Learning Causal Semantic Representation for Out-of-Distribution Prediction

Chang Liu[1]*, **Xinwei Sun**[1], **Jindong Wang**[1], **Haoyue Tang**[2]†, **Tao Li**[3]†,
Tao Qin[1], **Wei Chen**[1], **Tie-Yan Liu**[1]
[1] Microsoft Research Asia, Beijing, 100080.
[2] Tsinghua University, Beijing, 100084. [3] Peking University, Beijing, 100871.

Abstract

Conventional supervised learning methods, especially deep ones, are found to be sensitive to out-of-distribution (OOD) examples, largely because the learned representation mixes the semantic factor with the variation factor due to their domain-specific correlation, while only the semantic factor *causes* the output. To address the problem, we propose a Causal Semantic Generative model (CSG) based on a causal reasoning so that the two factors are modeled separately, and develop methods for OOD prediction from a *single* training domain, which is common and challenging. The methods are based on the causal invariance principle, with a novel design in variational Bayes for both efficient learning and easy prediction. Theoretically, we prove that under certain conditions, CSG can identify the semantic factor by fitting training data, and this semantic-identification guarantees the boundedness of OOD generalization error and the success of adaptation. Empirical study shows improved OOD performance over prevailing baselines.

1 Introduction

Deep learning has initiated a new era of artificial intelligence where the potential of machine learning models is greatly unleashed. Despite the great success, these methods heavily rely on the assumption that data from training and test domains follow the same distribution (*i.e.*, the IID assumption), while in practice the test domain is often out-of-distribution (OOD), meaning that the test data distribute differently from the training data. Popular models for predicting the output (or label, response, outcome) y from the input (or covariate) x have been found erroneous when confronted with a distribution change, even from an essentially irrelevant perturbation like a position shift or background change for images [91, 6, 102, 41, 2, 27]. These phenomena pose serious concerns on the robustness and trustworthiness of machine learning methods and severely impede them from risk-sensitive scenarios.

Looking into the problem, although deep learning models allow extracting abstract representation for prediction with their powerful approximation capacity, the representation may unconsciously mix up semantic factors s (*e.g.*, shape of an object) and variation factors v (*e.g.*, background, object position) due to a correlation between them (*e.g.*, desks often appear in a workspace background and beds in bedrooms), so the model also relies on the variation factors v for prediction via this correlation. However, this correlation tends to be superficial and spurious (*e.g.*, a desk can also appear in a bedroom, but this does not make it a bed), and may change drastically in a new domain, making the effect from v misleading. So it is desired to learn a representation that identifies s against v.

Formally, the essence of this goal is to leverage *causal relations* for prediction, since the fundamental distinction between s and v is that only s is the cause of y. Causal relations better reflect basic

*Correspondence to: Chang Liu <changliu@microsoft.com>.
†Work done during an internship at Microsoft Research Asia.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

mechanisms of nature. They bring the merit to machine learning that they tend to be universal and *invariant* across domains [97, 87, 93, 77, 16, 96, 98], thus provide the most transferable and reliable information to unseen domains. This causal invariance has been shown to lead to proper domain adaptation [97, 123], lower adaptation cost and lighter catastrophic forgetting [87, 9, 56].

In this work, we propose a Causal Semantic Generative model (CSG) following a causal consideration to separately model the semantic (cause of prediction) and variation latent factors, and develop OOD prediction methods with theoretical guarantees on identifiability and the boundedness of OOD prediction error. Addressing the complaint that OOD prediction and causality methods often require multi-domain or intervention data, we focus on the most common and also challenging tasks where only one *single* training domain is available, including *OOD generalization* and *domain adaptation*, where in the latter, unsupervised test-domain data are additionally available for training. The methods and theory are based on the causal invariance principle, which suggests to share generative mechanisms across domains, while the latent factor distribution (*i.e.*, the prior $p(s, v)$) changes. We argue that this causal invariance is more reliable than *inference invariance* in the other direction adopted by many existing methods [33, 101, 2, 66, 79]. For our method, we design novel and delicate reformulations of the ELBO objective so that we avoid the cost to build and learn two inference models. Theoretically, we prove that under certain conditions, CSG *can identify* the semantic factor on the single training domain, even in presence of an s-v correlation. We further prove the merits from this identification: prediction error is bounded for OOD generalization, and for domain adaptation, the test-domain prior is identifiable which leads to an accurate prediction. To sum up our contributions,

- Up to our knowledge, we are the first to show a theoretical guarantee (under appropriate conditions) to identify the latent cause of prediction (*i.e.*, the semantic factor) on a single training domain, and also the first to show the theoretical benefits of this identification for OOD prediction. The results also contribute to generative representation learning for revealing what is learned.
- We develop effective methods for OOD generalization and domain adaptation, and achieve mostly better performance than prevailing methods on real-world image classification tasks.

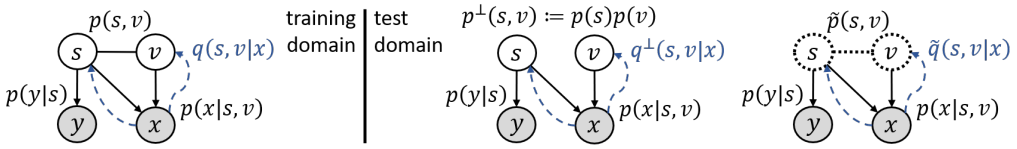
2 Related Work

OOD generalization with causality. There are trials that ameliorate discriminative models towards a causal behavior. Bahadori et al. [4] introduce a regularizer that reweights input dimensions based on their approximated causal effects to the output, and Shen et al. [102] reweight training samples by amortizing causal effects among input samples. Their linear input-output assumption is then extended [4, 41] by learning a representation. Some recent works require identity data (finer than label) and enforce inference invariance via variance minimization [42], or leverage a strong domain knowledge to augment images as an independent intervention on variation factors [79]. These methods introduce no additional generative modeling efforts, at the cost of limited capacity for invariant causal mechanisms.

Domain adaptation/generalization with causality. There are methods developed under various causal assumptions [97, 123] or using learned causal relations [93, 77]. Zhang et al. [123], Gong et al. [35, 36] also consider certain ways of mechanism change. The considered causality is among directly observed variables, which may not well suit general data like image pixels where causality rather lies in the conceptual latent level [75, 10, 59].

To consider latent factors, there are domain adaptation [83, 5, 33, 73, 74] and generalization methods [80, 101, 113] that learn a representation with a domain-invariant marginal distribution. Remarkable results have been achieved. Nevertheless, it is found that this invariance is neither sufficient nor necessary to identify the true semantics or lower the adaptation error ([54, 125]; see also Appx. E). Moreover, these methods are based on inference invariance, which may not be as reliable as causal invariance (see Sec. 3.2).

There are also generative methods for domain adaptation/generalization that model latent factors. Cai et al. [18] and Ilse et al. [49] introduce a semantic factor and a domain-feature factor. They assume the two latent factors are independent in both generative and inference models, which is unrealistic. Correlated factors are then considered [3]. But all these works do not adapt the prior for domain change thus resort to inference invariance. Zhang et al. [121] consider a partially observed manipulation variable, while still assuming its independence from the output in both the joint and posterior, and the adaptation is inconsistent with causal invariance. The above methods also do not show guarantees to identify their latent factors. Teshima et al. [108] leverage causal invariance and



(a) CSG (b) CSG-ind (c) CSG-DA

Figure 1: **(a)** Graphical structure of the proposed CSG. Solid arrows represent causal mechanisms $p(x|s,v)$ and $p(y|s)$, the undirected s-v clique represents a domain-specific prior $p(s,v)$, and the dashed bended arrows represent the inference model $q(s,v|x)$ for learning. **(b, c)** Graphical structures of CSG-ind and CSG-DA for prediction on the *test domain*. An independent prior $p^{\perp\!\!\!\perp}(s,v)$ (constructed from $p(s,v)$) and a new prior $\tilde{p}(s,v)$ (the dotted s-v clique) are introduced reflecting the intervention on the test domain. Respective inference models $q^{\perp\!\!\!\perp}(s,v|x)$ and $\tilde{q}(s,v|x)$ are also shown. All three models share the same causal mechanisms $p(x|s,v)$ and $p(y|s)$.

adapt the prior, yet also assume latent independence and do not separate the semantic factor. They require some supervised test-domain data, and their deterministic and invertible mechanism also indicates inference invariance. In addition, most domain generalization methods require *multiple* training domains, with exceptions [89] that still seek to augment domains. In contrast, CSG leverages causal invariance, and has *guarantee* to identify the semantic factor from a *single training domain*, even with a *correlation* to the variation factor.

Disentangled latent representations is also of interest in unsupervised learning. Despite empirical success [22, 43, 21], Locatello et al. [70] conclude that it is impossible to guarantee the disentanglement in unsupervised settings. Subsequent works then introduce ways of supervision like a few latent variable observations [71] or sample similarity [20, 72, 104]. Identifiable VAE [57] and extensions [58, 117] leverage the data of a cause variable of the latent variables and have established theoretical guarantees under a diversity condition. But the works do not depict domain change thus not suitable for OOD prediction. Instead of disentangling latent factors, we focus on identifying the semantic factor s (Sec. 5.1) and its benefit for OOD prediction. Appx. D shows more related work.

3 The Causal Semantic Generative Model

To develop the model soberly based on causality, we require its formal definition: *two variables have a causal relation, denoted as "*cause $\rightarrow$ effect*", if intervening the* cause *(by changing external variables out of the considered system) may change the* effect*, but not* vice versa [85, 88]. We follow this definition to build our model (Fig. 1a) by analyzing the example that an photographer takes a photo in a scene as x and labels it as y. Appx. C provides more explanations under other perspectives.

(1) It is likely that neither $y \rightarrow x$ (*e.g.*, intervening the label with noise by distracting the photographer does not change the image) nor $x \rightarrow y$ holds (*e.g.*, intervening an image by breaking a camera sensor unit does not change how the photographer labels it), as also argued in [88, Sec. 1.4; 59]. So we introduce a latent variable z to capture factors with causal relations. Also for this reason, we need a generative model (vs. discriminative model that only learns $x \rightarrow y$).

(2) The latent variable z as underlying generating factors (*e.g.*, object shape and texture, background and illumination during imaging) is plausible to cause both x (*e.g.*, changing object shape or background makes a different image, but breaking the camera does not change the shape or background) and y (*e.g.*, the photographer would give a different label if the object shape had been different, but noise-corrupting the label does not change the shape). So we orient the edges in the generative direction $z \rightarrow (x,y)$, as also adopted in [78, 88, 108]. This is in contrast to prior works [18, 49, 48, 19] that treat y as the cause of a semantic factor, which, when y is also a noisy observation, makes unreasonable implications (*e.g.*, adding noise to the labels in a dataset automatically changes object features and consequently the images, and changing the object features does not change the label). This difference is also discussed in [88, Sec. 1.4; 59].

(3) We attribute all x-y relation to the existence of some latent factor [68, "purely common cause"; 51] and exclude x-y edges. This can be achieved as long as z holds sufficient information of data (*e.g.*, with shape, background *etc.* fixed, breaking the camera does not change the label, and noise-corrupting the label does not change the image). Promoting this structure reduces arbitrariness in explaining x-y relation thus helps identify (part of) z. This is in contrast to prior works [63, 121, 19] that treat y as a cause of x as no latent variable is introduced between.

(4) Not all latent factors are the causes of y (*e.g.*, changing the shape may alter the label, while changing the background does not). We thus split the latent variable as $z = (s, v)$ and remove the $v \to y$ edge, where s represents the *semantic* factor that causes y, and v describes the *variation* or diversity in generating x. This formalizes the intuition on the concepts in Introduction (Sec. 1).

(5) The two factors s and v often have a relation (*e.g.*, a desk/bed shape tends to appear with a workspace/bedroom background), but it is usually a spurious correlation (*e.g.*, putting a desk in a bedroom does not automatically change the room as a workspace, nor does it turn the desk into a bed). So we keep the undirected s-v edge. This is in contrast to prior works [18, 49, 121, 108, 79] which assume independent latent variables. Although v is not a cause of y, modeling it explicitly is worth the effort since otherwise it would still be implicitly mixed into s anyway through the s-v correlation. We summarize these conclusions in the following definition.

Definition 1 (CSG)**.** A *Causal Semantic Generative Model* (CSG), $p := \langle p(s, v), p(x|s, v), p(y|s) \rangle$, is a generative model on data variables $x \in \mathcal{X} \subseteq \mathbb{R}^{d_\mathcal{X}}$ and $y \in \mathcal{Y}$ with semantic $s \in \mathcal{S} \subseteq \mathbb{R}^{d_\mathcal{S}}$ and variation $v \in \mathcal{V} \subseteq \mathbb{R}^{d_\mathcal{V}}$ latent variables, following the graphical structure shown in Fig. 1a.

3.1 The Causal Invariance Principle

Through the above process, we see that the s-v correlation embodied in the prior $p(s, v)$ tends to change across domains. Under a causal view, this means that the domain change comes from a (soft) intervention on s or v or both, leading to a different prior. On the other hand, the generative processes are likely causal mechanisms, so they enjoy the celebrated Independent Causal Mechanisms principle [88, 98] indicating that they are unaffected under the intervention on prior. This leads to the following causal invariance principle for CSG.

Principle 2 (causal invariance)**.** The causal generative mechanisms $p(x|s, v)$ and $p(y|s)$ in CSG are invariant across domains, and the change of prior $p(s, v)$ is the only source of domain change.

This invariance reflects the universality of basic laws of nature and is considered in some prior works [97, 88, 10, 16]. Other works instead introduce domain index [18, 49, 48, 19] or manipulation variables [121, 57, 58] to model distribution change explicitly. They then require multiple training domains or additional observations, while such changes can also be explained under causal invariance as long as the latent variables include all changing factors.

3.2 Comparison with Inference Invariance

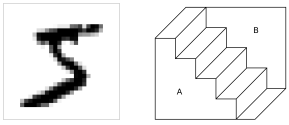


Figure 2: Examples of noisy (left) or degenerate (right) generating mechanisms that lead to ambiguity in inference. Left: handwritten digit that may be generated as either "3" or "5". Right: Schröder's stairs that may be generated with either A or B being the nearer surface. Inference results notably rely on the prior on the digits/surfaces, which is domain-specific.

Most domain adaptation and generalization methods (incl. domain-invariant-representation based [33, 101], invariant-latent-predictor based [2, 66, 79]) use a shared representation extractor across domains. This effectively assumes the invariance in the other direction, *i.e.* inferring latent factors z from observed data x. We note in its supportive examples (*e.g.*, inferring object position from image, extracting the fundamental frequency from audio), the causal mechanism $p(x|z)$ is nearly deterministic and invertible such that it preserves the information of z. Formally, for a given x, only one single z value achieves a positive $p(x|z)$ while all other values lead to zero. The inferred representation given by the posterior via the Bayes rule $p(z|x) \propto p(z)p(x|z)$ then concentrates on this z value, which is determined by the causal mechanism $p(x|z)$ alone, regardless of the domain-specific prior $p(z)$. Causal invariance then implies inference invariance.

In more general cases, the causal mechanism may be noisy or degenerate (Fig. 2), such that there are multiple z values that give a positive $p(x|z)$, *i.e.* they all could generate the same x. Inference is then ambiguous, and the posterior relies on the prior to choose from these z values. Since the prior changes across domains (*e.g.*, different labelers have different mindset), the inference rule then *changes by nature* and is not invariant,[3] while the causal invariance is rather more fundamental and reliable. To leverage causal invariance, we use a different prior for the test domain (CSG-ind and CSG-DA), which gives a different and more reliable prediction than following inference invariance.

[3]Particularly, although Mitrovic et al. [79] consider a similar causal structure and promote the invariance of $p(y|s)$, s actually depends on v for a given x, even when they are independent in the prior. So $p(s|x)$ must depend on the domain-specific $p(v)$, and a domain-invariant representation extractor does not exist.

4 Method

We now develop methods based on variational Bayes [55, 62] for OOD generalization and domain adaptation using CSG. Appx. F.1 shows all details.

4.1 Method for OOD Generalization

For OOD generalization, one only has supervised data from the underlying data distribution $p^*(x, y)$ on the *training domain*. Fitting a CSG $p := \langle p(s, v), p(x|s, v), p(y|s)\rangle$ to data by maximizing likelihood $\mathbb{E}_{p^*(x,y)}[\log p(x, y)]$ is intractable, since $p(x, y) := \int p(s, v, x, y)\,\mathrm{d}s\mathrm{d}v$ where $p(s, v, x, y) := p(s, v)p(x|s, v)p(y|s)$, is hard to estimate. The Evidence Lower BOund (ELBO) $\mathcal{L}_{p,\,q_{s,v|x,y}}(x, y) := \mathbb{E}_{q(s,v|x,y)}[\log \frac{p(s,v,x,y)}{q(s,v|x,y)}]$ [55, 112] is a tractable surrogate with the help of an inference model $q(s, v|x, y)$ that enjoys easy sampling and density evaluation. It is known that $\max_{q_{s,v|x,y}} \mathcal{L}_{p,\,q_{s,v|x,y}}(x, y)$ drives $q(s, v|x, y)$ towards the posterior $p(s, v|x, y) := \frac{p(s,v,x,y)}{p(x,y)}$, meanwhile makes $\mathcal{L}_{p,\,q_{s,v|x,y}}(x, y)$ a tighter lower bound of $\log p(x, y)$ for optimizing CSG p.

However, the subtlety with supervised learning is that prediction is still hard, as the introduced model $q(s, v|x, y)$ does not help estimate $p(y|x)$. To address this, we propose to employ an auxiliary model $q(s, v, y|x)$ targeting $p(s, v, y|x)$. It allows easy sampling of y given x for prediction, and can also serve as the required inference model: $q(s, v|x, y) = \frac{q(s,v,y|x)}{q(y|x)}$, where $q(y|x) := \int q(s, v, y|x)\,\mathrm{d}s\mathrm{d}v$ is also determined by $q(s, v, y|x)$. The ELBO objective $\mathbb{E}_{p^*(x,y)}[\mathcal{L}_{p,\,q_{s,v|x,y}}(x, y)]$ then becomes:

$$\mathbb{E}_{p^*(x)}\mathbb{E}_{p^*(y|x)}[\log q(y|x)] + \mathbb{E}_{p^*(x)}\mathbb{E}_{q(s,v,y|x)}\Big[\frac{p^*(y|x)}{q(y|x)}\log\frac{p(s, v, x, y)}{q(s, v, y|x)}\Big]. \tag{1}$$

As a functional of $q(s, v, y|x)$ (instead of $q(s, v|x, y)$) and the CSG p, this objective also drives them towards their targets: the first term is the negative of the standard cross entropy (CE) loss which drives $q(y|x)$ towards $p^*(y|x)$, and once this is achieved, the second term becomes the expected ELBO $\mathbb{E}_{p^*(x)}[\mathcal{L}_{p,\,q_{s,v,y|x}}(x)]$ that drives $q(s, v, y|x)$ towards $p(s, v, y|x)$ and $p(x)$ towards $p^*(x)$. Furthermore, as the target of $q(s, v, y|x)$ factorizes as $p(s, v, y|x) = p(s, v|x)p(y|s)$ (due to Fig. 1a) where $p(y|s)$ is already known (part of the CSG), we can instead employ a lighter inference model $q(s, v|x)$ for the minimally intractable component $p(s, v|x)$ therein, and use $q(s, v|x)p(y|s)$ as $q(s, v, y|x)$. This turns the objective Eq. (1) to:

$$\max_{p,\,q_{s,v|x}} \mathbb{E}_{p^*(x,y)}\Big[\log q(y|x) + \frac{1}{q(y|x)}\mathbb{E}_{q(s,v|x)}\Big[p(y|s)\log\frac{p(s, v)p(x|s, v)}{q(s, v|x)}\Big]\Big], \tag{2}$$

where $q(y|x) := \mathbb{E}_{q(s,v|x)}[p(y|s)]$. The expectations can be estimated by Monte Carlo after applying the reparameterization trick [62]. This is the basic CSG method.

CSG-ind To actively improve OOD generalization performance, we consider using an **ind**ependent prior $p^{\perp\!\!\!\perp}(s, v) := p(s)p(v)$ for prediction in the *test domain* (Fig. 1b), where $p(s)$ and $p(v)$ are the marginals of the training-domain prior $p(s, v)$. Intuitively, $p^{\perp\!\!\!\perp}(s, v)$ discards the spurious correlation between s and v on the training domain (*e.g.*, the "desk-workspace", "bed-bedroom" association), and promotes a cautious neutral belief on the unknown test-domain correlation in defence against all possibilities (*e.g.*, a "desk-bedroom", "bed-workspace" association). Formally, $p^{\perp\!\!\!\perp}(s, v)$ has a larger entropy than $p(s, v)$ [24, Thm. 2.6.6], so it reduces training-domain-specific information and encourages reliance on the causal mechanisms for better generalization. It also amounts to applying the do-operator [85] to Fig. 1a, representing a randomized experiment by independently soft-intervening s or v. In this way, causal invariance is properly leveraged, making a different and more reliable prediction than following inference invariance. Our theory below also shows that $p^{\perp\!\!\!\perp}(s, v)$ leads to a smaller generalization error bound (Thm. 6 Remark).

Methodologically, we need the test-domain inference model $q^{\perp\!\!\!\perp}(s, v|x)$ for prediction $p^{\perp\!\!\!\perp}(y|x) \approx \mathbb{E}_{q^{\perp\!\!\!\perp}(s,v|x)}[p(y|s)]$, but also need $q(s, v|x)$ for learning on the training domain. To save the cost of building and learning two inference models, we propose to use $q^{\perp\!\!\!\perp}(s, v|x)$ to represent $q(s, v|x)$. Noting that their targets are related by $p(s, v|x) = \frac{p(s,v)}{p^{\perp\!\!\!\perp}(s,v)}\frac{p^{\perp\!\!\!\perp}(x)}{p(x)}p^{\perp\!\!\!\perp}(s, v|x)$, we formulate $q(s, v|x) = \frac{p(s,v)}{p^{\perp\!\!\!\perp}(s,v)}\frac{p^{\perp\!\!\!\perp}(x)}{p(x)}q^{\perp\!\!\!\perp}(s, v|x)$ accordingly, so that this $q(s, v|x)$ achieves its target if and only if $q^{\perp\!\!\!\perp}(s, v|x)$

does. The objective Eq. (1) then becomes:

$$\max_{p,\, q^{\perp\!\!\!\perp}_{s,v|x}} \mathbb{E}_{p^*(x,y)}\Big[\log \pi(y|x) + \frac{1}{\pi(y|x)}\mathbb{E}_{q^{\perp\!\!\!\perp}(s,v|x)}\Big[\frac{p(s,v)}{p^{\perp\!\!\!\perp}(s,v)}p(y|s)\log\frac{p^{\perp\!\!\!\perp}(s,v)p(x|s,v)}{q^{\perp\!\!\!\perp}(s,v|x)}\Big]\Big], \quad (3)$$

where $\pi(y|x) := \mathbb{E}_{q^{\perp\!\!\!\perp}(s,v|x)}\big[\frac{p(s,v)}{p^{\perp\!\!\!\perp}(s,v)}p(y|s)\big]$. (Note $p^{\perp\!\!\!\perp}(s,v)$ is determined by $p(s,v)$ in the CSG p.)

4.2 Method for Domain Adaptation

In domain adaptation, one also has unsupervised data from the underlying data distribution $\tilde{p}^*(x)$ on the *test domain*. We can leverage them for better prediction. According to the causal invariance principle (2), we only need a new prior $\tilde{p}(s,v)$ for the test-domain CSG $\tilde{p} := \langle \tilde{p}(s,v), p(x|s,v), p(y|s)\rangle$ (Fig. 1c). Fitting test-domain data can be done through the standard ELBO objective with the test-domain inference model $\tilde{q}(s,v|x)$:

$$\max_{\tilde{p},\, \tilde{q}_{s,v|x}} \mathbb{E}_{\tilde{p}^*(x)}[\mathcal{L}_{\tilde{p},\, \tilde{q}_{s,v|x}}(x)], \text{where } \mathcal{L}_{\tilde{p},\, \tilde{q}_{s,v|x}}(x) = \mathbb{E}_{\tilde{q}(s,v|x)}\Big[\log\frac{\tilde{p}(s,v)p(x|s,v)}{\tilde{q}(s,v|x)}\Big]. \quad (4)$$

Prediction is given by $\tilde{p}(y|x) \approx \mathbb{E}_{\tilde{q}(s,v|x)}[p(y|s)]$. Similar to the CSG-ind case, we still need $q(s,v|x)$ for fitting training-domain data, and we can also avoid a separate $q(s,v|x)$ model by representing it using $\tilde{q}(s,v|x)$. Following the same relation between their targets, we let $q(s,v|x) = \frac{\tilde{p}(x)}{p(x)}\frac{p(s,v)}{\tilde{p}(s,v)}\tilde{q}(s,v|x)$, which reformulates the same training-domain objective Eq. (1) as:

$$\max_{p,\, \tilde{q}_{s,v|x}} \mathbb{E}_{p^*(x,y)}\Big[\log \pi(y|x) + \frac{1}{\pi(y|x)}\mathbb{E}_{\tilde{q}(s,v|x)}\Big[\frac{p(s,v)}{\tilde{p}(s,v)}p(y|s)\log\frac{\tilde{p}(s,v)p(x|s,v)}{\tilde{q}(s,v|x)}\Big]\Big], \quad (5)$$

where $\pi(y|x) := \mathbb{E}_{\tilde{q}(s,v|x)}\big[\frac{p(s,v)}{\tilde{p}(s,v)}p(y|s)\big]$. The resulting method, termed CSG-DA, solves both optimization problems Eqs. (4, 5) simultaneously.

4.3 Implementation and Model Selection

To implement the three CSG methods, we only need one inference model in each. Appx. F.2 shows its construction from a general discriminative model (*e.g.*, how to select its hidden nodes as s and v). In practice x often has a much larger dimension than y, making the first supervision term overwhelmed by the second unsupervised term in Eqs. (2,3,5). So we downscale the second term.

As recently emphasized [39], an OOD method should include a model selection method, since it is nontrivial and significantly affects performance [95, 120]. For our methods, we use a validation set from the *training domain* for model selection. This complies with the OOD setup, and is also suggested by our theory below which gives guarantees based on a good fit to the training-domain data distribution. For CSG-ind/DA, the learned predictor targets the *test* domain, so we *do not* use it directly for evaluating validation accuracy, but by normalizing $\pi(y|x)$. Appx. F.3 shows details.

5 Theory

We now establish theory for the identification of the semantic factor (cause of prediction) and subsequent merits for OOD generalization and domain adaptation. We focus on the distribution-level generalization instead of from finite samples to unseen samples under the same distribution, so we only consider the infinite-data regime. Appx. A shows all the proofs and auxiliary theory.

Latent variable identification is hard [65, 81, 116, 70] as it is beyond observational relations [51, 88]. Assumptions are thus required to draw definite conclusions.

Assumption 3. (**Additive noise**) There exist nonlinear functions f and g with bounded derivatives up to the third-order, and independent random variables μ and ν, such that $p(x|s,v) = p_\mu(x - f(s,v))$, and $p(y|s) = p_\nu(y - g(s))$ for continuous y or $p(y|s) = \mathrm{Cat}(y|g(s))$ for categorical y.
(**Bijectivity**) Assume f is bijective and g is injective.

The additive noise assumption is widely adopted in causal discovery [51, 17]. It disables expressing the same joint in the other direction [122, Thm. 8; 86, Prop. 23] so that CSG unnecessarily indicates inference invariance. For this reason, we exclude GAN [37] and flow-based [61] implementations. Bijectivity is a common assumption for identifiability [51, 100, 57, 68]. It is sufficient [86, Prop. 17; 88, Prop. 7.4] for the more fundamental [86, Prop. 7; 88, p.109] requirement of causal minimality [86, p.2012; 88, Def. 6.33]. Particularly, s and v may otherwise have dummy dimensions that f and g simply ignore, raising another ambiguity against identifiability. On the other hand, according to the

commonly acknowledged manifold hypothesis [115, 31], we can take $\mathcal{X}$ as the lower-dimensional data manifold and such a bijection exists as a coordinate map, which is an injection to the original data space and also allows $d_{\mathcal{S}} + d_{\mathcal{V}} < d_{\mathcal{X}}$.

5.1 Identifiability Theory

We first formalize the goal of identifying the semantic factor.

Definition 4 (semantic-identification)**.** We say a learned CSG p is *semantic-identified*, if there exists a homeomorphism[4] Φ on $\mathcal{S} \times \mathcal{V}$, such that **(i)** its output dimensions in $\mathcal{S}$ is constant of v: $\Phi^{\mathcal{S}}(s, v) = \Phi^{\mathcal{S}}(s, v'), \forall v, v' \in \mathcal{V}$ (hence denote $\Phi^{\mathcal{S}}(s, v)$ as $\Phi^{\mathcal{S}}(s)$), and **(ii)** it is a *reparameterization* of the ground-truth CSG p^*: $\Phi_{\#}[p^*_{s,v}] = p_{s,v}$, $p^*(x|s, v) = p(x|\Phi(s, v))$ and $p^*(y|s) = p(y|\Phi^{\mathcal{S}}(s))$.

Here, $\Phi_{\#}[p^*_{s,v}]$ denotes the pushed-forward distribution[5] of $p^*_{s,v}$ by Φ, *i.e.* the distribution of $\Phi(s, v)$ when $(s, v) \sim p^*_{s,v}$. As the ground-truth CSG could at most provide its information via the data distribution $p^*(x, y)$, a well-learned CSG that achieves $p(x, y) = p^*(x, y)$ still has the degree of freedom in parameterizing (s, v). This is described by this reparameterization Φ (Appx. Lemma 9). At the heart of the definition, the v-constancy of $\Phi^{\mathcal{S}}$ implies that Φ is *semantic-preserving*: the learned model *does not mix* the ground-truth v into its s, so that the learned s holds equivalent information to the ground-truth s. The definition can thus be seen as the semantic equivalence (Appx. Def. 10, Prop. 14) to the ground-truth CSG p^*.

For related concepts, this identification cannot be characterized by the *statistical independence* between s and v (vs. [18, 49, 121]), which is not sufficient [70] nor necessary (due to the existence of spurious correlation). It is also weaker than *disentanglement* [44, 11], which additionally requires the learned v to be constant of the ground-truth s. The following theorem shows that semantic-identification can be achieved on a single domain under certain conditions.

Theorem 5 (semantic-identifiability)**.** *With Assumption 3, a CSG p is semantic-identified, if it is well-learned such that $p(x, y) = p^*(x, y)$, under the conditions that $\log p(s, v)$ and $\log p^*(s, v)$ are bounded up to the second-order, and that*[6] **(i)** *$1/\sigma_\mu^2 \to \infty$ where $\sigma_\mu^2 := \mathbb{E}[\mu^\top \mu]$,* ***or*** **(ii)** *$p_\mu$ (*e.g.*, a Gaussian) has an a.e. non-zero characteristic function.*

Remarks. (1) (**Condition and Intuition**) Compared with the multi-domain case [87, 93, 2], identifiability on a single training domain comes at a cost and requires certain conditions. One may imagine that in some extreme cases *e.g.*, all desks appear in workspace and all beds in bedrooms, it is impossible to distinguish whether y labels the object or the background (unlearnable OOD problem [119]). The theorem finds an *appropriate condition* that excludes such cases: when $\log p^*(s, v)$ is bounded, deterministic s-v relations are not allowed as they concentrate $p^*(s, v)$ on a lower-dimensional subspace in $\mathcal{S} \times \mathcal{V}$ thus make it unbounded.

It also leads to the *intuition of identifiability*: a bounded $\log p^*(s, v)$ indicates a stochastic s-v relation, so mixing the ground-truth v into the learned s makes the inference of s more noisy due to the intrinsic diversity/uncertainty of this v. As prediction is made via the inferred s, this worsens prediction accuracy thus violates the "well-learned" requirement. Compared with discriminative models, CSG makes more faithful inference, and its causal structure leads to a proper description of domain change.

(2) In condition **(i)**, $1/\sigma_\mu^2$ measures the *intensity* of the causal mechanism $p(x|s, v)$. When it is large, the "strong" $p(x|s, v)$ helps disambiguating values of (s, v) in generating a given x. The formal version in Appx. Thm. 5' shows a quantitative reference for large enough intensity, and Appx. B gives a non-asymptotic extension showing how the intensity trades-off the tolerance of equalities in Def. 4. Condition **(ii)** goes beyond inference invariance. It roughly implies that different (s, v) values a.s. produce different $p(x|s, v)$, so their roles in generating x become clear which helps identification.

(3) The theorem does not contradict the impossibility result by Locatello et al. [70], which considers disentangling each latent dimension with an unconstrained $(s, v) \to (x, y)$, while we only identify s as a whole, with the $v \to y$ edge removed which breaks the s-v symmetry.

[4] A transformation is a homeomorphism if it is a continuous bijection with continuous inverse.

[5] The definition of $\Phi_{\#}[p^*_{s,v}]$ requires Φ to be measurable. This is satisfied by the continuity of Φ as a homeomorphism (as long as the Borel σ-field is considered) [13, Thm. 13.2].

[6] To be precise, the conclusions are that the equalities in Def. 4 hold asymptotically in the limit $1/\sigma_\mu^2 \to \infty$ for condition **(i)**, and hold a.e. for condition **(ii)**.

5.2 OOD Generalization Theory

Now we show the benefit of semantic-identification for OOD generalization that the prediction error is bounded. Note the optimal predictor $\tilde{\mathbb{E}}^*[y|x]$ [7] on the test domain is defined by the corresponding ground-truth CSG $\tilde{p}^*$, which differs from p^* only in the test-domain prior $\tilde{p}^*(s,v)$ (Principle 2).

Theorem 6 (OOD generalization error)**.** [8] *With Assumption 3, for a semantic-identified CSG p on the training domain with semantic-preserving reparameterization Φ, we have up to $O(\sigma_\mu^4)$,*

$$\mathbb{E}_{\tilde{p}^*(x)} \|\mathbb{E}[y|x] - \tilde{\mathbb{E}}^*[y|x]\|_2^2 \leqslant \sigma_\mu^4 B_{f^{-1}}'^4 B_g'^2 \, \mathbb{E}_{\tilde{p}_{s,v}} \left\| \nabla \log(\tilde{p}_{s,v}/p_{s,v}) \right\|_2^2, \tag{6}$$

where $B'_{f^{-1}}$ and B'_g bound the 2-norms[9] of the Jacobians of f^{-1} and g, respectively, and $\tilde{p}_{s,v} := \Phi_\#[\tilde{p}^*_{s,v}]$ is the test-domain prior under the parameterization of the CSG p.

In the bound, the term $\mathbb{E}_{\tilde{p}_{s,v}} \left\| \nabla \log(\tilde{p}_{s,v}/p_{s,v}) \right\|_2^2$ is the Fisher divergence measuring the difference between the two priors. As the prior change is the only source of domain change, this term also measures the "OODness" in terms of the effect on prediction. The bound also shows that when the causal mechanism $p(x|s,v)$ is strong (small σ_μ), it dominates prediction over the prior change, as the generalization error becomes small. Compared with other methods, using a CSG enforces causal invariance, so the boundedness of OOD generalization error becomes more plausible in practice.

Remark. The bound also shows the advantage of CSG-ind (Sec. 4.1). The Fisher divergence is revealed [28] to have a similar behavior as the forward KL divergence $p_{s,v} \mapsto \mathrm{KL}(\tilde{p}_{s,v} \| p_{s,v})$ that it is very sensitive to the insufficient coverage of $p_{s,v}$ on the support of $\tilde{p}_{s,v}$ [46, 109], since $\log(\tilde{p}_{s,v}/p_{s,v})$ is infinitely large on the uncovered region. As the independent prior $p^{\perp\!\!\!\perp}_{s,v}$ has a larger support than $p_{s,v}$, it is less likely to miss the support of $\tilde{p}_{s,v}$, so it induces a generally smaller Fisher divergence. CSG-ind thus generally has a smaller OOD generalization error bound than CSG.

5.3 Domain Adaptation Theory

CSG-DA (Sec. 4.2) learns a new prior $\tilde{p}_{s,v}$ by fitting unsupervised test-domain data, with causal mechanisms shared. If the mechanisms are semantic-identified, the ground-truth test-domain prior $\tilde{p}^*_{s,v}$ can also be identified under the learned parameterization, and prediction is made precise.

Theorem 7 (domain adaptation error)**.** *With conditions of Thm. 5, for a semantic-identified CSG p on the training domain with semantic-preserving reparameterization Φ, if its new prior $\tilde{p}_{s,v}$ is well-learned such that $\tilde{p}(x) = \tilde{p}^*(x)$, then $\tilde{p}_{s,v} = \Phi_\#[\tilde{p}^*_{s,v}]$, and $\tilde{\mathbb{E}}[y|x] = \tilde{\mathbb{E}}^*[y|x]$ for any $x \in \mathrm{supp}(\tilde{p}^*_x)$.*

Different from existing domain adaptation bounds (Appx. E), Theorems 6,7 allow different inference models in the two domains, thus go beyond inference invariance.

6 Experiments

For OOD generalization baselines, there is not much choice beyond the standard CE loss optimization, as domain adaptation methods require test-domain data and most domain generalization methods degenerate to CE with one training domain. The exception within our scope is a causal discriminative method CNBB [41]. For domain adaptation, we consider well-acknowledged methods DANN [33], DAN [73], CDAN [74] and recent compelling methods MDD [124] and BNM [25] (shown in Appx. Tables 2,3). Appx. G shows more details, results, and discussions. [10]

Shifted-MNIST. We first consider an OOD prediction task on MNIST to classify digits "0"s and "1"s. To make a spurious correlation, in the training data, we horizontally shift each "0" at random by $\delta_0 \sim \mathcal{N}(-5, 1^2)$ pixels, while each "1" by $\delta_1 \sim \mathcal{N}(5, 1^2)$ pixels. We consider two test domains with different digit-position distributions: each digit is not moved $\delta_0 = \delta_1 = 0$ in the first, and is shifted at random by $\delta_0, \delta_1 \sim \mathcal{N}(0, 2^2)$ pixels in the second. We implement all methods using a multilayer perceptron which is not naturally shift invariant. We use a larger architecture for non-generative methods to compensate the additional generative component of generative methods.

The performance is shown in Table 1(top 2 rows). For OOD generalization, CE is misled by the more noticeable position factor due to the spurious correlation to digits, and resorts to random guess (even

[7] For categorical y, the expectation of y is taken under the one-hot representation.

[8] See Appx. Thm. 6' for the formal version.

[9] As the induced operator norm for matrices (not the Frobenius norm).

[10] Codes are available at `https://github.com/changliu00/causal-semantic-generative-model`.

Table 1: Test accuracy (%) by various methods (ours in bold) for OOD generalization (left 4 cols) and domain adaptation (right 5 cols) on Shifted-MNIST (top 2 rows), ImageCLEF-DA (middle 4 rows) and PACS (bottom 4 rows) datasets. Averaged over 10 runs. Appx. Tables 2,3 show more results.

task	CE	CNBB	**CSG**	**CSG-ind**	DANN	DAN	CDAN	MDD	**CSG-DA**
$\delta_0 = \delta_1 = 0$	42.9±3.1	54.7±3.3	81.4±7.4	**82.6±4.0**	40.9±3.0	40.4±2.0	41.0±0.5	41.9±0.8	**97.6±4.0**
$\delta_0,\delta_1 \sim \mathcal{N}(0,2^2)$	47.8±1.5	59.2±2.4	61.7±3.6	**62.3±2.2**	46.2±0.7	45.6±0.7	46.3±0.6	45.8±0.3	**72.0±9.2**
C→P	65.5±0.3	72.7±1.1	73.6±0.6	**74.0±1.3**	74.3±0.5	69.2±0.4	74.5±0.3	74.1±0.7	**75.1±0.5**
P→C	91.2±0.3	91.7±0.2	92.3±0.4	**92.7±0.2**	91.5±0.6	89.8±0.4	**93.5±0.4**	92.1±0.6	**93.4±0.3**
I→P	74.8±0.3	75.4±0.6	76.9±0.3	**77.2±0.2**	75.0±0.6	74.5±0.4	76.7±0.3	76.8±0.4	**77.4±0.3**
P→I	83.9±0.1	88.7±0.5	90.4±0.3	**90.9±0.2**	86.0±0.3	82.2±0.2	90.6±0.3	90.2±1.1	**91.1±0.5**
others→**P**	**97.8±0.0**	96.9±0.2	97.7±0.2	**97.8±0.2**	97.6±0.2	97.6±0.4	97.0±0.4	97.6±0.3	**97.9±0.2**
others→**A**	88.1±0.1	73.1±0.3	**88.5±0.6**	**88.6±0.6**	85.9±0.5	84.5±1.2	84.0±0.9	88.1±0.8	**88.8±0.7**
others→**C**	77.9±1.3	50.2±1.2	84.4±0.9	**84.6±0.8**	79.9±1.4	81.9±1.9	78.5±1.5	83.2±1.1	**84.7±0.8**
others→**S**	79.1±0.9	43.3±1.2	80.7±1.0	**81.1±1.2**	75.2±2.8	77.4±3.1	71.8±3.9	80.2±2.2	**81.4±0.8**

worse) when position is not informative for prediction. CNBB ameliorates the position confusion, but not as thoroughly without modeling causal mechanisms. In contrast, our CSG gives more genuine predictions in unseen domains, thanks to the identification of the semantic factor. CSG-ind performs even better, justifying the merit of using an independent prior for prediction. For domain adaptation, CSG-DA achieves the best results. Existing adaptation methods even worsen the result (negative transfer), as the misleading position representation gets strengthened on the unsupervised test data. CSG is benefited from adaptation in a proper way that identifies the semantic factor.

ImageCLEF-DA is a standard benchmark for domain adaptation [1]. It has 12 classes and three domains of real-world images: **C**altech-256, **I**mageNet, **P**ascal VOC 2012. We select four OOD prediction tasks **C**↔**P**, **I**↔**P** that have not seen good enough results. We adopt the same setup as [74]. As shown in Table 1(middle 4 rows), CSG-ind again achieves the best OOD generalization results, and even outperforms some domain adaptation methods. Our CSG also outperforms the baselines mostly. For domain adaptation, CSG-DA is the best in most cases and on par with the best in others.

PACS is a more recent benchmark dataset [69]. It has 7 classes and is named after its four domains: **P**hoto, **A**rt, **C**artoon, **S**ketch; each contains images of a certain style. We follow the same setup as [39]; particularly, we pool together all domains but the test one as the single training domain. Results in Table 1(bottom 4 rows) show the same trend. CSG-DA even outperforms most domain generalization methods reported in [39], which are fed with more information. Appx. Tables 2,3 also show the results on an even larger dataset **VLCS** [30], which present a similar observation.

Visualization. Appx. Fig. 5 visualizes the learned models using LIME [91]. The results show our methods focus more on the semantic regions and shapes, indicating a causal representation is learned.

Dataset analysis. The results indicate our methods are more powerful on shifted-MNIST and PACS (and VLCS) than ImageCLEF-DA. This meets the intuition of identifiability (Thm. 5 Remark **(1)**): the random position or pooled training domain shows a diverse v for each s (while with a misleading spurious correlation), so identification is better guaranteed to overcome the spurious correlation.

Ablation study. To show the benefit of modeling s and v separately, we compare with a counterpart of CSG that treats s and v as a whole (equivalently, $v \to y$ is kept; see Appx. F.1.4 for method details). Appx. Tables 2,3 show that our methods outperform this baseline in all cases. This shows the separate modeling makes CSG consciously drive semantic representation into the dedicated variable s.

7 Conclusion and Discussion

We propose a Causal Semantic Generative model for single-domain OOD prediction tasks, which builds upon a causal reasoning, and models the semantic (cause of prediction) and variation factors separately. By the causal invariance principle, we develop novel and efficient learning and prediction methods, and prove the semantic-identifiability and the subsequent bounded generalization error and the success of adaptation. Experiments show the improved performance over prevailing baselines.

Notably, we answered the questions in the recent farseeing paper [98] on causal representation learning: we found an appropriate condition under which "causal variables can be recovered", and provided "compelling evidence on the advantages (of causal modeling) in terms of generalization". Also, separating semantics from variation extends to broader examples. Neural nets are found to

change their prediction under a different texture [34, 15]. Adversarial vulnerability [107, 38, 67] extends variation factors to human-imperceptible features, *i.e.* adversarial noise, which is found to have a strong correlation to the semantics [50]. The separation also matters for fairness when a sensitive variation factor may affect prediction. This work also inspires the dual connection between causal representation learning ("fill in the blanks" given a graph) and causal discovery ("link the nodes" given observed variables). Our theory shows the identifiability condition for causal discovery (the additive noise assumption) also makes causal representation identifiable. Studying the general connection between the two tasks is an interesting future work.

References

[1] The imageclef-da challenge 2014. https://www.imageclef.org/2014, 2014.

[2] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[3] Y. Atzmon, F. Kreuk, U. Shalit, and G. Chechik. A causal view of compositional zero-shot recognition. *Advances in Neural Information Processing Systems*, 33, 2020.

[4] M. T. Bahadori, K. Chalupka, E. Choi, R. Chen, W. F. Stewart, and J. Sun. Causal regularization. *arXiv preprint arXiv:1702.02604*, 2017.

[5] M. Baktashmotlagh, M. T. Harandi, B. C. Lovell, and M. Salzmann. Unsupervised domain adaptation by domain invariant projection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 769–776, 2013.

[6] S. Beery, G. Van Horn, and P. Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 456–473, 2018.

[7] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

[8] S. Ben-David, T. Lu, T. Luu, and D. Pál. Impossibility theorems for domain adaptation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 129–136, 2010.

[9] Y. Bengio, T. Deleu, N. Rahaman, N. R. Ke, S. Lachapelle, O. Bilaniuk, A. Goyal, and C. J. Pal. A meta-transfer objective for learning to disentangle causal mechanisms. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020.

[10] M. Besserve, N. Shajarisales, B. Schölkopf, and D. Janzing. Group invariance principles for causal generative models. In *International Conference on Artificial Intelligence and Statistics*, pages 557–565. PMLR, 2018.

[11] M. Besserve, A. Mehrjou, R. Sun, and B. Schölkopf. Counterfactuals uncover the modular structure of deep generative models. In *Proceedings of the International Conference on Learning Representations (ICLR 2020)*, 2020.

[12] I. Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2):115, 1987.

[13] P. Billingsley. *Probability and Measure*. John Wiley & Sons, New Jersey, 2012. ISBN 978-1-118-12237-2.

[14] C. M. Bishop. *Pattern recognition and machine learning*. springer, 2006.

[15] W. Brendel and M. Bethge. Approximating CNNs with bag-of-local-features models works surprisingly well on ImageNet. In *Proceedings of the International Conference on Learning Representations (ICLR 2019)*, 2019.

[16] P. Bühlmann. Invariance, causality and robustness. *arXiv preprint arXiv:1812.08233*, 2018.

[17] P. Bühlmann, J. Peters, J. Ernest, et al. CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 2014.

[18] R. Cai, Z. Li, P. Wei, J. Qiao, K. Zhang, and Z. Hao. Learning disentangled semantic representation for domain adaptation. In *Proceedings of the Conference of IJCAI*, volume 2019, page 2060. NIH Public Access, 2019.

[19] D. C. Castro, I. Walker, and B. Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):1–10, 2020.

[20] J. Chen and K. Batmanghelich. Weakly supervised disentanglement by pairwise similarities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3495–3502, 2020.

[21] R. T. Chen, X. Li, R. B. Grosse, and D. K. Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, pages 2610–2620, 2018.

[22] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2172–2180, 2016.

[23] C.-Y. Chuang, A. Torralba, and S. Jegelka. Estimating generalization under distribution shifts via domain-invariant representations. In *International Conference on Machine Learning*, pages 1984–1994. PMLR, 2020.

[24] T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2006.

[25] S. Cui, S. Wang, J. Zhuo, L. Li, Q. Huang, and Q. Tian. Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3941–3950, 2020.

[26] B. Dai and D. Wipf. Diagnosing and enhancing VAE models. In *International Conference on Learning Representations*, 2019.

[27] A. D'Amour, K. Heller, D. Moldovan, B. Adlam, B. Alipanahi, A. Beutel, C. Chen, J. Deaton, J. Eisenstein, M. D. Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*, 2020.

[28] C. Durkan and Y. Song. On maximum likelihood training of score-based generative models. *arXiv preprint arXiv:2101.09258*, 2021.

[29] D. M. Endres and J. E. Schindelin. A new metric for probability distributions. *IEEE Transactions on Information theory*, 49(7):1858–1860, 2003.

[30] C. Fang, Y. Xu, and D. N. Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1657–1664, 2013.

[31] C. Fefferman, S. Mitter, and H. Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

[32] Y. Gal and Z. Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the International Conference on Machine Learning*, pages 1050–1059, 2016.

[33] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17:1–35, 2016.

[34] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *Proceedings of the International Conference on Learning Representations (ICLR 2019)*, 2019.

[35] M. Gong, K. Zhang, T. Liu, D. Tao, C. Glymour, and B. Schölkopf. Domain adaptation with conditional transferable components. In *International Conference on Machine Learning*, pages 2839–2848, 2016.

[36] M. Gong, K. Zhang, B. Huang, C. Glymour, D. Tao, and K. Batmanghelich. Causal generative domain adaptation networks. *arXiv preprint arXiv:1804.04333*, 2018.

[37] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, Montréal, Canada, 2014. NIPS Foundation.

[38] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *Proceedings of the International Conference on Learning Representations (ICLR 2015)*, 2015.

[39] I. Gulrajani and D. Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

[40] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[41] Y. He, Z. Shen, and P. Cui. Towards non-i.i.d. image classification: A dataset and baselines. *arXiv preprint arXiv:1906.02899*, 2019.

[42] C. Heinze-Deml and N. Meinshausen. Conditional variance penalties and domain shift robustness. *stat*, 1050:13, 2019.

[43] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. Beta-VAE: Learning basic visual concepts with a constrained variational framework. In *Proceedings of the International Conference on Learning Representations (ICLR 2017)*, 2017.

[44] I. Higgins, D. Amos, D. Pfau, S. Racaniere, L. Matthey, D. Rezende, and A. Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.

[45] P. O. Hoyer, S. Shimizu, A. J. Kerminen, and M. Palviainen. Estimation of causal effects using linear non-gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49(2):362–378, 2008.

[46] F. Huszár. How (not) to train your generative model: Scheduled sampling, likelihood, adversary? *arXiv preprint arXiv:1511.05101*, 2015.

[47] A. Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(Apr):695–709, 2005.

[48] M. Ilse, J. M. Tomczak, and P. Forré. Designing data augmentation for simulating interventions. *arXiv preprint arXiv:2005.01856*, 2020.

[49] M. Ilse, J. M. Tomczak, C. Louizos, and M. Welling. DIVA: Domain invariant variational autoencoders. In *Medical Imaging with Deep Learning*, pages 322–348. PMLR, 2020.

[50] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, pages 125–136, 2019.

[51] D. Janzing, J. Peters, J. M. Mooij, and B. Schölkopf. Identifying confounders using additive noise models. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI 2009)*, pages 249–257. AUAI Press, 2009.

[52] D. Janzing, E. Sgouritsa, O. Stegle, J. Peters, and B. Schölkopf. Detecting low-complexity unobserved causes. In *27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, pages 383–391. AUAI Press, 2011.

[53] J. Jiang, B. Fu, and M. Long. Transfer-learning-library. `https://github.com/thuml/Transfer-Learning-Library`, 2020.

[54] F. D. Johansson, D. Sontag, and R. Ranganath. Support and invertibility in domain-invariant representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 527–536, 2019.

[55] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

[56] N. R. Ke, O. Bilaniuk, A. Goyal, S. Bauer, H. Larochelle, C. Pal, and Y. Bengio. Learning neural causal models from unknown interventions. *arXiv preprint arXiv:1910.01075*, 2019.

[57] I. Khemakhem, D. P. Kingma, R. P. Monti, and A. Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In S. Chiappa and R. Calandra, editors, *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pages 2207–2217, 2020.

[58] I. Khemakhem, R. P. Monti, D. P. Kingma, and A. Hyvärinen. ICE-BeeM: Identifiable conditional energy-based deep models. *arXiv preprint arXiv:2002.11537*, 2020.

[59] N. Kilbertus, G. Parascandolo, and B. Schölkopf. Generalization in anti-causal learning. *arXiv preprint arXiv:1812.00524*, 2018.

[60] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[61] D. P. Kingma and P. Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, 2018.

[62] D. P. Kingma and M. Welling. Auto-encoding variational Bayes. In *Proceedings of the International Conference on Learning Representations (ICLR 2014)*, Banff, Canada, 2014. ICLR Committee.

[63] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, pages 3581–3589, 2014.

[64] M. Kocaoglu, S. Shakkottai, A. G. Dimakis, C. Caramanis, and S. Vishwanath. Entropic latent variable discovery. *arXiv preprint arXiv:1807.10399*, 2018.

[65] T. C. Koopmans and O. Reiersol. The identification of structural characteristics. *The Annals of Mathematical Statistics*, 21(2):165–181, 1950.

[66] D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, R. L. Priol, and A. Courville. Out-of-distribution generalization via risk extrapolation (REx). *arXiv preprint arXiv:2003.00688*, 2020.

[67] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

[68] C. M. Lee, C. Hart, J. G. Richens, and S. Johri. Leveraging directed causal discovery to detect latent common causes. *arXiv preprint arXiv:1910.10174*, 2019.

[69] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017.

[70] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4114–4124, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[71] F. Locatello, M. Tschannen, S. Bauer, G. Rätsch, B. Schölkopf, and O. Bachem. Disentangling factors of variation using few labels. *arXiv preprint arXiv:1905.01258*, 2019.

[72] F. Locatello, B. Poole, G. Rätsch, B. Schölkopf, O. Bachem, and M. Tschannen. Weakly-supervised disentanglement without compromises. In *International Conference on Machine Learning*, pages 6348–6359. PMLR, 2020.

[73] M. Long, Y. Cao, J. Wang, and M. Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105, 2015.

[74] M. Long, Z. Cao, J. Wang, and M. I. Jordan. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pages 1640–1650, 2018.

[75] D. Lopez-Paz, R. Nishihara, S. Chintala, B. Schölkopf, and L. Bottou. Discovering causal signals in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6979–6987, 2017.

[76] C. Louizos, U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, and M. Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pages 6446–6456, 2017.

[77] S. Magliacane, T. van Ommen, T. Claassen, S. Bongers, P. Versteeg, and J. M. Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. In *Advances in Neural Information Processing Systems*, pages 10846–10856, 2018.

[78] J. D. Mcauliffe and D. M. Blei. Supervised topic models. In *Advances in Neural Information Processing Systems*, pages 121–128, Vancouver, Canada, 2008. NIPS Foundation.

[79] J. Mitrovic, B. McWilliams, J. C. Walker, L. H. Buesing, and C. Blundell. Representation learning via invariant causal mechanisms. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=9p2ekP904Rs`.

[80] K. Muandet, D. Balduzzi, and B. Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pages 10–18, 2013.

[81] K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[82] R. M. Neal. *Bayesian learning for neural networks*. PhD thesis, University of Toronto, 1995.

[83] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2010.

[84] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32:8026–8037, 2019.

[85] J. Pearl. *Causality*. Cambridge university press, 2009.

[86] J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(1):2009–2053, 2014.

[87] J. Peters, P. Bühlmann, and N. Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

[88] J. Peters, D. Janzing, and B. Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

[89] F. Qiao, L. Zhao, and X. Peng. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12556–12565, 2020.

[90] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In Y. Bengio and Y. LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[91] M. T. Ribeiro, S. Singh, and C. Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144, 2016.

[92] T. Richardson, P. Spirtes, et al. Ancestral graph Markov models. *The Annals of Statistics*, 30 (4):962–1030, 2002.

[93] M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1):1309–1342, 2018.

[94] J.-W. Romeijn and J. Williamson. Intervention and identifiability in latent variable modelling. *Minds and machines*, 28(2):243–264, 2018.

[95] D. Rothenhäusler, N. Meinshausen, P. Bühlmann, and J. Peters. Anchor regression: heterogeneous data meets causality. *arXiv preprint arXiv:1801.06229*, 2018.

[96] B. Schölkopf. Causality for machine learning. *arXiv preprint arXiv:1911.10500*, 2019.

[97] B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. M. Mooij. On causal and anticausal learning. In *International Conference on Machine Learning (ICML 2012)*, pages 1255–1262. International Machine Learning Society, 2012.

[98] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[99] E. Sgouritsa, D. Janzing, J. Peters, and B. Schölkopf. Identifying finite mixtures of nonparametric product distributions and causal inference of confounders. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence (UAI 2013)*, pages 556–575. AUAI Press, 2013.

[100] U. Shalit, F. D. Johansson, and D. Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3076–3085. JMLR.org, 2017.

[101] S. Shankar, V. Piratla, S. Chakrabarti, S. Chaudhuri, P. Jyothi, and S. Sarawagi. Generalizing across domains via cross-gradient training. In *Proceedings of the International Conference on Learning Representations (ICLR 2018)*, 2018.

[102] Z. Shen, P. Cui, K. Kuang, B. Li, and P. Chen. Causally regularized learning with agnostic data selection bias. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 411–419. ACM, 2018.

[103] I. Shpitser, R. J. Evans, T. S. Richardson, and J. M. Robins. Introduction to nested Markov models. *Behaviormetrika*, 41(1):3–39, 2014.

[104] R. Shu, Y. Chen, A. Kumar, S. Ermon, and B. Poole. Weakly supervised disentanglement with guarantees. In *International Conference on Learning Representations*, 2020.

[105] P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman. *Causation, prediction, and search*. MIT press, 2000.

[106] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[107] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR 2014)*, 2014.

[108] T. Teshima, I. Sato, and M. Sugiyama. Few-shot domain adaptation by causal mechanism transfer. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 9458–9469, 2020.

[109] L. Theis, A. van den Oord, and M. Bethge. A note on the evaluation of generative models. In *International Conference on Learning Representations (ICLR 2016)*, pages 1–10, 2016.

[110] T. Tieleman and G. Hinton. Lecture 6.5-RMSprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.

[111] T. Verma and J. Pearl. *Equivalence and synthesis of causal models*. UCLA, Computer Science Department, 1991.

[112] M. J. Wainwright, M. I. Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

[113] J. Wang, C. Lan, C. Liu, Y. Ouyang, and T. Qin. Generalizing to unseen domains: A survey on domain generalization. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4627–4635. International Joint Conferences on Artificial Intelligence Organization, 2021. Survey Track.

[114] Y. Wang and D. M. Blei. The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528):1574–1596, 2019.

[115] K. Q. Weinberger and L. K. Saul. Unsupervised learning of image manifolds by semidefinite programming. *International Journal of Computer Vision*, 70(1):77–90, 2006.

[116] Y. Yacoby, W. Pan, and F. Doshi-Velez. Learning deep bayesian latent variable regression models that generalize: When non-identifiability is a problem. *arXiv preprint arXiv:1911.00569*, 2019.

[117] M. Yang, F. Liu, Z. Chen, X. Shen, J. Hao, and J. Wang. CausalVAE: Structured causal disentanglement in variational autoencoder. *arXiv preprint arXiv:2004.08697*, 2020.

[118] L. Yao, S. Li, Y. Li, M. Huai, J. Gao, and A. Zhang. Representation learning for treatment effect estimation from observational data. In *Advances in Neural Information Processing Systems*, pages 2633–2643, 2018.

[119] H. Ye, C. Xie, T. Cai, R. Li, Z. Li, and L. Wang. Towards a theoretical framework of out-of-distribution generalization. *arXiv preprint arXiv:2106.04496*, 2021.

[120] K. You, X. Wang, M. Long, and M. Jordan. Towards accurate model selection in deep unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 7124–7133, 2019.

[121] C. Zhang, K. Zhang, and Y. Li. A causal view on robustness of neural networks. In *Advances in Neural Information Processing Systems*, 2020.

[122] K. Zhang and A. Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI 2009)*, pages 647–655. AUAI Press, 2009.

[123] K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, pages 819–827, 2013.

[124] Y. Zhang, T. Liu, M. Long, and M. Jordan. Bridging theory and algorithm for domain adaptation. In *International Conference on Machine Learning*, pages 7404–7413, 2019.

[125] H. Zhao, R. T. Des Combes, K. Zhang, and G. Gordon. On learning invariant representations for domain adaptation. In *International Conference on Machine Learning*, pages 7523–7532, 2019.

Appendix

A Proofs

We first introduce some handy concepts and results to make the proof succinct, meanwhile providing more information for understanding our model and theory. We begin with some extended discussions on CSG.

Definition 8. A homeomorphism Φ on $\mathcal{S}\times\mathcal{V}$ is called a *reparameterization* from CSG p to CSG p', if $\Phi_{\#}[p_{s,v}] = p'_{s,v}$, and $p(x|s,v) = p'(x|\Phi(s,v))$ and $p(y|s) = p'(y|\Phi^{\mathcal{S}}(s,v))$ for any $(s,v)\in\mathcal{S}\times\mathcal{V}$. A reparameterization Φ is called to be *semantic-preserving*, if its output dimensions in $\mathcal{S}$ is constant of v: $\Phi^{\mathcal{S}}(s,v) = \Phi^{\mathcal{S}}(s,v')$ for any $v, v'\in\mathcal{V}$ (hence denote $\Phi^{\mathcal{S}}(s,v)$ as $\Phi^{\mathcal{S}}(s)$ in this case).

Note that a reparameterization unnecessarily has its output dimensions in $\mathcal{S}$, *i.e.* $\Phi^{\mathcal{S}}(s,v)$, constant of v. The condition that $p(y|s) = p'(y|\Phi^{\mathcal{S}}(s,v))$ for any $v\in\mathcal{V}$ does not indicate that $\Phi^{\mathcal{S}}(s,v)$ is constant of v, since $p'(y|s')$ may ignore the change of $s' = \Phi^{\mathcal{S}}(s,v)$ from the change of v. The following lemma shows the meaning of a reparameterization: it allows a CSG to vary while inducing the same distribution on the observed data variables (x,y) (*i.e.*, holding the same effect on describing data).

Lemma 9. *If there exists a reparameterization Φ from CSG p to CSG p', then $p(x,y) = p'(x,y)$.*

Proof. By the definition of a reparameterization, we have:

$$\begin{aligned} p(x,y) &= \int p(s,v)p(x|s,v)p(y|s)\,\mathrm{d}s\mathrm{d}v = \int \Phi^{-1}_{\#}[p'_{s,v}](s,v)p'(x|\Phi(s,v))p'(y|\Phi^{\mathcal{S}}(s,v))\,\mathrm{d}s\mathrm{d}v \\ &= \int p'_{s,v}(s',v')p'(x|s',v')p'(y|s')\,\mathrm{d}s'\mathrm{d}v' = p'(x,y), \end{aligned}$$

where we used variable substitution $(s',v') := \Phi(s,v)$ in the second-last equality. Note that by the definition of pushed-forward distribution and the bijectivity of Φ, $\Phi_{\#}[p_{s,v}] = p'_{s,v}$ implies $p_{s,v} = \Phi^{-1}_{\#}[p'_{s,v}]$, and $\int f(s',v')p'_{s,v}(s',v')\,\mathrm{d}s'\mathrm{d}v' = \int f(\Phi(s,v))\Phi^{-1}_{\#}[p'_{s,v}](s,v)\,\mathrm{d}s\mathrm{d}v$ (can also be verified deductively using the rule of change of variables, *i.e.* Lemma 12 in the following). □

We can now define and verify an equivalent relation on CSGs so that the resulting equivalent class contains CSGs that induce the same (x,y) data distribution and hold the same semantic information in their s variables.

Definition 10 (semantic-equivalence)**.** We say two CSGs p and p' are *semantic-equivalent*, if there exists a homeomorphism[11] Φ on $\mathcal{S}\times\mathcal{V}$, such that **(i)** is *semantic-preserving*: its output dimensions in $\mathcal{S}$ is constant of v, $\Phi^{\mathcal{S}}(s,v) = \Phi^{\mathcal{S}}(s)$ for any $v\in\mathcal{V}$, and **(ii)** it acts as a *reparameterization* from p to p': $\Phi_{\#}[p_{s,v}] = p'_{s,v}$, $p(x|s,v) = p'(x|\Phi(s,v))$ and $p(y|s) = p'(y|\Phi^{\mathcal{S}}(s))$.

Proposition 14 in Appx. A.1 below shows that the defined binary relation is indeed an equivalence relation in common cases. As a reparameterization, Φ allows the two models to have different latent-variable parameterizations while inducing the same distribution on the observed data variables (x,y) (Lemma 9). The definition of semantic-identification (Def. 4) is then the semantic-equivalence of the ground-truth CSG p^* to the learned CSG p, which is also the semantic-equivalence of the learned CSG p to the ground-truth CSG p^* in common cases where it is an equivalence relation (Prop. 14).

This definition of semantic-equivalence can be rephrased as the *existence* of a semantic-preserving reparameterization. With proper model assumptions, we can show that *any* reparameterization between two CSGs is semantic-preserving, so that semantic-preserving CSGs cannot be converted to each other by a reparameterization that mixes s with v.

Lemma 11. *For two CSGs p and p', if $p'(y|s)$ has a statistics $M'(s)$ that is an injective function of s, then* any *reparameterization Φ from p to p', if exists, has its $\Phi^{\mathcal{S}}$ constant of v.*

Proof. Let $\Phi = (\Phi^{\mathcal{S}}, \Phi^{\mathcal{V}})$ be any reparameterization from p to p'. Then the condition that $p(y|s) = p'(y|\Phi^{\mathcal{S}}(s,v))$ for any $v\in\mathcal{V}$ indicates that $M(s) = M'(\Phi^{\mathcal{S}}(s,v))$. If there exist $s\in\mathcal{S}$ and $v^{(1)}\neq v^{(2)}\in\mathcal{V}$ such that $\Phi^{\mathcal{S}}(s,v^{(1)})\neq\Phi^{\mathcal{S}}(s,v^{(2)})$, then $M'(\Phi^{\mathcal{S}}(s,v^{(1)}))\neq M'(\Phi^{\mathcal{S}}(s,v^{(2)}))$

[11] A transformation is a homeomorphism if it is a continuous bijection with continuous inverse.

since M' is injective. This violates $M(s) = M'(\Phi^{\mathcal{S}}(s, v))$ which requires both $M'(\Phi^{\mathcal{S}}(s, v^{(1)}))$ and $M'(\Phi^{\mathcal{S}}(s, v^{(2)}))$ to be equal to $M(s)$. So $\Phi^{\mathcal{S}}(s, v)$ must be constant of v. □

We then introduce two mathematical facts.

Lemma 12 (rule of change of variables)**.** *Let z be a random variable on a Euclidean space $\mathbb{R}^{d_z}$ with density function $p_z(z)$, and let Φ be a homeomorphism on $\mathbb{R}^{d_z}$ whose inverse Φ^{-1} is differentiable. Then the distribution of the transformed random variable $z' = \Phi(z)$ has a density function $\Phi_{\#}[p_z](z') = p_z(\Phi^{-1}(z'))|J_{\Phi^{-1}}(z')|$, where $|J_{\Phi^{-1}}(z')|$ denotes the absolute value of the determinant of the Jacobian matrix $(J_{\Phi^{-1}}(z'))_{ia} := \frac{\partial}{\partial z'_i}(\Phi^{-1})_a(z')$ of Φ^{-1} at z'.*

Proof. See *e.g.*, Billingsley [13, Thm. 17.2]. Note that a homeomorphism is (Borel) measurable since it is continuous [13, Thm. 13.2], so the definition of $\Phi_{\#}[p_z]$ is valid. □

Lemma 13. *Let μ be a random variable whose characteristic function is a.e. non-zero. For two functions f and f' on the same space, we have: $f * p_\mu = f' * p_\mu \iff f = f'$ a.e., where $(f * p_\mu)(x) := \int f(x)p_\mu(x - \mu)\,\mathrm{d}\mu$ denotes convolution.*

Proof. The function equality $f * p_\mu = f' * p_\mu$ leads to the equality under Fourier transformation $\mathscr{F}[f * p_\mu] = \mathscr{F}[f' * p_\mu]$, which gives $\mathscr{F}[f]\mathscr{F}[p_\mu] = \mathscr{F}[f']\mathscr{F}[p_\mu]$. Since $\mathscr{F}[p_\mu]$ is the characteristic function of p_μ, the condition that it is a.e. non-zero indicates that $\mathscr{F}[f] = \mathscr{F}[f']$ a.e. thus $f = f'$ a.e. See also Khemakhem et al. [57, Thm. 1]. □

A.1 Proof of the Equivalence Relation

Proposition 14. *The semantic-equivalence in Def. 10 is an equivalence relation if $\mathcal{V}$ is connected and is either open or closed in $\mathbb{R}^{d_{\mathcal{V}}}$.*

Proof. Let Φ be a semantic-preserving reparameterization from one CSG $p = \langle p(s, v), p(x|s, v), p(y|s)\rangle$ to another $p' = \langle p'(s, v), p'(x|s, v), p'(y|s)\rangle$. It has its $\Phi^{\mathcal{S}}$ constant of v, so we can write $\Phi(s, v) = (\Phi^{\mathcal{S}}(s), \Phi^{\mathcal{V}}(s, v)) =: (\phi(s), \psi_s(v))$.

(1) We first show that ϕ, and ψ_s for any $s \in \mathcal{S}$, are homeomorphisms on $\mathcal{S}$ and $\mathcal{V}$, respectively, and that $\Phi^{-1}(s', v') = (\phi^{-1}(s'), \psi^{-1}_{\phi^{-1}(s')}(v'))$.

- Since $\Phi(\mathcal{S} \times \mathcal{V}) = \mathcal{S} \times \mathcal{V}$, so $\phi(\mathcal{S}) = \Phi^{\mathcal{S}}(\mathcal{S}) = \mathcal{S}$, so ϕ is surjective.
- Suppose that there exists $s' \in \mathcal{S}$ such that $\phi^{-1}(s') = \{s^{(i)}\}_{i\in\mathcal{I}}$ contains multiple distinct elements.
  1. Since Φ is surjective, for any $v' \in \mathcal{V}$, there exist $i \in \mathcal{I}$ and $v \in \mathcal{V}$ such that $(s', v') = \Phi(s^{(i)}, v) = (\phi(s^{(i)}), \psi_{s^{(i)}}(v))$, which means that $\bigcup_{i\in\mathcal{I}} \psi_{s^{(i)}}(\mathcal{V}) = \mathcal{V}$.
  2. Since Φ is injective, the sets $\{\psi_{s^{(i)}}(\mathcal{V})\}_{i\in\mathcal{I}}$ must be mutually disjoint. Otherwise, there would exist $i \neq j \in \mathcal{I}$ and $v^{(1)}, v^{(2)} \in \mathcal{V}$ such that $\psi_{s^{(i)}}(v^{(1)}) = \psi_{s^{(j)}}(v^{(2)})$ thus $\Phi(s^{(i)}, v^{(1)}) = (s', \psi_{s^{(i)}}(v^{(1)})) = (s', \psi_{s^{(j)}}(v^{(2)})) = \Phi(s^{(j)}, v^{(2)})$, which violates the injectivity of Φ since $s^{(i)} \neq s^{(j)}$.
  3. In the case where $\mathcal{V}$ is open, then so is any $\psi_{s^{(i)}}(\mathcal{V}) = \Phi(s^{(i)}, \mathcal{V})$ since Φ is continuous. But the union of disjoint open sets $\bigcup_{i\in\mathcal{I}} \psi_{s^{(i)}}(\mathcal{V}) = \mathcal{V}$ cannot be connected. This violates the condition that $\mathcal{V}$ is connected.
  4. A similar argument holds in the case where $\mathcal{V}$ is closed.

  So $\phi^{-1}(s')$ contains only one unique element for any $s' \in \mathcal{S}$. So ϕ is injective.
- The above argument also shows that for any $s' \in \mathcal{S}$, we have $\bigcup_{i\in\mathcal{I}} \psi_{s^{(i)}}(\mathcal{V}) = \psi_{\phi^{-1}(s')}(\mathcal{V}) = \mathcal{V}$. For any $s \in \mathcal{S}$, there exists $s' \in \mathcal{S}$ such that $s = \phi^{-1}(s')$, so we have $\psi_s(\mathcal{V}) = \mathcal{V}$. So ψ_s is surjective for any $s \in \mathcal{S}$.
- Suppose that there exist $v^{(1)} \neq v^{(2)} \in \mathcal{V}$ such that $\psi_s(v^{(1)}) = \psi_s(v^{(2)})$. Then $\Phi(s, v^{(1)}) = (\phi(s), \psi_s(v^{(1)})) = (\phi(s), \psi_s(v^{(2)})) = \Phi(s, v^{(2)})$, which contradicts the injectivity of Φ since $v^{(1)} \neq v^{(2)}$. So ψ_s is injective for any $s \in \mathcal{S}$.
- That Φ is continuous and $\Phi(s, v) = (\phi(s), \psi_s(v))$ indicates that ϕ and ψ_s are continuous. For any $(s', v') \in \mathcal{S} \times \mathcal{V}$, we have $\Phi(\phi^{-1}(s'), \psi^{-1}_{\phi^{-1}(s')}(v')) = (\phi(\phi^{-1}(s')), \psi_{\phi^{-1}(s')}(\psi^{-1}_{\phi^{-1}(s')}(v'))) = (s', v')$. Applying Φ^{-1} to both sides gives $\Phi^{-1}(s', v') = (\phi^{-1}(s'), \psi^{-1}_{\phi^{-1}(s')}(v'))$.
- Since Φ^{-1} is continuous, ϕ^{-1} and ψ_s^{-1} are also continuous.

(2) We now show that the relation is an equivalence relation. It amounts to showing the following three properties.

- Reflexivity. For two identical CSGs, we have $p(s,v) = p'(s,v)$, $p(x|s,v) = p'(x|s,v)$ and $p(y|s) = p'(y|s)$. So the identity map as Φ obviously satisfies all the requirements.
- Symmetry. Let Φ be a semantic-preserving reparameterization from $p = \langle p(s,v), p(x|s,v), p(y|s)\rangle$ to $p' = \langle p'(s,v), p'(x|s,v), p'(y|s)\rangle$. From the above conclusion in **(1)**, we know that $(\Phi^{-1})^{\mathcal{S}}(s',v') = \phi^{-1}(s')$ is semantic-preserving. Also, Φ^{-1} is a homeomorphism on $\mathcal{S}\times\mathcal{V}$ since Φ is. So we only need to show that Φ^{-1} is a reparameterization from p' to p for symmetry.
  1. From the definition of pushed-forward distribution, we have $\Phi^{-1}_{\#}[p'_{s,v}] = p_{s,v}$ if $\Phi_{\#}[p_{s,v}] = p'_{s,v}$. It can also be verified through the rule of change of variables (Lemma 12) when Φ and Φ^{-1} are differentiable. From $\Phi_{\#}[p_{s,v}] = p'_{s,v}$, we have for any (s',v'), $p_{s,v}(\Phi^{-1}(s',v'))|J_{\Phi^{-1}}(s',v')| = p'_{s,v}(s',v')$. Since for any (s,v) there exists (s',v') such that $(s,v) = \Phi^{-1}(s',v')$, this implies that for any (s,v), $p_{s,v}(s,v)|J_{\Phi^{-1}}(\Phi(s,v))| = p'_{s,v}(\Phi(s,v))$, or $p_{s,v}(s,v) = p'_{s,v}(\Phi(s,v))/|J_{\Phi^{-1}}(\Phi(s,v))| = p'_{s,v}(\Phi(s,v))|J_{\Phi}(s,v)|$ (inverse function theorem), which means that $p_{s,v} = \Phi^{-1}_{\#}[p'_{s,v}]$ by the rule of change of variables.
  2. For any (s',v'), there exists (s,v) such that $(s',v') = \Phi(s,v)$, so $p'(x|s',v') = p'(x|\Phi(s,v)) = p(x|s,v) = p(x|\Phi^{-1}(s',v'))$, and $p'(y|s') = p'(y|\Phi^{\mathcal{S}}(s)) = p(y|s) = p(y|(\Phi^{-1})^{\mathcal{S}}(s'))$.

  So Φ^{-1} is a reparameterization from p' to p.
- Transitivity. Given a third CSG $p'' = \langle p''(s,v), p''(x|s,v), p''(y|s)\rangle$ that is semantic-equivalent to p', there exists a semantic-preserving reparameterization Φ' from p' to p''. It is easy to see that $(\Phi'\circ\Phi)^{\mathcal{S}}(s,v) = \Phi'^{\mathcal{S}}(\Phi^{\mathcal{S}}(s,v)) = \Phi'^{\mathcal{S}}(\Phi^{\mathcal{S}}(s))$ is constant of v thus semantic-preserving. As the composition of two homeomorphisms Φ and Φ' on $\mathcal{S}\times\mathcal{V}$, $\Phi'\circ\Phi$ is also a homeomorphism. So we only need to show that $\Phi'\circ\Phi$ is a reparameterization from p to p'' for transitivity.
  1. From the definition of pushed-forward distribution, we have $(\Phi'\circ\Phi)_{\#}[p_{s,v}] = \Phi'_{\#}[\Phi_{\#}[p_{s,v}]] = \Phi'_{\#}[p'_{s,v}] = p''_{s,v}$ if $\Phi_{\#}[p_{s,v}] = p'_{s,v}$ and $\Phi'_{\#}[p'_{s,v}] = p''_{s,v}$. It can also be verified through the rule of change of variables (Lemma 12) when Φ^{-1} and Φ'^{-1} are differentiable. For any (s'',v''), we have
$$\begin{aligned}
&(\Phi'\circ\Phi)_{\#}[p_{s,v}](s'',v'') = p_{s,v}((\Phi'\circ\Phi)^{-1}(s'',v''))\big|J_{(\Phi'\circ\Phi)^{-1}}(s'',v'')\big| \\
&= p_{s,v}(\Phi^{-1}(\Phi'^{-1}(s'',v')))\big|J_{\Phi^{-1}}(\Phi'^{-1}(s'',v''))\big||J_{\Phi'^{-1}}(s'',v'')| \\
&= \Phi_{\#}[p_{s,v}](\Phi'^{-1}(s'',v''))|J_{\Phi'^{-1}}(s'',v'')| \\
&= p'_{s,v}(\Phi'^{-1}(s'',v''))|J_{\Phi'^{-1}}(s'',v'')| = \Phi'_{\#}[p'_{s,v}](s'',v'') = p''_{s,v}(s'',v'').
\end{aligned}$$
  2. For any (s,v), we have:
$$\begin{aligned}
&p(x|s,v) = p'(x|\Phi(s,v)) = p''(x|\Phi'(\Phi(s,v))) = p''(x|(\Phi'\circ\Phi)(s,v)), \\
&p(y|s) = p'(y|\Phi^{\mathcal{S}}(s)) = p''(y|\Phi'^{\mathcal{S}}(\Phi^{\mathcal{S}}(s))) = p''(y|(\Phi'\circ\Phi)^{\mathcal{S}}(s)).
\end{aligned}$$

  So $\Phi'\circ\Phi$ is a reparameterization from p to p''.

This completes the proof for an equivalence relation. □

A.2 Proof of the Semantic-Identifiability Thm. 5

We present a more general and detailed version of Thm. 5 and prove it. The conclusions in the theorem in the main context corresponds to conclusions **(ii)** and **(i)** below by taking the two CSGs p' and p as the well-learned CSG p and the ground-truth CSG p^*, respectively.

Theorem 5' (semantic-identifiability)**.** *Consider two CSGs p and p' that have Assumption 3 hold, with the bounded derivative conditions specified to be that for both CSGs, f^{-1} and g are twice and f thrice differentiable with mentioned derivatives bounded. Further assume that they have absolutely continuous priors whose log-densities $\log p(s,v)$ and $\log p'(s,v)$ are bounded up to the second-order. If the two CSGs induce the same distribution on data,* i.e. *$p(x,y) = p'(x,y)$, then they*

*are semantic-equivalent, under **one of** the following three conditions:* [12]
(i) *p_μ has an a.e. non-zero characteristic function (*e.g.*, a Gaussian distribution);*[13]
(ii) *$\frac{1}{\sigma_\mu^2} \to \infty$, where $\sigma_\mu^2 := \mathbb{E}[\mu^\top \mu]$;*
(iii) *$\frac{1}{\sigma_\mu^2} \gg B'^2_{f^{-1}} \max\{B'_{\log p} B'_g + \frac{1}{2} B''_g + \frac{3}{2} d B'_{f^{-1}} B''_f B'_g, B_p B'^d_{f^{-1}} (B'^2_{\log p} + B''_{\log p} + 3 d B'_{f^{-1}} B''_f B'_{\log p} + 3 d^{\frac{3}{2}} B'^2_{f^{-1}} B''^2_f + d^3 B'''_f B'^3_{f^{-1}})\}$, where $d := d_\mathcal{S} + d_\mathcal{V}$, and for both CSGs, the constant B_p bounds $p(s, v)$, $B'_{f^{-1}}, B'_g, B'_{\log p}$ and $B''_f, B''_g, B''_{\log p}$ bound the 2-norms[14] of the gradient/Jacobian and the Hessians of the respective functions, and B'''_f bounds all the 3rd-order derivatives of f.*

Proof. Without loss of generality, we assume that μ and ν (for continuous y) have zero mean. If it is not, we can redefine $f(s, v) := f(s, v) + \mathbb{E}[\mu]$ and $\mu := \mu - \mathbb{E}[\mu]$ (similarly for ν for continuous y) which does not alter the joint distribution $p(s, v, x, y)$ nor violates any assumptions. Also without loss of generality, we consider one scalar component (dimension) l of y, and abuse the use of symbols y and g for y_l and g_l to avoid unnecessary complication. Note that for continuous y, due to the additive noise structure $y = g(s) + \nu$ and that ν has zero mean, we also have $\mathbb{E}[y|s] = g(s)$ as the same as the categorical y case (under the one-hot representation). We sometimes denote $z := (s, v)$ for convenience.

First note that for both CSGs and both continuous and categorical y, by construction $g(s)$ is a sufficient statistics of $p(y|s)$ (not only the expectation $\mathbb{E}[y|s]$), and it is injective. So by Lemma 11, we only need to show that there exists a reparameterization from p to p'. We will show that $\Phi := f'^{-1} \circ f$ is such a reparameterization.

Since f and f' are bijective and continuous, we have $\Phi^{-1} = f^{-1} \circ f'$, so Φ is bijective and Φ and Φ^{-1} are continuous. So Φ is a homeomorphism. Also, by construction, we have:

$$p(x|z) = p_\mu(x - f(z)) = p_\mu(x - f'(f'^{-1}(f(z)))) = p_\mu(x - f'(\Phi(z))) = p'(x|\Phi(z)). \tag{7}$$

So we only need to show that $p(x, y) = p'(x, y)$ indicates $\Phi_\#[p_z] = p'_z$ and $p(y|s) = p'(y|\Phi^\mathcal{S}(s, v)), \forall v \in \mathcal{V}$ under the conditions.

Proof under condition (i). We begin with a useful reformulation of the integral $\int t(z) p(x|z) \, \mathrm{d}z$ for a general function t of z. We will encounter integrals in this form. By the additive noise Assumption 3, we have $p(x|z) = p_\mu(x - f(z))$, so we consider a transformation $\Psi_x(z) := x - f(z)$ and let $\mu = \Psi_x(z)$. It is invertible, $\Psi_x^{-1}(\mu) = f^{-1}(x - \mu)$, and $J_{\Psi_x^{-1}}(\mu) = -J_{f^{-1}}(x - \mu)$. By these definitions and the rule of change of variables, we have:

$$\begin{aligned}
\int t(z) p(x|z) \, \mathrm{d}z &= \int t(z) p_\mu(\Psi_x(z)) \, \mathrm{d}z = \int t(\Psi_x^{-1}(\mu)) p(\mu) \left| J_{\Psi_x^{-1}}(\mu) \right| \mathrm{d}\mu \\
&= \int t(f^{-1}(x - \mu)) p(\mu) \left| J_{f^{-1}}(x - \mu) \right| \mathrm{d}\mu \\
&= \mathbb{E}_{p(\mu)}[(\bar{t}V)(x - \mu)] \qquad (8) \\
&= (f_\#[t] * p_\mu)(x), \qquad (9)
\end{aligned}$$

where we have denoted functions $\bar{t} := t \circ f^{-1}$, $V := \left| J_{f^{-1}} \right|$, and abused the push-forward notation $f_\#[t]$ for a general function t to formally denote $(t \circ f^{-1}) \left| J_{f^{-1}} \right| = \bar{t} V$.

According to the graphical structure of CSG, we have:

$$\begin{aligned}
p(x) &= \int p(z) p(x|z) \, \mathrm{d}z, \qquad (10) \\
\mathbb{E}[y|x] &= \frac{1}{p(x)} \int y p(x, y) \, \mathrm{d}y = \frac{1}{p(x)} \iint y p(z) p(x|z) p(y|s) \, \mathrm{d}z \mathrm{d}y
\end{aligned}$$

[12] To be precise, the conclusions are that the equalities in Def. 10 hold a.e. for condition **(i)**, hold asymptotically in the limit $\frac{1}{\sigma_\mu^2} \to \infty$ for condition **(ii)**, and hold up to a negligible quantity for condition **(iii)**.

[13] This also requires that p and p' have the same p_μ, or that the ground-truth p_μ is known in learning. However, p_μ is easier to model/specify/learn than f, and f dominates $p(x|s, v)$ over p_μ when the causal mechanism tends to be strong. So learning or specifying p_μ in learning is not a significant violation of this requirement.

[14] As an induced operator norm for matrices (not the Frobenius norm).

$$= \frac{1}{p(x)} \int p(z)p(x|z)\mathbb{E}[y|s]\,\mathrm{d}z = \frac{1}{p(x)} \int g(s)p(z)p(x|z)\,\mathrm{d}z. \tag{11}$$

So from Eq. (9), we have:

$$p(x) = (f_{\#}[p_z] * p_\mu)(x), \qquad \mathbb{E}[y|x] = \frac{1}{p(x)}(f_{\#}[gp_z] * p_\mu)(x). \tag{12}$$

Matching the data distribution $p(x, y) = p'(x, y)$ indicates both $p(x) = p'(x)$ and $\mathbb{E}[y|x] = \mathbb{E}'[y|x]$. Using Lemma 13 under condition **(i)**, this further indicates:

$$f_{\#}[p_z] = f'_{\#}[p'_z] \text{ a.e.}, \qquad f_{\#}[gp_z] = f'_{\#}[g'p'_z] \text{ a.e.},$$

given that p and p' have the same p_μ. The former indicates $\Phi_{\#}[p_z] = p'_z$. The latter can be reformed as $\bar{g}f_{\#}[p_z] = \bar{g}'f'_{\#}[p'_z]$ a.e., so $\bar{g} = \bar{g}'$ a.e., where we have denoted $\bar{g} := g \circ (f^{-1})^{\mathcal{S}}$ and $\bar{g}' := g' \circ (f'^{-1})^{\mathcal{S}}$ similarly. From $\bar{g} = \bar{g}'$, we have for any $v \in \mathcal{V}$,

$$\begin{aligned} g(s) &= g((f^{-1} \circ f)^{\mathcal{S}}(s, v)) = g((f^{-1})^{\mathcal{S}}(f(s, v))) = \bar{g}(f(s, v)) \\ &= \bar{g}'(f(s, v)) = g'((f'^{-1})^{\mathcal{S}}(f(s, v))) = g'(\Phi^{\mathcal{S}}(s, v)). \end{aligned} \tag{13}$$

For both continuous and categorical y, $g(s)$ uniquely determines $p(y|s)$. So the above equality means that $p(y|s) = p'(y|\Phi^{\mathcal{S}}(s, v))$ for any $v \in \mathcal{V}$.

Proof under condition (ii). Applying Eq. (8) to Eqs. (10, 11) (or expanding Eq. (12)), we have:

$$p(x) = \mathbb{E}_{p(\mu)}[(\bar{p}_z V)(x - \mu)], \qquad \mathbb{E}[y|x] = \frac{1}{p(x)}\mathbb{E}_{p(\mu)}[(\bar{g}\bar{p}_z V)(x - \mu)],$$

where we have similarly denoted $\bar{p}_z := p_z \circ f^{-1}$. Under condition **(ii)**, $\mathbb{E}[\mu^\top \mu]$ is infinitesimal, so we can expand the expressions w.r.t μ. For $p(x)$, we have:

$$\begin{aligned} p(x) &= \mathbb{E}_{p(\mu)}\big[\bar{p}_z V - \nabla(\bar{p}_z V)^\top \mu + \frac{1}{2}\mu^\top \nabla\nabla^\top (\bar{p}_z V)\mu + O(\mathbb{E}[\|\mu\|_2^3])\big] \\ &= \bar{p}_z V + \frac{1}{2}\mathbb{E}_{p(\mu)}\big[\mu^\top \nabla\nabla^\top (\bar{p}_z V)\mu\big] + O(\sigma_\mu^3), \end{aligned}$$

where all functions are evaluated at x. For $\mathbb{E}[y|x]$, we first expand $1/p(x)$ using $\frac{1}{x+\varepsilon} = \frac{1}{x} - \frac{\varepsilon}{x^2} + O(\varepsilon^2)$ to get: $\frac{1}{p(x)} = \frac{1}{\bar{p}_z V} - \frac{1}{2\bar{p}_z^2 V^2}\mathbb{E}_{p(\mu)}\big[\mu^\top \nabla\nabla^\top (\bar{p}_z V)\mu\big] + O(\sigma_\mu^3)$. The second term is expanded as: $\bar{g}\bar{p}_z V + \frac{1}{2}\mathbb{E}_{p(\mu)}\big[\mu^\top \nabla\nabla^\top (\bar{g}\bar{p}_z V)\mu\big] + O(\sigma_\mu^3)$. Combining the two parts, we have:

$$\mathbb{E}[y|x] = \bar{g} + \frac{1}{2}\mathbb{E}_{p(\mu)}\big[\mu^\top \big((\nabla \log \bar{p}_z V)\nabla\bar{g}^\top + \nabla\bar{g}(\nabla \log \bar{p}_z V)^\top + \nabla\nabla^\top \bar{g}\big)\mu\big] + O(\sigma_\mu^3). \tag{14}$$

This equation holds for any $x \in \mathrm{supp}(p_x)$ since the expectation is taken w.r.t the distribution $p(x, y)$. Since $p(x, y) = p'(x, y)$, the considered x here is any value generated by the model. So up to $O(\sigma_\mu^2)$,

$$\begin{aligned} |p(x) - (\bar{p}_z V)(x)| &= \frac{1}{2}\big|\mathbb{E}_{p(\mu)}\big[\mu^\top \nabla\nabla^\top (\bar{p}_z V)\mu\big]\big| \leqslant \frac{1}{2}\mathbb{E}_{p(\mu)}\big[\big|\mu^\top \nabla\nabla^\top (\bar{p}_z V)\mu\big|\big] \\ &\leqslant \frac{1}{2}\mathbb{E}_{p(\mu)}\big[\|\mu\|_2 \big\|\nabla\nabla^\top (\bar{p}_z V)\big\|_2 \|\mu\|_2\big] = \frac{1}{2}\mathbb{E}[\mu^\top \mu]\big\|\nabla\nabla^\top (\bar{p}_z V)\big\|_2 \\ &= \frac{1}{2}\mathbb{E}[\mu^\top \mu]|\bar{p}_z V|\big\|\nabla\nabla^\top \log \bar{p}_z V + (\nabla \log \bar{p}_z V)(\nabla \log \bar{p}_z V)^\top\big\|_2 \\ &\leqslant \frac{1}{2}\mathbb{E}[\mu^\top \mu]|\bar{p}_z V|\big(\big\|\nabla\nabla^\top \log \bar{p}_z V\big\|_2 + \|\nabla \log \bar{p}_z V\|_2^2\big), \end{aligned} \tag{15}$$

$$\begin{aligned} |\mathbb{E}[y|x] - \bar{g}(x)| &= \frac{1}{2}\Big|\mathbb{E}_{p(\mu)}\big[\mu^\top \big((\nabla \log \bar{p}_z V)\nabla\bar{g}^\top + \nabla\bar{g}(\nabla \log \bar{p}_z V)^\top + \nabla\nabla^\top \bar{g}\big)\mu\big]\Big| \\ &\leqslant \frac{1}{2}\mathbb{E}_{p(\mu)}\big[\big|\mu^\top \big((\nabla \log \bar{p}_z V)\nabla\bar{g}^\top + \nabla\bar{g}(\nabla \log \bar{p}_z V)^\top + \nabla\nabla^\top \bar{g}\big)\mu\big|\big] \\ &\leqslant \frac{1}{2}\mathbb{E}_{p(\mu)}\big[\|\mu\|_2 \big\|(\nabla \log \bar{p}_z V)\nabla\bar{g}^\top + \nabla\bar{g}(\nabla \log \bar{p}_z V)^\top + \nabla\nabla^\top \bar{g}\big\|_2 \|\mu\|_2\big] \\ &\leqslant \frac{1}{2}\mathbb{E}[\mu^\top \mu]\big(\big\|(\nabla \log \bar{p}_z V)\nabla\bar{g}^\top\big\|_2 + \big\|\nabla\bar{g}(\nabla \log \bar{p}_z V)^\top\big\|_2 + \big\|\nabla\nabla^\top \bar{g}\big\|_2\big) \\ &= \mathbb{E}[\mu^\top \mu]\Big(\big|(\nabla \log \bar{p}_z V)^\top \nabla\bar{g}\big| + \frac{1}{2}\big\|\nabla\nabla^\top \bar{g}\big\|_2\Big). \end{aligned} \tag{16}$$

Given the bounding conditions in the theorem, the multiplicative factors to $\mathbb{E}[\mu^\top \mu]$ in the last expressions are bounded by a constant. So when $\frac{1}{\sigma_\mu^2} \to \infty$, *i.e.* $\mathbb{E}[\mu^\top \mu] \to 0$, we have $p(x)$ and $\mathbb{E}[y|x]$ converge uniformly to $(\bar{p}_z V)(x) = f_\#[p_z](x)$ and $\bar{g}(x)$, respectively. So $p(x, y) = p'(x, y)$ indicates $f_\#[p_z] = f'_\#[p'_z]$ and $\bar{g} = \bar{g}'$, which means $\Phi_\#[p_z] = p'_z$ and $p(y|s) = p'(y|\Phi^{\mathcal{S}}(s, v))$ for any $v \in \mathcal{V}$, due to Eq. (13) and the explanation that follows.

Proof under condition (iii). We only need to show that when $\frac{1}{\sigma_\mu^2}$ is much larger than the given quantity, we still have $p(x, y) = p'(x, y) \Longrightarrow \bar{p}_z V = \bar{p}'_z V', \bar{g} = \bar{g}'$ up to a negligible effect. This task amounts to showing that the residuals $|p(x) - (\bar{p}_z V)(x)|$ and $|\mathbb{E}[y|x] - \bar{g}(x)|$ controlled by Eqs. (15, 16) are negligible. To achieve this, we need to further expand the controlling functions using derivatives of f, g and p_z explicitly, and bound them by the bounding constants. In the following, we use indices a, b, c for the components of x and i, j, k for those of z. For functions of z appearing in the following (*e.g.*, f, g, p_z and their derivatives), they are evaluated at $z = f^{-1}(x)$ since we are bounding functions of x.

(1) Bounding $|\mathbb{E}[y|x] - \bar{g}(x)| \leqslant \mathbb{E}[\mu^\top \mu]\big(\big|(\nabla \log \bar{p}_z V)^\top \nabla \bar{g}\big| + \frac{1}{2}\big\|\nabla\nabla^\top \bar{g}\big\|_2\big)$ from Eq. (16).

From the chain rule of differentiation, it is easy to show that:

$$\nabla \log \bar{p}_z = J_{f^{-1}} \nabla \log p_z, \qquad \nabla \bar{g} = J_{(f^{-1})^{\mathcal{S}}} \nabla g = J_{f^{-1}} \nabla_z g, \tag{17}$$

where $\nabla_z g = (\nabla g^\top, 0_{d_{\mathcal{V}}}^\top)^\top$ (recall that g is a function only of s). For the term $\nabla \log V$, we apply Jacobi's formula for the derivative of the log-determinant:

$$\begin{aligned}\partial_a \log V(x) &= \partial_a \log\big|J_{f^{-1}}(x)\big| = \operatorname{tr}\Big(J_{f^{-1}}^{-1}(x)\big(\partial_a J_{f^{-1}}(x)\big)\Big) = \sum_{b,i} J_{f^{-1}}^{-1}(x)_{ib}\big(\partial_a J_{f^{-1}}(x)_{bi}\big) \\ &= \sum_{b,i} J_f(f^{-1}(x))_{ib} \partial_b \partial_a f_i^{-1}(x) = \sum_i \big(J_f(\nabla\nabla^\top f_i^{-1})\big)_{ia}. \end{aligned} \tag{18}$$

However, as bounding Eq. (17) already requires bounding $\big\|J_{f^{-1}}\big\|_2$, directly using this expression to bound $\|\nabla \log V\|_2$ would require to also bound $\|J_f\|_2$. This requirement to bound the first-order derivatives of both f and f^{-1} is a relatively restrictive one. To ease the requirement, we would like to express $\nabla \log V$ in terms of $J_{f^{-1}}$. This can be achieved by expressing $\nabla\nabla^\top f_i^{-1}$'s in terms of $\nabla\nabla^\top f_c$'s. To do this, first consider a general invertible-matrix-valued function $A(\alpha)$ on a scalar α. We have $0 = \partial_\alpha\big(A(\alpha)^{-1}A(\alpha)\big) = (\partial_\alpha A^{-1})A + A^{-1}\partial_\alpha A$, so we have $A^{-1}\partial_\alpha A = -(\partial_\alpha A^{-1})A$, consequently $\partial_\alpha A = -A(\partial_\alpha A^{-1})A$. Using this relation (in the fourth equality below), we have:

$$\begin{aligned}&\big(\nabla\nabla^\top f_i^{-1}\big)_{ab} = \partial_a\partial_b f_i^{-1} = \partial_a\big(J_{f^{-1}}\big)_{bi} = \big(\partial_a J_{f^{-1}}\big)_{bi} \\ =\ & -\Big(J_{f^{-1}}(\partial_a J_{f^{-1}}^{-1})J_{f^{-1}}\Big)_{bi} = -\Big(J_{f^{-1}}\big(\partial_a J_f\big)J_{f^{-1}}\Big)_{bi} \\ =\ & -\sum_{jc}(J_{f^{-1}})_{bj}\big(\partial_a(\partial_j f_c)\big)(J_{f^{-1}})_{ci} = -\sum_{jck}(J_{f^{-1}})_{bj}(\partial_k\partial_j f_c)(\partial_a f_k^{-1})(J_{f^{-1}})_{ci} \\ =\ & -\sum_c (J_{f^{-1}})_{ci}\sum_{jk}(J_{f^{-1}})_{bj}(\partial_k\partial_j f_c)(J_{f^{-1}})_{ak} = -\sum_c (J_{f^{-1}})_{ci}\big(J_{f^{-1}}(\nabla\nabla^\top f_c)J_{f^{-1}}^\top\big)_{ab},\end{aligned}$$

or in matrix form,

$$\nabla\nabla^\top f_i^{-1} = -\sum_c (J_{f^{-1}})_{ci} J_{f^{-1}}(\nabla\nabla^\top f_c)J_{f^{-1}}^\top =: -\sum_c (J_{f^{-1}})_{ci} K^c, \tag{19}$$

where we have defined the matrix $K^c := J_{f^{-1}}(\nabla\nabla^\top f_c)J_{f^{-1}}^\top$ which is symmetric. Substituting with this result, we can transform Eq. (18) into a desired form:

$$\begin{aligned}\nabla \log V(x) &= \sum_i \big(J_f(\nabla\nabla^\top f_i^{-1})\big)_{i:}^\top = -\sum_i \Big(J_f \sum_c (J_{f^{-1}})_{ci} J_{f^{-1}}(\nabla\nabla^\top f_c)J_{f^{-1}}^\top\Big)_{i:}^\top \\ &= -\sum_i \Big(\sum_c (J_{f^{-1}})_{ci} J_f J_f^{-1}(\nabla\nabla^\top f_c)J_{f^{-1}}^\top\Big)_{i:}^\top = -\sum_{ci}(J_{f^{-1}})_{ci}\Big((\nabla\nabla^\top f_c)J_{f^{-1}}^\top\Big)_{i:}^\top \\ &= -\sum_c \Big(J_{f^{-1}}(\nabla\nabla^\top f_c)J_{f^{-1}}^\top\Big)_{c:}^\top = -\sum_c (K_{c:}^c)^\top = -\sum_c K_{:c}^c, \end{aligned} \tag{20}$$

so its norm can be bounded by:

$$\begin{aligned}\|\nabla \log V(x)\|_2 &= \Big\| \sum_c K^c_{c:} \Big\|_2 = \Big\| \sum_c (J_{f^{-1}})_{c:} (\nabla\nabla^\top f_c) J_{f^{-1}}^\top \Big\|_2 \\ &\leqslant \sum_c \big\|(J_{f^{-1}})_{c:}\big\|_2 \big\|\nabla\nabla^\top f_c\big\|_2 \big\|J_{f^{-1}}\big\|_2 \leqslant B''_f B'^2_{f^{-1}} \sum_c \big\|(J_{f^{-1}})_{c:}\big\|_2 \\ &\leqslant dB'^2_{f^{-1}} B''_f, \end{aligned} \tag{21}$$

where we have used the following result in the last inequality:

$$\sum_c \big\|(J_{f^{-1}})_{c:}\big\|_2 \leqslant d^{1/2} \sqrt{\sum_c \big\|(J_{f^{-1}})_{c:}\big\|_2^2} = d^{1/2} \big\|J_{f^{-1}}\big\|_F \leqslant d \big\|J_{f^{-1}}\big\|_2 \leqslant dB'_{f^{-1}}. \tag{22}$$

Integrating Eq. (17) and Eq. (21), we have:

$$\begin{aligned}\big|(\nabla \log \bar{p}_z V)^\top \nabla \bar{g}\big| &= (J_{f^{-1}} \nabla \log p_z + \nabla \log V)^\top J_{f^{-1}} \nabla_z g \\ &\leqslant \big(\big\|J_{f^{-1}}\big\|_2 \|\nabla \log p_z\|_2 + \|\nabla \log V\|_2\big) \big\|J_{f^{-1}}\big\| \|\nabla g\|_2 \\ &\leqslant \big(B'_{f^{-1}} B'_{\log p} + dB'^2_{f^{-1}} B''_f\big) B'_{f^{-1}} B'_g \\ &= \big(B'_{\log p} + dB'_{f^{-1}} B''_f\big) B'^2_{f^{-1}} B'_g. \end{aligned} \tag{23}$$

For the Hessian of $\bar{g}$, direct calculus gives:

$$\begin{aligned}\nabla\nabla^\top \bar{g} &= J_{(f^{-1})^{\mathcal{S}}} (\nabla\nabla^\top g) J_{(f^{-1})^{\mathcal{S}}}^\top + \sum_{i=1}^{d_{\mathcal{S}}} (\nabla g)_{s_i} (\nabla\nabla^\top f_{s_i}^{-1}) \\ &= J_{f^{-1}} (\nabla_z \nabla_z^\top g) J_{f^{-1}}^\top + \sum_i (\nabla_z g)_i (\nabla\nabla^\top f_i^{-1}). \end{aligned}$$

To avoid the requirement of bounding both $\nabla\nabla^\top f_c$'s and $\nabla\nabla^\top f_i^{-1}$'s, we substitute $\nabla\nabla^\top f_i^{-1}$ using Eq. (19):

$$\begin{aligned}\nabla\nabla^\top \bar{g} &= J_{f^{-1}} (\nabla_z \nabla_z^\top g) J_{f^{-1}}^\top - \sum_i (\nabla_z g)_i \sum_c (J_{f^{-1}})_{ci} K^c \\ &= J_{f^{-1}} (\nabla_z \nabla_z^\top g) J_{f^{-1}}^\top - \sum_c \Big((J_{f^{-1}})_{c,:} (\nabla_z g) \Big) K^c. \end{aligned}$$

So its norm can be bounded by:

$$\begin{aligned}\big\|\nabla\nabla^\top \bar{g}\big\|_2 &\leqslant \big\|J_{f^{-1}}\big\|_2^2 \big\|\nabla\nabla^\top g\big\|_2 + \sum_c \big|(J_{f^{-1}})_{c:} (\nabla_z g)\big| \|K^c\|_2 \\ &\leqslant B'^2_{f^{-1}} B''_g + \sum_c \big|(J_{f^{-1}})_{c:} (\nabla_z g)\big| B'^2_{f^{-1}} B''_f \\ &\leqslant B'^2_{f^{-1}} \Big(B''_g + B''_f \sum_c \big\|(J_{f^{-1}})_{c:}\big\|_2 \|\nabla_z g\|_2 \Big) \\ &\leqslant B'^2_{f^{-1}} \Big(B''_g + B''_f B'_g \sum_c \big\|(J_{f^{-1}})_{c:}\big\|_2 \Big) \\ &\leqslant B'^2_{f^{-1}} \Big(B''_g + dB'_{f^{-1}} B''_f B'_g \Big), \end{aligned} \tag{24}$$

where we have used Eq. (22) in the last inequality. Assembling Eq. (23) and Eq. (24) into Eq. (16), we have:

$$|\mathbb{E}[y|x] - \bar{g}(x)| \leqslant \mathbb{E}[\mu^\top \mu] B'^2_{f^{-1}} \big(B'_{\log p} B'_g + \frac{1}{2} B''_g + \frac{3}{2} dB'_{f^{-1}} B''_f B'_g \big). \tag{25}$$

So given the condition **(iii)**, this residual can be neglected.

(2) Bounding $|p(x) - (\bar{p}_z V)(x)| \leqslant \frac{1}{2}\mathbb{E}[\mu^\top \mu] |\bar{p}_z V| \big(\|\nabla \log \bar{p}_z V\|_2^2 + \big\|\nabla\nabla^\top \log \bar{p}_z\big\|_2 + \big\|\nabla\nabla^\top \log V\big\|_2 \big)$ from Eq. (15).

To begin with, for any x, $\bar{p}_z(x) = p_z(f^{-1}(x)) \leqslant B_p$, and $V(x) = \left|J_{f^{-1}}(x)\right|$ is the product of absolute eigenvalues of $J_{f^{-1}}(x)$. Since $\left\|J_{f^{-1}}(x)\right\|_2$ is the largest absolute eigenvalue of $J_{f^{-1}}(x)$, so $V(x) \leqslant \left\|J_{f^{-1}}(x)\right\|_2^d \leqslant B'^d_{f^{-1}}$.

For the first norm in the bracket of the r.h.s of Eq. (15), we have:

$$\begin{aligned}\|\nabla \log \bar{p}_z V\|_2^2 &= \|\nabla \log \bar{p}_z\|_2^2 + 2(\nabla \log \bar{p}_z)^\top \nabla \log V + \|\nabla \log V\|_2^2 \\ &\leqslant \|\nabla \log \bar{p}_z\|_2^2 + 2\|\nabla \log \bar{p}_z\|_2 \|\nabla \log V\|_2 + \|\nabla \log V\|_2 \\ &\leqslant B'^2_{f^{-1}} B'^2_{\log p} + 2dB'^3_{f^{-1}} B'_f B'_{\log p} + \|\nabla \log V\|_2^2, \end{aligned} \tag{26}$$

where we have utilized Eq. (17) and Eq. (21) in the last inequality. We consider bounding $\|\nabla \log V\|_2^2$ separately. Using Eq. (20) (in the second equality below), we have:

$$\begin{aligned}\|\nabla \log V\|_2^2 &= \left|(\nabla \log V)^\top (\nabla \log V)\right| = \Big|\sum_c (K^c_{:c})^\top \sum_d K^d_{:d}\Big| \\ &= \Big|\sum_{cd} K^c_{c:} K^d_{:d}\Big| \leqslant \sum_{cd} \left|K^c_{c:} K^d_{:d}\right| \\ &= \sum_{cd} \Big|(J_{f^{-1}})_{c:} (\nabla\nabla^\top f_c) J^\top_{f^{-1}} J_{f^{-1}} (\nabla\nabla^\top f_d) (J_{f^{-1}})^\top_{d:}\Big| \\ &\leqslant \sum_{cd} \left|(J_{f^{-1}})_{c:} (J_{f^{-1}})^\top_{d:}\right| \Big\|(\nabla\nabla^\top f_c) J^\top_{f^{-1}} J_{f^{-1}} (\nabla\nabla^\top f_d)\Big\|_2 \\ &\leqslant \sum_{cd} \left|(J_{f^{-1}})_{c:} (J_{f^{-1}})^\top_{d:}\right| B'^2_{f^{-1}} B''^2_f = B'^2_{f^{-1}} B''^2_f \sum_{cd} \Big|(J_{f^{-1}} J^\top_{f^{-1}})_{cd}\Big| \\ &\leqslant d^{3/2} B'^2_{f^{-1}} B''^2_f \Big\|J_{f^{-1}} J^\top_{f^{-1}}\Big\|_2 \leqslant d^{3/2} B'^4_{f^{-1}} B''^2_f, \end{aligned} \tag{27}$$

where we have used the facts for general matrix A and (column) vectors α, β that

$$\left|\alpha^\top A \beta\right| = \left\|\alpha (A\beta)^\top\right\|_2 = \left\|\alpha \beta^\top A^\top\right\|_2 \leqslant \left\|\alpha\beta^\top\right\|_2 \|A\|_2 = \left|\alpha^\top \beta\right| \|A\|_2 \tag{28}$$

in the fifth last inequality, and that

$$\sum_{cd} |A_{cd}| \leqslant \sqrt{d^2} \sqrt{\sum_{cd} |A_{cd}|^2} = d\|A\|_F \leqslant d^{3/2} \|A\|_2 \tag{29}$$

in the second last inequality. Substituting Eq. (27) into Eq. (26), we have:

$$\|\nabla \log \bar{p}_z V\|_2^2 \leqslant B'^2_{f^{-1}} B'^2_{\log p} + 2dB'^3_{f^{-1}} B'_f B'_{\log p} + d^{3/2} B'^4_{f^{-1}} B''^2_f. \tag{30}$$

For the second norm in the bracket of the r.h.s of Eq. (15), similar to Eq. (24), we have:

$$\left\|\nabla\nabla^\top \log \bar{p}_z\right\|_2 \leqslant B'^2_{f^{-1}} \big(B''_{\log p} + dB'_{f^{-1}} B'_f B'_{\log p}\big). \tag{31}$$

The third norm $\left\|\nabla\nabla^\top \log V\right\|_2$ in the bracket of the r.h.s of Eq. (15) needs some more effort. From Eq. (20), we have $\partial_b \log V = -\sum_{cij} (J_{f^{-1}})_{ci} (\partial_i \partial_j f_c) (J_{f^{-1}})_{bj}$, thus

$$\begin{aligned}\partial_a \partial_b \log V =& -\sum_{cij} \partial_a (J_{f^{-1}})_{ci} (\partial_i \partial_j f_c) (J_{f^{-1}})_{bj} - \sum_{cij} (J_{f^{-1}})_{ci} (\partial_i \partial_j f_c) \partial_a (J_{f^{-1}})_{bj} \\ &- \sum_{cij} (J_{f^{-1}})_{ci} \partial_a (\partial_i \partial_j f_c) (J_{f^{-1}})_{bj} \\ =& -\sum_{cij} (\partial_a \partial_c f_i^{-1}) (\partial_i \partial_j f_c) (J_{f^{-1}})_{bj} - \sum_{cij} (J_{f^{-1}})_{ci} (\partial_i \partial_j f_c) (\partial_a \partial_b f_j^{-1}) \\ &- \sum_{cijk} (J_{f^{-1}})_{ci} (\partial_a f_k^{-1}) (\partial_k \partial_i \partial_j f_c) (J_{f^{-1}})_{bj} \end{aligned}$$

$$
\begin{aligned}
&= \sum_{cijd} (J_{f^{-1}})_{di} K^d_{ac} (\partial_i \partial_j f_c)(J_{-1})_{bj} + \sum_{cijd} (J_{f^{-1}})_{ci} (\partial_i \partial_j f_c)(J_{f^{-1}})_{dj} K^d_{ab} \\
&\quad - \sum_{cijk} (J_{f^{-1}})_{ci} (\partial_k \partial_i \partial_j f_c)(J_{f^{-1}})_{ak} (J_{f^{-1}})_{bj} \\
&= \sum_{cd} K^d_{ac} K^c_{db} + \sum_{cd} K^c_{cd} K^d_{ab} - \sum_{cijk} (J_{f^{-1}})_{ci} (\partial_k \partial_i \partial_j f_c)(J_{f^{-1}})_{ak} (J_{f^{-1}})_{bj},
\end{aligned}
$$

where we have used Eq. (19) in the third equality for the first two terms. In matrix form, we have:

$$
\nabla \nabla^\top \log V = \sum_{cd} K^d_{:c} K^c_{d:} + \sum_{cd} K^c_{cd} K^d - \sum_{cijk} (J_{f^{-1}})_{ci} (\partial_k \partial_i \partial_j f_c)(J_{f^{-1}})_{:k} (J_{f^{-1}})^\top_{:j}.
$$

We now bound the norms of the three terms in turn. For the first term,

$$
\begin{aligned}
&\Big\| \sum_{cd} K^d_{:c} K^c_{d:} \Big\|_2 \leqslant \sum_{cd} \big\| K^d_{:c} K^c_{d:} \big\|_2 = \sum_{cd} \big| K^c_{d:} K^d_{:c} \big| \\
&= \sum_{cd} \Big| (J_{f^{-1}})_{d:} (\nabla \nabla^\top f_c) J^\top_{f^{-1}} J_{f^{-1}} (\nabla \nabla^\top f_d) (J_{f^{-1}})^\top_{c:} \Big| \\
&\leqslant \sum_{cd} \big| (J_{f^{-1}})_{d:} (J_{f^{-1}})^\top_{c:} \big| \Big\| (\nabla \nabla^\top f_c) J^\top_{f^{-1}} J_{f^{-1}} (\nabla \nabla^\top f_d) \Big\|_2 \\
&\leqslant B'^2_{f^{-1}} B''^2_f \sum_{cd} \Big| (J_{f^{-1}} J^\top_{f^{-1}})_{dc} \Big| \leqslant d^{3/2} B'^2_{f^{-1}} B''^2_f \Big\| J_{f^{-1}} J^\top_{f^{-1}} \Big\|_2 \\
&\leqslant d^{3/2} B'^4_{f^{-1}} B''^2_f,
\end{aligned} \tag{32}
$$

where we have used Eq. (28) in the fourth last inequality and Eq. (29) in the second last inequality. For the second term,

$$
\begin{aligned}
&\Big\| \sum_{cd} K^c_{cd} K^d \Big\|_2 \leqslant \sum_{cd} |K^c_{cd}| \big\| K^d \big\|_2 \leqslant B'^2_{f^{-1}} B''_f \sum_{cd} |K^c_{cd}| \\
&\leqslant d^{1/2} B'^2_{f^{-1}} B''_f \sum_c \sqrt{\sum_d |K^c_{cd}|^2} = d^{1/2} B'^2_{f^{-1}} B''_f \sum_c \| K^c_{c:} \|_2 \\
&\leqslant d^{1/2} B'^2_{f^{-1}} B''_f \sum_c \big\| (J_{f^{-1}})_{c:} \big\|_2 \Big\| (\nabla \nabla^\top f_c) J^\top_{f^{-1}} \Big\|_2 \leqslant d^{1/2} B'^3_{f^{-1}} B''^2_f \sum_c \big\| (J_{f^{-1}})_{c:} \big\|_2 \\
&\leqslant d^{3/2} B'^4_{f^{-1}} B''^2_f,
\end{aligned} \tag{33}
$$

where we have used Eq. (22) in the last inequality. For the third term,

$$
\begin{aligned}
&\Big\| \sum_{cijk} (J_{f^{-1}})_{ci} (\partial_k \partial_i \partial_j f_c)(J_{f^{-1}})_{:k} (J_{f^{-1}})^\top_{:j} \Big\|_2 \\
&\leqslant \sum_{cijk} \big| (J_{f^{-1}})_{ci} (\partial_k \partial_i \partial_j f_c) \big| \big\| (J_{f^{-1}})_{:k} (J_{f^{-1}})^\top_{:j} \big\|_2 \leqslant B'''_f \sum_{ci} \big| (J_{f^{-1}})_{ci} \big| \sum_{jk} \big\| (J_{f^{-1}})_{:k} (J_{f^{-1}})^\top_{:j} \big\|_2 \\
&\leqslant d^{3/2} B'''_f \big\| J_{f^{-1}} \big\|_2 \sum_{jk} \big| (J_{f^{-1}})^\top_{:k} (J_{f^{-1}})_{:j} \big| \leqslant d^{3/2} B'''_f B'_{f^{-1}} \sum_{jk} \Big| (J^\top_{f^{-1}} J_{f^{-1}})_{kj} \Big| \\
&\leqslant d^3 B'''_f B'_{f^{-1}} \Big\| J^\top_{f^{-1}} J_{f^{-1}} \Big\|_2 \leqslant d^3 B'''_f B'^3_{f^{-1}},
\end{aligned} \tag{34}
$$

where we have used Eq. (29) in the fourth last and second last inequalities.

Finally, by assembling Eqs. (30, 31, 32, 33, 34) into Eq. (15), we have:

$$
\begin{aligned}
|p(x) - (\bar{p}_z V)(x)| &\leqslant \frac{1}{2} \mathbb{E}[\mu^\top \mu] B_p B'^d_{f^{-1}} \big(B'^2_{f^{-1}} B'^2_{\log p} + 2d B'^3_{f^{-1}} B''_f B'_{\log p} + d^{3/2} B'^4_{f^{-1}} B''^2_f \\
&\quad + B'^2_{f^{-1}} (B''_{\log p} + d B'_{f^{-1}} B''_f B'_{\log p}) + 2 d^{3/2} B'^4_{f^{-1}} B''^2_f + d^3 B'''_f B'^3_{f^{-1}} \big) \\
&= \frac{1}{2} \mathbb{E}[\mu^\top \mu] B_p B'^{d+2}_{f^{-1}} \big(B'^2_{\log p} + B''_{\log p} + 3 d B'_{f^{-1}} B''_f B'_{\log p} \\
&\quad + 3 d^{3/2} B'^2_{f^{-1}} B''^2_f + d^3 B'''_f B'_{f^{-1}} \big).
\end{aligned}
$$

So given the condition **(iii)**, this residual can be neglected. □

A.3 Proof of the OOD Generalization Error Bound Thm. 6

We give the following more detailed version of Thm. 6 and prove it. The theorem in the main context corresponds to conclusion **(ii)** below (*i.e.*, Eq. (38) below recovers Eq. (6)), by taking the CSGs p', p and $\tilde{p}$, as the semantic-identified CSG p on the training domain, and the ground-truth CSGs p^* and $\tilde{p}^*$ on the training and test domains, respectively. In the theorem in the main context, the semantic-identification requirement on the learned CSG p is to guarantee that it is semantic-equivalent to the ground-truth CSG p^* on the training domain, so that the condition in conclusion **(ii)** below is satisfied.

Theorem 6' (OOD generalization error)**.** *Let Assumption 3 hold.* **(i)** *Consider two CSGs p and $\tilde{p}$ that share the same generative mechanisms $p(x|s,v)$ and $p(y|s)$ but have different priors $p_{s,v}$ and $\tilde{p}_{s,v}$. Then up to $O(\sigma_\mu^2)$ where $\sigma_\mu^2 := \mathbb{E}[\mu^\top \mu]$, we have for any $x \in \operatorname{supp}(p_x) \cap \operatorname{supp}(\tilde{p}_x)$,*

$$\left|\mathbb{E}[y|x] - \tilde{\mathbb{E}}[y|x]\right| \leqslant \sigma_\mu^2 \|\nabla g\|_2 \left\|J_{f^{-1}}\right\|_2^2 \left\|\nabla \log(p_{s,v}/\tilde{p}_{s,v})\right\|_2\Big|_{(s,v)=f^{-1}(x)}, \tag{35}$$

where $J_{f^{-1}}$ is the Jacobian of f^{-1}. Further assume that the bounds B's defined in Thm. 5'**(iii)** *hold. Then the error is negligible for any $x \in \operatorname{supp}(p_x) \cap \operatorname{supp}(\tilde{p}_x)$ if $\frac{1}{\sigma_\mu^2} \gg B'_{\log p} B'_g B'^2_{f^{-1}}$, and:*

$$\begin{aligned}\mathbb{E}_{\tilde{p}(x)}\left|\mathbb{E}[y|x] - \tilde{\mathbb{E}}[y|x]\right|^2 &\leqslant \sigma_\mu^4 B'^2_g B'^4_{f^{-1}} \mathbb{E}_{\tilde{p}_{s,v}} \left\|\nabla \log(p_{s,v}/\tilde{p}_{s,v})\right\|_2^2 \\ &= \sigma_\mu^4 B'^2_g B'^4_{f^{-1}} \mathbb{E}_{\tilde{p}_{s,v}}[2\Delta \log p_{s,v} - \Delta \log \tilde{p}_{s,v} + \|\nabla \log p_{s,v}\|_2^2] \end{aligned} \tag{36}$$

if $\operatorname{supp}(p_x) = \operatorname{supp}(\tilde{p}_x)$, where Δ denotes the Laplacian operator.

(ii) *Let p' be a CSG that is semantic-equivalent to the CSG p introduced in* **(i)**. *Then up to $O(\sigma_\mu^2)$, we have for any $x \in \operatorname{supp}(p'_x) \cap \operatorname{supp}(\tilde{p}_x)$,*

$$\left|\mathbb{E}'[y|x] - \tilde{\mathbb{E}}[y|x]\right| \leqslant \sigma_\mu^2 \|\nabla g'\|_2 \left\|J_{f'^{-1}}\right\|_2^2 \left\|\nabla \log(p'_{s,v}/\tilde{p}'_{s,v})\right\|_2\Big|_{(s,v)=f'^{-1}(x)}, \tag{37}$$

where $\tilde{p}'_{s,v} := \Phi_\#[\tilde{p}_{s,v}]$ is the prior of CSG $\tilde{p}$ under the parameterization of CSG p', derived as the pushed-forward distribution by the reparameterization $\Phi := f'^{-1} \circ f$ from p to p'. Similarly,

$$\mathbb{E}_{\tilde{p}(x)}\left|\mathbb{E}'[y|x] - \tilde{\mathbb{E}}[y|x]\right|^2 \leqslant \sigma_\mu^4 B'^2_g B'^4_{f^{-1}} \mathbb{E}_{\tilde{p}'_{s,v}} \left\|\nabla \log(p'_{s,v}/\tilde{p}'_{s,v})\right\|_2^2 \tag{38}$$

$$= \sigma_\mu^4 B'^2_g B'^4_{f^{-1}} \mathbb{E}_{\tilde{p}'_{s,v}}[2\Delta \log p'_{s,v} - \Delta \log \tilde{p}'_{s,v} + \left\|\nabla \log p'_{s,v}\right\|_2^2]. \tag{39}$$

In the expected OOD generalization error in Eqs. (36, 39), the term $\mathbb{E}_{\tilde{p}_{s,v}}[2\Delta \log p_{s,v} - \Delta \log \tilde{p}_{s,v} + \|\nabla \log p_{s,v}\|_2^2]$ is actually the score matching objective (Fisher divergence) [47] that measures the difference between $\tilde{p}_{s,v}$ and $p_{s,v}$. For Gaussian priors $p(s,v) = \mathcal{N}(0,\Sigma)$ and $\tilde{p}(s,v) = \mathcal{N}(0,\tilde{\Sigma})$, the term reduces to the matrix trace, $\operatorname{tr}(-2\Sigma^{-1} + \tilde{\Sigma}^{-1} + \Sigma^{-1}\tilde{\Sigma}\Sigma^{-1})$. For $\Sigma = \tilde{\Sigma}$, the term vanishes.

For conclusion **(ii)**, note that since p and p' are semantic-equivalent, we have $p'_x = p_x$ and $\mathbb{E}'[y|x] = \mathbb{E}[y|x]$ (from Lemma 9). So Eqs. (35, 37) and Eqs. (36, 39) bound the same quantity. Equation (37) expresses the bound using the structures of the CSG p'. It is considered since recovering the exact CSG p from (x,y) data is impractical and we can only learn a CSG p' that is semantic-equivalent to p.

Proof. Following the proof A.2 of Thm. 5', we assume the additive noise variables μ and ν (for continuous y) have zero mean without loss of generality, and we denote $z := (s,v)$.

Proof under condition (i). Under the assumptions, we have Eq. (14) in the proof A.2 of Thm. 5' hold. Noting that the two CSGs share the same $\bar{g}$ and V (since they share the same $p(x|s,v)$ and $p(y|s)$ thus f and g), we have for any $x \in \operatorname{supp}(p_x) \cap \operatorname{supp}(\tilde{p}_x)$,

$$\begin{aligned}\mathbb{E}[y|x] &= \bar{g} + \frac{1}{2}\mathbb{E}_{p(\mu)}\left[\mu^\top \left((\nabla \log \bar{p}_z V)\nabla \bar{g}^\top + \nabla \bar{g}(\nabla \log \bar{p}_z V)^\top + \nabla \nabla^\top \bar{g}\right)\mu\right] + O(\sigma_\mu^3), \\ \tilde{\mathbb{E}}[y|x] &= \bar{g} + \frac{1}{2}\mathbb{E}_{p(\mu)}\left[\mu^\top \left((\nabla \log \tilde{\bar{p}}_z V)\nabla \bar{g}^\top + \nabla \bar{g}(\nabla \log \tilde{\bar{p}}_z V)^\top + \nabla \nabla^\top \bar{g}\right)\mu\right] + O(\sigma_\mu^3), \end{aligned} \tag{40}$$

where we have similarly defined $\bar{\tilde{p}}_z := \tilde{p}_z \circ f^{-1}$. By subtracting the two equations, we have that up to $O(\sigma_\mu^2)$,

$$\begin{aligned}
\Big|\mathbb{E}[y|x] - \tilde{\mathbb{E}}[y|x]\Big| &= \frac{1}{2}\Big|\mathbb{E}_{p(\mu)}\big[\mu^\top\big(\nabla \log(\bar{p}_z/\bar{\tilde{p}}_z)\nabla\bar{g}^\top + \nabla\bar{g}\nabla\log(\bar{p}_z/\bar{\tilde{p}}_z)^\top\big)\mu\big]\Big| \\
&\leqslant \frac{1}{2}\mathbb{E}_{p(\mu)}\big[\big|\mu^\top\big(\nabla \log(\bar{p}_z/\bar{\tilde{p}}_z)\nabla\bar{g}^\top + \nabla\bar{g}\nabla\log(\bar{p}_z/\bar{\tilde{p}}_z)^\top\big)\mu\big|\big] \\
&\leqslant \frac{1}{2}\mathbb{E}_{p(\mu)}\big[\|\mu\|_2^2\big(\big\|\nabla \log(\bar{p}_z/\bar{\tilde{p}}_z)\nabla\bar{g}^\top\big\|_2 + \big\|\nabla\bar{g}\nabla\log(\bar{p}_z/\bar{\tilde{p}}_z)^\top\big\|_2\big)\big] \\
&= \big|\nabla\bar{g}^\top\nabla\log(\bar{p}_z/\bar{\tilde{p}}_z)\big|\mathbb{E}[\mu^\top\mu]. \qquad (41)
\end{aligned}$$

The multiplicative factor to $\mathbb{E}[\mu^\top\mu]$ on the right hand side can be further bounded by:

$$\begin{aligned}
\big|\nabla\bar{g}^\top\nabla\log(\bar{p}_z/\bar{\tilde{p}}_z)\big| &= \big|(J_{(f^{-1})^\mathcal{S}}\nabla g)^\top(J_{f^{-1}}\nabla\log(p_z/\tilde{p}_z))\big| \\
&= \Big|\nabla g^\top J_{(f^{-1})^\mathcal{S}}^\top J_{f^{-1}}\nabla\log(p_z/\tilde{p}_z)\Big| \\
&= \Big|((\nabla g)^\top, 0_{d_\mathcal{V}}^\top)J_{f^{-1}}^\top J_{f^{-1}}\nabla\log(p_z/\tilde{p}_z)\Big| \\
&\leqslant \|\nabla g\|_2\big\|J_{f^{-1}}\big\|_2^2\|\nabla\log(p_z/\tilde{p}_z)\|_2, \qquad (42)
\end{aligned}$$

where ∇g and $\nabla\log(p_z/\tilde{p}_z)$ are evaluated at $z = f^{-1}(x)$. This gives:

$$\Big|\mathbb{E}[y|x] - \tilde{\mathbb{E}}[y|x]\Big| \leqslant \sigma_\mu^2\|\nabla g\|_2\big\|J_{f^{-1}}\big\|_2^2\|\nabla\log(p_z/\tilde{p}_z)\|_2,$$

i.e. Eq. (35) in conclusion **(i)**. When the bounds B's in Thm. 5'**(iii)** hold, we further have:

$$\begin{aligned}
\Big|\mathbb{E}[y|x] - \tilde{\mathbb{E}}[y|x]\Big| &\leqslant \sigma_\mu^2\|\nabla g\|_2\big\|J_{f^{-1}}\big\|_2^2\|\nabla\log p_z - \nabla\log\tilde{p}_z\|_2 \\
&\leqslant \sigma_\mu^2\|\nabla g\|_2\big\|J_{f^{-1}}\big\|_2^2(\|\nabla\log p_z\|_2 + \|\nabla\log\tilde{p}_z\|_2) \\
&\leqslant 2\sigma_\mu^2 B_g' B_{f^{-1}}'^2 B_{\log p}'.
\end{aligned}$$

So when $\frac{1}{\sigma_\mu^2} \gg B_{\log p}' B_g' B_{f^{-1}}'^2$, this difference is negligible for any $x \in \operatorname{supp}(p_x) \cap \operatorname{supp}(\tilde{p}_x)$.

We now turn to the expected OOD generalization error Eq. (36) in conclusion **(i)**. When $\operatorname{supp}(p_x) = \operatorname{supp}(\tilde{p}_x)$, Eq. (35) hold on $\tilde{p}_x$. Together with the bounds in Thm. 5'**(iii)**, we have:

$$\begin{aligned}
\mathbb{E}_{\tilde{p}(x)}\Big|\mathbb{E}[y|x] - \tilde{\mathbb{E}}[y|x]\Big|^2 &\leqslant \sigma_\mu^4 B_g'^2 B_{f^{-1}}'^4\mathbb{E}_{\tilde{p}(x)}\Big\|\nabla\log(p_z/\tilde{p}_z)\big|_{z=f^{-1}(x)}\Big\|_2^2 \\
&= \sigma_\mu^4 B_g'^2 B_{f^{-1}}'^4\mathbb{E}_{\tilde{p}_z}\|\nabla\log(p_z/\tilde{p}_z)\|_2^2,
\end{aligned}$$

where the equality holds due to the generating process of the model. Note that the term $\mathbb{E}_{\tilde{p}_z}\|\nabla\log(p_z/\tilde{p}_z)\|_2^2$ therein is the score matching objective (Fisher divergence). By Hyvärinen [47, Thm. 1], we can reformulate it as $\mathbb{E}_{\tilde{p}_z}[2\Delta\log p_z - \Delta\log\tilde{p}_z + \|\nabla\log p_z\|_2^2]$, so we have:

$$\mathbb{E}_{\tilde{p}(x)}\Big|\mathbb{E}[y|x] - \tilde{\mathbb{E}}[y|x]\Big|^2 \leqslant \sigma_\mu^4 B_g'^2 B_{f^{-1}}'^4\mathbb{E}_{\tilde{p}_z}[2\Delta\log p_z - \Delta\log\tilde{p}_z + \|\nabla\log p_z\|_2^2].$$

Proof under condition (ii). From Eq. (14) in the proof A.2 of Thm. 5', we have for CSG p' that for any $x \in \operatorname{supp}(p_x')$ or equivalently $x \in \operatorname{supp}(p_x)$,

$$\mathbb{E}'[y|x] = \bar{g}' + \frac{1}{2}\mathbb{E}_{p(\mu)}\big[\mu^\top\big((\nabla\log\bar{p}_z' V')\nabla\bar{g}'^\top + \nabla\bar{g}'(\nabla\log\bar{p}_z' V')^\top + \nabla\nabla^\top\bar{g}'\big)\mu\big] + O(\sigma_\mu^3), \quad (43)$$

where we have similarly defined $\bar{p}_z' := p_z' \circ f'^{-1}$ and $\bar{g}' := g' \circ (f'^{-1})^\mathcal{S}$. Since p and p' are semantic-equivalent with reparameterization Φ from p to p', we have $p(y|s) = p'(y|\Phi^\mathcal{S}(s,v))$ thus $g(s) = g'(\Phi^\mathcal{S}(s,v))$ for any $v \in \mathcal{V}$. So for any $x \in \operatorname{supp}(p_x)$ or equivalently $x \in \operatorname{supp}(p_x')$, we have $g((f^{-1})^\mathcal{S}(x)) = g'(\Phi^\mathcal{S}((f^{-1})^\mathcal{S}(x), (f^{-1})^\mathcal{V}(x))) = g'(\Phi^\mathcal{S}(f^{-1}(x))) = g'((f'^{-1})^\mathcal{S}(f(f^{-1}(x)))) = g'((f'^{-1})^\mathcal{S}(x))$, *i.e.*, $\bar{g} = \bar{g}'$. For another fact, since $\tilde{p}_z' := \Phi_\#[\tilde{p}_z] = (f'^{-1}\circ f)_\#[\tilde{p}_z]$ by definition, we have $f'_\#[\bar{\tilde{p}}_z'] = f_\#[\tilde{p}_z]$, *i.e.*, $\bar{\tilde{p}}_z' V' = \bar{\tilde{p}}_z V$. Subtracting Eqs. (43, 40) and applying these two facts, we have up to $O(\sigma_\mu^2)$, for any $x \in \operatorname{supp}(p_x') \cap \operatorname{supp}(\tilde{p}_x)$,

$$\Big|\mathbb{E}'[y|x] - \tilde{\mathbb{E}}[y|x]\Big| = \frac{1}{2}\Big|\mathbb{E}_{p(\mu)}\big[\mu^\top\big(\nabla\log(\bar{p}_z'/\bar{\tilde{p}}_z')\nabla\bar{g}'^\top + \nabla\bar{g}'\nabla\log(\bar{p}_z'/\bar{\tilde{p}}_z')^\top\big)\mu\big]\Big|$$

$$\leqslant \left|\nabla \bar{g}'^{\top} \nabla \log(\bar{p}'_z/\tilde{\bar{p}}'_z)\right| \mathbb{E}[\mu^{\top}\mu],$$

where the inequality follows Eq. (41). Using a similar result of Eq. (42), we have:

$$\left|\mathbb{E}'[y|x] - \tilde{\mathbb{E}}[y|x]\right| \leqslant \sigma_\mu^2 \|\nabla g'\|_2 \left\|J_{f'^{-1}}\right\|_2^2 \left\|\nabla \log(p'_z/\tilde{p}'_z)\right\|_2,$$

where $\nabla g'$ and $\nabla \log(p'_z/\tilde{p}'_z)$ are evaluated at $z = f'^{-1}(x)$. This gives Eq. (37). Derivation of Eqs. (38, 39) is similar as in conclusion **(i)**. □

A.4 Proof of the Domain Adaptation Error Thm. 7

To be consistent with the notation in the proofs, we prove the theorem by denoting the semantic-identified CSG p and the ground-truth CSG $\tilde{p}^*$ on the test domain as p' and $\tilde{p}$, respectively.

Proof. The new prior $\tilde{p}'(z)$ is learned by fitting unsupervised data from the test domain $\tilde{p}(x)$. Applying the deduction in the proof A.2 of Thm. 5' to the test domain, we have that under any of the three conditions in Thm. 5', $\tilde{p}(x) = \tilde{p}'(x)$ indicates $f_\#[\tilde{p}_z] = f'_\#[\tilde{p}'_z]$. This gives $\tilde{p}'_z = (f'^{-1} \circ f)_\#[\tilde{p}_z] = \Phi_\#[\tilde{p}_z]$.

From Eq. (12) in the same proof, we have that:

$$\tilde{p}(x)\tilde{\mathbb{E}}[y|x] = (f_\#[g\tilde{p}_z] * p_\mu)(x) = ((f_\#[\tilde{p}_z]\bar{g}) * p_\mu)(x),$$
$$\tilde{p}'(x)\tilde{\mathbb{E}}'[y|x] = (f'_\#[g'\tilde{p}'_z] * p_\mu)(x) = ((f'_\#[\tilde{p}'_z]\bar{g}') * p_\mu)(x).$$

From the proof A.3 of Thm. 6'**(ii)** (the paragraph under Eq. (43)), the semantic-equivalence between CSGs p and p' indicates that $\bar{g} = \bar{g}'$. So from the above two equations, we have $\tilde{p}(x)\tilde{\mathbb{E}}[y|x] = \tilde{p}'(x)\tilde{\mathbb{E}}'[y|x]$ (recall that $\tilde{p}(x) = \tilde{p}'(x)$ indicates $f_\#[\tilde{p}_z] = f'_\#[\tilde{p}'_z]$). Since $\tilde{p}(x) = \tilde{p}'(x)$ (that is how $\tilde{p}'_z$ is learned), we have for any $x \in \text{supp}(\tilde{p}_x)$ or equivalently $x \in \text{supp}(\tilde{p}'_x)$,

$$\tilde{\mathbb{E}}'[y|x] = \tilde{\mathbb{E}}[y|x]. \tag{44}$$

□

B Alternative Identifiability Theory for CSG

The presented identifiability theory, particularly Thm. 5, shows that the semantic-identifiability can be achieved in the deterministic limit ($\frac{1}{\sigma_\mu^2} \to \infty$), but does not quantitatively describe the extent of violation of the identifiability for a finite variance σ_μ^2. Here we define a "soft" version of semantic-equivalence and show that it can be achieved with a finite variance, with a trade-off between the "softness" and the variance.

Definition 15 (δ-semantic-dependency)**.** For $\delta > 0$ and two CSGs p and p', we say that they are δ-semantic-dependent, if there exists a homeomorphism Φ on $\mathcal{S} \times \mathcal{V}$ such that: **(i)** $p(x|s,v) = p'(x|\Phi(s,v))$, **(ii)** $\sup_{v\in\mathcal{V}} \left\|g(s) - g'(\Phi^{\mathcal{S}}(s,v))\right\|_2 \leqslant \delta$ where we have denoted $g(s) := \mathbb{E}[y|s]$, and **(iii)** $\sup_{v^{(1)},v^{(2)}\in\mathcal{V}} \left\|\Phi^{\mathcal{S}}(s,v^{(1)}) - \Phi^{\mathcal{S}}(s,v^{(2)})\right\|_2 \leqslant \delta$.

In the definition, we have released the prior conversion requirement, and relaxed the exact likelihood conversion for $p(y|s)$ in **(ii)** and the v-constancy of $\Phi^{\mathcal{S}}$ in **(iii)** to allow an error bounded by δ. When $\delta = 0$, the v-constancy of $\Phi^{\mathcal{S}}$ is exact, and under the additive noise Assumption 3 we also have the exact likelihood conversion $p(y|s) = p'(y|\Phi^{\mathcal{S}}(s,v))$ for any $v \in \mathcal{V}$. So 0-semantic-dependency with the prior conversion requirement reduces to the semantic-equivalence.

Due to the quantitative nature, the binary relation cannot be made an equivalence relation but only a dependency. Here, a dependency refers to a binary relation with reflexivity and symmetry, but no transitivity.

Proposition 16. *The δ-semantic-dependency is a dependency relation if the function $g := \mathbb{E}[y|s]$ is bijective and its inverse g^{-1} is $\frac{1}{2}$-Lipschitz.*

Proof. Showing a dependency relation amounts to showing the following two properties.

- Reflexivity. For two identical CSGs p and p', we have $p(x|s,v) = p'(x|s,v)$ and $p(y|s) = p'(y|s)$. So the identity map as Φ obviously satisfies all the requirements in Def. 15.

- Symmetry. Let CSG p be δ-semantic-dependent to CSG p' with homeomorphism Φ. Obviously Φ^{-1} is also a homeomorphism. For any $(s', v') \in \mathcal{S} \times \mathcal{V}$, we have $p'(x|s', v') = p'(x|\Phi(\Phi^{-1}(s', v'))) = p(x|\Phi^{-1}(s', v'))$, and $\left\|g'(s') - g((\Phi^{-1})^{\mathcal{S}}(s', v'))\right\|_2 = \left\|g'(\Phi^{\mathcal{S}}(s, v)) - g(s)\right\|_2 \leqslant \delta$ where we have denoted $(s, v) := \Phi^{-1}(s', v')$ here. So Φ^{-1} satisfies requirements **(i)** and **(ii)** in Def. 15.

  For requirement **(iii)**, we need the following fact: for any $s^{(1)}, s^{(2)} \in \mathcal{S}$, $\left\|s^{(1)} - s^{(2)}\right\|_2 = \left\|g^{-1}(g(s^{(1)})) - g^{-1}(g(s^{(2)}))\right\|_2 \leqslant \frac{1}{2}\left\|g(s^{(1)}) - g(s^{(2)})\right\|_2$, where the inequality holds since g^{-1} is $\frac{1}{2}$-Lipschitz. Then for any $s' \in \mathcal{S}$, we have:

$$
\begin{aligned}
&\sup_{v'^{(1)}, v'^{(2)} \in \mathcal{V}} \left\|(\Phi^{-1})^{\mathcal{S}}(s', v'^{(1)}) - (\Phi^{-1})^{\mathcal{S}}(s', v'^{(2)})\right\|_2 \\
\leqslant &\sup_{v'^{(1)}, v'^{(2)} \in \mathcal{V}} \frac{1}{2}\left\|g\big((\Phi^{-1})^{\mathcal{S}}(s', v'^{(1)})\big) - g\big((\Phi^{-1})^{\mathcal{S}}(s', v'^{(2)})\big)\right\|_2 \\
= &\sup_{v'^{(1)}, v'^{(2)} \in \mathcal{V}} \frac{1}{2}\left\|\Big(g\big((\Phi^{-1})^{\mathcal{S}}(s', v'^{(1)})\big) - g'(s')\Big) - \Big(g\big((\Phi^{-1})^{\mathcal{S}}(s', v'^{(2)})\big) - g'(s')\Big)\right\|_2 \\
\leqslant &\sup_{v'^{(1)}, v'^{(2)} \in \mathcal{V}} \frac{1}{2}\Big(\left\|g\big((\Phi^{-1})^{\mathcal{S}}(s', v'^{(1)})\big) - g'(s')\right\|_2 + \left\|g\big((\Phi^{-1})^{\mathcal{S}}(s', v'^{(2)})\big) - g'(s')\right\|_2\Big) \\
= &\frac{1}{2}\Big(\sup_{v'^{(1)} \in \mathcal{V}} \left\|g\big((\Phi^{-1})^{\mathcal{S}}(s', v'^{(1)})\big) - g'(s')\right\|_2 + \sup_{v'^{(2)} \in \mathcal{V}} \left\|g\big((\Phi^{-1})^{\mathcal{S}}(s', v'^{(2)})\big) - g'(s')\right\|_2\Big) \\
\leqslant &\,\delta,
\end{aligned}
$$

  where in the last inequality we have used the fact that Φ^{-1} satisfies requirement **(ii)**. So p' is δ-semantic-dependent to p via the homeomorphism Φ^{-1}.

□

The corresponding δ-semantic-identifiability result follows.

Theorem 17 (δ-semantic-identifiability)**.** *Assume the same as Thm. 5' and Prop. 16, and let the bounds B's defined in* Thm. 5'**(iii)** *hold. For two such CSGs p and p', if they have $p(x, y) = p'(x, y)$, then they are δ-semantic-dependent for any $\delta \geqslant \sigma_\mu^2 B_{f^{-1}}'^2\big(2B'_{\log p}B'_g + B''_g + 3dB'_{f^{-1}}B''_f B'_g\big)$, where $d := d_{\mathcal{S}} + d_{\mathcal{V}}$.*

Proof. Let $\Phi := f'^{-1} \circ f$, where f and f' are given by the two CSGs p and p' via the additive noise Assumption 3. We now show that p and p' are δ-semantic-dependent via this Φ for any δ in the theorem. Obviously Φ is a homeomorphism on $\mathcal{S} \times \mathcal{V}$, and it satisfies requirement **(i)** in Def. 15 by construction due to Eq. (7) in the proof A.2 of Thm. 5'.

Consider requirement **(ii)** in Def. 15. Based on the same assumptions as Thm. 5', we have Eq. (25) hold for both CSGs:

$$
\max\{\|\mathbb{E}[y|x] - \bar{g}(x)\|_2, \|\mathbb{E}'[y|x] - \bar{g}'(x)\|_2\} \leqslant \sigma_\mu^2 B_{f^{-1}}'^2\big(B'_{\log p}B'_g + \frac{1}{2}B''_g + \frac{3}{2}dB'_{f^{-1}}B''_f B'_g\big),
$$

where we have denoted $\sigma_\mu^2 := \mathbb{E}[\mu^\top \mu]$. Since both CSGs induce the same $p(y|x)$, so $\mathbb{E}[y|x] = \mathbb{E}'[y|x]$. This gives:

$$
\begin{aligned}
\|\bar{g}(x) - \bar{g}'(x)\|_2 &= \left\|\big(\mathbb{E}'[y|x] - \bar{g}'(x)\big) - \big(\mathbb{E}[y|x] - \bar{g}(x)\big)\right\|_2 \\
&\leqslant \|\mathbb{E}'[y|x] - \bar{g}'(x)\|_2 + \|\mathbb{E}[y|x] - \bar{g}(x)\|_2 \\
&\leqslant \sigma_\mu^2 B_{f^{-1}}'^2\big(2B'_{\log p}B'_g + B''_g + 3dB'_{f^{-1}}B''_f B'_g\big).
\end{aligned}
$$

So for any $(s, v) \in \mathcal{S} \times \mathcal{V}$, by denoting $x := f(s, v)$, we have:

$$
\begin{aligned}
&\left\|g(s) - g'(\Phi^{\mathcal{S}}(s, v))\right\|_2 = \left\|g((f^{-1})^{\mathcal{S}}(x)) - g'((f'^{-1})^{\mathcal{S}}(f(s, v)))\right\|_2 = \|\bar{g}(x) - \bar{g}'(x)\|_2 \\
\leqslant\, &\sigma_\mu^2 B_{f^{-1}}'^2\big(2B'_{\log p}B'_g + B''_g + 3dB'_{f^{-1}}B''_f B'_g\big).
\end{aligned}
$$

So the requirement is satisfied.

For requirement **(iii)**, note from the proof of Prop. 16 that when g is bijective and its inverse is $\frac{1}{2}$-Lipschitz, requirement **(ii)** implies requirement **(iii)**. So this Φ is a homeomorphism that makes p δ-semantic-dependent to p' for any $\delta \geqslant \sigma_\mu^2 B_{f^{-1}}'^2\big(2B'_{\log p}B'_g + B''_g + 3dB'_{f^{-1}}B''_f B'_g\big)$. □

Note that although the δ-semantic-dependency does not have transitivity, the above theorem is still informative: for any two CSGs sharing the same data distribution, particularly for a well-learned CSG p and the ground-truth CSG p^*, the likelihood conversion error $\sup_{(s,v)\in\mathcal{S}\times\mathcal{V}}\left\|g(s)-g'(\Phi^{\mathcal{S}}(s,v))\right\|_2$, and the degree of mixing v into s, measured by $\sup_{v^{(1)},v^{(2)}\in\mathcal{V}}\left\|\Phi^{\mathcal{S}}(s,v^{(1)})-\Phi^{\mathcal{S}}(s,v^{(2)})\right\|_2$, are bounded by $\sigma_\mu^2 B_{f^{-1}}'^2\left(2B'_{\log p}B'_g+B''_g+3dB'_{f^{-1}}B''_fB'_g\right)$.

C More Explanations on the Model

Explanations on our model. We see the data generating process as coming up with a conceptual latent factors (s,v) first, and then generating both x and y based on the factors. A prototyping example is that a photographer takes an image x of an object and meanwhile gives a label y to it, based on conceptual features (s,v) in the scene (*e.g.*, shape, color, texture, orientation and pose of the object, background objects and environment, illumination during imaging). The image x is produced by assembling these factors (s,v) in the scene and passing the reflected light through a camera, and the label y is produced by processing causally relevant factors s (*e.g.*, object shape, texture) by the photographer. Under this view, intervening the image x is to break the imaging process (*e.g.*, by malfunctioning the camera by breaking a sensor unit or making the sensor noisy), which does not alter the latent factors (s,v) and the labeling process, hence also the label y. Similarly, intervening the label y is to break the labeling process (*e.g.*, by reforming the labeling rule or randomly flipping the labels), which does not alter the latent factors (s,v) and the imaging process, hence also the image x. On the other hand, intervening the latent factors (s,v) (*e.g.*, by replacing the object with a different one at the imaging and labeling moment) may change both x and y through the imaging and labeling processes. This verifies the model in Fig. 1a by checking its causal implications.

This view of the data generating process is also adopted and promoted by popular existing works. Mcauliffe and Blei [78] treat both a document and its label be generated by the involved topics in the document (represented as a topic proportion), which is an abstract latent factor. Peters et al. [88, Sec. 1.4]; Kilbertus et al. [59] view the generation of an OCR dataset under a causal perspective as the writer first comes up with an intension to write a character, and then writes down the character and gives its label based on the intension. Teshima et al. [108] treat both an image and its label be produced from a set of latent factors. This view of the data generating process is also natural for medical image datasets, where the label may be diagnosed based on more fundamental features (*e.g.*, PCR test results showing the pathogen) that are not included in the dataset but actually cause the medical image.

On the labeling process from images that one would commonly think of, we also view it as a $s\to y$ process. Human directly knows the critical semantic feature s (*e.g.*, the shape and position of each stroke) by seeing the image, through the nature gift of the vision system [12]. The label is given by processing the feature (*e.g.*, the angle between two linear strokes, the position of a circular stroke relative to a linear stroke), which is a $s\to y$ process.

The causal graph in Fig. 1a implies that $x \perp\!\!\!\perp y \mid s$. However, this does not indicate that the semantic factor s generates an image x regardless of the label y. Given s, the generated image is dictated to hold the given semantics regardless of randomness, so the statistical independence does not mean semantic irrelevance. If an image x is given, the corresponding label is given by $p(y|x)$, which is $\int p(s|x)p(y|s)\,\mathrm{d}s$ by the causal graph. So the semantic concept to cause the label through $p(y|s)$, is inferred from the image through $p(s|x)$.

Comparison with the graph $y_{\mathrm{tx}}\to s\to x\to y_{\mathrm{rx}}$. One may consider this graph as a communication channel, where y_{tx} is a transmitted signal and y_{rx} is the received signal.

If the observed label y is treated as y_{tx}, the graph then implies $y\to s$. This is argued at the end of item **(2)** in Sec. 3 that it may make unreasonable implications. Moreover, the graph also implies that y is a cause of x, as is challenged in item **(1)** in Sec. 3. The unnatural implications arise since intervening y is different from intervening the "ground-truth" label. We consider y as an observation that may be noisy, while the "ground-truth label" is never observed: one cannot tell if the labels at hand are noise-corrupted, based on the dataset alone. For example, the label of either image in Fig. 2 may be given by a labeler's random guess. Our adopted causal direction $s\to y$ is consistent with these examples and is also argued and adopted by Mcauliffe and Blei [78]; Peters et al. [88, Sec. 1.4]; Kilbertus et al. [59]; Teshima et al. [108].

If the observed label y is treated as y_{rx}, the graph then implies $x \to y$, as is challenged in item **(1)** in Sec. 3. It is also argued by Schölkopf et al. [97]; Peters et al. [88, Sec. 1.4]; Kilbertus et al. [59]. Treating the observed label y as y_{rx} and y_{tx} as the "ground-truth" label may be the motivation of this graph. But the graph implies $y_{\mathrm{tx}} \perp\!\!\!\perp y_{\mathrm{rx}} \mid x$, that is, $p(y_{\mathrm{tx}}|x, y_{\mathrm{rx}}) = p(y_{\mathrm{tx}}|x)$ and $p(y_{\mathrm{rx}}|x, y_{\mathrm{tx}}) = p(y_{\mathrm{rx}}|x)$. So modeling y_{tx} (resp. y_{rx}) does not benefit predicting y_{rx} (resp. y_{tx}) from x.

D More Related Work

Generative supervised learning is not new [78, 63], but most works do not consider the encoded causality. Other works consider solving causality tasks, notably causal/treatment effect estimation [76, 118, 114]. The task does not focus on OOD prediction, and requires labels for both treated and controlled groups.

Causality with latent variable has been considered in a rich literature [111, 105, 92, 45, 103], while most works focus on the consequence on observation-level causality. Others consider identifying the latent variable. Janzing et al. [51], Lee et al. [68] show the identifiability under additive noise or similar assumptions. For discrete data, a "simple" latent variable can be identified under various specifications [52, 99, 64]. Romeijn and Williamson [94] consider using interventional datasets for identification. Over these works, we step further to separate and identify the latent variable as semantic and variation factors, and show the benefit for OOD prediction.

E Relation to Existing Domain Adaptation Theory

In this section, to align with the domain adaptation (DA) literature, we call "training/test domain" as "source/target domain", and use $p(x, y)$ and $\tilde{p}(x, y)$ to denote the underlying data-generating distributions $p^*(x, y)$ and $\tilde{p}^*(x, y)$ on the source and target domains, respectively. In a DA task, supervised data from $p(x, y)$ on the source domain are available, but on the target domain, only unsupervised data from $\tilde{p}(x) = \int \tilde{p}(x, y)\,\mathrm{d}y$ [15] are available. The goal is to find a labeling function $h : \mathcal{X} \to \mathcal{Y}$ within a hypothesis space $\mathcal{H}$ that minimizes the target-domain risk $\tilde{R}(h) := \mathbb{E}_{\tilde{p}(x,y)}[\ell(h(x), y)]$ defined by a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$.

General DA theory Since $\tilde{p}(x, y)$ is not accessible, it is of practical interest to consider the source-domain risk $R(h)$ and investigate its relation to $\tilde{R}(h)$. Ben-David et al. [7, Thm. 1] give a bound relating the two risks:

$$\begin{aligned}\tilde{R}(h) \leqslant R(h) &+ 2d_1(p_x, \tilde{p}_x) \\ &+ \min\{\mathbb{E}_{p(x)}[|h^*(x) - \tilde{h}^*(x)|], \mathbb{E}_{\tilde{p}(x)}[|h^*(x) - \tilde{h}^*(x)|]\}, \\ \text{where: } d_1(p_x, \tilde{p}_x) &:= \sup_{X \in \mathscr{X}} |p_x[X] - \tilde{p}_x[X]|\end{aligned} \tag{45}$$

is the *total variation* between the two distributions, $\mathscr{X}$ denotes the sigma-field on $\mathcal{X}$, and $h^* \in \operatorname{argmin}_{h \in \mathcal{H}} R(h)$ and $\tilde{h}^* \in \operatorname{argmin}_{\tilde{h} \in \mathcal{H}} \tilde{R}(\tilde{h})$ are the oracle labeling functions on the source and target domains, respectively (*e.g.*, $h^*(x) = \mathbb{E}[y|x]$ and $\tilde{h}^*(x) = \tilde{\mathbb{E}}[y|x]$ if $\operatorname{supp}(p_x) = \operatorname{supp}(\tilde{p}_x)$). Note that as oracle labeling functions, h^* and $\tilde{h}^*$ are two *certain* but not *any* risk minimizers. The second and third terms on the r.h.s measure the domain difference in terms of the distribution on x and the correspondence of y on x, respectively. Zhao et al. [125, Thm. 4.1] give a similar bound in the case of binary classification $\mathcal{Y} = \{0, 1\}$, in terms of the $\tilde{\mathcal{H}}$-divergence $d_{\tilde{\mathcal{H}}}$ in place of the total variance d_1, which is defined as $d_{\tilde{\mathcal{H}}}(p_x, \tilde{p}_x) := \sup_{X \in \mathscr{X}_{\tilde{\mathcal{H}}}} |p_x[X] - \tilde{p}_x[X]|$, where $\mathscr{X}_{\tilde{\mathcal{H}}} := \{h^{-1}(1) : h \in \tilde{\mathcal{H}}\}$ and $\tilde{\mathcal{H}} := \{\operatorname{sign}(|h(x) - h'(x)| - t) : h, h' \in \mathcal{H}, t \in [0, 1]\}$.

Ben-David et al. [7] also argue that in this bound, the total variation d_1 is overly strict (thus making the bound unnecessarily loose) and hard to estimate from finite data samples, so they develop another bound which is better known (7, Thm. 2; 54, Thm. 1) (only showing the asymptotic version here, *i.e.*, omitting the estimation error from finite samples):

$$\tilde{R}(h) \leqslant R(h) + d_{\mathcal{H}\Delta\mathcal{H}}(p_x, \tilde{p}_x) + \lambda_{\mathcal{H}}, \tag{46}$$

[15] Under the general definition of an integral (*e.g.*, Billingsley [13, p.211]), it also allows a discrete $\mathcal{Y}$, in which case $\mathrm{d}y$ is the counting measure and the integral reduces to a summation.

$$\text{where: } d_{\mathcal{H}\Delta\mathcal{H}}(p_x, \tilde{p}_x) := \sup_{h,h' \in \mathcal{H}} \left| \mathbb{E}_{p(x)}[\ell(h(x), h'(x))] - \mathbb{E}_{\tilde{p}(x)}[\ell(h(x), h'(x))] \right|,$$

$$\lambda_{\mathcal{H}} := \inf_{h \in \mathcal{H}} \left[R(h) + \tilde{R}(h) \right].$$

Here, $d_{\mathcal{H}\Delta\mathcal{H}}(p_x, \tilde{p}_x)$ is called the *$\mathcal{H}\Delta\mathcal{H}$-divergence* measuring the difference between $p(x)$ and $\tilde{p}(x)$, under the discriminative efficacy of the labeling function family $\mathcal{H}$ (thus not as strict as the total variation d_1), and $\lambda_{\mathcal{H}}$ is the *ideal joint risk* achieved by $\mathcal{H}$ measuring the richness or expressiveness of $\mathcal{H}$ for the two prediction tasks. The $\mathcal{H}\Delta\mathcal{H}$-divergence $d_{\mathcal{H}\Delta\mathcal{H}}$ is also estimable from finite data samples [7, Lemma 1]. Long et al. [73, Thm. 1] give a similar bound in terms of maximum mean discrepancy (MMD) d_K in place of $d_{\mathcal{H}\Delta\mathcal{H}}$.

For successful adaptation, some assumptions on the unknown distribution $\tilde{p}(x, y)$ are required. A commonly adopted one is:

$$\text{(covariate shift) } \tilde{h}^*(x) = h^*(x) \text{ or } p(y|x) = \tilde{p}(y|x), \forall x \in \operatorname{supp}(p_x, \tilde{p}_x) := \operatorname{supp}(p_x) \cup \operatorname{supp}(\tilde{p}_x).$$

DA-DIR Domain-invariant representation (DIR) based DA methods (DA-DIR) [83, 5, 73, 33] aims to learn a deterministic representation extractor $\eta : \mathcal{X} \to \mathcal{S}$ to some representation space $\mathcal{S}$, in order to achieve a domain-invariant representation:

$$\text{(DIR) } p(s) = \tilde{p}(s), \text{where } p(s) := \eta_{\#}[p_x](s) \text{ and } \tilde{p}(s) := \eta_{\#}[\tilde{p}_x](s)$$

are the representation distributions on the two domains. The motivation is that, once DIR is achieved, the distribution difference term (the second term on the r.h.s) of bound Eq. (45) or Eq. (46) diminishes on the representation space $\mathcal{S}$. So the bound on $\mathcal{S}$ is then controlled by the source risk (the first term), and driving h to let $R(h)$ approach $R(h^*)$ (*i.e.*, to minimize the source risk $R(h)$) effectively minimizes the target risk.

Let $g : \mathcal{S} \to \mathcal{Y}$ be a labeling function on the representation space $\mathcal{S}$. The end-to-end labeling function is then $h = g \circ \eta$. Combining the two desiderata of achieving DIR and $R(h^*)$, the typical objective of DA-DIR is in the following form:

$$\min_{\eta \in \mathcal{E}, g \in \mathcal{G}} R(g \circ \eta) + \lambda d(\eta_{\#}[p_x], \eta_{\#}[\tilde{p}_x]),$$

where $d(\cdot, \cdot)$ is a metric or discrepancy ($d(q, p) \geqslant 0$; $d(q, p) = 0 \iff q = p$) on distributions, λ is a weighting parameter, and $\mathcal{E}$ and $\mathcal{G}$ are the hypothesis spaces for η and g, respectively.

For the existence of the solution of this problem, Johansson et al. [54] consider the following assumption:

$$\text{(strong existence assumption) } \exists \eta^* \in \mathcal{E}, g^* \in \mathcal{G}, \text{ s.t. } \eta^*_{\#}[p_x] = \eta^*_{\#}[\tilde{p}_x], g^* \circ \eta^* = h^*.$$

They also mention that this is not guaranteed to hold in practice, since it is quite strong: both DIR and $R(h^*)$ can be simultaneously achieved.

Problem of DA-DIR Johansson et al. [54], Zhao et al. [125] give examples where even under the strong assumption of both covariate shift and the strong existence assumption [54, Assumption 3], simultaneously achieving both DIR and $R(h^*)$ still leads the target risk $\tilde{R}(g \circ \eta)$ to the worst value.

We first analyze the problem through the lens of the above DA bounds. We will show that when reducing the bounds on $\mathcal{S}$, they can be uselessly large.

(1) For the bound Eq. (45). Applying the bound on the representation space $\mathcal{S}$ gives:

$$\begin{aligned} \tilde{R}(g \circ \eta) \leqslant\ & R(g \circ \eta) + 2d_1(\eta_{\#}[p_x], \eta_{\#}[\tilde{p}_x]) \\ & + \min\{\mathbb{E}_{\eta_{\#}[p_x](s)}[|g^*_\eta(s) - \tilde{g}^*_\eta(s)|], \mathbb{E}_{\eta_{\#}[\tilde{p}_x](s)}[|g^*_\eta(s) - \tilde{g}^*_\eta(s)|]\}, \end{aligned} \tag{47}$$

where g^*_η and $\tilde{g}^*_\eta$ are the optimal labeling functions on top of the representation extractor η. It is shown that under the assumption of covariate shift [8, 35] or additionally strong existence [54], simultaneously achieving both DIR and $R(h^*)$ is not sufficient to guarantee $g^*_\eta = \tilde{g}^*_\eta$, so the bound may still be large.

In both examples of Johansson et al. [54] and Zhao et al. [125], the considered η, although achieving both desiderata, is not η^*, and this η renders different optimal representation-level labeling functions on the two domains: $g^*_\eta \neq \tilde{g}^*_\eta$, so the bound is still large. Johansson et al. [54] claim that it is

necessary to require η to be invertible to make $g_\eta^* = \tilde{g}_\eta^*$, and develop a bound (Thm. 2) that explicitly shows the effect of the invertibility of η. The η functions in the examples are not invertible.

(2) For the bound Eq. (46). Applying the bound on the representation space $\mathcal{S}$ gives:

$$\begin{aligned}\mathbb{E}_{\tilde{p}(s,y)}[\ell(g(s),y)] \leqslant \mathbb{E}_{p(s,y)}[\ell(g(s),y)] + d_{\mathcal{G}\Delta\mathcal{G}}(\eta_\#[p_x], \eta_\#[\tilde{p}_x]) \\ + \inf_{g\in\mathcal{G}} \left[\mathbb{E}_{\tilde{p}(s,y)}[\ell(g(s),y)] + \mathbb{E}_{p(s,y)}[\ell(g(s),y)]\right],\end{aligned}$$

where $p_{s,y} := (\eta, \mathrm{id}_y)_\#[p_{x,y}]$ with $\mathrm{id}_y : (x,y) \mapsto y$ and similarly $\tilde{p}_{s,y} := (\eta, \mathrm{id}_y)_\#[\tilde{p}_{x,y}]$. Note that $\mathbb{E}_{p(s,y)}[\ell(g(s),y)] = \mathbb{E}_{p(x,y)}[\ell(g(\eta(x)),y)] = R(g\circ\eta)$ and similarly $\mathbb{E}_{\tilde{p}(s,y)}[\ell(g(s),y)] = \tilde{R}(g\circ\eta)$. So the last term on the r.h.s becomes $\inf_{g\in\mathcal{G}}\left[\tilde{R}(g\circ\eta) + R(g\circ\eta)\right] = \lambda_{\mathcal{G}\circ\eta}$, where $\mathcal{G}\circ\eta := \{g\circ\eta : g\in\mathcal{G}\}$, and the bound then reformulates to:

$$\tilde{R}(g\circ\eta) \leqslant R(g\circ\eta) + d_{\mathcal{G}\Delta\mathcal{G}}(\eta_\#[p_x], \eta_\#[\tilde{p}_x]) + \lambda_{\mathcal{G}\circ\eta}. \tag{48}$$

This result is shown by Johansson et al. [54]. They argue that finding η that achieves both DIR and $R(h^*)$ simultaneously (with some g_η^*) cannot guarantee a tighter bound since the last term $\lambda_{\mathcal{G}\circ\eta}$ may be very large.

In both examples of Johansson et al. [54] and Zhao et al. [125], it holds that $\mathrm{supp}(p_x)\cap\mathrm{supp}(\tilde{p}_x) = \emptyset$. It may cause the problem that $g\circ\eta$ is very different from h^* on $\mathrm{supp}(\tilde{p}_x)$ even when $R(h^*)$ is achieved, since $R(g\circ\eta) = R(h^*)$ only constraints the behavior of $g\circ\eta$ on $\mathrm{supp}(p_x)$. The developed bound by Johansson et al. [54, Thm. 2] also explicitly shows the role of a support overlap, thus is called a support-invertibility bound. They also give an example showing that DIR (particularly implemented by minimizing MMD) is not necessary ("sometimes too strict") for learning the shared/invariant $p(y|x)$.

The problem of DA-DIR is also studied under more modern bounds **(3)** **(4)** and arguments **(5)**.

(3) A third bound. Zhao et al. [125] develop another bound for binary classification $\mathcal{Y} := \{0,1\}$, under the risk function $R(h) := \mathbb{E}_{p(x)}[|h^*(x) - h(x)|]$. The bound is expressed in terms of the JS distance [29] $d_{\mathrm{JS}}(p,q) := \sqrt{\mathrm{JS}(p,q)}$, where $\mathrm{JS}(p,q)$ is the JS divergence, which is bounded: $0 \leqslant \mathrm{JS}(p,q) \leqslant 1$ [16]. It is shown that [125, Lemma 4.8]:

$$d_{\mathrm{JS}}(p_y, \tilde{p}_y) \leqslant d_{\mathrm{JS}}(\eta_\#[p_x], \eta_\#[\tilde{p}_x]) + \sqrt{R(g\circ\eta)} + \sqrt{\tilde{R}(g\circ\eta)}.$$

If $d_{\mathrm{JS}}(p_y, \tilde{p}_y) \geqslant d_{\mathrm{JS}}(\eta_\#[p_x], \eta_\#[\tilde{p}_x])$[17], the bound is given as [125, Thm. 4.3]:

$$R(g\circ\eta) + \tilde{R}(g\circ\eta) \geqslant \frac{1}{2}\left(d_{\mathrm{JS}}(p_y, \tilde{p}_y) - d_{\mathrm{JS}}(\eta_\#[p_x], \eta_\#[\tilde{p}_x])\right)^2, \tag{49}$$

or when the two domains are allowed to have their own representation-level labeling functions g and $\tilde{g}$, we have [125, Corollary 4.1]:

$$R(g\circ\eta) + \tilde{R}(\tilde{g}\circ\eta) \geqslant \frac{1}{2}\left(d_{\mathrm{JS}}(p_y, \tilde{p}_y) - d_{\mathrm{JS}}(\eta_\#[p_x], \eta_\#[\tilde{p}_x])\right)^2. \tag{50}$$

When $p(y) \neq \tilde{p}(y)$, we have $d_{\mathrm{JS}}(p_y, \tilde{p}_y) > 0$, so DIR, which minimizes $d_{\mathrm{JS}}(\eta_\#[p_x], \eta_\#[\tilde{p}_x])$, becomes harmful to minimizing the target risk $\tilde{R}(\tilde{g}\circ\eta)$.

(4) Chuang et al. [23, Thm. 6] probe into the mysterious term $\lambda_{\mathcal{G}\circ\eta}$ in the bound Eq. (48) and show how it is affected by the complexity of $\mathcal{E}$ (the hypothesis space of η):

$$\tilde{R}(g\circ\eta) \leqslant R(g\circ\eta) + d_{\mathcal{G}\Delta\mathcal{G}}(\eta_\#[p_x], \eta_\#[\tilde{p}_x]) + d_{\mathcal{G}_{\mathcal{E}\Delta\mathcal{E}}}(p_x, \tilde{p}_x) + \lambda_{\mathcal{G}\circ\mathcal{E}}(\eta), \tag{51}$$

$$\text{where: } d_{\mathcal{G}_{\mathcal{E}\Delta\mathcal{E}}}(p_x, \tilde{p}_x) := \sup_{g\in\mathcal{G};\eta,\eta'\in\mathcal{E}} \left|\mathbb{E}_{p_x}[\ell(g\circ\eta, g\circ\eta')] - \mathbb{E}_{\tilde{p}_x}[\ell(g\circ\eta, g\circ\eta')]\right|,$$

$$\lambda_{\mathcal{G}\circ\mathcal{E}}(\eta) := \inf_{g'\in\mathcal{G},\eta'\in\mathcal{E}} 2R(g'\circ\eta) + R(g'\circ\eta') + \tilde{R}(g'\circ\eta').$$

[16] This bound is under the unit of bits, *i.e.*, base 2 logarithm is used in the KL divergence defining the JS divergence. Under the unit of nats, *i.e.*, the natural logarithm $\ln$ is used, the bound becomes $0 \leqslant \mathrm{JS}(p,q) \leqslant \ln 2$.

[17] Unfortunately, it seems that the opposite direction of the inequality holds when there exist η^* and g^* (unnecessarily the ones in the strong existence assumption or Assumption 3 of Johansson et al. [54]) such that $p_y = (g^*\circ\eta^*)_\#[p_x]$ and $\tilde{p}_y = (g^*\circ\eta^*)_\#[\tilde{p}_x]$ and that η is a reparameterization of η^*, due to the celebrated data processing inequality.

Here, $d_{\mathcal{G}\Delta\mathcal{G}}(\eta_{\#}[p_x], \eta_{\#}[\tilde{p}_x])$ measures the representation distribution difference, $d_{\mathcal{G}_{\mathcal{E}\Delta\mathcal{E}}}(p_x, \tilde{p}_x)$ measures the complexity of the representation-extractor family $\mathcal{E}$ w.r.t $\mathcal{G}$ [23, Def. 5], and $\lambda_{\mathcal{G}\circ\mathcal{E}}(\eta)$ is "a variant of the best in-class joint risk". For a given $\mathcal{G}$, although a more expressive $\mathcal{E}$ lowers $\lambda_{\mathcal{G}\circ\mathcal{E}}(\eta)$ and contains a more capable η to reduce $d_{\mathcal{G}\Delta\mathcal{G}}(\eta_{\#}[p_x], \eta_{\#}[\tilde{p}_x])$, such an $\mathcal{E}$ also incurs a larger $d_{\mathcal{G}_{\mathcal{E}\Delta\mathcal{E}}}(p_x, \tilde{p}_x)$, so there is a trade-off when choosing a proper $\mathcal{E}$. Chuang et al. [23] illustrate this trade-off by a toy example, and observe this trade-off in experiments. Similarly, there is also a trade-off in the complexity of $\mathcal{G}$ (a more expressive $\mathcal{G}$ lowers $\lambda_{\mathcal{G}\circ\mathcal{E}}(\eta)$ but increases $d_{\mathcal{G}\Delta\mathcal{G}}(\eta_{\#}[p_x], \eta_{\#}[\tilde{p}_x])$ and $d_{\mathcal{G}_{\mathcal{E}\Delta\mathcal{E}}}(p_x, \tilde{p}_x)$), but Chuang et al. [23] find the performance of DA-DIR much less sensitive to it empirically. They also point out the implication of this trade-off in choosing which layer in a neural network as the representation (Prop. 7) with an empirical study.

Chuang et al. [23] also propose a method to estimate the target-domain performance (*i.e.*, the OOD generalization performance) in terms of $\tilde{R}(h)$ of a supervised model h using a set of DA-DIR models $\hat{\mathcal{H}}^*$. The method is supported by its Lemma 4: $\left|\tilde{R}(h) - \sup_{h' \in \hat{\mathcal{H}}^*} \mathbb{E}_{\tilde{p}(x)}[\ell(h(x), h'(x))]\right| \leqslant \sup_{h' \in \hat{\mathcal{H}}^*} \tilde{R}(h')$. The supremum on the l.h.s can be estimated using unsupervised data on the target domain, and it is treated as an estimate to $\tilde{R}(h)$ given that the r.h.s is believed to be small for DA-DIR models $\hat{\mathcal{H}}^*$.

(5) Arjovsky et al. [2] point out that in the covariate shift case $p(y|s) = \tilde{p}(y|s)$, achieving DIR $p(s) = \tilde{p}(s)$ implies $p(y) = \tilde{p}(y)$ (since $p(s)p(y|s) = \tilde{p}(s)\tilde{p}(y|s)$). This may not hold in practice. When it does not hold, the bound Eq. (49) shows that DIR may limit the target-domain performance.

Comparison with CSG The key feature of our CSG is that it is based on causal invariance. In most of the above bounds, including Eqs. (45, 46) for general DA and Eqs. (47, 48, 49, 51) for DA-DIR, the same labeling function h or $g \circ \eta$ is used in both domains (the risks R and $\tilde{R}$ on both domains measure the same h or $g \circ \eta$). So for successful adaptation, covariate shift (invariant h^* or $p(y|x)$) is a basic assumption, which implies inference invariance (invariant η^* or $p(s|x)$) for DA-DIR. Yet, as explained in Sec. 3.2, since the data at hand is produced from a certain mechanism of nature anyway, the invariance in the causal generative direction $p(x|s, v)$ is more fundamental and reliable than covariate shift or inference invariance. The causal invariance allows $p(s) \neq \tilde{p}(s)$ and subsequently a difference in the inference direction: $p(s|x) \neq \tilde{p}(s|x)$ or $\eta^* \neq \tilde{\eta}^*$, and $p(y|x) \neq \tilde{p}(y|x)$ or $h^* \neq \tilde{h}^*$. Following this new philosophy, CSG-ind and CSG-DA use a different inference and prediction rule in the target domain, and Theorems 6 and 7 give OOD prediction guarantees for this different prediction rule. This is in contrast to most existing DA methods and theory.

Another advantage of CSG is that it has an identifiability guarantee (Thm. 5). In the above analyses **(1)** and **(2)**, we see that the problem of DA-DIR arises since achieving both DIR and $R(h^*)$ simultaneously cannot guarantee $\eta = \eta^*$ or $g = g^*$ or $g \circ \eta = h^*$ on $\operatorname{supp}(p_x, \tilde{p}_x)$, even in some sense of semantic or performance equivalence. This is essentially an identifiability problem. CSG achieves identifiability by fitting the entire data distribution $p(x, y)$. In contrast, DA-DIR is not a generative method, and only fits $p(y|x)$. Although DA-DIR also seeks to achieve DIR, it is a weaker goal than fitting $p(x)$ (DIR cannot give $p(x)$). So DA-DIR does not fully exploit the data distribution $p(x, y)$, and identifiability is a problem even with the strong assumption of both covariate shift and the strong existence assumption.

In terms of the considered quantity in the bounds, all the existing ones above bound the objective of the target risk $\tilde{R}(h)$ in terms of the accessible source risk $R(h)$ for an arbitrary labeling function h, while our bound Eq. (36) relates the target risks of the optimally-learned source-domain labeling function h'^* and of the target-domain oracle labeling function $\tilde{h}^*$, *i.e.*, it bounds $|\tilde{R}(h'^*) - \tilde{R}(\tilde{h}^*)|$. It measures the risk gap of the best source labeling function on the target domain. After adaptation, Thm. 7 (Eq. (44)) shows that CSG-DA achieves the optimal labeling function on the target domain.

Under bounds Eqs. (49, 50), we are not minimizing $d_{\mathrm{JS}}(\eta_{\#}[p(x)], \eta_{\#}[\tilde{p}(x)])$, so our method is good under that view. In fact, in CSG the representation distributions on the two domains are $p(s) = \int p(s, v)\,\mathrm{d}v$ and $\tilde{p}(s) = \int \tilde{p}(s, v)\,\mathrm{d}v$ (replacing $\eta_{\#}[p(x)]$ and $\eta_{\#}[\tilde{p}(x)]$). They are generally different and we do not seek to match them.

F Methodology Details

F.1 Derivation of Learning Objectives

F.1.1 The Evidence Lower BOund (ELBO).

A common and effective approach to let the model p match the data distribution $p^*(x, y)$ is maximizing likelihood, that is to maximize $\mathbb{E}_{p^*(x,y)}[\log p(x, y)]$. It is equivalent to minimizing $\mathrm{KL}(p^*(x, y)\|p(x, y))$ (since $\mathbb{E}_{p^*(x,y)}[\log p^*(x, y)]$ is constant of p), so it drives $p(x, y)$ towards $p^*(x, y)$. But the likelihood function $p(x, y) = \int p(s, v, x, y)\,\mathrm{d}s\mathrm{d}v$ involves an intractable integration, which is hard to estimate and optimize. To address this, the popular method of *variational expectation-maximization* (variational EM) introduces a tractable (has closed-form density function and easy to draw samples from it) distribution $q(s, v|x, y)$ of the latent variables given observed variables, and a lower bound of the likelihood function can be derived:

$$\begin{aligned}\log p(x, y) =\ & \log \mathbb{E}_{p(s,v)}[p(s, v, x, y)] = \log \mathbb{E}_{q(s,v|x,y)}\left[\frac{p(s, v, x, y)}{q(s, v|x, y)}\right] \\ \geqslant\ & \mathbb{E}_{q(s,v|x,y)}\left[\log \frac{p(s, v, x, y)}{q(s, v|x, y)}\right] =: \mathcal{L}_{p,\, q_{s,v|x,y}}(x, y),\end{aligned} \tag{52}$$

where the inequality follows Jensen's inequality and the concavity of the $\log$ function. The function $\mathcal{L}_{p,\, q_{s,v|x,y}}(x, y)$ is thus called *Evidence Lower BOund* (ELBO). The tractable distribution $q(s, v|x, y)$ is called *variational distribution*, and is commonly instantiated by a standalone model (from the generative model) called an *inference model*. Moreover, we have:

$$\begin{aligned}&\mathcal{L}_{p,\, q_{s,v|s,y}}(x, y) + \mathrm{KL}(q(s, v|x, y)\|p(s, v|x, y)) \\ =\ & \mathbb{E}_{q(s,v|x,y)}\left[\log \frac{p(s, v, x, y)}{q(s, v|x, y)}\right] + \mathbb{E}_{q(s,v|x,y)}\left[\log \frac{q(s, v|x, y)}{p(s, v|x, y)}\right] \\ =\ & \mathbb{E}_{q(s,v|x,y)}\left[\log \frac{p(s, v, x, y)}{p(s, v|x, y)}\right] = \mathbb{E}_{q(s,v|x,y)}[\log p(x, y)] \\ =\ & \log p(x, y),\end{aligned}$$

so maximizing $\mathcal{L}_{p,\, q_{s,v|x,y}}(x, y)$ w.r.t $q(s, v|x, y)$ is equivalent to minimizing $\mathrm{KL}(q(s, v|x, y)\|p(s, v|x, y))$ (since the r.h.s $\log p(x, y)$ is constant of $q(s, v|x, y)$), which drives $q(s, v|x, y)$ towards the true posterior (*i.e.*, the goal of *variational inference*), and once this is (perfectly) done, $\mathcal{L}_{p,\, q_{s,v|x,y}}(x, y)$ becomes a lower bound of $\log p(x, y)$ that is tight at the current model p, so maximizing $\mathcal{L}_{p,\, q_{s,v|x,y}}(x, y)$ w.r.t p effectively maximizes $\log p(x, y)$ (*i.e.*, the goal of maximizing likelihood). So the training objective becomes the expected ELBO $\mathbb{E}_{p^*(x,y)}[\mathcal{L}_{p,\, q_{s,v|x,y}}(x, y)]$. Optimizing it w.r.t $q(s, v|x, y)$ and p alternately drives $q(s, v|x, y)$ towards $p(s, v|x, y)$ and $p(x, y)$ towards $p^*(x, y)$ eventually. The derivations and conclusions above hold for general latent variable models, with (s, v) representing the latent variables, and (x, y) observed variables (data variables).

This standard form of ELBO gives the objective for fitting unsupervised test-domain data from the underlying data distribution $\tilde{p}^*(x)$. In this case, the observed variable is only x while the latent variable is still (s, v), so the required joint distribution for latent and observed variables is $\tilde{p}(s, v, x) = \tilde{p}(s, v)p(x|s, v)$, and the inference model is in the form $\tilde{q}(s, v|x)$. Following the form of Eq. (52), the ELBO objective for fitting $\tilde{p}^*(x)$ (*i.e.*, the lower bound for $\log \tilde{p}(x)$) is:

$$\mathcal{L}_{\tilde{p},\, \tilde{q}_{s,v|x}}(x) = \mathbb{E}_{\tilde{q}(s,v|x)}\Big[\log \frac{\tilde{p}(s, v, x)}{\tilde{q}(s, v|x)}\Big].$$

This leads to Eq. (4).

F.1.2 Variational EM for learning CSG.

In the supervised case, the expected ELBO objective $\mathbb{E}_{p^*(x,y)}[\mathcal{L}_{p,\, q_{s,v|x,y}}(x, y)]$ can also be understood as the conventional supervised learning loss, *i.e.* the cross entropy, regularized by a generative reconstruction term. As explained in the main text (Sec. 4), after training, we only have the model $p(s, v, x, y)$ and an approximation $q(s, v|x, y)$ to the posterior $p(s, v|x, y)$, and prediction using $p(y|x)$ is still intractable. So we employ a tractable distribution $q(s, v, y|x)$ to model the required variational distribution as $q(s, v|x, y) = q(s, v, y|x)/q(y|x)$, where $q(y|x) = \int q(s, v, y|x)\,\mathrm{d}s\mathrm{d}v$

is the derived marginal distribution of y from $q(s,v,y|x)$ (we will show that it can be effectively estimated and sampled from). With this instantiation, the expected ELBO becomes:

$$
\begin{aligned}
&\mathbb{E}_{p^*(x,y)}[\mathcal{L}_{p,\, q_{s,v|x,y}=\cdots(q_{s,v,y|x})}(x,y)] \\
=& \int p^*(x,y)\frac{q(s,v,y|x)}{q(y|x)}\log\frac{p(s,v,x,y)q(y|x)}{q(s,v,y|x)}\,\mathrm{d}s\mathrm{d}v\mathrm{d}x\mathrm{d}y \\
=& \int p^*(x,y)\frac{q(s,v,y|x)}{q(y|x)}\log q(y|x)\,\mathrm{d}s\mathrm{d}v\mathrm{d}x\mathrm{d}y + \int p^*(x,y)\frac{q(s,v,y|x)}{q(y|x)}\log\frac{p(s,v,x,y)}{q(s,v,y|x)}\,\mathrm{d}s\mathrm{d}v\mathrm{d}x\mathrm{d}y \\
=& \int p^*(x)\bigg(\int p^*(y|x)\frac{\int q(s,v,y|x)\,\mathrm{d}s\mathrm{d}v}{q(y|x)}\log q(y|x)\,\mathrm{d}y\bigg)\mathrm{d}x \\
&+ \int p^*(x)\bigg(\int \frac{p^*(y|x)}{q(y|x)}q(s,v,y|x)\log\frac{p(s,v,x,y)}{q(s,v,y|x)}\,\mathrm{d}s\mathrm{d}v\mathrm{d}y\bigg)\mathrm{d}x \\
=& \mathbb{E}_{p^*(x)}\mathbb{E}_{p^*(y|x)}[\log q(y|x)] + \mathbb{E}_{p^*(x)}\mathbb{E}_{q(s,v,y|x)}\left[\frac{p^*(y|x)}{q(y|x)}\log\frac{p(s,v,x,y)}{q(s,v,y|x)}\right],
\end{aligned}
$$

which is Eq. (1). Here, we use the shorthand "$q_{s,v|x,y} = \cdots(q_{s,v,y|x})$" for the above substitution $q(s,v|x,y) = q(s,v,y|x)/\int q(s,v,y|x)\,\mathrm{d}s\mathrm{d}v$ and highlight the argument therein. The first term is the (negative) expected cross entropy loss, which drives the inference model (predictor) $q(y|x)$ towards $p^*(y|x)$ for $p^*(x)$-a.e. x. Once this is (perfectly) done, the second term becomes $\mathbb{E}_{p^*(x)}\mathbb{E}_{q(s,v,y|x)}[\log\big(p(s,v,x,y)/q(s,v,y|x)\big)]$, which is the expected ELBO $\mathbb{E}_{p^*(x)}[\mathcal{L}_{p,\, q_{s,v,y|x}}(x,y)]$ for $q(s,v,y|x)$. It thus drives $q(s,v,y|x)$ towards $p(s,v,y|x)$ and $p(x)$ towards $p^*(x)$. It accounts for a regularization by fitting the input distribution $p^*(x)$ and align the inference model (predictor) with the generative model.

The target of $q(s,v,y|x)$, *i.e.* $p(s,v,y|x)$, adopts a factorization $p(s,v,y|x) = p(s,v|x)p(y|s)$ due to the graphical structure (Fig. 1a) of CSG (*i.e.*, $y \perp\!\!\!\perp (x,v) \mid s$). The factor $p(y|s)$ is known (the invariant causal mechanism to generate y in CSG), so we only need to employ an inference model $q(s,v|x)$ for the intractable factor $p(s,v|x)$, so $q(s,v,y|x) = q(s,v|x)p(y|s)$. Using this relation, we can reformulate Eq. (1) as:

$$
\begin{aligned}
&\mathbb{E}_{p^*(x,y)}[\mathcal{L}_{p,\, q_{s,v|x,y}=\cdots(q_{s,v|x},p_{y|s})}(x,y)] \\
=& \mathbb{E}_{p^*(x,y)}[\log q(y|x)] + \mathbb{E}_{p^*(x)}\left[\int q(s,v|x)p(y|s)\frac{p^*(y|x)}{q(y|x)}\log\frac{p(s,v,x)}{q(s,v|x)}\,\mathrm{d}s\mathrm{d}v\mathrm{d}y\right] \\
=& \mathbb{E}_{p^*(x,y)}[\log q(y|x)] + \mathbb{E}_{p^*(x)}\left[\int \frac{p^*(y|x)}{q(y|x)}\bigg(\int q(s,v|x)p(y|s)\log\frac{p(s,v,x)}{q(s,v|x)}\,\mathrm{d}s\mathrm{d}v\bigg)\mathrm{d}y\right] \\
=& \mathbb{E}_{p^*(x,y)}[\log q(y|x)] + \mathbb{E}_{p^*(x,y)}\left[\frac{1}{q(y|x)}\mathbb{E}_{q(s,v|x)}\Big[p(y|s)\log\frac{p(s,v,x)}{q(s,v|x)}\Big]\right],
\end{aligned}
\tag{53}
$$

which is Eq. (2). We used the shorthand "$q_{s,v|x,y} = \cdots(q_{s,v|x}, p_{y|s})$" for the substitution for $q(s,v|x,y)$ using $q(s,v|x)$ and $p(y|s)$. With this form of $q(s,v,y|x) = q(s,v|x)p(y|s)$, we have $q(y|x) = \mathbb{E}_{q(s,v|x)}[p(y|s)]$ which can also be estimated and optimized using reparameterization. For prediction, we can sample from the approximation $q(y|x)$ instead of the intractable $p(y|x)$. This can be done by ancestral sampling: first sample (s,v) from $q(s,v|x)$, and then use the sampled s to sample y from $p(y|s)$.

F.1.3 Variational EM for learning CSG with test-domain inference model (Learning CSG-ind and CSG-DA on the training domain).

See the main text in Sec. 4.1 and Sec. 4.2 for motivations and the basic idea of the methods. Methods for CSG-ind and CSG-DA are similar, so we mainly show the detailed derivation for CSG-ind.

Since the prior is the only difference between $p(s,v,x,y)$ and $p^{\perp\!\!\!\perp}(s,v,x,y)$, we have $\frac{p(s,v,x,y)}{p^{\perp\!\!\!\perp}(s,v,x,y)} = \frac{p(s,v)}{p^{\perp\!\!\!\perp}(s,v)}$. So $p(s,v,y|x) = \frac{p(s,v)}{p^{\perp\!\!\!\perp}(s,v)}\frac{p^{\perp\!\!\!\perp}(x)}{p(x)}p^{\perp\!\!\!\perp}(s,v,y|x)$. As explained, inference models now only need to approximate the posterior $(s,v) \mid x$. Since $p(s,v,y|x) = p(s,v|x)p(y|s)$ and $p^{\perp\!\!\!\perp}(s,v,y|x) = p^{\perp\!\!\!\perp}(s,v|x)p(y|s)$ share the same $p(y|s)$ factor, we have $p(s,v|x) = \frac{p(s,v)}{p^{\perp\!\!\!\perp}(s,v)}\frac{p^{\perp\!\!\!\perp}(x)}{p(x)}p^{\perp\!\!\!\perp}(s,v|x)$. The variational distributions $q(s,v|x)$ and $q^{\perp\!\!\!\perp}(s,v|x)$ target $p(s,v|x)$ and $p^{\perp\!\!\!\perp}(s,v|x)$ respectively, so we

can express the former with the latter:

$$q(s,v|x) = \frac{p(s,v)}{p^{\perp\!\!\!\perp}(s,v)}\frac{p^{\perp\!\!\!\perp}(x)}{p(x)}q^{\perp\!\!\!\perp}(s,v|x). \tag{54}$$

Once $q^{\perp\!\!\!\perp}(s,v|x)$ achieves its goal, such represented $q(s,v|x)$ also does so. So we only need to construct an inference model for $q^{\perp\!\!\!\perp}(s,v|x)$ and optimize it. With this representation, we have:

$$\begin{aligned} q(y|x) &= \mathbb{E}_{q(s,v|x)}[p(y|s)] = \mathbb{E}_{q^{\perp\!\!\!\perp}(s,v|x)}\left[\frac{p(s,v)}{p^{\perp\!\!\!\perp}(s,v)}\frac{p^{\perp\!\!\!\perp}(x)}{p(x)}p(y|s)\right] = \frac{p^{\perp\!\!\!\perp}(x)}{p(x)}\mathbb{E}_{q^{\perp\!\!\!\perp}(s,v|x)}\left[\frac{p(s,v)}{p^{\perp\!\!\!\perp}(s,v)}p(y|s)\right] \\ &= \frac{p^{\perp\!\!\!\perp}(x)}{p(x)}\pi(y|x), \end{aligned} \tag{55}$$

where $\pi(y|x) := \mathbb{E}_{q^{\perp\!\!\!\perp}(s,v|x)}\left[\frac{p(s,v)}{p^{\perp\!\!\!\perp}(s,v)}p(y|s)\right]$ as in the main text, which can be estimated and optimized using the reparameterization of $q^{\perp\!\!\!\perp}(s,v|x)$. From Eq. (2), the expected ELBO training objective can be reformulated as:

$$\begin{aligned} &\mathbb{E}_{p^*(x,y)}[\mathcal{L}_{p,\, q_{s,v|x,y}=\cdots(q^{\perp\!\!\!\perp}_{s,v|x},p)}(x,y)] \\ &= \mathbb{E}_{p^*(x,y)}\left[\log q(y|x) + \frac{1}{q(y|x)}\mathbb{E}_{q(s,v|x)}\left[p(y|s)\log\frac{p(s,v,x)}{q(s,v|x)}\right]\right] \\ &= \mathbb{E}_{p^*(x,y)}\Bigg[\log\frac{p^{\perp\!\!\!\perp}(x)}{p(x)} + \log\pi(y|x) \\ &\quad + \frac{p(x)}{p^{\perp\!\!\!\perp}(x)}\frac{1}{\pi(y|x)}\mathbb{E}_{q^{\perp\!\!\!\perp}(s,v|x)}\left[\frac{p(s,v)}{p^{\perp\!\!\!\perp}(s,v)}\frac{p^{\perp\!\!\!\perp}(x)}{p(x)}p(y|s)\log\frac{p(s,v)p(x|s,v)}{\frac{p(s,v)}{p^{\perp\!\!\!\perp}(s,v)}\frac{p^{\perp\!\!\!\perp}(x)}{p(x)}q^{\perp\!\!\!\perp}(s,v|x)}\right]\Bigg] \\ &= \mathbb{E}_{p^*(x,y)}\Bigg[\log\frac{p^{\perp\!\!\!\perp}(x)}{p(x)} + \log\pi(y|x) \\ &\quad + \frac{1}{\pi(y|x)}\mathbb{E}_{q^{\perp\!\!\!\perp}(s,v|x)}\left[\frac{p(s,v)}{p^{\perp\!\!\!\perp}(s,v)}p(y|s)\Big(\log\frac{p(x)}{p^{\perp\!\!\!\perp}(x)} + \log\frac{p^{\perp\!\!\!\perp}(s,v)p(x|s,v)}{q^{\perp\!\!\!\perp}(s,v|x)}\Big)\right]\Bigg] \\ &= \mathbb{E}_{p^*(x,y)}\left[\log\frac{p^{\perp\!\!\!\perp}(x)}{p(x)} + \log\pi(y|x) + \frac{1}{\pi(y|x)}\mathbb{E}_{q^{\perp\!\!\!\perp}(s,v|x)}\left[\frac{p(s,v)}{p^{\perp\!\!\!\perp}(s,v)}p(y|s)\right]\log\frac{p(x)}{p^{\perp\!\!\!\perp}(x)}\right. \\ &\quad \left. + \frac{1}{\pi(y|x)}\mathbb{E}_{q^{\perp\!\!\!\perp}(s,v|x)}\left[\frac{p(s,v)}{p^{\perp\!\!\!\perp}(s,v)}p(y|s)\log\frac{p^{\perp\!\!\!\perp}(s,v)p(x|s,v)}{q^{\perp\!\!\!\perp}(s,v|x)}\right]\right] \\ &= \mathbb{E}_{p^*(x,y)}\left[\log\frac{p^{\perp\!\!\!\perp}(x)}{p(x)} + \log\pi(y|x) + \frac{1}{\pi(y|x)}\pi(y|x)\log\frac{p(x)}{p^{\perp\!\!\!\perp}(x)}\right. \\ &\quad \left. + \frac{1}{\pi(y|x)}\mathbb{E}_{q^{\perp\!\!\!\perp}(s,v|x)}\left[\frac{p(s,v)}{p^{\perp\!\!\!\perp}(s,v)}p(y|s)\log\frac{p^{\perp\!\!\!\perp}(s,v,x)}{q^{\perp\!\!\!\perp}(s,v|x)}\right]\right] \\ &= \mathbb{E}_{p^*(x,y)}\left[\log\pi(y|x) + \frac{1}{\pi(y|x)}\mathbb{E}_{q^{\perp\!\!\!\perp}(s,v|x)}\left[\frac{p(s,v)}{p^{\perp\!\!\!\perp}(s,v)}p(y|s)\log\frac{p^{\perp\!\!\!\perp}(s,v,x)}{q^{\perp\!\!\!\perp}(s,v|x)}\right]\right], \end{aligned} \tag{56}$$

where in the second-last equality we have used the definition of $\pi(y|x)$. The shorthand "$q_{s,v|x,y} = \cdots(q^{\perp\!\!\!\perp}_{s,v|x}, p)$" represents the substitution using $q^{\perp\!\!\!\perp}(s,v|x)$ and $p = \langle p(s,v), p(x|s,v), p(y|s)\rangle$ for $q(s,v|x,y) = q(s,v|x)p(y|s)/\int q(s,v|x)p(y|s)\,\mathrm{d}s\mathrm{d}v$ where $q(s,v|x)$ is determined by $q^{\perp\!\!\!\perp}(s,v|x)$ and p via Eq. (54) (recall that $p^{\perp\!\!\!\perp}(s,v)$ is determined by $p(s,v)$, so $p^{\perp\!\!\!\perp}(x)$ is also determined by $p(s,v)$ and $p(x|s,v)$). This Eq. (56) gives Eq. (3) for CSG-ind. Note that $\pi(y|x)$ is not used in prediction, so there is no need to sample from it. Prediction is done by ancestral sampling from $q^{\perp\!\!\!\perp}(y|x)$, that is to first sample from $q^{\perp\!\!\!\perp}(s,v|x)$ and then from $p(y|s)$. Using this reformulation, we can train a CSG with independent prior even on data that manifests a correlated prior.

For CSG-DA, we only need to replace the independent prior $p^{\perp\!\!\!\perp}(s,v)$ hypothesized for the test domain with the standalone prior model $\tilde{p}(s,v)$ dedicated to learning the test-domain prior, and re-denote the test-domain inference model $q^{\perp\!\!\!\perp}(s,v|x)$ with $\tilde{q}(s,v|x)$. By doing so, Eq. (56) gives Eq. (5), *i.e.* the objective for CSG-DA on the training domain. For numerical stability, we employ the `log-sum-exp` trick to estimate the expectations and compute the gradients.

F.1.4 Methods for CSGz for ablation study.

The conclusions and methods can also be applied to general latent-variable generative models for supervised learning, by replacing (s, v) with their latent variables. Particularly, the method also applies to the counterpart of CSG in the ablation study experiment, which does not distinguish the two latent factors s and v and treats them as a united latent variable $z = (s, v)$. We thus call it **CSGz**. The essential difference from CSG is that CSGz keeps the $v \to y$ arrow, which is unlikely a causal relation as we argued in Sec. 3, item **(4)**. Formally, a CSGz model is defined as the tuple $p := \langle p(z), p(x|z), p(y|z) \rangle$, and the corresponding inference model is in the form $q(z|x)$.

Following a similar derivation of Eq. (53), we have the objective for fitting training-domain data:

$$\begin{aligned}&\mathbb{E}_{p^*(x,y)}[\mathcal{L}_{p,\, q_{z|x,y}=\cdots(q_{z|x},p_{y|z})}(x,y)]\\ &= \mathbb{E}_{p^*(x,y)}[\log q(y|x)] + \mathbb{E}_{p^*(x,y)}\left[\frac{1}{q(y|x)}\mathbb{E}_{q(z|x)}\left[p(y|z)\log\frac{p(z,x)}{q(z|x)}\right]\right],\end{aligned}$$

where $q(y|x) = \mathbb{E}_{q(z|x)}[p(y|z)]$. The shorthand "$q_{z|x,y} = \cdots(q_{z|x}, p_{y|z})$" is similarly for the substitution $q(z|x,y) = q(z|x)p(y|z)/\int q(z|x)p(y|z)\,\mathrm{d}z$ using $q(z|x)$ and $p(y|z)$.

As CSGz does not consider the distinction between s and v, there is no CSGz-ind version. The CSGz-DA version for domain adaptation is possible by using a standalone prior model $\tilde{p}(z)$ for the test domain, which is learned by optimizing the corresponding ELBO objective similar to Eq. (4):

$$\max_{\tilde{p},\, \tilde{q}_{z|x}} \mathbb{E}_{\tilde{p}^*(x)}[\mathcal{L}_{\tilde{p},\, \tilde{q}_{z|x}}(x)], \text{ where } \mathcal{L}_{\tilde{p},\, \tilde{q}_{z|x}}(x) = \mathbb{E}_{\tilde{q}(z|x)}\left[\log\frac{\tilde{p}(z)p(x|z)}{\tilde{q}(z|x)}\right].$$

To fit training-domain data using the test-domain inference model $\tilde{q}(z|x)$, following a similar derivation of Eq. (56), we have the objective on the training domain for CSG-DA:

$$\max_{p,\, \tilde{q}_{z|x}} \mathbb{E}_{p^*(x,y)}\left[\log\pi(y|x) + \frac{1}{\pi(y|x)}\mathbb{E}_{\tilde{q}(z|x)}\left[\frac{p(z)}{\tilde{p}(z)}p(y|z)\log\frac{\tilde{p}(z)p(x|z)}{\tilde{q}(z|x)}\right]\right],$$

where $\pi(y|x) := \mathbb{E}_{\tilde{q}(z|x)}\left[\frac{p(z)}{\tilde{p}(z)}p(y|z)\right]$.

F.2 Instantiating the Inference Model

Although motivated from learning a generative model, the method can be implemented using a general discriminative model (with hidden nodes) with causal behavior. By parsing some of the hidden nodes as s and some others as v, a discriminative model could formalize a distribution $q(s, v, y|x)$, which implements the inference model and the generative mechanism $p(y|s)$. The parsing mode is shown in Fig. 3, which is based on the following consideration.

(1) The graphical structure of CSG in Fig. 1a indicates that $(v, x) \perp\!\!\!\perp y \mid s$, so the hidden nodes for s should isolate y from v and x. The model then factorizes the distribution as $q(s, v, y|x) = q(s, v|x)q(y|s)$, and since the inference and generative models share the distribution on $y|s$ (see the main text for explanation), we can thus use the component $q(y|s)$ given by the discriminative model to implement the generative mechanism $p(y|s)$.

(2) The graphical structure in Fig. 1a also indicates that $s \not\perp\!\!\!\perp v \mid x$ due to the v-structure (collider) at x ("explain away"). The component $q(s, v|x)$ should embody this dependence, so the hidden nodes chosen as v should have an effect on those as s. Note that the arrows in Fig. 3 represent computation directions but not causal directions. We orient the computation direction $v \to s$ since all hidden nodes in a discriminative model eventually contribute to computing y.

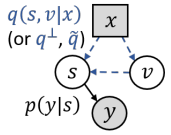


Figure 3: Parsing a general discriminative model as an inference model for CSG. The black solid arrow constructs $p(y|s)$ in the generative model, and the blue dashed arrows (representing computational but not causal directions) construct $q(s, v|x)$ (or $q^{\perp\!\!\!\perp}(s, v|x)$ or $\tilde{q}(s, v|x)$) as the inference model.

After parsing, the discriminative model gives a mapping $(s, v) = \eta(x)$. We implement the distribution by[18] $q(s, v|x) = \mathcal{N}(s, v|\eta(x), \Sigma_q)$. For all the three cases of CSG, CSG-ind and CSG-DA, only one inference model for $(s, v) \mid x$ is required. The component $(s, v) \mid x$ of the discriminative model thus parameterizes $q^{\perp\!\!\!\perp}(s, v|x)$ and $\tilde{q}(s, v|x)$ for CSG-ind and CSG-DA. The expectations in all objectives (except for expectations over p^* which are estimated by averaging over data) are all under the respective $(s, v) \mid x$. They can be estimated using $\eta(x)$ by the reparameterization trick [62], and the gradients can be back-propagated.

We need two more components beyond the discriminative model to implement the method, *i.e.* the prior $p(s, v)$ and the generative mechanism $p(x|s, v)$. The latter can be implemented using a generator or decoder architecture comparable to the component $q(s, v|x)$. The prior can be commonly implemented using a multivariate Gaussian distribution, $p(s, v) = \mathcal{N}(\binom{s}{v}|\binom{\mu_s}{\mu_v}, \Sigma = \left(\begin{smallmatrix}\Sigma_{ss} & \Sigma_{sv}\\ \Sigma_{vs} & \Sigma_{vv}\end{smallmatrix}\right))$. In implementation, the means μ_s and μ_v are fixed as zero vectors. We parameterize Σ via its Cholesky decomposition, $\Sigma = LL^\top$, where L is a lower-triangular matrix with positive diagonals, which is in turn parameterized as $L = \left(\begin{smallmatrix}L_{ss} & 0\\ M_{vs} & L_{vv}\end{smallmatrix}\right)$ with smaller lower-triangular matrices L_{ss} and L_{vv} and any matrix M_{vs}. Matrices L_{ss} and L_{vv} are parameterized by a summation of positive diagonals (guaranteed via an exponential map) and a lower-triangular (excluding diagonals) matrix. Training CSG-ind via Eq. (3) requires estimating the ratio $\frac{p(s,v)}{p^{\perp\!\!\!\perp}(s,v)} = \frac{p(s,v)}{p(s)p(v)} = \frac{p(v|s)}{p(v)}$, where $p(v) = \mathcal{N}(v|\mu_v, \Sigma_{vv})$ with $\Sigma_{vv} = L_{vv}L_{vv}^\top + M_{vs}M_{vs}^\top$, and the conditional distribution $p(v|s)$ is given by $p(v|s) = \mathcal{N}(v|\mu_{v|s}, \Sigma_{v|s})$ with $\mu_{v|s} = \mu_v + M_{vs}L_{ss}^{-1}(s - \mu_s)$, $\Sigma_{v|s} = L_{vv}L_{vv}^\top$ (see *e.g.*, Bishop [14]). This prior does not imply a causal direction between s and v (the linear Gaussian case of Zhang and Hyvärinen [122]) thus well serves as a prior for CSG.

F.3 Model Selection Details

We use a validation set on the training domain for hyperparameter selection, to avoid overfitting due to the finiteness of training data samples, and to guarantee a good fit to the training-domain data distribution $p^*(x, y)$ as the semantic-identifiability theorem 5 recommends. We note that model selection in OOD prediction tasks is itself controversial and nontrivial, and it is still an active research direction [120, 39]. It is argued that if a validation set from the test domain is available, the OOD setup that there is no supervision on the test domain is violated, and then a better choice would be to incorporate it in learning as the semi-supervised adaptation task, instead of using it just for validation. As our methods are designed to fit the training domain data and our theory shows guarantees under a good fit to the training-domain data distribution, model selection using a training-domain validation set is reasonable. This does not contradict the trade-off between training- and test-domain accuracies shown in some prior works (*e.g.*, [95]), since they consider arbitrary distribution change, and using the same prediction rule in both domains, while we leverage causal invariance and develop a different prediction rule in the test domain. In implementation, the training and validation sets are constructed by a 80%-20% random split of all training-domain data in each task.

More specifically, for hyperparameter selection, we align the scale of the supervision loss terms ($\mathbb{E}_{p^*(x,y)}[\log \pi(y|x)]$ for CSG-ind/-DA and CSGz-DA, and the CE loss term for others) in the objectives of all methods, and tune the coefficients of the ELBOs to be their largest values that make the accuracy near 1 on the validation set, so that they wield the most power on the test domain while being faithful to explicit supervision. The coefficients are preferred to be large to well fit $p^*(x)$ (and $\tilde{p}^*(x)$ for domain adaptation) to gain generalizability in the test domain, while they should not affect training accuracy, which is required for a good fit to the training distribution.

For CSG-ind/-DA and CSGz-DA, since their inference models target the test domain, it is not reasonable to evaluate validation accuracy directly using them in the form of $\mathbb{E}_{q^{\text{test}}(s,v|x)}[p(y|s)]$ (q^{test} here refers to $q^{\perp\!\!\!\perp}$ or $\tilde{q}$). Instead, Eq. (55) shows that $\pi(y|x) := \mathbb{E}_{q^{\text{test}}(s,v|x)}[\frac{p(s,v)}{p^{\text{test}}(s,v)}p(y|s)]$ ($p^{\text{test}}(s, v)$ refers to $p^{\perp\!\!\!\perp}(s, v)$ or $\tilde{p}(s, v)$) is an unnormalized density of $q(y|x)$, the training-domain predictor. So we evaluate $\pi(y|x)$ for every value of y (which is not too large for classification tasks) and normalize them for the validation accuracy.

[18] Other approaches to introducing randomness are also possible, such as employing stochasticity on the parameters/weights as in Bayesian neural networks [82], or using dropout [106, 32]. Here we adopt this simple treatment to highlight the main contribution.

Compared with recent model selection methods [120, 119], our method does not introduce additional hyperparameters or assumptions, and does not require multiple training domains. These advantages stem from the explicit description of domain change of our CSG model based on the causal invariance principle 2.

G Experiment Details

The CSGz baseline for ablation study. To show the benefit of modeling s and v separately, we consider a counterpart of CSG that does not separate its latent variable z into s and v; or equivalently, it does not remove the edge $v \to y$. This means that all its latent variables in z directly (*i.e.*, not mediated by s) affect the output y. We thus call it CSGz. Detailed methods for OOD generalization (CSGz; note it does not have a "-ind" version) and domain adaptation (CSGz-DA) are introduced in Appx. F.1.4. To align the model architecture for fair comparison, this means that the latent variable z of CSGz can only be taken as the latent variable s in CSG (see Appx. F.2, Fig. 3).

More about the baselines. The CSGz(-DA) baselines are implemented in our codebase along with the proposed CSG(-ind/-DA) methods. The CNBB method [41] as an OOD generalization baseline is also implemented, based on the description in the paper. For domain adaptation baseline methods DANN [33], DAN [73], CDAN [74] and MDD [124], we use their implementation in the `dalib` package[19] [53]. The BNM method [25] is integrated into our codebase based on its official implementation[20]. Results of CE, DANN, DAN and CDAN are taken from [74] for the ImageCLEF-DA dataset and from [39] except DAN for the PACS and VLCS datasets. All methods share the same optimization setup.

Note that we do not consider domain generalization baselines (*e.g.*, invariant risk minimization [2]) as they degenerate to the CE baseline (*i.e.*, the standard supervised learning method, or empirical risk minimization) when given only one training domain.

Computation infrastructure. Each run of the experiment is on a single Tesla P100 GPU. All the experiments are implemented in PyTorch [84].

More analysis on the results. Complete results including the MDD, CSGz and CSGz-DA baselines, as well as the VLCS [30] dataset, are shown in Table 2 for OOD generalization and in Table 3 for domain adaptation. The complete results support the same conclusions in the main text.

In addition, for the **ablation study**, we observe that our CSG methods outperform CSGz methods in all tasks, demonstrating the benefit of modeling the semantic and variation factors separately. Also, CSGz methods usually have a larger variance, possibly due to the lack of semantic-identifiability so the learned representation gets misled by the variation factor more or less from run to run. On the other hand, CSGz methods still outperform existing methods most of the time, which are discriminative methods. This shows the advantage of using a *generative model*: the invariance of generative mechanisms (causal invariance) is more reliable.

From the domain adaptation results in Table 3, we note that the advantage of CSG-DA on ImageCLEF-DA is not as significant as on other datasets (shifted-MNIST, PACS, VLCS); existing methods CDAN and BNM achieve a comparable or sometimes better result than CSG-DA on ImageCLEF-DA. This reveals the **suitable problem** that our CSG methods solve the best, as discussed in the main text. We expand the analysis below.

Generally speaking, most domain adaptation methods are designed to extract prediction-informative features that are also common across domains, but at the risk to end up with such a feature that leverages a spurious correlation and misleads prediction. In contrast, our CSG methods can be seen to filter out misleading candidates of such features, but with the requirement for identifiability that the training domain shows a diverse v for each s. This requirement comes from the bounded prior condition in the identifiability theorem 5, or the intuition to reduce the risk of extreme cases (Thm. 5 Remark **(1)**).

For the ImageCLEF-DA task, there is no severe spurious correlation, since the style factor as v has no preference on a particular class in any domain. So existing domain adaptation methods do not

[19] `https://github.com/thuml/Transfer-Learning-Library`

[20] `https://github.com/cuishuhao/BNM`

Table 2: Test accuracy (%) for **OOD generalization** by various methods (ours in bold and line separated; CSGz baseline included) on **Shifted-MNIST** (top two rows), **ImageCLEF-DA** (mid-top four rows), **PACS** (mid-bottom four rows) and **VLCS** (bottom four rows) datasets. Results of CE are taken from [74] for ImageCLEF-DA and from [39] for PACS and VLCS. Averaged over 10 runs.

task		CE	CNBB	CSGz	**CSG**	**CSG-ind**
Shifted-MNIST	$\delta_0 = \delta_1 = 0$	42.9±3.1	54.7±3.3	53.0±6.7	81.4±7.4	**82.6±4.0**
	$\delta_0, \delta_1 \sim \mathcal{N}(0,2^2)$	47.8±1.5	59.2±2.4	54.8±5.6	61.7±3.6	**62.3±2.2**
Image CLEF-DA	**C→P**	65.5±0.3	72.7±1.1	73.3±1.0	73.6±0.6	**74.0±1.3**
	P→C	91.2±0.3	91.7±0.2	91.6±0.9	92.3±0.4	**92.7±0.2**
	I→P	74.8±0.3	75.4±0.6	77.0±0.2	76.9±0.3	**77.2±0.2**
	P→I	83.9±0.1	88.7±0.5	90.4±0.3	90.4±0.3	**90.9±0.2**
PACS	others→**P**	**97.8±0.0**	96.9±0.2	97.7±0.3	97.7±0.2	**97.8±0.2**
	others→**A**	88.1±0.1	73.1±0.3	87.3±0.8	**88.5±0.6**	**88.6±0.6**
	others→**C**	77.9±1.3	50.2±1.2	84.3±0.9	84.4±0.9	**84.6±0.8**
	others→**S**	79.1±0.9	43.3±1.2	80.6±1.4	80.7±1.0	**81.1±1.2**
VLCS	others→**V**	76.4±1.5	75.5±0.9	79.4±1.0	79.3±1.1	**80.0±0.9**
	others→**L**	63.3±0.9	61.1±1.2	69.6±0.8	69.6±0.5	**70.1±0.8**
	others→**C**	97.6±1.0	97.1±0.4	99.2±0.3	**99.4±0.3**	**99.5±0.2**
	others→**S**	72.2±0.5	73.7±0.6	75.0±0.9	76.1±1.3	**76.9±1.2**

Table 3: Test accuracy (%) for **domain adaptation** by various methods (ours in bold and line separated; BNM and CSGz-DA baselines included) on **Shifted-MNIST** (top two rows), **ImageCLEF-DA** (mid-top four rows), **PACS** (mid-bottom four rows) and **VLCS** (bottom four rows) datasets. Results of DANN, DAN and CDAN on ImageCLEF-DA are taken from [74], and results of DANN and CDAN on PACS and VLCS are taken from [39]. Averaged over 10 runs.

task		DANN	DAN	CDAN	MDD	BNM	CSGz-DA	**CSG-DA**
Shifted-MNIST	$\delta_0 = \delta_1 = 0$	40.9±3.0	40.4±2.0	41.0±0.5	41.9±0.8	40.8±1.0	78.0±27.2	**97.6±4.0**
	$\delta_0, \delta_1 \sim \mathcal{N}(0,2^2)$	46.2±0.7	45.6±0.7	46.3±0.6	45.8±0.3	45.7±1.0	68.1±17.4	**72.0±9.2**
Image CLEF-DA	**C→P**	74.3±0.5	69.2±0.4	74.5±0.3	74.1±0.7	**75.2±1.4**	74.3±0.3	**75.1±0.5**
	P→C	91.5±0.6	89.8±0.4	**93.5±0.4**	92.1±0.6	**93.5±2.8**	92.7±0.4	**93.4±0.3**
	I→P	75.0±0.6	74.5±0.4	76.7±0.3	76.8±0.4	76.7±1.4	77.0±0.3	**77.4±0.3**
	P→I	86.0±0.3	82.2±0.2	90.6±0.3	90.2±1.1	**91.0±0.8**	90.6±0.4	**91.1±0.5**
PACS	others→**P**	97.6±0.2	97.6±0.4	97.0±0.4	97.6±0.3	87.6±4.2	97.6±0.4	**97.9±0.2**
	others→**A**	85.9±0.5	84.5±1.2	84.0±0.9	88.1±0.8	86.4±0.4	88.0±0.8	**88.8±0.7**
	others→**C**	79.9±1.4	81.9±1.9	78.5±1.5	83.2±1.1	83.6±1.7	**84.6±0.9**	**84.7±0.8**
	others→**S**	75.2±2.8	77.4±3.1	71.8±3.9	80.2±2.2	59.1±1.5	80.9±1.2	**81.4±0.8**
VLCS	others→**V**	78.3±0.3	74.6±0.8	76.9±0.2	79.0±1.1	70.0±2.5	79.1±1.4	**81.1±0.8**
	others→**L**	64.9±1.1	67.1±0.5	65.2±0.4	63.8±0.8	54.0±5.9	69.6±0.9	**70.2±0.7**
	others→**C**	98.5±0.2	98.5±0.6	97.5±0.1	99.3±0.3	96.5±5.1	99.3±0.3	**99.5±0.2**
	others→**S**	73.1±0.7	75.0±1.1	73.4±1.1	75.8±1.8	66.8±2.0	76.1±1.8	**77.1±1.1**

meet a serious problem. But the task is hard for identifiability: for each value of a semantic factor, a single elementary training domain cannot show a diverse variation factor. This weakens the power of CSG-DA. On other datasets (shifted-MNIST, PACS, VLCS), spurious correlation is stronger. Shifted-MNIST is deliberately constructed to show a strong digit-position correlation in the training domain while the correlation disappears in test domains. As for PACS and VLCS, whenever different domains have different class proportions, pooling them together introduces a class-style(domain) correlation, which does not hold in a test domain. On the other hand, the training domain of shifted-MNIST shows a noisy position for each digit, and the pooled training domains of PACS and VLCS show a diverse style for each class. So these datasets better satisfy the requirement of CSG-DA meanwhile ameliorating spurious correlation is the key problem. This makes the advantage of CSG-DA more salient.

G.1 Shifted-MNIST

Dataset. The dataset is based on the standard MNIST dataset[21], where only images of "0" and "1" are collected. The resulting training set has 5,923 (46.77%) "0"s and 6,742 (53.23%) "1"s (12,665 in total) and the test set has 980 (46.34%) "0"s and 1,135 (53.66%) "1"s (2,115 in total). As described in the main text, we horizontally shift each "0" in the training data at random by δ_0 pixels where $\delta_0 \sim \mathcal{N}(-5, 1^2)$, and each "1" by $\delta_1 \sim \mathcal{N}(5, 1^2)$ pixels. We construct two test sets, where in the first one, each digit from the test set is not moved $\delta_0 = \delta_1 = 0$, and is horizontally shifted randomly by $\delta_0, \delta_1 \sim \mathcal{N}(0, 2^2)$ pixels in the second. All domains have balanced classes.

Setup and implementation details. For generative methods (*i.e.*, CSGz(-DA) and our methods CSG(-ind/-DA)), we use a multilayer perceptron (MLP) with 784(for x)-400-200(first 100 for v)-50(for s or z)-1(for y) nodes in each layer for the inference model, and use an MLP with 50(for s)-(100(for v)+100)-400-784(for x) nodes in each layer for the generative component (*i.e.*, the mean function of the additive Gaussian $p(x|s, v)$). The activation function in the MLPs is the sigmoid function, and the variables s and v are taken after the activation. The expectation under $q(s, v|x)$ in ELBO is estimated by evaluating the function at the mode of the additive Gaussian with reparameterization. For discriminative methods (*i.e.*, CE, CNBB, DANN, DAN, CDAN, MDD, BNM), we use a larger MLP architecture with 784-600-300-75-1 nodes in each layer to compensate the additional parameters of the generative component in generative methods.

For all the methods, we use a mini-batch of size 128 in each optimization step, and use the RMSprop optimizer [110], with weight decay parameter 1×10^{-5}, and learning rate 1×10^{-3} for OOD generalization and 3×10^{-4} for domain adaptation. These hyperparameters are chosen by running and validating using CE and DANN. For generative methods, we take the additive Gaussian variance of the generative mechanism $p(x|s, v)$ as 0.03^2. The scale of the standard derivations of these additive Gaussian distributions are chosen small to meet the intense causal mechanism assumption in our theory.[22] For the Gaussian variances of s and v in $q(s, v|x)$, they are also outputs from the discriminative model through additional branches. Each of these branches is a fully-connected layer forked from the last layer of s or v, with a softplus activation to ensure positivity. Their weights are learned via the same objectives.

Hyperparameter configurations. For both OOD generalization and domain adaptation tasks on the two test domains, we train the models for 100 epochs (average runtime 10 minutes) when all the methods converge in terms of loss and validation accuracy. We align the scale of the supervision loss terms in the objectives of all methods, and scale the ELBO terms with the largest weight that makes training accuracy near 1 in OOD generalization. We then fix the tuned ELBO weight and scale the weight of adaptation terms in a similar way for domain adaptation. Other parameters are tuned similarly. For generative methods (*i.e.*, CSGz(-DA) and our methods CSG(-ind/-DA)), the ELBO weight is 1×10^{-4} selected from $\{1, 3\} \times 10^{\{-1,-2,\cdots,-6\}}$. For domain adaptation methods, the adaptation weight is 1×10^{-4} for DANN, 1×10^{-8} for DAN, 1×10^{-6} for CDAN, 1×10^{-6} for MDD, 1×10^{-7} for BNM, and 1×10^{-4} for CSGz-DA and CSG-DA, all selected from $1 \times 10^{\{-1,-2,\cdots,-8\}}$. For CNBB, we use regularization coefficients 1×10^{-4} and 3×10^{-6} to regularize the sample weight and learned representation, and run 4 inner gradient descent iterations with learning rate 1×10^{-3} to optimize the sample weight. These four parameters are selected from a grid search where the range of the parameters are: $\{1, 3\} \times 10^{\{-2,-3,-4\}}$, $\{1, 3\} \times 10^{\{-4,-5,-6\}}$, $\{4, 8\}$, $1 \times 10^{\{-1,-2,-3\}}$.

G.2 ImageCLEF-DA

Dataset. ImageCLEF-DA[23] is a standard benchmark dataset for the ImageCLEF 2014 domain adaptation challenge [1]. There are three domains in this dataset: **C**altech-256, **I**mageNet and **P**ascal VOC 2012. Each domain has 12 classes and 600 images. Each image is center-cropped to shape $(3, 224, 224)$ as x (also for PACS and VLCS experiments).

[21] `http://yann.lecun.com/exdb/mnist/`

[22] Choosing small variances is also supported by a direct analysis of additive Gaussian VAEs [26] for well learning the data manifold.

[23] `http://imageclef.org/2014/adaptation`

Setup and implementation details. We adopt the same setup as in Long et al. [74] [24] for a common practice and fair comparison with existing results. This means that we use the ResNet50 structure [40] pretrained on the ImageNet dataset as the backbone of the discriminative/inference model. For CSG(-ind/-DA), we select the first 128 dimensions of the bottleneck layer (*i.e.*, the layer that replaces the last fully-connected layer of the pretrained ResNet50; its output dimension is 1024) as the variable v, and take s as the 256-dimensional output of the two-layer MLP (with 1024 hidden nodes) built on the bottleneck layer. Both s and v are taken before activation. The logits for y is produced by a linear layer built on s.

For generative methods (*i.e.*, CSGz(-DA) and our methods CSG(-ind/-DA)), we construct an image decoder/generator for the mean function of the additive Gaussian $p(x|s,v)$ that uses the DCGAN generator model [90] pretrained on the Cifar10 dataset as the backbone. The pretrained DCGAN is taken from the PyTorch-GAN-Zoo[25]. The generator connects to the DCGAN backbone by an MLP with 384(dimension of (s,v))-128-120(input dimension of DCGAN) nodes in each layer, and generates images of desired size $(3,224,224)$ by appending to the output of DCGAN of size $(3,64,64)$ with an transposed convolution layer with kernel size 4, stride size 4, and padding size 16. The expectation under $q(s,v|x)$ in ELBO is estimated by evaluating the function at the mean of the conditional Gaussian with reparameterization.

Following Long et al. [74], we use a mini-batch of size $n_B = 32$ in each optimization step, and adopt the SGD optimizer with Nesterov momentum parameter 0.9, weight decay parameter 5×10^{-4}, and a shrinking step size scheme $\varepsilon_i = \varepsilon_0(1+\alpha n_B i)^{-\beta}$ for optimization iteration i, with initial scale $\varepsilon_0 = 1 \times 10^{-3}$, per-datum coefficient[26] $\alpha = 6.25 \times 10^{-6}$, and shrinking exponent $\beta = 0.75$. For the parameters of the backbone components, a 10 times smaller learning rate is used. For generative methods, the Gaussian variances of s and v in $q(s,v|x)$ are also outputs from the discriminative model through additional branches. Each of these branches is a fully-connected layer forked from the last layer of s or v, with a softplus activation to ensure positivity. Their weights are learned via the same objectives.

Hyperparameter configurations. For all the four OOD prediction tasks, we train the models for 30 epochs (average runtime 10 minutes) when all the methods converge in terms of loss and validation accuracy. For generative methods, the Gaussian variance of $p(x|s,v)$ is taken as 0.1, which is searched within $\{1,3\} \times 10^{\{-4,-2,-1,0,2,4\}}$. The ELBO weight is 1×10^{-7} for CSGz(-DA) and is 1×10^{-8} for our CSG(-ind/-DA), both selected from $1 \times 10^{\{-2,-4,-6\}} \cup \{1,3\} \times 10^{\{-7,-8,-9,-10\}}$. The adaptation weight is 1×10^{-8} selected from $1 \times 10^{\{-2,-4,-6\}} \cup \{1,3\} \times 10^{\{-7,-8,-9,-10\}}$ for both CSGz-DA and CSG-DA, 1×10^{-2} selected from $1 \times 10^{\{-1,-2,-4,-6\}}$ for MDD, and 1.0 selected from $1 \times 10^{\{1,0,-1,-2,-4\}}$ for BNM. Results of other domain adaptation baselines DANN, DAN and CDAN and the results of CE are taken from [74] under the same setting. For CNBB, we use regularization coefficients 1×10^{-6} and 3×10^{-6} to regularize the sample weight and learned representation, and run 4 inner gradient descent iterations with learning rate 1×10^{-4} to optimize the sample weight. These four parameters are selected from a grid search where the range of the parameters are: $1 \times 10^{\{-4,-5,-6,-7\}} \cup \{3 \times 10^{-6}\}$, $\{1,3\} \times 10^{\{-5,-6,-7\}}$, $\{4\}$, $1 \times 10^{\{-2,-3,-4,-5\}}$.

G.3 PACS

Dataset. The PACS dataset [69] has 7 classes. It is named after its four domains: **P**hoto, **A**rt, **C**artoon, **S**ketch; each contains images of a certain style. It contains 9,991 images in total. We use the dataset via the open-source `domainbed` repository[27] [39].

Setup and implementation details. We adopt the same setup as in Gulrajani and Lopez-Paz [39] for a common practice and fair comparison with existing results. This means for each domain as the test domain, the single training domain is constructed by merging/pooling the other three domains. This is done by merging the three mini-batches of size 32 from each of the three domains for optimization. The Adam optimizer [60] with learning rate 5×10^{-5} is adopted. Data augmentation

[24] `https://github.com/thuml/CDAN`

[25] `https://github.com/facebookresearch/pytorch_GAN_zoo`

[26] The coefficient α here is amortized onto each datum, so its value is different from that in Long et al. [74] and a batch size n_B is multiplied to the iteration number i.

[27] `https://github.com/facebookresearch/DomainBed`

Table 4: Test accuracy (%) for **OOD generalization** (middle 4 columns) and **domain adaptation** (right 3 columns) by various methods (ours in bold and line separated) on **PACS** with **single training domains**. Averaged over 10 runs.

task		CE	CSGz	**CSG**	**CSG-ind**	DAN	CSGz-DA	**CSG-DA**
PACS	**C→A**	**78.9**±1.1	78.2±1.8	78.4±1.2	**78.9**±1.3	**80.9**±1.2	79.1±0.7	79.1±0.8
	P→A	73.1±1.9	73.4±1.9	**73.5**±0.9	73.4±1.5	**76.6**±2.6	73.8±0.7	75.0±0.7
	S→A	64.2±2.8	63.4±1.6	63.7±1.7	**65.4**±2.1	62.4±1.8	64.7±2.2	**65.7**±2.0
others→**A**		88.1±0.1	87.3±0.8	**88.5**±0.6	**88.6**±0.6	84.5±1.2	88.0±0.8	**88.8**±0.7

is conducted by random flip and crop, gray-scaling and color-jitter (*i.e.*, randomly changing brightness, contrast, saturation and hue). Other setups are basically the same as in the ImageCLEF-DA experiment, except that the layer for variable s has 512 nodes, and that the backbone components use the same learning rate (*i.e.*, not multiplied by 0.1).

Hyperparameter configurations. For all methods we train for 40 epochs (average runtime 30 minutes) when they all converge in terms of loss and validation accuracy. For all generative methods (*i.e.*, CSGz(-DA) and our methods CSG(-ind/-DA)), the Gaussian variance of $p(x|s,v)$ is taken as 0.3. The ELBO weight is 1×10^{-7} for CSGz, CSG and CSG-ind, and is 1×10^{-8} for CSGz-DA and CSG-DA, both selected from $1\times10^{\{0,-2,-4,-5,-6,-7,-8,-9\}}$. The adaptation weight is 1×10^{-8} selected from $1\times10^{\{0,-2,-4,-6,-7,-8,-9\}}$ for CSGz-DA and CSG-DA, 1×10^{-2} selected from $1\times10^{\{0,-1,-2,-3,-4,-6\}}$ for DAN, and is the same as in the ImageCLEF-DA experiment for MDD and BNM. Results of other domain adaptation baselines DANN and CDAN and the results of CE are taken from [39] under the same setting. For CNBB, the hyperparameters are the same as in the ImageCLEF-DA experiment, except the regularization coefficients for sample weights is 1×10^{-4}. These hyperparameters are selected from the same range as used in the ImageCLEF-DA experiment.

Results using single training domains. We also conducted an experiment on PACS with single training domains, similar to the setup on ImageCLEF-DA. The results are presented in Table 4. We see that the advantage of our methods is not as significant as in the standard pooled training domain case. This agrees with the discussion in the "dataset analysis" in the main paper: our methods are more powerful in handling a misleading spurious s-v correlation but which needs to be diverse/stochastic enough to allow identification, following the intuition on the identifiability (Thm. 5 Remark **(1)**).

G.4 VLCS

The VLCS dataset [30] has 5 classes. It is also named after its four domains: **V**OC2007, **L**abelMe, **C**altech101, **S**UN09; each is an image dataset collected in a certain way. It contains 10,729 images in total. We use the dataset also via the `domainbed` repository. Setup, implementation details and hyperparameters are the same as in the PACS experiment. Results are shown at the last four rows in Table 2 for OOD generalization and in Table 3 for domain adaptation.

G.5 Visualization of the Learned Representation

To better understand how our methods work, we compare the visualization of the learned model by our methods with that by the corresponding baselines. Visualization is done by the *Local Interpretable Model-agnostic Explanation* (LIME) method [91][28], which uses an interpretable model, *e.g.* a linear model, to approximate the target model locally at the query image. The learned weight of the linear model then reflects the importance of the components/dimensions of the input, *i.e.* pixels in the image, which can be visualized after binarization as focused regions on the image. This gives a hint on the learned representation by the model for making prediction.

The visualization results are shown in Fig. 5. We see that in each case, the focused regions of our methods (CSG-ind and CSG-DA) are more relevant to the semantic of the image, and the boundary of the region reflects the characterizing shape of the object. In contrast, the baselines also involve much background regions. This result shows our CSG methods indeed better learn a causal semantic factor for prediction, which supports the motivation to introduce the CSG model, verifies the theory, and explains the better robustness for OOD prediction.

[28] We use the official codebase at `https://github.com/marcotcr/lime-experiments`.

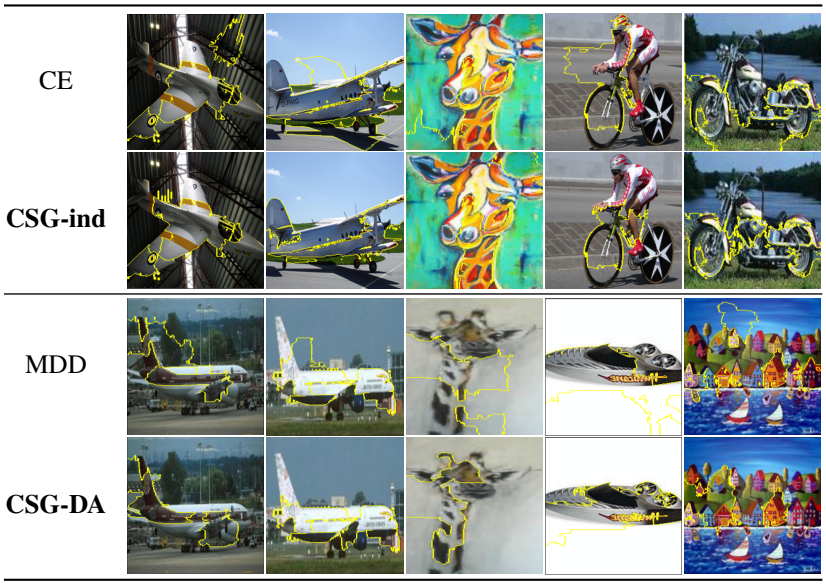


Figure 5: Visualization (via LIME [91]) of the learned representation by various methods (ours in bold). The top two rows are for OOD generalization and the bottom two rows are for domain adaptation.