

# Virtues and Pitfalls of Weak-to-Strong Generalization: From Intrinsic Dimensions to Spurious Correlations

Qi Lei

NYU Courant Math & CDS

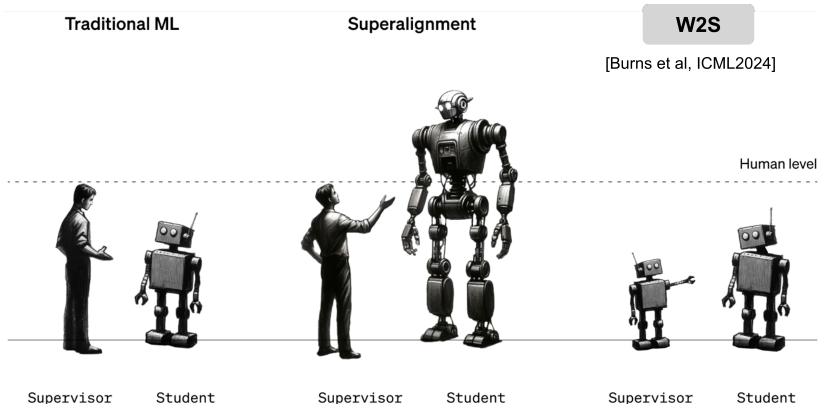
NJIT

<https://arxiv.org/abs/2502.05075>

<https://arxiv.org/abs/2509.24005>

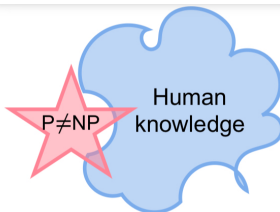
# Superalignment $\Rightarrow$ Weak-to-Strong (W2S)

- **Setup:** Strong, pre-trained student learns from *weaker* teacher via pseudo-labels.
- **Phenomenon:** Student often outperforms teacher (*weak-to-strong generalization*).
- **Question:** When and how does W2S happen? What governs its gain?



# Two explanations

- **Lower approximation error:** Student has new knowledge beyond teacher.



# Two explanations

- **Lower approximation error:** Student has new knowledge beyond teacher.  
(Lang et al., (2024), Shin et al., (2024), Ildiz et al., (2024), Wu & Sahai, (2024), Medvedev et al., (2025), Mulgund & Pabbaraju, (2025), Xue et al., (2025), Oh et al., (2025) and more)

# Two explanations

- **Lower approximation error:** Student has new knowledge beyond teacher.  
(Lang et al., (2024), Shin et al., (2024), Ildiz et al., (2024), Wu & Sahai, (2024), Medvedev et al., (2025), Mulgund & Pabbaraju, (2025), Xue et al., (2025), Oh et al., (2025) and more)
- **Lower estimation error:** Student uses knowledge more efficiently through fine-tuning.



# Two explanations

- **Lower approximation error:** Student has new knowledge beyond teacher.  
(Lang et al., (2024), Shin et al., (2024), Ildiz et al., (2024), Wu & Sahai, (2024), Medvedev et al., (2025), Mulgund & Pabbaraju, (2025), Xue et al., (2025), Oh et al., (2025) and more)
- **Lower estimation error:** Student uses knowledge more efficiently through fine-tuning.

1

---

<sup>1</sup>"Discrepancies are Virtue: Weak-to-Strong Generalization through Lens of Intrinsic Dimension", Yijun Dong, Yicheng Li, Yunai Li, Jason D Lee, Qi Lei, ICML 2025

# Intrinsic-dimension parameterization

## Intrinsic Dimension

The minimal parameter count needed to achieve (near-)optimal downstream performance.

$$\boxed{\theta^D} = \boxed{\theta_0^D} + \boxed{\Gamma} \boxed{\theta^d}$$

Model parameter of high dimension  $D$

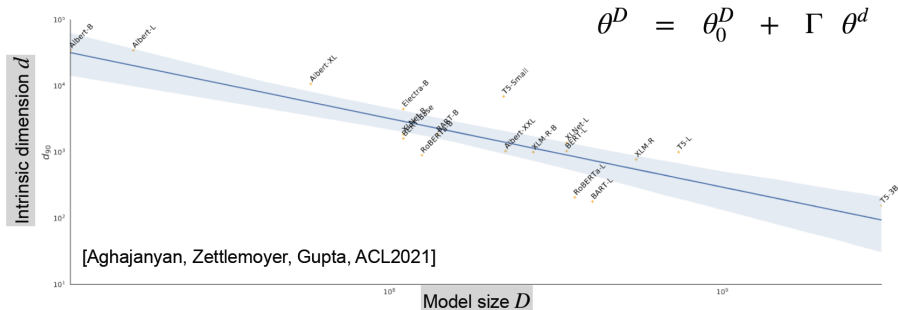
Pretrained initialization

Finetunable parameter of intrinsic dimension  $d < D$

$D \times d$  random projection

# Intrinsic Dimension

Learning over a well-pretrained model (e.g. finetuning) usually exhibits **low intrinsic dimensions**.



# Finetuning with low intrinsic dimensions

## Downstream task

- $(x, y) \sim \mathcal{D}(f_*)$  s.t.  $y = f_*(x) + z$  with i.i.d.  $z \sim \mathcal{N}(0, \sigma^2)$
- Learn  $f_* : \mathcal{X} \rightarrow \mathbb{R}$  from two datasets:

**Labeled** (small):  $\tilde{X} \in \mathcal{X}^n$  with noisy labels  $\tilde{y} \in \mathbb{R}^n$

**Unlabeled** (large):  $X \in \mathcal{X}^N$  with unknown labels  $y \in \mathbb{R}^N$

## Finetuning (FT) $\approx$ linear probing on low-rank gradient features

- Pretrained feature representations/gradient features for (weak) teacher and (strong) student:  $\phi_w, \phi_s$ .
- Kernel regime:  $f_\theta(x) = \phi(x)^\top \theta$  with finetunable  $\theta \in \mathbb{R}^D$ .

- **Weak** model  $\phi_w : \mathcal{X} \rightarrow \mathbb{R}^d$  produces

$$\tilde{\Phi}_w = \phi_w(\tilde{X}) \in \mathbb{R}^{n \times D}, \quad \Phi_w = \phi_w(X) \in \mathbb{R}^{N \times D}$$

$$\Sigma_w = \mathbb{E}[\phi_w(x)\phi_w(x)^\top] \\ \text{rank}(\Sigma_w) = d_w \ll D$$

- **Strong** model  $\phi_s : \mathcal{X} \rightarrow \mathbb{R}^D$  produces

$$\tilde{\Phi}_s = \phi_s(\tilde{X}) \in \mathbb{R}^{n \times D}, \quad \Phi_s = \phi_s(X) \in \mathbb{R}^{N \times D}$$

$$\Sigma_s = \mathbb{E}[\phi_s(x)\phi_s(x)^\top] \\ \text{rank}(\Sigma_s) = d_s \ll D$$

$$\text{rank}(\Sigma_w) = d_w \ll D \quad \text{rank}(\Sigma_s) = d_s \ll D$$

# W2S finetuning as regression



**Weak teacher**  $f_w(x) = \phi_w(x)^\top \theta_w$   
$$\theta_w = \arg \min_{\theta \in \mathbb{R}^d} \frac{1}{n} \|\tilde{\Phi}_w \theta - \tilde{y}\|_2^2 + \alpha_w \|\theta\|_2^2$$

**W2S**  $f_{w2s}(x) = \phi_s(x)^\top \theta_{w2s}$   
$$\theta_{w2s} = \arg \min_{\theta \in \mathbb{R}^d} \frac{1}{N} \|\Phi_s \theta - \Phi_w \theta_w\|_2^2 + \alpha_{w2s} \|\theta\|_2^2$$

**Strong SFT**  $f_s(x) = \phi_s(x)^\top \theta_s$   
$$\theta_s = \arg \min_{\theta \in \mathbb{R}^d} \frac{1}{n} \|\tilde{\Phi}_s \theta - \tilde{y}\|_2^2 + \alpha_s \|\theta\|_2^2$$

**Strong ceiling**  $f_c(x) = \phi_s(x)^\top \theta_c$   
$$\theta_c = \arg \min_{\theta \in \mathbb{R}^d} \frac{1}{n + N} \left\| \begin{bmatrix} \Phi_s \\ \Phi_s \end{bmatrix} \theta - \begin{bmatrix} \tilde{y} \\ y \end{bmatrix} \right\|_2^2 + \alpha_c \|\theta\|_2^2$$

## W2S v.s. SFT

How to evaluate the performance gain compared to the ideal case?

PGR (Performance Gap Recovery)

$$:= \frac{\Delta_{\text{Weak} \rightarrow \text{W2S}}}{\Delta_{\text{Weak} \rightarrow \text{Ceiling}}}$$

# Weak v.s. strong: model capacity + similarity

**Representation efficiency** — low intrinsic dimensions:

$$\text{rank}(\Sigma_w) = d_w \ll D, \quad \text{rank}(\Sigma_s) = d_s \ll D.$$

**Representation error** — FT approximation error:  $0 \leq \rho_s \leq \rho_w \leq 1$  where

$$\rho_s := \min_{\theta \in \mathbb{R}^d} \mathbb{E}[(\phi_s(x)^\top \theta - f_*(x))^2], \quad \rho_w := \min_{\theta \in \mathbb{R}^d} \mathbb{E}[(\phi_w(x)^\top \theta - f_*(x))^2].$$

We are interested in the **variance-dominated regime**  $\rho_s + \rho_w \ll \sigma^2$ .

**Representation similarity** — correlation dimension: Consider spectral decompositions:

$$\Sigma_s = V_s \Lambda_s V_s^\top \quad (D \times D), \quad \Sigma_w = V_w \Lambda_w V_w^\top \quad (D \times D).$$

The **correlation dimension** of  $(\phi_s, \phi_w)$  is

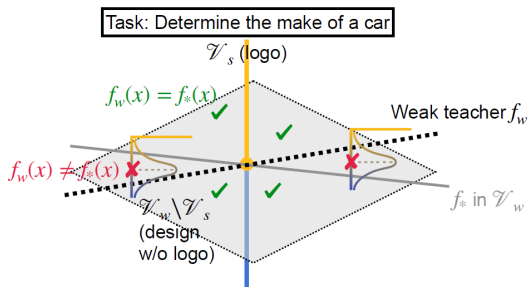
$$d_{s \wedge w} = \|V_s^\top V_w\|_F^2, \quad 0 \leq d_{s \wedge w} \leq \min\{d_s, d_w\}.$$

# Intuition: How does variance reduction in W2S happen?

$$\mathcal{V}_s = \text{Range}(\Sigma_s), \quad \mathcal{V}_w = \text{Range}(\Sigma_w)$$

$$\text{Var}(f_{w2s}) \approx \frac{d_{s \wedge w}}{n} + \frac{d_s}{N} \times \frac{d_w - d_{s \wedge w}}{n}$$

Var. in  $\mathcal{V}_w \cap \mathcal{V}_s$    W2S   Var. in  $\mathcal{V}_w \setminus \mathcal{V}_s$



Pseudolabel error in  $\mathcal{V}_w \setminus \mathcal{V}_s$  can be viewed as **independent label noise** w.r.t. the orthogonal strong features  $\mathcal{V}_s$ . The resulting variance *reduces proportionally* to  $d_s/N$ .

# Interpretation of Results: Performance Gap Recovery

## Definition of PGR

$$\text{Performance gap recovery (PGR)} = \frac{\text{ER}(f_w) - \text{ER}(f_{w2s})}{\text{ER}(f_w) - \text{ER}(f_c)}.$$



# Interpretation of Results: Performance Gap Recovery

## Definition of PGR

$$\text{Performance gap recovery (PGR)} = \frac{\text{ER}(f_w) - \text{ER}(f_{w2s})}{\text{ER}(f_w) - \text{ER}(f_c)}.$$



## Key Relationship

$$\text{PGR} \geq 1 - O\left(\frac{d_{s \wedge w}}{d_w}\right), \quad \text{where } d_{s \wedge w} = \|V_s^\top V_w\|_F^2,$$

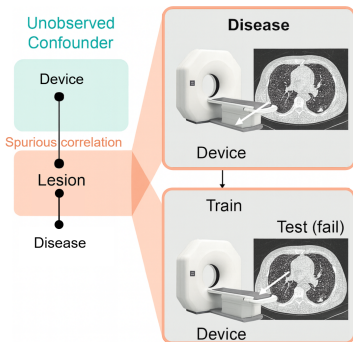
when the approximation error is negligible, and for large enough  $n, N$ .

### Interpretation:

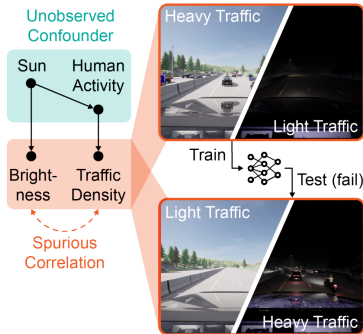
- Relatively smaller  $d_{s \wedge w} \Rightarrow$  better W2S recovery.
- 1) efficient student feature representation  $d_s \downarrow$ ;  
2) complementary student-teacher feature representation  $d_s - d_{s \wedge w} \uparrow$

# Beyond Intrinsic Dimension

- Real data often carry systematic biases (group imbalance, spurious features).



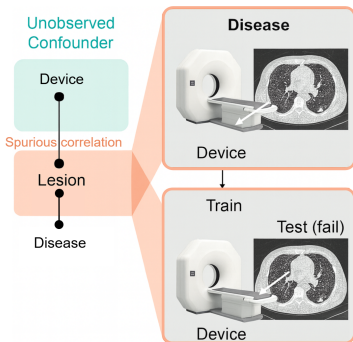
medical diagnosis



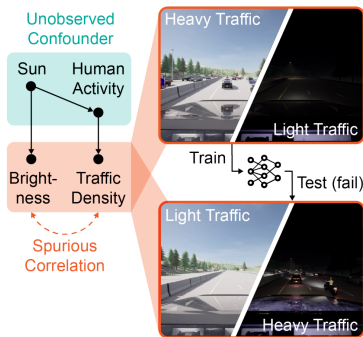
autonomous driving

# Beyond Intrinsic Dimension

- Real data often carry systematic biases (group imbalance, spurious features).
- Question: does W2S still hold under spurious correlations?



medical diagnosis



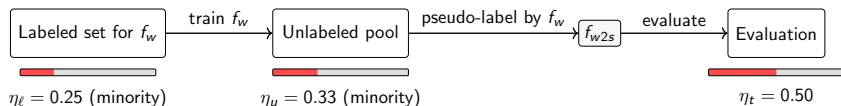
autonomous driving

# W2S under spurious correlations?<sup>2</sup>

- **General pretraining (diverse):** teacher  $f_w$  and student  $f_s$  originate from broad, heterogeneous corpora.
- **Specialized downstream task:** labels scarce; data acquisition induces selection/group bias  $\Rightarrow$  spurious features.
- **Two bias sources in W2S:** labeled set for  $f_w$  ( $\eta_\ell$ ) and unlabeled pool for pseudo-labels ( $\eta_u$ ); we study their effect.

## Specialized downstream task

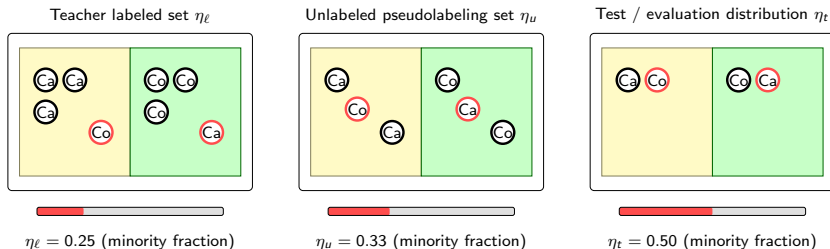
(label-scarce, biased)



<sup>2</sup>"Does Weak-to-strong Generalization Happen under Spurious Correlations?"  
Chenruo Liu, Yijun Dong, Qi Lei, Arxiv Preprint

# Setup: A Thought Experiment

- **Core feature**  $z(x) \in \mathbb{R}^{d_z}$ : Ca = camel, Co = cow
- **Majority**: Ca@desert, Co@grass    **Minority**: Ca@grass, Co@desert
- $\eta_\ell, \eta_u, \eta_t$ : minority fractions in labeled, unlabeled, and test sets



# Main Results: W2S under Spurious Correlation

**Proportional asymptotic limit:**

$$d_z, n, N \rightarrow \infty, d_z/n \rightarrow \gamma_z, d_z/N \rightarrow \nu_z$$



**Precise W2S gain:**

$$\Delta \mathcal{R}_{\eta_t} = \mathbb{E}_{\eta_\ell}[\mathbf{ER}_{\eta_t}(f_w)] - \mathbb{E}_{\eta_\ell, \eta_u}[\mathbf{ER}_{\eta_t}(f_s)]$$

**Theorem** (informal [LDL25]). When both teacher and student are unbiased over the population, assuming a large unlabeled data size  $\nu_z \ll 1$ ,

$$\begin{array}{lcl} \mathbb{E}_{\eta_\ell}[\mathbf{ER}_{\eta_t}(f_w)] & \xrightarrow{\mathbb{P}} & \begin{array}{c} \text{From label noise} \\ \mathbb{E}_{\eta_\ell=0.5}[\mathbf{ER}_{\eta_t}(f_w)] \end{array} + \gamma_z \Theta \left( \begin{array}{c} \text{From group imbalance} \\ (\eta_t - \eta_\ell)^2 \end{array} \right) \\ & & \text{VI} \\ \mathbb{E}_{\eta_\ell, \eta_u}[\mathbf{ER}_{\eta_t}(f_s)] & \xrightarrow{\mathbb{P}} & \begin{array}{c} \mathbb{E}_{\eta_\ell=\eta_u=0.5}[\mathbf{ER}_{\eta_t}(f_s)] \end{array} + \gamma_z \Theta \left( \begin{array}{c} \text{Negligible when } \eta_\ell = \eta_u \\ (\eta_u - \eta_\ell)^2 + \nu_z(\eta_t - \eta_u)^2 \end{array} \right) \end{array}$$

# Main Results: W2S under Spurious Correlation

**Proportional asymptotic limit:**

$$d_z, n, N \rightarrow \infty, d_z/n \rightarrow \gamma_z, d_z/N \rightarrow \nu_z$$



**Precise W2S gain:**

$$\Delta \mathcal{R}_{\eta_t} = \mathbb{E}_{\eta_\ell}[\mathbf{ER}_{\eta_t}(f_w)] - \mathbb{E}_{\eta_\ell, \eta_u}[\mathbf{ER}_{\eta_t}(f_s)]$$

For  $p_{s \wedge w} \ll p_s$  and  $\nu_z \ll 1$ :

😊  $\Delta \mathcal{R}_{\eta_t} > 0$  if  $\eta_\ell = \eta_u$

😞  $\Delta \mathcal{R}_{\eta_t} \searrow$  as  $(\eta_u - \eta_\ell)^2 \nearrow$

**Theorem** (informal [LDL25]). When both teacher and student are unbiased over the population, assuming a large unlabeled data size  $\nu_z \ll 1$ ,

From label noise

From **group imbalance**

$$\mathbb{E}_{\eta_\ell}[\mathbf{ER}_{\eta_t}(f_w)] \xrightarrow{\mathbb{P}} \mathbb{E}_{\eta_\ell=0.5}[\mathbf{ER}_{\eta_t}(f_w)] + \gamma_z \Theta \left( \frac{(\eta_t - \eta_\ell)^2}{\nu_z} \right)$$

VI

Negligible when  $\eta_\ell = \eta_u$

$$\mathbb{E}_{\eta_\ell, \eta_u}[\mathbf{ER}_{\eta_t}(f_s)] \xrightarrow{\mathbb{P}} \mathbb{E}_{\eta_\ell=\eta_u=0.5}[\mathbf{ER}_{\eta_t}(f_s)] + \gamma_z \Theta \left( \frac{(\eta_u - \eta_\ell)^2 + \nu_z(\eta_t - \eta_u)^2}{\nu_z} \right)$$

# Synthetic Experiments: Impact of Minority Ratio

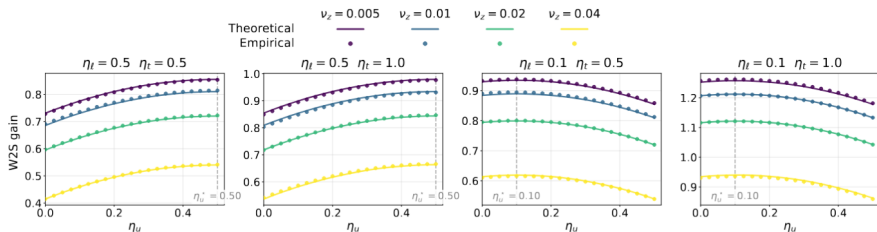


Figure 1: W2S gains across different combinations of  $\eta_\ell$  and  $\eta_t$ . Each panel shows theoretical (solid lines) and empirical (circles) results for W2S gain as a function of  $\eta_u$ , across different  $\nu_z$  values. Here we fix  $\mu_T, \mu_S, \Xi$ , and  $d_z$  with  $\|\mu_T\|_2^2 = 10.0$ ,  $\|\mu_S\|_2^2 = 0.1$ ,  $\|\Xi\|_F^2 = 0.1p_S$ . Vertical dashed lines indicate the theoretical optimal  $\eta_u^*$  values that maximize W2S gain.

# Real Experiments: Impact of Minority Ratio

Benchmarks: Waterbirds, BFFHQ, ImageNet-9, BG-COCO.

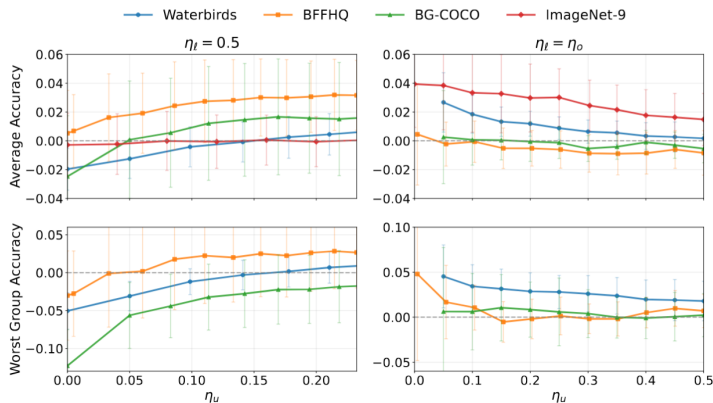


Figure 2: Average W2S gain across all teacher-student pairs as a function of  $\eta_u$  on all four datasets. Top row: average accuracy; bottom row: worst group accuracy. Left column fixes  $\eta_l = 0.5$ ; right column fixes  $\eta_l = \eta_o$ .

# Enhanced W2S under Spurious Correlations

**Motivation:** Vanilla W2S performance drops when

- $\eta_u \neq \eta_\ell$ : mismatch between unlabeled and labeled group proportions;
- pseudo-label noise is structured, often concentrated in minority groups.

**Key idea:** strengthen W2S by a *second-stage retraining* that focuses on more reliable signals and is robust to residual noise.

- (i) **Confidence-based selection:** choose a fraction  $p$  of unlabeled samples with highest student confidence (low-entropy predictions), filtering for clearer feature use.
- (ii) **Generalized cross-entropy (GCE):** replace CE with GCE on this subset, down-weighting occasional high-confidence but incorrect pseudo-labels.

# Enhanced-W2S Algorithm

## Effect:

- reduces over-reliance on spurious correlations;
- improves both average and worst-group accuracy;
- consistent gains across datasets and backbones, without group labels.

| Dataset    | $\eta_\ell$ | $\eta_u$ | Teacher–Student pair |                   |                    |               |                     |                      |                 |                     |                |                 |
|------------|-------------|----------|----------------------|-------------------|--------------------|---------------|---------------------|----------------------|-----------------|---------------------|----------------|-----------------|
|            |             |          | DINOv2<br>ConvNeXt   | DINOv2<br>Clipb32 | DINOv2<br>ResNet18 | DINOv2<br>MAE | ConvNeXt<br>Clipb32 | ConvNeXt<br>ResNet18 | ConvNeXt<br>MAE | Clipb32<br>ResNet18 | Clipb32<br>MAE | ResNet18<br>MAE |
| Waterbirds | 0.5         | $\eta_o$ | 6.60                 | 11.29             | 7.34               | 16.68         | 3.79                | 2.05                 | 6.28            | —                   | 2.07           | 0.77            |
|            | $\eta_o$    | 0.5      | 7.19                 | 13.86             | 11.73              | 11.62         | 2.85                | 2.02                 | 4.33            | —                   | 1.32           | 14.54           |
| BFFHQ      | 0.5         | $\eta_o$ | 6.85                 | 2.75              | 8.42               | 4.93          | 4.05                | —                    | —               | 6.54                | 5.12           | —               |
|            | $\eta_o$    | 0.5      | 3.92                 | 8.53              | 2.02               | 4.56          | 2.09                | —                    | —               | 2.06                | -1.37          | —               |
| BG-COCO    | 0.5         | $\eta_o$ | 5.38                 | 13.40             | 12.88              | 24.01         | 9.82                | 6.49                 | 15.25           | 3.39                | 12.43          | 2.05            |
|            | $\eta_o$    | 0.5      | 10.21                | 16.99             | 12.25              | -3.52         | 3.41                | 1.21                 | -3.07           | 3.48                | 0.31           | 3.70            |
| ImageNet-9 | 0.5         | $\eta_o$ | —                    | 6.03              | 7.45               | 24.11         | 4.74                | 5.30                 | 18.49           | 4.22                | 21.73          | 17.98           |
|            | $\eta_o$    | 0.5      | —                    | 8.21              | 11.28              | 22.00         | 3.77                | 1.81                 | 10.50           | 4.51                | 23.24          | 15.76           |

Table 1: Relative improvement of Enhanced-W2S over vanilla W2S (% , measured by average accuracy) across all datasets and teacher-student pairs

- Part I: W2S enabled by low intrinsic dimension + representation discrepancy.
- Part II: W2S affected by distribution mismatch + spurious correlations.
- Together: W2S governed by (i) representation efficiency, (ii) representation similarity, (iii) distribution alignment.

# Conclusion and Outlook

- Why W2S happens: intrinsic dimension + discrepancy.
- When W2S is vulnerable: spurious correlations, imbalanced groups.
- Outlook: multiple weak teachers, broader distribution shifts, alternative training (AI for education), fairness/safety.

# Conclusion and Outlook

- Why W2S happens: intrinsic dimension + discrepancy.
- When W2S is vulnerable: spurious correlations, imbalanced groups.
- Outlook: multiple weak teachers, broader distribution shifts, alternative training (AI for education), fairness/safety.

Thank you! Questions?

# Theoretical Setup: Regression under Spurious Correlation

- **Core feature**  $z(x) \in \mathbb{R}^{d_z}$ : semantic signal that drives the label

$$y = z(x)^\top \beta_* + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma_y^2).$$

- **Group feature**  $\xi(x) \in \mathbb{R}^{d_\xi}$ : depends only on group  $g \in \{0, 1\}$ ,

$$\xi(x) \sim \mathcal{N}(g\mu_\xi, \sigma_\xi^2 I).$$

Not predictive alone, but spurious correlation appears through *interaction terms*  $z(x) \otimes \xi(x)$ .

- **Teacher vs. student representations:**

$$\varphi_w(x) = [z; z \otimes (\mathbb{W}^\top \xi)], \quad \varphi_s(x) = [z; z \otimes (\S^\top \xi)],$$

with group-dimensions  $p_w - 1$  vs.  $p_s - 1$  ( $p_s \leq p_w$ ). Projection means:

$$\mu_w = \mathbb{W}^\top \mu_\xi, \quad \mu_s = \S^\top \mu_\xi.$$

- **Overlap:**  $\Xi = \mathbb{W}^\top \S, \quad p_{s \wedge w} = 1 + \|\Xi\|_F^2$ .

**Risk evaluation:** for test distribution  $\mathbb{D}(\eta_t)$

$$\mathbf{ER}_{\eta_t}(f) = \mathbb{E}_{(x,y) \sim \mathbb{D}(\eta_t)}[(f(x) - f^*)^2]$$

# Main Results: W2S under Spurious Correlation

Teacher (weak, after SFT):

$$\mathbf{ER}_{\eta_t}(f_w) \rightarrow \sigma_y^2 \frac{d_z}{n} \left( \underbrace{p_w}_{\text{variance term}} + \underbrace{\frac{\|(\eta_t - \eta_\ell)\mu_w\|_2^2}{\sigma_\xi^2}}_{\text{spurious term}} \right)$$

Student (strong, after W2S):

$$\mathbf{ER}_{\eta_t}(f_s) \rightarrow \sigma_y^2 \frac{d_z}{n} \left( \underbrace{p_{s \wedge w}}_{\text{variance} \leq p_w} + \underbrace{\frac{\|(\eta_u - \eta_\ell)\mu_w + (\eta_t - \eta_u)\Xi\mu_s\|_2^2}{\sigma_\xi^2}}_{\text{spurious term}} + \underbrace{\Theta\left(\frac{d_z}{N}\right)}_{\text{small term}} \right)$$