Provable Robustness to Spurious Correlations via Invariant Data For Robust Finetuning

Ruqi Bai[1], **Yao Ji**[2], **Mingyu Kim**[1], **Easton Currie**[1], **Zeyu Zhou**[1], **David I. Inouye**[1]
[1] Purdue University, [2] Georgia Institute of Technology

Abstract

Scaling models on simple predictive objectives is often insufficient to overcome spurious correlations that degrade out-of-distribution generalization. While Domain Generalization (DG) methods aim to learn invariant representations–often based on causality principles–they can be computationally expensive and underperform simple Empirical Risk Minimization (ERM). We propose a data-centric alternative: Geometric Robustness via Invariant Training (GRIT). Instead of explicit causal modeling, GRIT enforces a geometric constraint during finetuning based on a small set of *noisy invariant pairs*, which implicitly encode an invariance property. We provide the first finite-sample analysis of this setting, showing that our framework generalizes latent linear causal models. We prove GRIT achieves robust generalization that scales at a rate of $O(1/\sqrt{k})$ with the number of pairs k, offering a scalable alternative to ERM or explicit causal modeling for out-of-distribution robustness.

1 Introduction

Adapting pre-trained foundation models via finetuning has become ubiquitous. While scaling these feature spaces captures rich semantic information, they still suffer from spurious correlations—misleading patterns that hold in the training data but fail in new environments [Sagawa et al., 2019]. Addressing these failures reveals a critical limitation: simply scaling standard ERM on observational data is often insufficient to induce true out-of-distribution robustness.

Standard approaches fall under Robust Fine-Tuning or Domain Generalization (DG). Methods like WiSE-FT [Wortsman et al., 2022] require expensive retraining. Many DG approaches aim to learn invariant representations and are explicitly inspired by causality; for example, Invariant Risk Minimization (IRM) [Arjovsky et al., 2019] builds on invariant causal prediction (ICP) [Peters et al., 2016], and MatchDG [Mahajan et al., 2021] explicitly assumes a causal model. However, these methods are often unstable, underperform simple ERM [Gulrajani and Lopez-Paz, 2021], and rely on strong assumptions. While a few prior works explore out-of-distribution robustness by leveraging paired data [Mahajan et al., 2021, Nguyen et al., 2021, Surner et al., 2025], they lack theoretical guarantees addressing noisy and finite data pairs. (See Section A for expanded related works.)

We shift the focus from explicit causal modeling by asking: **Can we provide theoretic guarantees for the robustness of a predictor on a fixed feature space using a small number of *noisy invariant pairs*—samples that share the same semantic label but differ in spurious attributes (e.g., counterfactuals)?**

We operate under the geometric intuition that spurious correlations reside in specific subspaces of the feature embedding. Importantly, our framework generalizes latent linear causal models, demonstrating the connection to causal models. While grounded in linear analysis, our framework naturally extends to kernelized non-linear predictors. Crucially, our framework explicitly models pair noise and provides a finite-sample analysis with respect to the number of pairs. Our contributions are threefold:

- We propose GRIT, a fine-tuning approach that enforces geometric invariance via noisy pairs.
- We prove that GRIT robustness improves at a rate of $O(1/\sqrt{k})$ with the number of pairs k.
- We demonstrate that GRIT outperforms baselines on real-world datasets (Waterbirds, ColoredMNIST) using frozen CLIP embeddings.

2 Problem Setup

We analyze the robustness of a model trained on a frozen representation feature space $\boldsymbol{z} = \phi(\boldsymbol{x}) \in \mathbb{R}^d$.

Geometric Structure and Distribution Shifts. We assume the representation space possesses a geometric structure where predictive signals and spurious correlations reside in distinct subspaces.

Assumption 1 (Linear Subspace Decomposition). *The representation feature space $\mathbb{R}^d$ decomposes into two linearly independent subspaces: $\mathbb{R}^d = \mathcal{S}_{core} \oplus \mathcal{S}_{sp}$, where $\oplus$ denotes the direct sum of vector spaces and $\mathcal{S}_{core} \cap \mathcal{S}_{sp} = \{0\}$ where $\mathcal{S}_{core}$ is the stable feature subspace and $\mathcal{S}_{sp}$ is the spurious features subspace.*

Because of Assumption 1, any feature representation can be uniquely decomposed as $\boldsymbol{z} = \boldsymbol{z}_{\text{core}} + \boldsymbol{z}_{\text{sp}}$. To formalize out-of-distribution robustness, we model distribution shifts as preserving the stable core subspace and target distribution while allowing the spurious feature distribution $\mathbb{P}(\boldsymbol{z}_{\text{sp}}|\boldsymbol{z}_{\text{core}}, y)$ to change arbitrarily.

Assumption 2 (Spurious Shift). *We assume the distribution on the stable core subspace and the target remains the same, i.e., $\mathbb{P}_{test}(\boldsymbol{z}_{test,core}, y_{test}) = \mathbb{P}_{train}(\boldsymbol{z}_{core}, y)$.*

(See Remark 1 in appendix for discussion on relaxing this to covariate shift, i.e., only $\mathbb{P}_{\text{test}}(y_{\text{test}}|\boldsymbol{z}_{\text{test,core}}) = \mathbb{P}_{\text{train}}(y|\boldsymbol{z}_{\text{core}})$.) Given this, we can construct a probabilistic coupling to connect the train and test distributions and prove that its distribution is equal to the test distribution.

Definition 1 (Training and Test Coupling). *For any train and test distributions satisfying Assumption 2, we construct new coupled random variables $(\boldsymbol{z}', y')$ defined as:*

$$\boldsymbol{z}' := \boldsymbol{z} + \Delta, \quad \Delta := \boldsymbol{z}'_{sp} - \boldsymbol{z}_{sp}, \quad y' := y\,, \tag{1}$$

where $(\boldsymbol{z}, y) \sim \mathbb{P}_{train}(\boldsymbol{z}, y)$, $\Delta \in \mathcal{S}_{sp}$ is random shift vector, and $\boldsymbol{z}'_{sp}$ is drawn from the conditional distribution $\pi(\boldsymbol{z}'_{sp} \mid \boldsymbol{z}_{sp}, \boldsymbol{z}_{core}, y)$ derived from a conditional optimal transport coupling $\pi(\boldsymbol{z}_{sp}, \boldsymbol{z}'_{sp} \mid \boldsymbol{z}_{core}, y)$ between the train and test spurious conditionals.

Lemma 1 (Distributional Equivalence via Coupling). *The joint distribution of the coupled variables in Definition 1 exactly matches the test distribution: $\mathbb{P}(\boldsymbol{z}', y') = \mathbb{P}_{test}(\boldsymbol{z}_{test}, y_{test})$.*

Observation Model: Noisy Invariant Pairs. We do not observe the subspaces $\mathcal{S}_{\text{core}}$ or $\mathcal{S}_{\text{sp}}$ directly. Instead, we assume access to k noisy invariant pairs whose difference vectors span the spurious subspace. These pairs need not be out of distribution and can be constructed entirely within the training support.

Assumption 3 (Noisy Invariant Pairs). *We observe k invariant pairs $\{(\boldsymbol{x}_i, \boldsymbol{x}'_i)\}_{i=1}^k$, where we model the difference of their representations $\boldsymbol{\delta}_i = \phi(\boldsymbol{x}_i) - \phi(\boldsymbol{x}'_i)$ as a combination of a signal and random noise:*

$$\boldsymbol{\delta}_i = \boldsymbol{v}_i + \boldsymbol{\xi}_i\,, \tag{2}$$

where the corresponding uncentered covariance matrix is defined as $\tilde{\Sigma}_k := \frac{1}{k}\sum \delta_i\delta_i^\top$. We assume the following structural properties:

1. *Signal Diversity (Geometry): The clean difference vectors $\boldsymbol{v}_i$ lie strictly within the spurious subspace $\mathcal{S}_{sp}$. Furthermore, they fully span this subspace, such that the zero-noise, signal covariance $\Sigma_k := \frac{1}{k}\sum \boldsymbol{v}_i\boldsymbol{v}_i^\top$ has rank exactly $d_{sp} = \dim(\mathcal{S}_{sp})$.*
2. *Measurement Noise (Statistics): The noise vectors $\boldsymbol{\xi}_i$ are independent, zero-mean, isotropic sub-Gaussian random vectors with parameter σ_ξ.*

See Remark 2 for a discussion on these assumptions.

Connection to Causal Models. While presented through a geometric lens, our framework generalizes standard causal formulations as explained in Section C. Specifically, our problem setup generalizes Linear Latent Structural Causal Models (SCMs), and we prove that causally generated "spurious counterfactuals" perfectly satisfy our invariant pair conditions. Thus, while being inspired by causal modeling, we do not require any causal modeling assumptions.

3 Geometric Robustness via Invariant Training (GRIT)

Our geometric characterization suggests a direct algorithm for robustness: estimate $\mathcal{S}_{\text{sp}}$ and force the model to be orthogonal to it. We compute the eigendecomposition of the proxy covariance $\tilde{\Sigma}_k$ and estimate the spurious subspace $\hat{\mathcal{S}}_{\text{sp}}$ as the span of its top-r eigenvectors, denoted by $\tilde{Q}_r \in \mathbb{R}^{d\times r}$. The hyperparameter r controls the estimated dimension of the spurious subspace.

With the estimated basis $\tilde{Q}_r$, GRIT seeks a classifier w that minimizes the empirical risk while remaining strictly invariant to shifts in $\hat{\mathcal{S}}_{\text{sp}}$:

$$\min_w \mathbb{E}_{(\boldsymbol{x},y)\sim\hat{\mathbb{P}}_{\text{train}}}[\ell(w^\top\phi(\boldsymbol{x}), y)] \quad \text{s.t.} \quad w^\top\tilde{Q}_r = 0 \tag{3}$$

where $\ell(\cdot,\cdot)$ is a convex loss function. Geometrically, this restricts the classifier to the orthogonal complement $\hat{\mathcal{S}}_{\text{sp}}^\perp$.

This constrained problem can be solved efficiently by projecting all training samples onto the null space of $\hat{\mathcal{S}}_{\text{sp}}$ using $P_\perp = I - \tilde{Q}_r\tilde{Q}_r^\top$ prior to standard unconstrained ERM. The full pseudocode for GRIT and an extended discussion unifying GRIT with soft Lagrangian penalties (e.g., used in MatchDG [Mahajan et al., 2021]) are deferred to Remark 2.

4 Theoretical Guarantees for Robust Linear Probing

In this section, we provide finite-sample guarantees for GRIT in the context of robust linear probing. Our analysis proceeds in two steps: (1) decomposing the test error into the training error and a *spurious subspace misalignment* term, and (2) deriving an explicit finite-sample bound for this misalignment.

We analyze a linear predictor $h(\boldsymbol{z}) = w^\top \boldsymbol{z}$ trained on the fixed feature space. To quantify the deviation of a specific test domain from the training domain under the spurious shift model (Assumption 2), let $\Sigma_\Delta^{\text{te}}$ denote the uncentered covariance of the test shifts $\Delta = \boldsymbol{z}' - \boldsymbol{z}$ from Definition 1:

$$\Sigma_\Delta^{\text{te}} \coloneqq \mathbb{E}[\Delta\Delta^\top] = Q^{\text{te}}\Lambda^{\text{te}}Q^{\text{te}\top}. \tag{4}$$

Here, Q^{te} represents the orthonormal basis of the subspace spanned by these specific test shifts, and Λ^{te} is the diagonal matrix of descending eigenvalues λ_i^{te}.

Error Decomposition. We first establish a general error bound that holds for any estimated subspace $\hat{\mathcal{S}}_{\text{sp}}$, represented by the projection matrix onto its orthogonal complement.

Theorem 1 (Test-Domain Error Decomposition)**.** *Let w be any linear predictor satisfying the GRIT constraint $w^\top \tilde{Q}_r = 0$, where $\tilde{Q}_r \in \mathbb{R}^{d\times r}$ is the estimated spurious basis. Let $\tilde{Q}_{r,\perp}$ denote the basis for its orthogonal complement. Assuming Assumption 2, the test risk for Logistic Regression (Log Loss ℓ_{LL}) is bounded by:*

$$\mathbb{E}_{\text{test}}\left[\ell_{\text{LL}}(w^\top \boldsymbol{z}_{\text{test}}, y)\right] \le \underbrace{\mathbb{E}_{\text{train}}[\ell_{\text{LL}}(w^\top \boldsymbol{z}, y)]}_{\textit{Term I: Training error}} + \underbrace{\|w\|\|\tilde{Q}_{r,\perp}^\top Q^{te}\|_{\Lambda^{te}}}_{\textit{Term II: Subspace misalignment}}.$$

A similar error decomposition holds for Linear Regression (Squared Error).

We now unpack *Term II* to understand how the choice of r influences robustness by separating it into an estimation error (subspace distance) and an approximation error (tail eigenvalues).

Lemma 2 (Structural Upper Bound)**.** *Let $s \coloneqq \min\{r, \operatorname{rank}(\Sigma_\Delta^{te})\}$ be the overlap between the estimated rank and the test shift dimension, and let $\operatorname{dist}^2(Q, Q') \coloneqq \|QQ^\top - Q'Q'^\top\|^2$ denote the squared distance between subspaces. The misalignment term satisfies:*

$$\left\|\tilde{Q}_{r,\perp}^\top Q^{te}\right\|_{\Lambda^{te}}^2 \le \lambda_1^{te}\operatorname{dist}^2(\tilde{Q}_s, Q_s^{te}) + \lambda_{s+1}^{te}. \tag{5}$$

This highlights the explicit trade-off for choosing r: if $r < \operatorname{rank}(\Sigma_\Delta^{\text{te}})$, the approximation error $\lambda_{s+1}^{\text{te}}$ remains non-zero. Conversely, if r is too large, the in-domain training error may increase as informative core features are incorrectly penalized.

Finite-Sample Analysis. We now derive an explicit finite-sample bound when estimating the subspace from k noisy invariant pairs. By combining matrix perturbation theory with high-dimensional concentration inequalities, we quantify the impact of noise and sample size.

Theorem 2 (Finite-Sample Generalization Bound)**.** *Assume Assumptions 1, 2, and 3 and assume the selected rank r is greater than or equal to the true spurious dimension (i.e., $r \ge d_{sp}$). Further, assume the number of pairs $k \ge d$. With probability at least $1-\eta$, the test-domain error for log-loss is bounded by:*

$$\mathbb{E}_{\text{test}}[\ell_{LL}(w^\top \boldsymbol{z}, y)] \le \mathbb{E}_{\text{train}}[\ell_{LL}(w^\top \boldsymbol{z}, y)] + C\|w\|\sqrt{\lambda_1^{te}}\left[\frac{\sigma_\xi^2 + \sigma_\xi\sqrt{\lambda_1(\Sigma_k)}}{\lambda_{d_{sp}}(\Sigma_k)}\sqrt{\frac{d\log(2d/\eta)}{k}}\right]$$

where λ_1^{te} is the spectral norm of the test shift covariance, $\lambda_i(\Sigma_k)$ are the eigenvalues of the clean signal covariance Σ_k, σ_ξ is the sub-Gaussian parameter of the noise, and C is a universal constant. Similar bounds hold for regression with squared error loss.

This finite-sample bound yields several key geometric intuitions about GRIT's robustness. First, test risk scales linearly with the strength of the test domain shift ($\sqrt{\lambda_1^{\text{te}}}$), meaning large test shifts increase estimation errors. Second, robustness depends heavily on the signal-to-noise ratio and diversity of the invariant pairs (governed by the ratio $\frac{\sigma_\xi^2+\sigma_\xi\sqrt{\lambda_1(\Sigma_k)}}{\lambda_{d_{\text{sp}}}(\Sigma_k)}$, where σ_ξ is the noise and $\lambda_{d_{\text{sp}}}(\Sigma_k)$ is the signal). Finally, the error decays at a rate of $\mathcal{O}(\sqrt{d/k})$, confirming that collecting more noisy invariant pairs steadily recovers exact geometric invariance. Extended discussions regarding the implicit dependency on the noise dimensionality d, a relaxation of the invariant pair diversity constraint, and theoretical limitations (such as the strict reliance on $r \ge d_{\text{sp}}$ and sub-Gaussian isotropic noise assumptions) are discussed in Section E.

5 Empirical Evaluation

We evaluate GRIT's theoretical properties via synthetic data and its practical efficacy via robust linear probing on frozen CLIP features [Radford et al., 2021].

Synthetic Experiments. We validate our finite-sample bounds (Theorem 2) and the rank trade-off

Table 1: Main Results with in-domain validation. GRIT consistently outperforms baselines on Worst Group (WG) accuracy. Results with oracle validation and PACS are in the appendix.

			ColoredMNIST		Waterbirds	
	Data	Model	in acc	test acc	in acc	wg acc
ERM (CLIP)	DG	Probing	0.852	0.093	0.885	0.781
IRM	DG	Probing	0.799	0.118	0.838	0.707
REx	DG	Probing	0.797	0.121	0.891	0.617
GroupDRO	DG	Probing	0.798	0.127	0.906	0.684
Fish	DG	Probing	0.798	0.118	0.900	0.744
SWAD	DG	Probing	0.800	0.113	-	-
LISA	DG	Probing	0.705	**0.000**	0.904	0.722
MatchDG	1NN	CNN	0.698	0.361	0.970	0.080
MatchDG	1NN	Finetune	0.850	0.181	0.920	0.112
MatchDG	random	Probing	0.799	0.120	0.793	0.009
MatchDG	1NN	Probing	0.789	0.217	0.886	0.411
MatchDG	clean	Probing	0.793	0.181	0.906	0.536
GRIT	random	Probing	0.794	0.176	0.804	0.269
GRIT	1NN	Probing	0.736	0.649	0.892	0.521
GRIT	clean	Probing	0.740	**0.693**	**0.864**	**0.812**
random guess			0.500	0.500	-	-
ERM oracle			0.735	0.730	-	-
theory oracle			0.750	0.750	-	-

(Lemma 2) using a synthetic dataset with a known spurious subspace dimension $d_{\text{sp}} = 20$. As detailed in Appendix J.1, test accuracy monotonically improves at the predicted $O(1/\sqrt{k})$ rate as the number of pairs k increases. Furthermore, accuracy peaks precisely when the estimated rank $r \approx d_{\text{sp}}$, directly confirming our error decomposition and the theoretical "price of noise." (Full synthetic results and figures are deferred to Appendix J.1).

Real-world Datasets & Setup. We benchmark GRIT on tasks with significant spurious correlations: *ColoredMNIST* [Arjovsky et al., 2019] (color vs. digit) and *Waterbirds* [Sagawa et al., 2019] (background vs. bird species). To isolate the background shift in Waterbirds, we construct noisy invariant pairs ("counterfactuals") by pairing minority group samples (e.g., waterbirds on land) with randomly selected majority samples (e.g., waterbirds on water). We compare GRIT against standard ERM, matching heuristics (Random, 1-Nearest Neighbor), and strong Domain Generalization (DG) baselines. All extended dataset details, validation metrics, and hyperparameters are provided in Appendix J.

Results. As summarized in Table 1, GRIT consistently outperforms all baselines. On *ColoredMNIST*, GRIT achieves 81.2% worst-group accuracy using in-domain validation, surpassing ERM by 3.1%. Notably, standard DG methods often underperform ERM here due to optimization difficulties, whereas GRIT's closed-form projection remains highly stable. On *Waterbirds*, GRIT achieves 86.0% worst-group accuracy, outperforming the best baseline by 3.3%. We also observe that basic pairing heuristics (Random or 1NN) work adequately on the uniform shifts of ColoredMNIST but fail entirely on the high-variance Waterbirds dataset. This confirms that accurate $\hat{\mathcal{S}}_{\text{sp}}$ estimation relies critically on the semantic consistency of the noisy invariant pairs.

6 Extensions

Extension 1: Non-linear Predictors via Kernel GRIT. While GRIT naturally operates in linear spaces, it can be extended to non-linear predictors using an invariant kernel approach. Instead of learning a linear projection, Kernel GRIT applies a direct invariant L2 kernel correction to the Gram matrix before SVM classification. Evaluating this on the ColoredMNIST benchmark using frozen CLIP features, Kernel GRIT demonstrates strong robustness. While standard ERM achieves a high in-domain validation accuracy (84.86%) but collapses to 10.11% on out-of-distribution test data, Kernel GRIT maintains a strong 71.50% test accuracy. This performance operates near the dataset's theoretical ceiling caused by inherent label noise. Full details of the fully kernelized dual formulation, our efficient Preconditioned Conjugate Gradient (PCG) solver, and complete experimental results are deferred to Appendix G.

Extension 2: Semi-Supervised Domain Generalization (SSDG). GRIT naturally extends to the Semi-Supervised Domain Generalization (SSDG) setting, where models must generalize to unseen domains using limited labeled data and abundant unlabeled data. By leveraging data augmentations (e.g., AdaIN style transfer, Fourier amplitude mixing) on source-domain samples, we can construct noisy invariant pairs without requiring additional labels. Applying GRIT via post-hoc linear probing on frozen SSDG backbones consistently improves out-of-distribution performance. Notably, GRIT provides complementary gains even when applied on top of state-of-the-art SSDG algorithms like StyleMatch and UPCSC. Full methodology, related works, and comprehensive results are deferred to Appendix H.

References

Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-distribution generalization, 2022.

Asha Anoosheh, Eirikur Agustsson, Radu Timofte, and Luc Van Gool. Combogan: Unrestrained scalability for image domain translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 783–790, 2018.

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

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

Liang Chen, Yong Zhang, Yibing Song, Anton Van Den Hengel, and Lingqiao Liu. Domain generalization via rationale invariance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1751–1760, 2023.

Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020.

Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Katherine Hermann, Hossein Mobahi, Thomas FEL, and Michael Curtis Mozer. On the foundations of shortcut learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Tj3xLVuE9f.

Narges Honarvar Nazari and Adriana Kovashka. Domain generalization using shape representation. In *European Conference on Computer Vision*, pages 666–670. Springer, 2020.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *The Eleventh International Conference on Learning Representations*, 2023.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International conference on machine learning*, pages 5815–5826. PMLR, 2021.

Dongkwan Lee, Kyomin Hwang, and Nojun Kwak. Unlocking the potential of unlabeled data in semi-supervised domain generalization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30599–30608, 2025.

Sara Magliacane, Thijs Van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. *Advances in neural information processing systems*, 31, 2018.

Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In *International Conference on Machine Learning*, pages 7313–7324. PMLR, 2021.

A Tuan Nguyen, Toan Tran, Yarin Gal, and Atilim Gunes Baydin. Domain invariant representation learning with domain density transformations. *Advances in Neural Information Processing Systems*, 34:5264–5275, 2021.

Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM conference on health, inference, and learning*, pages 151–159, 2020.

Jonas Peters, Peter Bühlmann, and Nicolai 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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Mohammad Mahfujur Rahman, Clinton Fookes, Mahsa Baktashmotlagh, and Sridha Sridharan. Multi-component image translation for deep domain generalization. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 579–588. IEEE, 2019.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. *arXiv preprint arXiv:2010.05761*, 2020.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Kristof T Schütt, Huziel E Sauceda, P-J Kindermans, Alexandre Tkatchenko, and K-R Müller. Schnet–a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24), 2018.

Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

Adarsh Subbaswamy, Peter Schulam, and Suchi Saria. Preventing failures due to dataset shift: Learning predictive models that transport. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3118–3127. PMLR, 2019.

Martin Surner, Abdelmajid Khelil, and Ludwig Bothmann. Invariance pair-guided learning: Enhancing robustness in neural networks. *arXiv preprint arXiv:2502.18975*, 2025.

Joel A Tropp et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

Roman Vershynin. High-dimensional probability, 2009.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. *The Caltech-UCSD Birds-200-2011 Dataset*. California Institute of Technology, Jul 2011.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip S Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE transactions on knowledge and data engineering*, 35 (8):8052–8072, 2022.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7959–7971, 2022.

Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14383–14392, 2021.

Huaxiu Yao, Xinyu Yang, Xinyi Pan, Shengchao Liu, Pang Wei Koh, and Chelsea Finn. Improving domain generalization with domain relations. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Dc4rXq3HIA.

Yi Yu, Tengyao Wang, and Richard J Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15 (11):e1002683, 2018.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

Kaiyang Zhou, Chen Change Loy, and Ziwei Liu. Semi-supervised domain generalization with stochastic stylematch. *International Journal of Computer Vision*, 131(9):2377–2387, 2023.

A Expanded Related Works

Data augmentation and generation: Data augmentations can be seen as simple-to-generate counterfactual pairs, where the augmentations implicitly encode knowledge about desired invariances. For example, standard functions like rotation, scaling, and noise addition suggest that such transformations should not alter the predicted class [Honarvar Nazari and Kovashka, 2020, Shorten and Khoshgoftaar, 2019]. More sophisticated strategies follow this principle; LISA, for instance, is a Mixup-inspired method that learns domain-invariant predictors through intra-label and intra-domain mixing to encourage the model to respect class boundaries [Yao et al., 2024]. DIRT [Nguyen et al., 2021] suggests using StarGAN [Choi et al., 2018] to generate paired samples, while other work has used ComboGAN [Anoosheh et al., 2018] to generate new data [Rahman et al., 2019]. From our perspective, these generated samples can be seen as complex, class-preserving data augmentations to estimate pairs regarding some type of invariances. Our work provides a causal language and theoretical guarantee for this approach.

Distribution or sample matching in addressing spurious correlations: Invariant Risk Minimization (IRM) [Arjovsky et al., 2019] aim to mitigate spurious correlations by learning domain-invariant representations. Despite their theoretical appeal, IRM-based approaches often under perform in practice, prompting several works to analyze and refine them [Rosenfeld et al., 2020, Krueger et al., 2021, Ahuja et al., 2022]. Beyond distribution matching, MatchDG [Mahajan et al., 2021] introduces an iterative sample-level matching objective that aligns representations across domains in latent space. DIRT [Nguyen et al., 2021] first applies distribution matching to create data pairs and then applies a penalty to the squared difference similar to MatchDG. IPG [Surner et al., 2025] proposed to an invariant pair regularizer based on matching rational matrices [Chen et al., 2023] with an adaptive step size though [Surner et al., 2025] provide no theoretic analysis. Our method similarly employs sample-wise matching but crucially, we provide a theoretical robustness guarantee and deeper exploration on the properties of these pairs.

Causal inference seeking invariant predictor for robustness: The goal of domain generalization in a causal perspective is to find a representation Φ of $\boldsymbol{x}$ such that $y \perp \mathrm{e}|\Phi(\boldsymbol{x})$. Different approach to induce Φ has been heavily explored. Most of the causal inference type of work focusing on observable causal variables Magliacane et al. [2018] proposes to find subset of causal variable $\Phi(\boldsymbol{x})$ in $\boldsymbol{x}$, where $\boldsymbol{x}$ is the set of observable causal variables and $\Phi(\boldsymbol{x}) \subset \boldsymbol{x}$, such that $y \perp \mathrm{e}|\Phi(\boldsymbol{x})$ holds. Subbaswamy et al. [2019] considers the graph surgery estimator that finding the stable estimator by removing unstable mechanism from the joint factorization. However, it is difficult when the causal variables are latent.

B Remarks

Remark 1 (Relaxation to Covariate Shift on Stable Subspace)**.** We hypothesize that we can relax the assumption to a covariate shift assumption on the stable subspace, i.e., $\mathbb{P}_{\text{test}}(y_{\text{test}}|\boldsymbol{z}_{\text{test,core}}) = \mathbb{P}_{\text{train}}(y|\boldsymbol{z}_{\text{core}})$, while allowing the marginal distribution of stable features to change, i.e., $\mathbb{P}_{\text{test}}(\boldsymbol{z}_{\text{test,core}}) \neq \mathbb{P}_{\text{train}}(\boldsymbol{z}_{\text{core}})$. This change would require including a weighting function in the theory corresponding to the ratio of densities between the train and test on the stable features, i.e., $\omega(\boldsymbol{z}_{\text{core}}) := \mathbb{P}_{\text{test}}(\boldsymbol{z}_{\text{test,core}} = \boldsymbol{z}_{\text{core}})/\mathbb{P}_{\text{train}}(\boldsymbol{z}_{\text{core}} = \boldsymbol{z}_{\text{core}})$ but would otherwise have a similar form.

Remark 2 (Scope of Robustness and Invariant Pair Diversity)**.** Our theoretical guarantees rely on the signal covariance Σ_k derived from the clean invariant pairs. We note that Assumption 3 (Diversity) holds *without loss of generality* with respect to the *observable* spurious subspace. Any spurious feature that does not vary within the invariant pairs is mathematically indistinguishable from a stable core feature, as there is no signal in the invariant pairs to reveal its spurious nature. Consequently, if the pairs span only a subspace $\mathcal{S}' \subsetneq \mathcal{S}_{\text{sp}}$ (e.g., capturing *background* but missing *lighting*), our theory remains valid relative to $\mathcal{S}'$. In this case, the unobserved spurious directions are implicitly treated as part of the stable core $\mathcal{S}_{\text{core}}$. While the model remains vulnerable to shifts along these unobserved dimensions, this reflects a fundamental information-theoretic limit rather than a restrictive assumption: one can only be robust to spurious variations represented in the invariant pairs.

Remark 3 (Proxy Covariance vs. Test Shift Covariance)**.** It is important to emphasize that $\tilde{\Sigma}_k$ is not intended to approximate the covariance of the test-time shift, $\Sigma_\Delta^{\text{te}} := \mathbb{E}[\Delta(\boldsymbol{z}, y)\Delta(\boldsymbol{z}, y)^\top]$. The actual distribution of shifts in the test domain (captured by the eigenvalues of $\Sigma_\Delta^{\text{te}}$) may be completely different from the distribution of differences in our collected pairs. Our method relies only on the *principal subspace* of $\tilde{\Sigma}_k$ approximating the support of $\Sigma_\Delta^{\text{te}}$ (i.e., the spurious subspace $\mathcal{S}_{\text{sp}}$). As long as the eigenvectors of our proxy covariance span the spurious subspace, GRIT can enforce invariance.

C Causal Grounding: The Linear Latent SCM Special Case

While our framework is presented through a geometric lens, it is fully compatible with—and indeed generalizes—standard linear causal models used in the domain generalization literature. In this section, we show that a Linear Latent Structural Causal Model (SCM) is a special case of our geometric setup, and that *spurious counterfactuals* in the causal sense are equivalent to our *invariant pairs*.

We consider a data generation process involving latent variables $\mathbf{h}$, where the observed representation $\boldsymbol{z}$ is a linear mixing of these latents.

Assumption 4 (Linear Latent SCM)**.** *Let the latent space decompose into stable variables $\mathbf{h}_{inv} \in \mathbb{R}^{d_c}$ and spurious variables $\mathbf{h}_{sp} \in \mathbb{R}^{d_s}$. The data generation process follows:*

1. ***Invariant Target Mechanism:*** *The parents of the label, $Pa(y)$, are a subset of the stable latents $\mathbf{h}_{inv}$. The conditional distribution of the label, $P(y \mid \mathbf{h})$, depends on the latents solely through a linear transformation of these parents. That is, $P(y \mid \mathbf{h}) = f(y; \beta^\top \mathbf{h}_{inv})$ for some density function f and parameter vector β.*

2. ***Linear Observation:*** *The observed features $\boldsymbol{z}$ are generated by a linear transformation $M \in \mathbb{R}^{d \times (d_c + d_s)}$ of the latents:*
$$\boldsymbol{z} = M \begin{bmatrix} \mathbf{h}_{inv} \\ \mathbf{h}_{sp} \end{bmatrix} = M_{inv}\mathbf{h}_{inv} + M_{sp}\mathbf{h}_{sp}, \tag{6}$$
satisfying the condition that the stable and spurious subspaces are linearly independent: $Image(M_{inv}) \cap Image(M_{sp}) = \{0\}$.

3. ***Spurious Mechanism Shift:*** *Domain shifts are modeled as changes to the causal mechanism generating the spurious latents $\mathbf{h}_{sp}$. While the mechanism generating $\mathbf{h}_{inv}$ and the conditional distribution $P(y \mid \mathbf{h}_{inv})$ remain invariant, the structural equation generating $\mathbf{h}_{sp}$ may change arbitrarily across domains.*

Remark 4 (Generality of Assumptions)**.** Assumption 4 is highly general regarding the underlying causal structure:

- **Non-linear Latent Mechanisms:** We place *no* restrictions on the functional form of the mechanisms generating $\mathbf{h}_{\text{inv}}$ and $\mathbf{h}_{\text{sp}}$. They may be outcomes of complex, deep, non-linear structural equations. The linearity constraint applies only to the *observation* mapping (from $\mathbf{h}$ to $\boldsymbol{z}$) and the *target* dependency (from $\mathbf{h}$ to y).
- **Mechanism vs. Distribution:** By defining shifts as changes to the *mechanism* (structural equation), we capture a broader class of interventions than simple distributional shifts. This allows for soft interventions where the data generating process changes even if the marginal distribution $P(\mathbf{h}_{\text{sp}})$ were to remain the same, as well as hard interventions (e.g., $do(\mathbf{h}_{\text{sp}} = c)$).
- **Rank Relaxation:** We do not require M to be full rank. We only require that the spurious signals do not perfectly mimic stable signals (linear independence of subspaces). If M_{sp} is rank-deficient, it simply means the spurious subspace $\mathcal{S}_{\text{sp}}$ is lower-dimensional than the number of spurious latents, which our geometric framework handles naturally.

We now prove that this causal model is a specific instance of the geometric assumptions used in GRIT.

Proposition 1 (Causal Model implies Geometric Assumptions)**.** *If the data is generated according to Assumption 4, then Assumption 1 (Oblique Subspace Decomposition) and Assumption 2 (Spurious Geometric Shift) are satisfied.*

Proof. **1. Subspace Decomposition:** Define the subspaces $\mathcal{S}_{\text{core}} = \text{Image}(M_{\text{inv}})$ and $\mathcal{S}_{\text{sp}} = \text{Image}(M_{\text{sp}})$. By the condition in Assumption 4 part 2, $\mathcal{S}_{\text{core}} \cap \mathcal{S}_{\text{sp}} = \{0\}$. Since any observation is $\boldsymbol{z} = M_{\text{inv}}\mathbf{h}_{\text{inv}} + M_{\text{sp}}\mathbf{h}_{\text{sp}}$, it decomposes uniquely into $\boldsymbol{z}_{\text{core}} + \boldsymbol{z}_{\text{sp}}$ with $\boldsymbol{z}_{\text{core}} \in \mathcal{S}_{\text{core}}$ and $\boldsymbol{z}_{\text{sp}} \in \mathcal{S}_{\text{sp}}$.

2. Geometric Shift: A domain shift corresponds to a change in the mechanism generating $\mathbf{h}_{\text{sp}}$. Let $\boldsymbol{z}$ be a sample generated under the training mechanism, and $\boldsymbol{z}_{\text{test}}$ be a sample where $\mathbf{h}_{\text{sp}}$ is replaced by $\mathbf{h}'_{\text{sp}}$ (generated by the shifted mechanism) while keeping the realization of $\mathbf{h}_{\text{inv}}$ fixed. The shift vector is $\Delta = \boldsymbol{z}_{\text{test}} - \boldsymbol{z} = M_{\text{sp}}(\mathbf{h}'_{\text{sp}} - \mathbf{h}_{\text{sp}})$. Since the image of M_{sp} is exactly $\mathcal{S}_{\text{sp}}$, we have $\Delta \in \mathcal{S}_{\text{sp}}$. □

Finally, we formalize the relationship between causal counterfactuals and the invariant pairs used by GRIT.

Proposition 2 (Spurious Counterfactuals are Invariant Pairs)**.** *Let a **spurious counterfactual** be defined as a pair of samples* $(\boldsymbol{z}, \boldsymbol{z}')$ *sharing the same stable latent values* $\mathbf{h}_{inv}$ *(and thus the same conditional distribution of* y*) but realized with different spurious latents* $\mathbf{h}_{sp}$ *and* $\mathbf{h}'_{sp}$*. Under Assumption 4, such a counterfactual pair satisfies the **invariant pair** condition in Assumption 3, specifically* $\boldsymbol{z} - \boldsymbol{z}' \in \mathcal{S}_{sp}$*.*

Proof. Let $\boldsymbol{z} = M_{\text{inv}}\mathbf{h}_{\text{inv}} + M_{\text{sp}}\mathbf{h}_{\text{sp}}$ and $\boldsymbol{z}' = M_{\text{inv}}\mathbf{h}_{\text{inv}} + M_{\text{sp}}\mathbf{h}'_{\text{sp}}$. The difference vector is:

$$\begin{aligned} \boldsymbol{\delta} = \boldsymbol{z} - \boldsymbol{z}' &= (M_{\text{inv}}\mathbf{h}_{\text{inv}} + M_{\text{sp}}\mathbf{h}_{\text{sp}}) - (M_{\text{inv}}\mathbf{h}_{\text{inv}} + M_{\text{sp}}\mathbf{h}'_{\text{sp}}) \qquad (7) \\ &= M_{\text{sp}}(\mathbf{h}_{\text{sp}} - \mathbf{h}'_{\text{sp}}). \qquad (8) \end{aligned}$$

Since $\text{Image}(M_{\text{sp}}) = \mathcal{S}_{\text{sp}}$, the difference vector $\boldsymbol{\delta}$ lies strictly within the spurious subspace. Thus, causally generated spurious counterfactuals provide the precise geometric signal required by GRIT to identify $\mathcal{S}_{\text{sp}}$. □

D GRIT Expanded Discussion

D.1 GRIT Algorithm

We provide the algorithm in Algorithm 1.

Algorithm 1 Noisy Counterfactual-Matching

```
Input: Training Dataset $\mathcal{D}_{\text{train}}$; pair difference matrix $\tilde{\Delta}_{\boldsymbol{x}} \in \mathbb{R}^{d \times k}$; truncated SVD size $r$; epochs $T$; step size $\eta$; batch size $B$.
  // Phase I: Find projection matrix to remove estimated spurious subspace $\tilde{Q}_r$.
  $\tilde{Q}_r, \tilde{\Sigma}_r, \tilde{V}_r^\top = \text{TruncatedSVD}(\tilde{\Delta}_{\boldsymbol{x}}, r)$
  $P = I - \tilde{Q}_r \tilde{Q}_r^\top$
  // Phase II: Gradient descent with preprocessing.
  for $t = 1, 2, \ldots, T$ do
    for sample mini-batch $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^B \subset \mathcal{D}_{\text{train}}$ do
      $\theta \leftarrow \theta - \eta \nabla \frac{1}{B} \sum_{i=1}^B \ell(h(P\phi(\boldsymbol{x}_i); \theta), \boldsymbol{y}_i)$,
    end for
  end for
Output $\theta$
```

D.2 Relationship to MatchDG and Other Matching Methods

GRIT shares the high-level intuition of matching-based domain generalization methods like MatchDG [Mahajan et al., 2021], which also leverage pairs of data points to encourage invariance. However, there are key distinctions in our approach and scope:

- *Geometric vs. Representation Learning:* MatchDG operates by learning a complex non-linear representation $\Phi(\boldsymbol{x})$ such that matched pairs have similar embeddings. In contrast, GRIT operates in a fixed feature space (e.g., a pre-trained model's output) and explicitly identifies a *linear subspace* of variation. This geometric focus allows us to provide rigorous finite-sample guarantees that are difficult to obtain for deep representation learning objectives.
- *Explicit Handling of Noise:* A central contribution of our work is the explicit modeling of *noise* in the invariant pairs (Assumption 3). While general matching objectives typically treat pairs as ground-truth constraints (often leading to over-regularization if pairs are imperfect), GRIT uses dimensionality reduction (the rank-r truncation) to separate the coherent spurious signal from the random measurement noise ξ_i.
- *Scope and Guarantees:* We intentionally tackle a narrower problem setting—linear classifiers under geometric shifts—to achieve a deeper theoretical understanding. While MatchDG is a broad approach for robust learning and proves some high-level theoretic results, GRIT provides a provable mechanism for how and when matching improves robustness, quantifying the exact trade-off between the number of pairs, noise level, and subspace dimension (as shown in Section 4).

Relation to Lagrangian Penalties. It is instructive to unify the GRIT constraint and the standard Lagrangian penalty (used in MatchDG) under a common spectral framework to see exactly how they differ. Consider the eigendecomposition of the uncentered covariance matrix $\tilde{\Sigma}_k = \sum_{i=1}^{d} \tilde{\lambda}_i q_i q_i^\top$, where q_i are the eigenvectors and $\tilde{\lambda}_i$ are the eigenvalues. The standard matching penalty with regularization parameter γ used in MatchDG can be rewritten using this decomposition as:

$$\gamma \sum_{j=1}^{k} \|w^\top \delta_j\|^2 = k\gamma w^\top \tilde{\Sigma}_k w = \sum_{i=1}^{d} (k\gamma\tilde{\lambda}_i)(w^\top q_i)^2$$

This reveals that MatchDG applies a finite penalty proportional to the eigenvalue $\tilde{\lambda}_i$ along each principal direction q_i. In contrast, GRIT's optimization constraint $w^\top \tilde{Q}_r = 0$ forces the projection onto the top-r directions to be exactly zero. This is equivalent to a binary, infinite weighting scheme that puts infinite weight on the first r directions and zero weight on the rest. We can express both methods as (potentially infinite) penalties along principal directions: $\sum_{i=1}^{d} c_i (w^\top q_i)^2$, where $c_i^{\text{MatchDG}} = k\gamma\tilde{\lambda}_i$ and $c_i^{\text{GRIT}} = \infty$ if $i \leq r$ else 0. This comparison highlights the fundamental difference: MatchDG's soft penalty shrinks weights along certain directions (large σ_i), but never strictly enforces invariance ($c_i < \infty$) and incorrectly penalizes core directions if they contain noise ($c_i > 0$ for $i > r$). Implicitly, this penalizes some spurious directions more than others depending on how strongly they are represented by the invariant pairs. GRIT's thresholding (r) and hard constraints ($c_i = \infty$) provide the strict geometric invariance required for robustness against arbitrary shifts in the spurious subspace.

E Extended Discussion of Theory

Remark 5. The finite-sample bound provides explicit geometric intuition into the drivers of robustness:

1. **Test Shift Strength ($\sqrt{\lambda_1^{\text{te}}}$):** The error scales linearly with the spectral norm of the test shift covariance. This implies that even small estimation errors in $\hat{\mathcal{S}}_{\text{sp}}$ can be amplified if the downstream distribution shifts aggressively along spurious directions.

2. **Signal-to-Noise and Diversity:** The term $\frac{\sigma_\xi(\sqrt{\lambda_1(\Sigma_k)}+\sigma_\xi)}{\lambda_{d_{\text{sp}}}(\Sigma_k)}$ governs the quality of the invariant pairs. The error naturally increases with the noise level σ_ξ. Crucially, it depends on the signal covariance via the ratio $\sqrt{\lambda_1(\Sigma_k)}/\lambda_{d_{\text{sp}}}(\Sigma_k)$, suggesting that a "flat" eigenspectrum (diverse coverage of the subspace) is optimal. Furthermore, if the condition number is fixed, the bound scales with $1/\sqrt{\lambda_{d_{\text{sp}}}(\Sigma_k)}$, confirming that increasing the magnitude (variance) of the invariant differences effectively suppresses the relative impact of noise.

Remark 6 (Sample Complexity and Noise Dimensionality)**.** This result explicitly quantifies the benefit of collecting more pairs. Our bound indicates that the shift error decays at a rate of $\mathcal{O}(\sqrt{d/k})$. This confirms that GRIT can reliably recover the spurious subspace given a sufficient number of noisy invariant pairs, provided the signal-to-noise ratio is manageable. We note that the dependence on the ambient dimension d arises from the assumption that the measurement noise $\boldsymbol{\xi}_i$ is isotropic and spans the entire feature space. In settings where the noise is sparse or confined to a lower-dimensional manifold, this dependence could likely be relaxed to scale with the intrinsic dimension or sparsity level of the noise (e.g., via sparse PCA results), though we leave this extension to future work.

Remark 7 (Robustness Depends on Diversity of Pairs)**.** If the collected pairs lack diversity (i.e., the clean vectors $\boldsymbol{v}_i$ do not span the entire test shift subspace range(Q_Δ)), the "Approximation Error" term λ_{s+1} in Lemma 2 will remain large regardless of the estimation accuracy. Thus, robustness is fundamentally limited by the diversity of the provided pairs: we can only be robust to spurious features represented in our data.

Limitations. Our theoretical analysis relies on two simplifying assumptions. First, we assume the selected rank r satisfies $r \geq d_{\text{sp}}$; if $r < d_{\text{sp}}$, an additional bias term (approximation error) is introduced which is not captured by this variance analysis (see Lemma 2. Second, the analysis assumes the noise is well-behaved (sub-Gaussian and isotropic). In adversarial settings or cases with highly heteroscedastic noise aligned with the spurious subspace, the convergence rates may differ and leave to future work.

F Expanded Explanations

F.1 Test-Domain Error Bound for ERM

A direct consequence of the theory allows us to characterize the risk of standard Empirical Risk Minimization (ERM) by setting $r = 0$ for GRIT.

Corollary 3 (Test-Domain Error Bound for ERM)**.** *By setting $r = 0$ (no invariance constraint), GRIT reduces to ERM. Since $s = 0$, the distance term vanishes and the approximation error becomes λ_1. The test-domain error is bounded by:*

$$\mathbb{E}_{\text{test}}\left[\ell_{\text{LL}}(w^\top \boldsymbol{z}_{\text{test}}, y)\right] \leq \mathbb{E}_{\text{train}}[\ell_{\text{LL}}(w^\top \boldsymbol{z}, y)] + \|w\|\sqrt{\lambda_1},$$

and similarly for linear regression.

This explicitly shows how ERM performance degrades as the magnitude of the largest geometric shift (λ_1) increases.

F.2 Availability of invariant pairs

While invariant pairs can be hard to acquire, we argue that it is feasible and practical to obtain in certain scenarios.

For certain applications, obtaining such invariant pairs are both possible and effective. Below we summarizes a range of cases where there could be enough implicit knowledge of spurious correlations to collect them. We further outline these levels in detail below.

Level 3 - Explicit knowledge: In some scientific settings, spurious correlations can be coded as an explicit and mathematical modeling constraint. For example, SchNet [Schütt et al., 2018] builds molecule symmetries and invariance directly into the model structure. This case is straightforward but does not hold in general, so we do not consider it in our work.

Level 2 - Domain expert "soft" knowledge of spurious features: In some applications, domain experts can articulate which features are irrelevant, even if they cannot encode this knowledge as model constraints. For example, an x-ray technician knows that certain medical equipments should not affect their diagnosis of cancer or not [Zech et al., 2018, Oakden-Rayner et al., 2020]. In this case, invariant pairs can be either manually curated (via image editing or generative models) or collected (e.g., by obtaining paired x-rays with and without fluid lines). Simple image augmentation techniques like rotations, flips or color distortions may also fall under this category as they implicitly encode spurious features that are assumed to not affect the downstream tasks (like ColoredMNIST experiment (cf. Section 5)).

Level 1 - Implicit knowledge: At this level, the only differences between domains are assumed to be spurious features because of application-specific knowledge, but domain experts may not know the spurious features a priori. As one example, the differences between data coming from two similar microscopes can be assumed to be spurious since the measurement effects should not affect the underlying physical phenomena of interest. In this case, it is feasible to collect a small number of counterfactual pairs by measuring a small number of samples with both microscopes.

Level 0 - No knowledge: Without any hints or assumptions about spurious features as in levels 1-3, making a model robust to spurious features is likely infeasible. To illustrate, consider a simple causal structure without any knowledge on (latent) spurious features: $\boldsymbol{z}_1 \to y \to \boldsymbol{z}_2$ where only $\boldsymbol{z}_1$ is invariant. Without any knowledge, there is no information to distinguish between invariant feature $\boldsymbol{z}_1$ and spurious feature $\boldsymbol{z}_2$. Moreover, if $\boldsymbol{z}_2$ is more strongly correlated to y or related to y that is easier to extract from inputs $\boldsymbol{x}$, models are prone to shortcut learning [Hermann et al., 2024], the model prediction will rely heavily or nearly solely on $\boldsymbol{z}_2$.

We specifically target the hard and feasible levels 1 and 2, and suggest that in certain cases these pairs could feasibly be collected or created either via manual editing or generative AI tools [Rombach et al., 2022, Betker et al., 2023]. It seems that our method requires additional domain knowledge compared to the standard DG setting. We claim this is an *alternative* form of domain knowledge, as the standard DG setting also requires domain knowledge, as it is encoded in the multi-domain data collection (see Section F.4 for further explanation).

Noticing that this invariant pairs acquisition are still costly, so we ask: *if we have k estimated invariant pairs with noise ε, what kind of robustness guarantee can we get?*

F.3 Discussion on Sample Complexity

We wanted to further clarify the difference between the data requirements of IRM and GRIT by considering the data scaling requirements for a certain number of intervened/spurious features, i.e., $|\mathcal{I}(\mathcal{F}_{\mathcal{E}})|$. For this comparison, we will assume an theoretically ideal setting with an infinite number of samples for each domain, but a finite number of perfect counterfactual pairs, where the true data-generating process is compatible with logistic loss.

IRM [Arjovsky et al., 2019]: As proven in a prior study [Rosenfeld et al., 2020, Corollary 5.2], achieving optimal invariant predictors with IRM requires the number of training domains e to be greater than the number of spurious feature dimensions, i.e., e $> |\mathcal{I}(\mathcal{F}_{\mathcal{E}})|$. And this requirement is true even if there is an infinite number of samples in each of the domains.

Geometric Robustness Invariant Training (GRIT): Our method, GRIT, requires the number of linear independent invariant pairs, k, to be greater than the spurious feature dimension to achieve optimal invariant predictors, as shown in our paper's Corollary 4, i.e., $k \geq |\mathcal{I}(\mathcal{F}_{\mathcal{E}})|$. Linear independence could be satisfied if we assume full rank exogenous noise and soft intervention across domains, which are both common assumptions that are easy to satisfy.

This distinction of the data requirement has significant practical implications, particularly in high-dimensional applications where the spurious feature dimension could be large: IRM would require the number of domains e to scale with the spurious feature dimension, which may be infeasible or too costly in practice (e.g., x-ray machine example). Our GRIT, on the other hand, only requires the number of counterfactual pairs k to scale with the spurious feature dimension, which may be significantly more feasible (e.g., x-ray machine example).

F.4 Discussion of Implicit Domain Knowledge Requirement

Nearly all methods require domain knowledge to some extends. For instance, IRM implicitly uses expert's knowledge based on the specification of the domain labels. In IRM, domain labels are by definition the way of specifying what the predictions should be invariant to. Another example of using expert knowledge is data augmentations (See section 6 in our paper). While it seems that data augmentations don't use any domain knowledge, they actually implicitly use the domain knowledge that "the predictions should not change under this augmentation" (e.g., small rotations or color distortions). While not explicit, these data setups actually incorporate domain knowledge. As a concrete example, take the Rotated MNIST dataset. If the goal is to predict the digit, then rotation can be a domain label. However, if the goal is to predict the rotation, then the digit can be a domain label. Thus, knowledge about the task and what is irrelevant is key for even defining what parts of the data can be considered domain labels. We argue that the expert knowledge required for validating or creating domain counterfactuals is similar in spirit to the expert knowledge for defining domain labels or data augmentations. They are implicit ways of incorporating expert knowledge.

While our data setup differs from standard domain generalization tasks, we argue that expert knowledge is not required to employ the learning algorithm itself, but rather to construct the appropriate dataset. One approach is to use standard domain generalization methods like IRM, which require labeled data from multiple domains. In contrast, we present an alternative approach that requires only possibly noisy invariant pairs and labeled data from a single domain. Note that in our setting, the invariant pairs do not need to be labeled (for example, different x rays of the same patient even if the diagnosis is unknown). In all cases, whether using domain labels as in IRM, data augmentation as in LISA, or counterfactual pairing as in GRIT, our algorithms can be generically applied given the appropriate data constructed through expert knowledge.

G Kernel GRIT: Non-Linear Invariance

G.1 Primal and Expanded Dual Formulation

The primal optimization problem for Kernel GRIT applies a squared L2-norm penalty on the spurious projections, which acts as a continuous relaxation of the truncated SVD penalty used in the linear setting:

$$
\begin{aligned}
&\min_{\mathbf{w},\xi,\psi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i + D\psi \\
\text{subject to}\quad & y_i\mathbf{w}^\top\mathbf{x}_i \geq 1-\xi_i,\ \text{for } i=1,\ldots,N \\
& \xi_i \geq 0,\ \text{for } i=1,\ldots,N \\
& \sum_{j=1}^{K}(\mathbf{w}^\top(\mathbf{z}_j-\mathbf{z}'_j))^2 \leq \psi, \\
& \psi \geq 0
\end{aligned}
$$

By introducing Lagrange multipliers and applying the Woodbury matrix identity, we can fully kernelize this problem. The final expanded dual formulation relies solely on the kernel function $k(\cdot,\cdot)$ evaluated between raw data vectors:

$$
\begin{aligned}
\max_{\alpha,\beta}\quad & \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\Bigg[(\alpha\vec{y})^\top K_{xx}(\alpha\vec{y}) \\
& \qquad - (\alpha\vec{y})^\top(K_{xz}-K_{xz'})\left(\frac{1}{2\beta}\mathbf{I}_K + (K_{zz}-K_{zz'}-K_{z'z}+K_{z'z'})\right)^{-1}(K_{xz}-K_{xz'})^\top(\alpha\vec{y})\Bigg] \\
\text{subject to}\quad & 0\leq\alpha_i\leq C,\quad i=1,\ldots,N \\
& 0\leq\beta\leq D
\end{aligned}
$$

where K_{xx} is the $N\times N$ Gram matrix of the training data, K_{xz} and $K_{xz'}$ map training data to the K invariant pair endpoints, and $K_{zz}, K_{z'z'}, K_{zz'}$ represent the intra- and inter-pair kernels.

G.2 Efficient Objective and Gradient Computation

A naive approach to optimizing the dual objective requires computing the inverse of the $K\times K$ matrix $M_\beta = \left(\frac{1}{2\beta}\mathbf{I}+K_{\Delta z}\right)$ at every iteration, costing $\mathcal{O}(K^3)$. Because β is updated during optimization, this repeated inversion is prohibitively slow for non-trivial K.

To compute the objective and its gradients efficiently, we instead solve the linear system $M_\beta\mathbf{v}=\mathbf{u}$ (where $\mathbf{u}=\sum\alpha_i y_i\mathbf{x}_i$) using the Preconditioned Conjugate Gradient (PCG) method. Crucially, the condition number of M_β deteriorates as $\beta\to D$. We therefore construct a single, global "master" preconditioner based on the worst-case scenario: $P=M_D=\left(\frac{1}{2D}\mathbf{I}+K_{\Delta z}\right)$.

We compute the Cholesky factorization $P=LL^\top$ once offline. During the online L-BFGS optimization loop, the preconditioning step $P\mathbf{z}=\mathbf{r}$ is solved in a highly efficient $\mathcal{O}(K^2)$ via forward/backward substitution. For extreme scaling where K is exceptionally large, Nyström approximations can be applied to $K_{\Delta z}$ to further compress the PCG solve dimension.

G.3 Preliminary Results on ColoredMNIST

We evaluated Kernel GRIT on the ColoredMNIST benchmark using frozen CLIP features without end-to-end backbone fine-tuning. We utilized a configuration with 25% symmetric label noise up to a global class inversion. This specific noise structure inherently caps a purely invariant (digit-shape only) predictor at a maximum observable accuracy of 75%; accuracies climbing significantly above this threshold indicate exploitation of the spurious color correlation.

Table 2: Best configuration per method selected by strictly in-domain validation. ERM overfits to the spurious correlation, whereas Kernel GRIT maintains robust accuracy near the dataset's theoretical limit.

Method	In-Val Acc	Test Acc
ERM	0.8486	0.1011
Linear GRIT	0.8142	0.2213
Kernel GRIT	0.7320	0.7150

As shown in Table 2, the standard ERM linear probe achieves an artificially inflated in-domain validation accuracy (84.86%) by relying heavily on color, leading to catastrophic failure on the out-of-distribution test set (10.11%). Linear GRIT improves robustness but remains constrained by the linear feature space of the frozen CLIP embeddings. By applying the invariant L2 kernel correction, Kernel GRIT correctly aligns with the true invariant signal, achieving a stable 73.20% in-domain accuracy and a highly robust 71.50% test accuracy.

H Semi-Supervised Domain Generalization (SSDG)

H.1 Related Works in SSDG

Domain Generalization (DG) aims to train models that generalize well to unseen target domains [Wang et al., 2022]. Semi-supervised domain generalization (SSDG) addresses the practical scenario where a model must generalize given only a small amount of labeled data and abundant unlabeled data from multiple source domains [Zhou et al., 2023]. Existing SSDG frameworks build on FixMatch [Sohn et al., 2020], which enforces consistency between weakly and strongly augmented views via pseudo-labeling. StyleMatch [Zhou et al., 2023] extends FixMatch with stochastic classifier weights and style augmentation as a third view, while UPCSC [Lee et al., 2025] introduces plug-and-play modules to leverage unconfident unlabeled samples typically discarded by prior methods. Separately, Deep Feature Reweighting (DFR) [Kirichenko et al., 2023] shows that simply retraining the last linear layer of ERM-trained networks suffices to reweight features for improved robustness, motivating post-hoc linear probing as a lightweight alternative to end-to-end fine-tuning.

H.2 Method and Experimental Details

We evaluate our method on four domain generalization benchmarks: PACS, OfficeHome, DigitsDG, and miniDomainNet, using a leave-one-domain-out evaluation protocol with 10 labels per class. We use an ImageNet [Russakovsky et al., 2015] pretrained ResNet50 [He et al., 2016] backbone, fine-tuned with SGD.

For GRIT, invariant pairs are generated from both labeled and unlabeled source-domain samples using four augmentation strategies: strong (RandAugment [Cubuk et al., 2020] and Cutout [DeVries and Taylor, 2017]), style (AdaIN-based style transfer [Huang and Belongie, 2017]), Fourier (amplitude mixing [Xu et al., 2021]), and tangent Fourier (which applies the same augmentation but retains only the tangential component by projecting out the radial direction).

To select the rank r, we sweep all values using scikit-learn's `LogisticRegression` for cost-efficient evaluation, selecting r via an in-domain validation set to strictly avoid target-domain leakage. We additionally report an oracle upper bound where r is tuned on the target domain. Linear probing is performed on the frozen, fine-tuned backbones of each baseline for 10 epochs using SGD with a learning rate of 0.01 and cosine decay.

H.3 SSDG Results

As shown in Table 3, applying GRIT consistently improves over the base methods and vanilla linear probing ($r = 0$). The tangent Fourier augmentation generally achieves the best or near-best performance among the strictly in-domain validated variants (GRIT$_{val}$). These gains are most pronounced on weaker backbones (e.g., ERM, FixMatch) but importantly persist even when applied on top of stronger methods (e.g., StyleMatch, UPCSC), demonstrating that GRIT is highly complementary to existing SSDG improvements.

Table 3: GRIT performance across various SSDG backbones (10 labels per class). **Base**: original method; **LP ($r = 0$)**: linear probing without GRIT; **GRIT$_{val}$**: r selected via source-domain validation; **GRIT$_{oracle}$**: r tuned on target-domain (upper bound). Best non-oracle method is in bold. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Backbone	Variant	PACS	OfficeHome	DigitsDG	miniDomainNet
FixMatch	Base	78.64 ± 1.6	60.95 ± 0.6	75.20 ± 1.7	61.23 ± 0.2
	LP ($r = 0$)	78.10 ± 1.3	61.20 ± 0.4	75.53 ± 1.1	61.28 ± 0.3
	GRIT$_{val;str}$	77.67 ± 1.6	60.97 ± 0.5	75.10 ± 1.3	61.09 ± 0.3
	GRIT$_{val;style}$	77.97 ± 1.6	61.17 ± 0.5	75.46 ± 1.3	61.15 ± 0.4
	GRIT$_{val;fourier}$	77.76 ± 1.0	61.30 ± 0.4	75.75 ± 1.0	61.25 ± 0.4
	GRIT$_{val;tan-fourier}$	**78.24 ± 1.3**	**61.52 ± 0.4**	**75.88 ± 0.9**	**61.31 ± 0.4**
	GRIT$_{oracle}$	*79.31*** ± 1.6*	*61.76** ± 0.4*	*76.13 ± 1.0*	*61.35 ± 0.4*
StyleMatch	Base	83.29 ± 0.9	63.52 ± 0.4	78.40 ± 1.6	60.27 ± 0.4
	LP ($r = 0$)	83.32 ± 1.0	64.24 ± 0.2	78.41 ± 1.4	60.60 ± 0.5
	GRIT$_{val;str}$	82.57 ± 1.5	64.18 ± 0.2	78.04 ± 1.3	**60.77** ± 0.5**
	GRIT$_{val;style}$	82.07 ± 1.5	64.33 ± 0.1	78.00 ± 1.3	60.69 ± 0.5
	GRIT$_{val;fourier}$	82.81 ± 1.5	64.37 ± 0.1	78.61 ± 1.4	60.69 ± 0.5
	GRIT$_{val;tan-fourier}$	**83.55 ± 0.8**	**64.38 ± 0.1**	**78.74 ± 1.6**	60.68 ± 0.5
	GRIT$_{oracle}$	*83.85 ± 1.0*	*64.36 ± 0.2*	*79.29* ± 1.2*	*60.79 ± 0.5*
FixMatch + UPCSC	Base	81.34 ± 1.0	62.28 ± 0.4	78.84 ± 1.1	63.14 ± 0.5
	LP ($r = 0$)	80.96 ± 0.9	62.63 ± 0.5	79.41 ± 1.2	63.12 ± 0.4
	GRIT$_{val;str}$	80.65 ± 0.8	62.52 ± 0.4	79.08 ± 0.9	62.96 ± 0.4
	GRIT$_{val;style}$	**81.00 ± 1.0**	62.80 ± 0.4	79.60 ± 0.8	63.08 ± 0.5
	GRIT$_{val;fourier}$	80.75 ± 1.4	62.80 ± 0.4	**79.69 ± 0.9**	**63.13 ± 0.5**
	GRIT$_{val;tan-fourier}$	80.72 ± 1.1	**62.91 ± 0.3**	79.61 ± 0.9	63.09 ± 0.5
	GRIT$_{oracle}$	*81.19 ± 0.7*	*63.11** ± 0.2*	*80.12** ± 0.8*	*63.13 ± 0.5*
StyleMatch + UPCSC	Base	83.57 ± 0.4	63.75 ± 0.2	79.97 ± 1.6	60.61 ± 0.9
	LP ($r = 0$)	84.10 ± 0.7	64.53 ± 0.4	80.70 ± 1.2	61.24 ± 0.2
	GRIT$_{val;str}$	83.97 ± 0.7	64.60 ± 0.4	80.57 ± 1.4	**61.36** ± 0.2**
	GRIT$_{val;style}$	83.79 ± 0.7	64.62 ± 0.6	80.21 ± 1.0	61.16 ± 0.2
	GRIT$_{val;fourier}$	83.63 ± 0.3	**64.69 ± 0.5**	80.56 ± 1.6	61.30 ± 0.2
	GRIT$_{val;tan-fourier}$	83.68 ± 1.1	64.65 ± 0.5	80.53 ± 1.3	61.30 ± 0.3
	GRIT$_{oracle}$	*84.52 ± 0.7*	*64.67 ± 0.6*	*81.05 ± 1.1*	*61.38** ± 0.2*

I Proofs

I.1 Proof of Lemma 1

Proof of Lemma 1. We construct a set of dummy random variables, denoted as $(\boldsymbol{z}', y')$, and prove that their joint distribution exactly matches the test distribution, i.e., $\mathbb{P}(\boldsymbol{z}', y') = \mathbb{P}_{\text{test}}(\boldsymbol{z}_{\text{test}}, y_{\text{test}})$.

First, let us draw a sample from the training distribution: $(\boldsymbol{z}, y) \sim \mathbb{P}_{\text{train}}(\boldsymbol{z}, y)$. We define the core features and the label of our dummy variables to be identical to the training variables:

$$\boldsymbol{z}'_{\text{core}} := \boldsymbol{z}_{\text{core}} \tag{9}$$

$$y' := y\,. \tag{10}$$

Because $\mathbb{P}_{\text{test}}(\boldsymbol{z}_{\text{test,core}}, y_{\text{test}}) = \mathbb{P}_{\text{train}}(\boldsymbol{z}_{\text{core}}, y)$ by Assumption 2, it immediately follows that the joint distribution of our dummy core features and label matches the test distribution: $\mathbb{P}(\boldsymbol{z}'_{\text{core}}, y') = \mathbb{P}_{\text{test}}(\boldsymbol{z}_{\text{test,core}}, y_{\text{test}})$.

Next, we define the spurious features for our dummy variable. Let $\pi(\boldsymbol{z}_{\text{sp}}, \boldsymbol{z}'_{\text{sp}} \mid \boldsymbol{z}_{\text{core}}, y)$ be a conditional optimal transport coupling between the training spurious conditional $\mathbb{P}_{\text{train}}(\boldsymbol{z}_{\text{sp}} \mid \boldsymbol{z}_{\text{core}}, y)$ and the test spurious conditional $\mathbb{P}_{\text{test}}(\boldsymbol{z}_{\text{test,sp}} \mid \boldsymbol{z}_{\text{test,core}} = \boldsymbol{z}_{\text{core}}, y_{\text{test}} = y)$. We draw the dummy spurious features from the conditional distribution derived from this coupling:

$$\boldsymbol{z}'_{\text{sp}} \sim \pi(\boldsymbol{z}'_{\text{sp}} \mid \boldsymbol{z}_{\text{sp}}, \boldsymbol{z}_{\text{core}}, y)\,. \tag{11}$$

By the definition of a valid coupling, marginalizing over the training variable $\boldsymbol{z}_{\text{sp}}$ ensures that the resulting distribution of $\boldsymbol{z}'_{\text{sp}}$ exactly matches the target marginal of the coupling. Therefore, $\mathbb{P}(\boldsymbol{z}'_{\text{sp}} \mid \boldsymbol{z}'_{\text{core}}, y') = \mathbb{P}_{\text{test}}(\boldsymbol{z}_{\text{test,sp}} \mid \boldsymbol{z}_{\text{test,core}} = \boldsymbol{z}'_{\text{core}}, y_{\text{test}} = y')$.

Now, we can define the full dummy feature vector as $\boldsymbol{z}' := \boldsymbol{z}'_{\text{core}} + \boldsymbol{z}'_{\text{sp}}$. The joint distribution of our dummy variables factors as:

$$\mathbb{P}(\boldsymbol{z}', y') = \mathbb{P}(\boldsymbol{z}'_{\text{core}}, y')\mathbb{P}(\boldsymbol{z}'_{\text{sp}} \mid \boldsymbol{z}'_{\text{core}}, y') . \tag{12}$$

Substituting the established equalities, we get:

$$\begin{aligned} \mathbb{P}(\boldsymbol{z}', y') &= \mathbb{P}_{\text{test}}(\boldsymbol{z}_{\text{test,core}}, y_{\text{test}})\mathbb{P}_{\text{test}}(\boldsymbol{z}_{\text{test,sp}} \mid \boldsymbol{z}_{\text{test,core}}, y_{\text{test}}) \\ &= \mathbb{P}_{\text{test}}(\boldsymbol{z}_{\text{test}}, y_{\text{test}}) . \end{aligned} \tag{13}$$

Thus, $(\boldsymbol{z}', y')$ is equal in distribution to $(\boldsymbol{z}_{\text{test}}, y_{\text{test}})$.

Given this equivalence, we can define the geometric shift Δ between the coupled spurious features:

$$\Delta := \boldsymbol{z}'_{\text{sp}} - \boldsymbol{z}_{\text{sp}} . \tag{14}$$

Because both $\boldsymbol{z}'_{\text{sp}}$ and $\boldsymbol{z}_{\text{sp}}$ reside in $\mathcal{S}_{\text{sp}}$, we have $\Delta \in \mathcal{S}_{\text{sp}}$. Finally, substituting this into the definition of our dummy variable yields:

$$\boldsymbol{z}' = \boldsymbol{z}'_{\text{core}} + \boldsymbol{z}'_{\text{sp}} = \boldsymbol{z}_{\text{core}} + (\boldsymbol{z}_{\text{sp}} + \Delta) = \boldsymbol{z} + \Delta . \tag{15}$$

Since we have proven that $(\boldsymbol{z}', y')$ exactly follows the test distribution, we can formally evaluate any expectations over the test distribution equivalently as expectations over $\boldsymbol{z}' := \boldsymbol{z} + \Delta$ and $y' := y$. □

I.2 Proof of Theorem 1

We provide the proofs for both Linear Regression (Squared Error) and Logistic Regression (Log Loss) using the geometric shift model defined in Assumption 2. First, we provide the theorem with the linear regression with squared error loss and then we will provide the proofs.

Theorem 4 (Test-Domain Error Decomposition)**.** *Let w be any linear predictor satisfying the GRIT constraint $w^\top \tilde{Q}_r = 0$, where $\tilde{Q}_r \in \mathbb{R}^{d \times r}$ is the estimated spurious basis. Let $\tilde{Q}_{r,\perp}$ denote the basis for the orthogonal complement of $\tilde{Q}_r$ (i.e., the subspace the model is allowed to use). Assuming the test distribution is generated via geometric shifts $\Delta(\boldsymbol{z}, y)$ (Assumption 2), the test risk is bounded as follows:*

a) Logistic Regression (Log Loss ℓ_{LL}):

$$\begin{aligned} &\mathbb{E}_{\text{test}}\left[\ell_{\text{LL}}(w^\top \boldsymbol{z}_{\text{test}}, y)\right] \\ &\le \underbrace{\mathbb{E}_{\text{train}}[\ell_{\text{LL}}(w^\top \boldsymbol{z}, y)]}_{\textit{Term I: Training error}} \quad \underbrace{+\|w\|\left\|\tilde{Q}_{r,\perp}^\top Q^{te}\right\|_{\Lambda^{te}}}_{\textit{Term II: Spurious subspace misalignment}} . \end{aligned}$$

b) Linear Regression (Squared Error ℓ_{SE}):

$$\begin{aligned} &\mathbb{E}_{\text{test}}[\ell_{\text{SE}}(w^\top \boldsymbol{z}_{\text{test}}, y)] \\ &\le 2\,\underbrace{\mathbb{E}_{\text{train}}[\ell_{\text{SE}}(w^\top \boldsymbol{z}, y)]}_{\textit{Term I: Training error}} \quad \underbrace{+2\|w\|^2 \left\|\tilde{Q}_{r,\perp}^\top Q^{te}\right\|_{\Lambda^{te}}^2}_{\textit{Term II: Spurious subspace misalignment}} \qquad . \end{aligned}$$

Part 1: Linear Regression (Squared Error). We aim to bound the expected loss on the test distribution, $\mathbb{E}_{\text{test}}[\ell_{\text{SE}}(w^\top \boldsymbol{z}_{\text{test}}, y)]$. Under Assumption 2 and based on *Lemma* 1, an expectation over the test distribution can be written as $\boldsymbol{z}' = \boldsymbol{z} + \Delta(\boldsymbol{z}, y)$, where $\boldsymbol{z} \sim \mathbb{P}_{\text{train}}$ and $\Delta(\boldsymbol{z}, y) \in \mathcal{S}_{\text{sp}}$. Substituting this into the squared error loss:

$$\begin{aligned} \mathbb{E}_{\text{test}}[\ell_{\text{SE}}(w^\top \boldsymbol{z}_{\text{test}}, y_{\text{test}})] &= \mathbb{E}_{(\boldsymbol{z},y)\sim\mathbb{P}_{\text{train}}}\left[(w^\top(\boldsymbol{z} + \Delta(\boldsymbol{z}, y)) - y)^2\right] && (16) \\ &= \mathbb{E}_{(\boldsymbol{z},y)\sim\mathbb{P}_{\text{train}}}\left[((w^\top \boldsymbol{z} - y) + w^\top \Delta(\boldsymbol{z}, y))^2\right] . && (17) \end{aligned}$$

Using the inequality $(a+b)^2 \leq 2a^2 + 2b^2$, we decompose the error:

$$\mathbb{E}_{\text{test}}[\ell_{\text{SE}}] \leq 2 \underbrace{\mathbb{E}_{(\boldsymbol{z},y)\sim\mathbb{P}_{\text{train}}}[(w^\top \boldsymbol{z} - y)^2]}_{\text{Term I: Training Error}} + 2 \underbrace{\mathbb{E}_{(\boldsymbol{z},y)\sim\mathbb{P}_{\text{train}}}[(w^\top \Delta(\boldsymbol{z},y))^2]}_{\text{Shift Term}}. \tag{18}$$

We now analyze the Shift Term. Let $\Sigma_\Delta^{\text{te}} = \mathbb{E}[\Delta(\boldsymbol{z},y)\Delta(\boldsymbol{z},y)^\top]$ be the covariance of the geometric shifts. We can rewrite the term as a quadratic form:

$$\begin{aligned}\mathbb{E}[(w^\top \Delta(\boldsymbol{z},y))^2] &= \mathbb{E}[w^\top \Delta(\boldsymbol{z},y)\Delta(\boldsymbol{z},y)^\top w] &(19)\\ &= w^\top \Sigma_\Delta^{\text{te}} w. &(20)\end{aligned}$$

Recall that GRIT enforces the constraint $w^\top \tilde{Q}_r = 0$. This implies that w lies entirely in the subspace spanned by $\tilde{Q}_{r,\perp}$. Therefore, using the projection $P_\perp = \tilde{Q}_{r,\perp}\tilde{Q}_{r,\perp}^\top$, we have $w = P_\perp w = \tilde{Q}_{r,\perp}\tilde{Q}_{r,\perp}^\top w$.

Substituting the eigendecomposition $\Sigma_\Delta^{\text{te}} = Q^{\text{te}}\Lambda^{\text{te}}(Q^{\text{te}})^\top$:

$$\begin{aligned}w^\top \Sigma_\Delta^{\text{te}} w &= w^\top Q^{\text{te}}\Lambda^{\text{te}}(Q^{\text{te}})^\top w &(21)\\ &= (w^\top \tilde{Q}_{r,\perp}\tilde{Q}_{r,\perp}^\top) Q^{\text{te}}\Lambda^{\text{te}}(Q^{\text{te}})^\top (\tilde{Q}_{r,\perp}\tilde{Q}_{r,\perp}^\top w). &(22)\end{aligned}$$

Let $\boldsymbol{u} = \tilde{Q}_{r,\perp}^\top w$. Note that $\|\boldsymbol{u}\|_2 = \|w\|_2$ because $\tilde{Q}_{r,\perp}$ has orthonormal columns and w is in its range. The expression becomes:

$$\begin{aligned}w^\top \Sigma_\Delta^{\text{te}} w &= \boldsymbol{u}^\top (\tilde{Q}_{r,\perp}^\top Q^{\text{te}})\Lambda^{\text{te}}(Q^{\text{te}})^\top \tilde{Q}_{r,\perp})\boldsymbol{u} &(23)\\ &= \|(\Lambda^{\text{te}})^{1/2}(Q^{\text{te}})^\top \tilde{Q}_{r,\perp}\boldsymbol{u}\|_2^2 &(24)\\ &\leq \|(\Lambda^{\text{te}})^{1/2}(Q^{\text{te}})^\top \tilde{Q}_{r,\perp}\|_2^2 \|\boldsymbol{u}\|_2^2 &(25)\\ &= \|w\|^2 \|(\Lambda^{\text{te}})^{1/2}(Q^{\text{te}})^\top \tilde{Q}_{r,\perp}\|_2^2. &(26)\end{aligned}$$

Using the definition of the Mahalanobis-induced spectral norm $\|A\|_\Lambda := \|A^\top \Lambda^{1/2}\|$, and setting $A = \tilde{Q}_{r,\perp}^\top Q^{\text{te}}$ (implying $A^\top = (Q^{\text{te}})^\top \tilde{Q}_{r,\perp}$), we obtain:

$$w^\top \Sigma_\Delta^{\text{te}} w \leq \|w\|^2 \|\tilde{Q}_{r,\perp}^\top Q^{\text{te}}\|_{\Lambda^{\text{te}}}^2. \tag{27}$$

Combining this with the error decomposition yields the final bound for Linear Regression:

$$\mathbb{E}_{\text{test}}[\ell_{\text{SE}}] \leq 2\mathbb{E}_{\text{train}}[\ell_{\text{SE}}] + 2\|w\|^2 \|\tilde{Q}_{r,\perp}^\top Q^{\text{te}}\|_{\Lambda^{\text{te}}}^2. \tag{28}$$

□

Part 2: Logistic Regression (Log Loss). For Logistic Regression, the loss function $\ell_{\text{LL}}(z,y) = \log(1 + \exp(-yz))$ is 1-Lipschitz with respect to the margin z. Therefore:

$$\ell_{\text{LL}}(w^\top(\boldsymbol{a}+\boldsymbol{b}), y) \leq \ell_{\text{LL}}(w^\top \boldsymbol{a}, y) + |w^\top \boldsymbol{b}|. \tag{29}$$

We can bound the expectation with respect to the training distribution Lemma 1 and the above fact:

$$\begin{aligned}\mathbb{E}_{\text{test}}[\ell_{\text{LL}}(w^\top \boldsymbol{z}_{\text{test}}, y_{\text{test}})] &= \mathbb{E}_{\text{train}}[\ell_{\text{LL}}(w^\top(\boldsymbol{z}+\Delta), y)] &(30)\\ &\leq \mathbb{E}_{\text{train}}[\ell_{\text{LL}}(w^\top \boldsymbol{z}, y)] + \mathbb{E}_{\text{train}}[|w^\top \Delta(\boldsymbol{z},y)|]. &(31)\end{aligned}$$

We bound the second term using Jensen's inequality ($\mathbb{E}[|X|] \leq \sqrt{\mathbb{E}[X^2]}$):

$$\begin{aligned}\mathbb{E}[|w^\top \Delta(\boldsymbol{z},y)|] &\leq \sqrt{\mathbb{E}[(w^\top \Delta(\boldsymbol{z},y))^2]} &(32)\\ &= \sqrt{w^\top \Sigma_\Delta^{\text{te}} w}. &(33)\end{aligned}$$

From the Linear Regression proof (Part 1), we established that:

$$w^\top \Sigma_\Delta^{\text{te}} w \leq \|w\|^2 \|\tilde{Q}_{r,\perp}^\top Q^{\text{te}}\|_{\Lambda^{\text{te}}}^2. \tag{34}$$

Substituting this back yields the final bound for Logistic Regression:

$$\mathbb{E}_{\text{test}}[\ell_{\text{LL}}] \leq \mathbb{E}_{\text{train}}[\ell_{\text{LL}}] + \sqrt{\|w\|^2 \|\tilde{Q}_{r,\perp}^\top Q^{\text{te}}\|_{\Lambda^{\text{te}}}^2} \tag{35}$$

$$= \mathbb{E}_{\text{train}}[\ell_{\text{LL}}] + \|w\| \|\tilde{Q}_{r,\perp}^\top Q^{\text{te}}\|_{\Lambda^{\text{te}}}. \tag{36}$$

□

I.3 Proof of Lemma 2

Proof. We aim to bound the misalignment term $\|\tilde{Q}_{r,\perp}^\top Q^{\text{te}}\|_{\Lambda^{\text{te}}}^2$. Recall that $s = \min\{r, \text{rank}(\Sigma_\Delta^{\text{te}})\}$ and let $d_{\text{te}} = \text{rank}(\Sigma_\Delta^{\text{te}})$. The test shift covariance decomposes as $\Sigma_\Delta^{\text{te}} = Q^{\text{te}} \Lambda^{\text{te}} (Q^{\text{te}})^\top$. We partition the test eigenbasis Q^{te} into two blocks:

- Q_s^{te}: The top s eigenvectors (capturing the dominant shift directions).
- $Q_{>s}^{\text{te}}$: The remaining eigenvectors (indices $s+1$ to d_{te}).

Using the definition of the Mahalanobis-induced spectral norm, we expand the term:

$$\left\|\tilde{Q}_{r,\perp}^\top Q^{\text{te}}\right\|_{\Lambda^{\text{te}}}^2 = \left\|\tilde{Q}_{r,\perp}^\top Q^{\text{te}} (\Lambda^{\text{te}})^{1/2}\right\|_2^2 \tag{37}$$

$$= \left\|\tilde{Q}_{r,\perp}^\top \left[Q_s^{\text{te}} (\Lambda_s^{\text{te}})^{1/2} \mid Q_{>s}^{\text{te}} (\Lambda_{>s}^{\text{te}})^{1/2}\right]\right\|_2^2. \tag{38}$$

Because the blocks are orthogonal, the squared norm splits into two terms:

$$= \underbrace{\left\|\tilde{Q}_{r,\perp}^\top Q_s^{\text{te}} (\Lambda_s^{\text{te}})^{1/2}\right\|_2^2}_{\text{Estimation Error}} + \underbrace{\left\|\tilde{Q}_{r,\perp}^\top Q_{>s}^{\text{te}} (\Lambda_{>s}^{\text{te}})^{1/2}\right\|_2^2}_{\text{Approximation Error}}. \tag{39}$$

1. Bounding the Estimation Error: Using the sub-multiplicative property of the spectral norm:

$$\left\|\tilde{Q}_{r,\perp}^\top Q_s^{\text{te}} (\Lambda_s^{\text{te}})^{1/2}\right\|_2^2 \leq \left\|\tilde{Q}_{r,\perp}^\top Q_s^{\text{te}}\right\|_2^2 \left\|(\Lambda_s^{\text{te}})^{1/2}\right\|_2^2 \tag{40}$$

$$= \lambda_1^{\text{te}} \left\|\tilde{Q}_{r,\perp}^\top Q_s^{\text{te}}\right\|_2^2. \tag{41}$$

Since $s \leq r$, the estimated subspace range$(\tilde{Q}_r)$ contains the subspace range$(\tilde{Q}_s)$ (the top-s estimated components). Consequently, the null space of $\tilde{Q}_r$ is smaller than (or equal to) the null space of $\tilde{Q}_s$, i.e., range$(\tilde{Q}_{r,\perp}) \subseteq$ range$(\tilde{Q}_{s,\perp})$. Therefore, the projection norm is monotonic:

$$\left\|\tilde{Q}_{r,\perp}^\top Q_s^{\text{te}}\right\|_2 \leq \left\|\tilde{Q}_{s,\perp}^\top Q_s^{\text{te}}\right\|_2. \tag{42}$$

By the definition of subspace distance, $\text{dist}(\tilde{Q}_s, Q_s^{\text{te}}) = \|\tilde{Q}_s \tilde{Q}_s^\top - Q_s^{\text{te}} (Q_s^{\text{te}})^\top\|_2 = \|\tilde{Q}_{s,\perp}^\top Q_s^{\text{te}}\|_2$. Thus:

$$\text{Estimation Error} \leq \lambda_1^{\text{te}} \text{dist}^2(\tilde{Q}_s, Q_s^{\text{te}}). \tag{43}$$

2. Bounding the Approximation Error:

$$\left\|\tilde{Q}_{r,\perp}^\top Q_{>s}^{\text{te}} (\Lambda_{>s}^{\text{te}})^{1/2}\right\|_2^2 \leq \left\|\tilde{Q}_{r,\perp}^\top\right\|_2^2 \left\|Q_{>s}^{\text{te}} (\Lambda_{>s}^{\text{te}})^{1/2}\right\|_2^2 \tag{44}$$

$$\leq 1 \cdot \lambda_{s+1}^{\text{te}}. \tag{45}$$

Here, $\lambda_{s+1}^{\text{te}}$ is the largest eigenvalue in the tail block $\Lambda_{>s}^{\text{te}}$. (Note: If $r \geq \text{rank}(\Sigma_\Delta^{\text{te}})$, then $s = \text{rank}(\Sigma_\Delta^{\text{te}})$ and the tail block is empty, so $\lambda_{s+1}^{\text{te}} = 0$, which is consistent with the bound).

Conclusion: Summing the two bounds yields the result:

$$\left\|\tilde{Q}_{r,\perp}^\top Q^{\text{te}}\right\|_{\Lambda^{\text{te}}}^2 \leq \lambda_1^{\text{te}} \text{dist}^2(\tilde{Q}_s, Q_s^{\text{te}}) + \lambda_{s+1}^{\text{te}}. \tag{46}$$

□

I.4 Proof of Finite-Sample Bounds (Theorem 2)

We provide the complete derivation for the finite-sample guarantee of the spurious subspace estimation. To ensure clarity and avoid notation collisions, we adopt the following conventions for this proof:

- $\boldsymbol{\delta}_i$: The observed difference vectors (random vectors).
- η: The failure probability parameter (bounds hold with probability $1-\eta$).
- σ_ξ: The noise parameter.

I.4.1 Previous Results

We first state the two foundational results from high-dimensional probability and perturbation theory that our proof relies upon.

Lemma 3 (Davis-Kahan $\sin\Theta$ Theorem [Davis and Kahan, 1970, Yu et al., 2015])**.** *Let Σ and $\tilde{\Sigma}$ be symmetric matrices in $\mathbb{R}^{d\times d}$ with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d$ and $\tilde{\lambda}_1 \geq \cdots \geq \tilde{\lambda}_d$ respectively. Fix $1 \leq s \leq d-1$ and let Q and $\tilde{Q}$ be the orthonormal matrices whose columns are the top s eigenvectors of Σ and $\tilde{\Sigma}$. If the spectral gap $G := \lambda_s(\Sigma) - \lambda_{s+1}(\Sigma) > 0$ and the perturbation is bounded by $2\|\tilde{\Sigma}-\Sigma\| < G$, then:*

$$\operatorname{dist}(\tilde{Q}, Q) := \|\tilde{Q}\tilde{Q}^\top - QQ^\top\| \leq \frac{2\|\tilde{\Sigma}-\Sigma\|}{\lambda_s(\Sigma)-\lambda_{s+1}(\Sigma)}. \tag{47}$$

Lemma 4 (Matrix Bernstein Inequality (High-Probability Form) [Tropp et al., 2015])**.** *Consider a finite sequence $\{Z_i\}_{i=1}^k$ of independent, random, symmetric matrices of dimension $d \times d$. Assume that each random matrix satisfies $\mathbb{E}[Z_i] = 0$ and $\|Z_i\| \leq L$ almost surely. Let ν^2 be the matrix variance statistic:*

$$\nu^2 := \left\| \sum_{i=1}^k \mathbb{E}[Z_i^2] \right\|. \tag{48}$$

Then, for any $\eta \in (0,1)$, with probability at least $1-\eta$, it holds that

$$\left\| \sum_{i=1}^k Z_i \right\| \leq \sqrt{2\nu^2 \log(d/\eta)} + \frac{2L}{3}\log(d/\eta). \tag{49}$$

This lemma provides a bound on the difference between an empirical sum and its expectation (since $\mathbb{E}[\sum Z_i] = 0$). For sub-Gaussian variables where explicit boundedness L is not strictly defined, L represents the effective threshold of the tail decay with high probability (or one may use sub-Gaussian specific variants).

I.4.2 Geometric Stability Analysis

We first bridge the gap between covariance estimation and subspace estimation. We show that we can connect the distance to the test spurious subspace is bounded by the spurious subspace distance. Then, we show that for our specific noise model, the error in the spurious subspace is linearly bounded by the error in the covariance matrix, inversely scaled by the signal strength.

Lemma 5 (Geometric Stability via Population Noisy Covariance)**.** *Consider the setting of Theorem 4. Assume that $r \geq d_{sp}$. Let $\tilde{\Sigma}_k = \frac{1}{k}\sum_{i=1}^k \boldsymbol{\delta}_i\boldsymbol{\delta}_i^\top$ be the observed empirical noisy covariance matrix. Let $\tilde{\Sigma} := \mathbb{E}[\tilde{\Sigma}_k]$ be the population noisy covariance matrix. Let Q_s and $\tilde{Q}_s$ be the orthonormal matrices spanning the top-s eigenspaces of $\tilde{\Sigma}$ and $\tilde{\Sigma}_k$, respectively. Given these assumptions, we have that:*

$$\|\tilde{Q}_{r,\perp}^\top Q^{te}\|_{\Lambda^{te}} \leq \frac{2\sqrt{\lambda_1^{te}}\|\tilde{\Sigma}_k - \tilde{\Sigma}\|}{\lambda_{d_{sp}}(\Sigma_k)} \tag{50}$$

Proof. We aim to bound the misalignment between the learned invariant subspace and the test-time geometric shifts. The proof proceeds in two main stages: first, we use geometric containment arguments to relate the specific test-domain error to the general estimation error of the spurious subspace; second, we apply perturbation theory to bound this estimation error using the properties of the noisy covariance.

Stage 1: Geometric Containment and Monotonicity. Let $s = \mathrm{rank}(\Sigma_\Delta^{\mathrm{te}})$. By Assumption 2, the test shift lies strictly within the spurious subspace $\mathcal{S}_{\mathrm{sp}}$, so $s \leq d_{\mathrm{sp}}$. We are interested in the term $\|\tilde{Q}_{r,\perp}^\top Q^{\mathrm{te}}\|_{\Lambda^{\mathrm{te}}}$, which quantifies how much of the test shift remains after projection onto the estimated invariant subspace.

First, we utilize the definition of the Mahalanobis-induced spectral norm $\|\cdot\|_{\Lambda^{\mathrm{te}}}$ to separate the magnitude of the shift from the subspace alignment:

$$\begin{aligned}\|\tilde{Q}_{r,\perp}^\top Q^{\mathrm{te}}\|_{\Lambda^{\mathrm{te}}} &= \|\tilde{Q}_{r,\perp}^\top Q^{\mathrm{te}}(\Lambda^{\mathrm{te}})^{1/2}\|_2 \\ &= \|\tilde{Q}_{r,\perp}^\top Q_{d_{\mathrm{sp}}}^{\mathrm{te}}(\Lambda_{d_{\mathrm{sp}}}^{\mathrm{te}})^{1/2}\|_2 \\ &\leq \|\tilde{Q}_{r,\perp}^\top Q_{d_{\mathrm{sp}}}^{\mathrm{te}}\|_2 \, \|(\Lambda_{d_{\mathrm{sp}}}^{\mathrm{te}})^{1/2}\|_2 \\ &= \sqrt{\lambda_1^{\mathrm{te}}}\, \|\tilde{Q}_{r,\perp}^\top Q_{d_{\mathrm{sp}}}^{\mathrm{te}}\|_2 \, . \end{aligned} \tag{51}$$

The second equality is due to the fact that the test shift must be contained in the spurious subspace. Next, we exploit the assumption that the selected rank r is sufficient to cover the spurious subspace ($r \geq d_{\mathrm{sp}}$). This implies the following subspace inclusions.

1. **Estimator Monotonicity:** Since $r \geq d_{\mathrm{sp}}$, the top-r estimated eigenspace contains the top-d_{sp} estimated eigenspace: $\mathrm{range}(\tilde{Q}_{d_{\mathrm{sp}}}) \subseteq \mathrm{range}(\tilde{Q}_r)$. Consequently, the null space of $\tilde{Q}_r$ is contained within the null space of $\tilde{Q}_{d_{\mathrm{sp}}}$, implying $\mathrm{range}(\tilde{Q}_{r,\perp}) \subseteq \mathrm{range}(\tilde{Q}_{d_{\mathrm{sp}},\perp})$. The projection onto a smaller subspace yields a smaller norm, so:
$$\|\tilde{Q}_{r,\perp}^\top Q_{d_{\mathrm{sp}}}^{\mathrm{te}}\| \leq \|\tilde{Q}_{d_{\mathrm{sp}},\perp}^\top Q_{d_{\mathrm{sp}}}^{\mathrm{te}}\|. \tag{52}$$

2. **Test Shift Containment:** Since the test shifts are confined to the spurious subspace (Assumption 2), the basis vectors of the test shift Q^{te} lie within the true spurious subspace $\mathrm{range}(Q_{d_{\mathrm{sp}}})$. Therefore, the projection of Q^{te} onto any subspace is upper-bounded by the projection of the full spurious basis $Q_{d_{\mathrm{sp}}}$ onto that same subspace:
$$\|\tilde{Q}_{d_{\mathrm{sp}},\perp}^\top Q_{d_{\mathrm{sp}}}^{\mathrm{te}}\| \leq \|\tilde{Q}_{d_{\mathrm{sp}},\perp}^\top Q_{d_{\mathrm{sp}}}\|. \tag{53}$$

Combining these inequalities with (51), we link the specific test error to the general subspace estimation error:

$$\|\tilde{Q}_{r,\perp}^\top Q^{\mathrm{te}}\|_{\Lambda^{\mathrm{te}}} \leq \sqrt{\lambda_1^{\mathrm{te}}}\, \mathrm{dist}(\tilde{Q}_{d_{\mathrm{sp}}}, Q_{d_{\mathrm{sp}}}). \tag{54}$$

Stage 2: Perturbation Analysis (Davis-Kahan). We now bound the distance $\mathrm{dist}(\tilde{Q}_{d_{\mathrm{sp}}}, Q_{d_{\mathrm{sp}}})$ using the Davis-Kahan $\sin\Theta$ theorem (Lemma 3). We view the empirical noisy covariance $\tilde{\Sigma}_k$ as a perturbed version of the population noisy covariance $\tilde{\Sigma}$.

1. **Eigenstructure of $\tilde{\Sigma}$:** Recall that the difference vectors are $\boldsymbol{\delta} = \boldsymbol{v} + \boldsymbol{\xi}$, where $\boldsymbol{\xi}$ is isotropic noise ($\mathbb{E}[\boldsymbol{\xi}\boldsymbol{\xi}^\top] = \sigma_\xi^2 I$). The population covariance is $\tilde{\Sigma} = \mathbb{E}[\tilde{\Sigma}_k] = \Sigma_k + \sigma_\xi^2 I$. Adding a multiple of the identity matrix shifts eigenvalues but preserves eigenvectors. Thus, the eigenvectors of $\tilde{\Sigma}$ are identical to those of the clean signal covariance Σ_k. Specifically, the top d_{sp} eigenvectors of $\tilde{\Sigma}$ are exactly $Q_{d_{\mathrm{sp}}}$.

2. **Spectral Gap:** We determine the gap between the d_{sp}-th and $(d_{\mathrm{sp}}+1)$-th eigenvalues of $\tilde{\Sigma}$. Since $\mathrm{rank}(\Sigma_k) = d_{\mathrm{sp}}$, we have $\lambda_{d_{\mathrm{sp}}}(\Sigma_k) > 0$ and $\lambda_{d_{\mathrm{sp}}+1}(\Sigma_k) = 0$. The eigenvalues of $\tilde{\Sigma}$ are:
$$\lambda_{d_{\mathrm{sp}}}(\tilde{\Sigma}) = \lambda_{d_{\mathrm{sp}}}(\Sigma_k) + \sigma_\xi^2, \tag{55}$$
$$\lambda_{d_{\mathrm{sp}}+1}(\tilde{\Sigma}) = \lambda_{d_{\mathrm{sp}}+1}(\Sigma_k) + \sigma_\xi^2 = \sigma_\xi^2. \tag{56}$$
The spectral gap δ is therefore:
$$\delta := \lambda_{d_{\mathrm{sp}}}(\tilde{\Sigma}) - \lambda_{d_{\mathrm{sp}}+1}(\tilde{\Sigma}) = \lambda_{d_{\mathrm{sp}}}(\Sigma_k). \tag{57}$$

3. **Application of Davis-Kahan:** Let $E = \tilde{\Sigma}_k - \tilde{\Sigma}$. Provided that the perturbation is small ($2\|E\| < \delta$), Lemma 3 yields:
$$\mathrm{dist}(\tilde{Q}_{d_{\mathrm{sp}}}, Q_{d_{\mathrm{sp}}}) \leq \frac{2\|\tilde{\Sigma}_k - \tilde{\Sigma}\|}{\delta} = \frac{2\|\tilde{\Sigma}_k - \tilde{\Sigma}\|}{\lambda_{d_{\mathrm{sp}}}(\Sigma_k)}. \tag{58}$$

Substituting this bound back into (54) completes the proof. □

I.4.3 Decomposition and Exact Variance Calculation

Recall that the observed difference vectors are modeled as $\boldsymbol{\delta}_i = \boldsymbol{v}_i + \boldsymbol{\xi}_i$ (Assumption 3), where $\boldsymbol{v}_i$ represents the deterministic signal and $\boldsymbol{\xi}_i$ represents independent, zero-mean isotropic noise with covariance $\mathbb{E}[\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top] = \sigma_\xi^2 I$.

The empirical noisy covariance is given by $\tilde{\Sigma}_k = \frac{1}{k}\sum_{i=1}^k \boldsymbol{\delta}_i\boldsymbol{\delta}_i^\top$. We first derive the population counterpart $\tilde{\Sigma} := \mathbb{E}[\tilde{\Sigma}_k]$. Expanding the expectation:

$$\begin{aligned}\mathbb{E}[\boldsymbol{\delta}_i\boldsymbol{\delta}_i^\top] &= \mathbb{E}[(\boldsymbol{v}_i + \boldsymbol{\xi}_i)(\boldsymbol{v}_i + \boldsymbol{\xi}_i)^\top] \\ &= \boldsymbol{v}_i\boldsymbol{v}_i^\top + \boldsymbol{v}_i\mathbb{E}[\boldsymbol{\xi}_i]^\top + \mathbb{E}[\boldsymbol{\xi}_i]\boldsymbol{v}_i^\top + \mathbb{E}[\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top] \\ &= \boldsymbol{v}_i\boldsymbol{v}_i^\top + \sigma_\xi^2 I. \end{aligned} \tag{59}$$

Thus, the population noisy covariance is $\tilde{\Sigma} = \Sigma_k + \sigma_\xi^2 I$, where $\Sigma_k = \frac{1}{k}\sum_{i=1}^k \boldsymbol{v}_i\boldsymbol{v}_i^\top$ is the clean signal covariance.

We aim to bound the spectral norm of the deviation $E = \tilde{\Sigma}_k - \tilde{\Sigma}$. We decompose this deviation into a signal-noise cross term (E_1) and a pure noise term (E_2):

$$\begin{aligned}\tilde{\Sigma}_k - \tilde{\Sigma} &= \left(\frac{1}{k}\sum_{i=1}^k \boldsymbol{\delta}_i\boldsymbol{\delta}_i^\top\right) - \left(\frac{1}{k}\sum_{i=1}^k \boldsymbol{v}_i\boldsymbol{v}_i^\top + \sigma_\xi^2 I\right) \\ &= \frac{1}{k}\sum_{i=1}^k (\boldsymbol{v}_i + \boldsymbol{\xi}_i)(\boldsymbol{v}_i + \boldsymbol{\xi}_i)^\top - \frac{1}{k}\sum_{i=1}^k \boldsymbol{v}_i\boldsymbol{v}_i^\top - \sigma_\xi^2 I \\ &= \frac{1}{k}\sum_{i=1}^k \left(\boldsymbol{v}_i\boldsymbol{v}_i^\top + \boldsymbol{v}_i\boldsymbol{\xi}_i^\top + \boldsymbol{\xi}_i\boldsymbol{v}_i^\top + \boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top\right) - \frac{1}{k}\sum_{i=1}^k \boldsymbol{v}_i\boldsymbol{v}_i^\top - \sigma_\xi^2 I \\ &= \left(\frac{1}{k}\sum_{i=1}^k \boldsymbol{v}_i\boldsymbol{v}_i^\top\right) + \frac{1}{k}\sum_{i=1}^k (\boldsymbol{v}_i\boldsymbol{\xi}_i^\top + \boldsymbol{\xi}_i\boldsymbol{v}_i^\top) + \left(\frac{1}{k}\sum_{i=1}^k \boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top\right) - \frac{1}{k}\sum_{i=1}^k \boldsymbol{v}_i\boldsymbol{v}_i^\top - \sigma_\xi^2 I \\ &= \underbrace{\frac{1}{k}\sum_{i=1}^k (\boldsymbol{v}_i\boldsymbol{\xi}_i^\top + \boldsymbol{\xi}_i\boldsymbol{v}_i^\top)}_{E_1 \text{ (Cross Term)}} + \underbrace{\left(\frac{1}{k}\sum_{i=1}^k \boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top - \sigma_\xi^2 I\right)}_{E_2 \text{ (Noise Term)}} \end{aligned} \tag{60}$$

To bound $\|E_1\|$ using the Matrix Bernstein inequality (Lemma 4), we must calculate the matrix variance statistic ν^2. Let $Z_i = \frac{1}{k}(\boldsymbol{v}_i\boldsymbol{\xi}_i^\top + \boldsymbol{\xi}_i\boldsymbol{v}_i^\top)$. Note that $\mathbb{E}[Z_i] = 0$ because ξ_i are zero mean noise. The variance parameter is defined as $\nu^2 := \|\sum_{i=1}^k \mathbb{E}[Z_i^2]\|$.

First, we compute the expected square of the unscaled term $X_i = kZ_i = \boldsymbol{v}_i\boldsymbol{\xi}_i^\top + \boldsymbol{\xi}_i\boldsymbol{v}_i^\top$:

$$\begin{aligned} X_i^2 &= (\boldsymbol{v}_i\boldsymbol{\xi}_i^\top + \boldsymbol{\xi}_i\boldsymbol{v}_i^\top)(\boldsymbol{v}_i\boldsymbol{\xi}_i^\top + \boldsymbol{\xi}_i\boldsymbol{v}_i^\top) \\ &= \boldsymbol{v}_i\boldsymbol{\xi}_i^\top\boldsymbol{v}_i\boldsymbol{\xi}_i^\top + \boldsymbol{v}_i\boldsymbol{\xi}_i^\top\boldsymbol{\xi}_i\boldsymbol{v}_i^\top + \boldsymbol{\xi}_i\boldsymbol{v}_i^\top\boldsymbol{v}_i\boldsymbol{\xi}_i^\top + \boldsymbol{\xi}_i\boldsymbol{v}_i^\top\boldsymbol{\xi}_i\boldsymbol{v}_i^\top \\ &= (\boldsymbol{\xi}_i^\top\boldsymbol{v}_i)\boldsymbol{v}_i\boldsymbol{\xi}_i^\top + \|\boldsymbol{\xi}_i\|^2\boldsymbol{v}_i\boldsymbol{v}_i^\top + \|\boldsymbol{v}_i\|^2\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top + (\boldsymbol{\xi}_i^\top\boldsymbol{v}_i)\boldsymbol{\xi}_i\boldsymbol{v}_i^\top. \end{aligned} \tag{61}$$

We evaluate the expectation of each term term-by-term, relying on the isotropic property $\mathbb{E}[\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top] = \sigma_\xi^2 I$:

1. $\mathbb{E}[(\boldsymbol{\xi}_i^\top\boldsymbol{v}_i)\boldsymbol{v}_i\boldsymbol{\xi}_i^\top] = \mathbb{E}[\boldsymbol{v}_i(\boldsymbol{\xi}_i^\top\boldsymbol{v}_i)\boldsymbol{\xi}_i^\top] = \mathbb{E}[\boldsymbol{v}_i(\boldsymbol{v}_i^\top\boldsymbol{\xi}_i)\boldsymbol{\xi}_i^\top] = \mathbb{E}[\boldsymbol{v}_i\boldsymbol{v}_i^\top(\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top)] = \boldsymbol{v}_i\boldsymbol{v}_i^\top\mathbb{E}[\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top] = \sigma_\xi^2\boldsymbol{v}_i\boldsymbol{v}_i^\top$
2. $\mathbb{E}[\|\boldsymbol{\xi}_i\|^2\boldsymbol{v}_i\boldsymbol{v}_i^\top] = \mathbb{E}[\mathrm{Tr}(\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top)]\boldsymbol{v}_i\boldsymbol{v}_i^\top = \mathrm{Tr}(\sigma_\xi^2 I)\boldsymbol{v}_i\boldsymbol{v}_i^\top = d\sigma_\xi^2\boldsymbol{v}_i\boldsymbol{v}_i^\top$.
3. $\mathbb{E}[\|\boldsymbol{v}_i\|^2\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top] = \|\boldsymbol{v}_i\|^2\sigma_\xi^2 I$.
4. $\mathbb{E}[(\boldsymbol{\xi}_i^\top\boldsymbol{v}_i)\boldsymbol{\xi}_i\boldsymbol{v}_i^\top]$. This is the transpose of Term 1, so the expectation is $\sigma_\xi^2\boldsymbol{v}_i\boldsymbol{v}_i^\top$.

Summing these expectations yields:

$$\mathbb{E}[X_i^2] = \sigma_\xi^2\left((d+2)\boldsymbol{v}_i\boldsymbol{v}_i^\top + \|\boldsymbol{v}_i\|^2 I\right). \tag{62}$$

We now substitute back into the definition of ν^2. Since $Z_i = X_i/k$, we have $\mathbb{E}[Z_i^2] = \frac{1}{k^2}\mathbb{E}[X_i^2]$. Summing over $i = 1 \ldots k$:

$$\begin{aligned}\nu^2 &= \left\| \sum_{i=1}^{k} \frac{1}{k^2} \sigma_\xi^2 \left((d+2)\boldsymbol{v}_i \boldsymbol{v}_i^\top + \|\boldsymbol{v}_i\|^2 I \right) \right\| \\ &= \frac{\sigma_\xi^2}{k} \left\| (d+2) \left(\frac{1}{k} \sum_{i=1}^{k} \boldsymbol{v}_i \boldsymbol{v}_i^\top \right) + \left(\frac{1}{k} \sum_{i=1}^{k} \|\boldsymbol{v}_i\|^2 \right) I \right\|. \end{aligned} \tag{63}$$

Using the definition $\Sigma_k = \frac{1}{k}\sum \boldsymbol{v}_i \boldsymbol{v}_i^\top$ and the identity $\frac{1}{k}\sum \|\boldsymbol{v}_i\|^2 = \mathrm{Tr}(\Sigma_k)$, we obtain:

$$\nu^2 = \frac{\sigma_\xi^2}{k} \left\| (d+2)\Sigma_k + \mathrm{Tr}(\Sigma_k) I \right\|. \tag{64}$$

Since Σ_k is positive semi-definite, the spectral norm is the largest eigenvalue. We bound $\mathrm{Tr}(\Sigma_k) \leq d\|\Sigma_k\|$. This yields the final variance bound for the cross-term:

$$\begin{aligned}\nu^2 &\leq \frac{\sigma_\xi^2}{k} \left((d+2)\|\Sigma_k\| + \mathrm{Tr}(\Sigma_k)\|I\| \right) \\ &\leq \frac{\sigma_\xi^2}{k} \left((d+2)\|\Sigma_k\| + d\|\Sigma_k\| \right) \\ &= \frac{2(d+1)\sigma_\xi^2 \|\Sigma_k\|}{k}. \end{aligned} \tag{65}$$

This variance parameter explicitly captures the interplay between the signal geometry (Σ_k), the noise level (σ_ξ), and the dimensionality (d), and will be used to bound $\|E_1\|$. The pure noise term $\|E_2\|$ follows standard covariance concentration bounds (e.g., [Vershynin, 2009]).

I.4.4 Applying Concentration Inequalities

We now bound the norms of the cross term E_1 and the noise term E_2 using high-probability concentration inequalities. We fix a failure probability $\eta \in (0, 1)$ and allocate the error budget equally between the two terms.

Bounding the Cross Term (E_1): Recall from Step 1 that the variance parameter for E_1 is bounded by $\nu^2 \leq \frac{2(d+1)\sigma_\xi^2\|\Sigma_k\|}{k}$. We apply the Matrix Bernstein inequality (Lemma 4). For sub-Gaussian noise $\boldsymbol{\xi}_i$, the operator norm of the summands is not strictly bounded, but standard sub-Gaussian analysis allows us to control the effective bound L. In the regime where k is sufficiently large (signal dominates noise), the variance term dominates the bound. Thus, with probability at least $1 - \eta/2$:

$$\begin{aligned}\|E_1\| &\leq \sqrt{2\nu^2 \log\left(\frac{2d}{\eta}\right)} \\ &\leq \sqrt{\frac{4(d+1)\sigma_\xi^2 \|\Sigma_k\| \log(2d/\eta)}{k}} \\ &\leq 2\sigma_\xi \sqrt{\|\Sigma_k\|} \sqrt{\frac{(d+1)\log(2d/\eta)}{k}}. \end{aligned} \tag{66}$$

Bounding the Noise Term (E_2): The term $E_2 = \frac{1}{k}\sum_{i=1}^{k} \boldsymbol{\xi}_i \boldsymbol{\xi}_i^\top - \sigma_\xi^2 I$ represents the deviation of the sample covariance of isotropic sub-Gaussian vectors from their mean. We utilize the standard result for covariance estimation of sub-Gaussian distributions (see Vershynin [2009, Theorem 4.6.1]). There exists an absolute constant C_1 such that with probability at least $1 - \eta/2$:

$$\|E_2\| \leq C_1 \sigma_\xi^2 \left(\sqrt{\frac{d}{k}} + \frac{d}{k} \right) \sqrt{\log(2/\eta)}. \tag{67}$$

I.4.5 Union Bound and Final Estimation Error

We now combine the exact bounds derived for $\|E_1\|$ and $\|E_2\|$ to derive the total estimation error.

Total Covariance Error via Union Bound: Let $\mathcal{A}$ be the event that inequality (66) holds (the bound on the cross term), and $\mathcal{B}$ be the event that inequality (67) holds (the bound on the noise term). By the union bound, the probability that both hold simultaneously is at least $1-(\eta/2+\eta/2)=1-\eta$.

Conditional on this event, the total perturbation magnitude $\|E\|=\|\tilde{\Sigma}_k-\tilde{\Sigma}\|$ is bounded by:

$$\begin{aligned}\|E\| &\le \|E_1\|+\|E_2\|\\ &\le \underbrace{2\sigma_\xi\sqrt{\|\Sigma_k\|}\sqrt{\frac{(d+1)\log(2d/\eta)}{k}}}_{\text{Cross Term Bound}}+\underbrace{C_1\sigma_\xi^2\left(\sqrt{\frac{d}{k}}+\frac{d}{k}\right)\sqrt{\log(2/\eta)}}_{\text{Noise Term Bound}}.\end{aligned}\tag{68}$$

Subspace Distance Bound (Step 3) Substituting the concentration bounds into the geometric stability lemma yields:

$$\operatorname{dist}(\tilde{Q}_s,Q_s)\le\frac{4\sigma_\xi\sqrt{\lambda_1(\Sigma_k)}}{\lambda_{d_{\text{sp}}}(\Sigma_k)}\sqrt{\frac{(d+1)\log(2d/\eta)}{k}}+\frac{2C_1\sigma_\xi^2}{\lambda_{d_{\text{sp}}}(\Sigma_k)}\left(\sqrt{\frac{d}{k}}+\frac{d}{k}\right)\sqrt{\log(2/\eta)}.\tag{69}$$

Case I: General Regime. We use the facts that $\log(2/\eta)\le\log(2d/\eta)$ and $\sqrt{d+1}\le\sqrt{2d}$ to factor out the logarithmic term and group the terms by their decay rate ($\sqrt{d/k}$ vs d/k). We define a universal constant $C'=\max(2\sqrt{2},C_1)$:

$$\operatorname{dist}(\tilde{Q}_s,Q_s)\le\frac{C'\sigma_\xi\sqrt{\log(2d/\eta)}}{\lambda_{d_{\text{sp}}}(\Sigma_k)}\left[\left(2\sqrt{\lambda_1(\Sigma_k)}+\sigma_\xi\right)\sqrt{\frac{d}{k}}+\sigma_\xi\left(\frac{d}{k}\right)\right].\tag{70}$$

Case 2: Asymptotic Regime ($k\ge d$). Assuming sufficient samples ($k\ge d$), the linear term $\frac{d}{k}$ is dominated by the square root term $\sqrt{\frac{d}{k}}$. We factor out the common rate $\sqrt{\frac{d}{k}}$ and merge constants into C:

$$\operatorname{dist}(\tilde{Q}_s,Q_s)\le\frac{C\sigma_\xi(\sqrt{\lambda_1(\Sigma_k)}+\sigma_\xi)}{\lambda_{d_{\text{sp}}}(\Sigma_k)}\sqrt{\frac{d\log(2d/\eta)}{k}}.\tag{71}$$

Step 4: Combining with Generalization Bound We now substitute the finite-sample estimation bound into the error decomposition from Theorem 4. We assume the user has selected a sufficient rank $r\ge d_{\text{sp}}$ such that the approximation error vanishes ($\lambda_{s+1}^{\text{te}}=0$).

For Logistic Regression: From Theorem 4, the generalization error is bounded by the training error plus the misalignment penalty $\|w\|\|\tilde{Q}_{r,\perp}^\top Q^{\text{te}}\|_{\Lambda^{\text{te}}}$. Under the rank assumption, this simplifies to $\|w\|\sqrt{\lambda_1^{\text{te}}}\operatorname{dist}(\tilde{Q}_s,Q_s)$. Substituting the result from Step 3:

$$\mathbb{E}_{\text{test}}[\ell(w^\top\boldsymbol{z},y)]\le\mathbb{E}_{\text{train}}[\ell(w^\top\boldsymbol{z},y)]+\|w\|\sqrt{\lambda_1^{\text{te}}}\left[\frac{C\sigma_\xi(\sqrt{\lambda_1(\Sigma_k)}+\sigma_\xi)}{\lambda_{d_{\text{sp}}}(\Sigma_k)}\sqrt{\frac{d\log(2d/\eta)}{k}}\right].\tag{72}$$

For Linear Regression: From Theorem 4, the misalignment term is squared: $2\|w\|^2\|\tilde{Q}_{r,\perp}^\top Q^{\text{te}}\|_{\Lambda^{\text{te}}}^2\le 2\|w\|^2\lambda_1^{\text{te}}\operatorname{dist}^2(\tilde{Q}_s,Q_s)$. Substituting the result from Step 3:

$$\mathbb{E}_{\text{test}}[\ell(w^\top\boldsymbol{z},y)]\le2\mathbb{E}_{\text{train}}[\ell(w^\top\boldsymbol{z},y)]+2\|w\|^2\lambda_1^{\text{te}}\left[\frac{C\sigma_\xi(\sqrt{\lambda_1(\Sigma_k)}+\sigma_\xi)}{\lambda_{d_{\text{sp}}}(\Sigma_k)}\sqrt{\frac{d\log(2d/\eta)}{k}}\right]^2.\tag{73}$$

This confirms that for both models, the excess test risk decays with the number of pairs k, scaling as $O(k^{-1/2})$ for logistic regression and $O(k^{-1})$ for linear regression. □

J Detailed Experiment Setup and Hyperparameters

We provide a detailed description of our experiments setup and hyperparameter selection.

J.1 Synthetic Experiments Extended Details

We generate data where the true spurious subspace $\mathcal{S}_{\text{sp}}$ has dimension $d_{\text{sp}} = 20$. The results validate our finite-sample analysis:

- **Convergence with k (Theorem 2):** As shown in Figure 1a, GRIT achieves oracle-level accuracy (training directly on test distribution) when the number of invariant pairs k exceeds the spurious dimension d_{sp}. Even with noisy pairs ($\varepsilon > 0$), performance improves monotonically with k, confirming the theoretical convergence rate of $O(\sqrt{d/k})$.
- **Trade-off effect of r:** In Figure 1b, we fix $k = 100$ and vary the rank r of the estimated subspace $\hat{\mathcal{S}}_{\text{sp}}$. For low noise ($\varepsilon = 0, 1$), the test accuracy peaks exactly at the true spurious dimension $r \approx d_{\text{sp}}$, empirically validating the decomposition in Lemma 2. If $r < d_{\text{sp}}$, the approximation error (tail eigenvalues λ_{s+1}) dominates; if $r \gg d_{\text{sp}}$, the in-domain error increases as we discard informative features. Under higher noise ($\varepsilon = 5, 10$), a slightly larger r is required to capture the scattered spurious signal, illustrating the robustness-accuracy trade-off.

J.2 Detailed Construction of Waterbirds-CFs

The original Waterbirds dataset [Sagawa et al., 2019] combines bird images from the CUB dataset [Wah et al., 2011] with background images from the Places dataset [Zhou et al., 2017]. The task is to classify whether a given image depicts a waterbird ($y = 1$) or a landbird ($y = 0$). Waterbirds include seabirds (albatross, auklet, cormorant, frigatebird, fulmar, gull, jaeger, kittiwake, pelican, puffin, and tern) and waterfowl (gadwall, grebe, mallard, merganser, guillemot, and Pacific loon).

In the dataset, the invariant features are represented by the bird segments, while the spurious features are the backgrounds. In the training set, the background is highly correlated with the bird species: 95% of waterbirds appear in a water background (ocean or natural lake), and similarly, 95% of landbirds are shown against a land background (bamboo or broadleaf forest). The remaining 5% consist of counterfactual samples, which are random samples from the majority group. Counterfactual pairs share the same bird segment but differ in the background. The validation and test sets are identical to the original Waterbirds dataset, meaning the conditional distribution of the background given either waterbirds or landbirds is 50%.

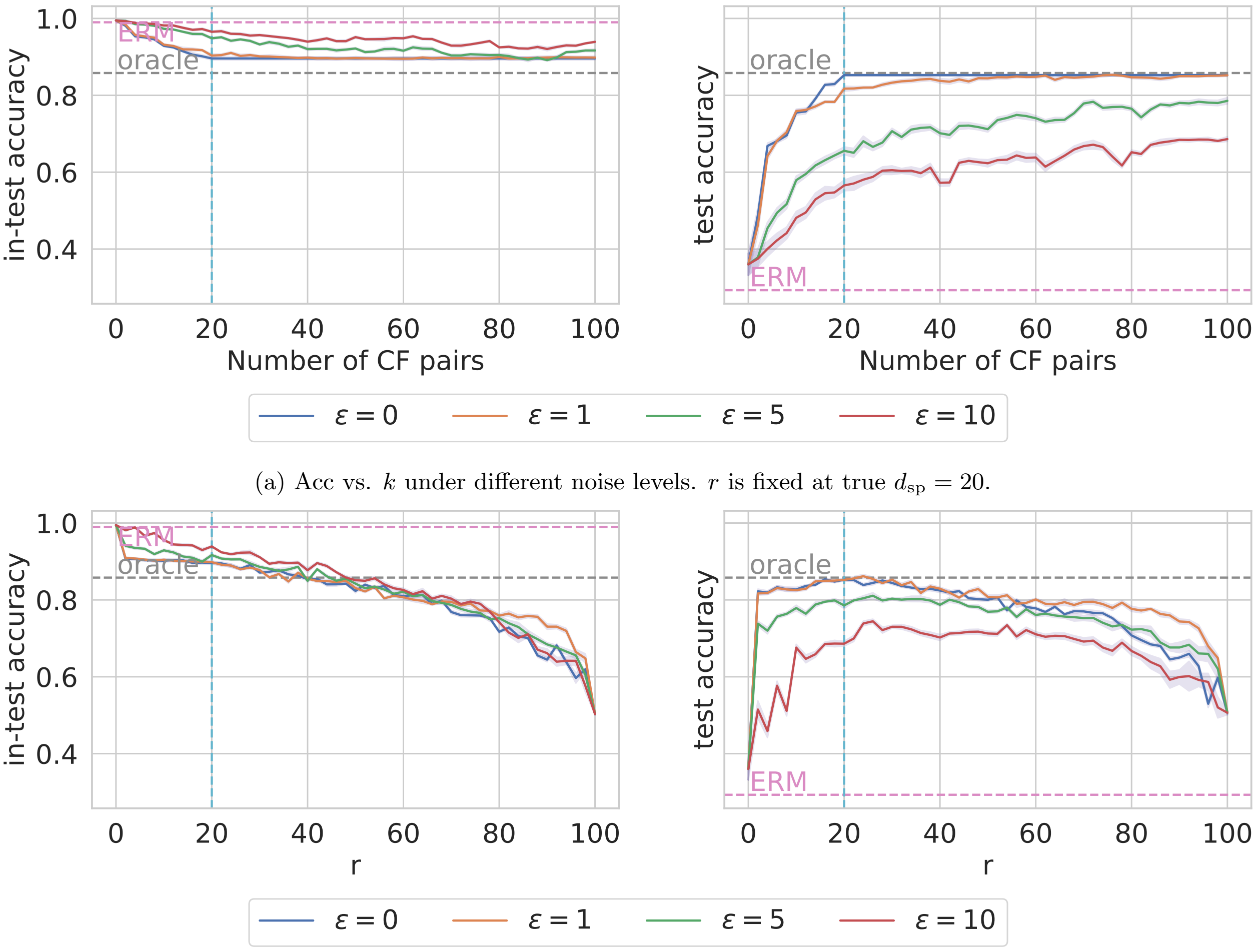
In summary, the modification made between our Waterbirds-CF dataset and the original waterbirds dataset only pertains to the minority groups in the training set. We randomly sampled 184 landbirds and 56 waterbirds from the majority group and replaced the backgrounds of these samples to generate the minority group counterfactuals. See Figure 2 for the illustration. We then applied ERM on both these two datasets to investigate the changing in the training distribution. We include the convergence curve in Section K.2.

J.3 Hyperparameter Selection

In this section, we present all the hyperparameters and evaluation used in our experiments to ensure reproducibility.

Synthetic Dataset Invariant features are sampled from a standard normal distribution, i.e., $\boldsymbol{z}_{\text{inv}} \sim \mathcal{N}(0, I)$, The observation function g_y is linear, with parameter $\theta_y \sim \mathcal{N}(0, \sigma I)$, and the label $y = \text{sign}(\boldsymbol{z}_{\text{inv}}\theta_y)$. The spurious features is correlated to the label y, i.e., $\boldsymbol{z}_{\text{spu}} \sim \mathcal{N}\left(\frac{y}{|\mathcal{I}(\mathcal{F}_{\mathcal{E}})|}, \sigma_s I\right)$ where σ_s varies across domains. The observation function g_x is a random orthonormal matrix. The dimension of $\boldsymbol{z}$ and $\boldsymbol{x}$ are both 100, i.e., $m = d = 100$.

We run 100 iterations of gradient descent using binary cross-entropy loss. We use the Adam optimizer [Kingma, 2014] with a learning rate of 0.01 for ERM, IRM, and GRIT. The Lagrange multiplier for both IRM and GRIT is set to $\lambda = 1000$ selected through grid search.



(a) Acc vs. k under different noise levels. r is fixed at true $d_{\text{sp}} = 20$.

(b) Acc vs. r under different noise levels with fixed $k = 100$.

Figure 1: Results on the synthetic dataset validating GRIT theoretical bounds. (a) **Sample Complexity:** Test accuracy improves with the number of pairs k, consistent with the $O(1/\sqrt{k})$ rate in Theorem 2. The vertical line denotes the true spurious dimension $d_{\text{sp}} = 20$. (b) **Rank Trade-off:** Varying the estimated rank r reveals the trade-off between removing spurious features and preserving core features, as predicted by Theorem 4. Shaded regions indicate standard deviation over 10 runs.

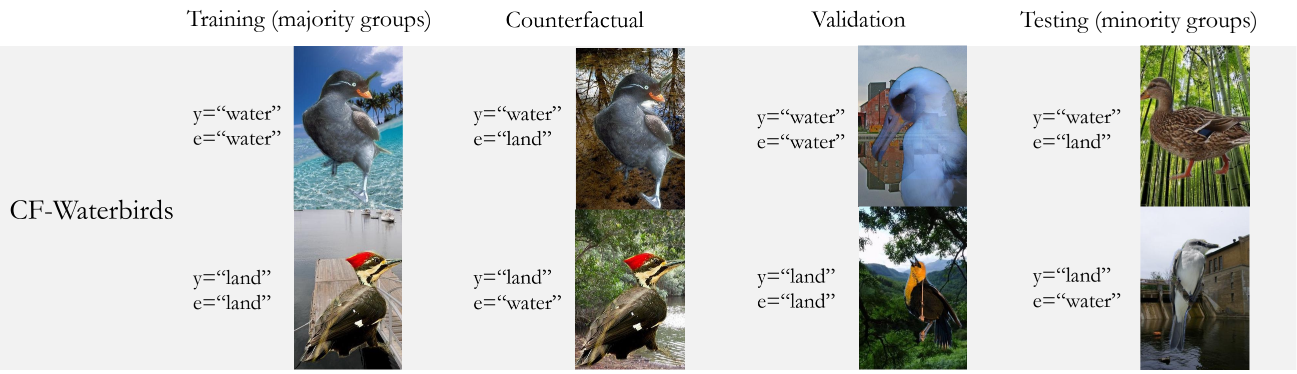


Figure 2: Illustration of the training samples, counterfactual pairs, validation examples, and test examples for the Waterbirds-CF dataset. The training set (majority groups), validation set, and test set are identical to those in the original Waterbirds dataset. Counterfactual pairs feature the same birds from the training set but with different backgrounds.

ColoredMNIST We use the Adam optimizer with a learning rate of 0.001 and a weight decay of 10^{-4}. The model is trained with a batch size of 256 for 40 epochs. We tune the hyperparameter r in the range [2, 24] using 256 counterfactual pairs.

Waterbirds-CF We use the Adam optimizer with a learning rate of 0.001 and a weight decay of 10^{-4}. The model is trained with a batch size of 256 for 100 epochs. We tune the hyperparameter r in the range [2, 24].

PACS We use the Adam optimizer with a learning rate of 0.01 and a weight decay of 10^{-4}. The model is trained with a batch size of 256 for 100 epochs. We tune the hyperparameter r in the range [2, 24].

Details on hyperparameter tuning and baseline method selection can be found in the code repository.

K Additional Experiments

Table 4: Main Results on synthetic dataset with $\varepsilon = 5$. GRIT outperforms baselines accuracy across datasets.

	In-domain	DG
ERM	0.99	0.31
IRM	0.98	0.29
REx	0.98	0.31
GroupDRO	0.99	0.35
Fish	0.99	0.37
SWAD	0.97	0.3
LISA	0.98	0.33
MatchDG	0.89	0.49
GRIT	0.86	0.65

In this section, we include more experiments results. We further illustrate the effectiveness of our method, as well as the hyperparameters sensitivity.

Table 5: Main Results on ColoredMNIST

	in-domain validation		oracle validation	
	in acc	test acc	in acc	test acc
ERM (CLIP)	0.852	0.093	0.753	0.253
IRM	0.799	0.118	0.724	0.469
REx	0.797	0.121	0.691	0.664
GroupDRO	0.798	0.127	0.786	0.201
Fish	0.798	0.118	0.495	0.486
SWAD	0.800	0.113	0.501	0.505
LISA	0.705	**0.693**	0.705	0.693
MatchDG w. random	0.799	0.120	0.511	0.512
MatchDG w. 1NN	0.789	0.217	0.728	0.662
MatchDG w. clean	0.793	0.181	0.742	0.672
GRIT w. random	0.794	0.176	0.680	0.706
GRIT w. 1NN	0.736	0.649	0.711	0.707
GRIT w. clean	0.740	**0.693**	0.727	**0.714**
random guess	0.500	0.500	0.500	0.500
ERM oracle	0.735	0.730	0.735	0.730
theory oracle	0.750	0.750	0.750	0.750

Table 6: Main Results on Waterbirds-CF

	In-domain Validation		Oracle Validation	
	in acc	wg acc	in acc	wg acc
ERM (CLIP)	0.885	0.781	0.882	0.800
ERM+UW	0.889	0.795	0.882	0.829
IRM	0.838	0.707	0.820	0.767
REx	0.891	0.617	0.878	0.729
GroupDRO	0.906	0.684	0.896	0.827
Fish	0.900	0.744	0.869	0.805
LISA	0.904	0.722	0.876	0.812
MatchDG w. random	0.793	0.009	0.785	0.149
MatchDG w. 1NN	0.886	0.411	0.886	0.411
MatchDG w. estimated CF	0.906	0.536	0.896	0.651
GRIT w. random	0.804	0.269	0.804	0.269
GRIT w. 1NN	0.892	0.521	0.882	0.560
GRIT w. estimated CF	0.864	**0.812**	0.854	**0.860**

Table 7: Main Results on PACS

	In-domain Validation					Oracle Validation				
	A	C	P	S	Avg	A	C	P	S	Avg
ERM (CLIP)	0.924	0.968	0.996	0.859	0.937	0.924	0.968	0.996	0.859	0.937
IRM	0.938	0.976	0.996	0.840	0.938	0.941	0.976	0.996	0.845	0.940
REx	0.953	0.963	0.993	0.836	0.936	0.953	0.975	0.996	0.845	0.942
GroupDRO	0.903	0.963	0.996	0.873	0.934	0.941	0.975	0.996	0.843	0.939
Fish	0.936	0.973	0.996	0.837	0.936	0.936	0.973	0.996	0.837	0.936
SWAD	0.941	0.976	0.996	0.838	0.938	0.941	0.977	0.996	0.838	0.938
LISA	0.926	**0.978**	0.997	0.848	0.937	0.940	**0.983**	0.997	0.864	0.946
MatchDG w. rand.	0.412	0.509	0.316	0.749	0.497	0.454	0.509	0.358	0.749	0.518
MatchDG w. 1NN.	**0.964**	0.971	0.995	0.880	**0.953**	**0.964**	0.973	0.996	**0.887**	**0.955**
GRIT w. rand.	0.591	0.609	0.577	0.833	0.653	0.592	0.625	0.583	0.843	0.661
GRIT w. 1NN.	0.957	0.974	**0.998**	**0.882**	**0.953**	**0.964**	0.974	**0.998**	0.885	**0.955**

K.1 Ablation Study

Sensitivity on truncated SVD parameter r. We empirically evaluate the trade-off effect of the hyper-parameter r on model performance during linear probing on the Waterbirds-CF dataset (cf. Figure 3), thus validating Theorem 4 comment **(iii)**: accuracy trade-off induced by r. This pattern reflects the model's shifting reliance from spurious to invariant features: when r is too small, spurious correlations dominate, resulting in high in-domain but low worst-group performance. As r increases and suppresses these spurious features, worst-group accuracy improves. However, beyond a certain point, further increases in r begin to remove invariant features as well, leading to a decline in both metrics.

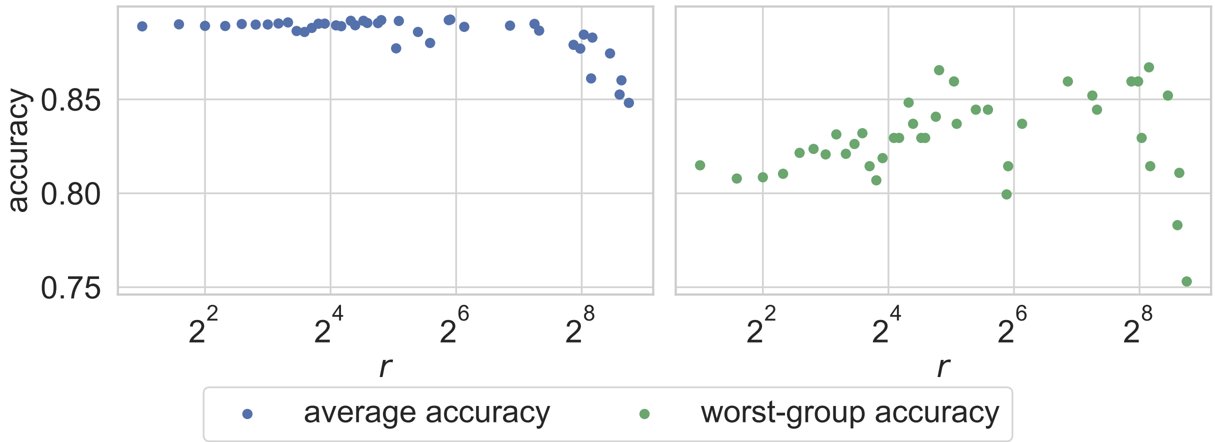


Figure 3: In-domain test and worst-group accuracy with changing hyperparameter r. In-domain accuracy remains stable for small values of r, but starts to drop at $r \approx 128$. Worst-group accuracy first increases then decreases as r grows.

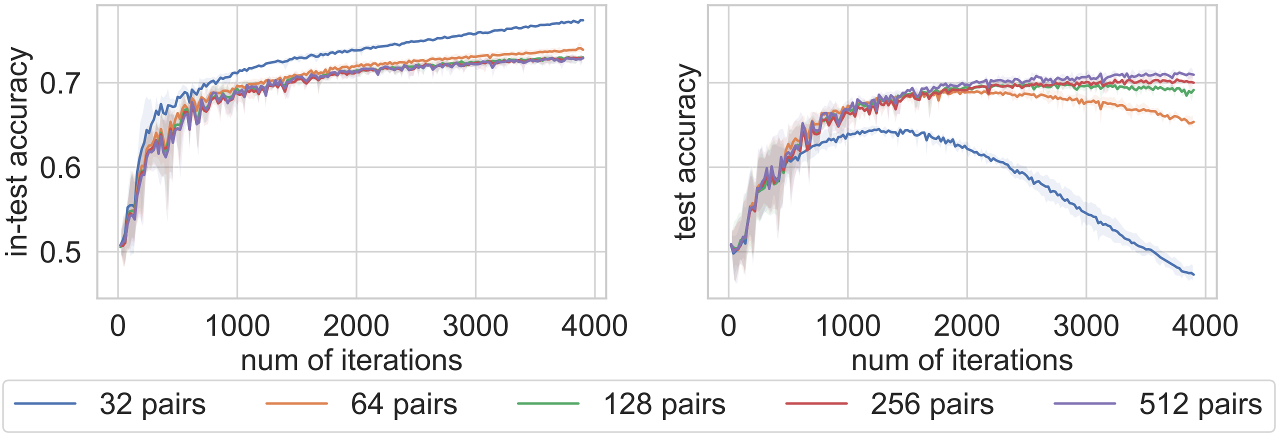


Figure 4: The number of counterfactuals vs. in-test accuracy curve and test accuracy curve on ColoredMNIST using the CLIP + Linear model. We conduct evaluations using 32, 64, 128, 256, and 512 data pairs.

Sensitivity on the number of invariant pairs. We evaluate the number of counterfactual pairs needed on ColoredMNIST dataset and report the in-domain test accuracy and test accuracy with 32,64,128,256,512 invariant pairs. The results show that with 32 counterfactual pairs, the number of pairs is insufficient for the model to eliminate spurious features, leading to spurious correlations (as indicated by an in-domain accuracy over 75%, meaning the classification relies on spurious features). However, when using 128 or 256 counterfactual pairs, the performance increases significantly and remains stable compared to the 32 counterfactual pairs. An insufficient number of pairs fails to eliminate the spurious feature, allowing the model to eventually rely on it, which leads to decreased accuracy on the test domain.

K.2 Comparison between waterbirds and Waterbirds-CF on ERM.

We run ERM on both waterbirds dataset and our Waterbirds-CF dataset. The results of ERM on both datasets are almost identical (cf. Figure 5).

K.3 Beyond linearity

Though our GRIT relies on linear assumption, our method could further work under nonlinear models empirically. In this section, we consider waterbirds-cf dataset using ResNet dataset. We apply mini-batch SGD with 300

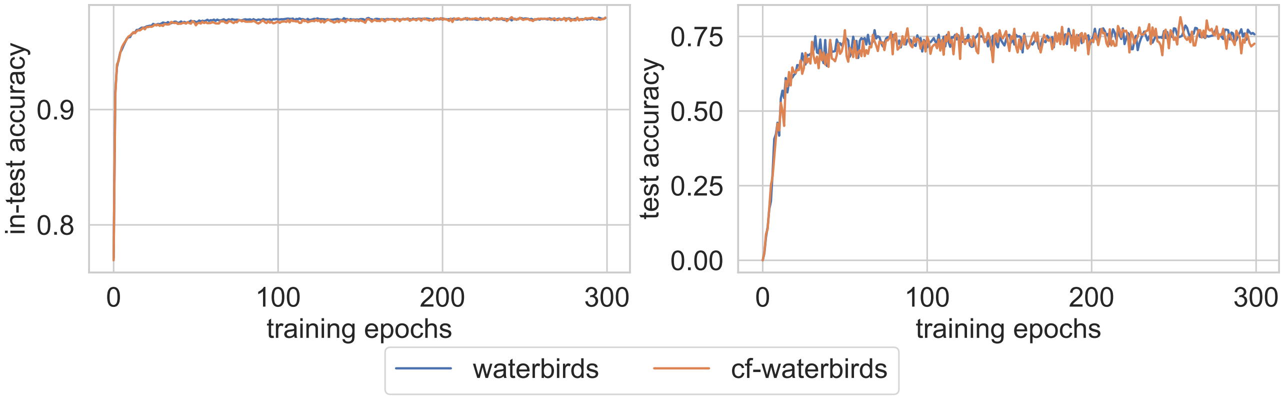


Figure 5: Compare the convergence curve of in-domain average test accuracy as well as the worst-case test accuracy on waterbirds and Waterbirds-CF datasets

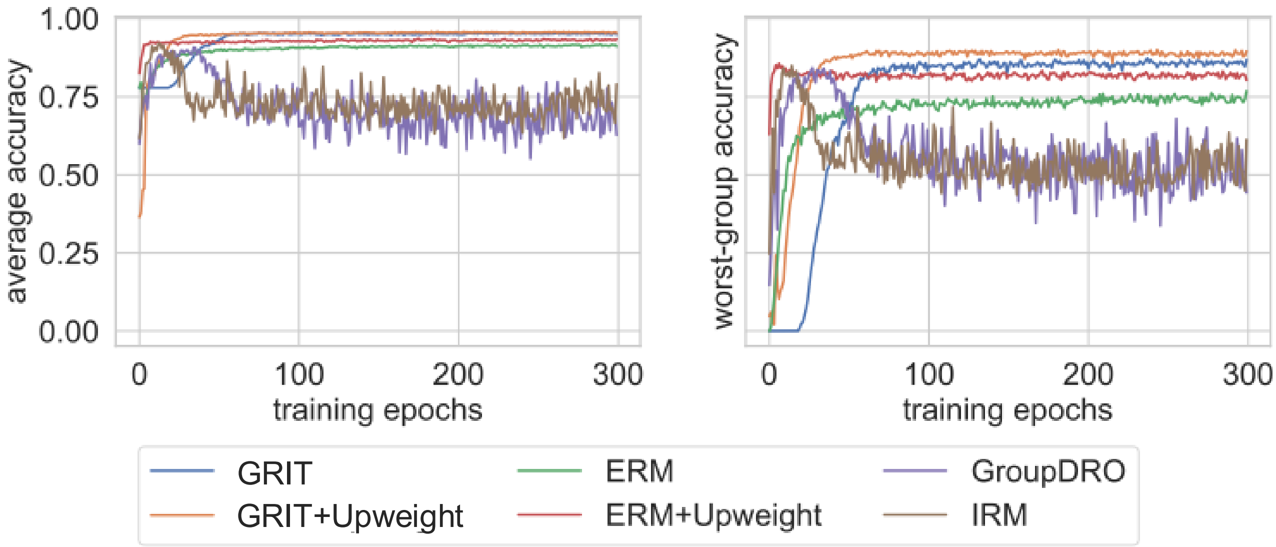


Figure 6: The convergence curve of Waterbirds-CF dataset. It shows that our method is significantly more stable than other DG methods without suffering overfitting.

epochs on the pretrained ResNet50 [He et al., 2016][1]. The optimizer used is SGD with a step size of 0.001, momentum of 0.9, and weight decay of 0.0001, as recommended for the Waterbirds dataset. The batch size is set to 128, For each batch, 128 counterfactual pairs are sampled to form the constraint term, which these pairs are matched prior to the linear classifier with MSE loss, The latent dimensionality is 64. The Lagrange multiplier is set to 500 (and 100 for IRM). For GroupDRO, we set the learning rate for updating the weight to be 0.01. All these hyperparameters are selected through grid search. We report the convergence curve of our methods as well as comparison to other baselines in Table 6. In the table, we report all the methods' best performance on average over 300 epochs of running. From the result we show that our GRIT using only 240 counterfactual pairs, outperforms ERM by 10.5% on the worst group accuracy. Further, we outperform other baselines like IRM and GroupDRO by 3.0% and 2.6%. Observe that Figure 6 shows that GRIT is much more stable compared to IRM and GroupDRO. As mentioned that GRIT is a causal data-centric approach, it could be combined with existed method to further improve domain generalization potentially. Here, we combine our method with up-weighting technique and we get 4.4% improvement over the ERM up-weighting counterpart. We further include experiments on the sensitivity of hyperparameters in Section J.

[1] pretrained model is IMAGENET1K_V1 from torchvision. Download here.

Table 8: Waterbirds-CF results on ResNet-50: best performance over 300 epochs, averaged across 5 runs. Adjusted accuracy is the reweighted metric to match the training distribution. Avg. Acc and WG. Acc denotes average accuracy and worst-group accuracy respectively.

	Oracle Validation		
	In-domain Acc	Test Acc	Worst domain Acc
ERM	0.978	0.917	0.767
ERM + UW	0.980	0.958	0.856
IRM	0.943	0.920	0.849
GroupDRO	0.934	0.907	0.842
GRIT w. oracle pairing	0.978	0.953	0.872
GRIT w. oracle pairing +UW	0.980	0.957	0.900