CAUSAL-DRIVEN FEATURE EVALUATION FOR CROSS-DOMAIN IMAGE CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Out-of-distribution (OOD) generalization remains a core challenge in real-world classification. Most existing methods for OOD and domain generalization focus on cross-domain stability. However, cross-domain stability alone does not imply causal effectiveness. A feature may appear stable across source domains, yet fail to truly drive classification decisions under distribution shift. We re-evaluate learned representations from the perspective of causal driving force. By intervening on feature representations, we explicitly measure the probability of necessity and sufficiency of each representation segment for cross-domain prediction and convert these measurements into causal effectiveness scores. Based on these scores, we select high-scoring segments to form a representation subspace and retrain the classifier on this subspace, encouraging the model to rely more on representation components that drive classification and thereby improving robustness on unseen domains. We validate the proposed method on multiple datasets and pretrained backbones, where it consistently improves classification performance on unseen domains. We further provide theoretical derivations for the proposed method.

1 INTRODUCTION

Existing image classification models are often developed under the assumption that training and test data are drawn from the same distribution. In real-world deployment, however, this assumption rarely holds. Test samples may be affected by noise, changes in capture devices, imaging conditions, scene styles, or preprocessing pipelines, leading to distribution shifts from the training data. Maintaining reliable classification performance under such shifts is a central problem in out-of-distribution (OOD) generalization and domain generalization (Quiñonero-Candela et al., 2008; Ben-David et al., 2010; Gulrajani & Lopez-Paz, 2021).

Traditional OOD and domain generalization methods often address distribution shifts by learning cross-domain invariant representations (Arjovsky et al., 2019; Sagawa* et al., 2020; Cha et al., 2021). Their underlying assumption is that features stable across multiple source domains are more likely to remain effective in unseen domains. However, cross-domain invariance is an observational form of stability and does not necessarily imply causal effectiveness in the classification mechanism. A feature may remain stable across source domains due to shared backgrounds, acquisition conditions, or preprocessing biases, yet still fail to drive classification decisions (Geirhos et al., 2020). As a result, representations learned solely from invariance-based criteria may contain stable but low-value components; if the classifier relies on such components, its generalization ability on unseen domains can be limited.

A simple example illustrates this issue. In visual classification, object structure usually provides a more semantic basis for classification, while low-level visual cues such as color, brightness, or texture may also correlate with labels across multiple source domains and provide effective predictive signals within those domains. For methods that focus only on cross-domain stability, these cues may all appear useful because they improve source-domain classification. However, when the target domain changes the color or intensity relationship, decision rules relying on these low-level cues may fail. What is needed are features that continue to support classification decisions after the distribution shifts.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To address this issue, we re-evaluate learned representations from the perspective of causal efficacy. We first construct a comprehensive feature space to serve as the foundation for subsequent interventions on and selection of representation segments. We employ pre-trained visual backbones, which typically encode rich semantic information across different layers, ranging from local appearance patterns to high-level object structures, making them well suited for segment-level causal evaluation on complex visual datasets. Specifically, to determine whether internal representation segments are both necessary and sufficient for cross-domain prediction, we formalize the segment evaluation task using the causal framework of the Probability of Necessity and Sufficiency (PNS) (Pearl, 2009) and apply controlled interventions to each segment to assess these properties. We then rank and select representation segments based on the resulting causal efficacy scores. Finally, we construct a representation subspace using the high-scoring segments and retrain the classifier within this subspace, thereby guiding the final model to rely more heavily on representation components that facilitate classification under distribution shifts. Furthermore, we provide a theoretical derivation of the proposed scoring mechanism.

We validate the proposed method on three datasets—PACS (Li et al., 2017), Office-Home (Venkateswara et al., 2017), and DomainNet (Peng et al., 2019)—using three frozen vision backbones: DINOv2 (Oquab et al., 2024), MAE (He et al., 2022), and ImageNet-supervised ViT (Dosovitskiy et al., 2021). Experimental results demonstrate that the method consistently improves performance across diverse datasets and backbone architectures.

The core contribution of this paper is a representation-level causal formulation for layer selection in out-of-distribution (OOD) settings. Specifically: (i) we treat frozen layer-wise features as causal nodes to construct a representation-level structural causal model (SCM), thereby providing a well-defined causal foundation for subsequent interventions and counterfactual evaluations; (ii) within this SCM, we explicitly define treatment and control interventions for each representation segment and further define its necessity and sufficiency for cross-domain prediction; (iii) we derive a two-sided characterization of the causal efficacy of representation segments, providing a sandwich-style bounding framework for layer selection that naturally corresponds to the lower and upper bounds of classical binary Probability of Necessity and Sufficiency (PNS). Empirically, the criteria induced by the two sides of this theoretical characterization yield highly consistent Top-K layer selections, showing a close correspondence between the theoretical formulation and the observed layer-selection behavior.

2 Related Work

Out-of-distribution (OOD) classification and domain generalization (DG) aim to maintain predictive performance when the test distribution differs from the training domains. Existing methods include Invariant Risk Minimization (IRM) (Arjovsky et al., 2019), Distributionally Robust Optimization (GroupDRO) (Sagawa* et al., 2020), domain alignment and moment-matching techniques (Li et al., 2018b), meta-learning-based domain generalization (Li et al., 2018a), and optimization-based approaches such as SWAD (Cha et al., 2021). Recent studies have further improved domain generalization from various perspectives. CaSN (Yang et al., 2023) distills the concept of Probability of Necessity and Sufficiency (PNS) into a training loss for OOD learning to characterize causal effects; Pro-RandConv generates diverse virtual domains by progressively stacking small-kernel random convolutions (Choi et al., 2023); SAGM (Wang et al., 2023) seeks flatter solutions by aligning the gradients of empirical and parameter-perturbed losses; M2-CL (Ballas & Diou, 2024) combines low- and high-level representations across different layers and scales through contrastive learning; GGA (Ballas & Diou, 2025) identifies regions with more consistent source-domain gradients through gradient-guided parameter search; and ARITH (Wang et al., 2025) uses arithmetic-weighted gradients to obtain an optimization direction that is better balanced across domain-specific optima.

3 Method

3.1 Problem setup

Let $\mathcal{X}$ denote the input image space and $\mathcal{Y} = \{1, \ldots, C\}$ denote the class label space. Each sample is written as (x, y, d), where $x \in \mathcal{X}$ is an image, $y \in \mathcal{Y}$ is its class label, and d denotes the domain

index. The model predicts the class label y, while d is used only to indicate the domain from which the sample is drawn.

We consider a multi-source domain generalization setting with source domains $\mathcal{S} = \{S_1, \ldots, S_M\}$. Each source domain S_m induces a joint distribution $P_m(X, Y)$ over images and labels. All source domains share the same label space, but their image distributions may differ because of noise, background, imaging conditions, style, or preprocessing. Training samples are drawn from the source mixture

$$P_{\mathcal{S}}(X, Y) = \sum_{m=1}^{M} \pi_m P_m(X, Y), \tag{1}$$

where π_m denotes the proportion of domain S_m in the source mixture.

Given source-domain data, the goal is to learn a classifier $f : \mathcal{X} \to \Delta^{C-1}$ that generalizes to an unseen target domain T with distribution $P_T(X, Y)$. The target domain shares the same label space as the source domains, but its image distribution may differ from any P_m. The target domain is not available during training and is used only for final evaluation. For any domain distribution Q, we define the classification risk as

$$R_Q(f) = \mathbb{E}_{(x,y)\sim Q} \left[\ell(f(x), y)\right], \tag{2}$$

where ℓ is the classification loss. OOD generalization aims to obtain low target-domain risk $R_T(f)$, or equivalently high target-domain accuracy.

In the rest of the paper, we decompose the classifier as $f(x) = g(E(x))$, where $E : \mathcal{X} \to \mathcal{Z}$ is a feature encoder and $g : \mathcal{Z} \to \Delta^{C-1}$ is a classification head. Our method evaluates and selects representation segments within $\mathcal{Z}$.

3.2 Representation-Level SCM and Segmentation

Our method performs segment-level causal analysis in the representation space of a frozen visual encoder. Let $E : \mathcal{X} \to \mathcal{Z}$ denote a pre-trained and frozen visual encoder containing L layer-wise representations. For an input image $x \in \mathcal{X}$, we write the complete representation as

$$z = E(x) = z^{(1)} \oplus z^{(2)} \oplus \cdots \oplus z^{(L)}, \tag{3}$$

where $z^{(k)} \in \mathcal{Z}^{(k)}$ denotes the representation produced at the k-th layer, $\oplus$ denotes concatenation, and $\mathcal{Z} = \mathcal{Z}^{(1)} \times \cdots \times \mathcal{Z}^{(L)}$. Pre-trained visual encoders typically encode different levels of visual information across layers, ranging from local appearance patterns to higher-level object structures. These layer-wise representations therefore provide a natural segmentation of the complete feature space for subsequent causal evaluation.

We construct a representation-level structural causal model (SCM) directly on these frozen features. Each layer-wise representation is treated as a separate causal node, $R_k := U_k$ for $k = 1, \ldots, L$, and the downstream prediction is modeled as $\widehat{Y} = g_{\text{score}}(R_1, \ldots, R_L)$, where U_k provides the initial value of R_k using the corresponding feature $z^{(k)}$ extracted by the frozen encoder. There are no directed structural relations among $R_1, \ldots, R_L$ in this SCM. Their initial values may be statistically correlated because they are jointly extracted from the same input, but such correlation is a property of the assigned feature values rather than a causal relation among the representation nodes. Consequently, each representation node can be separately manipulated without propagating changes through the other representation nodes. This node-level structure provides the causal basis for the treatment and control interventions introduced in the next section.

For later use, we denote the complement of the k-th representation segment by

$$z_{\setminus k} = z^{(1)} \oplus \cdots \oplus z^{(k-1)} \oplus z^{(k+1)} \oplus \cdots \oplus z^{(L)}. \tag{4}$$

3.3 Reference Construction and Complementary Interventions

Based on the representation-level SCM defined above, we construct controlled representation interventions for evaluating each segment. We first train a scoring classifier on the complete frozen representation:

$$g_{\text{score}} = \arg\min_{g} \sum_{D\in\mathcal{S}} \mathbb{E}_{(x,y)\sim P_D} \left[\ell(g(E(x)), y)\right]. \tag{5}$$

The same g_{score} is used throughout all subsequent evaluations, so that different representation configurations are compared under a common decision function.

Reference construction. To obtain replacement values for representation interventions, we construct a class- and domain-balanced reference distribution over the complete frozen representation:

$$Q = \frac{1}{MC} \sum_{m=1}^{M} \sum_{c=1}^{C} P_m(Z \mid Y = c). \tag{6}$$

In practice, we first uniformly sample a source domain and a class, and then sample one representation from the corresponding training group. This avoids biasing the reference distribution toward domains or classes containing more samples. A complete reference representation is denoted by

$$\bar{z} = \bar{z}^{(1)} \oplus \cdots \oplus \bar{z}^{(L)} \sim Q. \tag{7}$$

Complementary treatment and control interventions. For the k-th representation segment, we define two complementary interventions. The treatment intervention retains the observed value of segment k and assigns reference values to all remaining representation nodes:

$$x_1^{(k)} : \quad do\left(R_k = z^{(k)}, R_{\setminus k} = \bar{z}_{\setminus k}\right), \tag{8}$$

which produces the representation

$$z_1^{(k)} = z^{(k)} \oplus_k \bar{z}_{\setminus k}. \tag{9}$$

This configuration retains only the sample-specific information carried by segment k.

Conversely, the control intervention assigns the reference value to segment k while retaining the observed values of all remaining segments:

$$x_0^{(k)} : \quad do\left(R_k = \bar{z}^{(k)}, R_{\setminus k} = z_{\setminus k}\right), \tag{10}$$

which gives

$$z_0^{(k)} = \bar{z}^{(k)} \oplus_k z_{\setminus k}. \tag{11}$$

The two interventions are therefore complementary: $x_1^{(k)}$ keeps segment k and replaces its complement, whereas $x_0^{(k)}$ replaces segment k and keeps its complement. We additionally use the original full representation and the complete reference representation as two anchor configurations,

$$z_{\text{full}} = z, \qquad z_{\text{ref}} = \bar{z}. \tag{12}$$

For a source domain D, each reference draw r yields domain-level accuracy outcomes $Y_{x_1}^{(r)}$, $Y_{x_0}^{(r)}$, and $Y_{\text{ref}}^{(r)}$ for the treatment, control, and reference configurations, respectively. The same realized reference representation $\bar{z}^{(r)}$ is shared across these three evaluations, whereas Y_{full} is independent of the reference draw. Each $Y \in [0, 1]$ denotes classification accuracy on the evaluation set of domain D, rather than per-sample binary correctness.

We denote the corresponding accuracies, averaged over reference draws, by

$$A_D^{\text{full}}, \qquad A_D^{(k)}, \qquad A_D^{(-k)}, \qquad A_D^{\text{ref}}, \tag{13}$$

where

$$A_D^{\text{full}} = Y_{\text{full}}, \qquad A_D^{(k)} = \mathbb{E}_r[Y_{x_1}^{(r)}], \qquad A_D^{(-k)} = \mathbb{E}_r[Y_{x_0}^{(r)}], \qquad A_D^{\text{ref}} = \mathbb{E}_r[Y_{\text{ref}}^{(r)}].$$

These four quantities form the basis for the necessity and sufficiency characterization developed in the next section.

3.4 CONTINUOUS-PNS CHARACTERIZATION OF REPRESENTATION CAUSAL EFFICACY

For binary potential outcomes $Y_{x_1}, Y_{x_0} \in \{0, 1\}$, the classical Probability of Necessity and Sufficiency (PNS) is defined as

$$\mathrm{PNS} = P(Y_{x_1} = 1,\ Y_{x_0} = 0), \tag{14}$$

which measures the probability that the outcome occurs under treatment but not under control. Equivalently, for any threshold $t \in (0, 1]$, the same event can be written as

$$Y_{x_1} \geq t, \qquad Y_{x_0} < t.$$

Hence, binary PNS admits the threshold-integral representation

$$\mathrm{PNS} = \int_0^1 P(Y_{x_1} \geq t,\ Y_{x_0} < t)\, dt. \tag{15}$$

For bounded continuous outcomes $Y \in [a, b]$, we use the corresponding threshold-integrated extension

$$\mathrm{PNS}^c := \int_a^b P(Y_{x_1} \geq t,\ Y_{x_0} < t)\, dt = \mathbb{E}[(Y_{x_1} - Y_{x_0})_+], \tag{16}$$

where the equality follows from the layer-cake identity.

For a two-sided characterization, we require only the following two draw-wise anchor orderings:

$$Y_{x_0}^{(r)} \geq Y_{\mathrm{ref}}^{(r)}, \qquad Y_{x_1}^{(r)} \leq Y_{\mathrm{full}}. \tag{17}$$

For segment k in source domain D, the continuous PNS is

$$\mathrm{PNS}_{D,k}^c = \mathbb{E}_r \left[\left(Y_{x_1}^{(r)} - Y_{x_0}^{(r)} \right)_+ \right]. \tag{18}$$

The first ordering implies $Y_{x_1}^{(r)} - Y_{x_0}^{(r)} \leq Y_{x_1}^{(r)} - Y_{\mathrm{ref}}^{(r)}$, while the second implies $Y_{x_1}^{(r)} - Y_{x_0}^{(r)} \leq Y_{\mathrm{full}} - Y_{x_0}^{(r)}$. By monotonicity of the positive-part operator, it follows that $(Y_{x_1}^{(r)} - Y_{x_0}^{(r)})_+ \leq (Y_{x_1}^{(r)} - Y_{\mathrm{ref}}^{(r)})_+$ and $(Y_{x_1}^{(r)} - Y_{x_0}^{(r)})_+ \leq (Y_{\mathrm{full}} - Y_{x_0}^{(r)})_+$.

We therefore define

$$S_D^{+(k)} := \mathbb{E}_r \left[\left(Y_{x_1}^{(r)} - Y_{\mathrm{ref}}^{(r)} \right)_+ \right], \qquad N_D^{+(k)} := \mathbb{E}_r \left[\left(Y_{\mathrm{full}} - Y_{x_0}^{(r)} \right)_+ \right]. \tag{19}$$

We further define $U_D^{+(k)} := \min\{S_D^{+(k)}, N_D^{+(k)}\}$. The positive part is applied within each reference draw before averaging over r. The inequalities above directly give $\mathrm{PNS}_{D,k}^c \leq U_D^{+(k)}$.

For the lower side, since $a \leq (a)_+$ for any real a,

$$L_D^{(k)} := \mathbb{E}_r \left[Y_{x_1}^{(r)} - Y_{x_0}^{(r)} \right] = A_D^{(k)} - A_D^{(-k)} \leq \mathrm{PNS}_{D,k}^c. \tag{20}$$

Combining both sides gives

$$L_D^{(k)} \leq \mathrm{PNS}_{D,k}^c \leq U_D^{+(k)}. \tag{21}$$

In the unit-range binary case with $Y_{\mathrm{ref}} = 0$ and $Y_{\mathrm{full}} = 1$, $S_D^{+(k)} = P(Y_{x_1} = 1)$, $N_D^{+(k)} = P(Y_{x_0} = 0)$, and $L_D^{(k)} = P(Y_{x_1} = 1) + P(Y_{x_0} = 0) - 1$. Thus $U_D^{+(k)}$ reduces to the familiar marginal upper bound, while $L_D^{(k)}$ corresponds to the untruncated Fréchet lower-bound term for binary PNS (Tian & Pearl, 2000). The continuous formulation therefore preserves the same lower- and upper-side structure while extending the outcome from binary correctness to domain-level accuracy.

These two quantities characterize the causal efficacy of each representation segment from the lower and upper sides of the same continuous-PNS interval. For multiple source domains, we aggregate each side by its worst-domain value:

$$U^{+(k)} = \min_{D \in \mathcal{S}} U_D^{+(k)}, \qquad L^{(k)} = \min_{D \in \mathcal{S}} L_D^{(k)}. \tag{22}$$

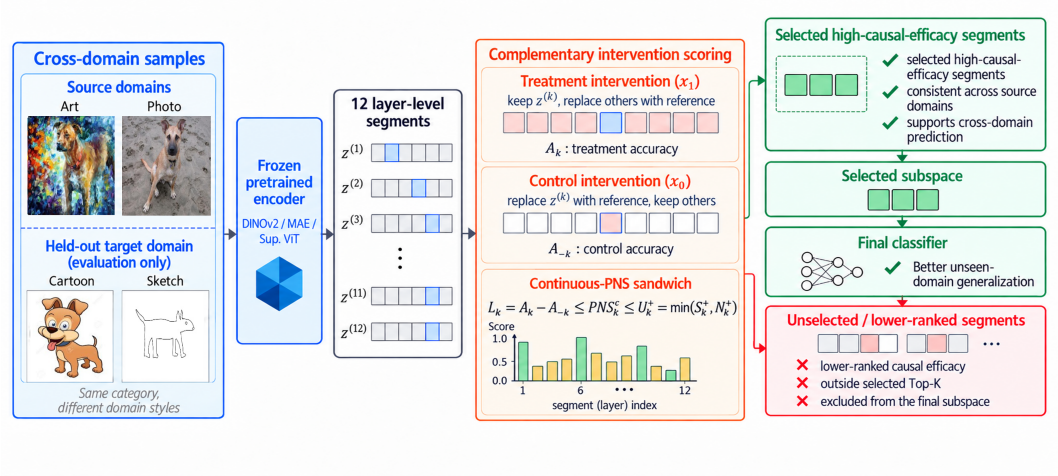


Figure 1: Overview of the proposed causal representation selection framework. Frozen layer-wise representations are evaluated through complementary treatment and control interventions. The resulting necessity and sufficiency quantities provide a two-sided continuous-PNS characterization, from which high-scoring segments are selected to construct the final representation subspace.

In practice, we use the upper-side criterion $U^{+(k)}$. The lower-side criterion $L^{(k)}$ provides the complementary side of the same characterization. As shown in Section 4.3, the two sides yield highly consistent Top-K selections in practice.

We use the selected Top-K segments to construct the final representation subspace

$$z_{\mathcal{K}^\star} = \bigoplus_{k \in \mathcal{K}^\star} z^{(k)}, \tag{23}$$

and retrain the final classifier exclusively on this subspace:

$$g_{\text{final}} = \arg\min_{g} \sum_{D \in \mathcal{S}} \mathbb{E}_{(x,y)\sim P_D} \left[\ell \left(g(E(x)_{\mathcal{K}^\star}), y \right) \right]. \tag{24}$$

The resulting predictor is

$$f^\star(x) = g_{\text{final}} \left(E(x)_{\mathcal{K}^\star} \right). \tag{25}$$

Figure 1 gives an overview of this pipeline.

Information-theoretic explanation. The difference between necessity and sufficiency can also be interpreted from an information-theoretic perspective. Consider an idealized Bayes classification setting with $Z = (B, V)$, where B denotes stable semantic cues such as shape, and V denotes source-effective but style-sensitive cues such as color, texture, or background.

First, V may have high sufficiency in the source domains. Let $P_{e,D}(V)$ be the Bayes classification error when only V is used. By Fano's inequality,

$$H_D(Y \mid V) \le h(P_{e,D}(V)) + P_{e,D}(V) \log(|\mathcal{Y}| - 1), \tag{26}$$

where $h(\cdot)$ is the binary entropy function. Therefore,

$$I_D(Y; V) = H_D(Y) - H_D(Y \mid V) \ge H_D(Y) - h(P_{e,D}(V)) - P_{e,D}(V) \log(|\mathcal{Y}| - 1). \tag{27}$$

Thus, when V performs well on its own, $I_D(Y; V)$ cannot be too small. Source-effective color or texture cues may therefore carry substantial marginal label information, which corresponds to high sufficiency.

However, high marginal label information does not imply high necessity. Necessity is related to the additional decision value of V once the other segments are already available. Define

$$A_D^\star(B) = \mathbb{E}_B \left[\max_y P_D(Y = y \mid B) \right], \qquad A_D^\star(B, V) = \mathbb{E}_{B,V} \left[\max_y P_D(Y = y \mid B, V) \right]. \tag{28}$$

For any given (b, v), the change in the Bayes optimal correct probability is bounded by

$$\left|\max_y P_D(Y = y \mid b, v) - \max_y P_D(Y = y \mid b)\right| \leq \mathrm{TV}\left(P_D(Y \mid b, v), P_D(Y \mid b)\right). \tag{29}$$

Taking expectation over (B, V), and applying Pinsker's inequality followed by Jensen's inequality, we obtain

$$\begin{aligned} A^\star_D(B, V) - A^\star_D(B) &\leq \mathbb{E}_{B,V}\mathrm{TV}\left(P_D(Y \mid B, V), P_D(Y \mid B)\right) \\ &\leq \sqrt{\frac{1}{2}\mathbb{E}_{B,V}\mathrm{KL}\left(P_D(Y \mid B, V) \| P_D(Y \mid B)\right)} \\ &= \sqrt{\frac{1}{2} I_D(Y; V \mid B)}. \end{aligned} \tag{30}$$

Therefore, even if $I_D(Y; V)$ is large, adding V provides little additional Bayes-optimal accuracy when $I_D(Y; V \mid B)$ is small. In terms of our intervention scores, V can have high sufficiency but low necessity once the stable semantic component B is already available. The upper-side score $U^+_D(V) = \min\{N^+_D(V), S^+_D(V)\}$ therefore penalizes such source-effective but low-necessity fragments.

Let P_S and P_T denote the source and target domain distributions, respectively. For a classifier that relies on the representation V', the target-domain risk can be related to the source-domain risk and a distribution-shift term such as $\mathrm{KL}(P_{T,V',Y} \| P_{S,V',Y})$. When $V' = (B, V)$, the KL chain rule yields

$$\begin{aligned} \mathrm{KL}(P_{T,B,V,Y} \| P_{S,B,V,Y}) &= \mathrm{KL}(P_{T,B,Y} \| P_{S,B,Y}) \\ &\quad + \mathbb{E}_{P_{T,B,Y}}\mathrm{KL}\left(P_T(V \mid B, Y) \| P_S(V \mid B, Y)\right). \end{aligned} \tag{31}$$

The second term captures source-to-target shifts in $P(V \mid B, Y)$, which can be large for style-sensitive cues. In the idealized case where selection retains the stable component while excluding V, this explicit conditional-shift term disappears, motivating reduced reliance on low-necessity style-sensitive features.

4 Experiments

4.1 Experimental Setup

Datasets and backbones. We evaluate on PACS (Li et al., 2017), Office-Home (Venkateswara et al., 2017), and DomainNet (Peng et al., 2019), using 3-to-1 protocols for PACS and Office-Home and 5-to-1 for DomainNet. We use three frozen pretrained backbones: DINOv2 ViT-B/14 (Oquab et al., 2024), MAE ViT-B/16 (He et al., 2022), and ImageNet-supervised ViT-B/16 (Dosovitskiy et al., 2021), with the 12 Transformer blocks treated as candidate representation segments.

Implementation and baselines. We compare with ERM, IRM, GroupDRO, MLDG, MMD, CDANN, CaSN, GGA and ARITH (Gulrajani & Lopez-Paz, 2021; Arjovsky et al., 2019; Sagawa* et al., 2020; Li et al., 2018a;b;c; Yang et al., 2023; Ballas & Diou, 2025; Wang et al., 2025). All methods use the same source train/validation splits and linear classifier, and results are averaged over three seeds. Further details are provided in the Appendix.

4.2 Main Results

Tables 1–3 report the results across the three pretrained backbones.

Across the nine dataset–backbone combinations, our method achieves the best performance in eight settings. With DINOv2, it outperforms the strongest competing method by 2.55, 5.71, and 2.27 percentage points on PACS, Office-Home, and DomainNet, respectively. Notably, these improvements are obtained on top of a strong self-supervised representation, suggesting that the proposed selection mechanism remains useful even when the pretrained features already provide strong OOD performance.

Table 1: OOD classification accuracy (%) using frozen DINOv2 ViT-B/14 representations. Results are mean ± standard deviation over three seeds.

Method	PACS 3→1	Office-Home 3→1	DomainNet 5→1
ERM	90.51±0.32	70.29±0.57	47.17±0.49
IRM	90.58±1.25	69.78±0.52	49.59±0.34
GroupDRO	90.42±0.40	69.19±0.96	44.07±1.78
MLDG	90.45±0.51	70.36±0.59	46.83±0.49
MMD	90.46±0.28	70.41±0.36	47.32±0.41
CDANN	87.18±2.86	68.57±0.86	47.59±0.66
CaSN	90.99±0.31	73.79±0.29	49.61±0.77
GGA	91.80±0.32	75.70±0.21	50.84±0.65
ARITH	90.38±0.58	71.92±0.40	47.21±0.94
Ours	**94.35±0.13**	**81.41±0.21**	**53.11±0.12**

Table 2: OOD classification accuracy (%) using frozen MAE ViT-B/16 representations. Results are mean ± standard deviation over three seeds.

Method	PACS 3→1	Office-Home 3→1	DomainNet 5→1
ERM	59.65±0.52	48.02±1.27	27.86±0.35
IRM	59.44±0.60	48.75±1.42	27.66±0.32
GroupDRO	58.99±0.46	46.37±0.60	25.39±0.53
MLDG	59.79±0.55	48.10±0.52	27.14±0.30
MMD	59.67±0.58	48.57±0.87	27.84±0.35
CDANN	60.31±1.32	47.65±0.54	27.82±0.38
CaSN	61.84±0.22	52.98±0.69	29.36±0.59
GGA	61.82±0.29	49.08±0.81	**31.67±0.54**
ARITH	59.35±0.68	50.60±0.61	30.81±0.60
Ours	**63.63±0.54**	**55.38±0.24**	31.08±0.19

Table 3: OOD classification accuracy (%) using frozen ImageNet-supervised ViT-B/16 representations. Results are mean ± standard deviation over three seeds.

Method	PACS 3→1	Office-Home 3→1	DomainNet 5→1
ERM	68.79±0.40	64.40±0.46	36.15±0.29
IRM	69.76±0.92	64.27±0.66	36.57±0.19
GroupDRO	68.61±0.24	62.21±0.68	36.25±0.41
MLDG	69.17±0.54	64.43±0.50	35.79±0.29
MMD	68.79±0.45	64.42±0.68	36.13±0.31
CDANN	65.84±1.49	63.66±0.44	36.10±0.41
CaSN	69.88±0.61	65.18±0.33	37.17±0.31
GGA	67.85±0.58	64.64±0.89	36.53±0.48
ARITH	70.45±0.98	66.27±0.20	37.61±0.55
Ours	**71.25±0.59**	**70.44±0.18**	**37.69±0.13**

4.3 Ablation Experiments

To examine whether the gains arise merely from representation compression or from selecting a layer that is already highly predictive on the source domains, we compare our method with two simple baselines on PACS across all three pretrained backbones. Random-K randomly retains the same number of representation segments as our method, while Source-best layer selects the individual layer with the highest source-domain predictive performance, without access to the held-out target domain.

Table 4: Comparison with simple representation-selection baselines on PACS. Random-K retains the same number of representation segments as our method. Source-best layer selects the individual layer with the highest source-domain predictive performance without access to the held-out target domain.

Backbone	Full 12 layers	Random-K	Source-best layer	Ours
DINOv2	90.51	64.25	92.29	**94.35**
MAE	59.65	53.63	63.38	**63.63**
Sup. ViT	68.79	65.71	67.56	**71.25**
Average	72.98	61.20	74.41	**76.41**

Random-K consistently underperforms the full representation, showing that the improvement is not a generic consequence of reducing representation dimensionality. In contrast, Source-best layer provides a substantially stronger baseline, indicating that layer choice itself is important for OOD generalization. Nevertheless, our method achieves higher accuracy across all three backbones, improving the average performance from 74.41% to 76.41%. These results suggest that the gains cannot be fully explained by either representation compression or simply choosing the most source-predictive individual layer.

Consistency of the two characterization sides. We further compare the Top-K selections produced by the lower criterion $L_D^{(k)}$ and the upper criterion $U_D^{+(k)}$ in Section 3.4.

Table 5: Top-K overlap (%) between the lower- and upper-side selection criteria.

Dataset	DINOv2	MAE	Sup. ViT
PACS	100%	75%	88%
Office-Home	100%	100%	100%
DomainNet	100%	100%	96%

The two criteria achieve an average Top-K overlap of 95.4%. In six of the nine dataset–backbone settings, they produce identical selections for every held-out target domain. Although the two criteria are not mathematically equivalent, their highly consistent selections empirically support the two-sided characterization used in our framework. Detailed disagreement cases are provided in the Appendix.

5 Conclusion

We propose a causal-effectiveness-based representation selection method for domain generalization. The method evaluates frozen representation segments through complementary treatment and control interventions, and characterizes their causal efficacy through a two-sided continuous-PNS formulation. High-scoring segments are selected to construct the final representation subspace. Experiments across multiple datasets and pretrained backbones show consistent improvements on unseen target domains. Ablation studies further demonstrate that the gains cannot be explained by representation compression alone, while the two sides of the proposed characterization yield highly consistent Top-K selections in practice.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript editing, organization and standardization of the presentation of author-developed theoretical arguments, and figure refinement. The core research ideas, theoretical formulation, and mathematical claims were developed by the authors. All AI-assisted content was reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Implementation details, hyperparameters, additional analyses, and sensitivity results are provided in the Appendix. Source code and supplementary materials will be submitted anonymously to support reproduction of the reported results.

REFERENCES

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Aristotelis Ballas and Christos Diou. Multiscale and multilayer contrastive learning for domain generalization. *IEEE Transactions on Artificial Intelligence*, 5(12):6253–6266, 2024. doi: 10.1109/TAI.2024.3377173.

Aristotelis Ballas and Christos Diou. Gradient-guided annealing for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20558–20568, June 2025.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2010.

Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. SWAD: Domain generalization by seeking flat minima. In *Advances in Neural Information Processing Systems*, volume 34, pp. 22405–22418. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/bcb41ccdc4363c6848a1d760f26c28a0-Paper.pdf.

Seokeon Choi, Debasmit Das, Sungha Choi, Seunghan Yang, Hyunsin Park, and Sungrack Yun. Progressive random convolutions for single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10312–10322, June 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=lQdXeXDoWtI.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, June 2022.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 5542–5550, 2017.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018a. doi: 10.1609/aaai.v32i1.11596.

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C. Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018b.

Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018c.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

Judea Pearl. *Causality*. Cambridge University Press, 2 edition, 2009.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence (eds.). *Dataset Shift in Machine Learning*. MIT Press, 2008. doi: 10.7551/mitpress/9780262170055.001.0001.

Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Jin Tian and Judea Pearl. Probabilities of causation: Bounds and identification. In Craig Boutilier and Moisés Goldszmidt (eds.), *UAI '00: Proceedings of the 16th Conference in Uncertainty in Artificial Intelligence, Stanford University, Stanford, California, USA, June 30 - July 3, 2000*, pp. 589–598. Morgan Kaufmann, 2000. URL https://dslpitt.org/uai/displayArticleDetails.jsp?mmnu=1&smnu=2&article_id=69&proceeding_id=16.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Pengfei Wang, Zhaoxiang Zhang, Zhen Lei, and Lei Zhang. Sharpness-aware gradient matching for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3769–3778, June 2023.

Xiran Wang, Jian Zhang, Lei Qi, and Yinghuan Shi. Balanced direction from multifarious choices: Arithmetic meta-learning for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30577–30587, June 2025.

Mengyue Yang, Zhen Fang, Yonggang Zhang, Yali Du, Furui Liu, Jean-Francois Ton, Jianhong Wang, and Jun Wang. Invariant learning via probability of sufficient and necessary causes. In *Advances in Neural Information Processing Systems*, volume 36, pp. 79832–79857. Curran Associates, Inc., 2023. doi: 10.52202/075280-3496. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/fc657b7fd7b9aaa462f2ef9f0362b273-Paper-Conference.pdf.

A Additional Experimental Details

Feature extraction and baseline setting. All experiments use frozen layer-wise representations extracted from the same pretrained backbone for both our method and all baselines. We consider DINOv2 ViT-B/14, MAE ViT-B/16, and ImageNet-supervised ViT-B/16, and treat the representations from their 12 Transformer blocks as candidate segments. For DINOv2 and supervised ViT, we use the CLS-token representation from each block, while for MAE we use the mean-pooled patch representation. All baselines use the same frozen representations and source-domain data splits as our method.

Training protocol. All methods use a linear classifier implemented with `nn.Linear` and optimized with AdamW using a learning rate of 10^{-3} and weight decay 10^{-4}. PACS and Office-Home are trained for 30 epochs, while DomainNet is trained for 10 epochs. Results are averaged over three random seeds. K is selected using source-domain information only and fixed before target evaluation.

Reference construction and segment scoring. Reference representations are sampled from the domain- and class-balanced reference distribution defined in Section 3.3. The same scoring classifier is used to evaluate the full, reference, treatment, and control configurations. After ranking the representation segments and selecting the Top-K subset, a final linear classifier is retrained on the selected representation subspace.

B Comparison of the Two Characterization Sides

As reported in the main text, the lower- and upper-side criteria achieve an average Top-K overlap of 95.4%, and therefore produce identical selections in most cases. Table 6 reports the few target-level cases in which the selected subsets differ, together with the corresponding target-domain accuracy after retraining the final classifier on each selected subspace.

Table 6: Cases where the upper- and lower-side criteria produce different Top-K selections. The corresponding target-domain accuracy (%) is reported. Results are mean $\pm$ standard deviation over three seeds.

Setting	Upper acc.	Lower acc.
PACS / MAE / Art	69.82±0.56	**70.08±0.51**
PACS / MAE / Photo	**87.49±0.15**	86.99±0.33
PACS / Sup. ViT / Art	76.82±0.64	**77.00±0.62**
PACS / Sup. ViT / Cartoon	**65.66±0.06**	65.12±0.09
DomainNet / Sup. ViT / Infograph	21.37±0.19	**21.41±0.08**

Even in these disagreement cases, the resulting performance differences remain small. Across the five cases, the mean accuracy difference between the upper and lower sides is only $+0.11$ percentage points, with a maximum absolute difference of 0.54 points. The upper side performs better in two cases, while the lower side performs better in three. Thus, neither criterion consistently dominates the other, supporting their interpretation as two complementary sides of the same continuous-PNS characterization rather than equivalent ranking functions.

C Validity of the Anchor Orderings

The two-sided characterization in Section 3.4 relies on the draw-wise orderings

$$Y_{x_0}^{(r)} \geq Y_{\mathrm{ref}}^{(r)}, \qquad Y_{x_1}^{(r)} \leq Y_{\mathrm{full}}.$$

These orderings have a natural interpretation in an idealized Bayes-optimal classification setting. For a fixed reference draw r, the treatment representation $x_1^{(r)}$ retains only the observed segment

$z^{(k)}$, while the remaining segments are replaced by fixed reference values. Relative to the full representation, this intervention therefore removes sample-specific information from $z_{\setminus k}$. Since additional information cannot decrease Bayes-optimal classification accuracy, the corresponding Bayes accuracies satisfy

$$A_D^\star(z^{(k)}, z_{\setminus k}) \geq A_D^\star(z^{(k)}, \bar{z}_{\setminus k}^{(r)}).$$

Likewise, $x_0^{(r)}$ retains the observed complement $z_{\setminus k}$, whereas the complete reference representation contains only reference values. Hence,

$$A_D^\star(\bar{z}^{(k,r)}, z_{\setminus k}) \geq A_D^\star(\bar{z}^{(k,r)}, \bar{z}_{\setminus k}^{(r)}).$$

Thus, in the Bayes-optimal limit, the two anchor orderings follow from the monotonicity of optimal prediction under information removal.

Our implementation, however, evaluates all configurations using the same finite-sample scoring classifier g_{score}, rather than re-optimizing a Bayes classifier separately for each intervention. Therefore, we treat the above orderings as empirically verifiable anchor conditions rather than universal properties of an arbitrary fixed classifier.

We verify these conditions over all 54,000 draw-level domain-accuracy evaluations used in our experiments. Both orderings hold in 100% of the evaluated cases.

D Distribution Shift Induced by Reference Recombination

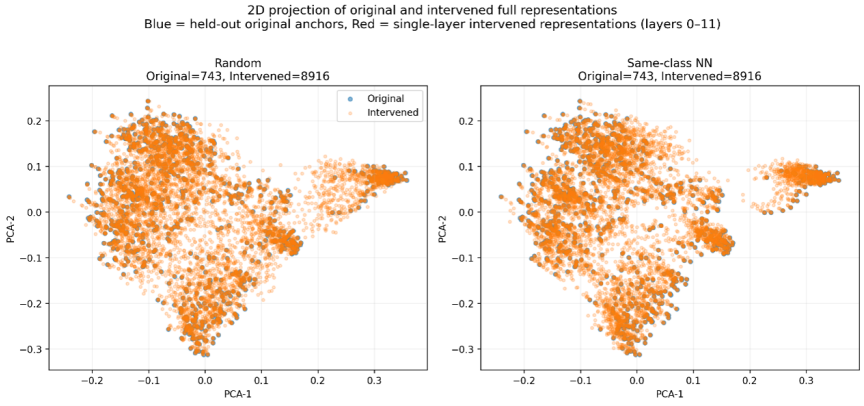


Figure 2: PCA visualization of natural and reference-recombined representations.

Our interventions construct recombined representations by replacing selected segments with reference features sampled from other source examples. Since the replacement segment is sampled without requiring its class label to match that of the anchor sample, such recombinations may be label-inconsistent or partially off-manifold. The scoring classifier g_{score} is trained only on natural full representations and is kept fixed throughout all intervention evaluations.

For comparison, we additionally visualize a class-consistent variant in which the replacement segment is taken from a same-class nearest neighbor. This variant is used only as a diagnostic to distinguish the effect of label inconsistency from the effect of representation recombination itself, and is not part of the proposed method.

We quantify the resulting distribution shift using the k-nearest-neighbor distance to a natural source-domain representation bank. The off-support threshold is defined as the 95th percentile of held-out natural kNN distances; an intervened representation is regarded as off-support when its distance exceeds this threshold. We first average the off-support indicator over intervened layers for each sample and then across samples. The corresponding 95% confidence intervals are estimated using 10,000 sample-level bootstrap resamples.

The natural off-support rate is 5.11%, close to the calibrated 5% baseline, whereas random reference replacement increases the rate to 13.86%. As shown in Figs. 2 and 3, PCA and t-SNE visualizations

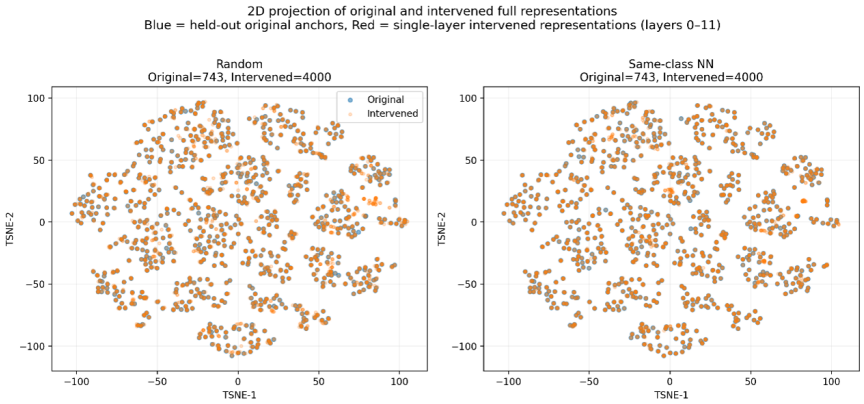


Figure 3: t-SNE visualization of natural and reference-recombined representations.

exhibit substantial overlap between natural and recombined representations, although some intervened samples move outside the natural support. Thus, reference recombination can induce partial label inconsistency and off-manifold displacement, but the shift remains limited and does not separate the recombined representations from the natural representation space. In addition, intervention scores are averaged over multiple independent reference draws, reducing sensitivity to any individual replacement.

Such displacement may also provide complementary evidence about representation structure: if replacing a segment changes both the classifier output and the position of the representation relative to the natural manifold, that segment may contribute to maintaining the organization of the representation space. We regard this observation only as supporting evidence rather than direct proof of segment importance.

E Computational Complexity

Let N_s, N_v, and N_t denote the numbers of source-training, source-validation, and target samples, respectively; let $N = N_s + N_v + N_t$ denote the total number of samples considered for feature extraction; L the number of candidate layers; d the representation dimension per layer; C the number of classes; E the number of source environments; R the number of reference replacements per intervention; K the number of retained layers; and B the cost of one frozen-backbone forward pass. Let I_b and I_K denote the effective optimization work of the full-representation and final Top-K classifiers, respectively. In our experiments, $L = 12$, $d = 768$, and $R = 5$.

The frozen backbone returns all intermediate representations in a single forward pass, so feature extraction costs

$$\mathcal{O}(NB),$$

while storing or loading all layer-wise representations costs

$$\mathcal{O}(NLd).$$

The full representation has dimension Ld, and fitting the scoring classifier costs

$$\mathcal{O}(I_b N_s C L d).$$

For PNS scoring, each of the L candidate layers is evaluated under both intervention directions using R reference replacements. A fixed-classifier evaluation over all source-validation samples costs $\mathcal{O}(N_v C L d)$. Therefore, the dominant PNS-scoring cost is

$$\mathcal{O}(R N_v C L^2 d),$$

where the constant factor of two from the two intervention directions is omitted in asymptotic notation. Cross-environment aggregation and layer ranking require $\mathcal{O}(EL)$ and $\mathcal{O}(L \log L)$ operations, respectively.

After the Top-K layers are selected, constructing the selected representation, fitting the final classifier, and evaluating the target domain cost

$$\mathcal{O}(N_s K d), \qquad \mathcal{O}(I_K N_s C K d), \qquad \mathcal{O}(N_t C K d),$$

respectively. Hence, the overall dominant complexity is

$$\mathcal{O}\big(NB + NLd + I_b N_s C L d + R N_v C L^2 d + EL + L \log L + N_s K d + I_K N_s C K d + N_t C K d\big).$$

Feature extraction, full-classifier fitting, PNS scoring, cross-environment aggregation, and layer ranking are independent of the final choice of K. Only subspace construction, final-classifier fitting, and target evaluation depend on K. For fixed L, d, C, E, R, K, with $N_s, N_v, N_t = \Theta(N)$ and I_b and I_K fixed or bounded under the chosen solver, the overall complexity scales linearly with dataset size:

$$T_{\text{ours}}(N) = \mathcal{O}(N).$$