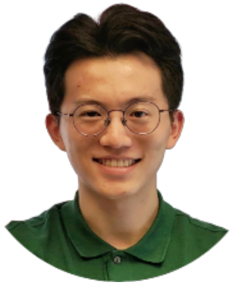
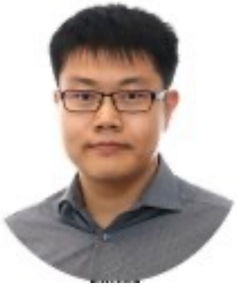
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

NEURAL INFORMATION PROCESSING SYSTEMS

Introspective Distillation for Robust Question Answering

“A training paradigm for robust QA (e.g., Visual QA and extractive QA) models that improve the OOD performance without sacrifice of ID performance."





Yulei Niu

Hanwang Zhang

MReaL Lab
School of Computer Science and Engineering
Nanyang Technological University

Question Answering (QA)

- **Answer** the **question** based on the **context**
  - Visual QA (VQA): vision context --- image
  - Extractive QA: language context --- passage

Q: What is the mustache made of?

A: Banana.

(VQA)

"... Other legislation followed, including the Migratory Bird Conservation Act of 1929, a 1937 treaty prohibiting the hunting of right and gray whales, and the Bald Eagle Protection Act of 1940. These later laws had a low cost to society—the species were relatively rare—and little opposition was raised ..."

Q: Which laws faced significant opposition?

A: Later laws.

(Extractive QA)

Goyal, Yash, et al. **Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering**. CVPR 2017.
Ko, Miyoung, et al. **Look at the First Sentence: Position Bias in Question Answering**. EMNLP 2020.

Training Bias in QA

language prior
correlation between QA pairs

(VQA)

position bias
spurious position cues

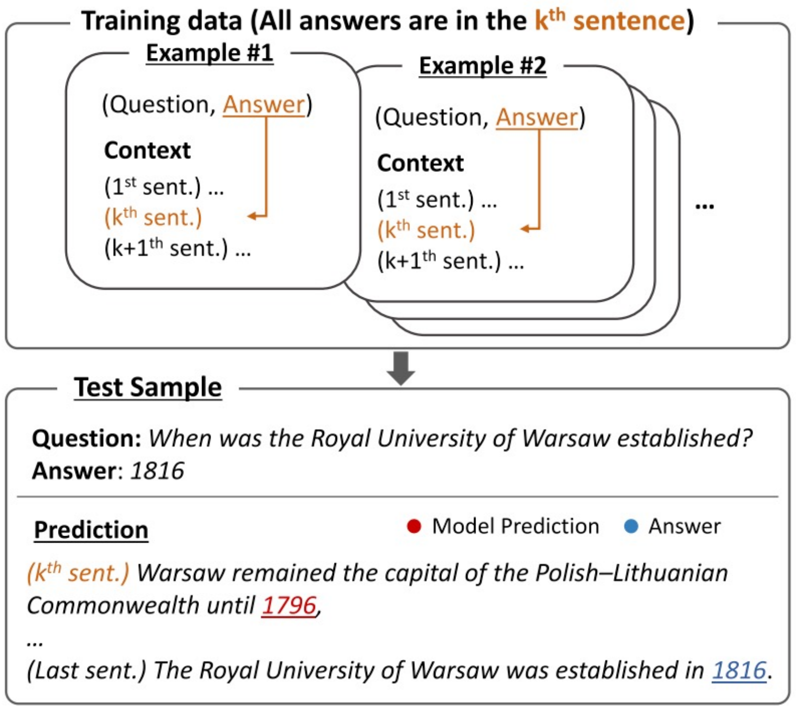
(Extractive QA)

Goyal, Yash, et al. **Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering**. CVPR 2017.
Ko, Miyoung, et al. **Look at the First Sentence: Position Bias in Question Answering**. EMNLP 2020.

Training Bias in QA

language prior
correlation between QA pairs

(VQA)

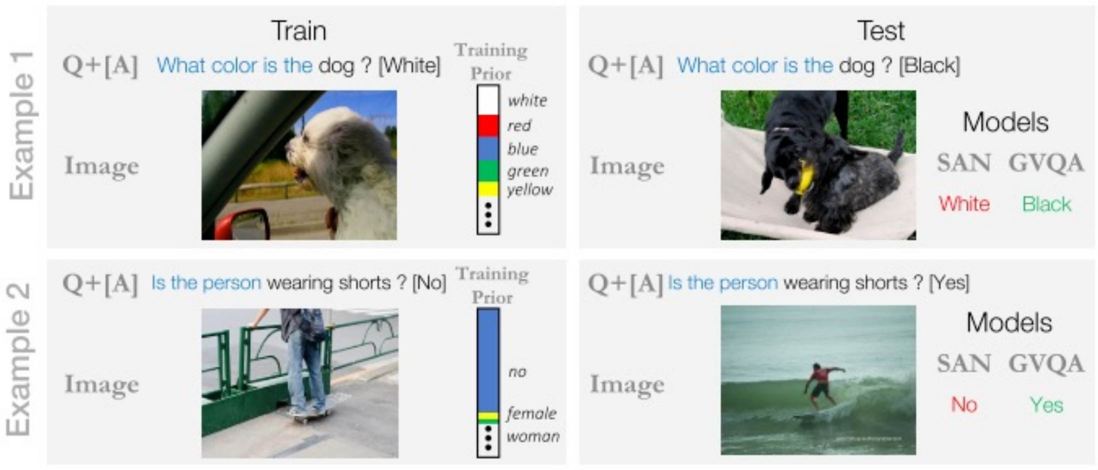


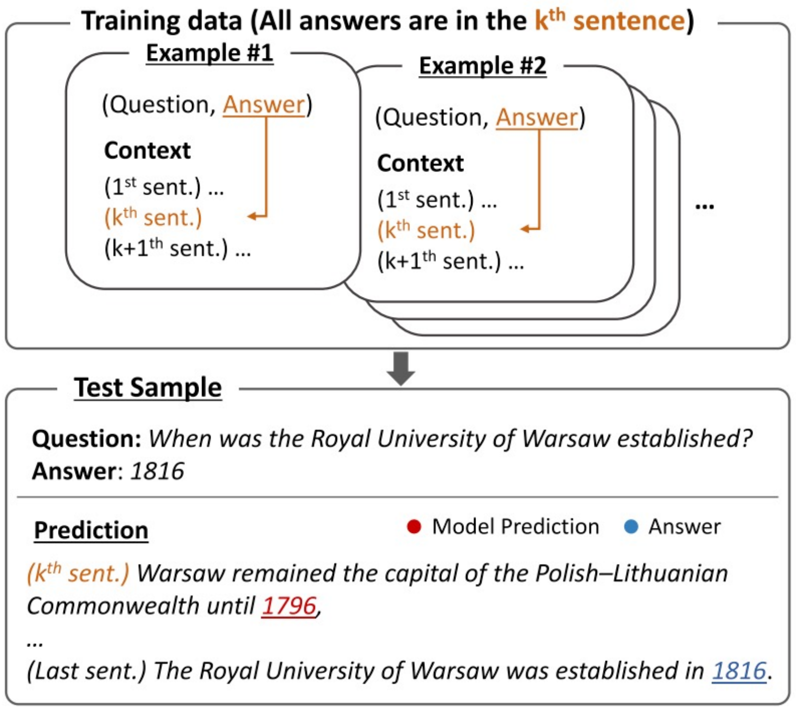
position bias
spurious position cues

(Extractive QA)

Goyal, Yash, et al. **Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering**. CVPR 2017.
Ko, Miyoung, et al. **Look at the First Sentence: Position Bias in Question Answering**. EMNLP 2020.

Training Bias in QA





language prior
correlation between QA pairs

(VQA)

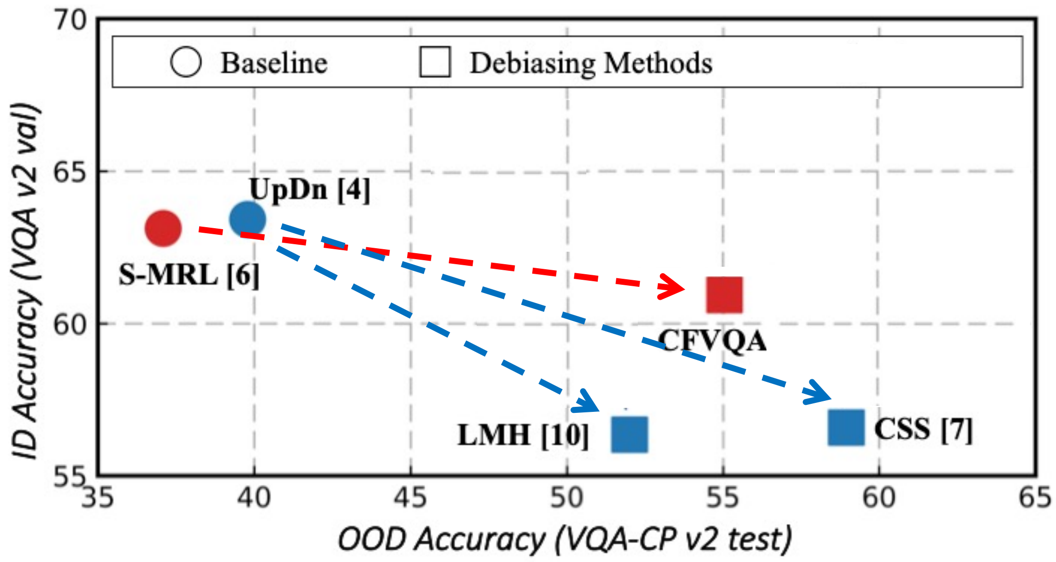
position bias
spurious position cues

(Extractive QA)

Goyal, Yash, et al. **Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering**. CVPR 2017.
Ko, Miyoung, et al. **Look at the First Sentence: Position Bias in Question Answering**. EMNLP 2020.

Overcoming Training Bias in QA

- Debiasing VQA Methods
  - Assume that training and test distribution are very different or even reversed
  - Improve out-of-distribution (OOD) performance by large margins 😊
  - Decrease in-distribution (ID) performance 😭

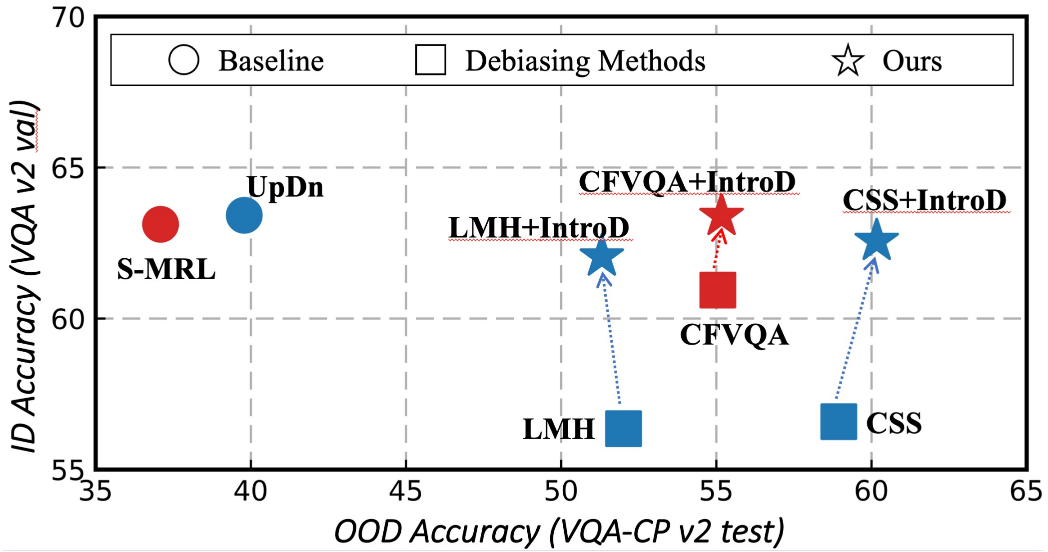


Goyal, Yash, et al. **Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering**. CVPR 2017.
Ko, Miyoung, et al. **Look at the First Sentence: Position Bias in Question Answering**. EMNLP 2020.

Overcoming Training Bias in QA

- Debiasing VQA Methods
  - Assume that training and test distribution are very different or even reversed
  - Improve out-of-distribution (OOD) performance by large margins 😊
  - Decrease in-distribution (ID) performance 😭

- Can we make the best of both worlds?
  - Yes! We did in this paper!



Goyal, Yash, et al. **Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering**. CVPR 2017.
Ko, Miyoung, et al. **Look at the First Sentence: Position Bias in Question Answering**. EMNLP 2020.

Ours: Introspective Distillation (IntroD)

- What happened?
  - Over-exploiting ID (OOD) inductive bias -> degraded OOD (ID) performance

Ours: Introspective Distillation (IntroD)

- What happened?
  - Over-exploiting ID (OOD) inductive bias -> degraded OOD (ID) performance
- How to solve?
  - Blend the ID and OOD inductive bias fairly

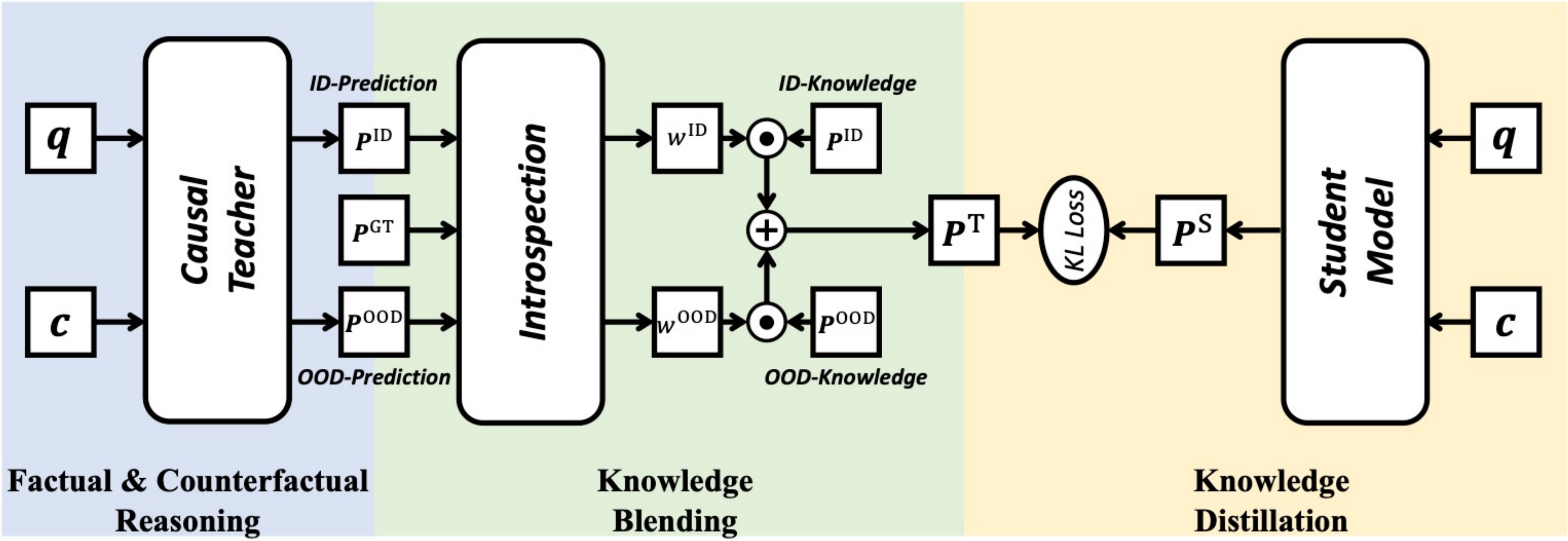
Ours: Introspective Distillation (IntroD)

- What happened?
  - Over-exploiting ID (OOD) inductive bias -> degraded OOD (ID) performance

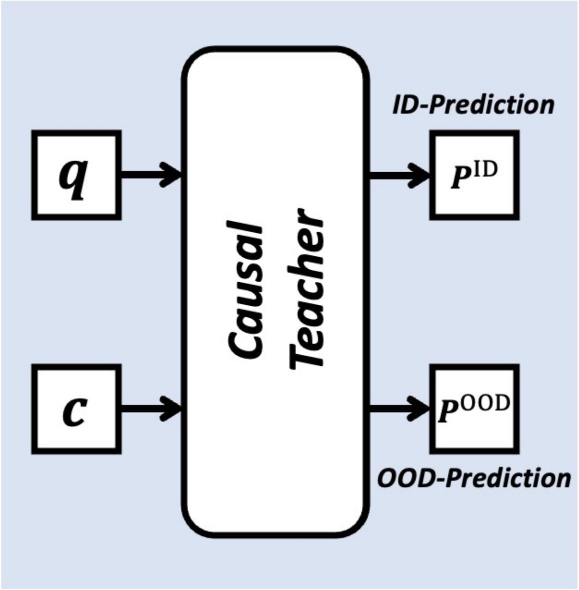
- How to solve?
  - Blend the ID and OOD inductive bias fairly

- How to implement?
  - Obtain ID-teacher and OOD-teacher
  - Introspect whether ID (OOD) bias dominates the learning
  - Blend the knowledge of ID-teacher and OOD-teacher
  - Distill the knowledge to a student

Training Paradigm



Step 1: Factual & Counterfactual Reasoning

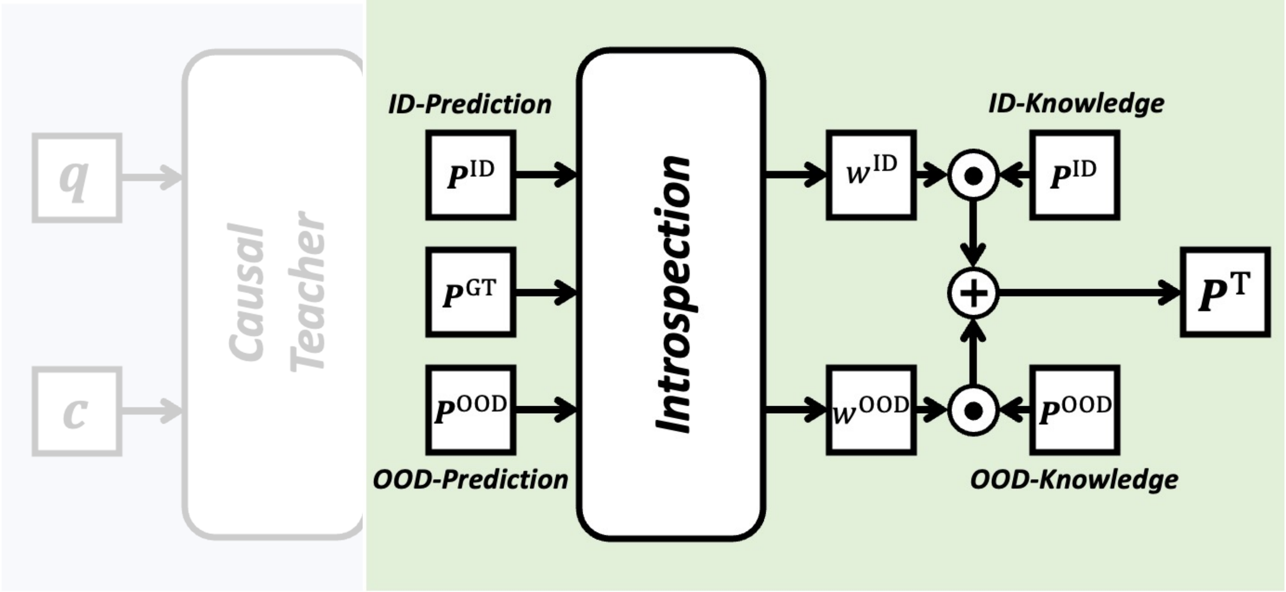


q question
c context
(image in VQA, passage in extractive QA)

- Obtain ID-teacher and OOD-teacher
  - Depict ID and OOD worlds, respectively
- Implemented as the same causal model [Niu et al, 2021]
  - es(Total Effect)
    - Include shortcut bias (Q->A in VQA, C->A in extractive QA)
  - Counterfactual reasoning -> OOD-teacher (Indirect Effect)
    - Eliminate shortcut bias

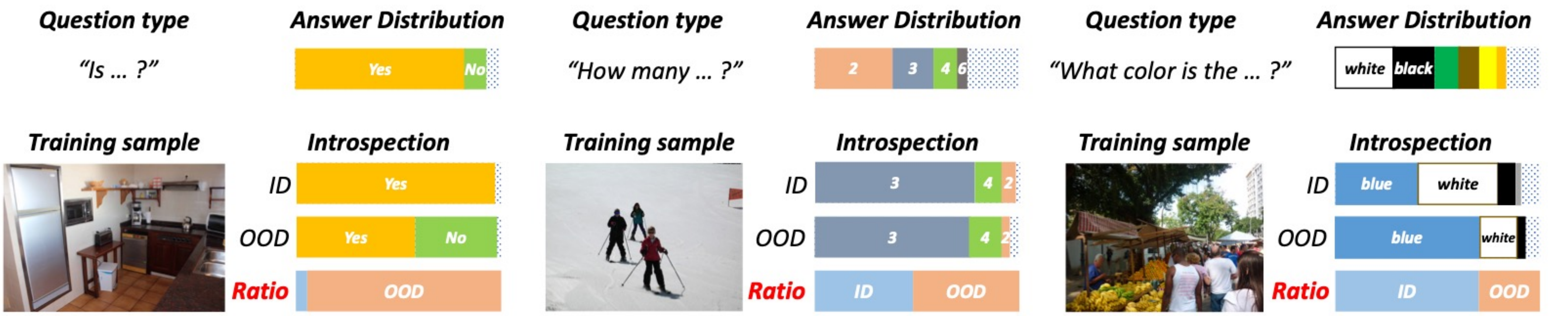
Niu, Yulei, et al. **Counterfactual VQA: A Cause-Effect Look at Language Bias**. CVPR 2021.

Step 2: Knowledge Blending



- Examine whether the inductive bias is over-exploited
- Blend ID and OOD inductive bias fairly

Step 2: Knowledge Blending



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

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

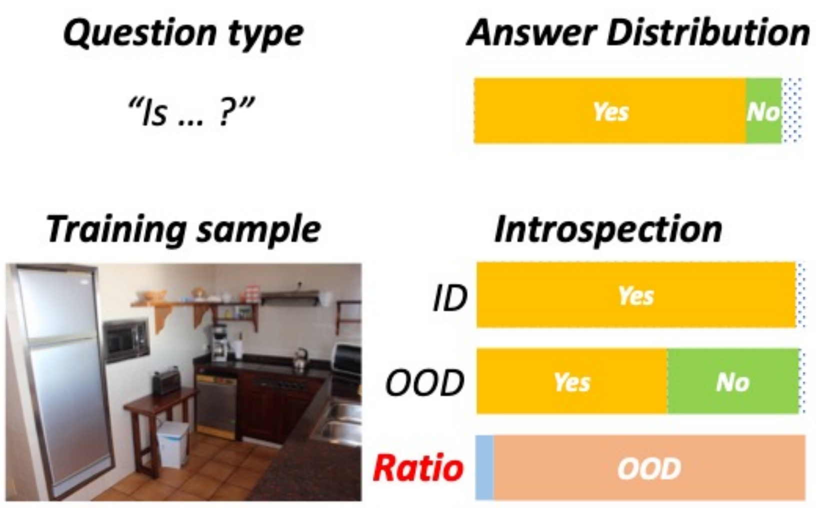
Q: What color is the older man's shirt? (GT: Blue.)

ID-bias > OOD-bias
↓
ID-teacher < OOD-teacher

ID-bias ≈ OOD-bias
↓
ID-teacher ≈ OOD-teacher

ID-bias < OOD-bias
↓
ID-teacher > OOD-teacher

Step 2: Knowledge Blending



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

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

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

ID-bias > OOD-bias
↓
ID-teacher < OOD-teacher

ID-bias ≈ OOD-bias
↓
ID-teacher ≈ OOD-teacher

ID-bias < OOD-bias
↓
ID-teacher > OOD-teacher

Step 2: Knowledge Blending

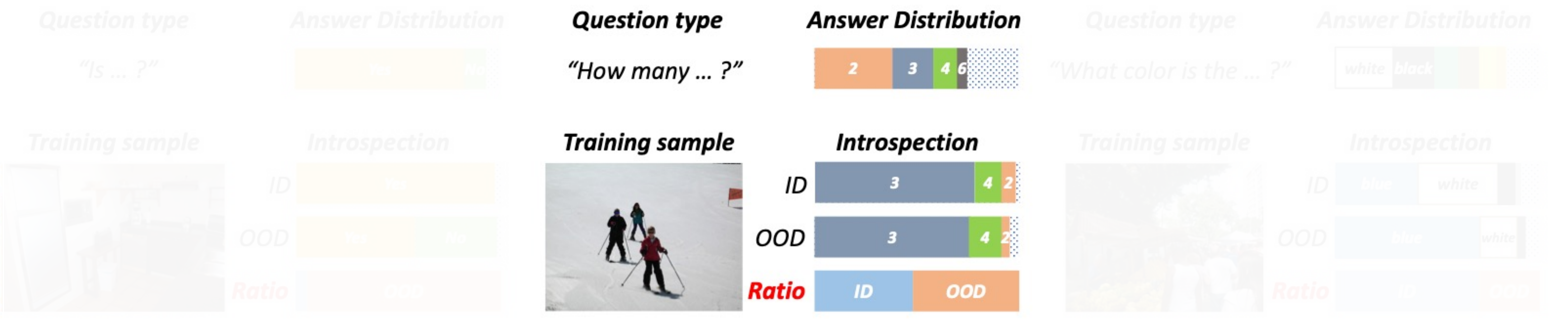
Q: <u>Is</u> that an electric oven? (GT: Yes.) Q: <u>How many</u> skiers? (GT: 3.) Q: <u>What color is the</u> older man's shirt? (GT: Blue.)

ID-bias > OOD-bias
↓
ID-teacher < OOD-teacher

ID-bias ≈ OOD-bias
↓
ID-teacher ≈ OOD-teacher

ID-bias < OOD-bias
↓
ID-teacher > OOD-teacher

Step 2: Knowledge Blending



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

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

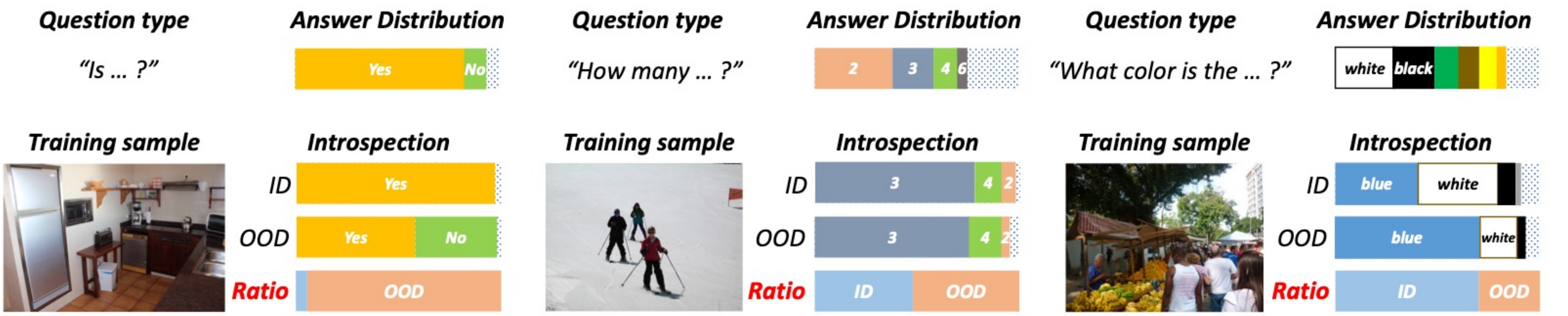
Q: What color is the older man's shirt? (GT: Blue.)

ID-bias > OOD-bias
↓
ID-teacher < OOD-teacher

ID-bias ≈ OOD-bias
↓
ID-teacher ≈ OOD-teacher

ID-bias < OOD-bias
↓
ID-teacher > OOD-teacher

Step 2: Knowledge Blending



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

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

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

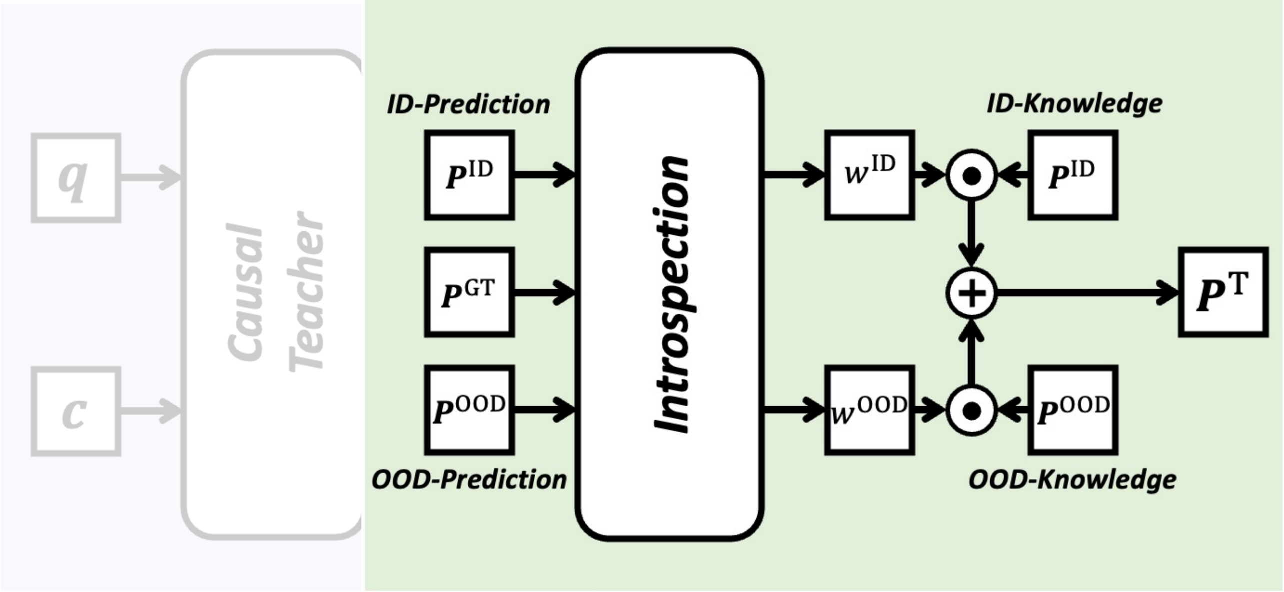
ID-bias > OOD-bias
↓
ID-teacher < OOD-teacher

ID-bias ≈ OOD-bias
↓
ID-teacher ≈ OOD-teacher

ID-bias < OOD-bias
↓
ID-teacher > OOD-teacher

larger confidence -> smaller weight

Step 2: Knowledge Blending



Confidence

$$s^{\mathrm{ID}} = \frac{1}{XE(\boldsymbol{P}^{\mathrm{GT}}, \boldsymbol{P}^{\mathrm{ID}})} = \frac{1}{\sum_{a\in\mathcal{A}} -\boldsymbol{P}^{\mathrm{GT}}(a)\log \boldsymbol{P}^{\mathrm{ID}}(a)},$$

$$s^{\mathrm{OOD}} = \frac{1}{XE(\boldsymbol{P}^{\mathrm{GT}}, \boldsymbol{P}^{\mathrm{OOD}})} = \frac{1}{\sum_{a\in\mathcal{A}} -\boldsymbol{P}^{\mathrm{GT}}(a)\log \boldsymbol{P}^{\mathrm{OOD}}(a)},$$

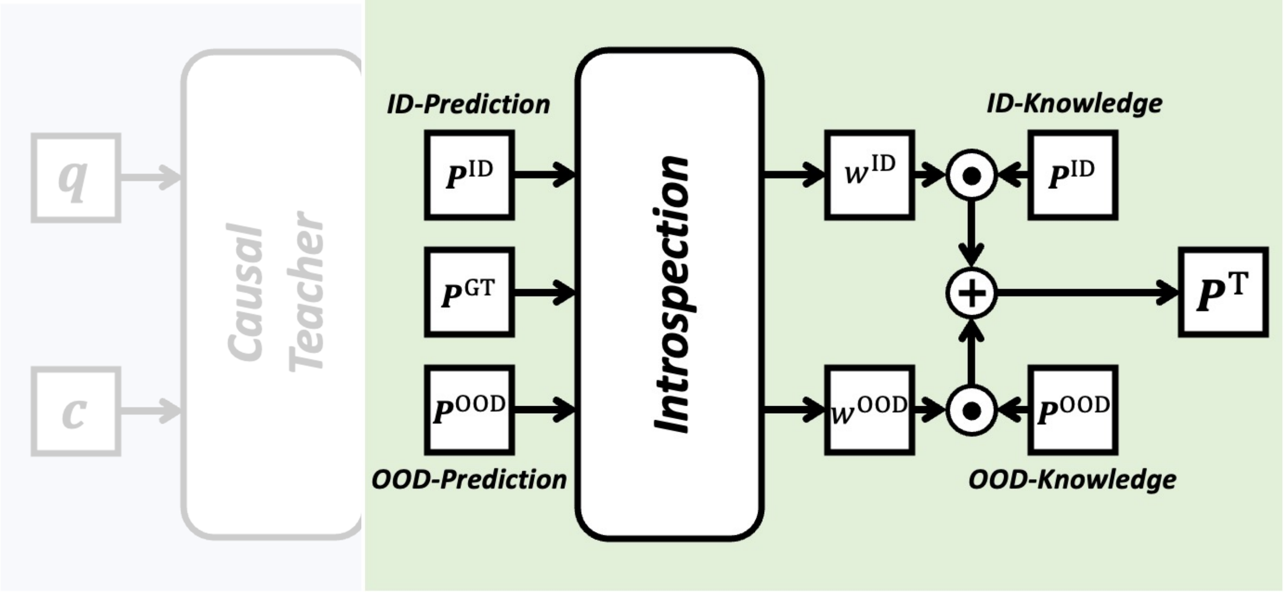
Weight

$$w^{\mathrm{ID}} = \frac{(s^{\mathrm{ID}})^{-1}}{(s^{\mathrm{ID}})^{-1} + (s^{\mathrm{OOD}})^{-1}} = \frac{s^{\mathrm{OOD}}}{s^{\mathrm{ID}} + s^{\mathrm{OOD}}},$$

$$w^{\mathrm{OOD}} = 1 - w^{\mathrm{ID}}.$$

- Examine whether the inductive bias is over-exploited

Step 2: Knowledge Blending



Confidence

$$s^{\mathrm{ID}} = \frac{1}{XE(\boldsymbol{P}^{\mathrm{GT}}, \boldsymbol{P}^{\mathrm{ID}})} = \frac{1}{\sum_{a\in\mathcal{A}} -\boldsymbol{P}^{\mathrm{GT}}(a)\log \boldsymbol{P}^{\mathrm{ID}}(a)},$$

$$s^{\mathrm{OOD}} = \frac{1}{XE(\boldsymbol{P}^{\mathrm{GT}}, \boldsymbol{P}^{\mathrm{OOD}})} = \frac{1}{\sum_{a\in\mathcal{A}} -\boldsymbol{P}^{\mathrm{GT}}(a)\log \boldsymbol{P}^{\mathrm{OOD}}(a)},$$

Weight

$$w^{\mathrm{ID}} = \frac{(s^{\mathrm{ID}})^{-1}}{(s^{\mathrm{ID}})^{-1} + (s^{\mathrm{OOD}})^{-1}} = \frac{s^{\mathrm{OOD}}}{s^{\mathrm{ID}} + s^{\mathrm{OOD}}},$$

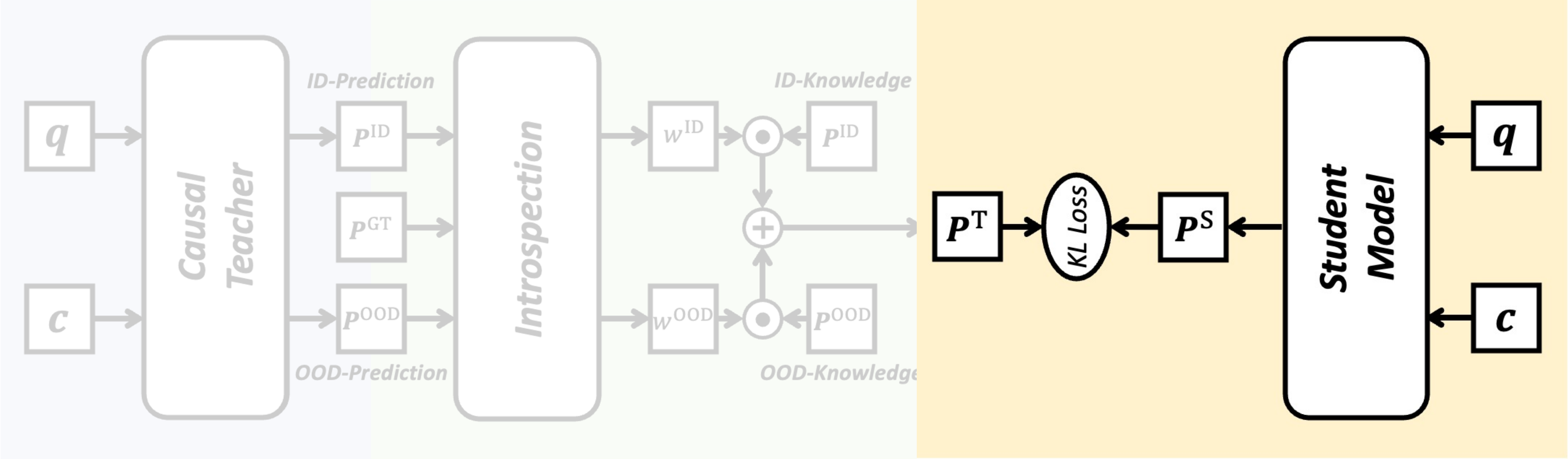
$$w^{\mathrm{OOD}} = 1 - w^{\mathrm{ID}}.$$

- Examine whether the inductive bias is over-exploited
- Blend ID and OOD inductive bias fairly

$$\boldsymbol{P}^{\mathrm{T}} = w^{\mathrm{ID}} \cdot \textbf{ID-Knowledge} + w^{\mathrm{OOD}} \cdot \textbf{OOD-Knowledge}.$$

  - ID-Knowledge: ground-truth labels; OOD-knowledge: OOD-prediction

Step 3: Knowledge Distillation

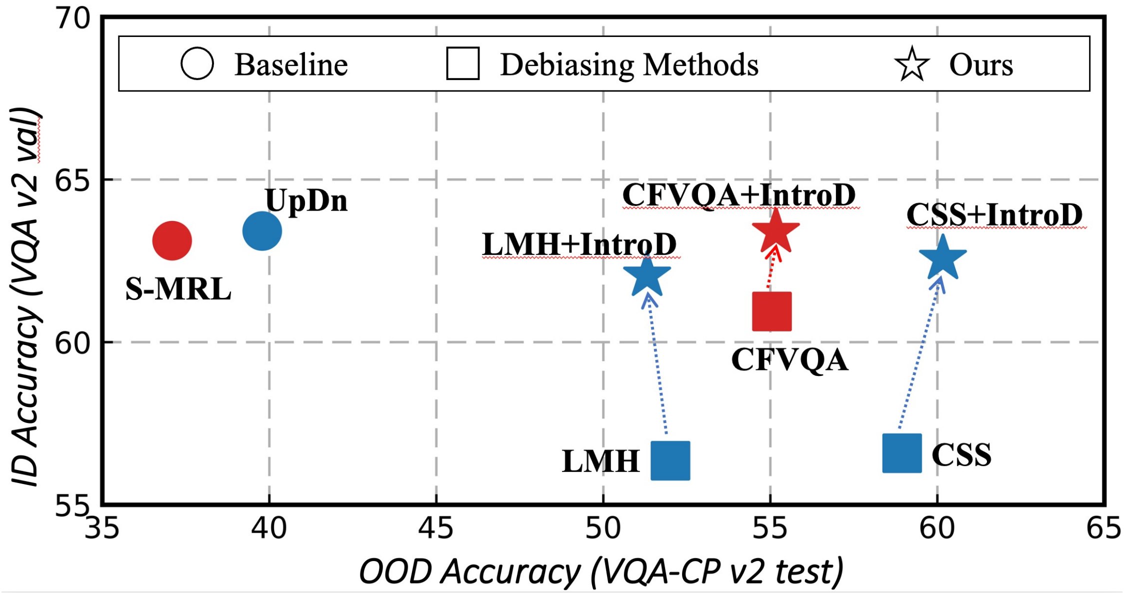


- Distill the blended knowledge to a student model
  - Same architecture with teacher model without shortcut branches (Q->A, C->A)

$$\mathcal{L} = KL(\boldsymbol{P}^{\mathrm{T}}, \boldsymbol{P}^{\mathrm{S}}) = \sum_{a \in \mathcal{A}} \boldsymbol{P}^{\mathrm{T}}(a) \log \frac{\boldsymbol{P}^{\mathrm{T}}(a)}{\boldsymbol{P}^{\mathrm{S}}(a)}$$

Experiments

- Visual QA
  - Bias: language prior
  - VQA v2, VQA-CP v2
- Extractive QA
  - Bias: position bias
  - SQuAD



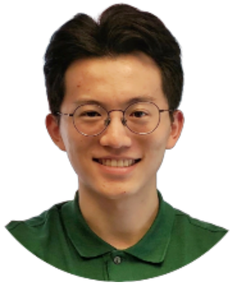
Methods	SQuAD$_{dev}^{k=1}$ (ID) EM	SQuAD$_{dev}^{k=1}$ (ID) F1	SQuAD$_{dev}^{k\neq1}$ (OOD) EM	SQuAD$_{dev}^{k\neq1}$ (OOD) F1	SQuAD$_{dev}$ (All) EM	SQuAD$_{dev}$ (All) F1
XLNet	79.65	87.48	30.17	35.91	47.20	53.65
LM [10]	78.31	85.97	61.04	**69.49**	66.98	75.16
+ **IntroD**	**81.08** +2.77	**88.55** +2.58	**61.52** +0.48	68.84 -0.65	**68.25** +1.27	**75.62** +0.46
BERT	77.87	86.41	10.95	16.17	33.95	40.34
LM [10]	77.18	85.15	71.31	79.79	73.33	81.64
+ **IntroD**	**79.21** +2.03	**87.04** +1.89	**72.14** +0.83	**79.97** +0.18	**74.58** +1.25	**82.40** +0.76

Goyal, Yash, et al. **Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering**. CVPR 2017.
Ko, Miyoung, et al. **Look at the First Sentence: Position Bias in Question Answering**. EMNLP 2020.

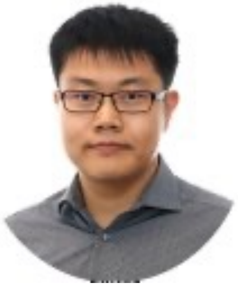
Conclusion

- Can we achieve the best of both ID and OOD worlds? Yes.
- The selection of ID-Knowledge? Ground-truth annotations are better than ID-prediction.
- Can the student learn more from the more (rather than less) confident teacher? No.
- Can the student equally learn from ID and OOD teachers? No.
- Can the student only learn from OOD-teacher? Yes, but worse than our IntroD.
- Is our IntroD a simple ensemble method? No.

Thank you for listening!



<u>Yulei Niu</u>



Hanwang Zhang