We thank the reviewers for their time and valuable feedback. We believe that a few misreadings of our results may have led to a somewhat negative evaluation, which we ask for reconsideration given the clarifications provided below.

Connection with previous works and contributions. Under the structural causal model framework (Pearl, 2000), there has been extensive work on the problem of causal effect identification (as cited in line 30), which determines whether the causal functional could be obtained from observed distribution given a causal graph (i.e., *identifiable*) and derives such functional whenever identifiable. One outstanding challenge to applying these identification results in practical settings is that there has been no sample and computationally efficient estimators working for *any* identifiable causal functional. This paper addresses this challenge by filling the gap between causal "*identification*" and "*estimation*". We develop a learning framework that could work for *any* identifiable causal functional beyond the ignorability assumption.

Clarity. The paper aims to fill the gap from causal effect identification to estimation and assumes a basic background in identification theory. The discussion regarding Eq. (1) in line 85-94 and 140-151 is to show that it's possible to *manually* convert a functional output by a standard identification algorithm, but not friendly for the WERM framework, into a weighted form using known identification techniques. Non-familiarity with the identification techniques will not impact the rest of the paper, as this work develops a systematic algorithm to achieve the task (ref. line 181-184).

Reply to Reviewer 1. *"if other causal functionals exist, such as regression adjustment and IPTW, what is the advantage of the WERM-ID?"* Good question. The only setting where regression-based and weighting based estimators both exist is when the ignorability assumption holds (e.g., Fig. 1(a)). In this case, the proposed WERM estimator reduces to the standard re-weighting based estimators, which one can estimate using any ML methods (cited in line 58). *"...only compared with plug-in estimands. What about other weighting methods?"* As noted in lines 39 and 319, the plug-in estimator is the only viable estimator known to date for arbitrary identifiable functionals. Other weighting methods are not applicable as mentioned in lines (39-40, 320-321). *"This claim excludes any counterexample; is it too strong?'* This claim is a major contribution of this paper. We show *any* identifiable causal functionals can be converted in the weighted distribution format, and estimated using the WERM framework (Thm. 1,2). *"What does the dash curve with double arrow in Figure 1 mean?"* As noted in line 112 and following the convention (Pearl, 2000), the dashed-bidirected arrows between (X, Y) encode unobserved causes between (X, Y); i.e., $X \leftarrow U \rightarrow Y$, where U is unobserved. *"Why is Eq. (1) true?"* As stated in line 79, one could derive Eq. (1) by running a standard identification algorithm (e.g., [48] or [45]). *"There are also many skipped steps in the derivation above Lemma 1 for $P(y|do(x)) = P(x, y|do(r))/P(x|do(r)) = P(y|do(r), x) = P^W(y|x, r)$"* The first equality is due to Eq. (1), and explanation in line 140-146; the 2nd is definition of conditional probability; the last is from the equation in line 148. *"The connection with previous weighting methods..."* Existing weighting estimators were developed for settings where the ignorability assumption holds. Our work proposes a novel method working for *any identifiable* causal functionals.

Reply to Reviewer 2. *"doesn't explain that the causal graph needs to be provided beforehand"* We respectfully disagree since we explicitly state that the identification problem assumes a given causal graph in line 25-28, and in the subsequent example line 31-35. *"no empirical evidence of performance is given on large graphs... the algorithm can only handle 3-4 nodes"* We respectfully note that neither the theorems nor the empirical simulations limit the proposed algorithm to small graphs. The time complexity is *polynomial* in the size of the graph (Thm. 3) and empirically demonstrated in Fig. 3(d,e,f). In the experiments, the covariates W is set to be a vector of D binary variables (line 302), with $D = 15$ in Fig. 3(a,b,c) (stated in line 304, 307, 310).

Reply to Reviewer 3. *"I am a bit curious ..."* Great suggestion. A comparison example is given in Fig. 1. As expected, the performance of the PO framework based estimator is inferior to the proposed estimator ('WERM-ID-R'). This result implies adjusting covariates without taking into account the causal graph might yield inaccurate estimates of the causal effect; we'll add this to the paper, thanks.

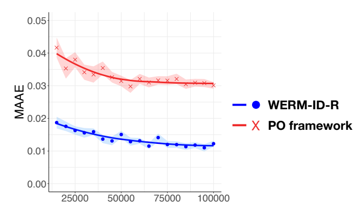

Figure 1: (For Reviewer 3) A MAAE plot comparing the proposed vs. PO-based estimator for Example 3 ($D = 15$). Shades are standard deviations.

Reply to Reviewer 7. *"Eq. (1): What happened to the variable r?"* That the r.h.s of (1) is independent of the value r is known as a Verma constraint on the observed distribution implied by the causal graph. This issue is discussed in Appendix A line 61-68. *"Lines 89-90: Why does $P(x, y, w|do(r))$ equal $(1/P(r|w))P(x, y, w, r)$?"* This can be derived by a standard identification algorithm (e.g., in [48] or [45]), or directly using Theorem 1 in [48]. *Algorithm 1, line 8.5: What is "T"?* $\mathbf{T}$ is an arbitrary set. Procedure $\texttt{wIdentify}(\mathbf{C}, \mathbf{T}, Q[\mathbf{T}], \mathbf{r}, \mathcal{W})$ outputs $Q[\mathbf{C}]$ for $\mathbf{C} \subseteq \mathbf{T}$ given input $\mathbf{T}$ and $Q[\mathbf{T}] = P^{\mathcal{W}}(\mathbf{t}|\mathbf{r})$. *"Learning low variance weights is not novel as (Swaminathan and Joachims, 2015) have already addressed..."* We appreciated the great work of SJ15 [47] and cited it in line (58, 200, 216). We adopted the idea in SJ15 of learning low variance weights. However, the results (which deals with ignorable cases / propensity score weighting) are not directly applicable, and properties such as learning guarantee in Thm. 2 needs to be re-derived in our context. *"there is no explicit discussion on how this work differs from the prior work"* The prior work on applying WERM to causal inference is limited to settings contingent on the ignorability/backdoor condition (line 56-61). This work developed a general learning framework that fully brings together the causal identification theory and WERM methods (as summarized in line 95-106).