A Truly Unbiased Model

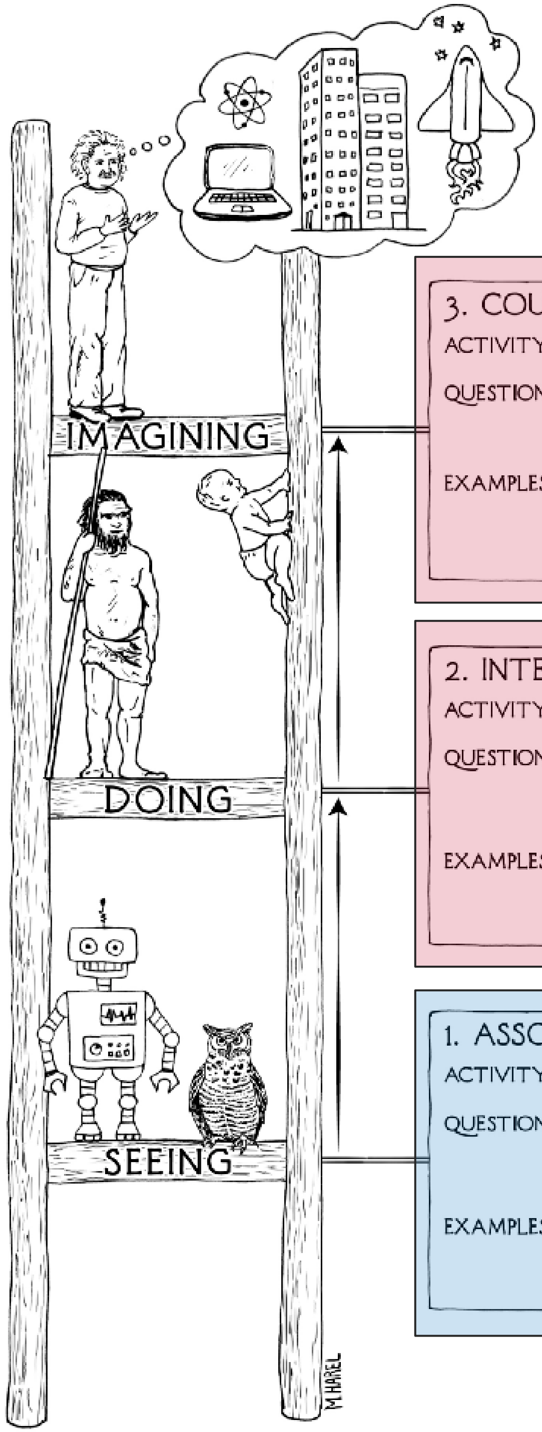
Recent Progress @ MReaL

Hanwang Zhang 张含望

https://mreallab.github.io/

hanwangzhang@ntu.edu.sg

Causality vs. Correlation

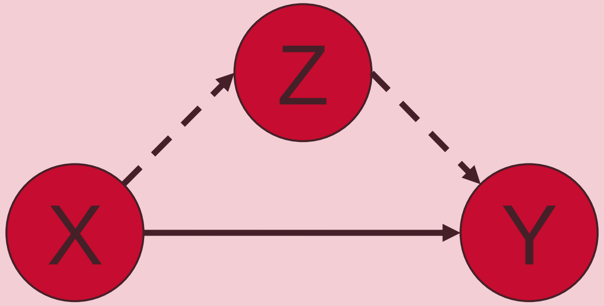


3. COUNTERFACTUALS

ACTIVITY: Imagining, Retrospection, Understanding

QUESTIONS: *What if I had done ...? Why?*
(Was it X that caused Y? What if X had not occurred? What if I had acted differently?)

EXAMPLES: Was it the aspirin that stopped my headache? Would Kennedy be alive if Oswald had not killed him? What if I had not smoked for the last 2 years?

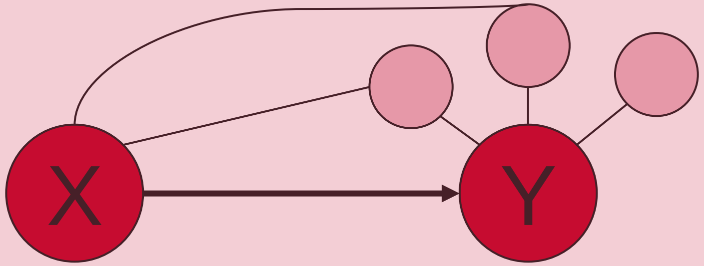


2. INTERVENTION

ACTIVITY: Doing, Intervening

QUESTIONS: *What if I do ...? How?*
(What would Y be if I do X?
How can I make Y happen?)

EXAMPLES: If I take aspirin, will my headache be cured?
What if we ban cigarettes?



1. ASSOCIATION

ACTIVITY: Seeing, Observing

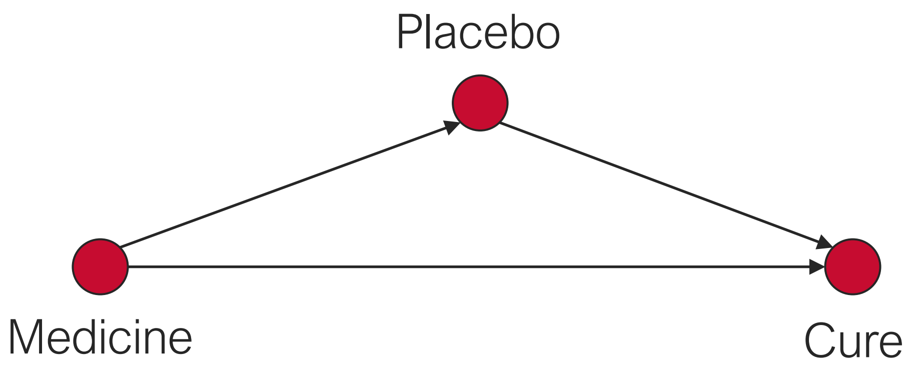
QUESTIONS: *What if I see ...?*
(How are the variables related?
How would seeing X change my belief in Y?)

EXAMPLES: What does a symptom tell me about a disease?
What does a survey tell us about the election results?

NANYANG TECHNOLOGICAL UNIVERSITY

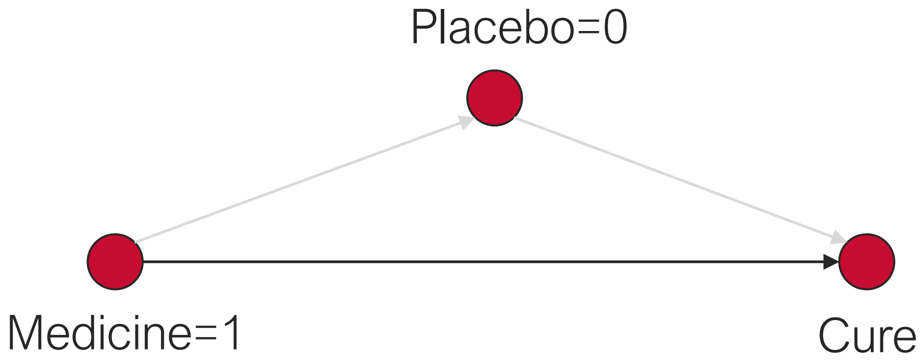
Mediation Effect

- How to remove Placebo Effect



Mediation Effect

- How to remove Placebo Effect?
- Challenge: Med = 1 and Placebo = 1 always co-occur; or, illegal to realize the following graph



Ideal case

Mediation Effect: TDE (the minus trick)

- How to remove Placebo Effect?
- Solution: counterfactual→cheating→Med = 0 but Placebo = 1

Do and CF in debiasing methods

- Assumption: train $\neq$ test (OOD)
- Do: CSS, CVL, Re-weighting/Re-sample
- CF: RUBi, CF-VQA, LMH

CSS: Chen et al. Counterfactual Samples Synthesizing for Robust Visual Question Answering. CVPR’20

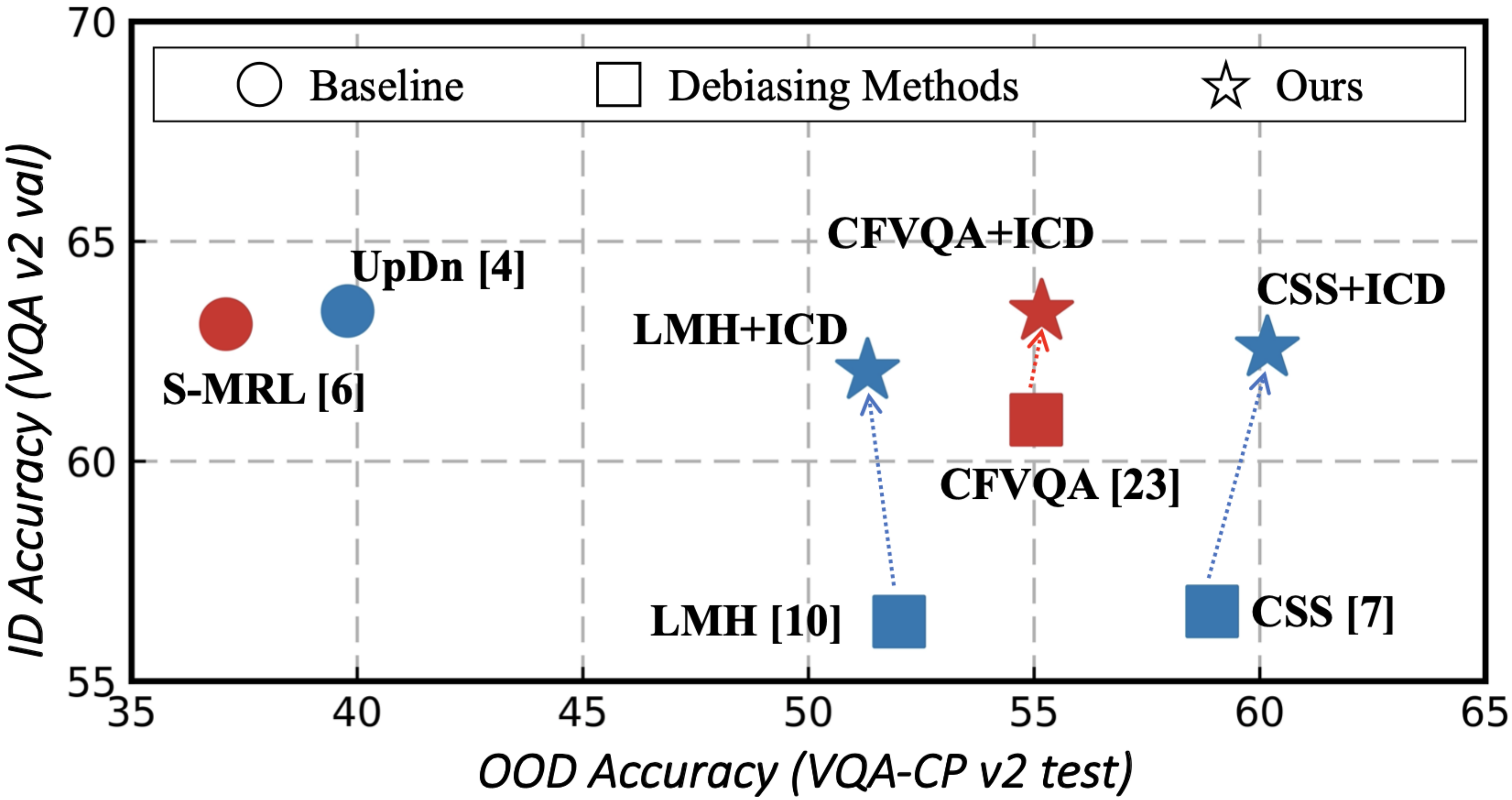
CVL: Abbasnejad et al. Counterfactual Vision and Language Learning. CVPR’20

RUBi: Cadene et al. RUBi: Reducing Unimodal Biases in Visual Question Answering. NeurIPS’19

LMH: Clark et al. Don’t Take the Easy Way Out: Ensemble based Methods for Avoiding Known Dataset Biases. EMNLP’19

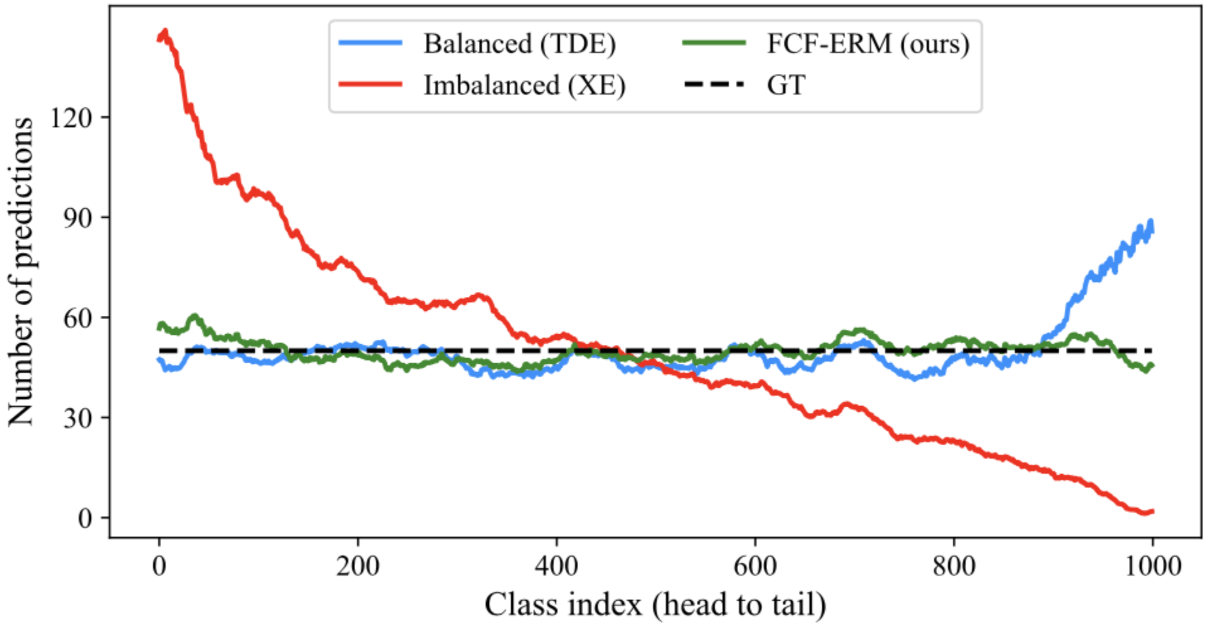
NANYANG TECHNOLOGICAL UNIVERSITY

VQA OOD

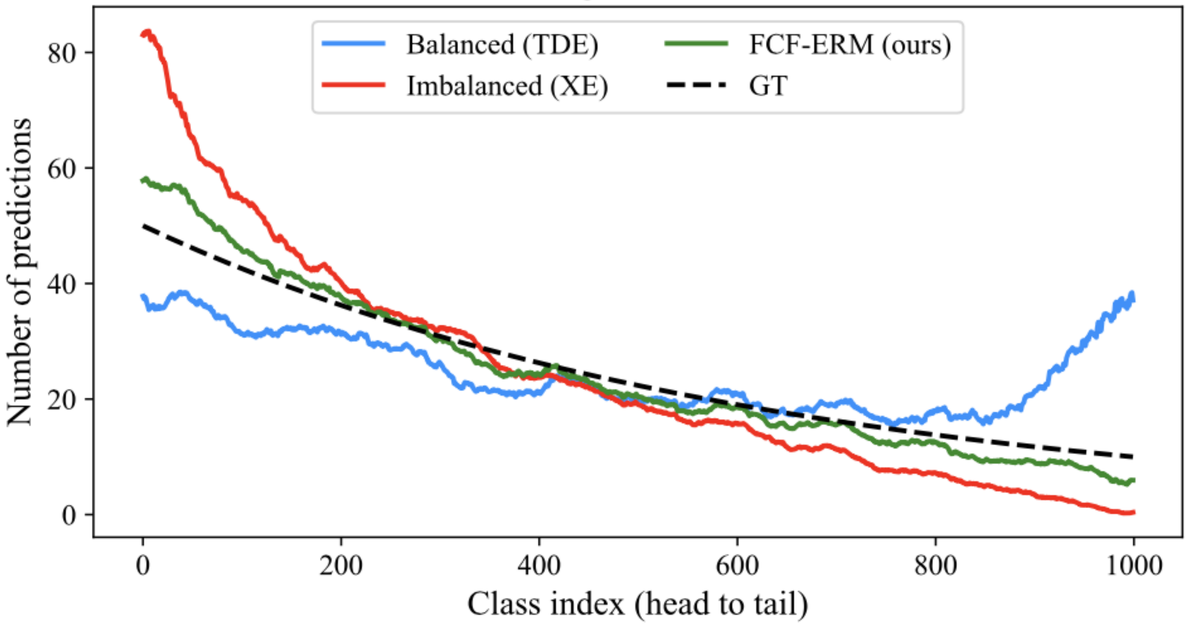


Baseline
Debiasing Methods
Ours
UpDn [4]
S-MRL [6]
CFVQA+ICD
LMH+ICD
CSS+ICD
CFVQA [23]
LMH [10]
CSS [7]
70
65
60
55
35
40
45
50
55
60
65
ID Accuracy (VQA v2 val)
OOD Accuracy (VQA-CP v2 test)

Long-tail OOD



(a) Balanced Test



(b) Imbalanced Test

What’s new?

- A best of two worlds VQA model
- A best of two worlds long-tailed model

NANYANG TECHNOLOGICAL UNIVERSITY

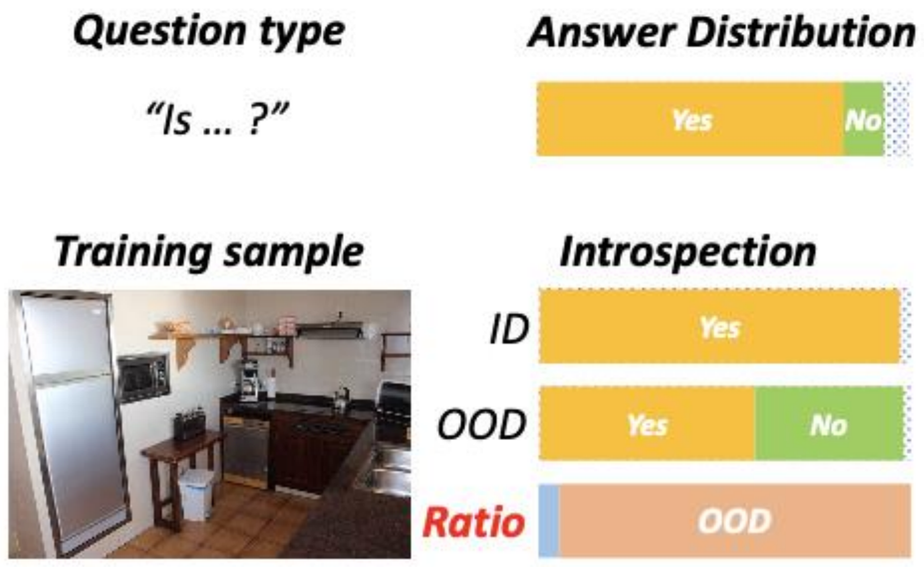
Introspective Distillation for VQA: Key Idea

- **ID-Teacher**: Good @ Train = Test, Bad @ Train != Test
- **OOD-Teacher**: Good @ Train != Test, Bad @ Train = Test
- A **Student** learns the best of the two teachers
- By **ONLY** given the train, how does the student know to whom she should listen（oracle)?

Introspective Distillation for VQA: Key Idea

Introspection: Case 1

- if *ID-bias* > *OOD-bias*, then *ID-teacher* < *OOD-teacher*

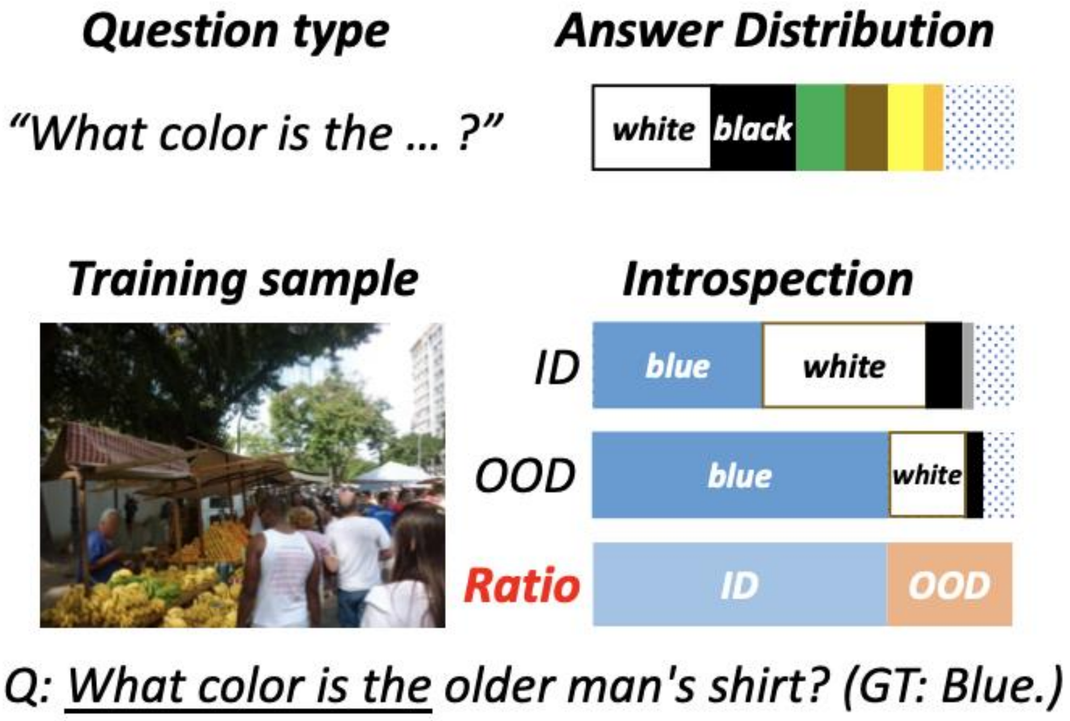


Q: Is that an electric oven? (GT: Yes.)

For each sample,
If ID-Teacher is too good to be true
OOD-Teacher not so good,
W(OOD) ∝ XE(OOD)/XE(ID)

Introspection: Case 2

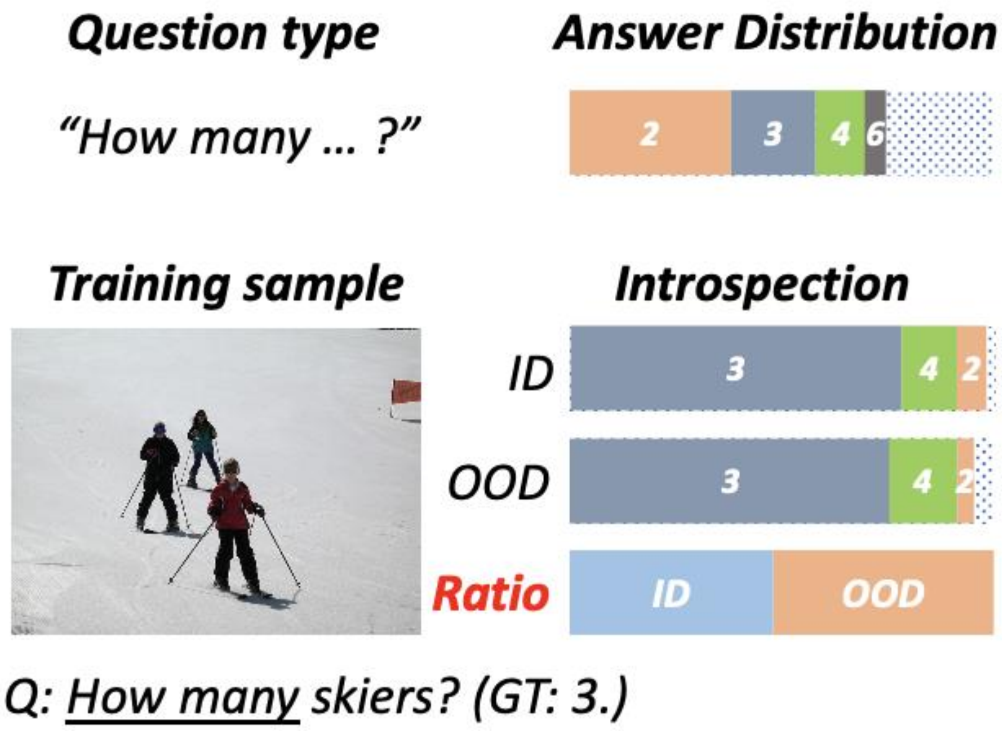
- if *ID-bias* < *OOD-bias*, then *ID-teacher* > *OOD-teacher*



For each sample,
If ID-Teacher is not so good,
OOD-Teacher is too good to be true,
W(ID) $\propto$ XE(ID)/XE(OOD)

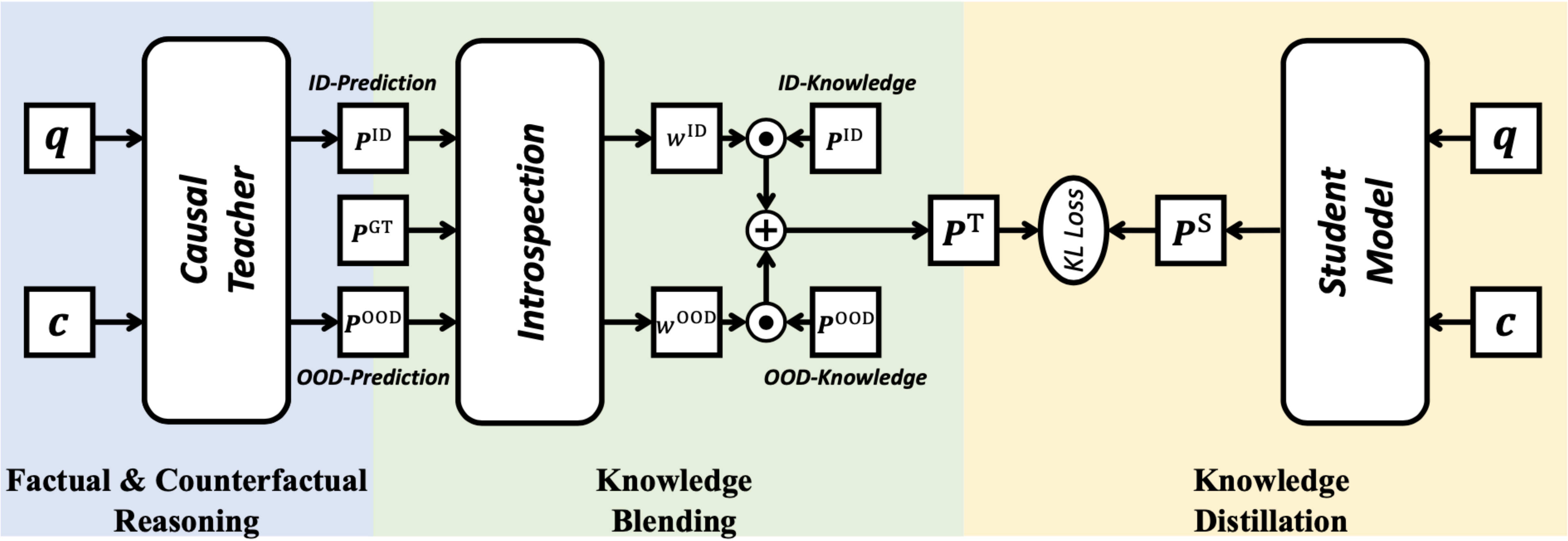
Introspection: Case 3

- if *ID-bias* ≈ *OOD-bias*, then *ID-teacher* ≈ *OOD-teacher*



For each sample,
If ID/OOD-teachers are similar,
W(ID) ≈ W(OOD) as
XE(ID) ≈ XE(OOD)

The Introspective Pipeline



How does Introspection look like?

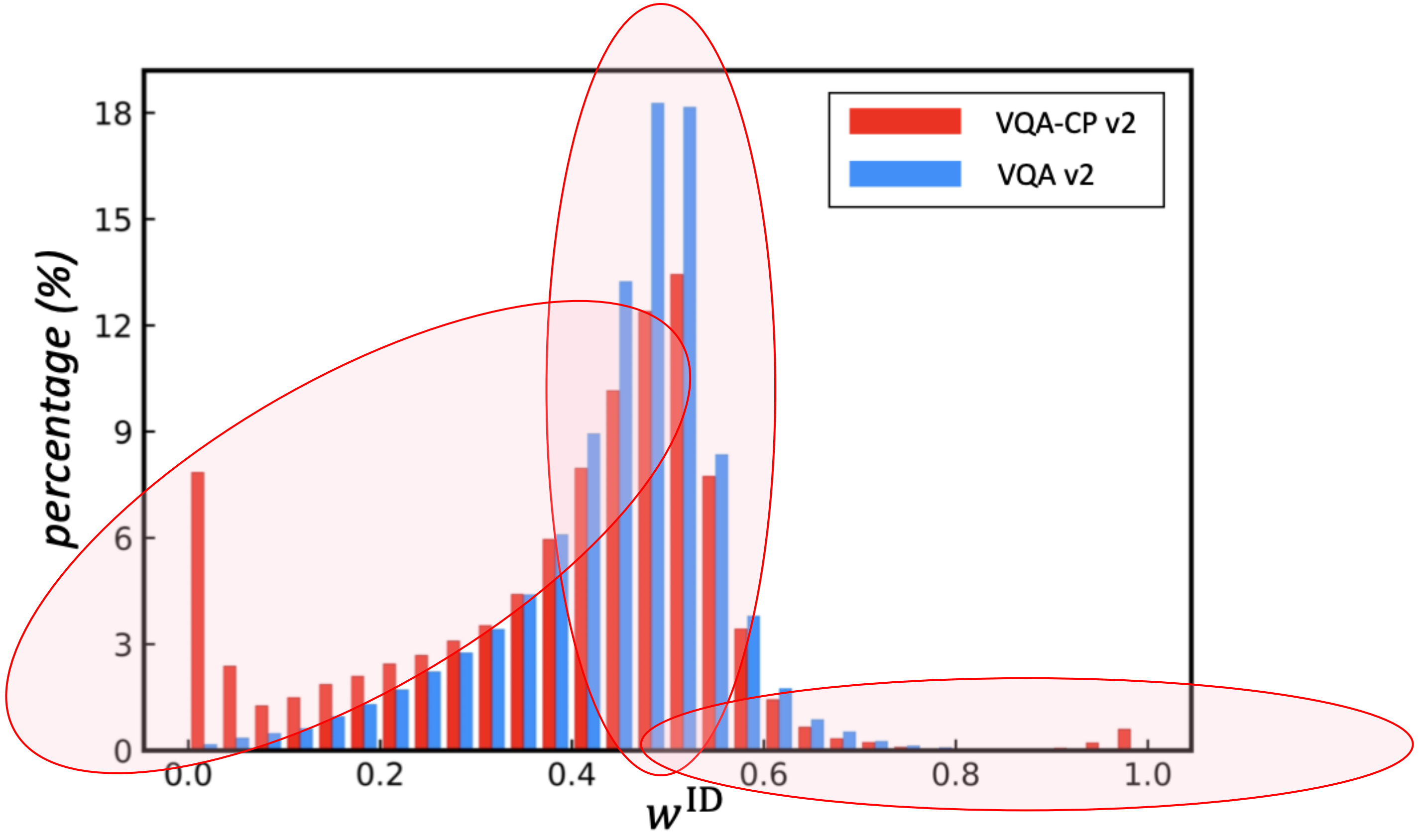


Figure 4: The distribution of w^{ID} on the VQA-CP v2 and VQA v2 training sets.

NANYANG TECHNOLOGICAL UNIVERSITY

How does Introspection look like? Both are mostly Case 3

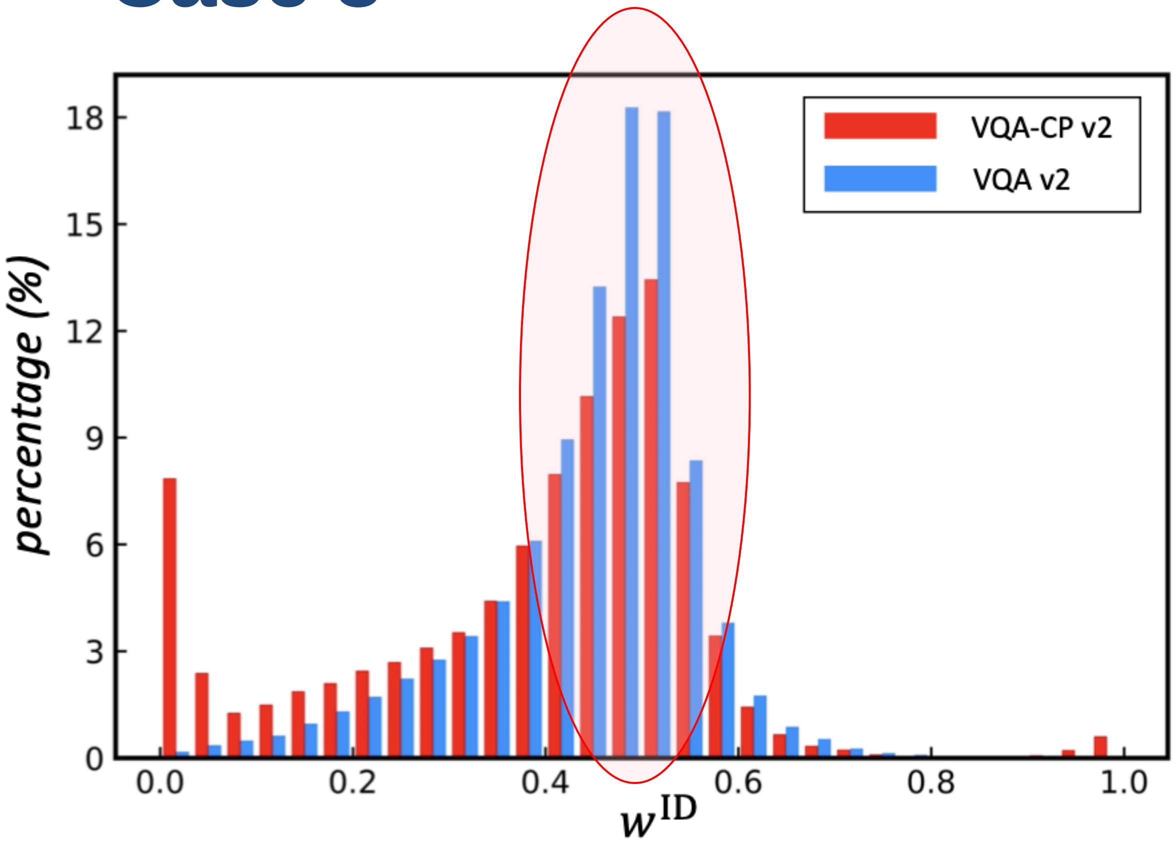
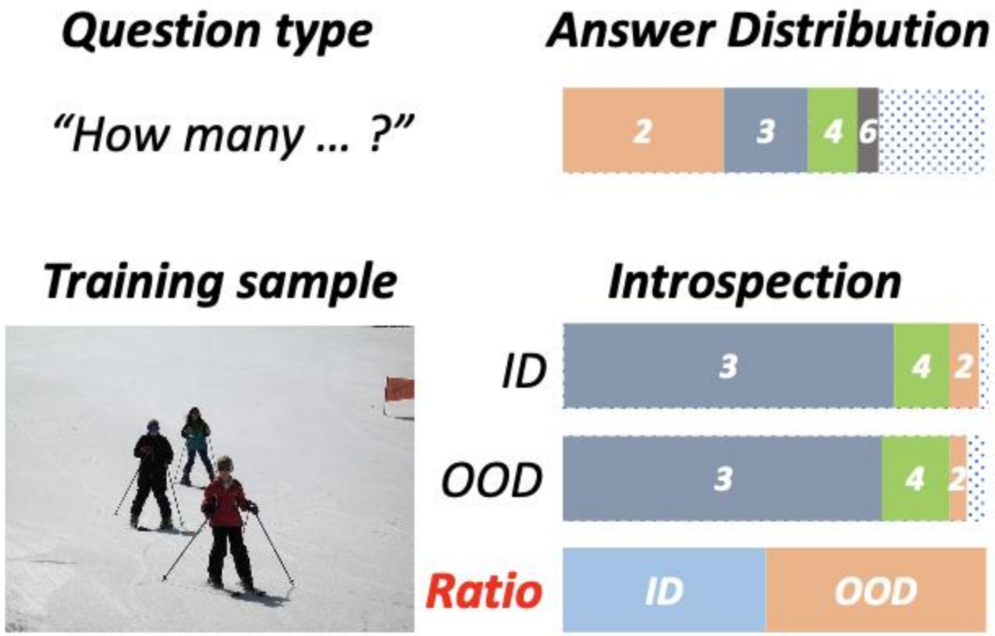


Figure 4: The distribution of w^{ID} on the VQA-CP v2 and VQA v2 training sets.

ID-teacher ≈ OOD-teacher



Q: How many skiers? (GT: 3.)

How does Introspection look like? VQA-CP has more Case 1 than VQA

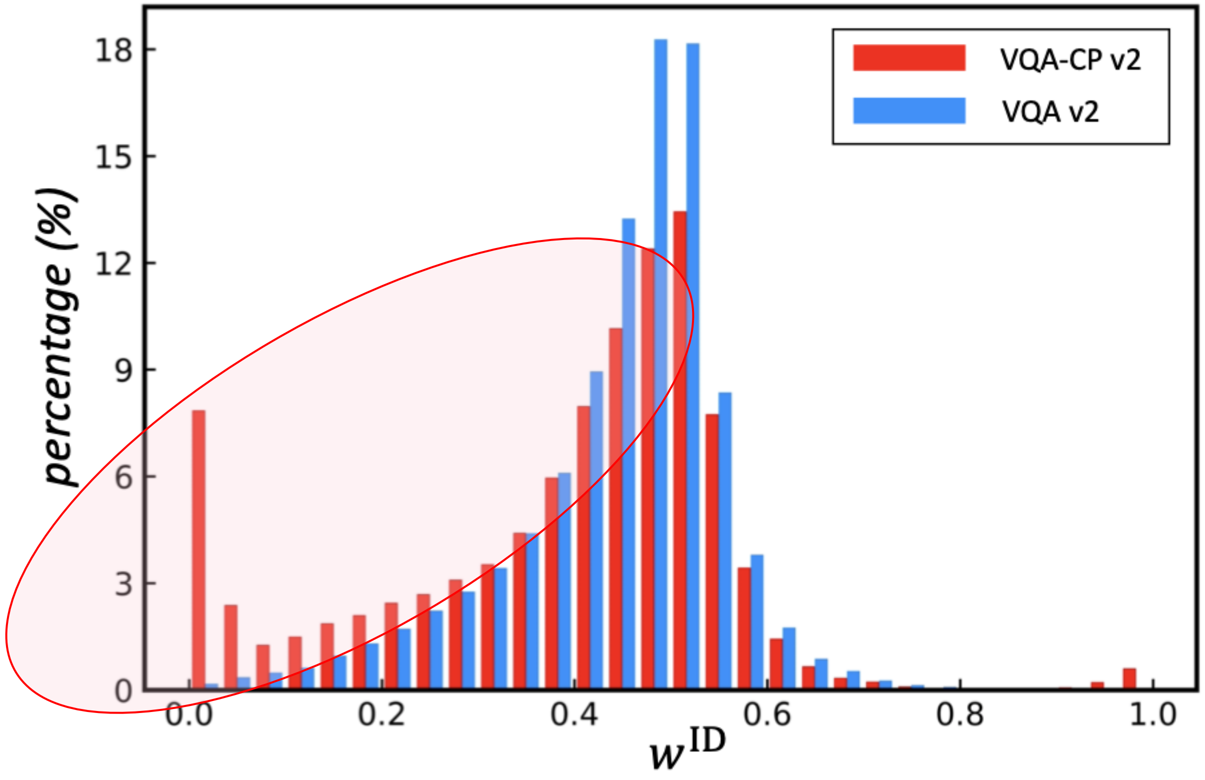
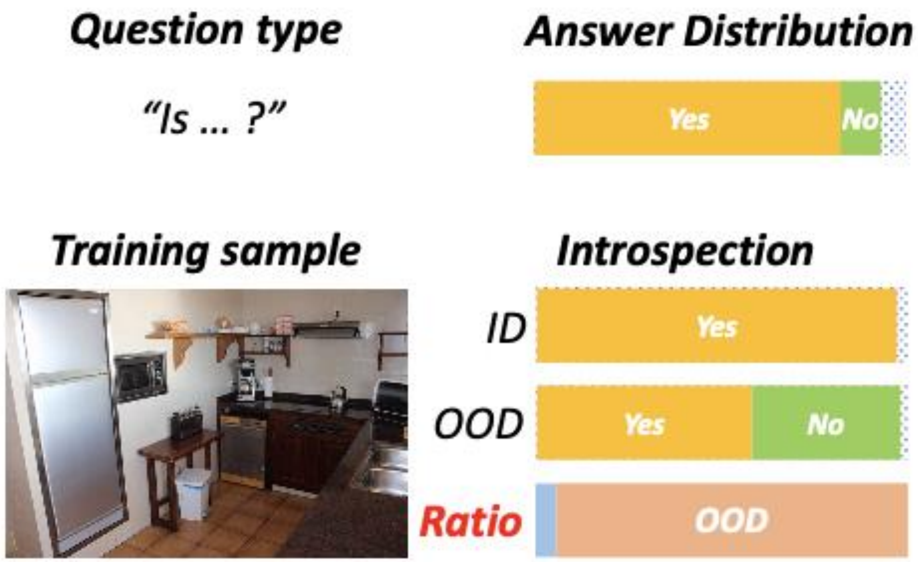


Figure 4: The distribution of w^{ID} on the VQA-CP v2 and VQA v2 training sets.

ID-teacher < OOD-teacher



Q: Is that an electric oven? (GT: Yes.)

How does Introspection look like? VQA has more Case 2 than VQA-CP

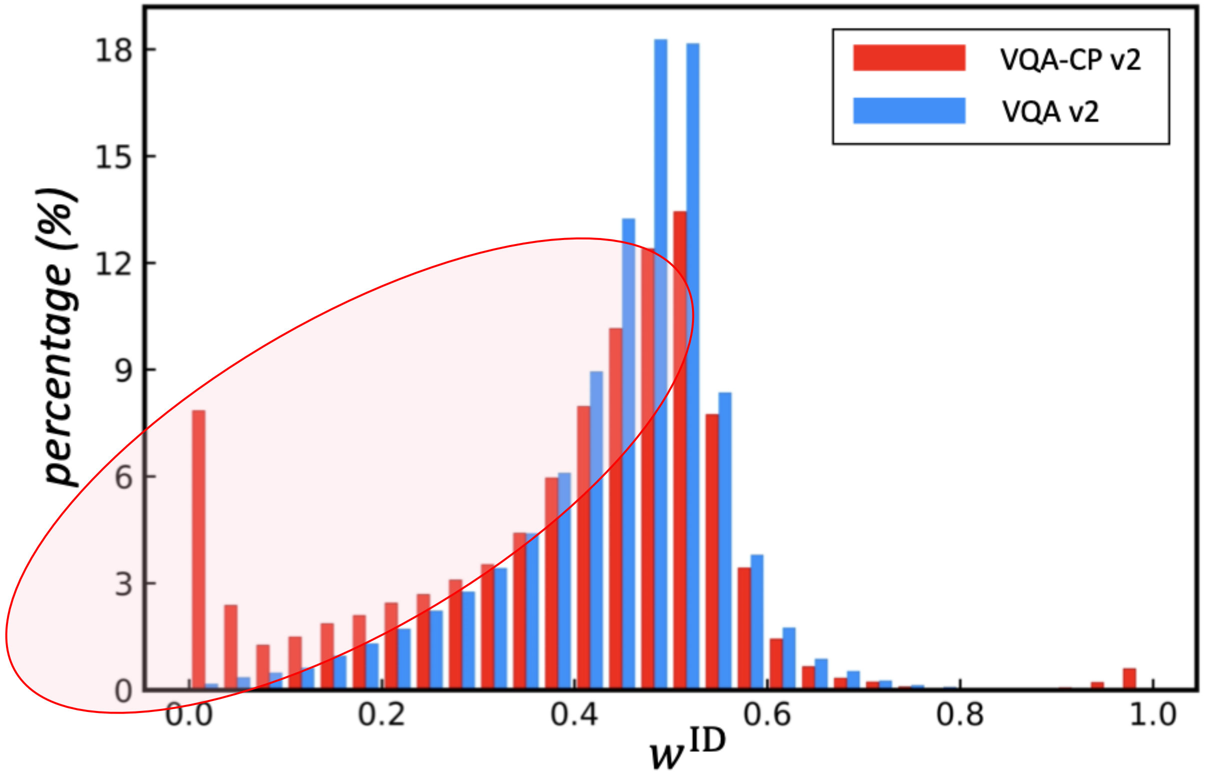
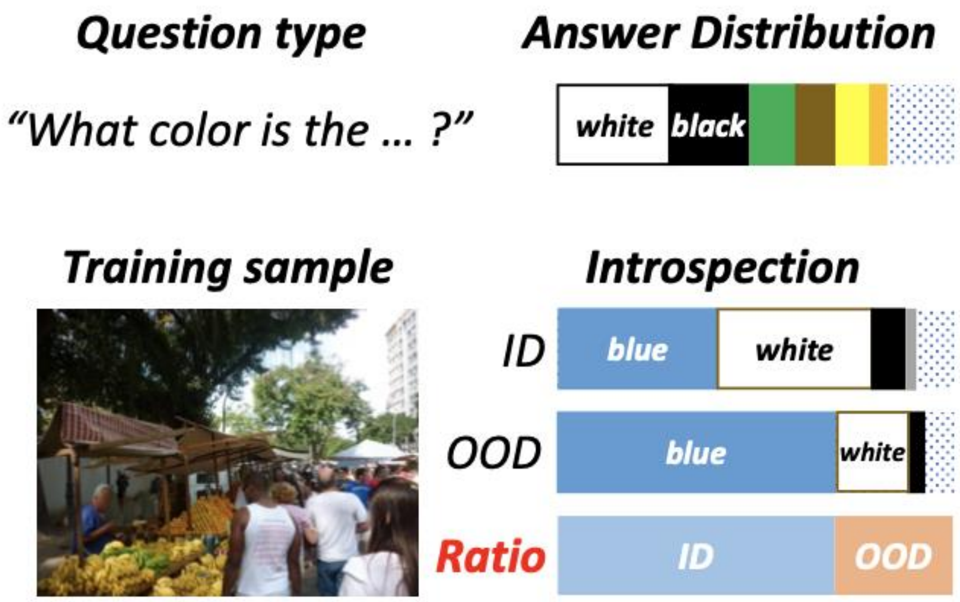


Figure 4: The distribution of w^{ID} on the VQA-CP v2 and VQA v2 training sets.

ID-teacher > OOD-teacher



Q: What color is the older man's shirt? (GT: Blue.)

How does Introspection look like? Both ID-Teachers are weaker (more biased than OOD-Teachers)

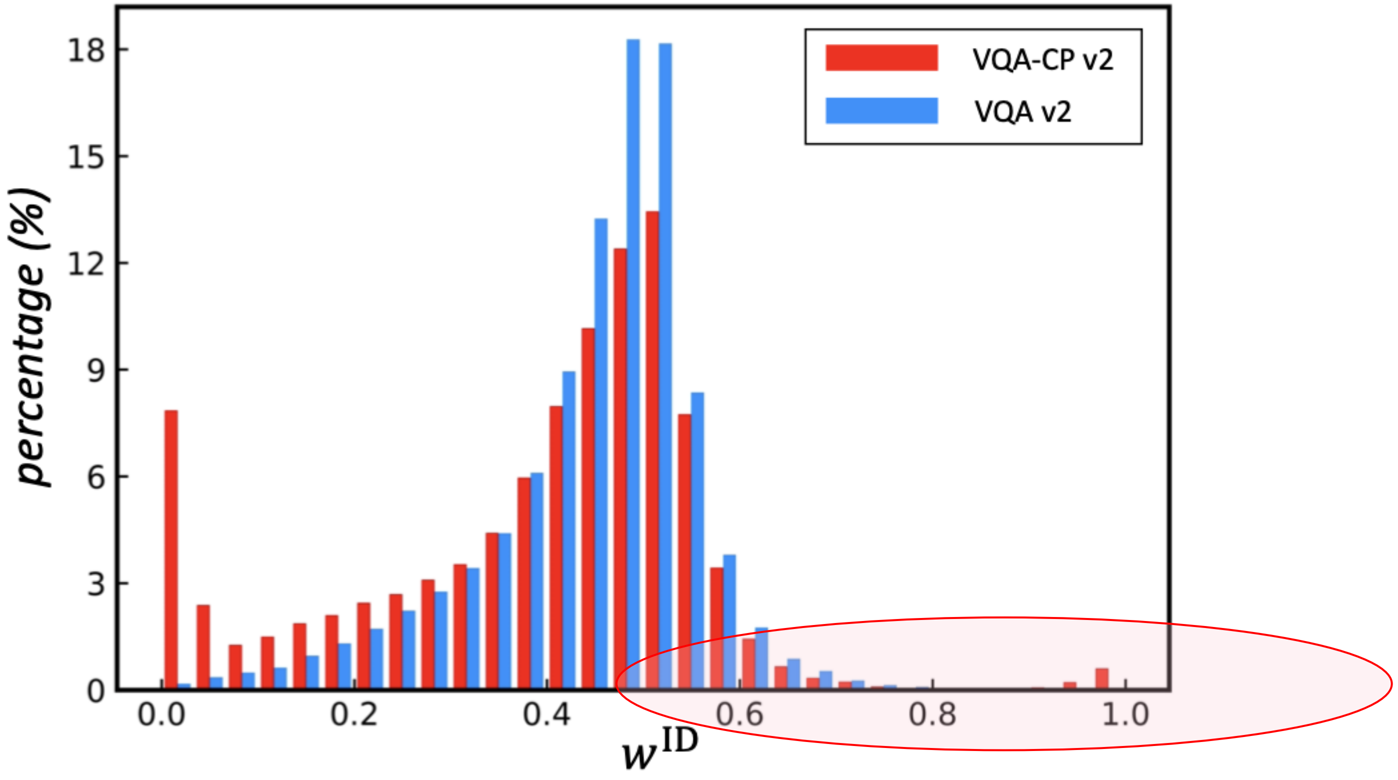


Figure 4: The distribution of w^{ID} on the VQA-CP v2 and VQA v2 training sets.

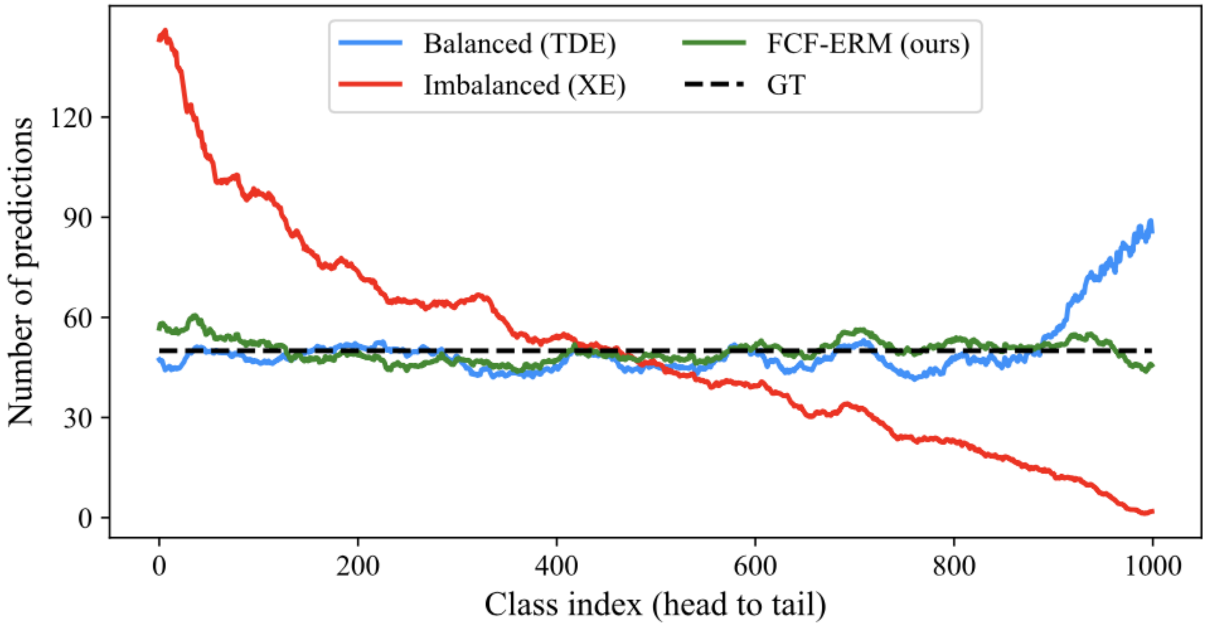
Homework

NANYANG TECHNOLOGICAL UNIVERSITY

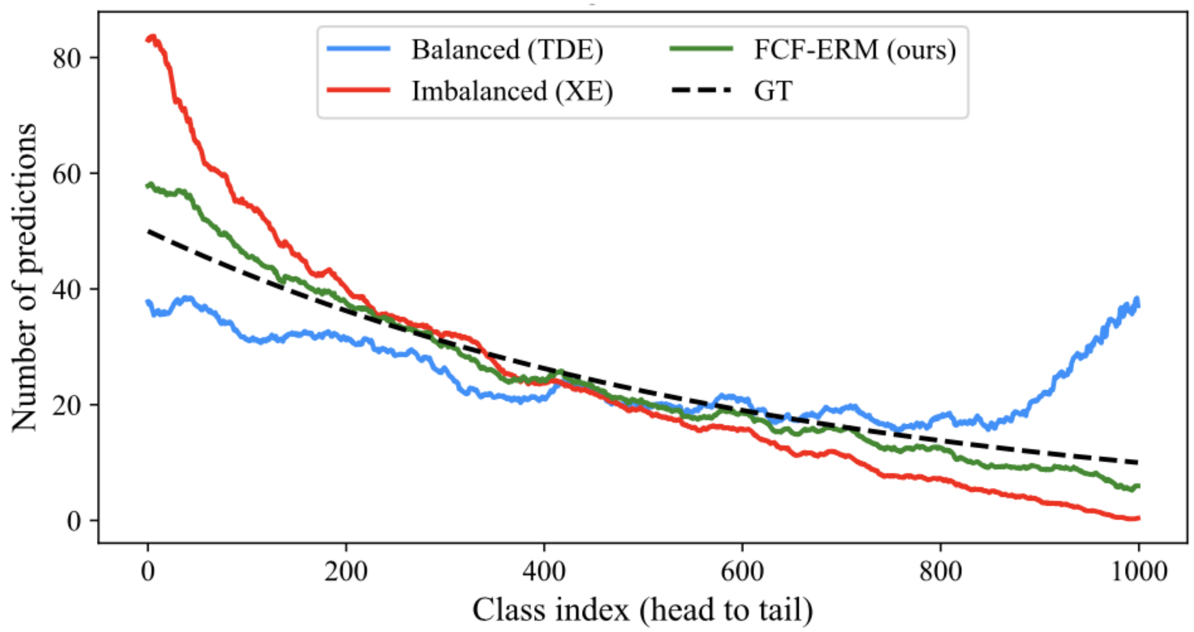
The best of the two worlds

Methods	VQA-CP v2 test (OOD) All	Y/N	Num.	Other	VQA v2 val (ID) All	Y/N	Num.	Other	HM
UpDn [4]	39.79	43.23	12.28	45.54	63.42	81.19	42.43	55.47	48.90
LMH [10]	**52.01**	**72.58**	**31.12**	46.97	56.35	65.06	37.63	54.69	54.09
+ IntroD	51.31 -0.70	71.39	27.13	**47.41**	**62.05** +5.70	**77.65**	**40.25**	**55.97**	**56.17** +2.08
CSS [7]	58.95	84.37	**49.42**	48.21	56.98	65.90	38.19	55.18	57.95
+ IntroD	**60.17** +1.22	**89.17**	46.91	**48.62**	**62.57** +5.59	**78.57**	**41.42**	**56.00**	**61.35** +3.40
S-MRL [6]	37.09	41.39	12.46	41.60	63.12	81.83	45.95	53.43	46.72
RUBi [6]	47.60	70.48	**20.33**	43.09	61.16	81.97	44.86	49.65	53.53
+ IntroD	**48.54** +0.96	**73.94**	19.43	**43.21**	**61.86** +0.70	**82.40**	**45.40**	**50.58**	**54.40** +0.87
RUBi-CF [23]	54.90	90.26	**34.33**	42.01	60.53	81.39	42.87	49.34	57.58
+ IntroD	**54.92** +0.02	**90.84**	25.17	**44.26**	**63.15** +2.62	**82.44**	**45.12**	**53.25**	**58.75** +1.17
CF-VQA [23]	55.05	90.61	**21.50**	45.61	60.94	81.13	43.86	50.11	57.85
+ IntroD	**55.17** +0.12	**90.79**	17.92	**46.73**	**63.40** +2.46	**82.48**	**46.60**	**54.05**	**58.99** +1.14

Current LT is just a "bias flip" game



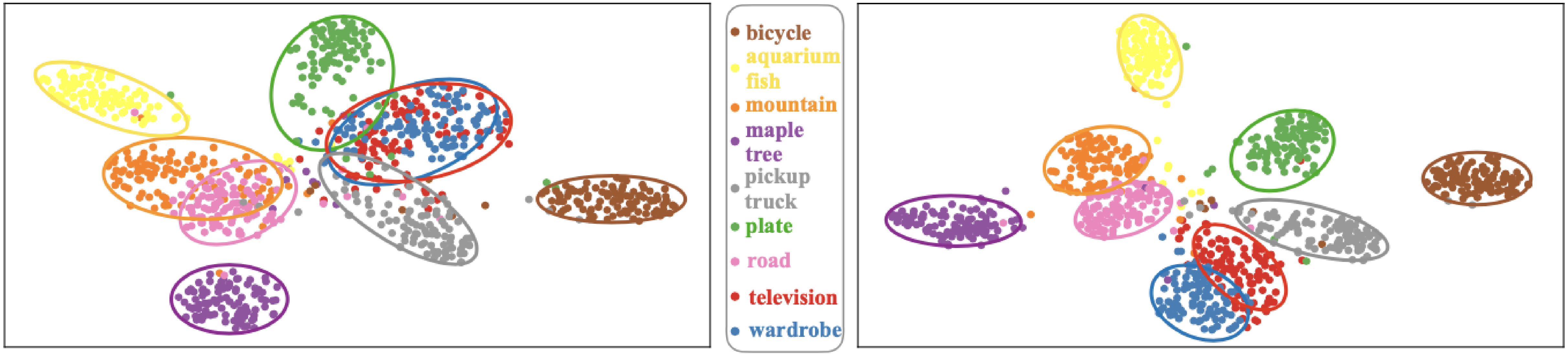
(a) Balanced Test



(b) Imbalanced Test

NANYANG TECHNOLOGICAL UNIVERSITY

So, it does not truly improve the feature



(c) t-SNE of balanced model (TDE)

(d) t-SNE of FCF-ERM (ours)

Factual and Counterfactual ERMs Blend: 3 Steps

Step 1

- Learn a conventional classifier on the imbalanced training data as the ***factual*** model

- Learn a balanced classifier as the ***counterfactual*** model

Factual and Counterfactual ERMs Blend: 3 Steps

Step 2: ER Weights

(Factual ER weight) $$w^{\mathrm{f}} = \frac{(XE^{\mathrm{f}})^{\gamma}}{(XE^{\mathrm{f}})^{\gamma} + (XE^{\mathrm{cf}})^{\gamma}},$$

(Counterfactual ER weight) $$w^{\mathrm{cf}} = 1 - w^{\mathrm{f}} = \frac{(XE^{\mathrm{cf}})^{\gamma}}{(XE^{\mathrm{f}})^{\gamma} + (XE^{\mathrm{cf}})^{\gamma}}.$$

Factual and Counterfactual ERMs Blend: 3 Steps

Step 3: Blended ERM

(Factual ER) $$\mathcal{R}^{\mathrm{f}}(f) = -w^{\mathrm{f}} \sum_i y_i \log f_i(x),$$

where y_i and f_i are the ground-truth and the predicted label for i-th class, respectively.

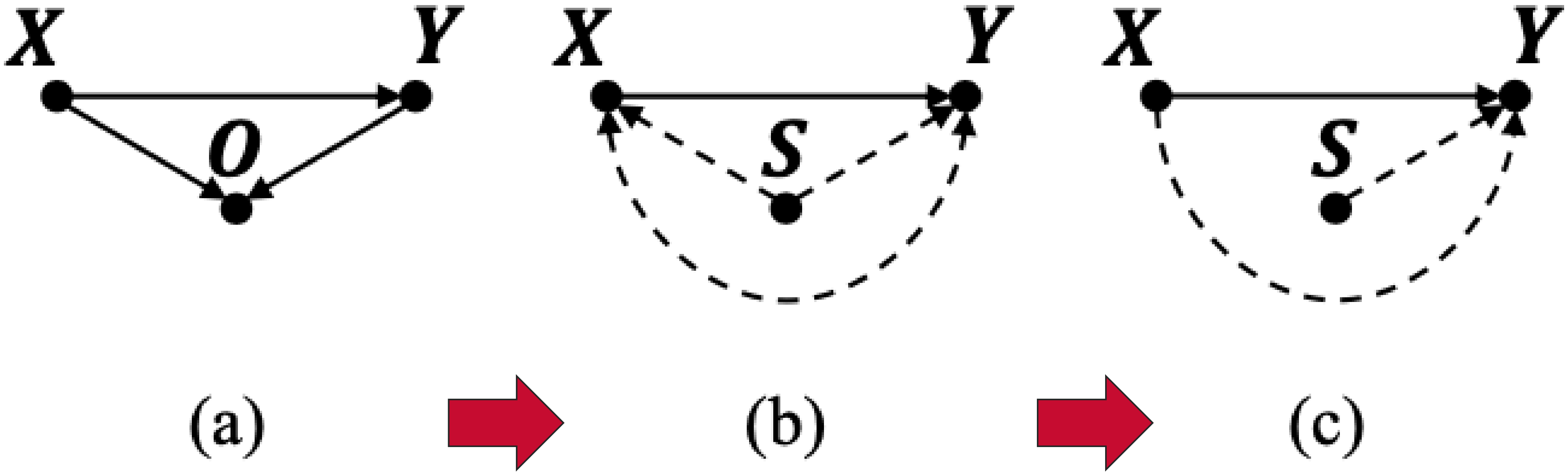
(Counterfactual ER) $$\mathcal{R}^{\mathrm{cf}}(f) = -w^{\mathrm{cf}} \sum_i \hat{y}_i \log f_i(x),$$

where $\hat{y}_i = p^{\mathrm{cf}}(y_i|x)$ denotes the balanced prediction for i-th class.
The overall empirical risk minimization:

$$\mathcal{R}(f) = \mathcal{R}^{\mathrm{f}}(f) + \mathcal{R}^{\mathrm{cf}}(f).$$

NANYANG TECHNOLOGICAL UNIVERSITY

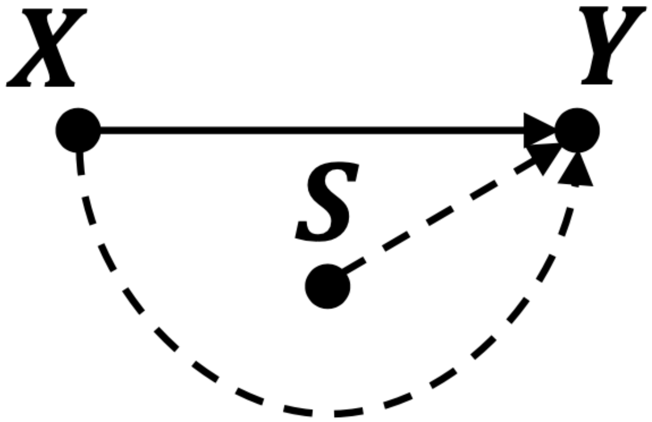
Why? Selection Bias Removal



Reichenbach Principle [raikin-ba:k]

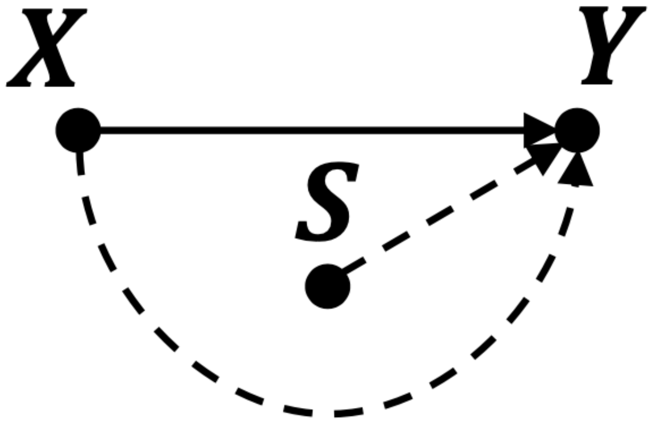
Do-operator

ERM on the Do-modified graph



$$\mathcal{R}(f) = \mathbb{E}_{x\sim P(X),y\sim P(Y|do(X=x))}\mathcal{L}(y, f(x)) = \sum_x \sum_y \mathcal{L}(y, f(x))P(y|do(x))P(x)$$

Backdoor Adjustment: from "interventional" distribution to "observational" distribution



$$P(y|do(x)) = \sum_{S=s\in\{0,1\}} P(y|x, S=s)P(S=s) = \frac{P(x,y,S=1)}{P(x|S=1)} + \frac{P(x,y,S=0)}{P(x|S=0)}.$$

More math

$$\mathcal{R}(f) = \mathbb{E}_{x\sim P(X),y\sim P(Y|do(X=x))}\mathcal{L}(y, f(x)) = \sum_x \sum_y \mathcal{L}(y, f(x))\boxed{P(y|do(x))}P(x)$$

$$P(y|do(x)) = \sum_{S=s\in\{0,1\}} P(y|x, S = s)P(S = s) = \frac{P(x, y, S = 1)}{P(x|S = 1)} + \frac{P(x, y, S = 0)}{P(x|S = 0)}.$$

Overall ERM

(x, y, 1) means factual sample, drawn from training data

(x, y, 0) means cf sample, drawn from balanced model

$$\mathcal{R}(f) = \sum_{(x,y)} \sum_{s\in\{0,1\}} \mathcal{L}(y_s, f(x)) \frac{P(x)}{P(x|S=s)} P(x,y,S=s)$$

$$= \frac{1}{N} \sum_{(x,y)} [\underbrace{\boxed{\mathcal{L}(y_{s=1}, f(x))} \frac{P(x)}{P(x|S=1)}}_{\text{factual ER}} + \underbrace{\boxed{\mathcal{L}(y_{s=0}, f(x))} \frac{P(x)}{P(x|S=0)}}_{\text{counterfactual ER}}].$$

XE loss

Overall ERM: it explains all

$$\mathcal{R}(f) = \sum_{(x,y)} \sum_{s\in\{0,1\}} \mathcal{L}(y_s, f(x)) \frac{P(x)}{P(x|S=s)} P(x,y,S=s)$$

$$= \frac{1}{N} \sum_{(x,y)} [\underbrace{\mathcal{L}(y_{s=1}, f(x)) \boxed{\frac{P(x)}{P(x|S=1)}}}_{\text{factual ER}} + \underbrace{\mathcal{L}(y_{s=0}, f(x)) \boxed{\frac{P(x)}{P(x|S=0)}}}_{\text{counterfactual ER}}].$$

$$w^{\mathrm{f}} = \frac{P(x)}{P(x|S=1)} \propto \frac{(XE^{\mathrm{f}})^\gamma}{(XE^{\mathrm{cf}})^\gamma}, \quad w^{\mathrm{cf}} = \frac{P(x)}{P(x|S=0)} \propto \frac{(XE^{\mathrm{cf}})^\gamma}{(XE^{\mathrm{f}})^\gamma}.$$

The best of the two worlds: balanced test

Methods	Acc	Recall			Precision			F1		
		Many	Med	Few	Many	Med	Few	Many	Med	Few
XE	49.0	68.6	42.9	15.0	46.9	**59.1**	**60.7**	55.7	49.7	24.1
τ-Norm [17]	49.6	61.8	46.2	27.4	52.2	48.5	43.7	56.6	47.3	33.7
LWS [17]	49.9	60.2	47.2	30.3	53.0	49.1	41.3	56.4	48.1	35.0
LADE [13]	51.7	62.6	49.0	30.4	55.3	50.5	41.2	58.7	49.7	34.9
DiVE [11]	53.1	64.1	50.4	**31.5**	-	-	-	-	-	-
DisAlign [40]	53.4	61.3	**52.2**	31.4	-	-	-	-	-	-
PC [13]	48.9	60.4	46.7	23.8	56.3	49.7	32.0	58.3	48.2	27.3
TDE [16]	51.8	62.7	49.0	31.4	**57.3**	52.3	39.5	59.9	50.6	35.0
FCF-ERM$_{PC}$	53.2	67.6	49.8	24.0	53.1	55.0	52.4	59.3	51.9	33.0
FCF-ERM$_{TDE}$	**54.1**	**68.6**	50.0	27.5	53.5	57.3	52.0	**60.1**	**53.4**	**36.0**

The best of the two worlds: imbalanced test

Imbalanced ratio	50	25	10	5
τ-Norm [17]	59.6	58.2	56.2	54.6
LWS [17]	60.6	59.2	57.0	55.0
PC [13]	58.2	56.8	54.5	52.7
LADE [13]	61.8	60.6	58.6	56.8
TDE [16]	63.0	61.6	59.5	57.6
XE	**67.7**	65.2	61.4	58.0
FCF-ERM$_{PC}$	66.8	65.3	62.5	60.1
FCF-ERM$_{TDE}$	**67.7**	**66.0**	**63.5**	**60.9**

NANYANG TECHNOLOGICAL UNIVERSITY

The best of two worlds: improved feature (LT data trained backbone. Normal classification on balanced data

Backbone	Acc	Recall			Precision			F1		
		Many	Med	Few	Many	Med	Few	Many	Med	Few
CIFAR100										
XE (PC [13])	52.6	60.3	51.9	44.4	59.6	51.1	44.4	60.0	51.5	44.4
TDE [16]	52.6	60.4	51.7	44.4	59.5	51.0	44.5	60.0	51.4	44.5
LADE [13]	53.9	58.7	53.8	47.8	60.2	54.5	47.1	59.4	54.1	47.4
FCF-ERM$_{TDE}$	55.1	62.8	54.5	46.7	61.7	53.9	48.1	62.3	54.2	47.4
FCF-ERM$_{PC}$	**55.3**	**60.9**	**56.0**	**48.0**	**63.7**	**54.3**	**48.3**	**62.3**	**55.1**	**48.1**
Places365										
XE (PC [13])	43.8	43.8	44.0	43.5	39.9	43.5	49.3	41.7	43.7	46.2
TDE [16]	43.8	43.8	43.9	43.6	39.7	43.6	48.7	41.6	43.8	46.0
LADE [13]	44.3	42.9	45.9	43.1	43.4	45.1	45.7	43.1	45.5	44.4
FCF-ERM$_{TDE}$	44.6	44.1	45.3	44.0	40.4	44.9	49.5	42.1	45.1	46.6
FCF-ERM$_{PC}$	**46.6**	**45.1**	**48.2**	**46.0**	**44.2**	**49.0**	**53.3**	**44.6**	**48.6**	**49.4**
ImageNet										
XE (PC [13])	56.5	64.5	53.8	43.2	59.8	55.1	50.6	62.1	54.4	46.6
TDE [16]	56.5	64.4	53.8	43.7	60.2	55.2	49.8	62.2	54.5	46.6
LADE [13]	57.9	62.6	55.7	52.2	62.4	56.5	52.9	62.5	56.1	52.5
FCF-ERM$_{TDE}$	58.9	66.5	56.4	46.2	62.1	57.8	**63.2**	64.2	57.1	49.4
FCF-ERM$_{PC}$	**60.2**	**64.8**	**58.2**	**53.8**	**64.9**	**58.3**	53.9	**64.8**	**58.2**	**53.8**

NANYANG TECHNOLOGICAL UNIVERSITY


SOCRATES
PLATO
ARISTOTLE
Q
&
A