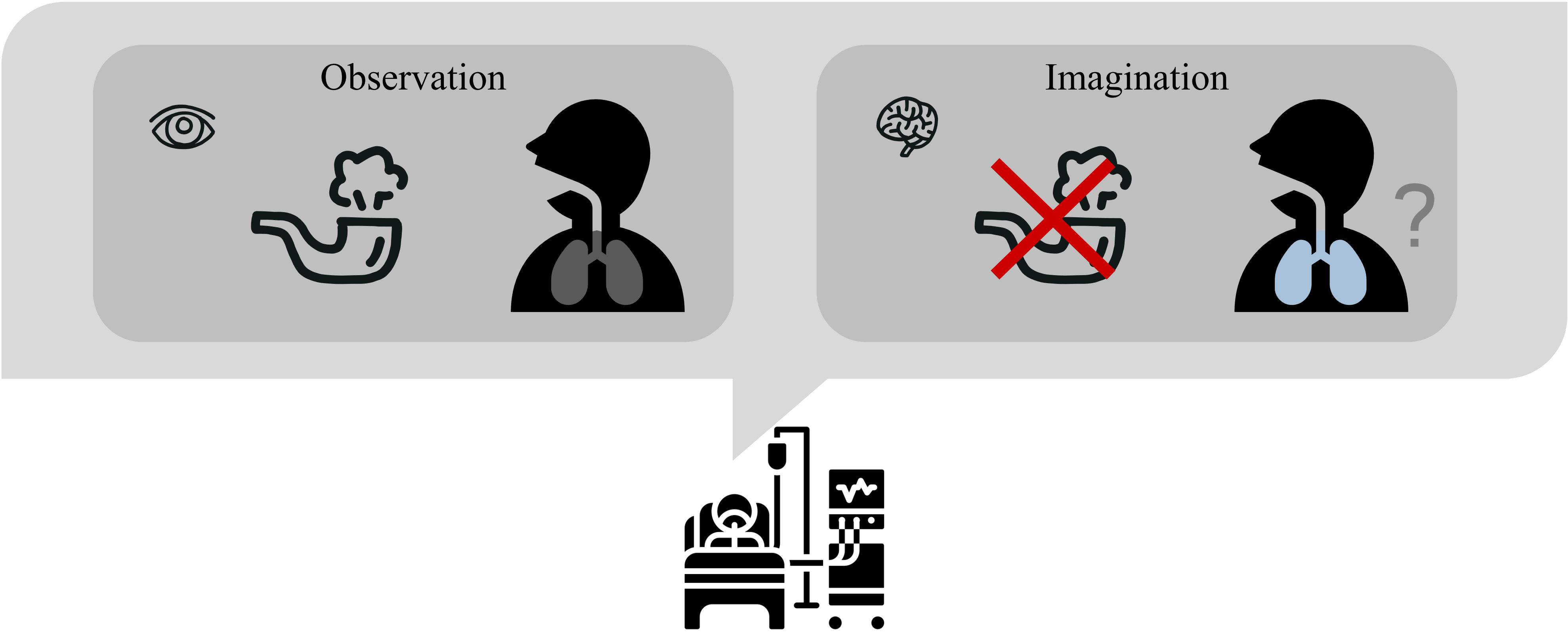
Exogenous Isomorphism for Counterfactual Identifiability

Yikang Chen[1,†], Dehui Du[1,*]

[1] Shanghai Key Laboratory of Trustworthy Computing, East China Normal University
* Corresponding author (dhdu@sei.ecnu.edu.cn)
† Presenter (yikangc620@gmail.com)

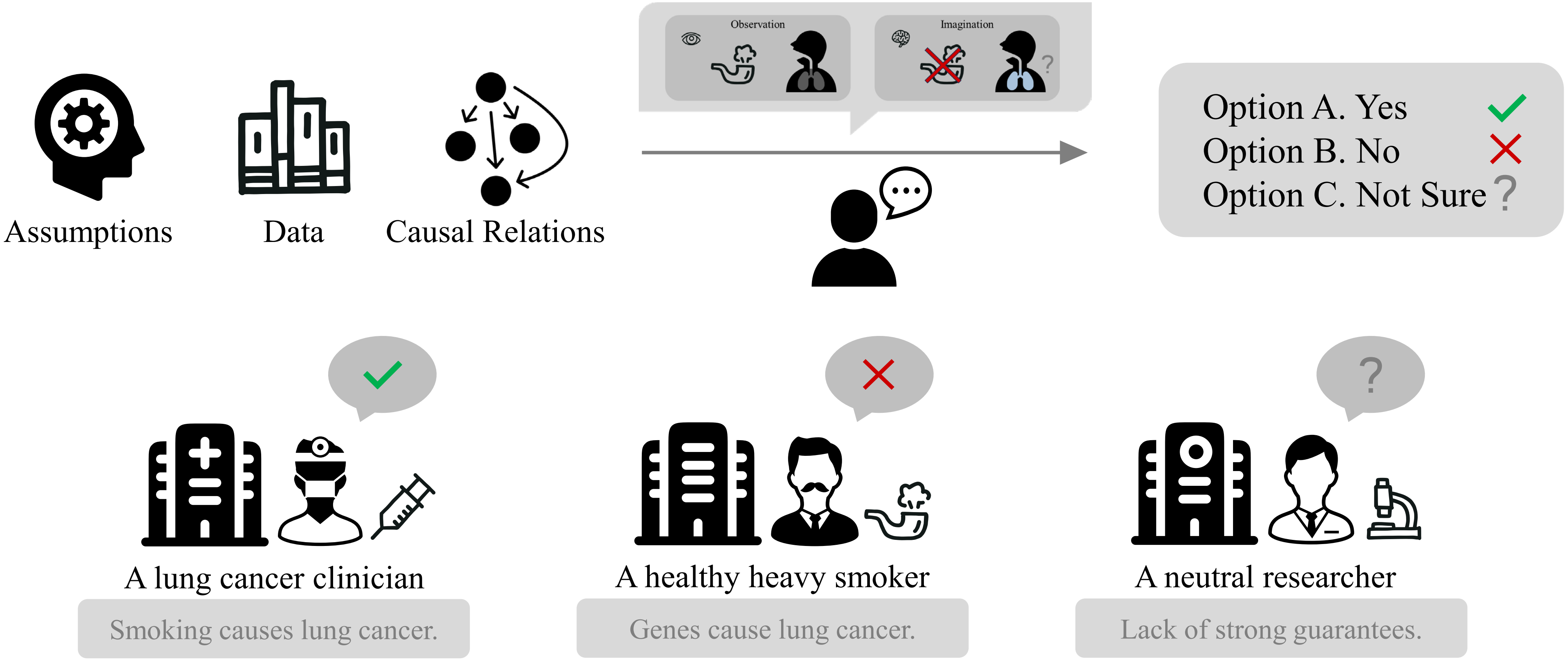
Counterfactual Identifiability

- Counterfactual is about answering “what-if” questions



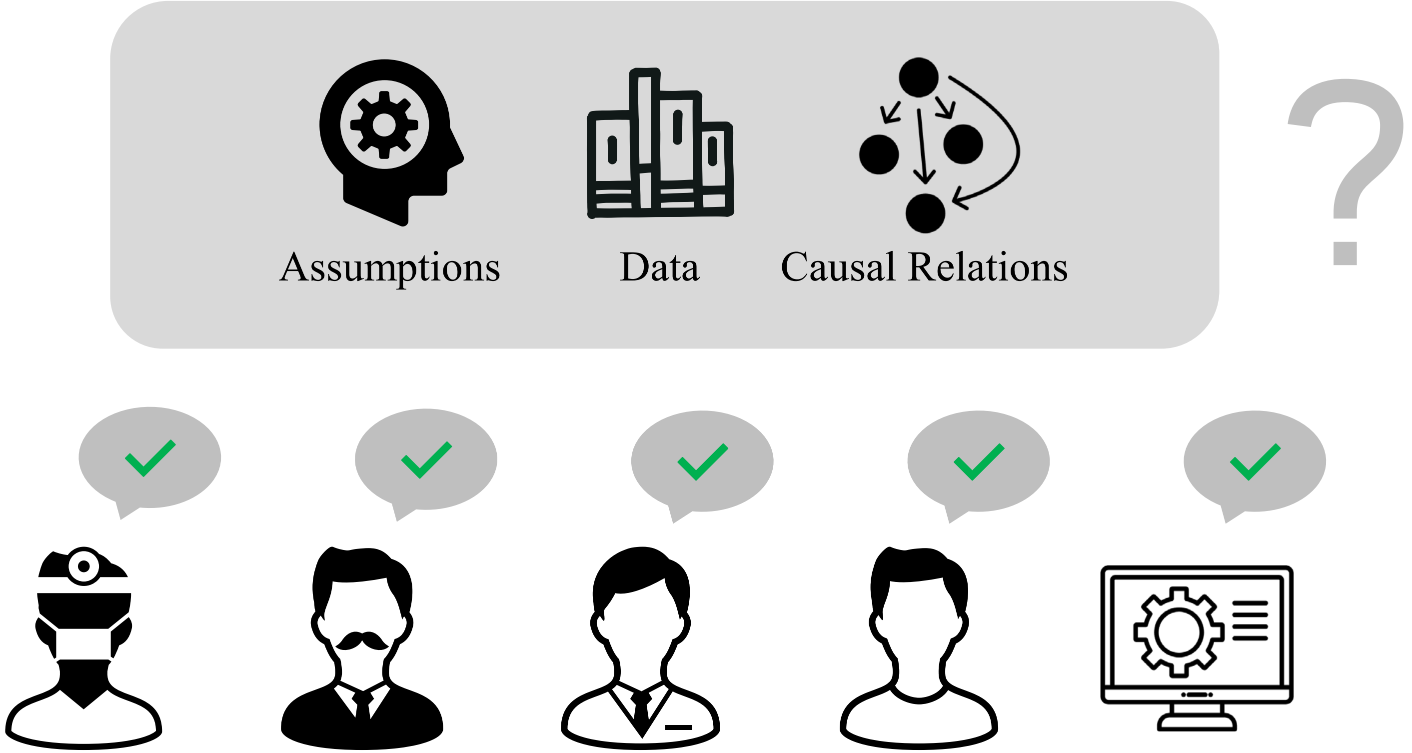
Counterfactual Identifiability

- Inconsistent answers may be produced.

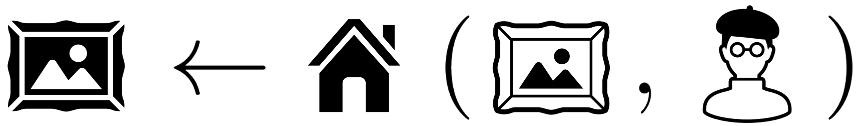


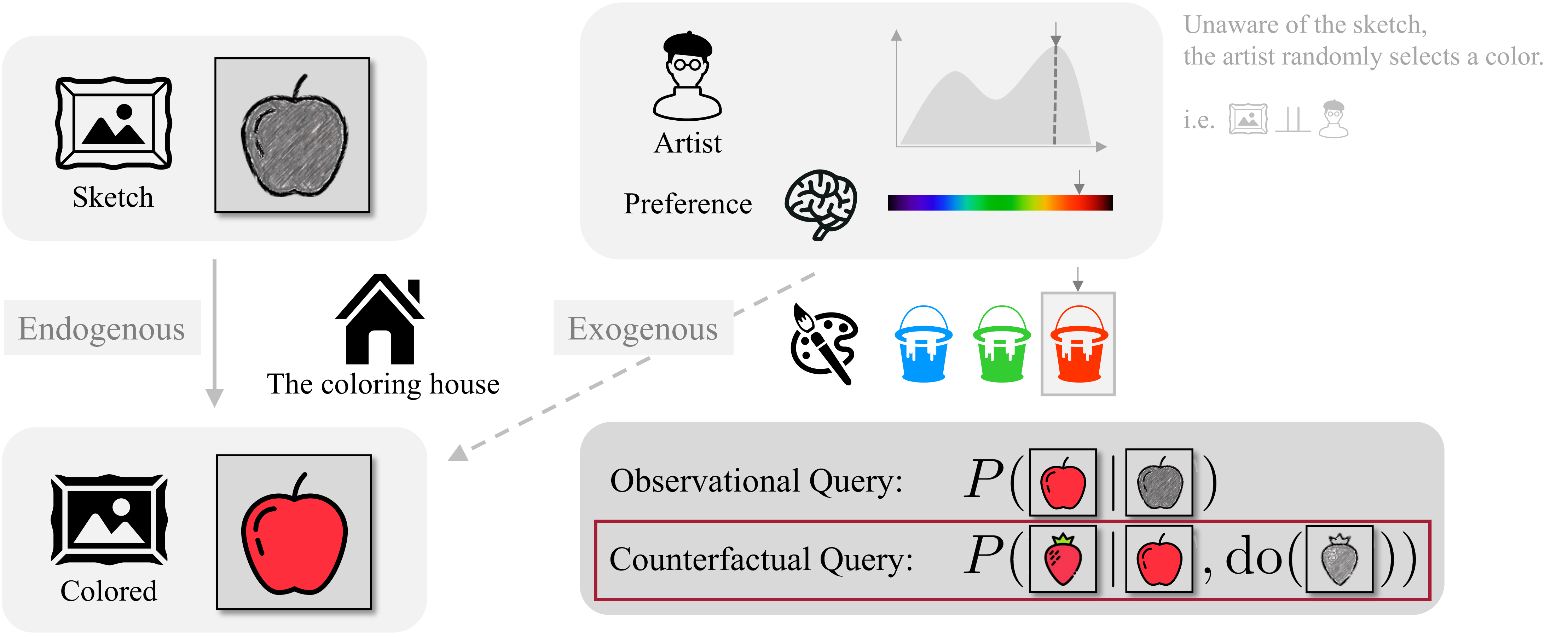
Counterfactual Identifiability

- Counterfactual Identifiability
  - How strong must the guarantees be for consistent answers?

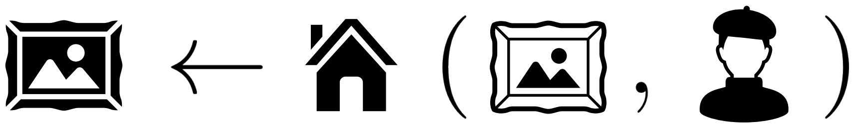


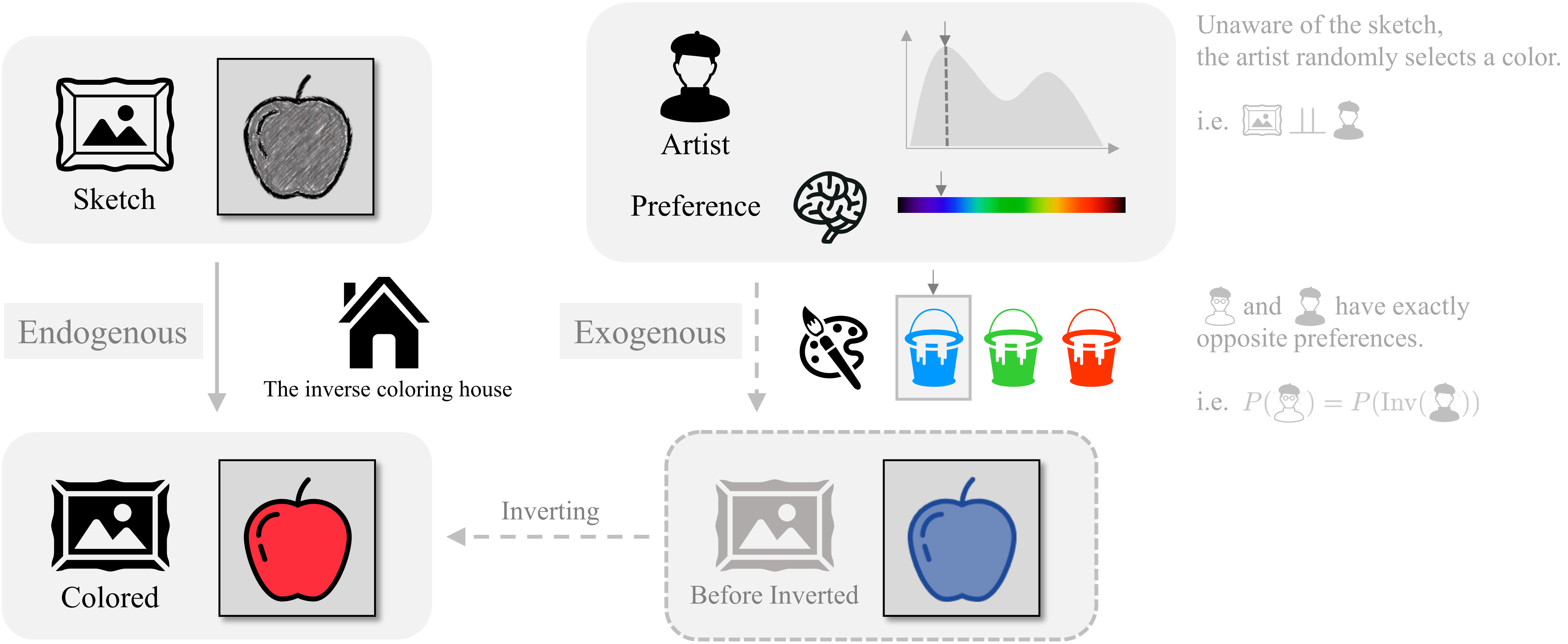
A Thought Experiment

- The coloring house 



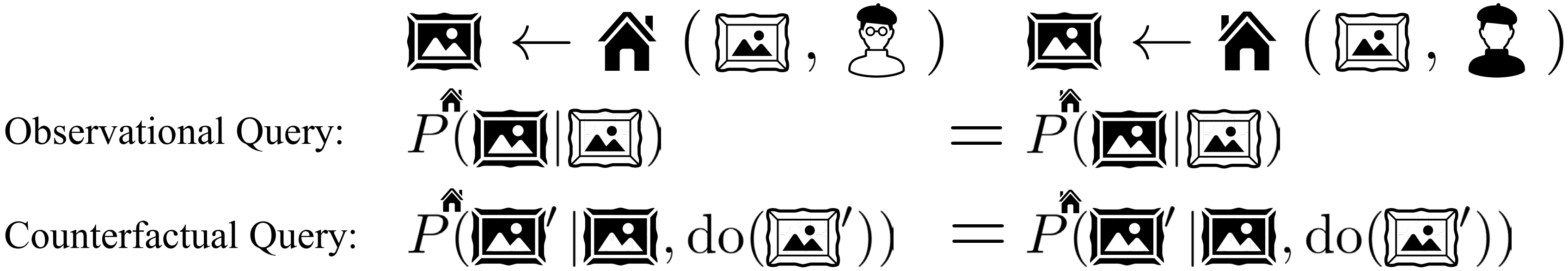
A Thought Experiment

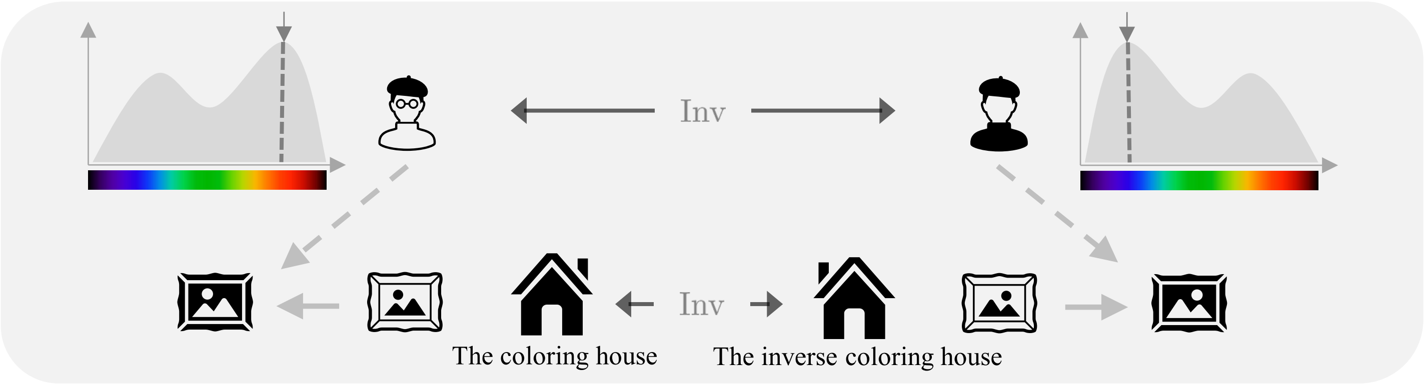
- The inverse coloring house 

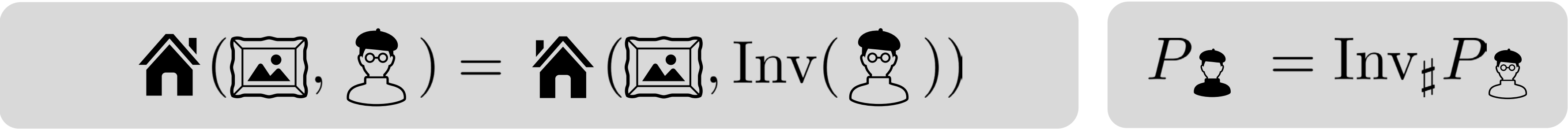


A Thought Experiment

- The consistency between two coloring houses

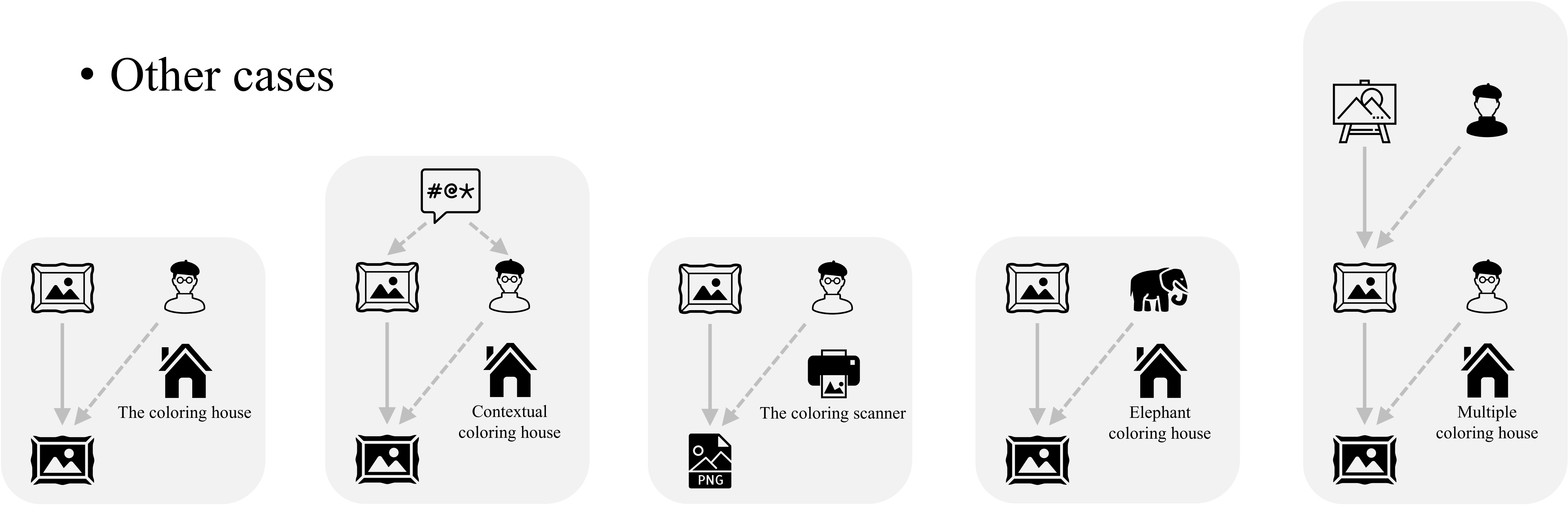






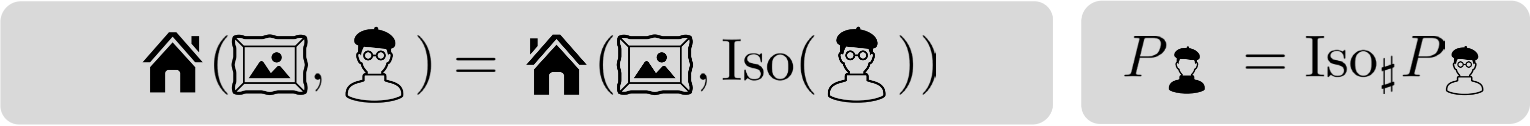
A Thought Experiment

- Other cases



Markovian | Not Markovian | Different Endogenous Domain | Different Exogenous Domain | Multiple Variables

Identifiability from Isomorphisms

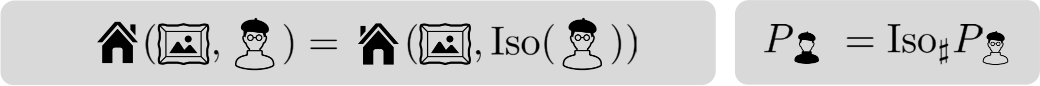


Isomorphism	Mechanism Isomorphism	Distributional Isomorphism	Mechanistic Requirement	Structural Requirement
Counterfactual Equivalence (Peters et al., 2017)	$f^{(1)}(\mathbf{v}_{\mathrm{pa}^{(1)}}, u^{(1)}) = f^{(2)}(\mathbf{v}_{\mathrm{pa}^{(2)}}, u^{(1)})$	$P^{(2)}_{\mathbf{U}} = P^{(1)}_{\mathbf{U}}$	-	order
BGM Equivalence (Nasr-Esfahany et al. 2023)	$f^{(1)}(\mathbf{v}_{\mathrm{pa}^{(1)}}, u^{(1)}) = f^{(2)}(\mathbf{v}_{\mathrm{pa}^{(2)}}, g(u^{(1)}))$	-	bijection	graph
LCM Isomorphism (Brehmer et al. 2022)	$f^{(1)}(\mathbf{v}_{\mathrm{pa}^{(1)}}, u^{(1)}) = f^{(2)}(\mathbf{v}_{\mathrm{pa}^{(2)}}, \varphi(u^{(1)}))$	$P^{(2)}_{\mathbf{U}} = \boldsymbol{\varphi}_{\sharp} P^{(1)}_{\mathbf{U}}$	diffeomorphism	graph isomorphism
Domain Counterfactual Equivalence (Zhou et al. 2024)	$f^{(1)}(\mathbf{v}_{\mathrm{pa}^{(1)}}, h_1(u^{(1)})) = f^{(2)}(\mathbf{v}_{\mathrm{pa}^{(2)}}, h_2(u^{(2)}))$	$P^{(k)}_{\mathbf{U}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$	bijection	domain label
Exogenous Isomorphism (Ours)	$f^{(1)}(\mathbf{v}_{\mathrm{pa}^{(1)}}, u^{(1)}) = f^{(2)}(\mathbf{v}_{\mathrm{pa}^{(2)}}, h(u^{(1)}))$	$P^{(2)}_{\mathbf{U}} = \boldsymbol{h}_{\sharp} P^{(1)}_{\mathbf{U}}$	-	order

[1] Peters, J., Janzing, D., & Scholkopf, B. (2017). *Elements of causal inference: Foundations and Learning Algorithms*. MIT Press.
[2] Nasr-Esfahany, A., Alizadeh, M., & Shah, D. (2023). Counterfactual Identifiability of Bijective Causal Models. In Proceedings of the 40th International Conference on Machine Learning (pp. 25733–25754). PMLR.
[3] Brehmer, J., Haan, P., Lippe, P., & Cohen, T. (2022). Weakly supervised causal representation learning. In Advances in Neural Information Processing Systems (pp. 38319–38331). Curran Associates, Inc..
[4] Zhou, Z., Bai, R., Kulinski, S., Kocaoglu, M., & Inouye, D. (2024). Towards Characterizing Domain Counterfactuals for Invertible Latent Causal Models. In The Twelfth International Conference on Learning Representations.

Identifiability from Isomorphisms

- Exogenous isomorphism



Recursive SCMs $\mathcal{M}^{(1)}$ and $\mathcal{M}^{(2)}$ are said to be **exogenously isomorphic**, denoted $\mathcal{M}^{(1)} \sim_{\mathrm{EI}} \mathcal{M}^{(2)}$, if there exists a shared causal ordering $\leq$ and function $\mathbf{h} : \Omega_{\mathbf{U}}^{(1)} \to \Omega_{\mathbf{U}}^{(2)}$ satisfying:

- Component-wise bijection $\quad \mathbf{h} = (h_i)_{i \in \mathcal{I}}$, where each $h_i : \Omega_{U_i}^{(1)} \to \Omega_{U_i}^{(2)}$ is a bijection
- Exogenous distribution isomorphism $\quad P_{\mathbf{U}}^{(2)} = \boldsymbol{h}_\sharp P_{\mathbf{U}}^{(1)}$
- Causal mechanism isomorphism $\quad f^{(1)}(\mathbf{v}_{\mathrm{pa}^{(1)}}, u^{(1)}) = f^{(2)}(\mathbf{v}_{\mathrm{pa}^{(2)}}, h(u^{(1)}))$

- Exogenous isomorphism implies counterfactual consistency

$$\mathcal{M}^{(1)} \sim_{\mathrm{EI}} \mathcal{M}^{(2)} \implies \mathcal{M}^{(1)} \sim_{\mathcal{L}_3} \mathcal{M}^{(2)}.$$

$\sim_{L_3}$ is an equivalence relation over Structural Causal Models (SCMs), indicating that the models yield the same answers to any counterfactual statement.

- This includes counterfactual outcomes, counterfactual effects, joint counterfactuals, and nested counterfactuals.
- From the perspective of the Pearl Causal Hierarchy (PCH), $\sim_{\mathcal{L}_3}$ implies that the complete counterfactual layer is identical, meaning the models are indistinguishable across the entire hierarchy.

To Achieve Exogenous Isomorphism

- EI-identifiability problem

$$
\begin{aligned}
\boldsymbol{\mathcal{A}} \models \quad & \mathcal{M}^{(1)} \sim_{\square} \mathcal{M}^{(2)} \\
\implies & \mathcal{M}^{(1)} \sim_{\mathrm{EI}} \mathcal{M}^{(2)} \\
\implies & \mathcal{M}^{(1)} \sim_{\mathcal{L}_3} \mathcal{M}^{(2)}
\end{aligned}
$$

Under what assumptions $\boldsymbol{\mathcal{A}}$ does the model equivalence class $\sim_{\square}$ imply $\sim_{\mathrm{EI}}$?

Bijective SCMs

Assumption (a). The causal mechanisms are **bijective** with respect to the exogenous variables.
Assumption (b). The structural causal model is **Markovian**.
Assumption (c). The **causal order** is given.
Assumption (d). The **observational distribution** is available.
Assumption (e). The model induces identical **counterfactual transport** (e.g., Knothe–Rosenblatt transport).

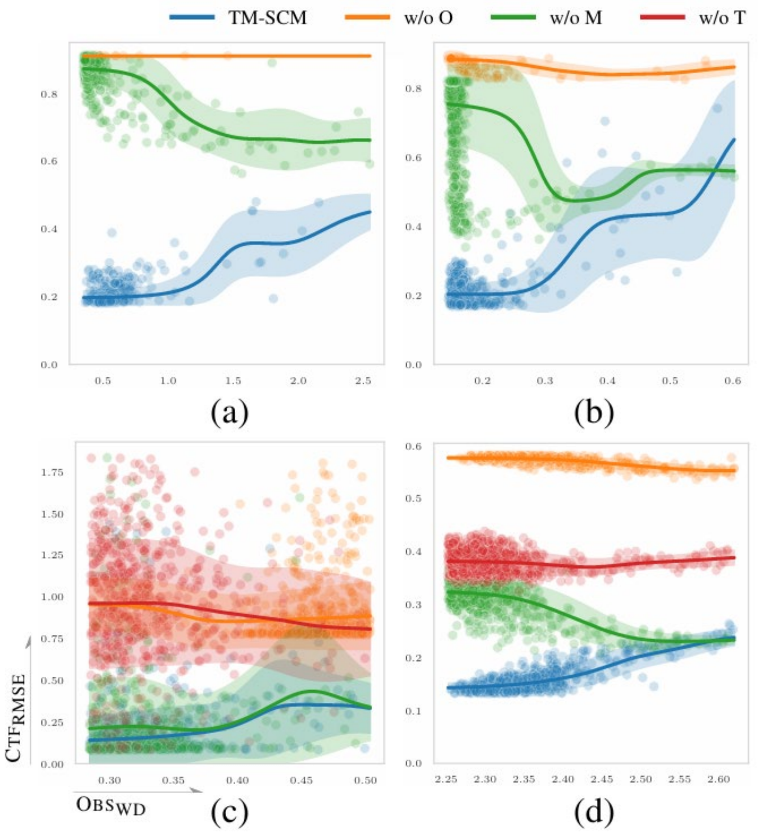
Triangular Monotonic SCMs

Assumption (a). The causal mechanisms are **monotonic** with respect to the exogenous variables.
Assumption (b). The structural causal model is **Markovian**.
Assumption (c). The **causal order** is given.
Assumption (d). The **observational distribution** is available.

- The model equivalence classes that satisfy the above assumptions entails EI-identifiability.
- As a result, counterfactual identifiability are ensured within these classes.

Ablation Study

- Ablation experiments on synthetic datasets with neural network-based Triangular Monotonic SCMs provide empirical support for the theoretical results.



METHOD		ER-DIAG-50	ER-TRIL-50
DNME	-	$\mathbf{0.53}_{\pm\mathbf{0.05}}$	$\mathbf{0.51}_{\pm\mathbf{0.12}}$
	w/o O	$0.78_{\pm0.05}$	$0.89_{\pm0.10}$
	w/o M	$0.62_{\pm0.04}$	$0.58_{\pm0.10}$
TNME	-	$\mathbf{0.47}_{\pm\mathbf{0.05}}$	$\mathbf{0.55}_{\pm\mathbf{0.12}}$
	w/o O	$11.24_{\pm20.98}$	$6.41_{\pm9.84}$
	w/o M	$0.62_{\pm0.04}$	$0.73_{\pm0.21}$
CMSM	-	$\mathbf{0.37}_{\pm\mathbf{0.05}}$	$\mathbf{0.42}_{\pm\mathbf{0.12}}$
	w/o O	$2.64_{\pm3.72}$	$2.12_{\pm2.49}$
	w/o M	$1.69_{\pm2.60}$	$0.75_{\pm0.49}$
	w/o T	$0.64_{\pm0.05}$	$1.25_{\pm1.29}$
TVSM	-	$\mathbf{0.46}_{\pm\mathbf{0.05}}$	$\mathbf{0.50}_{\pm\mathbf{0.12}}$
	w/o O	$0.79_{\pm0.04}$	$0.88_{\pm0.10}$
	w/o M	$0.53_{\pm0.05}$	$0.53_{\pm0.11}$
	w/o T	$0.67_{\pm0.05}$	$0.78_{\pm0.12}$

Summary

- We introduced exogenous isomorphism ensuring consistency across all counterfactual queries.
- We identified and proved that two specific classes of models satisfy EI-identifiability.
- Empirical results on neural SCMs and synthetic datasets support our theoretical claims.
- This work provides a principled foundation for learning counterfactually reliable models.

THANKS

Full paper is available at:
https://arxiv.org/abs/2505.02212

Code is available at:
https://github.com/cyisk/tmscm