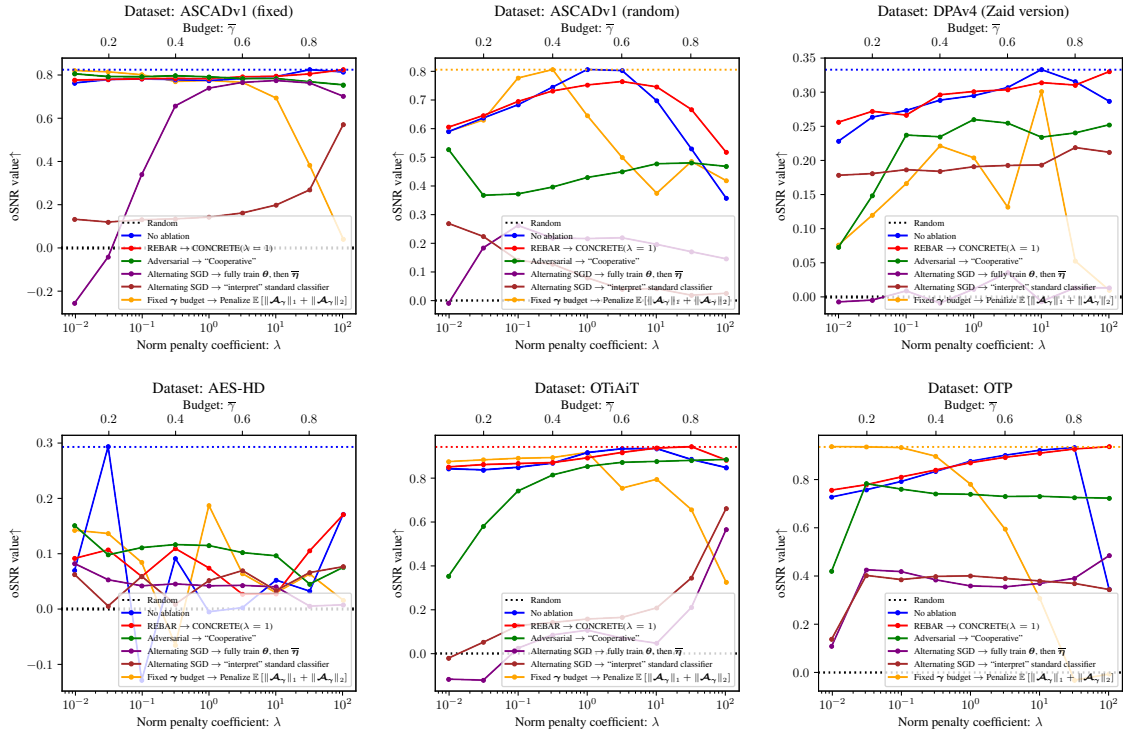

Figure 1: Ablation studies where we track the performance of ALL (ours) while changing various design decisions of our method to make it closer in spirit to Cheng (2023) := [1] and Yang (2021) := [2] as referenced by Reviewer Zg1h. Performance is measured using the oSNR metric (**larger is better**). For each ablated version of ALL we do a grid search over the 'amount of noise' hyperparameter (the norm penalty strength λ for the yellow trace and the budget $\overline{\gamma}$ for all others) while leaving all other hyperparameters fixed at their optimal values chosen through random search as described in Appendix C.3.3. The black dotted line denotes the performance of random guessing and is provided for reference. The blue trace denotes the unmodified version of ALL as described in the paper. The red trace denotes optimizing our erasure probabilities using CONCRETE with temperature 1, instead of REBAR (as done in [2]); this does not significantly change performance, and is an attractive option for future work due to its simplicity. The yellow trace denotes removing our budget constraint on $\boldsymbol{\gamma}$ and instead adding the term $\lambda\mathbb{E}\left[\|\boldsymbol{\mathcal{A}}_{\boldsymbol{\gamma}}\|_1 + \|\boldsymbol{\mathcal{A}}_{\boldsymbol{\gamma}}\|_2\right]$ to our objective function (as done in [2]); this appears to work similarly well, but with greater variability in performance and a less intuitively-meaningful hyperparameter. The green trace corresponds to doing 'cooperative' training where the noise distribution is trained to distribute noise to *minimize* the loss of the classifiers, and less-noisy timesteps are interpreted as leakier (closer in spirit to [1, 2] than our adversarial objective); this consistently moderately degrades performance. The brown and purple traces correspond to 'interpreting' a fixed classifier instead of doing alternating SGD where we train the classifiers to be optimal with respect to the noise distribution; both lead to significant performance degradation. In the former case we simply replace Φ_θ as defined in equation 8 with the fixed classifier used for the neural net attribution baselines (close to what [1, 2] did), and in the latter case we keep Φ_θ unchanged but fully solve the inner maximization problem followed by the outer minimization problem of equation 8 (similar in spirit, and more-consistent with our theory). *Note: due to time constraints these trials were performed with 1 random seed. We can increase this to 5 random seeds prior to publication.