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

Qi Lei

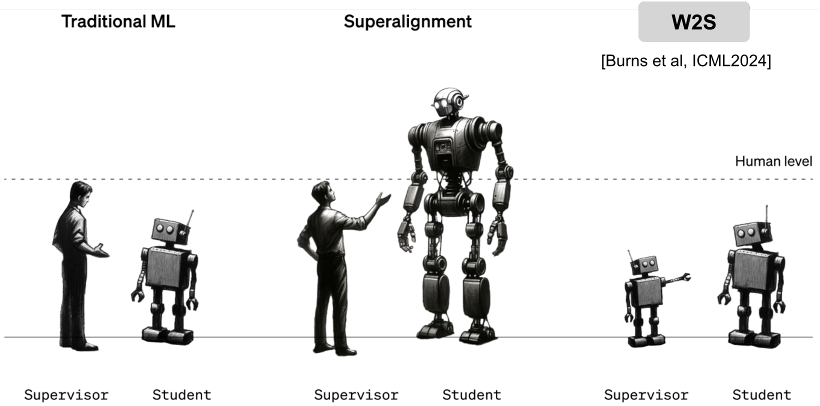
Courant Math & CDS

CDS Seminar

https://arxiv.org/abs/2502.05075
https://arxiv.org/abs/2509.24005

Superalignment ⇒ 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, 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, and more)
- **Lower estimation error:** Student uses knowledge more efficiently during FT. [1]

[1] "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.

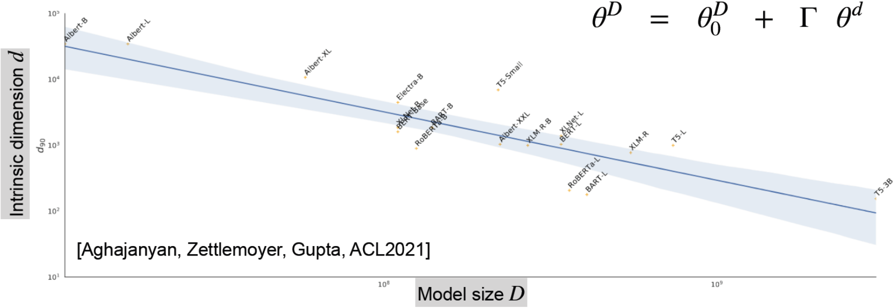
$$\theta^D = \theta_0^D + \Gamma \; \theta^d$$

- θ^D: Model parameter of high dimension D
- θ_0^D: Pretrained initialization
- Γ: $D \times d$ random projection
- θ^d: Finetunable parameter of intrinsic dimension $d < D$

Intrinsic Dimension

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

$$\theta^D = \theta_0^D + \Gamma \, \theta^d$$



[Aghajanyan, Zettlemoyer, Gupta, ACL2021]

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 $f_* : \mathcal{X} \to \mathbb{R}$ 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: ϕ_w, ϕ_s.
- Kernel regime: $f_\theta(x) = \phi(x)^\top \theta$ with finetunable $\theta \in \mathbb{R}^D$.
- **Weak** model $\phi_w : \mathcal{X} \to \mathbb{R}^d$ produces

  $\tilde{\Phi}_w = \phi_w(\tilde{X}) \in \mathbb{R}^{n \times D}, \ \Phi_w = \phi_w(X) \in \mathbb{R}^{N \times D}$
- **Strong** model $\phi_s : \mathcal{X} \to \mathbb{R}^D$ produces

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

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

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

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

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 $f_c(x) = \phi_s(x)^\top \theta_c$

$$\theta_c = \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{n+N} \left\| \begin{bmatrix} \tilde{\Phi}_s \\ \Phi_s \end{bmatrix} \theta - \begin{bmatrix} \tilde{y} \\ y \end{bmatrix} \right\|_2^2 + \alpha_c \|\theta\|_2^2$$

W2S v.s. SFT

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

PGR (Performance Gap Recovery)

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

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: $0 \leqslant \rho_s \leqslant \rho_w \leqslant 1$ where

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

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

Representation similarity — correlation dimension: Consider spectral decompositions:

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

The correlation dimension of (ϕ_s, ϕ_w) is

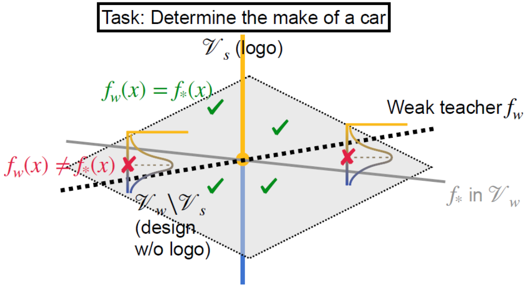
$$d_{s \wedge w} = \|V_s^\top V_w\|_F^2, \qquad 0 \leqslant d_{s \wedge w} \leqslant \min\{d_s, d_w\}.$$

Intuition: How does variance reduction in W2S happen?

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

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

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

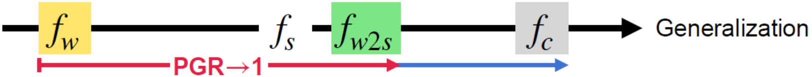


Pseudolabel error in $\mathcal{V}_w \setminus \mathcal{V}_s$ can be viewed as **independent label noise** w.r.t. the orthogonal strong features $\mathcal{V}_s$. The resulting variance *reduces proportionally* to d_s/N.

Interpretation of Results: Performance Gap Recovery

Definition of PGR

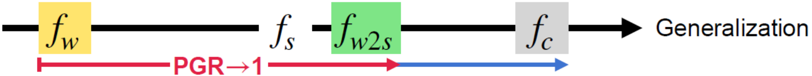
$$\text{Performance gap recovery (PGR)} = \frac{\text{ER}(f_w) - \text{ER}(f_{w2s})}{\text{ER}(f_w) - \text{ER}(f_c)}.$$



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

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

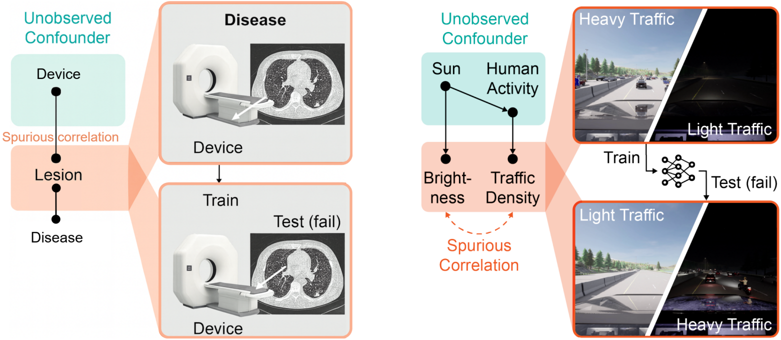
when the approximation error is negligible, and for large enough n, N.

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

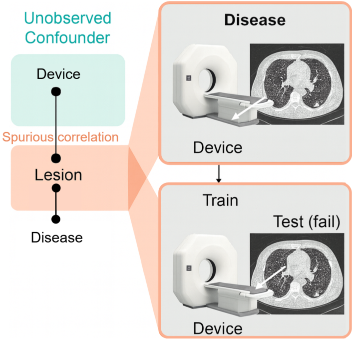
- Real data often carry systematic biases (group imbalance, spurious features).



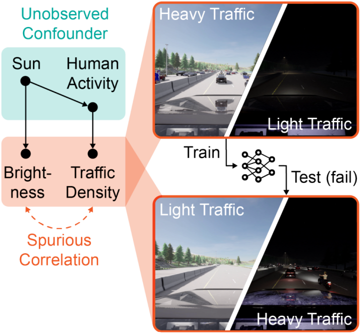
medical diagnosis autonomous driving

Beyond Intrinsic Dimension

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



medical diagnosis

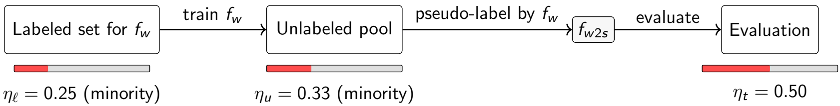


autonomous driving

Why study W2S under spurious correlations?[2]

- **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 (η_ℓ) and unlabeled pool for pseudo-labels (η_u); we study their effect.

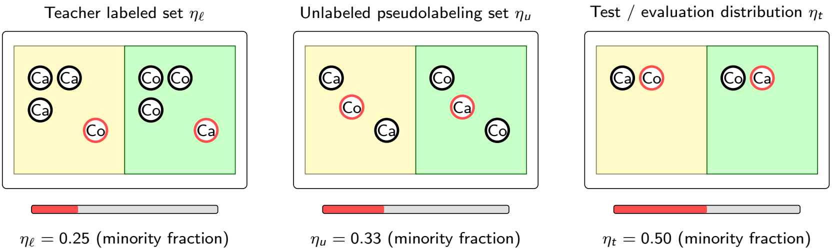
Specialized downstream task
(label-scarce, biased)



[2] "Does Weak-to-strong Generalization Happen under Spurious Correlations?" Chenruo Liu, Yijun Dong, Qi Lei, Arxiv Preprint

Setup: A Thought Experiment

- **Core feature** $z(x)$: 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



Theoretical Setup: Regression under Spurious Correlation

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

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

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

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

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

- **Teacher vs. student representations:**

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

  with group-dimensions $p_w - 1$ vs. $p_s - 1$ ($p_s \leqslant p_w$). Projection means: $\mu_w = \mathbb{W}^\top \mu_\xi,\ \mu_s = \S^\top \mu_\xi$.

- **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):

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

Student (strong, after W2S):

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

When Does W2S Work under Spurious Correlation?

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

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

- If $\eta_u = \eta_\ell$: W2S always happens with enough data.
- If $\eta_u \neq \eta_\ell$: W2S may fail, gain shrinks with mismatch.
- Teacher–student representation similarity Ξ also controls robustness.

Synthetic Experiments: Impact of Minority Ratio

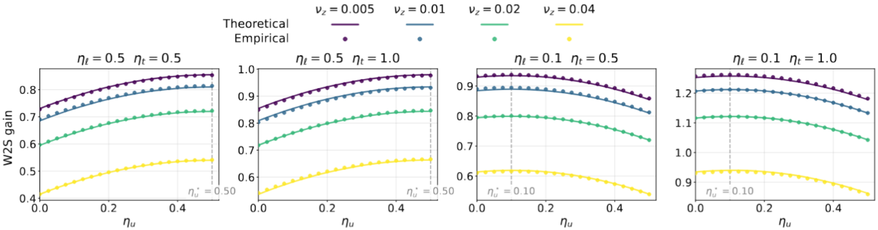


Figure 1: W2S gains across different combinations of η_ℓ and η_t. Each panel shows theoretical (solid lines) and empirical (circles) results for W2S gain as a function of η_u, across different ν_z values. Here we fix $\boldsymbol{\mu}_T$, $\boldsymbol{\mu}_S$, Ξ, and d_z with $\|\boldsymbol{\mu}_T\|_2^2 = 10.0$, $\|\boldsymbol{\mu}_S\|_2^2 = 0.1$, $\|\Xi\|_F^2 = 0.1p_S$. Vertical dashed lines indicate the theoretical optimal $\eta_u^\star$ values that maximize W2S gain.

Real Experiments: Impact of Minority Ratio

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

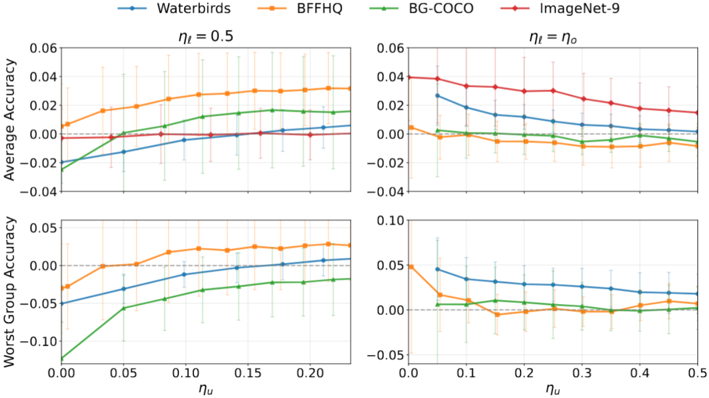


Figure 2: Average W2S gain across all teacher-student pairs as a function of η_u on all four datasets. Top row: average accuracy; bottom row: worst group accuracy. Left column fixes $\eta_\ell = 0.5$; right column fixes $\eta_\ell = \eta_o$.

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	η_ℓ	η_u	Teacher–Student pair									
			DINOv2 ConvNeXt	DINOv2 Clipb32	DINOv2 ResNet18	DINOv2 MAE	ConvNeXt Clipb32	ConvNeXt ResNet18	ConvNeXt MAE	Clipb32 ResNet18	Clipb32 MAE	ResNet18 MAE
Waterbirds	0.5	η_o	6.60	11.29	7.34	16.68	3.79	2.05	6.28	—	2.07	0.77
	η_o	0.5	7.19	13.86	11.73	11.62	2.85	2.02	4.33	—	1.32	14.54
BFFHQ	0.5	η_o	6.85	2.75	8.42	4.93	4.05	—	—	6.54	5.12	—
	η_o	0.5	3.92	8.53	2.02	4.56	2.09	—	—	2.06	-1.37	—
BG-COCO	0.5	η_o	5.38	13.40	12.88	24.01	9.82	6.49	15.25	3.39	12.43	2.05
	η_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	η_o	—	6.03	7.45	24.11	4.74	5.30	18.49	4.22	21.73	17.98
	η_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

Unifying View

- 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?