CleanPose: Category-Level Object Pose Estimation via Causal Learning and Knowledge Distillation






Xiao Lin[1] Yun Peng[1] Liuyi Wang[1] Minghao Zhu[1] Chengju Liu[1,2] Qijun Chen[1,2]

[1]College of Electronic and Information Engineering, Tongji University, Shanghai, China

[2]State Key Laboratory of Autonomous Intelligent Unmanned Systems, China

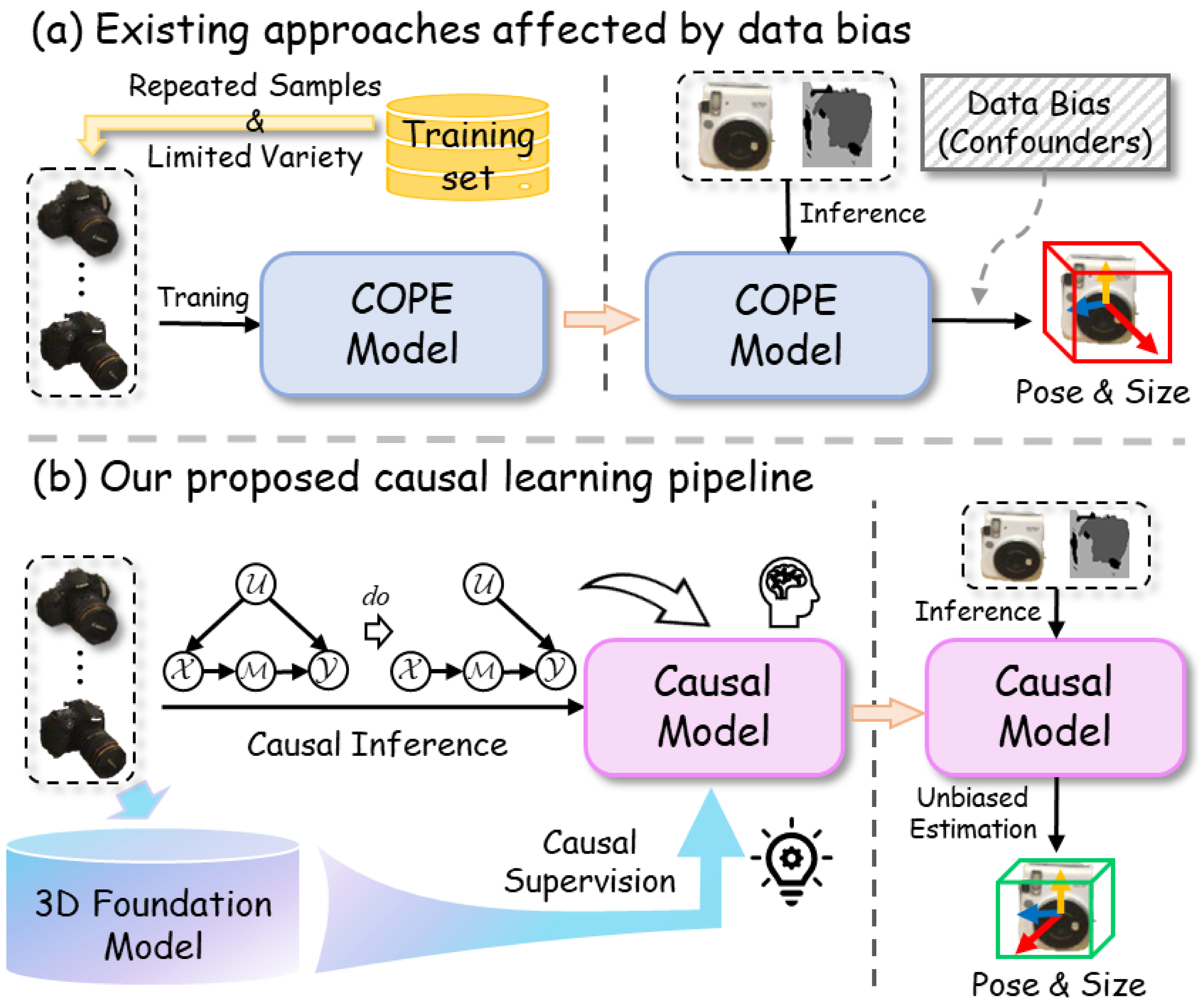
October 21st, 2025

➤ Motivation

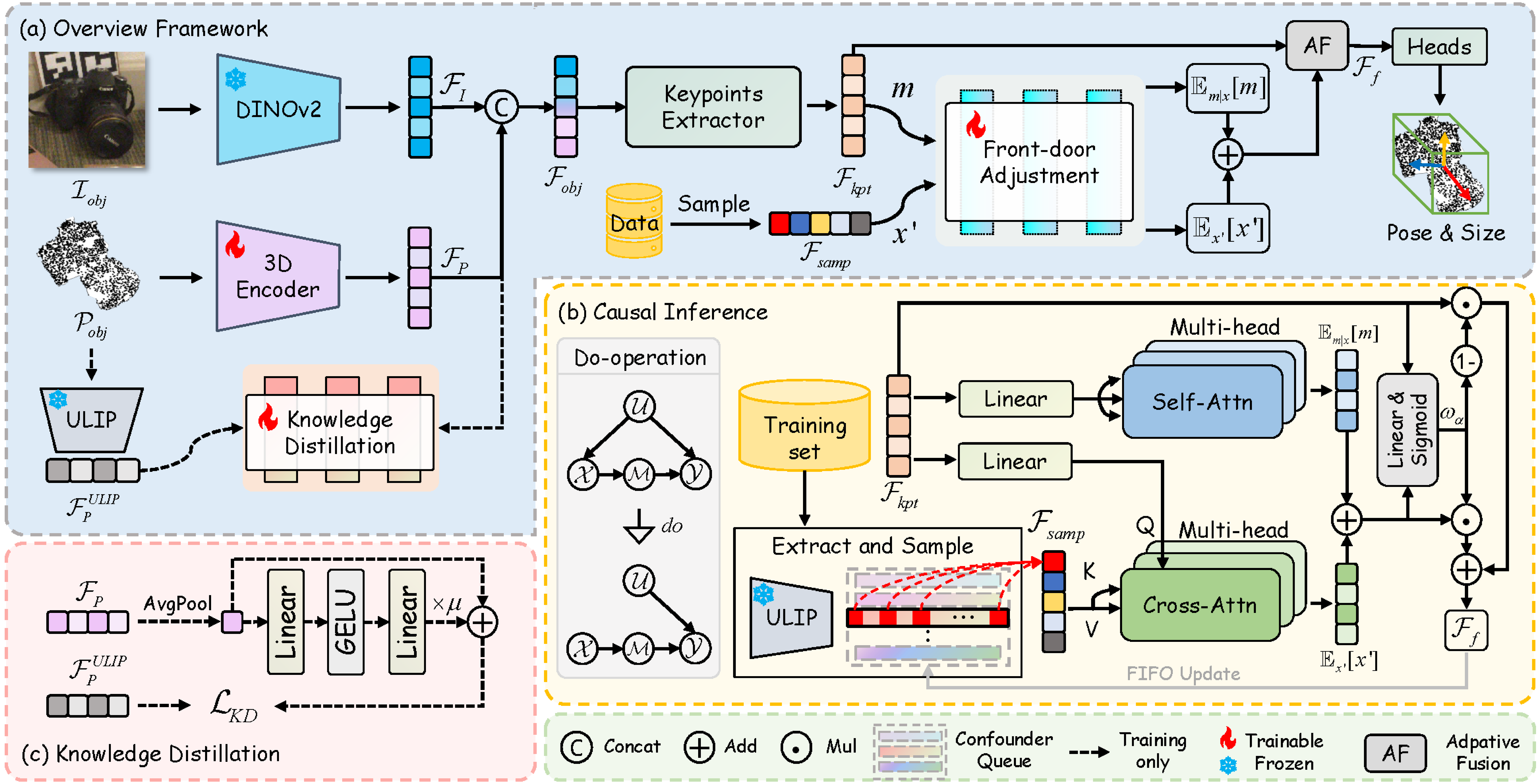
- The **inherent biases** are pervasive in current datasets, *e.g.*, repeated training samples and limited pose variety, which may mislead COPE models to overfit to familiar object's appearance and poses during data fitting.

- The dataset's scale is still significantly constrained by the cost of 3D data annotation. Moreover, achieving a perfectly balanced dataset free of bias remains nearly impossible.

- Therefore, the extension of datasets does not fundamentally solve the hindrance, developing a causal COPE models that can effectively confront and alleviate biases becomes a primary challenge.

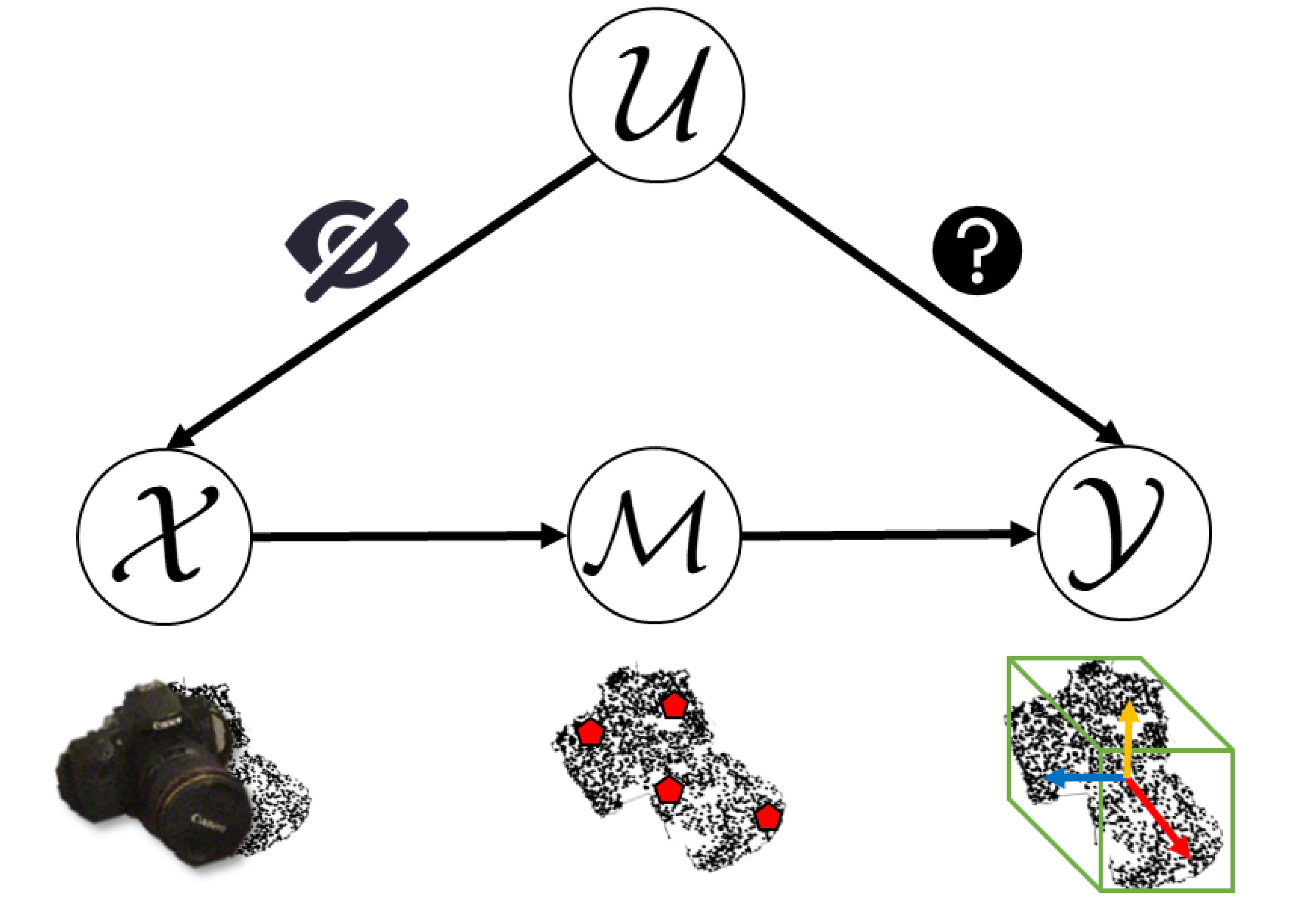


➢ Contributions



• For the first time, we propose to leverage **causal inference** and a residual-based **knowledge distillation** to alleviate the negative effects raised by <u>confounders.</u>

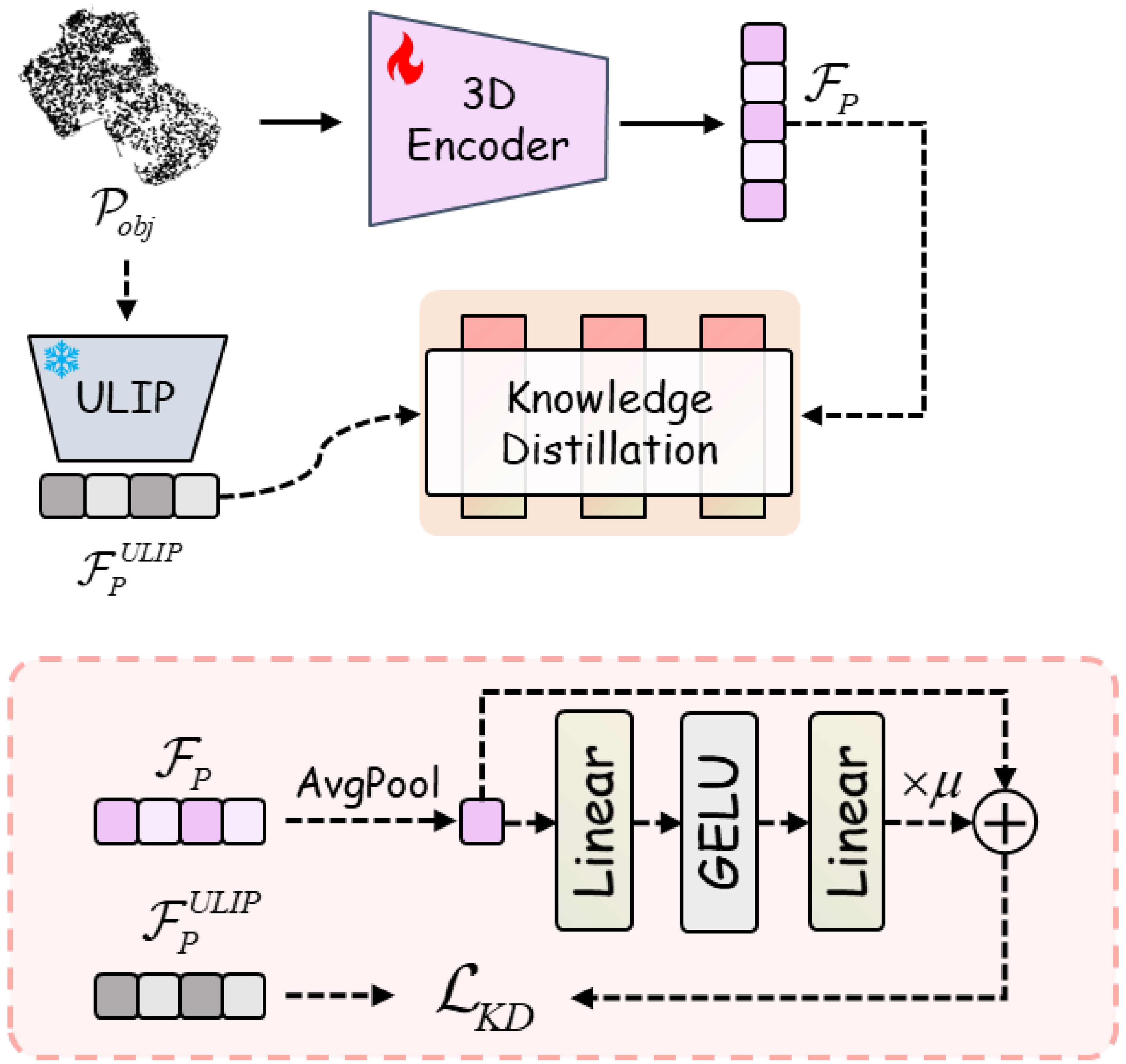
Structural Causal Model



- $\mathcal{X} \rightarrow \mathcal{M} \rightarrow \mathcal{Y}$(*front door path*): Human first recognize the keypoints, and then determine the pose.

- $\mathcal{X} \leftarrow \mathcal{U} \rightarrow \mathcal{Y}$(*hidden confounders*): The confounders are extraneous variables that influence both inputs and outputs.

Knowledge Distillation



$$\mathcal{L}_{KD} = \frac{1}{B}\sum_{i}^{B}\left\| \mathcal{F}_P^{ULIP} - \varphi\left(\widehat{\mathcal{F}}_P^{avg}\right)\right\|_2$$

Front-door Adjustment Causal Inference

$$P(\mathcal{Y}, \mathcal{X}, m, u) = P(u)P(m \mid \mathcal{X})P(\mathcal{Y} \mid m, u)P(\mathcal{X} \mid u)$$

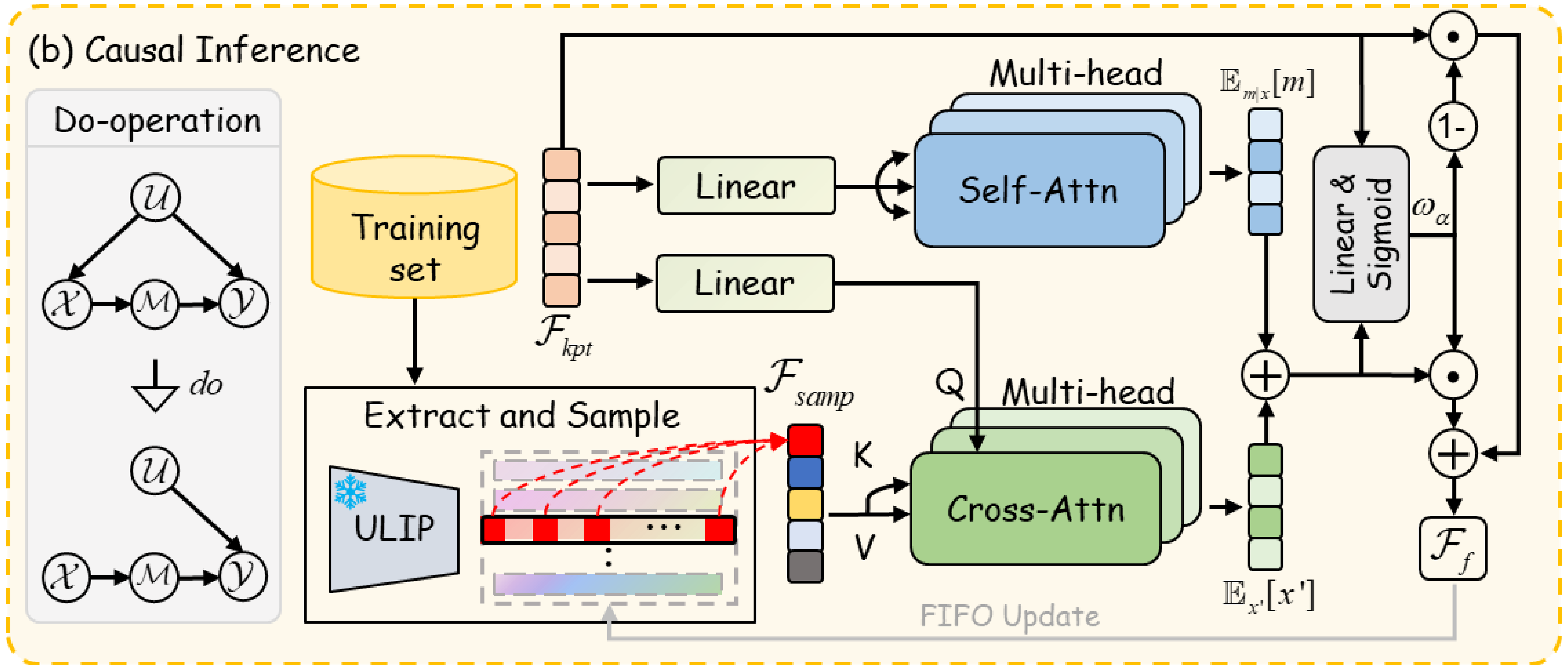
$$P(\mathcal{Y}, m, u \mid do(\mathcal{X})) = P(u)P(m \mid \mathcal{X})P(\mathcal{Y} \mid m, u)$$

$$P(\mathcal{Y} \mid do(\mathcal{X})) = \sum_m P(m \mid \mathcal{X}) \sum_u P(\mathcal{Y} \mid m, u)P(u)$$

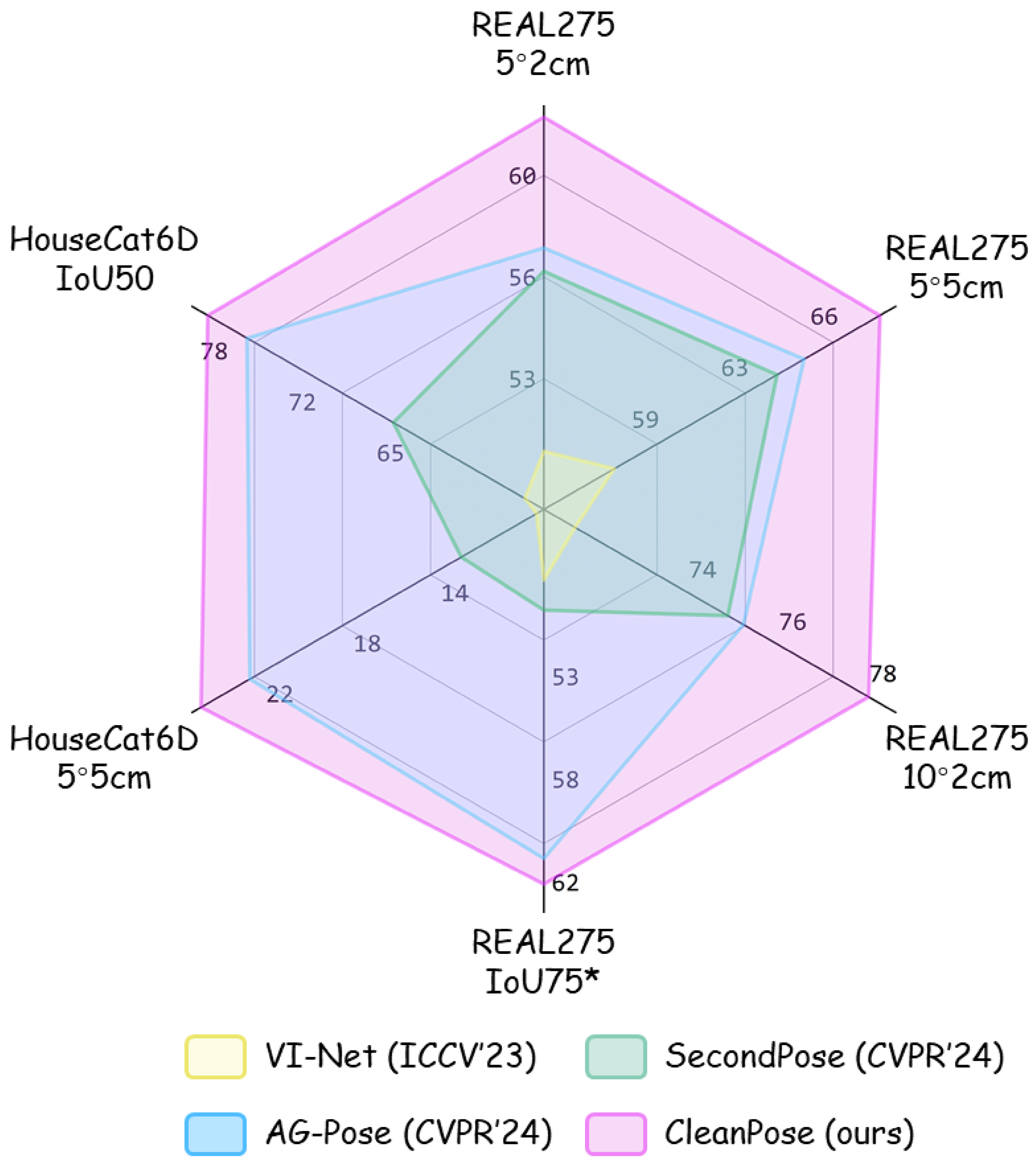
$$= \sum_m P(m \mid \mathcal{X}) \sum_x \sum_u P(\mathcal{Y} \mid m, u)P(u \mid \mathcal{X})P(\mathcal{X})$$

$$= \sum_{x'} P(x') \sum_m P(\mathcal{Y} \mid m, x')P(m \mid \mathcal{X})$$

$$= \mathbb{E}_{x'}\mathbb{E}_{m|x}\left[P(\mathcal{Y} \mid x', m)\right] = \mathbb{E}_{x'}[x'] + \mathbb{E}_{m|x}[m]$$



☐ Quantitative Results

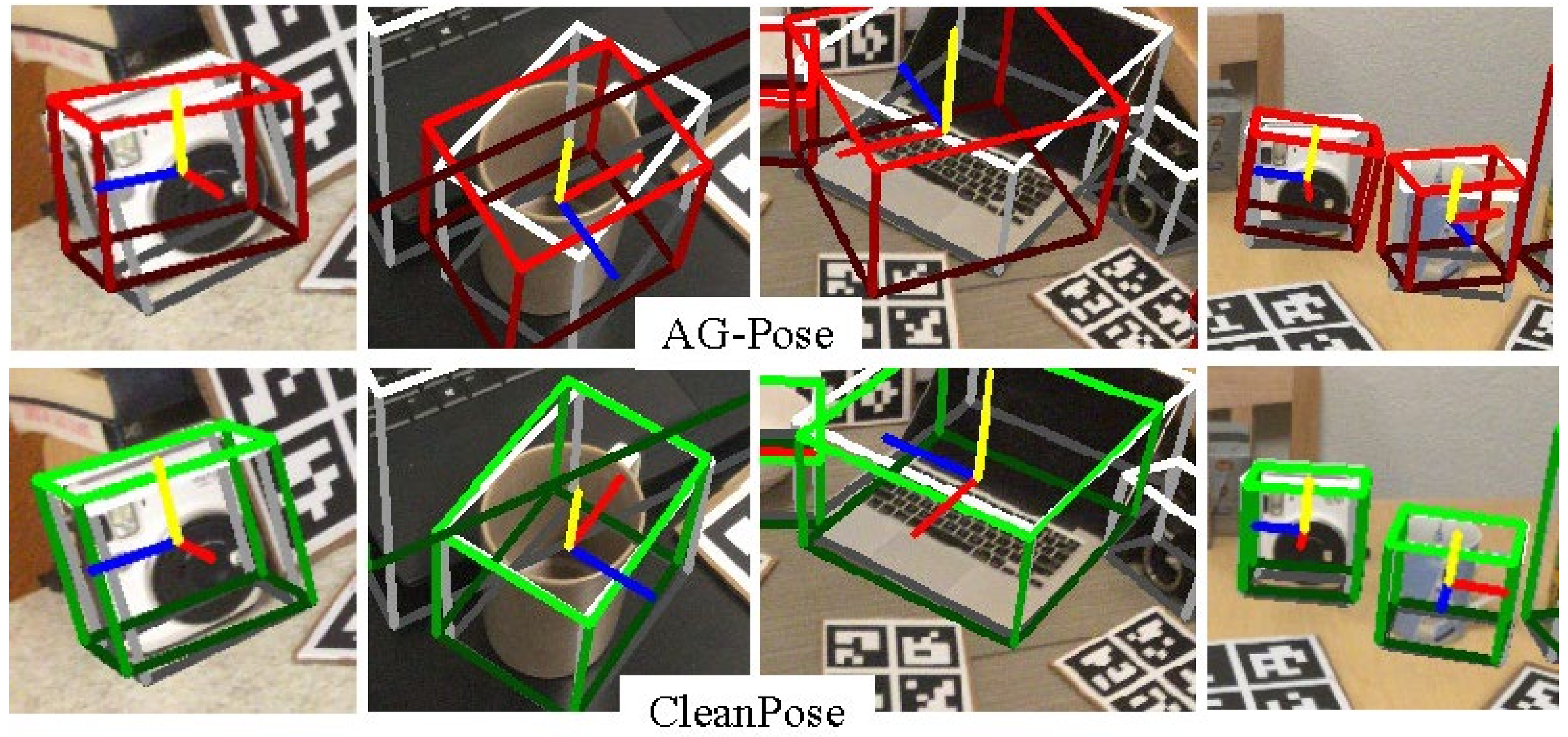


Methods	Venue/Source	Shape Prior	IoU^{*}_{75}↑	$5°2cm$↑	$5°5cm$↑	$10°2cm$↑	$10°5cm$↑
DPDN[19]	ECCV'22	✓	54.0	46.0	50.7	70.4	78.4
MH6D[26]	TNNLS'24	✓	54.2	53.0	61.1	72.0	82.0
GCE-Pose[17]	CVPR'25	✓	-	57.0	65.1	75.6	**86.3**
HS-Pose[56]	CVPR'23	✗	39.1	45.3	54.9	68.6	83.6
VI-Net[20]	ICCV'23	✗	48.3	50.0	57.6	70.8	82.1
CLIPose[23]	TCSVT'24	✗	-	48.5	58.2	70.3	85.1
GenPose[52]	NeurIPS'23	✗	-	52.1	60.9	72.4	84.0
SecondPose[3]	CVPR'24	✗	49.7	56.2	63.6	74.7	86.0
AG-Pose[22]	CVPR'24	✗	61.3	57.0	64.6	75.1	84.7
CleanPose (ours)		✗	**62.7**	**61.7**	**67.6**	**78.3**	**86.3**

Methods	IoU^{*}_{75}	$5°2cm$	$5°5cm$	$10°2cm$	$10°5cm$
HS-Pose[56]	-	73.3	80.5	80.4	89.4
CLIPose[23]	-	74.8	82.2	82.0	91.2
GeoReF[57]	79.2	77.9	84.0	83.8	90.5
AG-Pose[22]	**81.2**	79.5	83.7	87.1	92.6
CleanPose (ours)	80.7	**80.3**	**84.2**	**87.7**	**92.7**

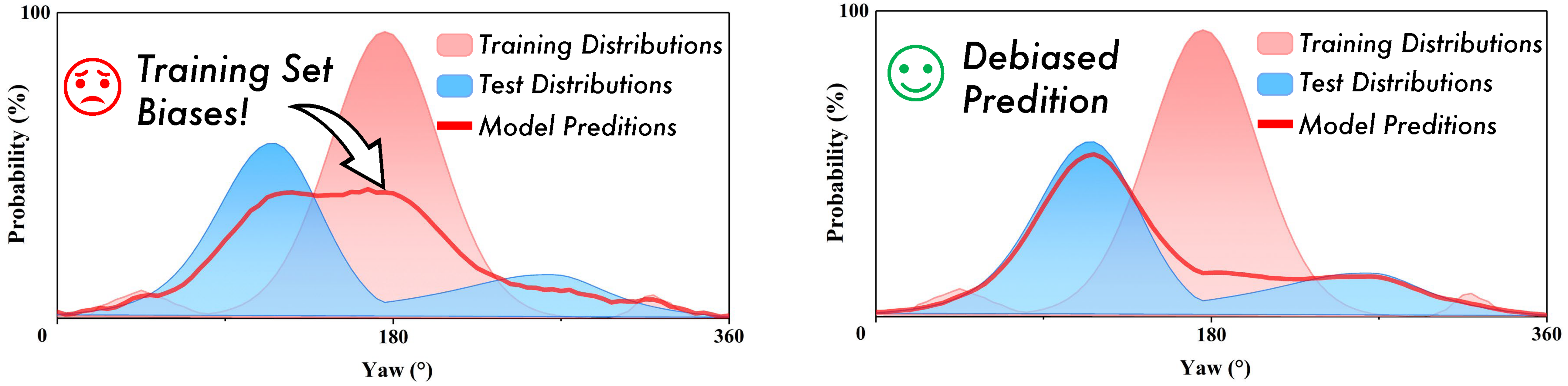
Methods	IoU_{25}	IoU_{50}	$5°2cm$	$5°5cm$	$10°2cm$	$10°5cm$
FS-Net[2]	74.9	48.0	3.3	4.2	17.1	21.6
GPV-Pose[4]	74.9	50.7	3.5	4.6	17.8	22.7
VI-Net[20]	80.7	56.4	8.4	10.3	20.5	29.1
SecondPose[3]	83.7	66.1	11.0	13.4	25.3	35.7
AG-Pose[22]	88.1	76.9	21.3	22.1	51.3	54.3
CleanPose (ours)	**89.2**	**79.8**	**22.4**	**24.1**	**51.6**	**56.5**

Qualitative Comparison



- The qualitative results of AG-Pose and proposed CleanPose shows that our method achieves significantly higher precision.

Illustration of Debiasing



- The predictions of baseline model are clearly biased toward the training set distributions, while the debiased model primarily unaffected.

➢ Conclusion and limitation

- We present CleanPose, **the first solution** that addresses the dataset biases in category-level pose estimation from the perspective of causal learning.

- We formulate the modeling of crucial causal variables and develop **a causal inference framework** in Category-level pose estimation task.

- We devise a residual knowledge distillation network to transfer unbiased semantics knowledge from 3D foundation model, providing comprehensive causal guidance to achieve unbiased estimation.

- Limitation: The investigate on the application of causal learning methods remains incomplete.

THANK YOU

Paper

Github Code

Acknowledgments: This work is supported by the National Natural Science Foundation of China under Grants (62233013, 62173248, 62333017, 624B2105).