmin}_\theta \left[\frac{1}{|\mathcal{T}|} \sum_{i=1}^{|\mathcal{T}|} \mathcal{L}(f_\theta(x_i + \delta_i), y_i) \right],$$

where f_{θ_*} denotes a victim model which itself is trained on the poisons included in the victim's training set, $\mathcal{T}$. Note that $\delta_i = \vec{0}$ for any $i > N$ (included for simplicity of presentation).

Because of the complexity of this problem, we need to use approximations to solve it. The current gold standard for clean-label poisoning for both targeted and backdoor attacks generally involves *gradient alignment* (Geiping et al., 2020; Souri et al., 2022). Gradient alignment was introduced as a method to poison victim models in Geiping et al. (2020), where, in plain words, the attacker imperceptibly modifies poisons so that the gradient of a surrogate model evaluated at these poisons *aligns* with an *adversarial gradient*. In Geiping et al. (2020), this adversarial gradient was generated to cause targeted misclassification of a particular image. However, the strategy also generalizes to other poisoning objectives, like a hidden trigger backdoor attack (Souri et al., 2022). More formally, let f_θ be a surrogate model available to the attacker, $P = \{x_i, y_i\}_{i=1}^N$ be a set of data samples with ground-truth labels which the attacker poisons, and $\mathcal{L}$ be the standard categorical cross-entropy loss, then poisons are crafted by minimizing the alignment objective $\mathcal{O}$ (over the perturbations $\{\delta_i\}$):

$$\mathcal{O} = 1 - \frac{1}{N} \sum_{i=1}^N \frac{\langle \nabla_\theta \mathcal{L}(f_\theta(x_i + \delta_i), y_i), \nabla_\theta \mathcal{A} \rangle}{||\nabla_\theta \mathcal{L}(f_\theta(x_i + \delta_i), y_i)|| \cdot ||\nabla_\theta \mathcal{A}||}. \quad (1)$$

In the case of targeted poisoning, the attacker wishes to induce misclassification of a target image x' into an incorrect class y', so they set

$$\mathcal{A} = \mathcal{L}(f_\theta(x', y')).$$

2

For backdoor attacks, the attacker aims for *any* image which has trigger t applied to be misclassified with label y', so, they choose an adversarial objective

$$\mathcal{A} = \mathbb{E}_{x \sim D} \mathcal{L}(f_\theta(x + t), y'),$$

where the expectation is approximated over a handful of samples available to the adversary. In practice, the label y' is usually called the *target class*, and the perturbations are usually applied to the *source class*. As an example, an attacker may modify images of dogs imperceptibly so that, when trained on, these dogs cause a victim network to correctly classify a clean dog at inference time, but misclassify that same dog when a trigger patch is applied.

2.2. Universal Guidance

Diffusion models have become pivotal in generative modeling, especially for image generation. They operate through a T-step process involving both forward and reverse phases. The forward phase incrementally infuses Gaussian noise into an original data point x_0, as described by the equation:

$$x_t = \sqrt{\alpha_t} x_0 + \sqrt{1 - \alpha_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}), \quad (2)$$

where ϵ represents a standard normal random variable, and α_t are predefined noise scales.

Conversely, the reverse phase methodically removes noise, aiming to retrieve x_0. This is achieved via a denoising network ϵ_θ, trained to estimate the noise ϵ in x_t at any step t:

$$\epsilon_\theta(x_t, t) \approx \epsilon = \frac{x_t - \sqrt{\alpha_t} x_0}{\sqrt{1 - \alpha_t}}. \quad (3)$$

The Denoising Diffusion Implicit Model (DDIM) (Song et al., 2021) is a prominent reverse process method. It starts by estimating a clean data point $\hat{x}_0$:

$$\hat{x}_0 = \frac{x_t - \sqrt{1 - \alpha_t}\epsilon_\theta(x_t, t)}{\sqrt{\alpha_t}}. \quad (4)$$

Then, x_{t-1} is sampled from $q(x_{t-1}|x_t, \hat{x}_0)$, replacing x_0 with $\hat{x}_0$ in the sampling formula:

$$\hat{x}_{t-1} = \sqrt{\alpha_t}\hat{x}_0 + \sqrt{1 - \alpha_t}\epsilon_\theta(x_t, t). \quad (5)$$

The universal guidance algorithm (Bansal et al., 2023) leverages $\hat{x}_0$ to create a guidance signal. It includes forward and backward guidance; the latter modifies $\hat{x}_0$, while the former adapts classifier guidance to suit any general guidance function by calculating the gradient relative to x_t.

In forward guidance, an external function f and a loss function ℓ guide image generation. This process is encapsulated by:

$$\hat{\epsilon}_\theta(x_t, t) = \epsilon_\theta(x_t, t) + s(t) \cdot \nabla_{x_t} \ell(c, f(\hat{x}_0)), \quad (6)$$

where $s(t)$ adjusts the guidance strength at each step. This approach ensures images align with the guidance while maintaining a trajectory within the data manifold. Additionally, the paper introduces universal stepwise refinement, a technique that repeats steps to align gradients, enhancing guidance and image fidelity.

For a detailed review of related work, refer to Appendix A.

3. Method

We weakly guide diffusion models with a gradient-matching poisoning objective to create base samples that lie very close to potent poisons, but simultaneously preserve image quality and appear similar to samples drawn from the training distribution. The latter properties ensure that the resulting poisons will be difficult to detect via visual inspection. Diffusion-generated base samples must also be *clean-label* meaning that their assigned labels are aligned with the semantic content of the images. We then use these base samples as an initialization for existing data poisoning and backdoor attack algorithms to boost their effectiveness. We now detail our simple three-step process for generating base samples and using them for downstream poison and backdoor generation:

(1) Generating base samples with guided diffusion: We use a diffusion model to generate base samples, leveraging a pretrained classifier. On each step of diffusion, we compute a cross-entropy loss using the probability the classifier assigns to the noisy image iterate being in the poison class. We include this classification loss in the universal guidance algorithm. For ImageNet (Deng et al., 2009) experiments, we use the classifier-guided diffusion method proposed in Dhariwal & Nichol (2021) to generate base samples. Guidance increases the extent to which the generated image looks like the poison class according to the classifier. We can adjust the effect of class-conditioning on the generated samples by scaling the guidance strength. In practice, our procedure will reliably produce clean-label base samples (see Section 4.5).

In addition to the classifier loss, we also add a data poisoning or backdoor loss function in the universal guidance algorithm when generating base samples, but with a very low guidance strength to preserve image quality and prevent the generated images from moving into the target class. In our experiments, we specifically use the Witches' Brew (Geiping et al., 2020) and Sleeper Agent (Souri et al., 2022) gradient matching loss functions for targeted data poisoning and backdoor attacks, respectively. We modify the gradient matching objectives from both of these works to optimize per-sample gradient matching instead of matching the average gradient computed over a batch of images. We find that this modification improves image quality and yields effective base samples in practice.

Table 1: **Targeted data poisoning. GDP achieves a far higher success rate than existing targeted data poisoning attacks, even with only a small budget**. Experiments are conducted on CIFAR-10 with ResNet-18 models. Perturbations are bounded by ℓ_∞-norm $16/255$. Poison budget refers to the number of images poisoned. We see that injection of only 25-50 poisoned samples is enough for the attack to be effective.

Poison budget	Poison Frogs	Bullseye	Witches' Brew	GDP (ours)
25 images (0.05%)	0%	0%	0%	**30%**
50 images (0.1%)	0%	0%	20%	**70%**

Table 2: **Targeted data poisoning. GDP achieves a far higher success rate than Witches' Brew, even with only a small budget**. Experiments are conducted on ImageNet with ResNet-18 models. We see that injection of only 50-100 poisoned samples is enough for the attack to be effective.

Poison budget	Perturbation ℓ_∞-norm	Witches' Brew	GDP (ours)
50 images ($\sim 0.004\%$)	8/255	0%	**40%**
100 images ($\sim 0.008\%$)	8/255	10%	**50%**
50 images ($\sim 0.004\%$)	16/255	0%	**50%**
100 images ($\sim 0.008\%$)	16/255	20%	**80%**

Table 3: **Backdoor attacks. GDP achieves a far higher success rate than existing backdoor attacks, even with only a small budget**. Experiments are conducted on CIFAR-10 with ResNet-18 models. Perturbations are bounded in ℓ_∞-norm by $16/255$. We see that injection of only 25-50 poisoned samples is enough for the attack to be effective.

Poison budget	CLBA	HTBA	Sleeper Agent	GDP (ours)
25 images (0.05%)	0.85%	1.09%	13.44%	**35.28%**
50 images (0.1%)	0.98%	1.27%	29.11%	**47.71%**

(2) Initializing poisoning and backdoor attacks with GDP base samples: After generating base samples, we use them as an initialization for state-of-the-art targeted data poisoning and backdoor attacks. Since these attacks typically constrain their perturbations via ℓ_∞ norm in order to ensure that the images are high quality and preserve the poison label, we constrain the attacks accordingly around the initial base samples. The validity of this procedure depends on the quality and clean-label of the generated base samples, which we will verify. We observe that this procedure leads to poisons which are much more potent and result in stronger attacks than ones which begin with randomly selected clean images.

(3) Filtering poisons: Given the randomness in diffusion modeling, some base samples do not result in effective poisons. After generating a collection of poisoned training samples, we filter them by selecting the samples that exhibit the lowest values of the corresponding downstream poisoning or backdoor loss. We select a number of samples equal to the poisoning budget and replace them with randomly selected clean samples from the victim's training set, thereby maintaining the original size of the training set. We exclude and ignore the poisoned samples with high poisoning loss.

Algorithm 1 (see Appendix) provides a summary of the proposed GDP attack. It outlines the key steps of the attack, offering an overview of how it is carried out. A detailed description of the threat model can be found in Appendix D.1. For visualization of the GDP base samples and final perturbed poisons on CIFAR-10 and ImageNet for poisoning and backdoor attacks, please refer to Appendix F.

4. Experimental Evaluations

In this section, we evaluate the proposed poisoning pipeline for poisoning image classification models trained on CIFAR-10 and ImageNet. We perform experiments on both targeted data poisoning, following the threat model of Shafahi et al. (2018), and backdoor attacks, following the threat model of Saha et al. (2020) and Souri et al. (2022) (see Appendix D.1).

Details regarding the experimental setup can be found in Appendix D. Further experimental evaluations are provided in Appendix E.

4.1. Potent Poisons, Even in Small Quantities

As baselines, we compare to the strongest existing targeted data poisoning attack, Witches' Brew (Geiping et al., 2020), along with Poison Frogs (Shafahi et al., 2018) and Bullseye Polytope (Aghakhani et al., 2021) in Table 1. These attacks usually require several hundred poisoned samples to be effective on CIFAR-10; providing only a budget of 25 and 50 poisoned images was previously insufficient to attack the model. However, we find that potent poisons developed with our approach described in Section 3 are far more effective, even though 25 images constitute only 0.05% of the CIFAR-10 training set.

We further find that these poisoning results scale to large-scale experiments on ImageNet, as shown in Table 2. Again modifying only a tiny subset of training images (0.004%-0.008%) is sufficient to poison the model, exceeding the effectiveness of existing approaches by a very wide margin.

This success also extends to backdoor attacks, where we show results in Table 3, noting the effectiveness of the attack compared to a recent state-of-the-art hidden trigger backdoor attack, Sleeper Agent (Souri et al., 2022), as well as Clean-Label Backdoor Attacks (CLBA) (Turner et al., 2018) and Hidden-Trigger Backdoor Attacks (HTBA) (Saha et al., 2020). GDP outperforms all of the above backdoor attacks.

Table 4: **Improving existing targeted data poisoning attacks with GDP base samples.** Experiments are conducted on CIFAR-10 with 6-layer ConvNets, and perturbations are bounded in ℓ_∞-norm by $32/255$. GDP base samples are generated using the corresponding downstream poisoning loss functions.

Poison budget	Poison Frogs	Bullseye	Poison Frogs + GDP	Bullseye + GDP
100 images (0.2%)	10%	20%	**20%**	**40%**
200 images (0.4%)	10%	20%	**30%**	**60%**

Table 5: **Improving existing backdoor attacks with GDP base samples.** Experiments are conducted on CIFAR-10 with ResNet-18 models, and perturbations are bounded in ℓ_∞-norm by $16/255$. GDP base samples are generated using the Sleeper Agent loss function and the victim is fine-tuned on the poisoned training set following Saha et al. (2020).

Poison budget	HTBA	HTBA + GDP
50 images (0.1%)	1.54%	**17.58%**

Backdoor attack evaluations on ImageNet can be found in Appendix E.2.

We can also use our diffusion-generated images as base samples for other existing targeted poisoning and backdoor attacks for significant boosts in effectiveness. In Table 4 and Table 5, we see that GDP base samples massively improve the success rates of Poisons Frogs and Bullseye Polytope targeted data poisoning attacks as well as Hidden-Trigger Backdoor Attacks.

4.2. Not Only Potent, but Also Stealthy

In contrast to previous work, GDP attacks are also successful with smaller ℓ_∞ constraints, as we highlight in Tables 2 and 6 for targeted data poisoning and Table 7 for backdoor attacks. Small perturbation attacks are a crucial application scenario, as higher perturbation budgets, such as $16/255$ are more easily detectable during data inspection, e.g. when the data points are labeled by an annotator.

4.3. Not Only Potent, but Also Transferable

We now investigate how transferable these poisons are in the hard black-box setting, where the architecture of the victim model is unknown to the attacker when crafting poisons. Tables 11 and 12 (see Appendix) highlight the transferability of GDP in targeted poisoning and backdoor settings, respectively. Compared to Witches' Brew and Sleeper Agent, our approach boosts transferability. In Appendix E.3, we present additional experiments with an ensemble of surrogate models.

Table 6: **Small perturbations. GDP enables stronger targeted data poisoning under small ℓ_∞-norm bound perturbations.** Experiments are conducted on CIFAR-10 with ResNet-18 models and poison budget of 50 images (0.1%).

Perturbation ℓ_∞-norm	Witches' Brew	GDP (ours)
$8/255$	0%	**10%**
$12/255$	0%	**40%**

Table 7: **Small perturbations. GDP enables stronger backdoor attacks under small ℓ_∞-norm bound perturbations.** Experiments are conducted on CIFAR-10 with ResNet-18 models and poison budget of 50 images (0.1%).

Perturbation ℓ_∞-norm	Sleeper Agent	GDP (ours)
$8/255$	6.91%	**12.89%**
$12/255$	13.58%	**25.58%**

4.4. Defenses and Mitigation Strategies

During a data poisoning or backdoor attack scenario, the victim may deploy a defense mechanism to either filter out suspected poisons or modify the training routine to alleviate the impact of poisoning. We therefore test our potent poisons against several widely adopted existing defense methods.

Spectral Signatures (Tran et al., 2018) is a representative defense approach that computes the top right singular vector of the covariance matrix of the representation vectors (features) and uses these vectors to compute an outlier score for each input. Inputs that have scores exceeding the outlier threshold are eliminated from the training set. Differentially private training (DP-SGD) (Abadi et al., 2016; Hong et al., 2020) provides protection against data poisoning by adding calibrated noise to the gradients during the training. This defense ensures that the influence of individual training samples is reduced. However, as demonstrated in Geiping et al. (2020) and Souri et al. (2022), this defense mechanism encounters a substantial trade-off between reducing poisoning accuracy and preserving validation accuracy. Recent works suggest that strong data augmentations can be implemented during training to mitigate the poisoning success (Borgnia et al., 2021; Schwarzschild et al., 2021). We evaluate our poisoning and backdoor attacks against mixup (Zhang et al., 2018), one of the most effective data augmentation techniques for countering data poisoning attacks (Borgnia et al., 2021).

Additionally, we test our GDP backdoor attack against STRIP (Gao et al., 2019), Neural Cleanse (Wang et al., 2019), Adversarial Neuron Pruning (ANP) (Wu & Wang,

5

Table 8: **Defenses against targeted data poisoning** on CIFAR-10 with ResNet-18 models. Perturbations have ℓ_∞-norm bounded above by $16/255$ and the attacker can poison 0.1% of training images (50 images).

	Undefended	Spectral Signatures	DP-SGD	Data Augmentation
Avg. Poison Success	70%	50%	50%	40%
Avg. Validation Acc.	92.16%	90.24%	91.34%	91.39%

Table 9: **Defenses against backdoor attacks** on CIFAR-10 with ResNet-18 models. Perturbations have ℓ_∞-norm bounded above by $16/255$ and the attacker can poison 0.1% of training images (50 images).

	Undefended	Spectral Signatures	DP-SGD	Data Augmentation	STRIP	NeuralCleanse	ANP	ABL
Attack Success Rate	47.71%	16.21%	27.65%	24.40%	39.21%	47.71%	29.89%	27.56%
Avg. Validation Acc.	92.15%	90.13%	91.47%	91.66%	91.96%	92.15%	82.52%	89.95%

2021), and Anti-Backdoor Learning (ABL) (Li et al., 2021). STRIP detects incoming backdoor-triggered inputs during testing by deliberately perturbing them and analyzing the entropy of the predicted class distribution. A low entropy suggests the presence of a backdoor input, leading to rejection. Neural Cleanse approximates the backdoor trigger using adversarial perturbations. We leverage this defense mechanism to identify the backdoored class in our attacks through outlier detection. ANP proposes to defend against backdoors by pruning the sensitive neurons through adversarial perturbations to model weights. ABL is another method for mitigating backdoor attacks that identifies suspected training samples with the smallest losses. It then proceeds to unlearn these identified poisoned samples to mitigate the backdoor attack.

Our experimental results, presented in Tables 8 and 9, indicate that our poisoning and backdoor attacks breach these defenses, consistently maintaining high poisoning accuracy. It must be noted that, as demonstrated in Geiping et al. (2020) and Souri et al. (2022), it is possible for these defensive techniques to highly degrade the poisoning accuracy but simultaneously imposing a significant decrease in validation accuracy. Therefore, in our defense experiments, we adjust the corresponding parameters to ensure a sufficiently high validation accuracy.

4.5. Human Evaluation: GDP Poisons Are Clean-Label

To ensure that the diffusion-generated base samples are indeed clean-label, we also conduct a human evaluation. We assemble a dataset of 500 GDP base samples crafted with the Witches' Brew and Sleeper Agent objectives and 500 randomly sampled natural images from the CIFAR-10 dataset, encompassing 50 samples for each class in both sets. We ask annotators to classify these images into the 10 CIFAR-10 classes, and annotators are not told which images are synthetic and which are natural. In Table 10, we present the accuracies observed on these sets across annotators. Our analysis demonstrates that their accuracy

Table 10: **GDP base samples are clean-label according to human evaluators.** Human evaluation, accuracy at performing CIFAR-10 10-class classification on real and GDP base samples crafted with the Witches' Brew and Sleeper Agent objectives. Humans perform at least as well classifying GDP base samples as they do on clean CIFAR-10 training samples.

	Natural images	GDP base samples
Classification accuracy	93%	96.8%

on GDP base samples is comparable to or even exceeds their accuracy on the natural images, indicating that the synthesized base samples maintain their assigned label as desired.

5. Conclusion

In this work, we showed that the base samples used for poisoning have a very strong impact on the effectiveness of the resulting poisons. With this principle in mind, we synthesize base samples from scratch specifically so that they lie near potent poisons. Our guided diffusion approach amplifies the effects of state-of-the-art targeted data poisoning and backdoor attacks across multiple datasets.

Acknowledgements

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Aghakhani, H., Meng, D., Wang, Y.-X., Kruegel, C., and Vigna, G. Bullseye polytope: A scalable clean-label poisoning attack with improved transferability. In *2021 IEEE European symposium on security and privacy (EuroS&P)*, pp. 159–178. IEEE, 2021.

Bansal, A., Borgnia, E., Chu, H.-M., Li, J. S., Kazemi, H., Huang, F., Goldblum, M., Geiping, J., and Goldstein, T. Cold diffusion: Inverting arbitrary image transforms without noise. *arXiv preprint arXiv:2208.09392*, 2022.

Bansal, A., Chu, H.-M., Schwarzschild, A., Sengupta, S., Goldblum, M., Geiping, J., and Goldstein, T. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 843–852, 2023.

Biggio, B., Nelson, B., and Laskov, P. Poisoning attacks against support vector machines. *arXiv preprint arXiv:1206.6389*, 2012.

Borgnia, E., Geiping, J., Cherepanova, V., Fowl, L., Gupta, A., Ghiasi, A., Huang, F., Goldblum, M., and Goldstein, T. Dp-instahide: Provably defusing poisoning and backdoor attacks with differentially private data augmentations. *arXiv preprint arXiv:2103.02079*, 2021.

Chen, X., Liu, C., Li, B., Lu, K., and Song, D. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

Chung, H., Kim, J., Mccann, M. T., Klasky, M. L., and Ye, J. C. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022a.

Chung, H., Sim, B., Ryu, D., and Ye, J. C. Improving diffusion models for inverse problems using manifold constraints. *arXiv preprint arXiv:2206.00941*, 2022b.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Fowl, L., Goldblum, M., Chiang, P.-y., Geiping, J., Czaja, W., and Goldstein, T. Adversarial examples make strong poisons. *Advances in Neural Information Processing Systems*, 34:30339–30351, 2021.

Gao, Y., Xu, C., Wang, D., Chen, S., Ranasinghe, D. C., and Nepal, S. Strip: A defence against trojan attacks on deep neural networks. In *Proceedings of the 35th Annual Computer Security Applications Conference*, pp. 113–125, 2019.

Geiping, J., Fowl, L. H., Huang, W. R., Czaja, W., Taylor, G., Moeller, M., and Goldstein, T. Witches' brew: Industrial scale data poisoning via gradient matching. In *International Conference on Learning Representations*, 2020.

Goldblum, M., Tsipras, D., Xie, C., Chen, X., Schwarzschild, A., Song, D., Mądry, A., Li, B., and Goldstein, T. Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45 (2):1563–1580, 2022.

Graikos, A., Malkin, N., Jojic, N., and Samaras, D. Diffusion models as plug-and-play priors. *arXiv preprint arXiv:2206.09012*, 2022.

Gu, T., Dolan-Gavitt, B., and Garg, S. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 32, 2020.

Hong, S., Chandrasekaran, V., Kaya, Y., Dumitraş, T., and Papernot, N. On the effectiveness of mitigating data poisoning attacks with gradient shaping. *arXiv preprint arXiv:2002.11497*, 2020.

Huang, H., Ma, X., Erfani, S. M., Bailey, J., and Wang, Y. Unlearnable examples: Making personal data unexploitable. *arXiv preprint arXiv:2101.04898*, 2021.

Huang, W. R., Geiping, J., Fowl, L., Taylor, G., and Goldstein, T. Metapoison: Practical general-purpose clean-label data poisoning. *Advances in Neural Information Processing Systems*, 33:12080–12091, 2020.

Kawar, B., Elad, M., Ermon, S., and Song, J. Denoising diffusion restoration models. *arXiv preprint arXiv:2201.11793*, 2022.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Li, Y., Lyu, X., Koren, N., Lyu, L., Li, B., and Ma, X. Anti-backdoor learning: Training clean models on poisoned data. *Advances in Neural Information Processing Systems*, 34:14900–14912, 2021.

Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., and Lee, Y. J. Gligen: Open-set grounded text-to-image generation. *arXiv preprint arXiv:2301.07093*, 2023.

Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., and Van Gool, L. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11461–11471, 2022.

Muñoz-González, L., Biggio, B., Demontis, A., Paudice, A., Wongrassamee, V., Lupu, E. C., and Roli, F. Towards poisoning of deep learning algorithms with back-gradient optimization. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pp. 27–38, 2017.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Saha, A., Subramanya, A., and Pirsiavash, H. Hidden trigger backdoor attacks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 11957–11965, 2020.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

Schwarzschild, A., Goldblum, M., Gupta, A., Dickerson, J. P., and Goldstein, T. Just how toxic is data poisoning? a unified benchmark for backdoor and data poisoning attacks. In *International Conference on Machine Learning*, pp. 9389–9398. PMLR, 2021.

Shafahi, A., Huang, W. R., Najibi, M., Suciu, O., Studer, C., Dumitras, T., and Goldstein, T. Poison frogs! targeted clean-label poisoning attacks on neural networks. *Advances in neural information processing systems*, 31, 2018.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. *International Conference on Learning Representations*, 2021.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

Souri, H., Fowl, L., Chellappa, R., Goldblum, M., and Goldstein, T. Sleeper agent: Scalable hidden trigger backdoors for neural networks trained from scratch. *Advances in Neural Information Processing Systems*, 35: 19165–19178, 2022.

Steinhardt, J., Koh, P. W. W., and Liang, P. S. Certified defenses for data poisoning attacks. *Advances in neural information processing systems*, 30, 2017.

Tran, B., Li, J., and Madry, A. Spectral signatures in backdoor attacks. *Advances in neural information processing systems*, 31, 2018.

Turner, A., Tsipras, D., and Madry, A. Clean-label backdoor attacks. 2018.

Wang, B., Yao, Y., Shan, S., Li, H., Viswanath, B., Zheng, H., and Zhao, B. Y. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 707–723. IEEE, 2019.

Wang, W., Bao, J., Zhou, W., Chen, D., Chen, D., Yuan, L., and Li, H. Semantic image synthesis via diffusion models. *arXiv preprint arXiv:2207.00050*, 2022a.

Wang, Y., Yu, J., and Zhang, J. Zero-shot image restoration using denoising diffusion null-space model. *arXiv preprint arXiv:2212.00490*, 2022b.

Whang, J., Delbracio, M., Talebi, H., Saharia, C., Dimakis, A. G., and Milanfar, P. Deblurring via stochastic refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16293–16303, 2022.

Wu, D. and Wang, Y. Adversarial neuron pruning purifies backdoored deep models. *Advances in Neural Information Processing Systems*, 34:16913–16925, 2021.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

Zhang, L. and Agrawala, M. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023.

Zhu, C., Huang, W. R., Li, H., Taylor, G., Studer, C., and Goldstein, T. Transferable clean-label poisoning attacks on deep neural nets. In *International Conference on Machine Learning*, pp. 7614–7623. PMLR, 2019.

A. Related Work

A.1. Data Poisoning and Backdoor Attacks

Data poisoning attacks can be roughly grouped by their threat model and the adversary's objective. Early poisoning attacks demonstrated the vulnerability of simple models, like linear classifiers and logistic regression, to malicious training data modifications (Biggio et al., 2012; Muñoz-González et al., 2017; Steinhardt et al., 2017; Goldblum et al., 2022). These early poisoning attacks often focused on degrading the overall accuracy of the victim model, and occasionally employed label flips in addition to data modifications. Such an attack is *dirty* label, as it assumes the attacker has some control over the victim's labeling scheme. In contrast, *clean* label attacks do not assume control over the victim's labeling method and usually modify data in a visually minimal way - maintaining their original semantic label. This objective (degrading overall victim accuracy) is sometimes referred to as *indiscriminate* poisoning or an *availability* attack.

Newer availability poisoning attacks have demonstrated the ability to degrade the accuracy of modern deep networks, but often require modifying a large proportion of the training data (Huang et al., 2021; Fowl et al., 2021). These attacks often rely on creating "shortcuts" for the victim model to minimize training loss without learning meaningful features of the clean data distribution.

In contrast to availability attacks, *targeted* poisoning attacks aim to cause a victim model to misclassify a particular target data point, or a small set of target data points, at inference time. Several works successfully poisoned deep neural networks in a transfer learning setting wherein a backbone is frozen, and a linear layer is retrained on top of the feature extractor (Shafahi et al., 2018; Zhu et al., 2019; Aghakhani et al., 2021). Newer, more powerful attacks successfully poison victim models trained from *scratch* by using ℓ_∞ bounded perturbations (Geiping et al., 2020; Huang et al., 2020). However, these attacks often require poisoning a non-negligible amount of training data - for example, Geiping et al. (2020) requires several hundred poisons to reliably cause misclassification of a single target CIFAR-10 (Krizhevsky et al., 2009) image at a relatively large $\varepsilon = 16/255$ perturbation bound (Geiping et al., 2020).

Instead of fixing particular samples to target with poisoning attacks, another class of attacks, known as *backdoor* attacks, aim to poison a victim so that *any* sample, with the addition of a trigger, will be misclassified at inference time. Triggers can be additive colorful patches, or small pixel modifications. Early attacks showed backdoor vulnerability in several settings, including transfer learning, and, like their availability attack counterparts, often employed label flips as an attack tool (Gu et al., 2017; Chen et al., 2017).

More recent backdoor attacks have advanced the capability of backdoor attacks that do not rely on any label flips. Such attacks have proven successful in transfer learning settings (Saha et al., 2020), as well as from-scratch settings (Souri et al., 2022). Recent clean-label attacks can operate with a hidden trigger - a trigger which is not present in any of the poisoned training data, but still effectively induces targeted misclassification when applied at inference time. However, it is worth noting that existing backdoor attack success deteriorates as the number of poisons reaches below 100. For example, Souri et al. (2022) achieve just over 10% attack success rate when 25 poisons are included in the victim's training data.

This class of clean-label, hidden trigger backdoor attacks, along with similar targeted poisoning attacks, can be quite pernicious as they generally require a smaller proportion of data to be modified by the attacker (compared to availability attacks), and are harder to detect by hand inspection of the model's performance on holdout sets. Thus, in this work, we focus on pushing the limits of targeted and backdoor attacks in the *low poison budget* regime wherein a victim might only scrape a handful of samples poisoned by a malicious attacker.

A.2. Guidance in Diffusion Models

Diffusion models (Song & Ermon, 2019; Ho et al., 2020) in machine learning have seen a remarkable evolution, particularly in their application to image generation tasks. These models, which simulate the process of transforming a random noise distribution into a specific data distribution, have incorporated various guidance mechanisms to direct this transformation process more precisely. The concept of condition (Ho & Salimans, 2022; Bansal et al., 2022; Nichol et al., 2021; Whang et al., 2022; Wang et al., 2022a; Li et al., 2023; Zhang & Agrawala, 2023) or guidance (Dhariwal & Nichol, 2021; Kawar et al., 2022; Wang et al., 2022b; Chung et al., 2022a; Lugmayr et al., 2022; Chung et al., 2022b; Graikos et al., 2022; Bansal et al., 2023) in diffusion models is crucial for achieving outputs that adhere to specific characteristics or criteria, a necessity in applications demanding high precision.

Initially, guidance within diffusion models was predominantly achieved through two methods: classifier guidance (Dhariwal & Nichol, 2021) and classifier-free guidance (Ho & Salimans, 2022). Classifier guidance (Dhariwal & Nichol, 2021) involves training a separate classifier, adept at handling noisy image inputs. This classifier generates a guidance signal during the diffusion process, steering the generative model toward desired outcomes. However, this method necessitates the training of a specialized classifier, often a resource-intensive task. In contrast, classifier-free guidance (Ho & Salimans, 2022) internalizes the guidance mecha-
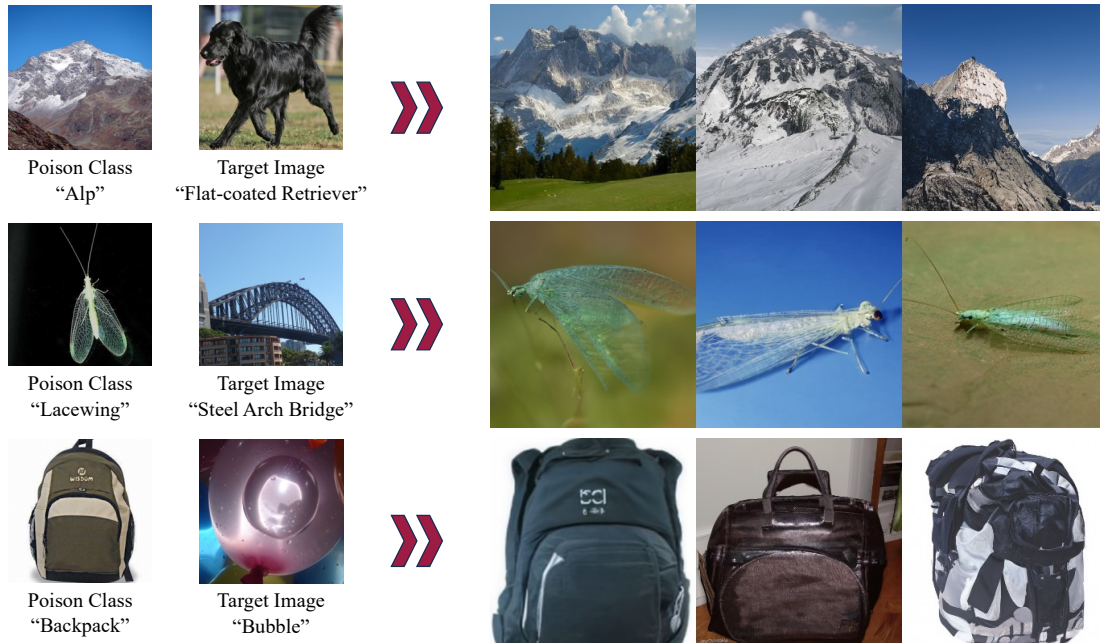
Figure 2: **GDP base samples are clean-label and high quality (ImageNet).** In each panel, the leftmost column contains a random sample from the poison class, the second column contains the target image, and the subsequent three columns contain GDP base samples. Experiments conducted using the Witches' Brew gradient-matching objective with a ResNet-18 model on ImageNet over randomly sampled poison class and target image pairs. Additional visualizations are found in Appendix F.

nism within the model's architecture. This method, while eliminating the need for an external classifier, comes with its limitation: once trained, its adaptability is restricted, unable to accommodate different classifiers or evolving guidance criteria.

To address these constraints, the Control Net (Zhang & Agrawala, 2023) approach was introduced, representing a significant development in guided diffusion models. Control Net, though requiring less training than traditional classifier guidance, still necessitates some degree of model training. Moreover, its utility is predominantly confined to image-to-image signal guidance, limiting its scope. On the other hand, Universal guidance (Bansal et al., 2023) takes a different approach and completely eschews the need for training new models or classifiers for guidance. Instead, it utilizes signals that can be derived from clean images, employing existing models or loss functions. This strategy significantly enhances the flexibility and efficiency of guidance in diffusion models. However, compared to Control Net, it is computationally expensive during inference.

B. Limitations and Future Work

- Our method requires a diffusion model trained specifically on the particular data distribution which is compu-

tationally expensive, as is guided diffusion generation. Can we craft equally effective base samples on a tight budget, perhaps without diffusion models at all?

- We require the entire training set, or at least a large subset, for training the diffusion model. Can we instead use a general purpose text-to-image diffusion model with appropriate prompts to avoid training a dataset-specific diffusion model?

- We generate significantly more poisons than we deploy and then filter them, but generating poisons is expensive, so this procedure is inefficient. Can we devise a more reliable optimization strategy to avoid filtering?

C. Potential Negative Societal Impacts

Similarly to adversarial and data poisoning attacks, Guided Diffusion Poisons could be leveraged by an attacker to compromise the security of real deep learning systems in the wild. As with other attack literature, our hope is to alert the community to the possibility of this threat and to encourage the development of robust defenses.
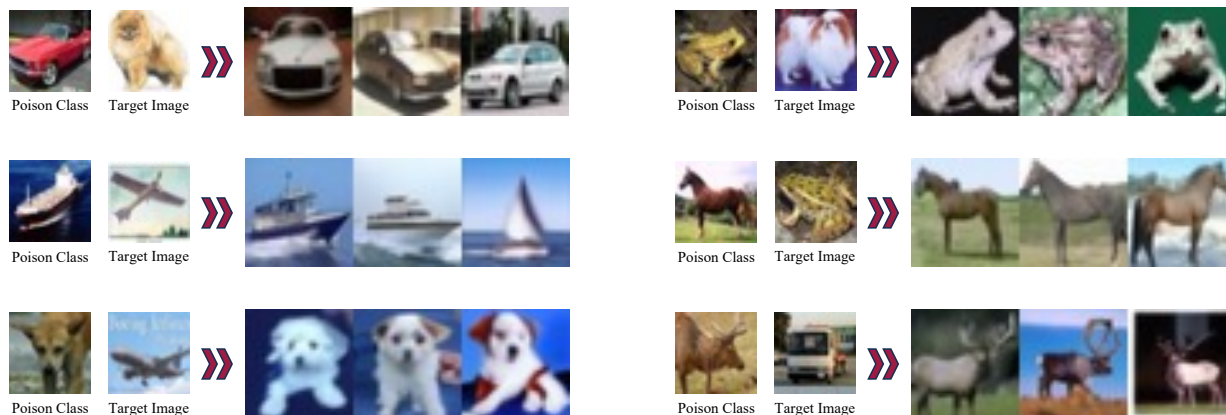
Figure 3: **GDP base samples are clean-label and high quality (CIFAR-10).** In each panel, the leftmost column contains a random sample from the poison class, the second column contains the target image, and the subsequent three columns contain GDP base samples. Experiments conducted using the Witches' Brew gradient-matching objective with a ResNet-18 model on CIFAR-10 over randomly sampled poison class and target image pairs. Additional visualizations are found in Appendix F.

D. Experimental Details

D.1. Threat Model

We adhere to the standard threat model commonly employed in targeted data poisoning and backdoor attacks (Geiping et al., 2020; Souri et al., 2022; Saha et al., 2020). Assuming the attacker possesses access to the training set, it can manipulate a small subset of the training data by perturbing images within an ℓ_∞-norm bound from clean non-perturbed images. In our baseline experiments, we consider a gray-box scenario, where the attacker is aware of the model architecture of the victim but lacks knowledge of the parameters. In addition, we assume the attack must be clean-label. Our backdoor attacks operate on a one-to-one basis, focusing on manipulating a single class and being triggered solely by a single patch during testing. The trigger, which is a random patch in our setup, must remain *hidden* from the training set and is chosen arbitrarily. After crafting the poisons, the victim trains their model from scratch on the poisoned training data. We then measure the *Poison Success* rate and *Attack Success Rate* for targeted data poisoning and backdoor attacks, respectively.

D.2. Experimental Setup

In this section, we provide the details of our experimental evaluations. In all poisoning experiments discussed in this paper, we consider *targeted data poisoning attacks* which are designed to cause a particular target test image to be misclassified with an intended label (Shafahi et al., 2018). In our poisoning attacks, the *Poison Success* rate denotes the

percentage of instances where the victim network assigns the *target image* the intended label. Consistent with the methodology outlined by Geiping et al. (2020), we choose the poison class to be the same as the intended class. Our reported poison success is based on the average performance across 10 randomly selected target-poison pairs.

For backdoor attacks, the *Attack Success Rate* is the rate at which the victim model misclassifies test images from the *target* class, which have been manipulated to include the trigger, with the intended label. It is noteworthy that, in previous backdoor attack studies, this scenario is often referred to as source-target pairs, indicating that images from the source class are misclassified as the target class (Saha et al., 2020; Souri et al., 2022). However, to prevent confusion with the terminology used in the targeted data poisoning task (i.e., target-intended), we abstain from employing the source-target analogy in our paper. Furthermore, in our backdoor attack experiments, following the approach outlined by Souri et al. (2022), we exclusively select poisons from the intended class. Consequently, we consistently use the term "*target-poison*" for all our data poisoning and backdoor attack experiments. We report the average attack success rate across 10 trials, with randomly selected target-poison pairs.

Algorithm 1 provides a summary of the proposed GDP attack. It outlines the key steps of the attack, offering an overview of how it is carried out.

Figure 4: Visualizations of the triggered test images from the ImageNet dataset.

Algorithm 1 Guided Diffusion Poisoning (GDP)

Input: Diffusion model G, surrogate model F_θ, downstream poisoning loss L, poison class c, number of base samples N, poison budget $P \leq N$

Begin:

1: **for** $i = 1, 2, ..., N$ **do**
2:     Generate base sample b_i with class label c from G using forward universal guidance as in Equation (6)
3: **end for**
4: Randomly initialize perturbations $\delta_{i=1}^N$
5: Calculate $\delta_{i=1}^N$ by optimizing $L(F_\theta, (b_i + \delta_i, c)_{i=1}^N)$
6: Select P poisons from $(b_i + \delta_i, c)_{i=1}^N$ with lowest value of $L(F_\theta, (b_i + \delta_i, c))$
7: Replace P randomly selected training images from class c with $(b_i + \delta_i, c)_{i=1}^P$

D.3. Implementation Details

In the backdoor attack experiments, adhering to the experimental setup outlined by Saha et al. (2020) and Souri et al. (2022), the trigger is a random patch as illustrated in Figure 4. The patch size is set to 8×8 for CIFAR-10 experiments and 30×30 for ImageNet experiments. In our baseline experiments, we employ ResNet-18 (He et al., 2016), while transfer experiments involve ResNet-34, MobileNet-v2, and VGG11 networks (He et al., 2016; Sandler et al., 2018; Simonyan & Zisserman, 2014). Additionally, we consider a 6-layer ConvNet in our Bullseye and Poison Frogs poisoning attacks following Geiping et al. (2020). Our 6-layer ConvNet consists of 5 convolutional layers succeeded by a linear layer. The initial learning rate is set to 0.1 for ResNet-18 and ResNet-34, and 0.01 for ConvNet, MobileNet-v2, and VGG11. All models undergo training for 40 epochs, with the learning rate reduced by a factor of 0.1 at epochs 14, 24, and 35. In line with the approach of Souri et al. (2022), backdoor attack experiments involve training the victim model for 80 epochs during validation. For all models, we employ SGD with Nesterov momentum and a momentum coefficient of 0.9. Additionally, data augmentation is applied to enhance classifier accuracy, including horizontal flipping with a probability of 0.5 and random crops of size 32×32 with zero-padding of 4. For ImageNet,

images are resized to 256×256, followed by a central crop of size 224×224, horizontal flip with a probability of 0.5, and random crops of size 224×224 with zero-padding of 28. In all experiments, the victim model is trained from scratch during validation.

To generate GDP base samples, we employ pretrained diffusion models from Ho et al. (2020) for CIFAR-10. For ImageNet, we utilize classifier-guided diffusion checkpoints from Dhariwal & Nichol (2021), with the classifier scale set to 1. Specifically, for CIFAR-10, we guide the diffusion model to generate samples from the poison class by employing cross-entropy loss calculated using a pretrained ResNet-18 classifier.

For all experiments, including generating GDP base samples, downstream poisoning, and backdoor attacks, we use one NVIDIA RTX A5000 GPU.

E. Additional Experiments

E.1. Effect of GDP on Validation Accuracy

To demonstrate the effectiveness of GDP, we present the validation accuracy of both clean and poisoned models in both targeted data poisoning and backdoor attack baseline experiments. As shown in Table 13, it is evident that GDP data poisoning does not lead to a degradation in the validation accuracy of the poisoned model. Furthermore, the validation accuracy of images from the poison class itself remains unaffected.

E.2. More Evaluations on ImageNet

In addition to the backdoor attack experiments detailed in Section 4.1, we further explore backdoor attacks on the ImageNet dataset, a task acknowledged as challenging in prior literature (Saha et al., 2020; Souri et al., 2022). As demonstrated in Table 14, our GDP backdoor attack proves effective even with a small budget of only 100 poisons, while Sleeper Agent yields a negligible attack success rate. It is noteworthy that 100 poisons represents approximately 0.008% of the ImageNet dataset.

Table 11: **Black-box targeted poisoning attacks. GDP improves targeted poisoning success rates in the harder black-box setting.** Perturbations have ℓ_∞-norm bounded above by $16/255$. Poisons crafted using ResNet-18 surrogate and transferred to different victim architectures. Experiments are conducted on CIFAR-10.

Victim architecture	Poison budget	Witches' Brew	GDP (ours)
ResNet-34	25 images (0.05%)	2.50%	**30.00%**
ResNet-34	50 images (0.1%)	20.00%	**37.50%**
ResNet-34	100 images (0.2%)	35.00%	**52.50%**
MobileNet-V2	100 images (0.2%)	0%	**13.75%**

Table 12: **Black-box backdoor attacks. GDP improves backdoor attack success rates in the harder black-box setting.** Perturbations have ℓ_∞-norm bounded above by $16/255$. Poisons crafted using ResNet-18 surrogate and transferred to different victim architectures. Experiments are conducted on CIFAR-10.

Attack	Poison budget	MobileNet-V2	VGG11	ResNet-34	Average
Sleeper Agent	25 images (0.05%)	5.52%	7.71%	11.32%	8.18%
GDP (ours)	25 images (0.05%)	**10.74%**	**9.77%**	**26.16%**	**15.58%**
Sleeper Agent	50 images (0.1%)	9.20%	11.20%	15.74%	12.38%
GDP (ours)	50 images (0.1%)	**13.24%**	**12.62%**	**39.64%**	**21.38%**

Table 13: **Validation accuracy of GDP**. Experiments are conducted on CIFAR-10 with ResNet-18 models, and perturbations are bounded in ℓ_∞-norm by $16/255$. Poison budget is 50 images (0.1%).

	Model	Targeted Data Poisoning	Backdoor Attack
Validation accuracy	Clean	92.08%	91.96%
Validation accuracy	Poisoned	92.16%	92.15%
Validation accuracy poison class	Clean	92.15%	92.36%
Validation accuracy poison class	Poisoned	92.50%	92.36%

Table 14: **Backdoor attacks. GDP achieves a far higher success rate than Sleeper Agent, even with only a small budget**. Experiments are conducted on ImageNet with ResNet-18 models. Perturbations are bounded in ℓ_∞-norm by $16/255$. We see that injection of only 100 poisoned samples is enough for the attack to be effective.

Poison budget	Sleeper Agent	GDP (ours)
100 images ($\sim 0.008\%$)	3.50%	**14.00%**

E.3. Additional Transfer Experiments

In Section 4.3, we discussed how GDP enhances the transferability of poisons across various architectures. To further illustrate this point, we conduct backdoor attack experiments on CIFAR-10 using an ensemble of six models, comprising two ResNet-18, two MobileNet-V2, and two VGG11 models. As depicted in Table 15, it is evident that, in comparison to the Sleeper Agent, GDP achieves a higher attack success rate by employing ensembling, achieving an average success rate of 19.37% with the budget of only 50 poisons.

F. Visualizations

In this section, we present additional visualizations of GDP attacks on CIFAR-10 and ImageNet datasets. In Figure 6, we observe the influence of a specific target image on the corresponding base samples, even within a fixed poison class. Figures 7, 8, 9, 10 depict GDP base samples along with their corresponding poisons in backdoor attacks on ImageNet and CIFAR-10. Additionally, GDP base samples and their corresponding poisons in targeted data poisoning attacks on ImageNet and CIFAR-10 are shown in Figures 11, 12, 13, 14, 15.

Table 15: **Transferring backdoor attacks using ensembles.** Experiments are conducted on CIFAR-10 and perturbations have ℓ_∞-norm bounded above by $16/255$. Poisons crafted using an ensemble of 6 models. S denotes the size of the ensemble.

Attack	Poison budget	ResNet-18	MobileNet-V2	VGG11	Average
Sleeper Agent ($S = 6$)	25 images (0.05%)	10.20%	6.35%	10.38%	8.97%
GDP ($S = 6$)	25 images (0.05%)	**19.73%**	**14.56%**	**15.95%**	**16.74%**
Sleeper Agent ($S = 6$)	50 images (0.1%)	20.25%	7.35%	14.89%	14.16%
GDP ($S = 6$)	50 images (0.1%)	**24.16%**	**14.65%**	**19.30%**	**19.37%**
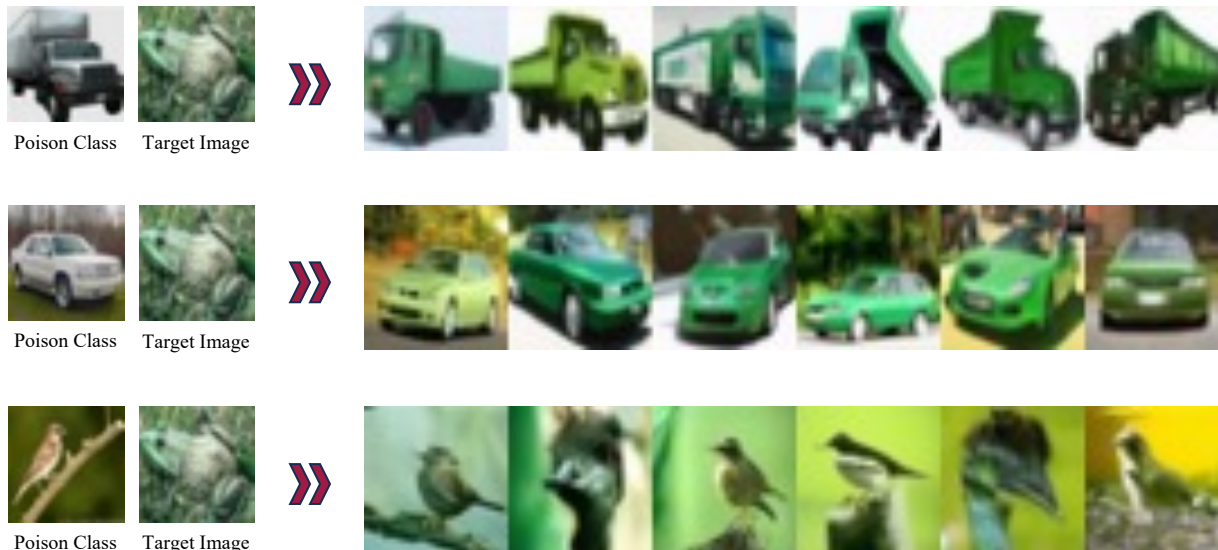
Figure 5: **GDP produces base samples that look like the target image while still remaining in the poison class.** We generate base samples from different poison classes while the target image is fixed. We see that all resulting GDP base samples contain similar colors to the target image but remain clean-label. Experiments conducted on the CIFAR-10 dataset using the Witches' Brew poisoning objective along with a ResNet-18 model.
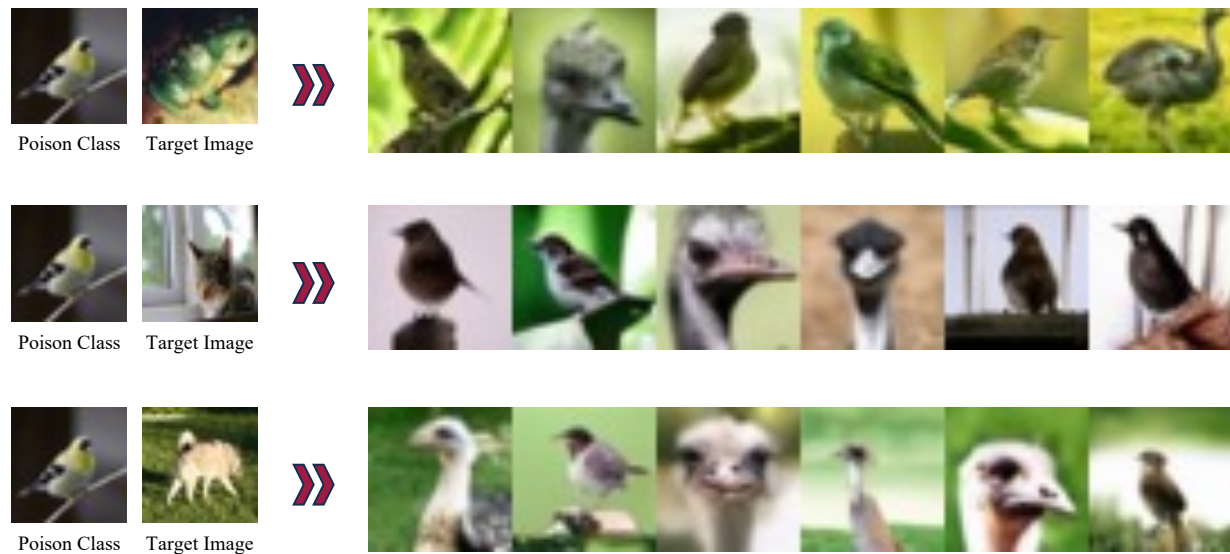
Figure 6: **The target image visibly influences the corresponding base samples.** We generate base samples from the fixed bird class but using different target images. We see that the resulting GDP base samples look similar to the target image but remain birds. Experiments conducted on the CIFAR-10 dataset using the Witches' Brew poisoning objective along with a ResNet-18 model.

Figure 7: **GDP base samples and their corresponding poisons (ImageNet).** Experiments conducted using the Sleeper Agent gradient-matching objective with a ResNet-18 model on ImageNet over randomly sampled poison class and target class pairs. Perturbations have ℓ_∞-norm bounded above by $16/255$, and the patch size is 30×30.
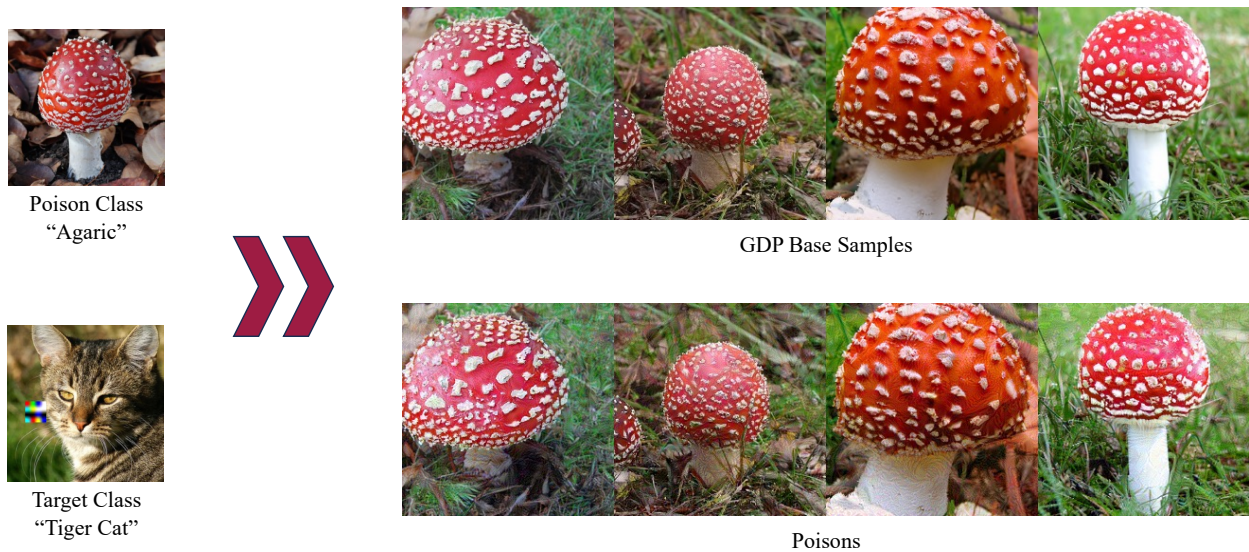
Figure 8: **GDP base samples and their corresponding poisons (ImageNet).** Experiments conducted using the Sleeper Agent gradient-matching objective with a ResNet-18 model on ImageNet over randomly sampled poison class and target class pairs. Perturbations have ℓ_∞-norm bounded above by $16/255$, and the patch size is 30×30.

Figure 9: **GDP base samples and their corresponding poisons (CIFAR-10).** Experiments conducted using the Sleeper Agent gradient-matching objective with a ResNet-18 model on CIFAR-10 over randomly sampled poison class and patched class pairs. Perturbations have ℓ_∞-norm bounded above by $16/255$, and the patch size is 8×8.
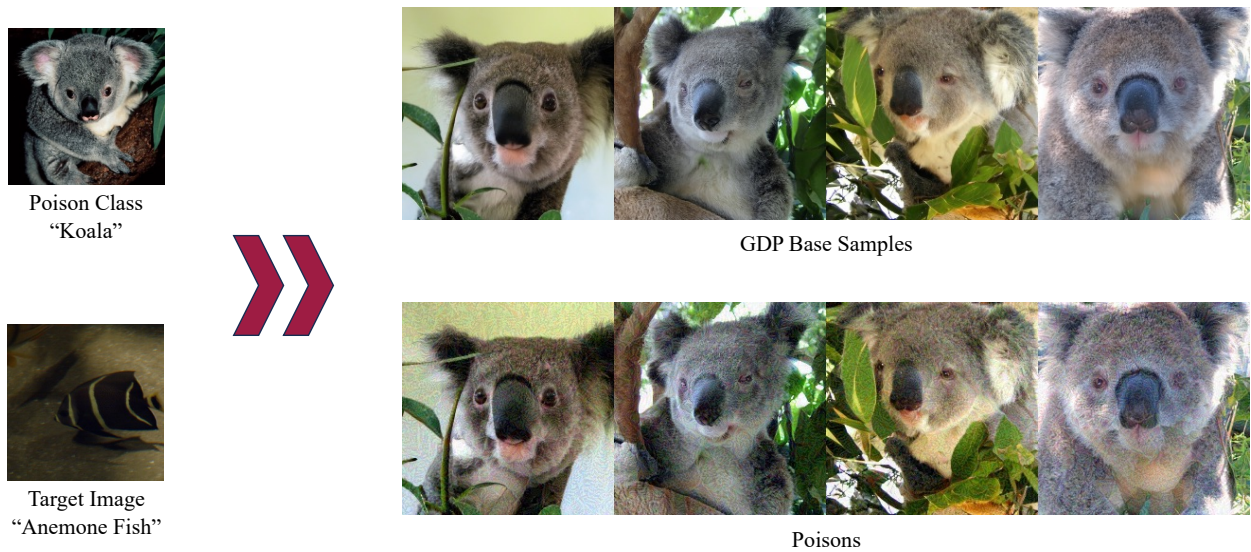
Figure 10: **GDP base samples and their corresponding poisons (CIFAR-10).** Experiments conducted using the Sleeper Agent gradient-matching objective with a ResNet-18 model on CIFAR-10 over randomly sampled poison class and patched class pairs. Perturbations have ℓ_∞-norm bounded above by $16/255$, and the patch size is 8×8.
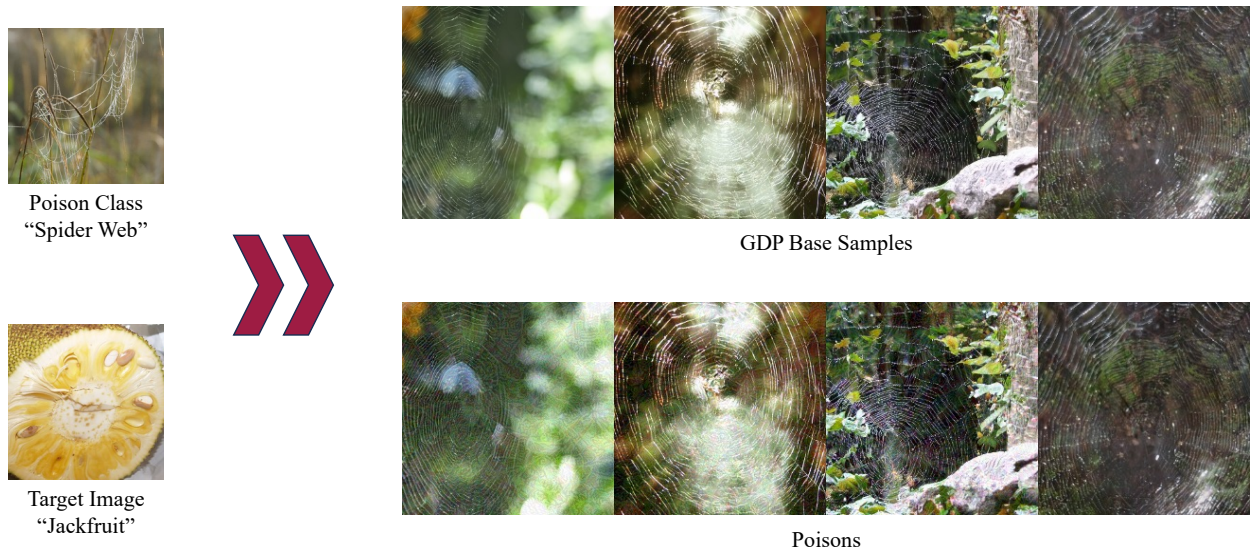

Figure 11: **GDP base samples and their corresponding poisons (ImageNet).** Experiments conducted using the Witches' Brew gradient-matching objective with a ResNet-18 model on ImageNet over randomly sampled poison class and target image pairs. Perturbations have ℓ_∞-norm bounded above by $16/255$.
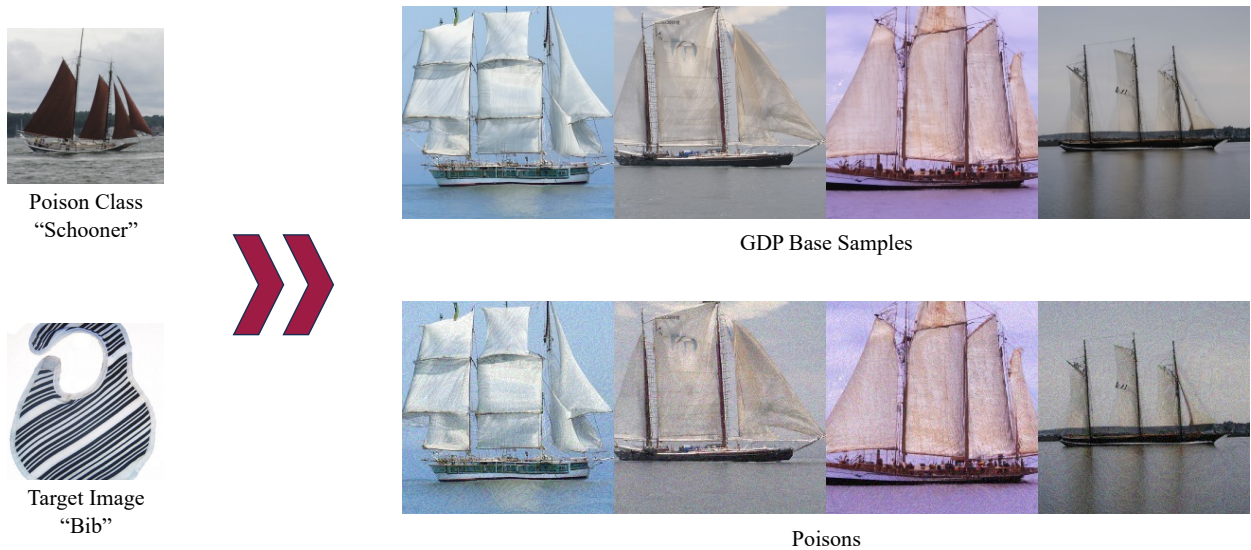
Figure 12: **GDP base samples and their corresponding poisons (ImageNet).** Experiments conducted using the Witches' Brew gradient-matching objective with a ResNet-18 model on ImageNet over randomly sampled poison class and target image pairs. Perturbations have ℓ_∞-norm bounded above by $16/255$.

Figure 13: **GDP base samples and their corresponding poisons (ImageNet).** Experiments conducted using the Witches' Brew gradient-matching objective with a ResNet-18 model on ImageNet over randomly sampled poison class and target image pairs. Perturbations have ℓ_∞-norm bounded above by $16/255$.
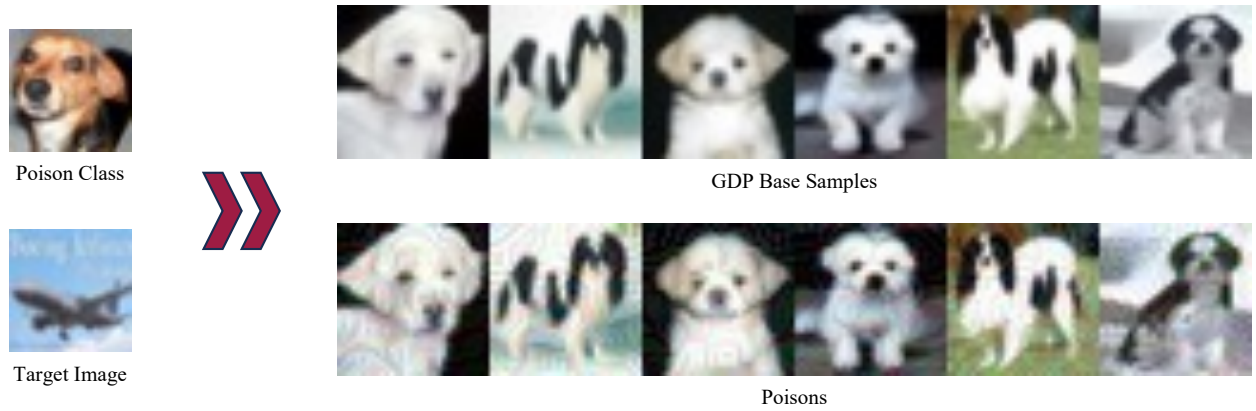
Poison Class

Target Image

GDP Base Samples

Poisons

Figure 14: **GDP base samples and their corresponding poisons (CIFAR-10).** Experiments conducted using the Witches' Brew gradient-matching objective with a ResNet-18 model on CIFAR-10 over randomly sampled poison class and target image pairs. Perturbations have ℓ_∞-norm bounded above by $16/255$.

Poison Class

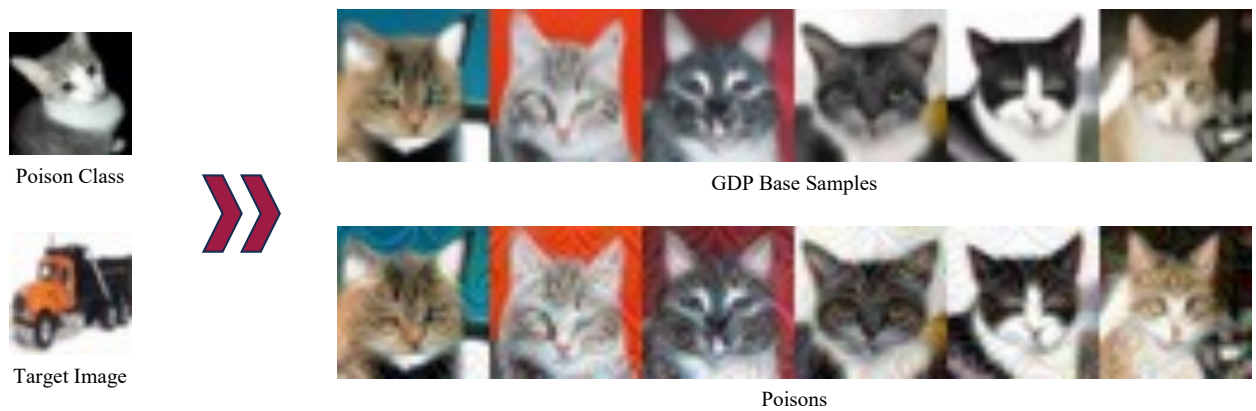Target Image

GDP Base Samples

Poisons

Figure 15: **GDP base samples and their corresponding poisons (CIFAR-10).** Experiments conducted using the Witches' Brew gradient-matching objective with a ResNet-18 model on CIFAR-10 over randomly sampled poison class and target image pairs. Perturbations have ℓ_∞-norm bounded above by $16/255$.