# RrED: Black-box Unsupervised Domain Adaptation via Rectifying-reasoning Errors of Diffusion

Yuwu Lu ✉, Chunzhi Liu
South China Normal University

## Motivation

Recent diffusion models have significantly advanced the ability to generate images from texts. In OOD scenarios, some approaches propose employing diffusion models to augment the dataset with additional synthetic samples. While they can produce realistic visuals across diverse prompts and demonstrate impressive compositional generalization, these diffusion-based methods focus solely on composition, overlooking their sensitivity to textual nuances. Moreover, they fail to address the emerging risk of data leakage caused by reverse generation from generative models.

To solve these problems, we propose Rectifying-reasoning Errors of Diffusion (RrED), which is the first work that applies the diffusion model to high-security BUDA tasks innovatively.

## Contributions

- We observe some weaknesses in existing DA methods and address them by proposing a novel method, named RrED, which introduces the diffusion model into the BUDA setup and strengthens the target model's reasoning ability through our two-stage learning.
- Inspired by the improved human decision-making process, RrED is designed to consist of two stages, namely DTR and SRM.
- We present the theoretical analysis of the conditions under which the BUDA task is feasible, addressing the lack of theoretical justification in existing work. Further, we elucidate the underlying mechanism of our method RrED.

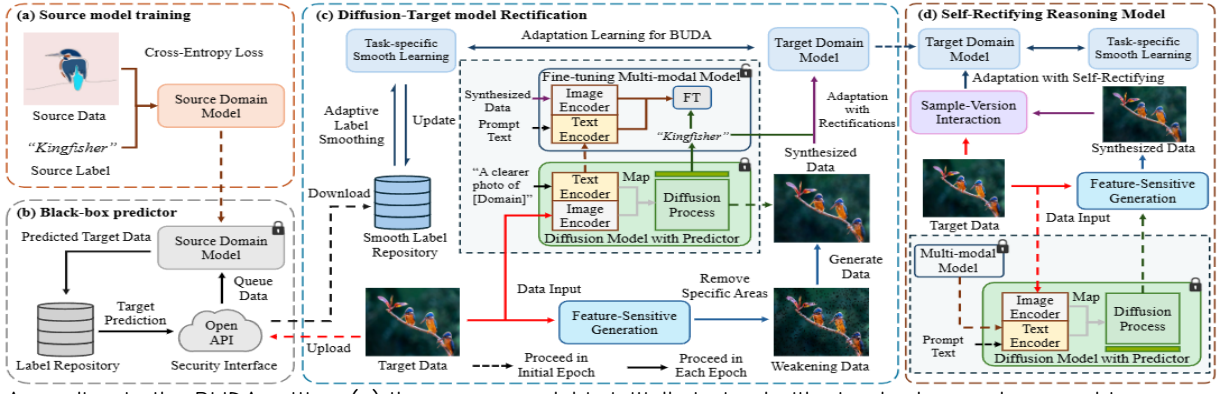
## Definition and Method

**What is Black-box Unsupervised Domain Adaptation (BUDA), and what are its key advantages ?**

| Setting | Source data | Source model | Predicted target labels | Target data | External prompt | Privacy risk |
|---|---|---|---|---|---|---|
| DG | ✓ | ✓ | × | × | × | Medium |
| Traditional DA | ✓ | ✓ | ✓ | ✓ | × | High |
| Source-free DA | × | ✓ | ✓ | ✓ | × | Medium |
| Black-box DA | × | × | ✓ | ✓ | × | Low |
| Diffusion-based DA | ✓ | ✓ | ✓ | ✓ | ✓ | High |
| Our RrED | × | × | ✓ | ✓ | ✓ | Low |

Diffusion-based DA relies on both labeled source and unlabeled target data, guided by an external diffusion model. However, it follows the Traditional DA, incurring high computational costs and posing significant risks of data leakage. Even though SFDA methods lower the possibility of privacy leaks by utilizing the pre-trained source model rather than source data, found that certain generation techniques have the potential to reconstruct the source data through learning from the source model. Black-box DA only relies on the unlabeled target data and the predicted labels from a black-box predictor, thus offering better data privacy at the cost of partial performance. Our RrED follows Black-box DA setting for training with a diffusion model incorporated, achieving performance improvement while maintaining high-level data privacy protection.

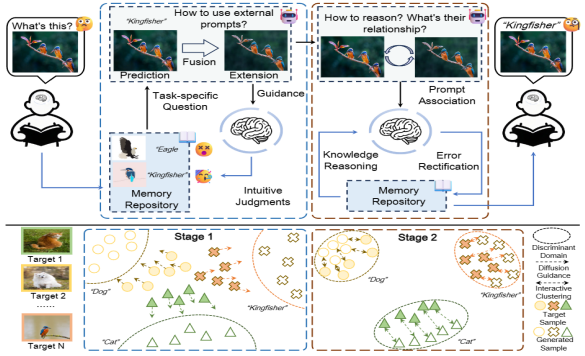
**How does RrED leverage a diffusion model to guide the training process and address the BUDA ?**




According to the BUDA setting, (a) the source model is initially trained with standard procedures and transferred to a black-box predictor; (b) the black-box predictor then exposes a restricted API, allowing external clients to query only batches of hard target predictions through iterative requests. In our RrED, (c) DTR guides the target model's learning by correcting reasoning errors from the diffusion model and leveraging its semantic knowledge; (d) SRM corrects the reasoning error of the target model by leveraging the model's reasoning from predictive differences across versions.

## Inspiration

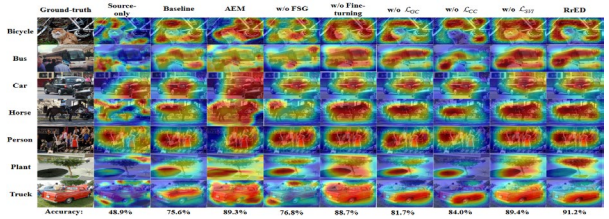
**Why does RrED adopt a two-stage learning framework ?**




Conceptual figure of our RrED. Above: our RrED simulates the human decision-making process under the guidance of external knowledge. Below: in stage 1, RrED aligns the decision boundaries of the diffusion model under guidance; in stage 2, RrED enhances the model's self-reasoning ability by rectifying the errors among different versions.

## Visualization


Generation bias rectification for diffusion model




Reasoning error rectification for target model