Supplementary Materials to "Invariance Learning based on Label Hierarchy"

A Proof of Theorem 2

$$\begin{aligned}\mathcal{R}^{(X,Y)}(p_\theta \circ \Phi) - \mathcal{R}^{(X,g(Y))}(p_\theta \circ \Phi) &= \int -\log p_\theta(Y|\Phi(X)) dP_{Y,\Phi(X)} \\ &\quad + \int \log p_\theta(g(Y)|\Phi(X)) dP_{g(Y),\Phi(X)} \\ &= -\int \log \frac{p_\theta(Y|\Phi(X))}{p_\theta(g(Y)|\Phi(X))} dP_{(Y,\Phi(X))} \\ &= -\int dP_{g(Y)} \int \log \frac{p_\theta(Y|\Phi(X))}{p_\theta(g(Y)|\Phi(X))} dP_{(Y,\Phi(X))|g(Y)} \qquad (8)\end{aligned}$$

By the definition of $p_\theta(y|\Phi(x), g(Y) = z)$ in Theorem 2, $\frac{p_\theta(y|\Phi(x))}{p_\theta(g(y)|\Phi(x))} = p_\theta(y|\Phi(x), g(Y) = z)$ holds, where $z = g(y)$. Therefore, we obtain

$$\begin{aligned}(6) &= -\int dP_{g(Y)} \int \log \frac{p_\theta(Y|\Phi(X))}{p_\theta(g(Y)|\Phi(X))} dP_{(Y,\Phi(X))|g(Y)} \\ &= -\int dP_{g(Y)} \int \log p_\theta(Y|\Phi(X), g(Y) = z) dP_{(Y,\Phi(X))|g(Y)=z} \\ &= -\sum_{z \in \mathcal{Z}} P(g(Y) = z) \int \log p_\theta(Y|\Phi(X), g(Y) = z) dP_{(Y,\Phi(X))|g(Y)=z} \\ &= -\sum_{z^{\not\hookrightarrow} \in \mathcal{Z} \not\hookrightarrow} P(g(Y) = z^{\not\hookrightarrow}) \int \log p_\theta(Y|\Phi(X), g(Y) = z^{\not\hookrightarrow}) dP_{(Y,\Phi(X))|g(Y)=z^{\not\hookrightarrow}} \\ &\qquad + \sum_{z^{\hookrightarrow} \in \mathcal{Z} - \mathcal{Z} \not\hookrightarrow} P(g(Y) = z^{\hookrightarrow}) \int \log p_\theta(Y|\Phi(X), g(Y) = z^{\hookrightarrow}) dP_{(Y,\Phi(X))|g(Y)=z^{\hookrightarrow}}. \qquad (9)\end{aligned}$$

Noting that, for any $z^{\hookrightarrow} \in \mathcal{Z} - \mathcal{Z} \not\hookrightarrow$ and $y := g^{-1}(z^{\hookrightarrow})$[1], $p_\theta(y|\Phi(x), g(Y) = z^{\hookrightarrow}) = 1$ holds, we can see that $\log p_\theta(y|\Phi(x), g(Y) = z^{\hookrightarrow}) = 0$. The second term in the last line thus equals to zero, which concludes the proof. □

B Proof of Theorem 3

Before proving Theorem 3, we recap the problem setting of variable selection for invariance learning discussed in the first paragraph of Section 4. Let $\mathcal{X} := \mathcal{X}_1 \times \mathcal{X}_2$ where $\mathcal{X}_1 := \mathbb{R}^{n_1}$ and $\mathcal{X}_2 := \mathbb{R}^{n_2}$ with $n_1, n_2 \in \mathbb{N}$, so that $\mathcal{X} = \mathbb{R}^n$ with $n = n_1 + n_2$. Throughout our theoretical analysis, to avoid discussing the non-trivial effects of nonlinear Φ, we focus on the simplified case of variable selections, where the feature map Φ is chosen from the projections of x to a subset of its components. For example, Φ may be $\Phi(x_1, x_2, x_3) = (x_1, x_3)$ when x is three-dimensional. Recall that Φ_i denote the $\mathcal{X}_i$-component of Φ ($i = 1, 2$) and $\mathrm{Im}\Phi_2 \neq \emptyset$ means that the range of Φ has a $\mathcal{X}_2$-component.

We rephrase the problem simplification (※) in Section 4. Throughout our theoretical analysis, the domain set $\mathcal{E}$ is defined by all the probability distributions with the fixed marginal distribution $P_{X_1^I, Y^I}$ of (X_1, Y); namely, all domains $T_{all} := \{(X^e, Y^e)\}_{e \in \mathcal{E}}$ are defined by

$$T_{all} := \left\{ (X, Y) : \text{a random variable on } \mathcal{X} \times \mathcal{Y} \,\middle|\, P_{\Phi^{\mathcal{X}_1}(X), Y} = P_{X_1^I, Y^I} \right\}. \qquad (※)$$

In this case, any variable $(X^e, Y^e) \in T_{all}$ satisfies (i) $P_{Y^e|\Phi^{\mathcal{X}_1}(X^e)}$ equals to $P_{Y^I|X_1^I}$, and (ii) the marginal distribution $P_{\Phi^{\mathcal{X}_1}(X)}$ of the invariant feature $\Phi^{\mathcal{X}_1}(X)$ equals to $P_{X_1^I}$. The above setting and definition persist through our theoretical analysis.

[1] $z^{\hookrightarrow} \in \mathcal{Z} - \mathcal{Z} \not\hookrightarrow$ implies that $|g^{-1}(z^{\hookrightarrow})| = 1$ and therefore, $g^{-1}(z^{\hookrightarrow})$ is determined uniquely. Note that there is no chance that $|g^{-1}(z^{\hookrightarrow})| = 0$ by the surjectivity of g.

We prepare some additional notations to state Theorem 3 and its proof more clearly and briefly. Recall that the single training domain e^* for the target task and the domains $\mathcal{E}_{ad}$ for the additional task play important roles in our problem setting (see Section 2.2). Throughout the section, the domains are abbreviated as follows. The single training domain $(X^{e^*}, Y^{e^*}) \in T_{all}$ for the target task is abbreviated by (X^*, Y^*). For the domains $\mathcal{E}_{ad}$ of the additional task with higher class labels, $\{X^e, Y^e\}_{e\in\mathcal{E}_{ad}}$ is abbreviated by a subclass $T_{ad} \subset T_{all}$. For a projection $\Phi : \mathcal{X} \to \mathbb{R}^{n_\Phi}$ with its range n_Φ variables, let $p^{*,\Phi} : \mathbb{R}^{n_\Phi} \to \mathcal{P}_\mathcal{Y}$ denote the conditional probability density functions (p.d.f.) of $P(Y^*|\Phi(X^*))$. With a slight abuse of notation, for any probability P_θ on $\mathcal{X} \times \mathcal{Y}$ and a projection Φ, the density function of the conditional distribution $P_\theta(Y|\Phi(X))$ is denoted by $p_\theta \circ \Phi$.

We add some additional explanations and interpretations about the definition (※). Throughout this section, the projection to the components of $\mathcal{X}_1$ is denoted by $\Phi^{\mathcal{X}_1}$, which is the desired projection to give the invariant predictor. From the condition of T_{all}, for the projection $\Phi^{\mathcal{X}_1}$, the conditional probability $P_{Y|\Phi^{\mathcal{X}_1}(X)}$ for *any* random variable $(X, Y) \in T_{all}$ is the same; namely, letting $p^I : \mathcal{X}_1 \to \mathcal{P}_\mathcal{Y}$ denote the conditional p.d.f. of the invariant predictor $P_{Y^I|X_1^I}$, we have

$$p^e \circ \Phi^{\mathcal{X}_1} = p^I \tag{10}$$

for any $(X^e, Y^e) \in T_{all}$, where p^e is the conditional p.d.f. of $P_{Y^e|\Phi^{\mathcal{X}_1}(X^e)}$.

We restate Theorem 3 as follows.

Theorem 6 (Theorem 3 in the main body, with some notation arrangements)**.** *Assume that all domains $T_{all} := \{(X^e, Y^e)\}_{e\in\mathcal{E}}$ are fixed as (※); namely,*

$$T_{all} := \left\{ (X, Y) : \textit{a random variable on } \mathcal{X} \times \mathcal{Y} \,\middle|\, P_{\Phi^{\mathcal{X}_1}(X),Y} = P_{X_1^I,Y^I} \right\}. \tag{11}$$

Additionally, assume that the following condition holds:

(A) For any projection Φ with $\mathrm{Im}\Phi_2 \neq \emptyset$, there exist $(X^{e_1}, Y^{e_1}), (X^{e_2}, Y^{e_2}) \in T_{ad}$ such that $P(g(Y^{e_1})|\Phi(X^{e_1})) \neq P(g(Y^{e_2})|\Phi(X^{e_2}))$.

Then, there exists $\lambda^* \in \mathbb{R}$ such that a minimizer $(\theta^\dagger, \theta_{ad}^\dagger, \Phi^\dagger)$ of the objective function

$$\min_{\theta,\theta_{ad},\Phi} \left\{ \mathcal{R}^{(X^*,Y^*)}(p_\theta \circ \Phi) + \lambda^* \cdot \sum_{(X^e,Y^e)\in T_{ad}} \|\nabla_{\hat{\theta}_{ad}=\theta_{ad}} \mathcal{R}^{(X^e,g(Y^e))}(p^{\mathcal{Z}|\mathcal{H}}_{\hat{\theta}_{ad}} \circ \Phi)\|^2 \right\} \tag{12}$$

is o.o.d. optimial, i.e.,

$$p_{\theta^\dagger} \circ \Phi^\dagger \in \operatorname*{argmin}_{p_\theta:\mathcal{X}\to\mathcal{P}_\mathcal{Y}} \mathcal{R}^{o.o.d.}(p_\theta),$$

where p_θ and $p^{\mathcal{Z}|\mathcal{H}}_{\theta_{ad}}$ in $\min_{\theta,\theta_{ad},\Phi}$ run all the p.d.f.s, and Φ runs all the variable selections. The gradient $\nabla_{\theta_{ad}}$ should be understood as the functional derivative on the space of p.d.f.

Before proving Theorem 6, we prepare one lemma, which asserts that, if $\mathrm{Im}\Phi_2 \neq \emptyset$, at least one domain in T_{ad} has non-trivial gradient:

Lemma 7.

$$\min_{\theta_{ad},\Phi:\mathrm{Im}\Phi_2\neq\emptyset} \sum_{(X^e,Y^e)\in T_{ad}} \|\nabla_{\hat{\theta}_{ad}=\theta_{ad}} \mathcal{R}^{(X^e,g(Y^e))}(p^{\mathcal{Z}|\mathcal{H}}_{\hat{\theta}_{ad}} \circ \Phi)\|^2 \;>\; 0.$$

Proof. It suffices to prove that, for any projection Φ with $\mathrm{Im}\Phi_2 \neq \emptyset$ and $p^{\mathcal{Z}|\mathcal{H}}_{\theta_{ad}}$, there is $(X^e, Y^e) \in T_{ad}$ such that $\|\nabla_{\hat{\theta}_{ad}=\theta_{ad}} \mathcal{R}^{(X^e,g(Y^e))}(p^{\mathcal{Z}|\mathcal{H}}_{\hat{\theta}_{ad}} \circ \Phi)\|^2 \neq 0$. We prove this by contradiction. Suppose that there exist a projection Φ with $\mathrm{Im}\Phi_2 \neq \emptyset$ and $p^{\mathcal{Z}|\mathcal{H}}_{\theta_{ad}}$ which satisfy

$$\|\nabla_{\hat{\theta}_{ad}=\theta_{ad}} \mathcal{R}^{(X^e,g(Y^e))}(p^{\mathcal{Z}|\mathcal{H}}_{\hat{\theta}_{ad}} \circ \Phi)\|^2 = 0 \quad (\forall (X^e, Y^e) \in T_{ad}).$$

From Assumption (A), take (X^{e_1}, Y^{e_1}) and (X^{e_2}, Y^{e_2}) in T_{ad} such that $P(g(Y^{e_1})|\Phi(X^{e_1})) \neq P(g(Y^{e_2})|\Phi(X^{e_2}))$.

Note that the risk is defined by the cross-entropy loss:

$$\mathcal{R}^{(X^e,g(Y^e))}(p^{\mathcal{Z}|\mathcal{H}}_{\hat{\theta}_{ad}} \circ \Phi) = -\int \log p^{\mathcal{Z}|\mathcal{H}}_{\hat{\theta}_{ad}}(g(Y^e)|\Phi(X^e)) dP_{X^e,Y^e}.$$

It is well known that this is minimized in the space of probability distributions if and only if $p^{\mathcal{Z}|\mathcal{H}}_{\hat{\theta}_{ad}}$ equals to $P(Y^e|\Phi(X^e))$. From $\|\nabla_{\hat{\theta}_{ad}=\theta_{ad}} \mathcal{R}^{(X^e,g(Y^e))}(p^{\mathcal{Z}|\mathcal{H}}_{\hat{\theta}_{ad}} \circ \Phi)\|^2 = 0$ for (X^{e_1}, Y^{e_1}) and (X^{e_2}, Y^{e_2}), we can conclude that $p^{\mathcal{Z}|\mathcal{H}}_{\theta_{ad}}$ should equal to both of $P(g(Y^{e_1})|\Phi(X^{e_1}))$ and $P(g(Y^{e_2})|\Phi(X^{e_2}))$. This contradicts with the assumption $P(g(Y^{e_1})|\Phi(X^{e_1})) \neq P(g(Y^{e_2})|\Phi(X^{e_2}))$. □

Proof of Theorem 6
Let Φ^{id} denote the identity map of $\mathcal{X}$. Define the constants C_1, C_2, and C_3 by

$$C_1 := \mathcal{R}^{(X^*,Y^*)}(p^{*,\Phi^{id}} \circ \Phi^{id}) = H(Y^*|X^*),$$
$$C_2 := \mathcal{R}^{(X^*,Y^*)}(p^{*,\Phi^I} \circ \Phi^I) = H(Y^*|X_1^*) = H(Y^I|X_1^I),$$
$$C_3 := \frac{C_2 - C_1}{\min_{\theta_{ad},\Phi:\mathrm{Im}\Phi_2 \neq \emptyset} \sum_{(X^e,Y^e)\in T_{ad}} \|\nabla_{\hat{\theta}_{ad}=\theta_{ad}} \mathcal{R}^{(X^e,g(Y^e))}(p^{\mathcal{Z}|\mathcal{H}}_{\hat{\theta}_{ad}} \circ \Phi)\|^2},$$

where $H(Y^I|X_1^I)$ and $H(Y^*|X^*)$ denote the conditional entropy. Note that C_3 is well-defined because of the positivity result of Lemma 7.

Take λ^* such that $\lambda^* > C_3$. For notational simplicity, the objective function (12) is denoted by $O(\theta, \theta_{ad}, \Phi)$; namely,

$$O(\theta,\theta_{ad},\Phi) := \mathcal{R}^{(X^*,Y^*)}(p_\theta \circ \Phi) + \lambda^* \cdot \sum_{(X^e,Y^e)\in T_{ad}} \|\nabla_{\hat{\theta}_{ad}=\theta_{ad}} \mathcal{R}^{(X^e,g(Y^e))}(p^{\mathcal{Z}|\mathcal{H}}_{\hat{\theta}_{ad}} \circ \Phi)\|^2.$$

We prove the theorem in three steps.

Step 1 $\min_{p:\mathcal{X}\to\mathcal{P}_\mathcal{Y}} \mathcal{R}^{o.o.d.}(p) = H(Y^I|X_1^I)$.

proof of Step 1
We will prove $p^I \in \operatorname*{argmin}_{p:\mathcal{X}\to\mathcal{P}_\mathcal{Y}} \mathcal{R}^{o.o.d.}(p)$. From the definition

$$\mathcal{R}^{o.o.d.}(p) = \max_{(X^e,Y^e)\in T_{all}} -\int \log p(Y^e|X^e) dP_{Y^e,X^e},$$

$p^I \in \operatorname*{argmin}_{p:\mathcal{X}\to\mathcal{P}_\mathcal{Y}} \mathcal{R}^{o.o.d.}(p)$ holds if and only if

$$\max_{(X^e,Y^e)\in T_{all}} -\int \log p_\theta(Y^e|X^e) dP_{Y^e,X^e} \geq \max_{(X^e,Y^e)\in T_{all}} -\int \log p^I(Y^e|X_1^e) dP_{Y^e,X^e}$$

for any $p_\theta : \mathcal{X} \to \mathcal{P}_\mathcal{Y}$. Note that, as discussed before Theorem 6, for any $(X^e, Y^e) \in T_{all}$, we have $P_{Y^e|X_1^e} = P_{Y^I|X_1^I}$. Then, it suffices to prove that for any p_θ there exists $(X^{e'}, Y^{e'}) \in T_{all}$ such that

$$\int -\log p_\theta(Y^{e'}|X^{e'}) dP_{Y^{e'},X^{e'}} \geq \int -\log p^I(Y^I|X_1^I) dP_{X^e,Y^e}. \tag{13}$$

Define $(X^{e'}, Y^{e'}) \in T_{all}$ such that its distribution is the direct product $P_{X_1^I,Y^I} \otimes P_{X_2^{e'}}$, where $P_{X_2^{e'}}$ is an arbitrary distribution on $\mathcal{X}_2$. In this case, the left hand side of (13) is given by

$$\begin{aligned}\int -\log p_\theta(Y^{e'}|X^{e'}) dP_{Y^{e'},X^{e'}} &= \int -\log p_\theta(Y^{e'}|X_1^{e'}, X_2^{e'}) dP_{Y^{e'},X^{e'}} \\ &= \int dP_{X_2^{e'}} \int -\log p_\theta(Y^I|X_1^I, X_2^{e'}) dP_{X_1^I,Y^I}. \end{aligned} \tag{14}$$

We can see that, for any $x_2 \in \mathcal{X}_2$, the inequality

$$\int -\log p_\theta(Y^I|X_1^I, X_2^{e'} = x_2)) dP_{X_1^I,Y^I} \geq \int -\log p^I(Y^I|X_1^I) dP_{X_1^I,Y^I}$$

holds, since the minimum of the cross entropy loss is attained at the conditional p.d.f. p^I. Integrating this inequality with $P_{X_2^{e'}}$, we have

$$\int dP_{X_2^e} \int -\log p_\theta(Y^I|X_1^I, X_2^e) dP_{X_1^I,Y^I} \geq \int -\log p^I(Y^I|X_1^I) dP_{X_1^I,Y^I}. \tag{15}$$

Eqs. (14) and (15) show (13), from which the assertion is obtained by $-\int \log p^I(Y^I|X_1^I) dP_{X_1^I,Y^I} = H(Y^I|X_1^I)$.

Step 2 Any minimizer of the objective function,

$$(\theta^\dagger, \theta_{ad}^\dagger, \Phi^\dagger) \in \operatorname*{argmin}_{\theta,\theta_{ad},\Phi} O(\theta,\theta_{ad},\Phi),$$

satisfies $\mathrm{Im}\Phi_2^\dagger = \emptyset$.

proof of Step 2
It suffices to prove that $\min_{\Phi:\mathrm{Im}\Phi_2=\emptyset,\theta,\theta_{ad}} O(\theta,\theta_{ad},\Phi) < \min_{\Phi:\mathrm{Im}\Phi_2\neq\emptyset,\theta,\theta_{ad}} O(\theta,\theta_{ad},\Phi)$. First, we have

$$
\begin{aligned}
&\min_{\Phi:\mathrm{Im}\Phi_2\neq\emptyset,\theta,\theta_{ad}} O(\theta,\theta_{ad},\Phi)\\
&= \min_{\Phi:\mathrm{Im}\Phi_2\neq\emptyset,\theta,\theta^{ad}} \left\{ \mathcal{R}^{(X^*,Y^*)}(p_\theta\circ\Phi) + \lambda^* \cdot \sum_{(X^e,Y^e)\in T_{ad}} \|\nabla_{\hat{\theta}_{ad}=\theta_{ad}} \mathcal{R}^{(X^e,g(Z^e))}(p^{\mathcal{Z}|\mathcal{H}}_{\hat{\theta}_{ad}}\circ\Phi)\|^2 \right\}.\\
&> \min_{\Phi:\mathrm{Im}\Phi_2\neq\emptyset,\theta,\theta_{ad}} \Bigg\{ \mathcal{R}^{(X^*,Y^*)}(p_\theta\circ\Phi) + \frac{C_2-C_1}{\min_{\theta_{ad},\Phi:\mathrm{Im}\Phi_2\neq\emptyset}\sum_{(X^e,Y^e)\in T_{ad}} \|\nabla_{\hat{\theta}_{ad}=\theta_{ad}} \mathcal{R}^{(X^e,g(Y^e))}(p^{\mathcal{Z}|\mathcal{H}}_{\hat{\theta}_{ad}}\circ\Phi)\|^2}\\
&\qquad\qquad \times \sum_{(X^e,Y^e)\in T_{ad}} \|\nabla_{\hat{\theta}_{ad}=\theta_{ad}} \mathcal{R}^{(X^e,g(Y^e))}(p^{\mathcal{Z}|\mathcal{H}}_{\hat{\theta}_{ad}}\circ\Phi)\|^2 \Bigg\}\\
&\geq \min_{\Phi:\mathrm{Im}\Phi_2\neq\emptyset,\theta,\theta_{ad}} \{\mathcal{R}^{(X^*,Y^*)}(p_\theta\circ\Phi) + C_2 - C_1\}\\
&= \min_{\Phi:\mathrm{Im}\Phi_2\neq\emptyset,\theta,\theta_{ad}} \{\mathcal{R}^{(X^*,Y^*)}(p_\theta\circ\Phi)\} + C_2 - C_1\\
&\geq \mathcal{R}^{(X^*,Y^*)}(p^{*,\Phi^{id}}\circ\Phi^{id}) + C_2 - C_1 = C_2.
\end{aligned}
$$

On the other hand, by taking $\Phi=\Phi^I$, we obtain

$$
\begin{aligned}
&\min_{\Phi:\mathrm{Im}\Phi_2=\emptyset,\theta,\theta_{ad}} O(\theta,,\theta_{ad},\Phi)\\
&\qquad\qquad \leq \mathcal{R}^{(X^*,Y^*)}(p^I) + \lambda^* \cdot \sum_{(X^e,Y^e)\in T_{ad}} \|\nabla_{\hat{\theta}_{ad}=\theta^*_{ad}} \mathcal{R}^{(X^e,g(Y^e))}(p^{\mathcal{Z}|\mathcal{H}}_{\hat{\theta}_{ad}}\circ\Phi^I)\|^2.
\end{aligned}
$$

Since $p^{\mathcal{Z}|\mathcal{H}}_{\hat{\theta}_{ad}}\circ\Phi^I = p^I(g(Y^I)|X^I_1)$ does not depend on θ_{ad}, the gradient is zero, and therefore

$$
\min_{\Phi:\mathrm{Im}\Phi_2=\emptyset,\theta,\theta_{ad}} O(\theta,,\theta_{ad},\Phi) \leq \mathcal{R}^{(X^*,Y^*)}(p^I) = C_2.
$$

We thus obtain

$$
\min_{\Phi:\mathrm{Im}\Phi_2=\emptyset,\theta,\theta_{ad}} O(\theta,\theta_{ad},\Phi) \leq C_2 < \min_{\Phi:\mathrm{Im}\Phi_2\neq\emptyset,\theta,\theta_{ad}} O(\theta,\theta_{ad},\Phi),
$$

which completes the proof.

Step 3 If $(p_{\theta^\dagger}, p^{\mathcal{Z}|\mathcal{H}}_{\theta^\dagger_{ad}}, \Phi^\dagger) \in \operatorname{argmin}_{\theta,\theta_{ad},\Phi} O(\theta,\theta^{ad},\Phi)$, then $\mathcal{R}^{o.o.d.}(p_{\theta^\dagger}\circ\Phi^\dagger) = H(Y^I|X^I_1)$.

proof of Step 3
From Step 1, we have $H(Y^I|X^I_1) \leq \mathcal{R}^{o.o.d.}(p_{\theta^\dagger}\circ\Phi^\dagger)$. We will probe the converse inequality.

From Step 2, we have $\mathrm{Im}\Phi^\dagger_2 = \emptyset$. This tells $\mathcal{R}^{o.o.d.}(p_{\theta^\dagger}\circ\Phi^\dagger) = \mathcal{R}^{e^*}(p_{\theta^\dagger}\circ\Phi^\dagger)$, since $P_{X_1,Y}$ are the same for all elements in T_{all}. Therefore,

$$
\begin{aligned}
&\mathcal{R}^{o.o.d.}(p_{\theta^\dagger}\circ\Phi^\dagger) = \mathcal{R}^{(X^*,Y^*)}(p_{\theta^\dagger}\circ\Phi^\dagger)\\
&\leq \min_{\theta_{ad}} \left\{ \mathcal{R}^{(X^*,Y^*)}(p_{\theta^\dagger}\circ\Phi^\dagger) + \lambda^* \cdot \sum_{(X^e,Y^e)\in T_{ad}} \|\nabla_{\hat{\theta}_{ad}=\theta^*_{ad}} \mathcal{R}^{(X^e,g(Y^e))}(p^{\mathcal{Z}|\mathcal{H}}_{\hat{\theta}_{ad}}\circ\Phi)\|^2 \right\}\\
&= \min_{\Phi,\theta,\theta_{ad}} \left\{ \mathcal{R}^{(X^*,Y^*)}(p_\theta\circ\Phi) + \lambda^* \cdot \sum_{(X^e,Y^e)\in T_{ad}} \|\nabla_{\hat{\theta}_{ad}=\theta_{ad}} \mathcal{R}^{(X^e,g(Y^e))}(p^{\mathcal{Z}|\mathcal{H}}_{\hat{\theta}_{ad}}\circ\Phi)\|^2 \right\}\\
&\leq C_2 = H(Y^I|X^I_1).
\end{aligned}
$$

Final step for the proof of Theorem 6
For $(\theta^\dagger,\theta^\dagger_{ad},\Phi^\dagger) \in \operatorname{argmin}_{\theta,\theta_{ad},\Phi} O(\theta,\theta_{ad},\Phi)$, Step 1 and Step 3 show

$$
\mathcal{R}^{o.o.d.}(p_{\theta^\dagger}\circ\Phi^\dagger) = H(Y^I|X^I_1) = \min_{p:\theta\to\mathcal{P}_\mathcal{Y}} \mathcal{R}^{o.o.d.}(p),
$$

which completes the proof.

C Proof of Theorem 4

Before the proof, let us rearrange some notations introduced in Section 4.2. Notations are the same as in Appendix B. Recall that we assume, given hyperparameter λ, the minimization of (5) achieves the global optimum perfectly, which yields the projection (variable selection) $\Phi^\lambda(x) : \mathcal{X} \to \mathbb{R}^{n_\lambda}$ ($n_\lambda \leq n_1 + n_2$) and the conditional p.d.f. of $P_{Y^{e^*}|\Phi^\lambda(X^{e^*})}$, denoted by $p^{*,\lambda}(y|\Phi^\lambda(x))$. The $\mathcal{X}_1$ and $\mathcal{X}_2$ components of $\Phi^\lambda(X)$ are denoted by $\Phi_1^\lambda(X)$ and $\Phi_2^\lambda(X)$, respectively.

We rephrase the o.o.d. risk (1) and its evaluation (6) by Method I with some notational rearrangements. For $\lambda \in \Lambda$ and the training variable (X^*, Y^*) for the target task, the conditional p.d.f. of $P(Y^*|\Phi^\lambda(X^*))$ given the selected variables is denoted by $p^{*,\lambda} : \mathbb{R}^{n_\lambda} \to \mathcal{P}_\mathcal{Y}$. Then, the the o.o.d. risk $\mathcal{R}^{o.o.d.}(\lambda)$ of $p^{*,\lambda} \circ \Phi^\lambda$ and its evaluation $\mathcal{R}^I(\lambda)$ ((6) in the main body) are represented as

$$\mathcal{R}^{o.o.d.}(\lambda) := \max_{(X,Y)\in T_{all}} \mathcal{R}^{(X,Y)}(p^{*,\lambda} \circ \Phi^\lambda),$$

$$\mathcal{R}^I(\lambda) := \max\left\{ \max_{(X,Y)\in T_{ad}} \mathcal{R}^{(X,g(Y))}(p^{*,\lambda} \circ \Phi^\lambda), \mathcal{R}^{(X^*,Y^*)}(p^{*,\lambda} \circ \Phi^\lambda) \right\},$$

respectively. We restate Theorem 4 with some notation arrangements:

Theorem 8 (Theorem 4 in the main body, with some notational arrangements)**.** *Assume that all domains $T_{all} := \{(X^e, Y^e)\}_{e\in\mathcal{E}}$ are fixed as (※) in Appendix B; namely,*

$$T_{all} := \left\{ (X, Y) : \textit{a random variable on } \mathcal{X} \times \mathcal{Y} \,\middle|\, P_{\Phi^{\mathcal{X}_1}(X),Y} = P_{X_1^I, Y^I} \right\}.$$

Additionally, assume the following two conditions:

(I) there is $\lambda^I \in \Lambda$ such that $\Phi^{\lambda^I} = \Phi^{\mathcal{X}_1}$, where $\Phi^{\mathcal{X}_1}$ is the projection to the $\mathcal{X}_1$-components.

(II) Let p^* be the p.d.f of $P_{X^*,g(Y^*)}$. For any λ with $\mathrm{Im}\Phi_2^\lambda \neq \emptyset$, there is $(X^{e_\lambda}, Y^{e_\lambda}) \in T_{ad}$ such that

$$(x, z) \sim P_{X^{e_\lambda},g(Y^{e_\lambda})} \textit{ satisfies } p^*(z|\Phi^\lambda(x)) \leq e^{-\beta} - \varepsilon \textit{ with probability } 1 \textit{ in } P_{X^{e_\lambda},g(Y^{e_\lambda})}.$$

Here, $\varepsilon \in \mathbb{R}_{>0}$ is a sufficiently small positive real number (that is, $0 < \varepsilon \ll 1$) and $\beta := H(Y^*|(X_1^*))$ is the conditional entropy of $((X_1^*), Y^*)$. Then, we have

$$\operatorname*{argmin}_{\lambda\in\Lambda} \mathcal{R}^I(\lambda) \subset \operatorname*{argmin}_{\lambda\in\Lambda} \mathcal{R}^{o.o.d.}(\lambda).$$

To prove Theorem 8, we prepare three lemmas, in which the notations are the same as in Theorem 8 and conditions (I) and (II) in Theorem 8 are also imposed.

Lemma 9. $\lambda^I \in \operatorname*{argmin}_{\lambda\in\Lambda} \mathcal{R}^{o.o.d.}(\lambda)$.

Lemma 10. *If $\hat{\lambda} \in \operatorname*{argmin}_{\lambda\in\Lambda} \mathcal{R}^I(\lambda)$, then $\mathrm{Im}\Phi_2^{\hat{\lambda}} = \emptyset$.*

Lemma 11. *If $\hat{\lambda} \in \Lambda$ satisfies $\mathrm{Im}\Phi_2^{\hat{\lambda}} = \emptyset$, then $\mathcal{R}^I(\hat{\lambda}) = \mathcal{R}^{o.o.d.}(\hat{\lambda})$.*

we prove Theorem 8 based on the above lemmas, before proving them.

proof of Theorem 8.
Take $\hat{\lambda} \in \operatorname{argmin} \mathcal{R}^I(\lambda)$. Then, $\mathrm{Im}\Phi_2^{\hat{\lambda}} = \emptyset$ holds by Lemma 10 and therefore, $\mathcal{R}^I(\hat{\lambda}) = \mathcal{R}^{o.o.d.}(\hat{\lambda})$ holds by Lemma 11. Moreover, $\mathcal{R}^{o.o.d.}(\hat{\lambda}) \geq \mathcal{R}^{o.o.d.}(\lambda^I)$ holds by Lemma 9 and $\mathcal{R}^{o.o.d.}(\lambda^I) = \mathcal{R}^I(\lambda^I)$ holds by Lemma 11 (since Φ^{λ^I} is the projection onto $\mathcal{X}_1$, $\mathrm{Im}\Phi_2^{\lambda^I} = \emptyset$). By the assumption $\hat{\lambda} \in \operatorname*{argmin}_{\lambda\in\Lambda} \mathcal{R}^I(\lambda)$, $\mathcal{R}^I(\lambda^I) \geq \mathcal{R}^I(\hat{\lambda})$ holds. Arranging these inequalities, we obtain

$$\mathcal{R}^I(\hat{\lambda}) = \mathcal{R}^{o.o.d.}(\hat{\lambda}) \geq \mathcal{R}^{o.o.d.}(\lambda^I) = \mathcal{R}^I(\lambda^I) \geq \mathcal{R}^I(\hat{\lambda}), \tag{16}$$

in which the inequalities must be equalities. Hence, we obtain $\mathcal{R}^{o.o.d.}(\hat{\lambda}) = \mathcal{R}^{o.o.d.}(\lambda^I)$. Because λ^I achieves the minimum of $\mathcal{R}^{o.o.d.}$ (Lemma 9), so does $\hat{\lambda}$, which concludes the proof. □

Since Lemma 9 is proven in the proof of Theorem 6 (especially, the proof in Step 1), we may prove the others.

proof of Lemma 10.
Let us prove the contraposition of Lemma 10. Take $\hat{\lambda} \in \Lambda$ with $\mathrm{Im}\Phi_2^{\hat{\lambda}} \neq \emptyset$. To prove that $\hat{\lambda} \notin \operatorname{argmin} \mathcal{R}^I(\lambda)$,

we may prove that $\mathcal{R}^I(\hat{\lambda}) > \mathcal{R}^I(\lambda^I)$ since $\lambda^I \in \Lambda$ (Assumption (I) in the statement). It then suffices to prove the following:

$$\text{there exists } (\bar{X}, \bar{Y}) \in T_{ad} \text{ such that } \int -\log p^{*,\hat{\lambda}}(g(\bar{Y})|\Phi^{\hat{\lambda}}(\bar{X}))dP_{\bar{X},g(\bar{Y})} > \mathcal{R}^I(\lambda^I). \tag{17}$$

From Condition (II), we can take $(X^{e_{\hat{\lambda}}}, Y^{e_{\hat{\lambda}}}) \in T_{ad}$ such that

$$(x, z) \sim P_{X^{e_{\hat{\lambda}}},g(Y^{e_{\hat{\lambda}}})} \text{ satisfies } p^*(z|\Phi^{\hat{\lambda}}(x)) \leq e^{-\beta} - \varepsilon \text{ with probability 1.}$$

To prove (17), we prepare one supplementary inequality:

Supplementary Inequality

$$\int -\log p^{*,\hat{\lambda}}\big(g(Y^{e_{\hat{\lambda}}})|\Phi^{\hat{\lambda}}(X^{e_{\hat{\lambda}}})\big)dP_{X^{e_{\hat{\lambda}}},g(Y^{e_{\hat{\lambda}}})} \geq -\log\left\{e^{-\beta} - \epsilon\right\}.$$

This inequality can be easily seen; from the way of taking $e_{\hat{\lambda}}$, we have

$$-\log p^*(z|\Phi^{\hat{\lambda}}(x)) \geq -\log\{e^{-\beta} - \varepsilon\}$$

with probability 1 with respect to $(x, z) \sim P_{X^{e_{\hat{\lambda}}},g(Y^{e_{\hat{\lambda}}})}$, and thus the integration proves the inequality.

Proof of Inequality (17).
It follows from the above supplementary inequality that

$$\int -\log p^{*,\hat{\lambda}}(g(Y^{e_{\hat{\lambda}}})|\Phi^{\hat{\lambda}}(X^{e_{\hat{\lambda}}}))dP_{X^{e_{\hat{\lambda}}},g(Y^{e_{\hat{\lambda}}})} \geq -\log\left\{e^{-\beta} - \epsilon\right\} > \beta = H(Y^*|X_1^*). \tag{18}$$

Since $\Phi^{\lambda^I} = \Phi^{\mathcal{X}_1}$ by Condition (I), the discussion at (10) tells that $\mathcal{R}^I(\lambda^I) = H(Y^I|X_1^I) = H(Y^*|X_1^*)$, which concludes (17) and the proof.

Proof of Lemma 11.
Take $\hat{\lambda} \in \Lambda$ that satisfies $\mathrm{Im}\Phi_2^{\hat{\lambda}} = \emptyset$. Then, $P_{\Phi^{\hat{\lambda}}(X),Y} = P_{\Phi^{\hat{\lambda}}(X^I),Y^I}$ holds for any $(X, Y) \in T_{all}$ because of $P_{X_1,Y} = P_{X_1^I,Y}$, and therefore, $\mathcal{R}^{(X,g(Y))}(p^{*,\hat{\lambda}} \circ \Phi^{\hat{\lambda}}) = \mathcal{R}^{(X^I,g(Y^I))}(p^{*,\hat{\lambda}} \circ \Phi^{\hat{\lambda}})$ and $\mathcal{R}^{(X^*,Y^*)}(p^{*,\hat{\lambda}} \circ \Phi^{\hat{\lambda}}) = \mathcal{R}^{(X^I,Y^I)}(p^{*,\hat{\lambda}} \circ \Phi^{\hat{\lambda}})$ hold. These two equalities lead the following equality:

$$\begin{aligned}\mathcal{R}^I(\hat{\lambda}) &= \max\left\{\max_{(X,Y)\in T_{ad}} \mathcal{R}^{(X,g(Y))}(p^{*,\hat{\lambda}} \circ \Phi^{\hat{\lambda}}), \mathcal{R}^{(X^*,Y^*)}(p^{*,\hat{\lambda}} \circ \Phi^{\hat{\lambda}})\right\} \\ &= \max\left\{\mathcal{R}^{(X^I,g(Y^I))}(p^{*,\hat{\lambda}} \circ \Phi^{\hat{\lambda}}), \mathcal{R}^{(X^I,Y^I)}(p^{*,\hat{\lambda}} \circ \Phi^{\hat{\lambda}})\right\}\end{aligned} \tag{19}$$

It follows from Theorem 2 that

$$\begin{aligned}&R^{(X^I,Y^I)}(p^{*,\hat{\lambda}} \circ \Phi^{\hat{\lambda}}) \\ &= \mathcal{R}^{(X^I,g(Y^I))}(p^{*,\hat{\lambda}} \circ \Phi^{\hat{\lambda}}) \\ &\quad + \sum_{z^{\not\prec} \in \mathcal{Z}^{\not\prec}} P(Y^I = g^{-1}(z^{\not\prec})) \int -\log p^{*,\hat{\lambda}}(Y^I|\Phi^{\hat{\lambda}}(X^I), g(Y^I) = z^{\not\prec})dP_{X^I,Y^I|g(Y^I)=z^{\not\prec}} \\ &\geq \mathcal{R}^{(X^I,g(Y^I))}(p^{*,\hat{\lambda}} \circ \Phi^{\hat{\lambda}})\end{aligned}$$

Therefore, from (19), we have $\mathcal{R}^I(\hat{\lambda}) = \mathcal{R}^{(X^I,Y^I)}(p^{*,\hat{\lambda}} \circ \Phi^{\hat{\lambda}})$. Since $P_{\Phi^{\hat{\lambda}}(X),Y}$ are the same for any elements in T_{all}, we obtain

$$\mathcal{R}^{(X^I,Y^I)}(p^{*,\hat{\lambda}} \circ \Phi^{\hat{\lambda}}) = \max_{(X,Y)\in T_{all}} \mathcal{R}^{(X,Y)}(p^{*,\hat{\lambda}} \circ \Phi^{\hat{\lambda}}) = \mathcal{R}^{o.o.d.}(p^{*,\hat{\lambda}} \circ \Phi^{\hat{\lambda}}),$$

which concludes the proof.

D Proof of Theorem 5

Before proving Theorem 5, we rephrase the evaluation (7) of the o.o.d. risk by Method II with some notation rearrangements. By using notation simplifications in Appendices B and C, the evaluation $\mathcal{R}^{II}(\lambda)$ by method II (corresponding to (7) in the main body) is represented as

$$\mathcal{R}^{II}(\lambda) := \max_{(X,Y)\in T_{ad}\cup\{(X^*,Y^*)\}} \left\{\mathcal{R}^{(X,g(Y))}(p^{*,\lambda} \circ \Phi^{\lambda}) + D_{\lambda}(Y)\right\},$$

where the correction term $D_\lambda(Y)$ is defined by

$$D_\lambda(Y) := \sum_{z^{\not\prec} \in \mathcal{Z}^{\not\prec}} \left\{ P(g(Y) = z^{\not\prec}) \int -\log p^{*,\lambda}(Y^*|\Phi^\lambda(X^*), g(Y^*) = z^{\not\prec}) dP_{(X^*,Y^*)|g(Y^*)=z^{\not\prec}} \right\}$$

Note that, in $D_\lambda(Y)$, although the random variable Y is given by $(X, Y) \in T_{all}$, the marginal distributions of Ys are the same by the assumption of T_{all}. Thus, hereafter, we use D_λ for the notation, and

$$D_\lambda = \sum_{z^{\not\prec} \in \mathcal{Z}^{\not\prec}} \left\{ P(g(Y^*) = z^{\not\prec}) \int -\log p^{*,\lambda}(Y^*|\Phi^\lambda(X^*), g(Y^*) = z^{\not\prec}) dP_{(X^*,Y^*)|g(Y^*)=z^{\not\prec}} \right\}.$$

Note also that $\beta_\lambda = H(Y^*|X_1^*) - D_\lambda$. We restate Theorem 5 with some notation arrangements:

Theorem 12 (Theorem 5 in the main body, with some notation arrangements)**.** *Assume that all domains $T_{all} := \{(X^e, Y^e)\}_{e\in\mathcal{E}}$ are fixed as (※) in Appendix B; namely,*

$$T_{all} := \left\{ (X, Y) : \textit{a random variable on } \mathcal{X} \times \mathcal{Y} \,\middle|\, P_{\Phi^{\mathcal{X}_1}(X),Y} = P_{X_1^I,Y^I} \right\}.$$

Notations are the same as in the statement of Theorem 8. In addition to the condition (I), assume the following condition (II)':

(II)' Let p^* be the p.d.f. of $P_{X^*,g(Y^*)}$. For any λ with $\mathrm{Im}\Phi_2^\lambda \neq \emptyset$, there is $(X^{e_\lambda}, Y^{e_\lambda}) \in T_{ad}$ such that

$$(x, z) \sim P_{X^{e_\lambda},g(Y^{e_\lambda})} \textit{ satisfies } p^*(z|\Phi^\lambda(x)) \leq e^{-\beta_\lambda} - \varepsilon \textit{ with probability } 1 \textit{ in } P_{X^{e_\lambda},g(Y^{e_\lambda})}.$$

Here, ε is some positive real number and

$$\beta_\lambda := H(Y^*|X_1^*) - D_\lambda(Y^*).$$

Then, we have

$$\operatorname*{argmin}_{\lambda\in\Lambda} \mathcal{R}^{II}(\lambda) \subset \operatorname*{argmin}_{\lambda\in\Lambda} \mathcal{R}^{o.o.d.}(\lambda).$$

We first show lemmas before the proof of the theorem.

Lemma 13. *If $\hat{\lambda} \in \operatorname*{argmin}_{\lambda\in\Lambda} \mathcal{R}^{II}(\lambda)$, then $\mathrm{Im}\Phi_2^{\hat{\lambda}} = \emptyset$.*

Lemma 14. *If $\hat{\lambda} \in \Lambda$ satisfies $\mathrm{Im}\Phi_2^{\hat{\lambda}} = \emptyset$, then $\mathcal{R}^{II}(\hat{\lambda}) = \mathcal{R}^{o.o.d.}(\hat{\lambda})$.*

proof of Theorem 12
Combining the above two lemmas and Lemma 9, we can derive the required assertion in essentially the same manner as in the proof of Theorem 8.

proof of Lemma 13.
Let us prove the contraposition of Lemma 13. Take $\hat{\lambda} \in \Lambda$ with $\mathrm{Im}\Phi_2^{\hat{\lambda}} \neq \emptyset$. To prove that $\hat{\lambda} \notin \operatorname{argmin} \mathcal{R}^{II}(\lambda)$, we may prove that $\mathcal{R}^{II}(\hat{\lambda}) > \mathcal{R}^{II}(\lambda^I)$ since $\lambda^I \in \Lambda$ (Assumption (I) in the statement). To show this, it suffices to prove the following statement:

$$\text{there is } (\bar{X}, \bar{Y}) \in T_{ad} \text{ such that } \mathcal{R}^{(\bar{X},g(\bar{Y}))}(\hat{\lambda}) + D_{\hat{\lambda}} > \mathcal{R}^{II}(\lambda^I). \tag{20}$$

Take $(X^{e_{\hat{\lambda}}}, Y^{e_{\hat{\lambda}}}) \in T_{ad}$ as in Condition (II)'. Then, in the same way as the proof of Lemma 10, we have the following inequality:

$$\int -\log p^{*,\hat{\lambda}}\big(g(Y^{e_{\hat{\lambda}}})|\Phi^{\hat{\lambda}}(X^{e_{\hat{\lambda}}})\big) dP_{X^{e_{\hat{\lambda}}},g(Y^{e_{\hat{\lambda}}})} \geq -\log\left\{e^{-\beta_{\hat{\lambda}}} - \epsilon\right\},$$

which leads us to obtain

$$\mathcal{R}^{(X^{e_{\hat{\lambda}}},g(Y^{e_{\hat{\lambda}}}))}(\hat{\lambda}) + D_{\hat{\lambda}} > \beta_{\hat{\lambda}} + D_{\hat{\lambda}} = H(Y^*|X_1^*) = \mathcal{R}^{(Y^*,X^*)}(p^{*,\lambda^I} \circ \Phi^{\lambda^I}). \tag{21}$$

On the other hand, for any $(X, Y) \in T_{all}$ the marginal distribution of $(Y, \Phi^I(X))$ is the same as that of (Y^*, X_1^*). Noting that $\mathcal{R}^{(X,g(Y))}(p^{*,\lambda^I} \circ \Phi^{\lambda^I}) + D_{\lambda^I}$ depends only on (Y, X_1), we have

$$\mathcal{R}^{II}(\lambda^I) = \mathcal{R}^{(X^*,g(Y^*))}(p^{*,\lambda^I} \circ \Phi^{\lambda^I}) + D_{\lambda^I}. \tag{22}$$

Now, Lemma 2 implies

$$\mathcal{R}^{(Y^*,X^*)}(p^{*,\lambda^I} \circ \Phi^{\lambda^I}) = \mathcal{R}^{(X^*,g(Y^*))}(p^{*,\lambda^I} \circ \Phi^{\lambda^I}) + D_{\lambda^I}. \tag{23}$$

From (21), (22), and (23), we thus have

$$\mathcal{R}^{(X^{e_{\hat{\lambda}}}, g(Y^{e_{\hat{\lambda}}}))}(p^{*,\lambda^I} \circ \Phi^{\lambda^I}) + D_{\hat{\lambda}} > \mathcal{R}^{II}(\lambda^I),$$

which shows (20) and completes the proof.

proof of Lemma 14.
Take $\hat{\lambda} \in \Lambda$ such that $\mathrm{Im}\Phi_2^{\hat{\lambda}} = \emptyset$. It follows from $\mathrm{Im}\Phi_2^{\hat{\lambda}} = \emptyset$ that $P_{\Phi^{\hat{\lambda}}(X),Y} = P_{\Phi^{\hat{\lambda}}(X^*),Y^*}$ holds for all $(X, Y) \in T_{all}$. Therefore,

$$\mathcal{R}^{o.o.d.}(\hat{\lambda}) = \max_{(X,Y)\in T_{all}} \mathcal{R}^{(X,Y)}(p^{*,\hat{\lambda}} \circ \Phi^{\hat{\lambda}}) = \mathcal{R}^{(X^*,Y^*)}(p^{*,\hat{\lambda}} \circ \Phi^{\hat{\lambda}}).$$

Likewise, from the condition of $\hat{\lambda}$, the definition of $\mathcal{R}^{II}(\hat{\lambda})$ involves the same distribution for $(Y, \Phi^{\hat{\lambda}}(X))$, and thus

$$\mathcal{R}^{II}(\hat{\lambda}) = \mathcal{R}^{(X^*,g(Y^*))}(p^{*,\hat{\lambda}} \circ \Phi^{\hat{\lambda}}) + D_{\hat{\lambda}}.$$

In a similar way to the proof of Lemma 13, Theorem 2 tells

$$\mathcal{R}^{(X^*,Y^*)}(p^{*,\hat{\lambda}} \circ \Phi^{\hat{\lambda}}) = \mathcal{R}^{(X^*,g(Y^*))}(p^{*,\hat{\lambda}} \circ \Phi^{\hat{\lambda}}) + D_{\hat{\lambda}}.$$

This completes the proof.

E Sufficient Conditions of (ii) and (ii)'

In the section, we reveal sufficient conditions of e^* for there to exist $(X^{e_\lambda}, Y^{e_\lambda})$ that satisfies (ii) and (ii)' in Theorems 4 and 5, respectively.

Theorem 15. *Notations are the same as in Theorem 8. Assume that (X^*, Y^*) satisfies the following condition:*

(A2) For a sufficiently small $\varepsilon \ll 1$, any λ with $\mathrm{Im}\Phi_2^\lambda \neq \emptyset$, any $a \in \mathrm{Im}\Phi_1^\lambda$, and any $b \in \mathcal{Y}$, there exists $c(\lambda, a, b)$[2] such that

$$P(Y^* = b | \Phi_1^\lambda(X^*) = a, \Phi_2^\lambda(X^*) = c) \geq (1 - e^{-\beta}) + \varepsilon.$$

Then, for any λ with $\mathrm{Im}\Phi_2^\lambda \neq \emptyset$, there exists $(X^{e_\lambda}, Y^{e_\lambda}) \in T_{all}$ such that the inequality in Theorem 8 (ii) holds.

Remark. The condition (A2) means that, in the domain $e = e^*$, the affection of domain-specific factors ($= \mathcal{X}_2$) to the response variable Y^{e^*} is large; indeed, the inequality in (A2) means that, if λ fails to remove domain-specific factors ($i.e.$, $\mathrm{Im}\Phi_2^\lambda \neq \emptyset$), we can control the probability of $Y^{e^*} = b$ by selecting c for any $b \in \mathcal{Y}$. Note also that the inequality (A2) is a lower bound of the likelihood, while the condition in (ii), Theorem 8, is an upper bound of the likelihood. Although imposing an upper bound might look reasonable to reflect non-fitting of the projection Φ^λ, Theorem 15 shows that we can use lower bound as a sufficient condition.

Proof. Fix λ with $\mathrm{Im}\Phi_2^\lambda \neq \emptyset$. Take $(\bar{X}, \bar{Y}) \in T_{all}$ such that its probability measure corresponds to $\bar{P}_{X_2|Y,X_1} \times P_{Y^I,X_1^I}$, where $\bar{P}_{X_2|Y,X_1}$ is defined by, setting $\hat{c}(\hat{\lambda}, a, b)$ by

$$\hat{c}(\hat{\lambda}, a, b) \in \operatorname*{argmin}_{c\in\mathcal{X}_2} P(g(Y^*) = g(b) | \Phi_1^{\hat{\lambda}}(X^*) = \Phi_1^{\hat{\lambda}}(a), \Phi_2^{\hat{\lambda}}(X^*) = \Phi_2^{\hat{\lambda}}(c)),$$

$\bar{P}_{X_2|Y=b,X_1=a} := \delta_{X_2=\hat{c}(\hat{\lambda},a,b)}$. Here, for $c \in \mathcal{X}_2$, the probability measure $\delta_{X_2=c}$ on $\mathcal{X}_2$ denotes a Dirac measure at $c \in \mathcal{X}_2$.

Before proving Theorem 15, we prepare the following inequalities:
Supplementary Inequality 1.

$$\forall a \in \mathcal{X}_1, \forall b \in \mathcal{Y}, P\left(g(Y^*) = g(b) \,\middle|\, \Phi_1^{\hat{\lambda}}(X^*) = \Phi_1^{\hat{\lambda}}(a), \Phi_2^{\hat{\lambda}}(X^*) = \Phi_2^{\hat{\lambda}}\big(\hat{c}(\hat{\lambda}, a, b)\big)\right) \leq e^{-\beta} - \epsilon.$$

To see the fact, take $b^* \in \mathcal{Y}$ such that $g(b^*) \neq g(b)$[3]. Then, by the condition (ii) of Theorem 5 and $\mathrm{Im}\Phi_2^{\hat{\lambda}} \neq \emptyset$, there exists $c(\hat{\lambda}, a, b) \in \mathcal{X}_2$ such that

$$P\left(Y^* = b^* \,\middle|\, \Phi_1^{\hat{\lambda}}(X^*) = \Phi_1^{\hat{\lambda}}(a), \Phi_2^{\hat{\lambda}}(X^*) = \Phi_2^{\hat{\lambda}}\big(c(\hat{\lambda}, a, b)\big)\right) \geq 1 - e^{-\beta} + \epsilon.$$

[2] $c(\lambda, a, b)$ means $c \in \mathcal{X}_2$ is determined by given $\lambda \in \Lambda$, $a \in \mathcal{X}_1$, $b \in \mathcal{Y}$.

[3] Such b^* always exists by the following reason if $|\mathcal{Z}| \geq 2$ by the following reason. Take $\mathcal{Z} \ni z^* \neq g(b)$. By the surjectivity of g, $g^{-1}(z^*) \neq \emptyset$. Taking $b^* \in g^{-1}(z^*)$, $g(b^*) = z^* \neq g(b)$.

Therefore,

$$
\begin{aligned}
&P\left(g(Y^*)=g(b)\,\middle|\,\Phi_1^{\hat{\lambda}}(X^*)=\Phi_1^{\hat{\lambda}}(a),\Phi_2^{\hat{\lambda}}(X^*)=\Phi_2^{\hat{\lambda}}\big(\hat{c}(\hat{\lambda},a,b)\big)\right)\\
&=\min_{c\in\mathcal{X}_2}P\big(g(Y^*)=g(b)|\Phi_1^{\hat{\lambda}}(X^*)=\Phi_1^{\hat{\lambda}}(a),\Phi_2^{\hat{\lambda}}(X^*)=\Phi_2^{\hat{\lambda}}(c)\big)\\
&\le P\left(g(Y^*)=g(b)\,\middle|\,\Phi_1^{\hat{\lambda}}(X^*)=\Phi_1^{\hat{\lambda}}(a),\Phi_2^{\hat{\lambda}}(X^*)=\Phi_2^{\hat{\lambda}}\big(c(\hat{\lambda},a,b)\big)\right)\\
&=1-\sum_{\bar{z}\neq g(b)}P\left(g(Y^*)=\bar{z}\,\middle|\,\Phi_1^{\hat{\lambda}}(X^*)=\Phi_1^{\hat{\lambda}}(a),\Phi_2^{\hat{\lambda}}(X^*)=\Phi_2^{\hat{\lambda}}\big(c(\hat{\lambda},a,b)\big)\right)\\
&\le 1-P\left(g(Y^*)=g(b^*)\,\middle|\,\Phi_1^{\hat{\lambda}}(X^*)=\Phi_1^{\hat{\lambda}}(a),\Phi_2^{\hat{\lambda}}(X^*)=\Phi_2^{\hat{\lambda}}\big(c(\hat{\lambda},a,b)\big)\right)\\
&\le 1-P\left(Y^*=b^*\,\middle|\,\Phi_1^{\hat{\lambda}}(X^*)=\Phi_1^{\hat{\lambda}}(a),\Phi_2^{\hat{\lambda}}(X^*)=\Phi_2^{\hat{\lambda}}\big(c(\hat{\lambda},a,b)\big)\right)\\
&\le 1-(1-e^{-\beta}+\epsilon)\\
&\le e^{-\beta}-\epsilon.
\end{aligned}
$$

Proof of Theorem 15
We may prove that $P_{\bar{X},\bar{Y}}(A)=1$ where

$$
\left\{(x,y)\in\mathcal{X}\times\mathcal{Y}\,\middle|\,P\left(g(Y^*)=g(b)\,\middle|\,\Phi^{\hat{\lambda}}(X^*)=\Phi^{\hat{\lambda}}(x)\right)\le e^{-\beta}-\epsilon\right\}.
$$

Then,

$$
\begin{aligned}
P_{\bar{X},\bar{Y}}(A)&=\int 1_A dP_{\bar{X},\bar{Y}}=\int 1_A d(\bar{P}_{X_2|Y,X_1}\times P_{Y^I,X_1^I})\\
&=\int dP_{Y^I,X_1^I}\int 1_A d\bar{P}_{X_2|Y,X_1}=\int dP_{Y^I,X_1^I}(x_1,y)\delta_{X_2=\hat{c}(\hat{\lambda},x_1,y)}(A_{(x_1,y)})
\end{aligned}
$$

holds where $A_{(x_1,y)}:=\{x_2\in\mathcal{X}_2|((x_1,x_2),y)\in\mathcal{X}\times\mathcal{Y}\}$. By the Supplementary Inequality 1, $\hat{c}(\hat{\lambda},x_1,y)\in A_{(x_1,y)}$ holds and therefore, $\delta_{X_2=\hat{c}(\hat{\lambda},x_1,y)}(A_{(x_1,y)})=1$, which leads us to the equation $\int dP_{Y^I,X_1^I}(x_1,y)\delta_{X_2=\hat{c}(\hat{\lambda},x_1,y)}(A_{(x_1,y)})=1$.

□

Theorem 16. *Notations are same as in Theorem 8 and 12. (X^*,Y^*) satisfies the following condition:*

(A2)' For a sufficiently small $\varepsilon\ll 1$, the following statement holds:
$\forall\lambda$ with $\mathrm{Im}\Phi_2^\lambda\neq\emptyset$, $\forall\alpha\in\mathrm{Im}\Phi_1^\lambda$, $\forall b\in\mathcal{Y}$, $\exists c(\lambda,a,b)$ s.t. $P(Y^*=b|\Phi_1^\lambda(X^*)=a,\Phi_2^\lambda(X^*)=c)\ge(1-e^{-\beta_\lambda})+\varepsilon$.

Then, $\forall\lambda$ with $\mathrm{Im}\Phi_2^\lambda\neq\emptyset$, there exists $(X^{e_\lambda},Y^{e_\lambda})\in T_{all}$ such that the inequality in (ii)' holds.

The proof of Theorem 16 is essentially same as the one of Theorem 15 and therefore, we omit.

F The real-world feasibility of (ii) and (ii)'

In the subsection, we discuss the feasibility of (ii) and (ii)', and show these conditions are not necessarily strong.

First, we discuss the Condition (ii). Since $\beta=H(Y^e|\Phi^{\mathcal{X}_1}(X^e))$ is the conditional entropy, we have

$$0\le\beta\le\log|\mathcal{Y}|$$

and hence

$$\frac{1}{|\mathcal{Y}|}-\varepsilon\le e^{-\beta}-\varepsilon\le 1-\varepsilon$$

holds. We can see that Condition (ii) is weak if $e^{-\beta}-\varepsilon$ approaches 1, or if β is small. Recall that $\Phi^{\mathcal{X}_1}(X^e)$ is the bias-removed feature of X^e (digit of CMNIST, or object of ImageNet, for example). We can then expect that, in many real-world settings, $\beta=H(Y^e|\Phi^{\mathcal{X}_1}(X^e))$ is often small, since the bias-removed feature $\Phi^{\mathcal{X}_1}(X^e)$ should have a large amount of information on the labels. Condition (ii) is satisfied if the likelihood $p^{e^*}(z|\Phi^\lambda(x))$ evaluated at a random point $(x,z)\sim P_{X^e,g(Y^e)}$ is bounded by the large value $e^{-\beta}-\varepsilon$ for at least one $e\in\mathcal{E}_{ad}$, so that the inequality in (ii) is likely to hold. Noting that (ii)' is weaker than (ii), the feasibility of (ii)' is concluded from one of (ii).

G Additional Experiment

G.1 Additional Experiments of Colored MNIST in Section 6

Although the Colored MNIST experiment in Section 6 fixes its flip rate to 25%, we additionally demonstrate by changing its flip rate among $\{10\%, 15\%, 20\%, 25\%\}$.

Table 4: Test Acc. of Hierarchical Colored MNIST (5runs)

flip rate	Test Acc. on	Best possible	Oracle	ERM	FT	FF	DSAN	Ours + CV1	Ours + CV2	Ours+TDV
0.25	$e = 0.1$	.750	.715(.001)	.693(.001)	.676(.003)	.677(.002)	.593(.007)	.706(.005)	.664 (.013)	.690 (.008)
	$e = 0.9$			.433 (.004)	.250 (.020)	.248(.015)	.073(.003)	.753(.011)	**.618 (.018)**	.657(.008)
0.20	$e = 0.1$	.800	.769(.001)	.800(.001)	.727(.002)	.725(.004)	.639(.003)	.752(.006)	.721 (.015)	.745 (.007)
	$e = 0.9$			.525 (.004)	.368 (.029)	.364(.011)	.080(.004)	.576(.014)	**.685 (.019)**	.719 (.004)
0.15	$e = 0.1$	.850	.822(.000)	.802(.002)	.782(.006)	.786(.003)	.682(.002)	.806(.006)	.794 (.008)	.794 (.008)
	$e = 0.9$			.630 (.006)	.493 (.038)	.512(.019)	.091(.005)	.673(.006)	**.774 (.006)**	.774 (.006)
0.10	$e = 0.1$	.900	.872(.001)	.848(.002)	.827(.004)	.829(.003)	.593(.007)	.857(.005)	.842 (.008)	.834 (.001)
	$e = 0.9$			.719 (.004)	.611(.016)	.623(.021)	.073(.003)	.756(.007)	**.800 (.007)**	.821 (.006)

Table 5: Baselines of CV methods

	Tr-CV	LOD-CV
0.25	.702 (.002) .597 (.006)	.590 (.004) .460 (.197)
0.20	.754 (.004) .678 (.008)	.716 (.018) .692 (.010)
0.15	.801 (.016) .678 (.008)	.787 (.004) .774 (.006)
0.10	.854 (.005) .751 (.013)	.836 (.004) .819 (.008)

Table 6: Means and SEs of $\{$(Accuracy of TDV on $e = 0.9$) -(Accuracy of Each CV on $e = 0.9$) $\}$ (5runs).

	CV I	CV II	Tr-CV	LOD-CV
0.25	.051 (.053)	.040 (.017)	.163 (.006)	.197 (.205)
0.20	.143 (.012)	.034 (.017)	.132 (.008)	.023 (.018)
0.15	.102 (.006)	.000 (.000)	.102 (.007)	.003 (.002)
0.10	.065 (005)	.021 (.010)	.075 (.010)	.005 (.002)

Table 4 and 5 show that, among several CV methods, our method II keeps a high predictive performance regardless of flipping rates. Table 6 the difference between accuracies by TDV and each CV for the same data set with $e = 0.9$. The result verifies that CVII selects preferable hyperparameters with smaller errors.

G.2 Additional Experiments: Colored MNIST II

We conduct an additional Colored MNIST experiment, changing annotation and coloring rules from ones in Section 6. Setting $\mathcal{Y} = [3]$ and $\mathcal{Z} := [2]$, we aim to predict Y^e from digit image data X^e, which is in the three categories $0-2$ ($Y^e = 0$), 3 or 4 ($Y^e = 1$) and $5-9$ ($Y^e = 2$). The label is changed randomly to one of the rest with a some probability ranging from $\{10\%, 15\%, 20\%, 25\%\}$. The domain index $e \in [0.0, 1.0]$ controls the color of the digit; for $Y^e = 0, 1$, the digit is colored in red with probability e and for $Y^e = 2$ colored in green with probability e. In the experiment, $\mathcal{D}^{e^*} \sim P_{X^{0.1}, Y^{0.1}}$ is drawn with sample size $n^{e^*} = 5000$, and Y^e is predicted based on X^e for $e = 0.1$ and 0.9. Regarding Z^e, we consider the task where we predict $Z^e = 0$ for X^e in $0-2$ and $Z^e = 1$ for $3-9$ (that is, $g(0) = 0$ and $g(1) = g(2) = 1$). We obtain the final label by flipping with some probability. As the domain-specific factor, we color the digit red for $Z^e = 0$ with probability e and green for $Z^e = 1$ with probability e. We set $\mathcal{E}_{ad} = \{0.1, 0.3, 0.5, 0.7, 0.9\}$ with $n^e = 5000$ for $\forall e \in \mathcal{E}_{ad}$. We model Φ by a 3-layer neural net. With the maximum epoch 500, we select (t, λ_{after}) from 3×10 candidates with $t \in \{0, 100, 200\}$, $\lambda_{after} \in \{10^0, 10^1, ..., 10^9\}$ by each CV method.

Table 7 and 8 shows test accuracies for 2000 random samples in the domains $e = 0.1$ and $e = 0.9$. The results demonstrate that the proposed methods significantly outperform the others for $e = 0.9$. Among the

Table 7: Test Accuracy for Hierarchical Colored MNIST (5runs)

flip rate	Test Acc. on	Best passible	Oracle	ERM	FT	FF	DSAN	Ours + CVI	Ours + CVII	Ours+ TDV
0.25	$e = 0.1$	.750	.729 (.004)	.771 (.001)	.771 (.001)	..771 (.001)	.767 (.004)	.727 (.004)	.714 (.013)	.673 (.006)
	$e = 0.9$			.125 (.003)	.128 (.002)	.131(.002)	.085 (.003)	.622 (.015)	**.644 (.019)**	.690 (.009)
0.20	$e = 0.1$	.800	.780 (.002)	.796 (.000)	.800 (.001)	.796 (.001)	.789 (.004)	.773 (.003)	.745 (.008)	.738 (.018)
	$e = 0.9$			.177 (.006)	.201 (.004)	.200(.007)	.091 (.005)	.644 (.011)	**.707 (.012)**	.732 (.008)
0.15	$e = 0.1$	.850	.828 (.004)	.822 (.000)	.823 (.001)	.824 (.002)	.815 (.002)	.814 (.007)	.797 (.011)	.822 (.001)
	$e = 0.9$			.277 (.007)	.323 (.006)	.312(.012)	.091 (.002)	.724 (.037)	**.743 (.020)**	.782 (.012)
0.10	$e = 0.1$	.900	.880 (.004)	.852 (.002)	.855(.001)	.856 (.001)	.833(.003)	.848 (.005)	.848 (.005)	.857 (.005)
	$e = 0.9$			.468 (.002)	.497 (.005)	.500(.007)	.106 (.010)	**.792 (.005)**	**.792 (.005)**	.829 (.005)

Table 8: Baselines of CV methods

	Tr-CV	LOD-CV
0.25	.759 (.008)	.362 (.059)
	.459 (.012)	.372 (.037)
0.20	.794 (.004)	.338 (.048)
	.541 (.007)	.334 (.029)
0.15	.834 (.002)	.348 (.031)
	.634 (.008)	.358 (.024)
0.10	.876 (.003)	.502 (.196)
	.708 (.006)	.497 (.194)

two proposed methods, CV II yields the higher test accuracy. Table 9 shows the difference between accuracies by TDV and each CV for the same data set with $e = 0.9$. The results verify that CVII selects preferable hyperparameters with smaller errors.

G.3 Additional Experiment: Synthesized Data

We compared the proposed method with the other approaches using synthesized data with $\mathcal{X} = \mathbb{R}^2$, $\mathcal{Y} = [3]$ and $\mathcal{Z} := [2]$. We used distributions $N_0 := \mathcal{N}(0, 10^2) \times \mathcal{N}(e, 10^2)$, $N_1 := \mathcal{N}(30, 10^2) \times \mathcal{N}(-4e, 10^2)$ and $N_2 := \mathcal{N}(-30, 10^2) \times \mathcal{N}(-e, 10^2)$, where $\mathcal{N}(a, b)$ denotes a normal distribution with its (mean, variance) $= (a, b)$. Given $x \sim N_i$, the task is to predict N_i among $i = 0, 1, 2$. The aim of IL is to ignore the second component of x, as it works as a domain-specific factor. Given $e^* \in \mathbb{N}_{\geq 0}$ ranging from 0 to 50, each experiment draws $\mathcal{D}^{e^*} \sim P_{X^{e^*}, Y^{e^*}}$ with its sample size $n^{e^*} = 2000$, and then predicts Y^{-e^*} from X^{-e^*}. Setting g by $g(0) = 0$ and $g(1) = g(2) = 1$, we draw $\mathcal{D}^e_{ad} \sim P_{X^e, Z^e}$ from $\mathcal{E}_{ad} = \{-100, -50, 0, 50, 100\}$ with its sample size $n^e = 1000$ ($\forall e \in \mathcal{E}_{ad}$). We model Φ by a 3-layer neural net. Setting the maximum epoch 500, we select (t, λ_{after}) from 3×5 candidates with $t \in \{0, 100, 200\}$ and $\lambda_{after} \in \{10^0, 10^1, ..., 10^4\}$ by each of the CV methods.

Table 10 shows the test accuracy of the estimates for $e = -e^*$ over 2000 random samples $(x, y) \sim P_{X^{-e^*}, Y^{-e^*}}$. When $e^* = 0$ and 5, the domain bias of training (e^*) are similar to the one of test ($-e^*$), and hence, the fine-tuning methods yield high performances, which may use biased correlation. As e^* increases, the difference between the training (e^*) and test ($-e^*$) distributions becomes larger, and the previous methods fail to achieve high accuracy. The proposed methods (Ours) keep higher performance than the others even for large e^*. Among the CV methods, our two methods (CVI, CVII) significantly outperform the others for larger e^*. For this data set, CVI and CVII show almost the same performance.

G.4 Additional Experiment: Bird recognition

Our method is applied to the Bird recognition problem [4], which aims to predict three labels Y^e of images X^e: *waterbird* (Y^e= 0), *landbird* (Y^e= 1) and *no bird* (Y^e= 2). The dataset is made by combining background images from the Places dataset [6] and bird images from the CUB dataset [5] in two different ways $\mathcal{E} := \{e_1, e_2\}$. In domain e_1, we prepare three types of image: landbird image with land background, waterbird image with water background, and no bird with land background (Figure 3, left). In domain e_2, we have landbird images with water background, waterbird images with land background, and no bird with water background (Figure 3, right). For the sample of the target task, we used the domain $e^* = e_1$ and generated $n^{e^*} = 8649$ data $\mathcal{D}^{e^*} \sim P_{X^{e_1}, Y^{e_1}}$. The sample in the higher level $\mathcal{D}^e_{ad}$ of (X^e, Z^e), whose label is

Table 9: Means and SEs of {(Accuracy of TDV on $e = 0.9$) -(Accuracy of Each CV on $e = 0.9$) } (5runs).

	CV I	CV II	Tr-CV	LOD-CV
0.25	.068 (.007)	.046 (.023)	.231 (.013)	.319 (.033)
0.20	.088 (.004)	.025 (.006)	.191 (.014)	.398 (.025)
0.15	.059 (.038)	.039 (.022)	.148 (.019)	.430 (.028)
0.10	.037 (.010)	.037 (.010)	.121 (.008)	.332 (.196)

Table 10: Average Test ACCs and SEs of Synthesized Data on $e = -e^*$ (5 runs): *Oracle* shows the results of the experiments with the first component. The best scores are **bolded**.

	$e^* = 0$	$e^* = 5$	$e^* = 10$	$e^* = 15$	$e^* = 20$	$e^* = 25$	$e^* = 30$	$e^* = 35$	$e^* = 40$	$e^* = 45$	$e^* = 50$
Oracle	906 (.007)										
ERM	.789 (.218)	.791 (.174)	.637 (.188)	.329 (.201)	.324 (.328)	.311 (.260)	.159 (.193)	.140 (.171)	.132 (.161)	.166 (.147)	.051 (.101)
FT	**.899 (.000)**	**.863 (.001)**	.575 (.002)	.568 (.001)	.673 (.103)	.583 (.088)	.402 (.004)	.350 (.001)	.003 (.000)	.000 (.000)	.000 (.000)
FF	**.899 (.000)**	.861 (.002)	.540 (.102)	.568 (.001)	.673 (.102)	.628 (.001)	.401 (.001)	.351 (.002)	.066 (.132)	.000 (.000)	.000 (.000)
DSAN	.684 (.008)	.367 (.016)	.195 (.015)	.112 (.008)	.045 (.008)	.013 (.003)	.006 (.001)	.001(.001)	.000 (.000)	000 (.000)	.000 (.000)
Ours + Our CV I	.799 (.232)	.784 (.231)	**.884 (.021)**	**.875 (.044)**	**.815 (.098)**	**.738 (.209)**	**.865 (.047)**	**.659 (.233)**	**.666 (.285)**	**.776 (.080)**	**.699 (.255)**
Ours + Our CV II	.799 (.232)	.783 (.231)	**.884 (.021)**	**.875 (.044)**	**.815 (.098)**	**.738 (.209)**	**.865 (.047)**	**.659 (.233)**	.563 (.291)	**.776 (.080)**	**.699 (.255)**
Ours + Tr-CV	.790 (.230)	.776 (.225)	.609 (.163)	.491 (.095)	.366 (.147)	.248 (.192)	.376 (.033)	.215 (.168)	.148 (.127)	.189 (.108)	.031 (.138)
Ours + LOD-CV	.662 (.180)	.521 (.145)	.569 (.204)	.538 (.168)	.450 (.158)	.371 (.213)	.641 (.221)	.571 (.221)	.380 (.196)	.423 (.218)	.316 (.127)
Ours + TDV	.915 (.005)	.905 (.006)	.896 (.002)	.895 (.010)	.848 (.059)	.849 (.069)	.887 (.030)	.764 (.152)	.796 (.174)	.848 (.055)	.775 (.179)

landbird ($Z^e = 0$) and *no landbird* ($Z^e = 1$) (*i.e.*, $g(1) = 0$ and $g(0) = g(2) = 1$), is drawn from both e_1 and e_2 with $n^{e_1} = n^{e_2} = 8649$. Here, we use $\mathcal{D}^{e^*}$ as $\mathcal{D}^{e_1}_{ad}$ with labels of $\mathcal{D}^{e^*}$ re-annotated by g. We made a predictor of Y^e based on X^e, and evaluated the test accuracy in the two domains $e = e_1, e_2$. We model Φ by ResNet50 [1]. Setting the maximum epoch 5, we select (t, λ_{after}) from 5×5 candidates with $t \in [5], \lambda_{after} \in \{10^0, 10^1, ..., 10^4\}$ by each CV method.

Figure 3: Visualization of Bird recognition problem

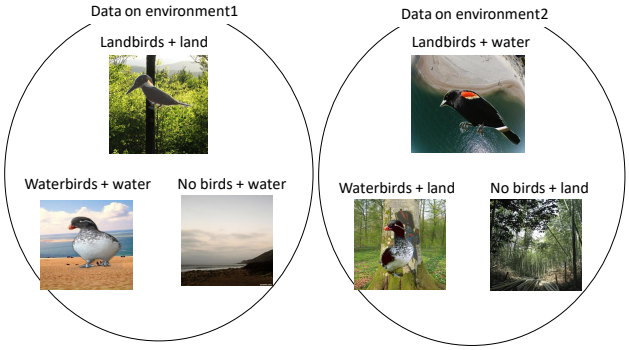


Table 11 shows test accuracies with 2162 random samples for e_1 and e_2. We can see that the proposed framework together with CV methods succeeded in capturing the predictor invariant to the change of background, while the other methods failed. ERM and FT show much higher accuracy for e_1 than Oracle and worst results for e_2, which implies that these methods learn biased correlation in $\mathcal{D}^{e^*}$.

Table 11: Average Test Accuracies and SEs of Bird recognition problem (5 runs). *Oracle* shows a result of ERM with samples from both e_1 and e_2 given. TDV selects λ which yields the highest performance on e_2. Best scores are **bolded**.

	Test Acc. on e_1	Test Acc. on e_2
Oracle	.875 (.018)	
ERM	.902 (.008)	.317 (.044)
FT	.909 (.012)	.364 (.028)
FE	.767 (.024)	.052 (.013)
Ours +Our CV I	.897 (.020)	**.727 (.062)**
Ours +Our CV II	.897 (.020)	**.727 (.062)**
Ours +Tr-CV	.919 (.006)	.651 (.031)
Ours +LOD CV	.338 (.048)	.334 (.029)
Ours +TDV	.886 (.035)	.782 (.020)

G.5 Additional Experiment: ImageNet

In the main body, only test accuracies on e_2 are shown. The result adding test accuracies on the training domain e_1 are as follows:

ImageNet: $\mathcal{Y} = [3], \mathcal{Z} = [2]$.

	Test Acc. on e_1	Test Acc. on e_2
random guess	.333	
Oracle	.743 (.018)	
ERM	.750 (.016)	.417 (.016)
FT	.793 (.018)	.463 (.030)
FF	.439 (.002)	.482 (.117)
DSAN	.288 (.012)	.278 (.004)
Ours + CV I	.843 (.024)	.652 (.028)
Ours + CV II	.852 (.009)	**.666 (.027)**
Ours + Tr-CV	.873 (.009)	.641 (.033)
Ours + LOD CV	.857 (.012)	.525 (.028)
Ours + TDV	.857 (.012)	.673 (.035)

ImageNet: $\mathcal{Y} = [7], \mathcal{Z} = [2]$.

	Test Acc. on e_1	Test Acc. on e_2
random guess	.143	
Oracle	.749 (.008)	
ERM	.740 (.017)	.507 (.020)
FT	.626 (.028)	.409 (.020)
FF	.191 (.004)	.226 (.046)
DSAN	.184 (.012)	.293 (.008)
Ours + CV I	.853 (.006)	**.622 (.011)**
Ours + CV II	.853 (.006)	**.622 (.011)**
Ours + Tr-CV	.850 (.004)	.612 (.012)
Ours + LOD CV	.825 (.017)	.572 (.022)
Ours + TDV	.837 (.019)	.634 (.003)

ImageNet: $\mathcal{Y} = [3], \mathcal{Z} = [2]$.

	Test Acc. on e_1	Test Acc. on e_2
random guess	.333	
Oracle	.743 (.018)	
ERM	.750 (.016)	.417 (.016)
FT	.793 (.018)	.463 (.030)
FF	.439 (.002)	.482 (.117)
DSAN	.288 (.012)	.278 (.004)
Ours + CV I	.843 (.024)	.652 (.028)
Ours + CV II	.852 (.009)	**.666 (.027)**
Ours + Tr-CV	.873 (.009)	.641 (.033)
Ours + LOD CV	.857 (.012)	.525 (.028)
Ours + TDV	.857 (.012)	.673 (.035)

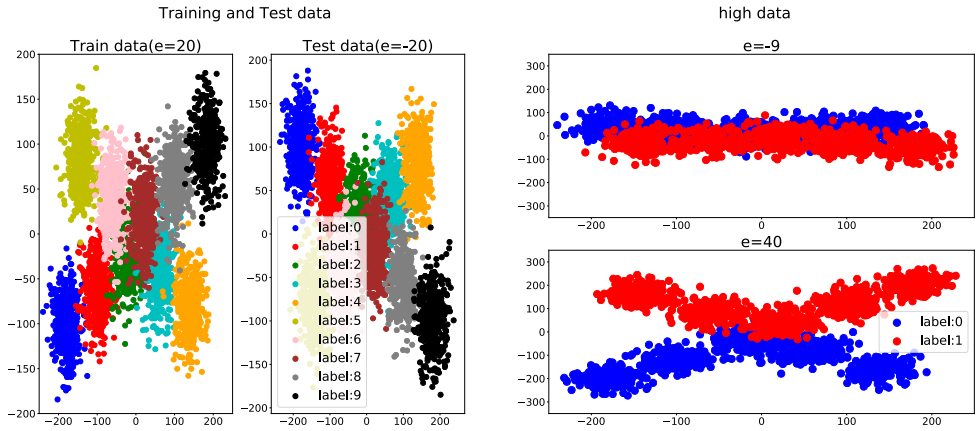
H Experimental Details

H.1 Visualizations of Experiment Results

Synthesized data in Appendix G.3 is visualized as in Figure 4. Synthesized data in the CVs comparison experiment (Section 6) is visualized as in Figure 5.

Figure 4: Synthesized Data in Appendix G.3. Left and middle figures illustrate training and test data on $e^* = 5$ and 50, respectively. As e^* increases, the test data and train data are more different, and therefore ERM yields lower performance. Right figure illustrates $\mathcal{D}_{ad}^{e}$.

Figure 5: Synthesized Data in Section 6. Left figure illustrates the training and test data of second experiment. Right figure illustrates $\mathcal{D}_{ad}^{40}$ and $\mathcal{D}_{ad}^{e_{ad}}$ with $e_{ad} = -9$.



H.2 Explicit representation of Synthetic data used in CV comparison experiment

$$
\begin{aligned}
N_1 &= \mathcal{N}(-180, 20^2) \times \mathcal{N}(-5e, 30^2), \\
N_2 &= \mathcal{N}(-100, 20^2) \times \mathcal{N}(-3e, 30^2), \\
N_3 &= \mathcal{N}(-20, 20^2) \times \mathcal{N}(-1e, 30^2), \\
N_4 &= \mathcal{N}(60, 20^2) \times \mathcal{N}(-2e, 30^2), \\
N_5 &= \mathcal{N}(140, 20^2) \times \mathcal{N}(-4e, 30^2), \\
N_6 &= \mathcal{N}(-140, 20^2) \times \mathcal{N}(4e, 30^2), \\
N_7 &= \mathcal{N}(-60, 20^2) \times \mathcal{N}(2e, 30^2), \\
N_8 &= \mathcal{N}(20, 20^2) \times \mathcal{N}(1e, 30^2),
\end{aligned}
$$

$$N_9 = \mathcal{N}(100, 20^2) \times \mathcal{N}(3e, 30^2),$$
$$N_{10} = \mathcal{N}(180, 20^2) \times \mathcal{N}(5e, 30^2).$$

H.3 Detail of ImageNet experiment data set

In the ImageNet experiment in Section 6, $\mathcal{Y}$ is set as follows:

- $\mathcal{Y} = [3]$:{**bird**, **turtle** ,snake }
- $\mathcal{Y} = [7]$: {**bird**, **turtle**, snake, cat, food, **vehicle**, **building** },
- $\mathcal{Y} = [17]$: {**bird**, **turtle**, snake, cat, **dog**, monkey, spider, butterfly, **food**, **vehicle**, **building**, **shoes**, hat, instrument, tellephone, **bottle**, chair}.

Images of **bolded** labels are composed of different species among e_1 and e_2. Explicitly, dataset are composed as follows:

$$\mathcal{Y} = [3]$$

label	e_1	e_2
bird	ruffed grouse, indigo bunting	albatross, water ouzel
turtle	loggerhead, leathback	.box turtle, mud turtle
snake	thunder snake, garther snake, ringneck. snake	

$$\mathcal{Y} = [7]$$

label	e_1	e_2
bird	ruffed grouse, indigo bunting	albatross, water ouzel
turtle	loggerhead, leathback	.box turtle, mud turtle
snake	thunder snake, garther snake, ringneck. snake	
cat	persian cat, siamese cat, egyptian cat	
food	cucumber, strawberry, pizza	
vechicle	submarine, container ship	golfcart, jeep
building	lighthouse, fountaink	castle, water tower

$\mathcal{Y} = [17]$

label	e_1	e_2
bird	ruffed grouse, indigo bunting	albatross, water ouzel
turtle	loggerhead, leathback	.box turtle, mud turtle
snake	thunder snake, garther snake, ringneck. snake	
cat	persian cat, siamese cat, egyptian cat	
dog	eskimo dog, dalmatian	newfoundland, German shepherd
monkey	guenon, colobus, titi	
spider	wolf spider, garden spider, barn spider	
butterfly	ringlet, monarch, cabbage butterfly	
food	pizza, strawberry	cucumber, broccoli
vechicle	submarine, container ship	golfcart, jeep
building	lighthouse, fountaink	castle, water tower
shoes	clog, sandal	running shoe, loafer
hat	pickelhaube, crash helmet, hat with a wide brim	
instrument	acoustic guitar, electric guitar, violin	
tellephone	cellular telephone, dial telephone, pay-phone	
bottle	pill bottle, pop bottle	beer bottle, wine bottle
chair	barber chair, folding chair, rocking chair	

H.4 model architectures and optimization procedures

Through the experiment in the present paper, all models of competitors are composed of neural networks where its loss function, activation function, and optimizer are cross entropy, Relu Networks and Adam [3]. In the following explanation, NN with its model architecture $a \to h_1 \to \cdots h_k \to h_n \to \mathcal{P}_{[m]}$ means that its input and hidden dimensions are a and $(h_1, ..., h_n)$ respectively, and its output is probability density functions on $[m]$. NN with its model architecture $a \to h_1 \to \cdots h_k \to h_n \to b$ means that its input, hidden and output dimensions are a, $(h_1, ..., h_n)$ and b respectively. All the experiment, we add L^2-reguralized term to our objective function.

We add explanations of previous CV methods. Tr-CV implements cross-validation with using only $\mathcal{D}^*$. In LOD-CV, a model is learnt with excluding one of the $\mathcal{D}^e \subset \mathcal{D}_{ad}$ from $\mathcal{D}_{ad}$, and evaluate its performance by $\mathcal{D}^e$. Changing the role of $e \in \mathcal{E}_{ad}$, and taking their mean, we evaluate final CV-value.

H.4.1 Colored MNIST

We set model architecture of Φ used in our method $2 \to 440 \to 440 \to 440$. We set model architecture of and ERM $2 \to 440 \to 440 \to \mathcal{P}_{[3]}$. When we use FT and FF, its model architecture on pre-train phase and retraining phase are $2 \to 440 \to 440 \to \mathcal{P}_{[2]}$ and $2 \to 440 \to 440 \to \mathcal{P}_{[3]}$ respectively. We set running rate and hyperparameter of L^2-regularized term 0.0004 and 0.002 respectively. When we use $DSAN$ [7], we inherit learning condition in the *Amazon Review dataset* experiment. When training, we use batch learning. We set $K = 10$ of our CV method.

H.4.2 ImageNet

We set model architecture of Φ used in our method ResNet50 [1] with changing its output dimension 256. We set model architecture of and ERM ResNet50 [1] with changing its output $\mathcal{P}_{[|\mathcal{Y}|]}$. When we use FT and FF, its model architecture on pre-train phase and retraining phase are ResNet50 [1] with changing its output dimension 2 and $|\mathcal{Y}|$ respectively. We set running rate and hyperparameter of L^2-regularized term 0.00004 and 0.001 respectively. When training, we use minibatch learning with a minibatch size 56. When we use $DSAN$ [7], we inherit learning condition in the *Amazon Review dataset* experiment. We set $K = 5$ of each CV method.

H.4.3 CV comparison experiment

We set model architecture of Φ used in our method $2 \to 8 \to 8 \to 1$. We set running rate and hyperparameters of L^2-regularized term 0.05 and 0.001 respectively. When training, we use minibatch learning with dividing $\mathcal{D}^*$, $\mathcal{D}^{e^{ad}}$ and $\mathcal{D}^{40}$ into 50 equal parts respectively. We set $K = 10$ of each CV method.

H.4.4 Appendix: Colored MNIST II

We set model architecture of Φ used in our method $2 \to 440 \to 440 \to 440$. We set model architecture of ERM $2 \to 440 \to 440 \to \mathcal{P}_{[3]}$. When we use FT and FF, its model architecture on pre-train phase and retraining phase are $2 \to 440 \to 440 \to \mathcal{P}_{[2]}$ and $2 \to 440 \to 440 \to \mathcal{P}_{[3]}$ respectively. We set running rate and hyperparameter of L^2-regularized term 0.0004 and 0.002 respectively. When we use $DSAN$ [7], we inherit learning condition in the *Amazon Review dataset* experiment. When training, we use batch learning. We set $K = 10$ of our CV method.

H.4.5 Appendix: Synthesized Data

We set model architecture of Φ used in our method $2 \to 20 \to 20 \to 1$. We set model architecture of ERM $2 \to 20 \to 20 \to \mathcal{P}_{[3]}$. When we use FT and FF, its model architecture on pre-train phase and retraining phase are $2 \to 20 \to 20 \to \mathcal{P}_{[2]}$ and $2 \to 20 \to 20 \to \mathcal{P}_{[3]}$ respectively. We set running rate and hyperparameters of L^2-regularized term 0.0115 and 0.01 respectively. When we use $DSAN$ [7], we inherit learning condition in the *Amazon Review dataset* experiment. When training, we use batch learning. We set $K = 10$ of each CV method.

H.4.6 Appendix: Birds recognition

We set model architecture of Φ used in our method ResNet50 [1] with changing its output dimension 256. We set model architecture of ERM ResNet50 [1] with changing its output $\mathcal{P}_{[3]}$. When we use FT and FF, its model architecture on pre-train phase and retraining phase are ResNet50 [1] with changing its output dimension 2 and 3 respectively. We set running rate and hyperparameter of L^2-regularized term 0.00004 and 0.001 respectively. When training, we use minibatch learning with a minibatch size 56. We set $K = 5$ of each CV method.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *In CVPR*, 2016.

[2] M. Rojas-Carulla, B. Scholkopf, R. Turner, and J. Peters. Invariant models for causal transfer learning. *JMLR*, 19(36):1–34, 2018.

[3] D. P. Kingma, J. L. Ba. Adam: A Method for Stochastic Optimization. *In ICLR*, 2015.

[4] S. Sagawa, P. W. Koh, T. B. Hashimoto, P. Liang. Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for worst-case generalization. *In ECCV*, 2019.

[5] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 dataset. Technical report, California Institute of Technology, 2011.

[6] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6): 1452–1464, 2017

[7] P. Stojanov, Z. Li, M. Gong, Ruichu Cai, J. G. Carbonell, K. Zhang. Domain Adaptation with Invariant Representation Learning: What Transformations to Learn? *In NeurIPS*, 2021.