A. How to Choose the Environments

The choice of environment is a central challenge not only in our invariant rationalization framework, but also generally in causal analyses with environments. Currently, there is not principled guidance on how to select environments, which remains an open research field. However, we have some general ideas in practice. For example, in NLP tasks, the identities of the people who write the texts, or a clustering of different writing styles, *e.g.*, word usage, can serve as environments. If the instances come with time labels of when the text was created, we can partition the instances into environments according to time.

B. Proof to Theorem 1

Proof. $\forall \boldsymbol{Z}$, partition $\boldsymbol{Z}$ into an invariant variable $\boldsymbol{Z}_I$ and a non-invariant variable $\boldsymbol{Z}_V$:

$$\boldsymbol{Z}_I = \boldsymbol{Z} \cap \{\boldsymbol{X}_1\}, \quad \boldsymbol{Z}_V = \boldsymbol{Z} \cap \{\boldsymbol{X}_2, \boldsymbol{X}_3\}.$$

Given an arbitrary π_1, we construct a specific $\pi_2 = \pi_2^*$ and $\pi_3 = \pi_3^*$ such that

$$\pi_2^*(\boldsymbol{x}_2|\boldsymbol{x}_1, y) = \pi_2^*(\boldsymbol{x}_2), \quad \pi_3^*(\boldsymbol{x}_3|\boldsymbol{x}_1) = \pi_3^*(\boldsymbol{x}_3). \tag{14}$$

In other words, set these two priors such that the all the non-invariant variables are uninformative of Y. Since the test adversary is allowed to choose any distribution, these set of priors is within the feasible set of the test adversary.

Under the set of priors in equation (14), the non-invariant features are not predicative of Y, and only the invariant features are predicative of Y, *i.e.*

$$p(Y|\boldsymbol{Z}, e_a) = p(Y|\boldsymbol{Z}_I, e_a). \tag{15}$$

Therefore

$$\begin{aligned}
\mathcal{L}^*_{\text{test}}(\boldsymbol{Z}; \pi_1, \pi_2^*, \pi_3^*) &= H(p(Y|\boldsymbol{Z}, e_a); p(Y|\boldsymbol{Z}, e_t)) \\
&\overset{(i)}{=} H(p(Y|\boldsymbol{Z}_I, e_a); p(Y|\boldsymbol{Z}, e_t)) \\
&\overset{(ii)}{\geq} H(p(Y|\boldsymbol{Z}_I, e_a)) \\
&\overset{(iii)}{\geq} H(p(Y|\boldsymbol{X}_1, e_a)) \\
&\overset{(iv)}{=} H(p(Y|\boldsymbol{X}_1, e_a); p(Y|\boldsymbol{X}_1, e_t)) \\
&= \mathcal{L}^*_{\text{test}}(\boldsymbol{X}_1; \pi_1, \pi_2^*, \pi_3^*),
\end{aligned} \tag{16}$$

where (i) is from equation (15); (ii) is from the relationship between cross entropy and entropy; (iii) is because $\boldsymbol{X}_1$ is the minimizer of conditional entropy of Y on $\boldsymbol{Z}_I$ and e_a, among all the invariant variables; (iv) is because, by the definition of invariant variables, $p(Y|\boldsymbol{X}_1, E) = p(Y|\boldsymbol{X}_1)$. Here, we use $\mathcal{L}^*_{\text{test}}(\boldsymbol{Z}; \pi_1, \pi_2^*, \pi_3^*)$ to emphasize that $\mathcal{L}^*_{\text{test}}(\boldsymbol{Z})$ is computed under the distribution of π_1, π_2^*, π_3^*. Therefore, if we optimize over π_2 and π_3, we have the following

$$\begin{aligned}
&\max_{\pi_2, \pi_3} \mathcal{L}^*_{\text{test}}(\boldsymbol{Z}; \pi_1, \pi_2, \pi_3) \geq \mathcal{L}^*_{\text{test}}(\boldsymbol{Z}; \pi_1, \pi_2^*, \pi_3^*), \\
&\max_{\pi_2, \pi_3} \mathcal{L}^*_{\text{test}}(\boldsymbol{X}_1; \pi_1, \pi_2, \pi_3) = \mathcal{L}^*_{\text{test}}(\boldsymbol{X}_1; \pi_1, \pi_2^*, \pi_3^*),
\end{aligned} \tag{17}$$

where the second line is because $p(Y|\boldsymbol{X}_1, e_a)$ does not depend on π_2 and π_3. Combining equations (16) and (17), we have

$$\max_{\pi_2, \pi_3} \mathcal{L}^*_{\text{test}}(\boldsymbol{Z}; \pi_1, \pi_2, \pi_3) \geq \max_{\pi_2, \pi_3} \mathcal{L}^*_{\text{test}}(\boldsymbol{X}_1; \pi_1, \pi_2, \pi_3). \tag{18}$$

Note that the above discussions holds for all π_1. Therefore, taking the maximum over π_1 of equation (18) preserves the inequality.

$$\max_{\pi_1, \pi_2, \pi_3} \mathcal{L}^*_{\text{test}}(\boldsymbol{Z}; \pi_1, \pi_2, \pi_3) \geq \max_{\pi_1, \pi_2, \pi_3} \mathcal{L}^*_{\text{test}}(\boldsymbol{X}_1; \pi_1, \pi_2, \pi_3),$$

which implies

$$\boldsymbol{X}_1 = \operatorname*{argmin}_{\boldsymbol{Z}} \max_{\pi_1, \pi_2, \pi_3} \mathcal{L}^*_{\text{test}}(\boldsymbol{Z}; \pi_1, \pi_2, \pi_3).$$

□

C. Additional Experiment

We investigate using of the differentiable training algorithm from Bastings et al. (2019), which performs binary selection of rationale words and achieves state-of-the-art results on the decorrelated beer review data from Lei et al. (2016). We use the out-of-box model from the paper[4] with our regularizer in equation (12). Table 3 gives the F1 scores on our more challenging beer review task.

The results reflect a very typical failure mode, where the algorithm always selects parts of the starting sentences as rationales, regardless of what aspect it is explaining. Because the comments of the appearance aspect usually appear in the beginning of the reviews, this results in a high F1 score in the appearance aspect, but very low in the other aspects.

This failure mode happens because our beer review task has very high correlations among different aspects. With span selection (equation (13)), our baselines, *i.e.*, RNP and 3PLAYER, can achieve reasonable highlighting in this setting. However, the method proposed by Bastings et al. (2019) uses binary selection, which is analogous to equation 12. The binary selection has resulted in a similar worse results in RNP and 3Player.

Table 3. F1 scores of rationale selection via (Bastings et al., 2019) on our dataset. The results are not directly comparable with the numbers in the main paper, because we use span selection and (Bastings et al., 2019) is restricted to binary selection.

Len	Appearance	Aroma	Palate
10	46.97	26.78	6.67
20	56.77	14.17	6.64
30	58.82	29.82	10.43

[4] https://github.com/bastings/interpretable_predictions.