Appendix

A. Proof of Lemma 1

For brevity we use $[K]$ to denote the integer set $\{1, 2, \ldots, K\}$. We have

$$\mathbb{E}_{x\sim\mathcal{D},\ x'\sim\mathcal{D}'}\|f(x)-f(x')\|_2 \overset{(a)}{=} \mathbb{E}_{x\sim\mathcal{D},\ x'\sim\mathcal{D}'}\|\eta(z(x))-\eta(z(x'))\|_2 \tag{5}$$

$$\overset{(b)}{=} \int_{x\sim\mathcal{D},\ x'\sim\mathcal{D}'} \|\eta(z(x))-\eta(z(x'))\|_2 \Phi_{\mathcal{D},\mathcal{D}'}(x,x')dxdx' \tag{6}$$

$$\overset{(c)}{=} \int_{x\sim\mathcal{D},\ x'\sim\mathcal{D}'} \|\eta(z(x))-\eta(z(x'))\|_2 \Phi_{\mathcal{D}}(x)\cdot\Phi_{\mathcal{D}'}(x')dxdx' \tag{7}$$

$$\overset{(d)}{\leq} \sqrt{K}\cdot\int_{x\sim\mathcal{D},\ x'\sim\mathcal{D}'} \max_{k\in[K]} |[\eta(z(x))]_k-[\eta(z(x'))]_k|\Phi_{\mathcal{D}}(x)\cdot\Phi_{\mathcal{D}'}(x')dxdx' \tag{8}$$

$$\overset{(e)}{\leq} 2\sqrt{K}\cdot \sup_{g:\mathbb{R}^K\mapsto\mathbb{R},\|g\|_{\text{Lip}}\leq 1} \mathbb{E}_{x\sim\mathcal{D}}[g(z(x))]-\mathbb{E}_{x'\sim\mathcal{D}'}[g(z(x'))] \tag{9}$$

$$\overset{(f)}{=} 2\sqrt{K}\cdot\mathcal{W}_1(\mu_z,\mu_z') \tag{10}$$

(a) follows the neural network model, (b) follows the definition of expectation, (c) follows the assumption of independent data drawing, and (d) follows that $\|x\|_2=\sqrt{\sum_i^d x_i^2}\leq\sqrt{d\cdot\max_i[x_i^2]}=\sqrt{d}\cdot\max_i|x_i|$, and thus $\|\eta-\eta'\|_2\leq\sqrt{K}\cdot\max_k|[\eta-\eta']_k|$. (e) holds by setting $k^+=\arg\max_{k\in[K]}[\eta(z(x))]_k-[\eta(z(x'))]_k$ and $k^-=\arg\max_{k\in[K]}[\eta(z(x'))]_k-[\eta(z(x))]_k$. Then by definition $\max_{k\in[K]}|[\eta(z(x))]_k-[\eta(z(x'))]_k|\leq[\eta(z(x))]_{k^+}-[\eta(z(x'))]_{k^+}+[\eta(z(x'))]_{k^-}-[\eta(z(x))]_{k^-}$. We further make the following three notes: (i) $[\eta(z(x))]_{k^+}-[\eta(z(x'))]_{k^+}\geq 0$ and $[\eta(z(x))]_{k^-}-[\eta(z(x'))]_{k^-}\geq 0$; (ii) $|a|=\max\{a,-a\}$, and if $a,b\geq 0$, $\max\{a,b\}\leq a+b$; (iii) There exist at least one k such that $[\eta(x)]_k-[\eta(x)]_k\geq 0$. One can use proof by contradiction to show (iii) is true. If $[\eta(x)]_k-[\eta(x)]_k<0$ for every k, then summing over k we get a contradiction that $1<1$. Therefore,

$$\int_{x\sim\mathcal{D},\ x'\sim\mathcal{D}'} \max_{k\in[K]} |[\eta(z(x))]_k-[\eta(z(x'))]_k|\Phi_{\mathcal{D}}(x)\cdot\Phi_{\mathcal{D}'}(x')dxdx' \tag{11}$$

$$\leq \int_{x\sim\mathcal{D},\ x'\sim\mathcal{D}'} \left([\eta(z(x))]_{k^+}-[\eta(z(x))]_{k^-}+[\eta(z(x'))]_{k^-}-[\eta(z(x'))]_{k^+}\right)\Phi_{\mathcal{D}}(x)\cdot\Phi_{\mathcal{D}'}(x')dxdx' \tag{12}$$

$$=\mathbb{E}_{x\sim\mathcal{D}}[[[\eta(z(x))]_{k^+}-[\eta(z(x))]_{k^-}]-\mathbb{E}_{x'\sim\mathcal{D}'}[[\eta(z(x'))]_{k^+}-[\eta(z(x'))]_{k^-}] \tag{13}$$

$$\leq 2\cdot \sup_{g:\mathbb{R}^K\mapsto\mathbb{R},\|g\|_{\text{Lip}}\leq 1} \mathbb{E}_{x\sim\mathcal{D}}[g(z(x))]-\mathbb{E}_{x'\sim\mathcal{D}'}[g(z(x'))] \tag{14}$$

where $\|g\|_{\text{Lip}}$ is defined as $\sup_{x,x'}|g(x)-g(x')|/\|x-x'\|_2$, and we use the fact that $[\eta(z)]_k$ is 1-Lipschitz for any $k\in[K]$ (Gao & Pavel, 2017) (so $[\eta]_{k^+}-[\eta]_{k^-}$ is 2-Lipschitz). Finally, (f) follows the Kantorovich-Rubinstein theorem (Kantorovich & Rubinstein, 1958) of the dual representation of the Wasserstein-1 distance.

B. Proof of Theorem 1

First, we decompose the target risk function as

$$\ell_{\mathcal{T}}(x_t+\delta^*,y_t) \overset{(a)}{=} \ell_{\mathcal{S}}(x_t+\delta^*,y_s) \tag{15}$$

$$\overset{(b)}{=} \|f_{\mathcal{S}}(x_t+\delta^*)-y_s\|_2 \tag{16}$$

$$\overset{(c)}{=} \|f_{\mathcal{S}}(x_t+\delta^*)-f_{\mathcal{S}}(x_s)+f_{\mathcal{S}}(x_s)-y_s\|_2 \tag{17}$$

$$\overset{(d)}{\leq} \underbrace{\|f_{\mathcal{S}}(x_t+\delta^*)-f_{\mathcal{S}}(x_s)\|_2}_{A}+\underbrace{\|f_{\mathcal{S}}(x_s)-y_s\|_2}_{B} \tag{18}$$

(a) is based on Assumption 3, (b) is based on the definition of risk function, (c) is by subtracting and adding the same term $f_{\mathcal{S}}(x_s)$, and (d) is based on the triangle inequality.

Note that by Assumption 1, $\mathbb{E}_{\mathcal{D}_\mathcal{S}} B = \mathbb{E}_{\mathcal{D}_\mathcal{S}}[\ell(x_s, y_s)] = \epsilon_\mathcal{S}$. Next, we proceed to bound $\mathbb{E}_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{T}} A \triangleq \mathbb{E}_{x_s\sim\mathcal{D}_\mathcal{S},x_t\sim\mathcal{D}_\mathcal{T}} A$. Using Lemma 1, we have

$$\mathbb{E}_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{T}} A \leq 2\sqrt{K} \cdot \mathcal{W}_1(\mu(z_\mathcal{S}(x_t + \delta^*)), \mu(z_\mathcal{S}(x_s)))_{x_t\sim\mathcal{D}_\mathcal{T},\, x_s\sim\mathcal{D}_\mathcal{S}} \tag{19}$$

Finally, take $\mathbb{E}_{\mathcal{D}_\mathcal{S},\mathcal{D}_\mathcal{T}}$ on both sides of equation (18) completes the proof.

C. Pre-Trained Model Studies

We provide advanced studies over different pre-trained acoustic architectures and the associated time series classification performance. In particular, we select the models below, which have attained competitive performance tested on Google Speech Commands version 2 dataset (Warden, 2018) or shown cutting-edge performance (ResNet (He et al., 2016)) in the acoustic scene (VGGish (Hershey et al., 2017)) and time series classification (TCN (Lea et al., 2016)). These models will be compared with V2S$_a$ (recurrent Attention (de Andrade et al., 2018)) and V2S$_u$ (V2S$_a$ enhanced by U-Net (Yang et al., 2021a)) used in the main paper.

ResNet: Deep residual network (He et al., 2016) (ResNet) is a popular deep architecture to resolve the gradient vanish issues by passing latent features with a residual connection, and it has been widely used in acoustic modeling tasks. We select a 34-layer ResNet model training from the scratch for V2S (denoted as V2S$_r$), which follows the identical parameter settings in (Hu et al., 2020; 2021) for reproducible studies.

VGGish: VGGish (Hershey et al., 2017) is a deep and wide neural network architecture with multi-channel convolution layers, which has been proposed for speech and acoustic modeling. VGGish is also well-known for the large-scale acoustic embedding studies with Audio-Set (Gemmeke et al., 2017) from 2 million Youtube audios. We use the same architecture and train two models: (i) training from scratch (denoted as V2S$_v$) and (ii) selecting an Audio-Set pretrained VGGish and then fine-tuning (denoted as V2S$_p$) on the Google Speech commands dataset for V2S.

Temporal Convolution Network (TCN): TCN (Lea et al., 2016) is an efficient architecture using temporal convolution with causal kernel for sequence classification tasks. We select the TCN architecture and train it from scratch as a baseline for V2S (denoted as V2S$_t$).

OpenL3: OpenL3 (Cramer et al., 2019) is a much recent embedding method with a deep fusion layer for acoustic modeling. We use pretrained OpenL3 embeddings and a dense layer for classification as another baseline for V2S (denoted as V2S$_o$).

C.1. V2S Performance and Sliced Wasserstein Distance

Table 4 shows different neural architectures for acoustic modeling to be used with the proposed V2S method, where acoustic models are pre-trained with the Google Speech Commands dataset (Warden, 2018) version two with 32 commands (denoted as GSCv2.) From the first three rows of Table 4, we observe the recurrent attention models and TCN perform better in mean prediction accuracy, validation loss of the source task. For the target task (same as Table 2), recurrent attention models (V2S$_a$ and V2S$_u$) attain the best performance.

Table 4. V2S ablation studies with different pre-trained acoustic models.

Model	V2S$_a$	V2S$_u$	V2S$_r$	V2S$_v$	V2S$_p$	V2S$_t$	V2S$_o$
Parameters (↓)	**0.2M**	0.3M	1M	62M	62M	1M	4.7M
Source Acc. (↑)	96.90	96.92	96.40	95.40	95.19	**96.93**	92.34
Source Loss (↓)	**0.1709**	0.1734	0.1786	0.1947	0.1983	0.1756	0.2145
Mean SWD (↓)	**1.829**	1.873	1.892	4.713	4.956	1.901	5.305
Mean Target Acc. (↑)	**89.91**	87.92	87.22	67.12	63.23	86.45	60.34
Target MPCE (↓)	**2.03**	2.10	2.23	33.4	38.3	2.34	41.34

Both VGGish based architectures (V2S$_v$ and V2S$_p$) show degraded performance on the target tasks prediction, which can be explained by the recent findings (Kloberdanz, 2020) on the degraded performance of using VGG (Simonyan & Zisserman, 2015) and MobileNet-V2 (Sandler et al., 2018) based "wide" and deep convolutional neural architectures for reprogramming visual models. These findings could be also explained in the sense that the wide neural architectures fail to adapt the source

domain distribution according to the sliced Wasserstein distance (SWD) results (fourth row in Table 4), as indicated by Theorem 1. We observe that using the pretrained models associated with higher source accuracy does not always guarantee higher target accuracy (e.g, V2S$_t$ and V2S$_u$)

D. Additional Ablation Studies

Based on the discussion in Section C, we further select three efficient V2S models with different model capacity (0.2M/1M/4.7M), including V2S$_a$, V2S$_r$, and V2S$_t$, to study the mean target accuracy in different training settings and to provide some insights into effective design of V2S models.

D.1. Pretrained Models from Different Dataset

In the previous model reprogramming studies (Tsai et al., 2020; Elsayed et al., 2019), little has been discussed regarding the effectiveness of using different datasets to train the pretrained models. We study three other public acoustic classification benchmark datasets (source tasks), (1) TAU Urban Acoustic Scenes 2020 Mobile (Heittola et al., 2020) (denoted as TAU-UAC), from the annual IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE), (2) AudioSet using in (Hershey et al., 2017), and (3) ESC-50 (Piczak, 2025), a dataset for environmental sound classification, for providing a preliminary V2S study. We extract acoustic features by Mel-spectrogram using Kapre audio layer following the same resolution and frequency setup in (Yang et al., 2020; 2021b) for a fair and reproducible comparison.

As shown in Table 5, the V2S models pretrained from GSCv2 show a higher mean prediction accuracy and lower SWD than the other models, which could be due to the shorter sample length (less than one second) of its source acoustic inputs.

Table 5. V2S performance (mean prediction accuracy) on time series classification (same as Table 2) with different pretrained neural acoustic models and the mean sliced Wasserstein distance for each dataset.

Pretrained Dataset	GSCv2	TAU	AudioSet	ESC
# Training Samples	105k	13.9k	2.08M	2k
# Output Classes	35	10	527	50
Audio length per sample clips	1 sec.	10 sec.	10 sec.	5 sec.
Mean Target Acc. w/ V2S$_a$ (↑)	**89.91**	82.61	80.68	84.48
Mean Target Acc. w/ V2S$_r$ (↑)	87.22	83.57	79.96	83.05
Mean Target Acc. w/ V2S$_t$ (↑)	86.45	80.1	81.81	84.58
Mean SWD per dataset (↓)	**1.874**	2.267	2.481	2.162

D.2. Different V2S Mapping Settings

In (Tsai et al., 2020), frequency mapping techniques show improved performance for black-box (e.g., zeroth order gradient estimation (Chen et al., 2017; Liu et al., 2020) based) adversarial reprogramming models for image classification. We also follow the setup in (Tsai et al., 2020) to compare the many-to-one frequency mapping and the many-to-one random mapping for time series classification. However, the frequency mapping based V2S results show equal or slightly worse (-0.013%) mean prediction accuracy and WSD (+0.0028) performance with 100 runs, which may be owing to the differences of the dimensions and scales between the tasks of image and time series reprogramming.

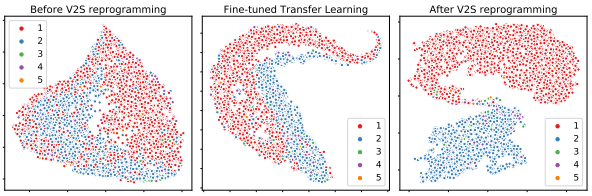
D.3. More tSNE Visualization

We provide more tSNE visualization over different test datasets to better understand the embedding results of V2S models discussed in Section 5.5. In Figure 6 (a) to (e), the reprogrammed representations (rightmost side) show better disentangled results in both 2D and 3D tSNE plots.

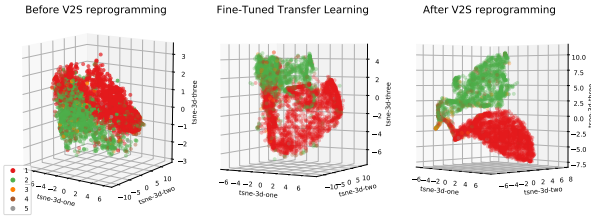
D.4. Hardware Setup and Energy Cost Discussion

We use Nvidia GPUs (2080-Ti and V100) for our experiments with Compute Unified Device Architecture (CUDA) version 10.1. To conduct the results shown in Table 2, it takes around 40 min to run 100 epochs (maximum) with a batch size 32 for each time series prediction dataset considering the hyper-parameters tuning (e.g., dropout rate) described in Section 5.2

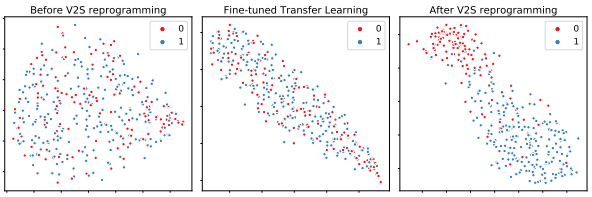
of the main paper. In total, the experiments presented (30 datasets and its ablation studies) in this paper took around 120 computing hours with a 300W power supplier. As another advantage, the V2S reprogramming techniques freeze pretrained neural models and only used a reprogramming layer for training new tasks. The proposed method could potentially recycle well-trained models for an additional task to alleviate extra energy costs toward deploying responsible ML systems.



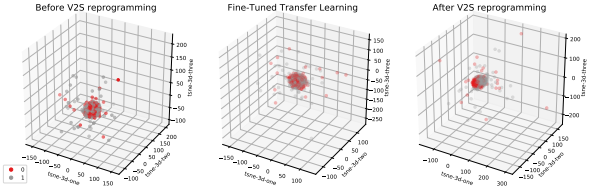
(a) Task: ECG 5000 with 2D tSNE



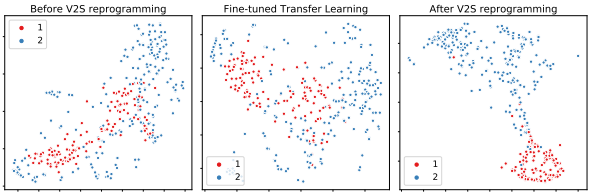
(b) Task: ECG 5000 with 3D tSNE



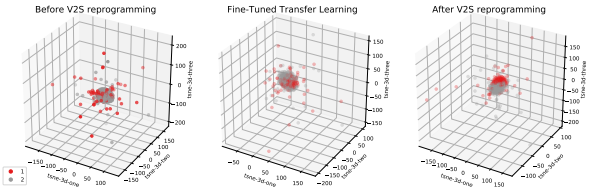
(c) Task: HandOutlines with 2D tSNE.



(d) Task: HandOutlines with 3D tSNE.



(e) Task: Strawberry 2D tSNE.



(f) Task: Strawberry 3D tSNE

Figure 6. More tSNE visualization. Numbers in the legend are class label indices.