CAUSAL ATTRIBUTION FROM VIDEO VIA COUNTERFACTUAL ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Causal attribution from video asks whether an outcome would have changed if one of the observed actions had been different. Such attribution is essential for embodied agents that must reason about past actions to act safely next time. It is difficult because several concurrent actions can explain the same outcome. Pretrained models may be misled by correlations under confounding, and counterfactual outcomes are rarely labelled. We formulate video causal attribution as counterfactual inference given an observed video. To evaluate methods on this problem, we build TongSim-CF, a household video benchmark with controlled confounding from concurrent actions and scene conditions. We identify the resulting attribution target from repeated videos of the same environment without counterfactual supervision. We then estimate a local average of the target around the observed scene. Our estimator combines outcome predictions with a weighted residual correction derived from the efficient influence function. In the hardest simulated and real settings, it reduces the error of the strongest baseline by 64% and 5%.

1 INTRODUCTION

Video understanding supports applications including video surveillance (Zhang et al., 2025), autonomous driving (Cheng et al., 2025), and embodied intelligence (Gupta et al., 2021). In these applications, a central task is to identify which of several simultaneous actions caused a change in the scene. This task is known as causal attribution. Consider a household robot that learns to open a door by watching people (Fig. 1). In its videos, a person often flicks a wall switch while walking toward the door, and then the door opens. In fact, the door opens automatically on approach, and the switch only controls the lights. If the robot attributes the opening to the switch, it will flick the switch, and the door will stay closed. Such misattribution is known as causal confusion (de Haan et al., 2019).

This example shows why causal attribution is difficult. Suppose people often switch on the lights as they enter. Flicking the switch and approaching the door then tend to occur together. Flicking the switch thus becomes strongly associated with the opening, although it has no effect on the door. Methods that learn from such associations may attribute the opening to the switch, and so does a frontier vision language model (VLM) in Fig. 1. To attribute the opening correctly, one must know whether the door would still open if the person approached without flicking the switch. The video does not show this outcome (Bareinboim et al., 2022). The main challenge is thus to separate the effect of one action from the actions that occur with it.

Existing approaches fall into three groups. Causal inference methods identify attribution from observed data, but under strong assumptions (Tian & Pearl, 2000; Lu et al., 2023; Fire & Zhu, 2016). For example, monotonicity requires that an action never prevents the outcome, whereas the effect of an action often depends on the scene. Counterfactually supervised methods learn from labels of the alternative outcome (Baradel et al., 2020; Patel et al., 2022). They cannot be used in scenes without such labels. Annotating counterfactual outcomes is also difficult and requires experts who know the mechanism of each scene. Foundation models such as VLMs (Comanici et al., 2025; Bai et al., 2025) instead answer from pretrained commonsense. This commonsense encodes correlations, such as switches operating devices, and confounding makes such correlations misleading. The models also tend to rely on such priors rather than on visual evidence (Guan et al., 2024; Chen et al., 2024b).

We address this challenge by learning the mechanism of an environment from the videos recorded in it. The physical mechanisms of an environment stay the same over time. A home camera or a robot working there can therefore collect many videos under the same mechanism. Since action choices vary across these videos, some of them show the outcomes that the video of interest lacks. For the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

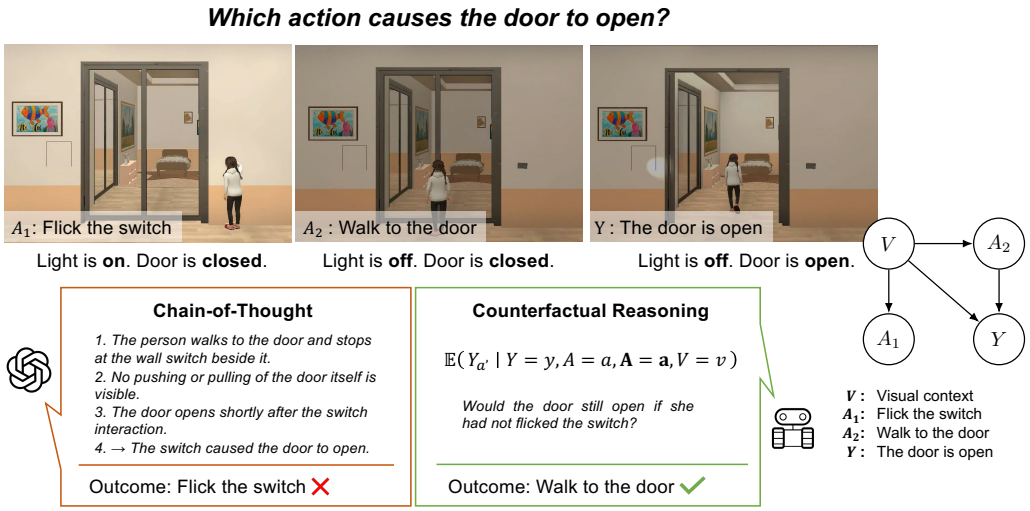

Figure 1: Causal attribution in a scene with an automatic door. **Top:** a person flicks a wall switch (A_1) and walks toward a door (A_2); the door opens (Y). The door opens automatically on approach; the switch only controls the lights. The right panel shows the causal graph over the visual context V, the two actions, and the outcome. **Bottom left:** a vision language model incorrectly attributes the opening to switch use. **Bottom right:** the door would still open without switch use because the approach triggers it. This answer follows from the scene mechanism.

door in Fig. 1, some videos show people approaching without flicking the switch, and the door still opens. Each video also records the initial scene and the other actions. They may be confounders between the target action and the outcome. Comparing videos with similar scenes and the same other actions holds them fixed. Combining such comparisons across videos therefore deconfounds the effect of the target action.

To address this problem, we formulate video causal attribution as *counterfactual inference given an observed video*. To evaluate methods on this problem, we build *TongSim-CF*, a household video benchmark. Its controlled confounding and counterfactual ground truth test counterfactual attribution rather than correlation. Combining the above properties of video with standard causal assumptions, we identify the target *without counterfactual supervision*. To estimate the target efficiently, we derive the *efficiency bound* of a localized version of the target and an estimator that attains it. To our knowledge, this is the first efficiency result for causal attribution in video.

Our contributions are as follows:

- We formulate causal attribution in video as counterfactual inference given an observed video, conditioned on the scene and the other actions.
- We build TongSim-CF, a household video benchmark with controlled confounding and counterfactual ground truth, to test action effects rather than correlations.
- We identify the target from videos of the same environment without counterfactual supervision, and derive an estimator that attains the efficiency bound of a localized target.
- In the hardest simulated and real settings, our estimator reduces the error of the strongest baseline in each by 64% and 5%.

2 Preliminaries

We are interested in attributing the outcome of a video to one of its actions. We denote each video by a tuple $(Y, A, \bar{\mathbf{A}}, V)$. $Y \in \{0, 1\}$ is the outcome of interest. $A \in \mathcal{A}$ is the target action whose causal role we want to assess. $\bar{\mathbf{A}} = (\bar{A}_1, \ldots, \bar{A}_d) \in \prod_{j=1}^{d} \mathcal{A}_j$ collects the other actions in the same window. $V \in \mathbb{R}^{d_V}$ is the visual observation of the scene before the action, taken from the first frame. A complete list of notation is given in Sec. B. For an observed video, we define the attribution target as

$$\psi(a'; y, a, \bar{\mathbf{a}}, v) \;=\; \mathbb{E}\big[\, Y_{a'} \,\big|\, Y = y,\; A = a,\; \bar{\mathbf{A}} = \bar{\mathbf{a}},\; V = v \,\big]\,, \tag{1}$$

where Y_a denotes the potential outcome under action $A = a$ (Rubin, 1974). That is, given what the video shows, ψ is the probability that the outcome would have occurred had the target action been a'.

Our target is related to the classical probabilities of necessity (PN), sufficiency (PS), and necessity and sufficiency (PNS) (Pearl, 1999; Tian & Pearl, 2000). For a single binary action, $\mathrm{PN} = \Pr(Y_0 = 0 \mid Y = 1, A = 1)$ is the probability that removing an observed action would prevent its outcome. $\mathrm{PS} = \Pr(Y_1 = 1 \mid Y = 0, A = 0)$ is the probability that adding an absent action would produce the missing outcome. The joint probability $\mathrm{PNS} = \Pr(Y_1 = 1, Y_0 = 0)$ measures how often the outcome would occur with the action but not without it. In contrast, Eq. (1) also conditions on $\bar{\mathbf{A}} = \bar{\mathbf{a}}$ and $V = v$, since the contribution of an action depends on the scene and other actions.

Like our target, the posterior causal effects of Lu et al. (2023) condition a counterfactual on observed evidence. Li et al. (2024) extend them to multiple outcomes. However, their evidence consists of binary causes and outcomes, so it can be matched exactly across records. In video, the evidence also includes the frame before the action. This frame is continuous and high dimensional, and no two videos share exactly the same frame. To condition on such evidence, we extract the relevant scene information and estimate the target around the observed scene (Sec. 4).

Estimating Eq. (1) is difficult because the video of interest shows only the outcome of the action taken. As in Sec. 1, the outcome under a' must therefore be learned from other videos of the same environment. We use n i.i.d. videos $\{(y_i, a_i, \bar{\mathbf{a}}_i, v_i)\}_{i=1}^{n}$ and make the following assumptions.

Assumption 1 (Consistency)**.** If $A = a$, then $Y = Y_a$.

Assumption 2 (Positivity)**.** $\Pr(A = a \mid \bar{\mathbf{A}} = \bar{\mathbf{a}}, V = v) > 0$ for every $a \in \mathcal{A}$ and every $(\bar{\mathbf{a}}, v)$ in the support of $(\bar{\mathbf{A}}, V)$.

Assumption 3 (No unmeasured confounding between the potential outcomes and A)**.** $(Y_a)_{a \in \mathcal{A}} \perp\!\!\!\perp A \mid \bar{\mathbf{A}}, V$.

Assumption 4 (Conditional independence of potential outcomes)**.** For every $a \neq a'$ in $\mathcal{A}$, $Y_a \perp\!\!\!\perp Y_{a'} \mid \bar{\mathbf{A}}, V$.

Assumption 5 (Sufficient representation)**.** There exists a scene representation $Z = g(V)$ such that, for every $a \in \mathcal{A}$, $(Y_a, A, \bar{\mathbf{A}}) \perp\!\!\!\perp V \mid Z$.

Discussion of these assumptions in video understanding. Ass. 1–3 are standard in causal inference (Rubin, 1974; Hernán & Robins, 2020; Imbens & Rubin, 2015). Ass. 3 requires action selection to be independent of the potential outcomes given the scene and other actions. It is violated when a physical condition affects both the choice of action and its outcome. In video, however, such conditions are often visible before the action. For example, a power indicator shows whether a lamp has power. The power state affects both how often the switch is used and whether pressing it lights the lamp. Since the indicator is visible in the first frame, conditioning on this frame compares switch use under the same power state. Such conditioning on the observed scene is common in physical video prediction (Lerer et al., 2016; Finn et al., 2016).

Ass. 4 means that, given $(\bar{\mathbf{A}}, V)$, one potential outcome provides no further information about another. Following Shen et al. (2013) and Tian & Wu (2025), we use it in place of monotonicity (Tian & Pearl, 2000; Pearl, 1999). Monotonicity requires that a binary action never prevents the outcome. It thus excludes actions whose effect reverses with the scene. For example, pressing a switch turns a lit lamp off but an unlit lamp on. Ass. 4 still holds in this example, because the initial state fixes both potential outcomes. More generally, it holds whenever the scene and other actions determine the outcome under each action. Sec. 5.4 further examines how the error grows when it is violated.

Under Ass. 5, Z retains all information in V relevant to the actions and potential outcomes. Z must therefore encode the physical conditions that determine action outcomes. Examples include object positions and support relations, which govern contact and stability. Recent studies show that learned representations capture such conditions. For example, object-centric VAEs encode object positions (Greff et al., 2019), and DINOv2 features encode support relations (Zhan et al., 2024). Moreover, video models commonly assume that representations of object states and interactions suffice to predict future or counterfactual outcomes (Baradel et al., 2020; Li et al., 2020). Sec. 4.3 describes how we construct Z.

Efficient estimation. Influence functions are a standard tool to construct efficient estimators (Bickel et al., 1993; Robins et al., 1994). Let P be the distribution of an observation O, and let $\Psi(P)$ be the quantity to estimate. The influence function $\phi(O; P)$ has mean zero and gives the pathwise derivative of Ψ. Specifically, for every smooth path $\{P_\varepsilon\}$ through P with score $s(O)$, it satisfies

the first equation below. Given a fitted distribution $\widehat{P}$ and observations $O_1, \ldots, O_n$, the one-step estimator adds the average of the estimated influence function to the plug-in estimate:

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\Psi(P_\varepsilon)\bigg|_{\varepsilon=0} = \mathbb{E}_P[\phi(O;P)\, s(O)]\,, \qquad \widehat{\Psi} = \Psi(\widehat{P}) + \frac{1}{n}\sum_{i=1}^{n}\phi(O_i;\widehat{P}). \tag{2}$$

Because ϕ is the first order derivative of Ψ, this average corrects the first order error of $\Psi(\widehat{P})$.

3 BENCHMARK

To test whether methods can attribute outcomes under confounding without counterfactual supervision, we introduce TongSim-CF, a household video benchmark. We first discuss two limitations of existing benchmarks and then describe how TongSim-CF improves on them.

3.1 LIMITATIONS OF EXISTING BENCHMARKS

Existing video benchmarks make it difficult to evaluate causal attribution without counterfactual supervision, for two reasons.

First, they lack overlapping videos. The video of interest never shows the outcome under the alternative action, so it must be learned from other videos of the same mechanism. In these videos, the target action should take different values under a comparable scene and the same other actions. CLEVRER (Yi et al., 2020) and CRIPP-VQA (Patel et al., 2022) pose counterfactual questions about individual videos, but they do not provide such videos. A method must then answer from commonsense or counterfactual labels, which makes it difficult to test attribution from videos.

Second, they lack controlled confounding. Confounding arises when other actions or scene conditions affect both whether the target action is taken and the outcome. The target action then occurs together with the outcome more often than its effect implies. Without confounding, this association equals the effect, so a benchmark cannot separate attribution from correlation. CLEVRER and CRIPP-VQA do not vary these common causes. MECD (Chen et al., 2024c) studies confounding among events but lacks overlapping videos. TongSim-CF improves on both points (Sec. A).

3.2 BENCHMARK CONSTRUCTION

Data generation. To control the mechanisms and compute counterfactual outcomes, we build TongSim-CF on TongSim (Sun et al., 2025; Peng et al., 2024), a household simulator with controllable objects, actions, and scene states. Each scenario specifies an interaction template with objects, candidate actions, and a binary outcome, such as a lamp lighting. To generate an episode, we initialise the arrangement and physical states of the objects, execute sampled actions, and record the outcome under the specified mechanism. The video and annotations of the episode give a sample $(Y, A, \bar{\mathbf{A}}, V)$ as defined in Sec. 2.

Scenarios and settings. To ensure that a method learns each mechanism from its videos rather than from appearance, we create variants of each scenario. The variants differ in how actions are selected or how the outcome responds to them. For example, one door variant opens only when a nearby button is pressed, whereas another opens on approach regardless of the button. The benchmark contains 17 scenarios, 31 variants, and 18,464 videos. Within each variant, we fix the mechanism, vary the initial scene, and sample whether each action is taken. The target action therefore varies under comparable scenes and the same other actions. The scenario templates and equations are given in Sec. H.1.

To address the lack of controlled confounding, we group the scenarios into three settings by the source of confounding. *Setting 1* has no confounding. For example, a person opens a drawer at random, and no other factor affects whether the drawer ends up open. In *Setting 2*, another action often occurs together with the target action and also affects the outcome. For example, a button is often pressed while a person approaches a door that opens on approach. In *Setting 3*, both scene conditions and other actions may confound the target action. For example, a switch is used more often when a lamp has power, and it lights the lamp only in that case.

Counterfactual ground truth. For each test video, the simulator restores the initial scene and the other actions and replaces the target action. It then computes the outcome probability from the equations of the variant. This probability serves as the counterfactual ground truth. Its construction is detailed in Sec. H.

4 METHODOLOGY

In this section, we develop an estimator of Eq. (1) that is efficient for a localized target at fixed bandwidth. We first smooth the target so that the correction in Eq. (2) applies. Sec. 4.1 then identifies the smoothed target. Sec. 4.2 constructs the estimator and proves its efficiency, and Sec. 4.3 adapts it to video.

The correction in Eq. (2) requires the target to be pathwise differentiable. The target in Eq. (1), however, conditions on an exact scene $V = v$ in a continuous space. Hardly any other video shares this scene. Such a pointwise target is generally not pathwise differentiable in a nonparametric model, so Eq. (2) does not apply directly (Kennedy, 2024). We therefore average over nearby scenes while retaining the observed action and outcome conditions (Abrevaya et al., 2015). For video, we approximate scene matching with discrete attributes and continuous visual features (Sec. 4.3). With a kernel $K_h(\cdot) = h^{-d_V} K(\cdot/h)$ and bandwidth $h > 0$, we define

$$\psi_h(a'; y, a, \bar{\mathbf{a}}, v) \;=\; \frac{\mathbb{E}\big[\,\mathbb{I}\{Y = y, A = a, \bar{\mathbf{A}} = \bar{\mathbf{a}}\}\, K_h(V - v)\, Y_{a'}\,\big]}{\mathbb{E}\big[\,\mathbb{I}\{Y = y, A = a, \bar{\mathbf{A}} = \bar{\mathbf{a}}\}\, K_h(V - v)\,\big]}. \tag{3}$$

Throughout, K is bounded and nonnegative with unit integral, and the denominator in Eq. (3) is finite and positive. For fixed h, this local target is pathwise differentiable under the regularity conditions in Sec. G. The following proposition quantifies how it approximates the pointwise target as $h \to 0$.

Proposition 1 (Local approximation)**.** *Under the second order kernel and smoothness conditions in Sec. G.4,*

$$\psi_h(a'; y, a, \bar{\mathbf{a}}, v) - \psi(a'; y, a, \bar{\mathbf{a}}, v) = O(h^2), \qquad h \to 0.$$

The proof is given in Sec. G.4.

4.1 IDENTIFICATION

Since the local target in Eq. (3) involves the unobserved outcome $Y_{a'}$, we express it through the distribution of videos. We first regress the outcome on the actions and scene representation. We then average the predictions under the alternative action over videos that share the observed evidence.

Proposition 2 (Outcome regression)**.** *Under Ass. 1–5, for every $a' \neq a$,*

$$\psi_h(a'; y, a, \bar{\mathbf{a}}, v) \;=\; \frac{\mathbb{E}\big[\,\mathbb{I}\{Y = y, A = a, \bar{\mathbf{A}} = \bar{\mathbf{a}}\}\, K_h(V - v)\, \mu(Z; a', \bar{\mathbf{a}})\,\big]}{\mathbb{E}\big[\,\mathbb{I}\{Y = y, A = a, \bar{\mathbf{A}} = \bar{\mathbf{a}}\}\, K_h(V - v)\,\big]}, \tag{4}$$

where $\mu(z; a', \bar{\mathbf{a}}) \coloneqq \mathbb{E}[Y \mid A = a', \bar{\mathbf{A}} = \bar{\mathbf{a}}, Z = z]$ is the outcome regression.

The proof is in Sec. D. Classical outcome regression averages predictions over the covariate distribution of the target population (Robins, 1986; Hernán & Robins, 2020). In Eq. (4), this distribution additionally conditions on the observed outcome $Y = y$. Under Ass. 4, conditioning on $Y = y$ changes only this averaging distribution, not the regression $\mu(Z; a', \bar{\mathbf{a}})$. The scene v sets a neighbourhood of scenes, and the outcome y weights the scenes within it (Sec. D.3).

However, estimating Eq. (4) relies on a fitted outcome regression and is sensitive to its misspecification. We therefore give a second identification that reweights the videos with action a'.

Proposition 3 (Inverse propensity weighting)**.** *Under Ass. 1–5, for every $a' \neq a$,*

$$\psi_h(a'; y, a, \bar{\mathbf{a}}, v) \;=\; \frac{\mathbb{E}\big[\,\mathbb{I}\{A = a', \bar{\mathbf{A}} = \bar{\mathbf{a}}\}\, f(Z; y, a, \bar{\mathbf{a}}, v, a')\, Y\,\big]}{\mathbb{E}\big[\,\mathbb{I}\{Y = y, A = a, \bar{\mathbf{A}} = \bar{\mathbf{a}}\}\, K_h(V - v)\,\big]}. \tag{5}$$

Here, $f(z; y, a, \bar{\mathbf{a}}, v, a') \coloneqq r(z; v)\, \eta(z; y, a, \bar{\mathbf{a}})\, \pi(z; a, \bar{\mathbf{a}}) / \pi(z; a', \bar{\mathbf{a}})$. Its components are the propensity $\pi(z; a, \bar{\mathbf{a}}) \coloneqq \Pr(A = a \mid \bar{\mathbf{A}} = \bar{\mathbf{a}}, Z = z)$, the outcome likelihood $\eta(z; y, a, \bar{\mathbf{a}}) \coloneqq \Pr(Y = y \mid A = a, \bar{\mathbf{A}} = \bar{\mathbf{a}}, Z = z)$, and the conditional scene weight $r(z; v) \coloneqq \mathbb{E}[K_h(V - v) \mid Z = z]$.

The proof is in Sec. D. In f, the classical inverse propensity factor $1/\pi(z; a', \bar{\mathbf{a}})$ corrects for how often the alternative action is taken in each scene (Hernán & Robins, 2020). In addition, $\pi(z; a, \bar{\mathbf{a}})$ and $\eta(z; y, a, \bar{\mathbf{a}})$ weight scenes by how likely the observed action and outcome are. The remaining factor $r(z; v)$ localizes the average around the observed scene. These weights require the alternative action to occur in the relevant scenes (Ass. 2). For example, if a person always flicks the switch when approaching a particular door, no video shows that door approached without flicking the switch.

4.2 EFFICIENT ESTIMATOR

An estimator based on Eq. (4) alone inherits the error of the fitted outcome regression. Similarly, one based on Eq. (5) inherits the error of the fitted weights. We combine the two through the one-step correction in Eq. (2). We then construct a cross-fitted estimator and prove that it is efficient.

To apply Eq. (2) to our target, we write $W = (Y, A, \bar{\mathbf{A}}, V)$ for a random observation and $w = (y, a, \bar{\mathbf{a}}, v)$ for the fixed video of interest. The local target is then $\psi_h(a'; w)$, and its efficient influence function is given below.

Proposition 4 (Efficient influence function)**.** *For fixed g and $h > 0$, suppose Ass. 1–5 hold and K_h and f are bounded. For $a' \neq a$, the efficient influence function in the unrestricted observed-data model of Sec. F is*

$$\phi(W, Z; w, a') = \frac{1}{\mathbb{E}[\rho_h(W; w)]} \Big[\underbrace{\rho_h(W; w)\{\mu(Z; a', \bar{\mathbf{a}}) - \psi_h(a'; w)\}}_{\textit{centred prediction term}} + \underbrace{\mathbb{I}\big\{A = a', \bar{\mathbf{A}} = \bar{\mathbf{a}}\big\}\, f(Z; w, a')\{Y - \mu(Z; a', \bar{\mathbf{a}})\}}_{\textit{weighted residual term}} \Big], \quad (6)$$

where $\rho_h(W; w) = \mathbb{I}\big\{Y = y, A = a, \bar{\mathbf{A}} = \bar{\mathbf{a}}\big\}\, K_h(V - v)$.

The derivation is given in Sec. F. The main difference from the influence function of an intervention mean (Robins et al., 1994) is that both terms of Eq. (6) are conditioned on the observed evidence. The weight $\rho_h(W; w)$ restricts the prediction term to videos with the observed actions and outcome near the observed scene. Meanwhile, the residual term uses the videos that took a'. Their outcomes need not equal y. The weight $f(Z; w, a')$ reweights these videos by how likely the observed evidence is in their scenes. Moreover, the influence functions for PN and PS (Tian & Wu, 2025) average over all covariates, whereas Eq. (6) is localized around a single observed scene.

To avoid fitting and evaluating the nuisance functions on the same videos, we apply Eq. (2) with cross-fitting (Chernozhukov et al., 2018). Specifically, we partition the videos into L folds $\{I_\ell\}_{\ell=1}^{L}$ and fit $\widehat{\mu}_\ell$ and $\widehat{f}_\ell$ outside I_ℓ. Evaluating the prediction and correction on each held out fold and combining them gives

$$\widehat{\psi}_h(a'; w) = \frac{1}{\sum_{\ell=1}^{L} \sum_{(W,Z)\in I_\ell} \rho_h(W; w)} \sum_{\ell=1}^{L} \sum_{(W,Z)\in I_\ell} \Big[\rho_h(W; w)\, \widehat{\mu}_\ell(Z; a', \bar{\mathbf{a}}) + \mathbb{I}_{a', \bar{\mathbf{a}}}(W)\, \widehat{f}_\ell(Z; w, a')\{Y - \widehat{\mu}_\ell(Z; a', \bar{\mathbf{a}})\} \Big], \quad (7)$$

where $\mathbb{I}_{a', \bar{\mathbf{a}}}(W) \coloneqq \mathbb{I}\big\{A = a',\ \bar{\mathbf{A}} = \bar{\mathbf{a}}\big\}$. The construction for each fold is detailed in Alg. 1.

The following theorem gives the asymptotic guarantees of this estimator.

Theorem 5 (Asymptotic normality and efficiency at fixed bandwidth)**.** *Fix $h > 0$ and g, and let $a' \neq a$. Suppose Ass. 1–5 and the regularity conditions in Sec. G hold, including, for each fold,*

$$\|\widehat{\mu}_\ell - \mu\|_{L_2(P_Z)}\, \|\widehat{f}_\ell - f\|_{L_2(P_Z)} = o_{\mathbb{P}}\Big(n^{-1/2}\Big).$$

Then Eq. (7) *is asymptotically normal around the local target:*

$$\sqrt{n}\,[\widehat{\psi}_h(a'; w) - \psi_h(a'; w)] \xrightarrow{\mathcal{D}} \mathcal{N}\big(0, \tau_h^2(w, a')\big).$$

The estimator is asymptotically linear with influence function $\phi(W, Z; w, a')$ and attains the efficiency bound in the model of Prop. 4. The variance $\tau_h^2(w, a')$ is specified in Sec. G.3.

The proof is in Sec. G. The theorem shows that the estimator fluctuates around the local target $\psi_h(a'; w)$ at scale $n^{-1/2}$ as $n \to \infty$, with h fixed. Both nuisance estimates must be consistent, but the rate condition concerns only the product of their errors. Neither nuisance estimator needs to reach the target estimator's $n^{-1/2}$ rate by itself: an $o_{\mathbb{P}}(n^{-1/4})$ error for each suffices, and unequal rates are allowed when their product is $o_{\mathbb{P}}(n^{-1/2})$. This makes asymptotic normality attainable with flexible nuisance models (Chernozhukov et al., 2018; Kennedy, 2024). Sufficient nuisance rates are discussed in Sec. G.

4.3 ADAPTING THE ESTIMATOR TO VIDEO

The estimator in Eq. (7) needs a scene representation, labels of actions and outcomes, and a kernel over scenes. In this section, we describe how we obtain each from videos.

Scene representation. Ass. 5 requires Z to retain the scene information that determines actions and outcomes. We compute Z from the first frame, before any action takes place. Later frames depend on the actions, so conditioning on them may introduce selection bias (Hernán & Robins, 2020). Continuous features Z^{c} come from a pretrained visual encoder. Such encoders capture object layout and support relations (Oquab et al., 2024; Zhan et al., 2024). Discrete features Z^{d} record semantic states that switch the mechanism, such as whether a device has power or a door is locked. Foundation models can read such states from the frame (OpenAI, 2026b; Google DeepMind, 2026), and annotations can also provide them.

Labels from videos. Each video gives one training tuple $(y_i, a_i, \bar{\mathbf{a}}_i, z_i)$. Action labels can come from annotations or from action recognition models adapted to the scene (Tong et al., 2022). Outcome labels can come from annotations or, for visible object states, from state recognition models applied after the action (Zang et al., 2025). On these tuples, we fit μ and η with a classifier of Y on $(A, \mathbf{A}, Z)$, and π with a classifier of A on $(\bar{\mathbf{A}}, Z)$ (Sec. I.1).

Kernel over scenes. The kernel compares the observed scene with other videos in representation space. For discrete features, we use an indicator kernel that requires exact agreement. It is the discrete kernel of Aitchison & Aitken (1976) without smoothing, as used in kernel regression with mixed data (Racine & Li, 2004). A discrete state such as power changes the mechanism itself, so scenes with different states should not be averaged together. For continuous features, we use a Gaussian kernel. Visual features lie in a high dimensional space, and scenes concentrate near nonlinear manifolds within it (Bengio et al., 2013). A kernel that weights only scenes within a fixed radius would often find no neighbours there. The Gaussian kernel instead decays smoothly with distance and gives positive weight to every scene with the same discrete state (Silverman, 1986; Wand & Jones, 1995). The bandwidth h sets how different the matched scenes may be. A small h keeps only nearly identical scenes, but few videos share a scene. A large h keeps more videos but mixes scene conditions and can increase bias relative to the pointwise target. We choose h to balance the two effects (Sec. J.3). We approximate r by averaging this feature-space kernel over samples around each video (Sec. I.1).

5 EXPERIMENTS

In this section, we evaluate our estimator on TongSim-CF and on real videos from Physics 101 (Wu et al., 2016). We also test its sensitivity to violations of the identification assumptions. Ablation studies and the effects of sample size, bandwidth, and other hyperparameters are reported in Sec. J.

5.1 EXPERIMENT SETUP

Data. On TongSim-CF, each variant provides 551 to 601 videos. Each video records the first frame, the actions, and the outcome. All methods except the counterfactually supervised model below learn only from these records. Each video serves as a test video for each of its target actions, which gives $31{,}534$ test tasks. Querying the VLMs on all of them would cost over $3{,}000$ USD, so we evaluate them on a random subset of 53 tasks from one test video per variant. On this subset, our estimator has overall MAE 0.048. For each task, a method predicts the outcome probability under the alternative value of the target action. We compare this prediction with the ground truth. The Physics 101 data are described in Sec. 5.3.

Compared baselines. We compare four groups of baselines. **(i)** *Causal methods*: **Fire-Zhu** (Fire & Zhu, 2016) estimates attribution from action statistics. Our adaptation of **LLCP** (Chen et al., 2024a) to scene features infers latent states from frame pairs before predicting counterfactual outcomes. **(ii)** *Prompted VLMs*: **GPT-5.5** (OpenAI, 2026a) and **Gemini-3-Flash** (Google, 2025) receive the first frame, the observed actions and outcome, and five demonstrations. **(iii)** *VLMs with retrieval*: the same two models additionally receive retrieved records covering both values of the target action. **(iv)** *Finetuned foundation models*: we train **Qwen2.5-VL-7B** (Bai et al., 2025) with LoRA (Hu et al., 2022) and an outcome predictor on frozen **V-JEPA 2** features (Assran et al., 2025). In addition, we train a **counterfactually supervised model** with the outcome probabilities under alternative actions as labels. This model serves as a reference with counterfactual supervision.

Table 1: MAE (mean ± SEM across variants, or clustered by object on Physics 101). *Subset of TongSim-CF. Shading marks counterfactual supervision; bold marks the best unshaded result.

		TongSim-CF				Physics 101		
Category	Method	Setting 1	Setting 2	Setting 3	Overall	Setting 1	Setting 2	Overall
Causal methods	Fire-Zhu	$\mathbf{0.017}_{\pm 0.005}$	$0.190_{\pm 0.015}$	$0.218_{\pm 0.037}$	$0.182_{\pm 0.018}$	$0.234_{\pm 0.017}$	$0.213_{\pm 0.016}$	$0.232_{\pm 0.016}$
	LLCP	$0.070_{\pm 0.008}$	$0.226_{\pm 0.024}$	$0.229_{\pm 0.033}$	$0.212_{\pm 0.019}$	$\mathbf{0.056}_{\pm 0.012}$	$0.182_{\pm 0.049}$	$\mathbf{0.068}_{\pm 0.012}$
Prompted VLMs	GPT-5.5*	$0.157_{\pm 0.096}$	$0.318_{\pm 0.048}$	$0.392_{\pm 0.070}$	$0.326_{\pm 0.038}$	$0.151_{\pm 0.012}$	$0.261_{\pm 0.043}$	$0.162_{\pm 0.012}$
	Gemini-3-Flash*	$0.162_{\pm 0.092}$	$0.350_{\pm 0.042}$	$0.403_{\pm 0.065}$	$0.349_{\pm 0.034}$	$0.093_{\pm 0.014}$	$0.205_{\pm 0.039}$	$0.103_{\pm 0.014}$
VLMs with retrieval	GPT-5.5 + retrieval*	$0.082_{\pm 0.030}$	$0.046_{\pm 0.007}$	$0.199_{\pm 0.068}$	$0.099_{\pm 0.025}$	$0.076_{\pm 0.011}$	$0.159_{\pm 0.025}$	$0.084_{\pm 0.011}$
	Gemini-3-Flash + retrieval*	$0.019_{\pm 0.004}$	$0.064_{\pm 0.006}$	$0.198_{\pm 0.048}$	$0.103_{\pm 0.020}$	$0.090_{\pm 0.014}$	$0.153_{\pm 0.023}$	$0.096_{\pm 0.013}$
Finetuned models	Qwen2.5-VL + LoRA	$0.071_{\pm 0.016}$	$0.111_{\pm 0.013}$	$0.222_{\pm 0.032}$	$0.143_{\pm 0.016}$	$0.059_{\pm 0.012}$	$0.157_{\pm 0.024}$	$0.069_{\pm 0.012}$
	V-JEPA 2 + head	$0.058_{\pm 0.004}$	$0.132_{\pm 0.011}$	$0.169_{\pm 0.027}$	$0.137_{\pm 0.012}$	$0.152_{\pm 0.017}$	$0.175_{\pm 0.052}$	$0.155_{\pm 0.016}$
Proposed	Ours	$0.019_{\pm 0.002}$	$\mathbf{0.041}_{\pm 0.005}$	$\mathbf{0.061}_{\pm 0.009}$	$\mathbf{0.045}_{\pm 0.005}$	$0.060_{\pm 0.012}$	$\mathbf{0.146}_{\pm 0.018}$	$\mathbf{0.068}_{\pm 0.012}$
Counterfactually supervised model		$0.002_{\pm 0.000}$	$0.005_{\pm 0.001}$	$0.061_{\pm 0.013}$	$0.023_{\pm 0.006}$	$0.060_{\pm 0.014}$	$0.147_{\pm 0.034}$	$0.068_{\pm 0.013}$

Implementation details. On TongSim-CF, we fit the logistic nuisance models with L-BFGS (Liu & Nocedal, 1989). Our scene representation combines discrete attributes recorded by the simulator with continuous features of the first frame. We extract these features with a pretrained VA-VAE aligned with DINOv2 (Yao et al., 2025) and compress them to two to four dimensions by PCA. We use bandwidth $h = 0.3$. We approximate the scene weight r by averaging Gaussian kernel weights over feature samples centred at the features of each video. On Physics 101, we extract features of object photographs with CLIP (Radford et al., 2021) and select the feature dimension and bandwidth on validation objects. For the prompted VLMs, we use five demonstrations from the nearest scenes, excluding the test video. Sec. I describes our estimator and all baselines in detail.

Metrics. We report the mean absolute error (MAE) between predictions and the ground truth. On TongSim-CF, the ground truth is the probability computed from the simulator (Sec. 3.2). We average errors within variants and then equally across variants, with standard errors of the mean (SEM) across variants. On Physics 101, the ground truth is the binary outcome observed in repeated trials (Sec. 5.3). We average errors over trials and cluster SEM by object.

5.2 Results on TongSim-CF

The left part of Tab. 1 shows the TongSim-CF results. Our estimator achieves the lowest overall MAE among methods without counterfactual supervision. Its advantage is largest in Setting 3, where its error is 0.061 and every baseline exceeds 0.16. **Fire-Zhu** estimates attribution from action statistics, and the **finetuned models** predict outcomes directly from images and actions. Neither adjusts for confounders, so both degrade in Settings 2 and 3. **LLCP** also does not adjust for how confounders drive action selection and remains inaccurate under confounding. The **prompted VLMs** fail in all settings, because pretrained knowledge does not reveal the mechanism of an unfamiliar environment. The **VLMs with retrieval** compare records with the same other actions and thus reduce error in Settings 1 and 2. However, they do not account for confounding scene conditions and fail in Setting 3. In contrast, **our estimator** conditions on a scene representation that includes these conditions. We also study whether our estimator can use the finetuned models as outcome predictors (Sec. J.5).

5.3 Evaluation on real videos

To test our estimator beyond simulation, we evaluate it on the ramp subset of Physics 101.

Data preparation. This subset records 101 real objects sliding down a ramp at 10 or 20 degrees. The action is the ramp angle, the outcome is whether the object slides, and the scene is a photograph of the object. Each object is tested several times at both angles. For a trial at one angle, we average the absolute error against the binary outcomes of the same object at the other angle. We hold out each object in turn and fit all methods on the other objects. To distinguish difficulty, we split the objects into two settings by whether their repeated trials at the same angle agree. In *Setting 1* (92 objects), an object either always or never slides at a given angle, so its outcome appears deterministic. In *Setting 2* (9 objects), the same object at the same angle sometimes slides and sometimes does not. Its outcome thus appears random.

Results. The right part of Tab. 1 shows the Physics 101 results. Our estimator matches the counterfactually supervised model overall. Unlike on TongSim-CF, **LLCP** and **finetuned Qwen** reach a similar overall error. This is because *Setting 1* contains about 90% of the trials, and its outcomes appear deterministic. A model that learns the association between object appearance and sliding therefore performs well there. The methods differ in *Setting 2*, where our estimator reduces the error of the strongest baseline by 5%. In 85 trials of this setting, the outcomes at the other angle are mixed. On these trials, all our predictions favour the more frequent outcome. In contrast, finetuned Qwen favours the less frequent outcome in 19% of them. LLCP errs even when the outcomes at the other angle agree. Other baselines are less accurate in both settings.

5.4 Sensitivity to identification assumptions

Since Ass. 3 and 4 cannot be verified from videos, we test the robustness of our estimator to their violations. We include LLCP, which also conditions on the observed outcome, and the stronger finetuned model, V-JEPA 2. We omit the VLM baselines, because repeated API calls across all violation levels are costly. To violate Ass. 3, we add a hidden confounder that affects both the actions and the outcome with strength γ. As shown in Fig. 2(a), the error of our estimator grows gradually with γ. This growth is expected, because the recorded scene cannot reveal the hidden confounder. Notably, LLCP improves under strong confounding ($\gamma \geq 1.5$). It reweights its latent samples by the observed outcome, and this outcome is more informative under stronger confounding. To violate Ass. 4, we couple the potential outcomes with correlation parameter ρ. In Fig. 2(b), our error grows as ρ departs from 0, and LLCP again improves as ρ increases up to 0.9. An oracle with the true outcome probabilities degrades in the same way as our estimator. The growth in error therefore reflects the violation itself. Detailed protocols and analyses are given in Sec. J.6.

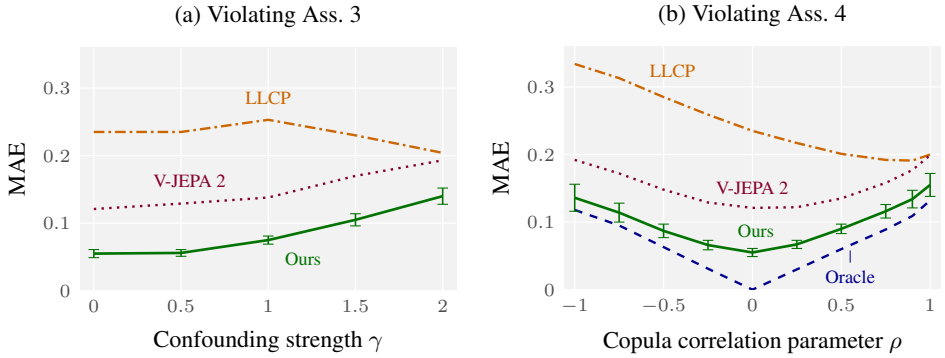

Figure 2: Sensitivity to (a) unmeasured action confounding and (b) dependence between potential outcomes, on test predictions that our estimator can form. Error bars show SEM across scenario variants for *Ours*. *Oracle* uses the true outcome probability under the alternative action.

6 Conclusion

In this paper, we formulate causal attribution in video as counterfactual inference given an observed video. To evaluate this problem, we build TongSim-CF, a household video benchmark with controlled confounding and counterfactual ground truth. Without counterfactual supervision, we identify the target from videos of the same environment and derive an estimator that attains the efficiency bound of a localized target. In the hardest simulated and real settings, our estimator reduces the error of the strongest baseline in each by 64% and 5%.

7 Limitations and Future Work

Our framework assumes that the physical conditions affecting both action selection and outcomes are visible before the action. Such a condition may be hidden, for example inside a device or outside the camera view. Its influence is then mixed with the effect of the target action, and the attribution is biased. Additional views and sensors could reveal more of these conditions. When a condition remains hidden, future work could derive partial identification bounds or use targeted interventions to separate its effect.

AI USE STATEMENT

We used large language model assistants for writing and coding. For writing, they polished the manuscript and checked formatting and bibliography entries. For coding, they helped write and run the scripts that evaluate the baselines, including the API queries to the prompted VLMs and the VLMs with retrieval, and the scripts that compute the reported metrics. The authors formulated the research problem, developed the theoretical results and proofs, designed the benchmark and the experiments, and analysed the results. The benchmark videos are rendered by a simulator, not generated by AI. Separately, the paper evaluates GPT-5.5, Gemini-3-Flash, and Qwen2.5-VL-7B as baselines, and Gemini-3-Flash also reads the discrete scene attributes from the first frame on TongSim-CF (Sec. I). The authors checked all AI-assisted text and code and take responsibility for the content of this paper.

ETHICS STATEMENT

All authors have read and adhere to the ICLR Code of Ethics. TongSim-CF is rendered by a simulator and contains no real people. Physics 101 is a public research dataset of objects sliding down a ramp. This work involves no human subjects or personal data. Our method learns from many videos of the same environment, such as recordings from a home camera. Such use requires the consent of the people recorded and protection of their privacy.

REPRODUCIBILITY STATEMENT

Sec. 2 states the identification assumptions. Secs. D, F and G give the proofs and regularity conditions. Sec. 3.2 and Sec. H specify the benchmark scenarios, structural equations, and ground truth. Sec. I describes the representations, nuisance models, and bandwidth of our estimator, and the inputs, prompts, and decoding of all baselines.

REFERENCES

Jason Abrevaya, Yu-Chin Hsu, and Robert P Lieli. Estimating conditional average treatment effects. *Journal of Business & Economic Statistics*, 33(4):485–505, 2015.

J. Aitchison and C. G. G. Aitken. Multivariate binary discrimination by the kernel method. *Biometrika*, 63(3):413–420, 1976. doi: 10.1093/biomet/63.3.413.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025. URL https://arxiv.org/abs/2506.09985.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. URL https://arxiv.org/abs/2502.13923.

Fabien Baradel, Natalia Neverova, Julien Mille, Greg Mori, and Christian Wolf. CoPhy: Counterfactual learning of physical dynamics. In *International Conference on Learning Representations (ICLR)*, 2020.

Elias Bareinboim, Juan Correa, Duligur Ibeling, and Thomas Icard. On Pearl's hierarchy and the foundations of causal inference. In Hector Geffner, Rita Dechter, and Joseph Y. Halpern (eds.), *Probabilistic and Causal Inference: The Works of Judea Pearl*, pp. 507–556. ACM Books, 2022. doi: 10.1145/3501714.3501743.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013. doi: 10.1109/TPAMI.2013.50.

Peter J Bickel, Chris AJ Klaassen, Ya'acov Ritov, and Jon A Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, 1993.

Florian Bordes, Quentin Garrido, Justine T. Kao, Adina Williams, Michael Rabbat, and Emmanuel Dupoux. IntPhys 2: Benchmarking intuitive physics understanding in complex synthetic environments. *arXiv preprint arXiv:2506.09849*, 2025.

Guangyi Chen, Yuke Li, Xiao Liu, Zijian Li, Eman Al Suradi, Donglai Wei, and Kun Zhang. LLCP: Learning latent causal processes for reasoning-based video question answer. In *International Conference on Learning Representations (ICLR)*, 2024a. URL https://openreview.net/forum?id=Cu5wJa5LGO.

Meiqi Chen, Bo Peng, Yan Zhang, and Chaochao Lu. CELLO: Causal evaluation of large vision-language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22353–22374, 2024b. doi: 10.18653/v1/2024.emnlp-main.1247.

Tieyuan Chen, Huabin Liu, Tianyao He, Yihang Chen, Chaofan Gan, Xiao Ma, Cheng Zhong, Yang Zhang, Yingxue Wang, Hui Lin, et al. MECD: Unlocking multi-event causal discovery in video reasoning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 92554–92580, 2024c. doi: 10.52202/079017-2939.

Tieyuan Chen, Huabin Liu, Yi Wang, Yihang Chen, Tianyao He, Chaofan Gan, Huanyu He, and Weiyao Lin. MECD+: Unlocking event-level causal graph discovery for video reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):2628–2645, 2026. doi: 10.1109/TPAMI.2025.3626549.

Yuefei Chen, Jiang Liu, Xiaodong Lin, and Ruixiang Tang. Distilling counterfactual reasoning from language to vision: Causal graph guided post-training for video understanding. *arXiv preprint arXiv:2511.19923*, 2025.

Zhenfang Chen, Kexin Yi, Yunzhu Li, Mingyu Ding, Antonio Torralba, Joshua B. Tenenbaum, and Chuang Gan. ComPhy: Compositional physical reasoning of objects and events from videos. In *International Conference on Learning Representations (ICLR)*, 2022.

Tongtong Cheng, Rongzhen Li, Yixin Xiong, Tao Zhang, Jing Wang, and Kai Liu. MCAM: Multimodal causal analysis model for ego-vehicle-level driving video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5479–5489, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Cheng_MCAM_Multimodal_Causal_Analysis_Model_for_Ego-Vehicle-Level_Driving_Video_Understanding_ICCV_2025_paper.html.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Pim de Haan, Dinesh Jayaraman, and Sergey Levine. Causal confusion in imitation learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. In *Advances in Neural Information Processing Systems*, volume 29, pp. 64–72, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/d9d4f495e875a2e075a1a4a6e1b9770f-Abstract.html.

Amy Fire and Song-Chun Zhu. Learning perceptual causality from video. *ACM Transactions on Intelligent Systems and Technology*, 7(2):23:1–23:22, 2016. doi: 10.1145/2809782.

Aaron Foss, Chloe Evans, Sasha Mitts, Koustuv Sinha, Ammar Rizvi, and Justine T. Kao. CausalVQA: A physically grounded causal reasoning benchmark for video models. *arXiv preprint arXiv:2506.09943*, 2025.

Google. Gemini 3 Flash: Frontier intelligence built for speed. https://blog.google/products-and-platforms/products/gemini/gemini-3-flash/, 2025. Accessed 2026-05-03.

Google DeepMind. Gemini 3.8 Flash model card, September 2026. URL https://deepmind.google/models/model-cards/gemini-3-8-flash/. Published September 2, 2026. Accessed September 24, 2026.

Klaus Greff, Raphaël Lopez Kaufman, Rishabh Kabra, Nick Watters, Christopher Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. Multi-object representation learning with iterative variational inference. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2424–2433, 2019. URL https://proceedings.mlr.press/v97/greff19a.html.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14375–14385, 2024. doi: 10.1109/CVPR52733.2024.01363.

Agrim Gupta, Silvio Savarese, Surya Ganguli, and Li Fei-Fei. Embodied intelligence via learning and evolution. *Nature Communications*, 12(1):5721, 2021. doi: 10.1038/s41467-021-25874-z.

Miguel A. Hernán and James M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC, Boca Raton, 2020. Available at https://miguelhernan.org/whatifbook.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

Edward H. Kennedy. Semiparametric doubly robust targeted double machine learning: A review. In *Handbook of Statistical Methods for Precision Medicine*, pp. 207–236. Chapman and Hall/CRC, 2024. doi: 10.1201/9781003216223-10.

Adam Lerer, Sam Gross, and Rob Fergus. Learning physical intuition of block towers by example. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 430–438, 2016. URL https://proceedings.mlr.press/v48/lerer16.html.

Jiangtong Li, Li Niu, and Liqing Zhang. From representation to reasoning: Towards both evidence and commonsense reasoning for video question-answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21273–21282, 2022.

Wei Li, Zitong Lu, Jinzhu Jia, Min Xie, and Zhi Geng. Retrospective causal inference with multiple effect variables. *Biometrika*, 111(2):573–589, 2024. doi: 10.1093/biomet/asad056.

Yunzhu Li, Antonio Torralba, Animashree Anandkumar, Dieter Fox, and Animesh Garg. Causal discovery in physical systems from videos. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Dong C. Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1–3):503–528, 1989. doi: 10.1007/BF01589116.

Zitong Lu, Zhi Geng, Wei Li, Shengyu Zhu, and Jinzhu Jia. Evaluating causes of effects by posterior effects of causes. *Biometrika*, 110(2):449–465, 2023. doi: 10.1093/biomet/asac038.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

OpenAI. GPT-5.5 model. https://developers.openai.com/api/docs/models/gpt-5.5, 2026a. Accessed September 24, 2026.

OpenAI. GPT-6 Astra system card, September 2026b. URL https://deploymentsafety.openai.com/gpt-6-astra. Published September 3, 2026; updated September 22, 2026. Accessed September 24, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Paritosh Parmar, Eric Peh, Ruirui Chen, Ting En Lam, Yuhan Chen, Elston Tan, and Basura Fernando. CausalChaos! dataset for comprehensive causal action question answering over longer causal chains grounded in dynamic visual scenes. In *Advances in Neural Information Processing Systems*, volume 37, pp. 92769–92802, 2024. doi: 10.52202/079017-2946.

Maitreya Patel, Tejas Gokhale, Chitta Baral, and Yezhou Yang. CRIPP-VQA: Counterfactual reasoning about implicit physical properties via video question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9856–9870, 2022.

Nick Pawlowski, Daniel C. Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://papers.nips.cc/paper/2020/hash/0987b8b338d6c90bbedd8631bc499221-Abstract.html.

Judea Pearl. Probabilities of causation: Three counterfactual interpretations and their identification. *Synthese*, 121(1–2):93–149, 1999. doi: 10.1023/A:1005233831499.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009.

Yujia Peng, Jiaheng Han, Zhenliang Zhang, Lifeng Fan, Tengyu Liu, Siyuan Qi, Xue Feng, Yuxi Ma, Yizhou Wang, and Song-Chun Zhu. The Tong test: Evaluating artificial general intelligence through dynamic embodied physical and social interactions. *Engineering*, 34:12–22, 2024. doi: 10.1016/j.eng.2023.07.006.

Jeff Racine and Qi Li. Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics*, 119(1):99–130, 2004. doi: 10.1016/S0304-4076(03)00157-X.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

James Robins. A new approach to causal inference in mortality studies with a sustained exposure period–application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7 (9–12):1393–1512, 1986. doi: 10.1016/0270-0255(86)90088-6.

James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427): 846–866, 1994. doi: 10.1080/01621459.1994.10476818.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974. doi: 10.1037/h0037350.

Changyu Shen, Jaesik Jeong, Xiaochun Li, Peng-Sheng Chen, and Alfred E. Buxton. Treatment benefit and treatment harm rate to characterize heterogeneity in treatment effect. *Biometrics*, 69(3): 724–731, 2013.

Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 9487–9499, 2019.

B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London, 1986.

Zhe Sun, Kunlun Wu, Chuanjian Fu, Zeming Song, Langyong Shi, Zihe Xue, Bohan Jing, Ying Yang, Xiaomeng Gao, Aijia Li, et al. TongSIM: A general platform for simulating intelligent machines. *arXiv preprint arXiv:2512.20206*, 2025.

Jin Tian and Judea Pearl. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28(1–4):287–313, 2000. doi: 10.1023/A:1018912507879.

Zhaoqing Tian and Peng Wu. Semiparametric efficient inference for the probability of necessary and sufficient causation. *Statistics in Medicine*, 44(18–19):e70242, 2025. doi: 10.1002/sim.70242.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, volume 35, pp. 10078–10093, 2022. doi: 10.52202/068431-0732.

Anastasios A Tsiatis. *Semiparametric Theory and Missing Data*. Springer, 2006.

Mark J van der Laan and Sherri Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, 2011.

Victor Veitch, Dhanya Sridhar, and David Blei. Adapting text embeddings for causal inference. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 919–928, 2020.

M. P. Wand and M. C. Jones. *Kernel Smoothing*. Chapman and Hall, London, 1995. doi: 10.1201/b14876.

Jiajun Wu, Joseph J. Lim, Hongyi Zhang, Joshua B. Tenenbaum, and William T. Freeman. Physics 101: Learning physical object properties from unlabeled videos. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2016.

Te-Lin Wu, Zi-Yi Dou, Qingyuan Hu, Yu Hou, Nischal Chandra, Marjorie Freedman, Ralph Weischedel, and Nanyun Peng. ACQUIRED: A dataset for answering counterfactual questions in real-life videos. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 11753–11770, 2023. doi: 10.18653/v1/2023.emnlp-main.719.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15703–15712, 2025.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. CLEVRER: Collision events for video representation and reasoning. In *International Conference on Learning Representations (ICLR)*, 2020.

Yuan Zang, Zitian Tang, Junho Cho, Jaewook Yoo, and Chen Sun. SAGE: A unified framework for generalizable object state recognition with state-action graph embedding. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.neurips.cc/paper_files/paper/2025/hash/5623c35f3ab5e2c72aeb3abce27dc28f-Abstract-Conference.html.

Guanqi Zhan, Chuanxia Zheng, Weidi Xie, and Andrew Zisserman. A general protocol to probe large vision models for 3D physical understanding. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/4ca9b0906de992ae32c0001915fbea22-Abstract-Conference.html.

Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Xiaonan Huang, Changxin Gao, Shanjun Zhang, Li Yu, and Nong Sang. Holmes-VAU: Towards long-term video anomaly understanding at any granularity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13843–13853, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhang_Holmes-VAU_Towards_Long-term_Video_Anomaly_Understanding_at_Any_Granularity_CVPR_2025_paper.html.

A Related Work

Video causal reasoning. Video question answering benchmarks ask how an outcome would change under a hypothetical change to the scene. CLEVRER (Yi et al., 2020) and CRIPP-VQA (Patel et al., 2022) grade such answers with simulator reruns, and ComPhy (Chen et al., 2022) adds reference videos of the same objects. Later benchmarks extend this format to long animations (Parmar et al., 2024; Chen et al., 2025) and real videos (Li et al., 2022; Wu et al., 2023; Foss et al., 2025). Event graph methods instead predict causal relations between events (Chen et al., 2024c; 2026), and Fire-Zhu attributes an outcome to the actions that occur just before it (Fire & Zhu, 2016). Dynamics models infer hidden physical properties or interaction graphs from motion (Baradel et al., 2020; Li et al., 2020). Deep structural causal models generate counterfactual images by reusing inferred noise (Pawlowski et al., 2020). These methods learn from counterfactual answers or model the full dynamics. In contrast, we estimate the probability of one outcome from observed videos without counterfactual labels.

Attribution and efficient estimation. Probabilities of necessity and sufficiency formalise attribution (Tian & Pearl, 2000; Pearl, 2009). Their identification requires a restriction on the joint potential outcomes, such as monotonicity or conditional independence (Tian & Wu, 2025). Tian & Wu (2025) derive efficient estimators of these probabilities averaged over covariates. We instead target attribution at a given scene, conditional on the other observed actions. Our estimator builds on influence functions (Robins et al., 1994; Bickel et al., 1993; Tsiatis, 2006) and double machine learning (Chernozhukov et al., 2018). These tools have been applied to text (Veitch et al., 2020) and biomedical signals (Shi et al., 2019).

A.1 Comparison of benchmark protocols

Tab. 2 compares benchmarks on the two properties in Sec. 3.1. CLEVRER and CRIPP-VQA provide counterfactual answers but no observed alternatives of the target action. ComPhy fixes the physical properties of objects across its reference videos, but not the other actions or scene conditions. IntPhys 2 (Bordes et al., 2025) pairs possible and impossible events, which tests physical plausibility rather than action effects. MECD studies confounding among events but has no videos in which the target action varies under comparable conditions. TongSim-CF provides such videos and adds controlled confounding.

Table 2: Benchmark protocols. Overlap means observed alternatives of the target action under comparable scenes and other actions. For MECD, confounding concerns relations among events. ✓: provided or tested; △: partly provided; ×: not provided.

Benchmark	Confounding	Overlap
CLEVRER, CRIPP-VQA	×	×
ComPhy	×	△
IntPhys 2	×	△
MECD	✓	×
TongSim-CF (ours)	✓	✓

B Notation

Tab. 3 lists the symbols used in the appendix in addition to those of the main text.

C Transfer Lemmas from V to Z

The proofs condition on the representation Z instead of the frame V. Under Ass. 5, the two lemmas below transfer the conditions of Ass. 3 and 4 from V to Z. All conditional identities below hold almost surely on the relevant conditioning support; nuisance functions may be extended by zero off that support.

Lemma 6 (Transfer of no unmeasured confounding)**.** *Suppose Ass. 3 and 5 hold. Then for every* $a \in \mathcal{A}$, $Y_a \perp\!\!\!\perp A \mid \bar{\mathbf{A}}, Z$.

Table 3: Notation used in the appendix.

Symbol	Meaning
$W \coloneqq (Y, A, \bar{\mathbf{A}}, V)$	Observed tuple without the representation
$O \coloneqq (W, Z)$	Full observation
Y_a	Potential outcome under action $a \in \mathcal{A}$
$w = (y, a, \bar{\mathbf{a}}, v)$	Observed evidence of the query
$\psi_h(a'; w)$	Kernel localised target in Eq. (3)
$\rho_h(W; w) \coloneqq \mathbb{I}\{Y = y, A = a, \bar{\mathbf{A}} = \bar{\mathbf{a}}\}\, K_h(V - v)$	Localisation weight
μ, π, η, r, f	Nuisance functions in Sec. 4.1
$\phi(W, Z; w, a')$	Efficient influence function in Prop. 4
$\mathbb{P}_n[g] \coloneqq n^{-1} \sum_{i=1}^{n} g(O_i)$	Empirical mean
$\mathbb{P}[g] \coloneqq \mathbb{E}[g(O)]$	Population expectation

Proof. Fix $a \in \mathcal{A}$. We compute the conditional joint law using $(Y_a, A) \perp\!\!\!\perp V \mid \bar{\mathbf{A}}, Z$, which follows by conditioning the joint independence in Ass. 5 on $\bar{\mathbf{A}}$:

$$\begin{aligned}
\mathbb{P}\big(Y_a, A \mid \bar{\mathbf{A}}, Z\big) &= \mathbb{P}\big(Y_a, A \mid \bar{\mathbf{A}}, V, Z\big) && \text{by Ass. 5} \\
&= \mathbb{P}\big(Y_a, A \mid \bar{\mathbf{A}}, V\big) && \text{since } Z \text{ is a function of } V \\
&= \mathbb{P}\big(Y_a \mid \bar{\mathbf{A}}, V\big)\, \mathbb{P}\big(A \mid \bar{\mathbf{A}}, V\big) && \text{by Ass. 3} \\
&= \mathbb{P}\big(Y_a \mid \bar{\mathbf{A}}, V, Z\big)\, \mathbb{P}\big(A \mid \bar{\mathbf{A}}, V, Z\big) && \text{since } Z \text{ is a function of } V \\
&= \mathbb{P}\big(Y_a \mid \bar{\mathbf{A}}, Z\big)\, \mathbb{P}\big(A \mid \bar{\mathbf{A}}, Z\big). && \text{by Ass. 5}
\end{aligned}$$

□

Lemma 7 (Transfer of cross-world independence)**.** *Suppose Ass. 4 and 5 hold. Then for every pair of distinct actions $a, a' \in \mathcal{A}$, $Y_a \perp\!\!\!\perp Y_{a'} \mid \bar{\mathbf{A}}, Z$.*

Proof. Fix distinct actions $a, a' \in \mathcal{A}$. We compute the joint law by first applying Ass. 4, and then using Ass. 5 separately for each marginal:

$$\begin{aligned}
\mathbb{P}\big(Y_a, Y_{a'} \mid \bar{\mathbf{A}}, Z\big) &= \mathbb{E}\big[\mathbb{P}\big(Y_a, Y_{a'} \mid \bar{\mathbf{A}}, V, Z\big) \mid \bar{\mathbf{A}}, Z\big] && \text{tower property} \\
&= \mathbb{E}\big[\mathbb{P}\big(Y_a, Y_{a'} \mid \bar{\mathbf{A}}, V\big) \mid \bar{\mathbf{A}}, Z\big] && \text{since } Z \text{ is a function of } V \\
&= \mathbb{E}\big[\mathbb{P}\big(Y_a \mid \bar{\mathbf{A}}, V\big)\, \mathbb{P}\big(Y_{a'} \mid \bar{\mathbf{A}}, V\big) \mid \bar{\mathbf{A}}, Z\big] && \text{by Ass. 4} \\
&= \mathbb{E}\big[\mathbb{P}\big(Y_a \mid \bar{\mathbf{A}}, Z\big)\, \mathbb{P}\big(Y_{a'} \mid \bar{\mathbf{A}}, Z\big) \mid \bar{\mathbf{A}}, Z\big] && \text{by Ass. 5} \\
&= \mathbb{P}\big(Y_a \mid \bar{\mathbf{A}}, Z\big)\, \mathbb{P}\big(Y_{a'} \mid \bar{\mathbf{A}}, Z\big). && \text{measurability}
\end{aligned}$$

□

D Proof of Props. 2 and 3

We fix the query tuple $(y, a, a', \bar{\mathbf{a}}, v)$ and bandwidth h throughout. The positive normalisation in Eq. (3) implies $\Pr(Y = y, A = a, \bar{\mathbf{A}} = \bar{\mathbf{a}}) > 0$, hence $\Pr(\bar{\mathbf{A}} = \bar{\mathbf{a}}) > 0$. Moreover, Ass. 2 transfers to Z: for $P_{\bar{\mathbf{a}}}$-almost every z,

$$\pi(z; a', \bar{\mathbf{a}}) = \mathbb{E}[\Pr(A = a' \mid \bar{\mathbf{A}} = \bar{\mathbf{a}}, V) \mid \bar{\mathbf{A}} = \bar{\mathbf{a}}, Z = z] > 0,$$

where $P_{\bar{\mathbf{a}}}$ denotes the law of Z given $\bar{\mathbf{A}} = \bar{\mathbf{a}}$.

D.1 Proof of Prop. 2

By the definition of ψ_h in Eq. (3),

$$\psi_h(a'; y, a, \bar{\mathbf{a}}, v) = \frac{\mathbb{E}[\rho_h(W; w)\, Y_{a'}]}{\mathbb{E}[\rho_h(W; w)]}.$$

It therefore suffices to show that $\mathbb{E}[\rho_h(W; w)\, Y_{a'}] = \mathbb{E}[\rho_h(W; w)\, \mu(Z; a', \bar{\mathbf{a}})]$. Let $\widetilde{\rho} \coloneqq \mathbb{I}\big\{Y_a = y,\, A = a,\, \bar{\mathbf{A}} = \bar{\mathbf{a}}\big\}\, K_h(V - v)$. By Ass. 1, $\rho_h(W; w) = \widetilde{\rho}$, which is measurable with

respect to $\big(Y_a, A, \bar{\mathbf{A}}, V\big)$. In the second equality below, the joint independence in Ass. 3 gives $Y_{a'} \perp\!\!\!\perp A \mid Y_a, \bar{\mathbf{A}}, V$; independence of each potential outcome from A separately would not suffice for this step. We compute

$$\begin{aligned}
\mathbb{E}[\rho_h(W;w)\, Y_{a'}] &= \mathbb{E}\big[\widetilde{\rho}\,\mathbb{E}\big[Y_{a'} \mid Y_a,\ A,\ \bar{\mathbf{A}},\ V\big]\big] && \text{consistency and iterated expectation}\\
&= \mathbb{E}\big[\widetilde{\rho}\,\mathbb{E}\big[Y_{a'} \mid Y_a,\ \bar{\mathbf{A}},\ V\big]\big] && \text{by Ass. 3}\\
&= \mathbb{E}\big[\widetilde{\rho}\,\mathbb{E}\big[Y_{a'} \mid \bar{\mathbf{A}},\ V\big]\big] && \text{by Ass. 4, using } a' \neq a\\
&= \mathbb{E}\big[\rho_h(W;w)\,\mathbb{E}\big[Y_{a'} \mid A = a',\ \bar{\mathbf{A}},\ V\big]\big] && \text{by Ass. 1 and 3}\\
&= \mathbb{E}[\rho_h(W;w)\,\mu(Z;a',\bar{\mathbf{a}})]\,. && \text{by Ass. 1 and 5}
\end{aligned}$$

Ass. 2 ensures that $\mu(\cdot;a',\bar{\mathbf{a}})$ is well defined on the support reached by ρ_h. The last equality uses Ass. 5 conditional on $A = a', \bar{\mathbf{A}} = \bar{\mathbf{a}}$, consistency, and the fact that ρ_h vanishes unless $\bar{\mathbf{A}} = \bar{\mathbf{a}}$.

D.2 Proof of Prop. 3

We start from the outcome regression form proved above. Define

$$q(z;v) \coloneqq \mathbb{E}\big[\mathbb{I}\{Y = y,\ A = a\}\, K_h(V - v) \mid \bar{\mathbf{A}} = \bar{\mathbf{a}},\ Z = z\big]\,.$$

By the tower property, $\mathbb{E}[\rho_h(W;w)\,\mu(Z;a',\bar{\mathbf{a}})] = \mathbb{E}\big[\mathbb{I}\big\{\bar{\mathbf{A}} = \bar{\mathbf{a}}\big\}\, q(Z;v)\,\mu(Z;a',\bar{\mathbf{a}})\big]$. Consistency and Ass. 5 imply $(Y, A, \bar{\mathbf{A}}) \perp\!\!\!\perp V \mid Z$ and $\bar{\mathbf{A}} \perp\!\!\!\perp V \mid Z$. Consequently,

$$\begin{aligned}
q(z;v) &= \mathbb{E}[K_h(V - v)\Pr(Y = y, A = a \mid \bar{\mathbf{A}} = \bar{\mathbf{a}}, V, Z = z) \mid \bar{\mathbf{A}} = \bar{\mathbf{a}}, Z = z]\\
&= \eta(z;y,a,\bar{\mathbf{a}})\pi(z;a,\bar{\mathbf{a}})\mathbb{E}[K_h(V - v) \mid \bar{\mathbf{A}} = \bar{\mathbf{a}}, Z = z]\\
&= r(z;v)\eta(z;y,a,\bar{\mathbf{a}})\pi(z;a,\bar{\mathbf{a}}).
\end{aligned}$$

Hence, by the definition of f,

$$\begin{aligned}
\mathbb{E}\big[\mathbb{I}\big\{\bar{\mathbf{A}} = \bar{\mathbf{a}}\big\}\, q(Z;v)\,\mu(Z;a',\bar{\mathbf{a}})\big] &= \mathbb{E}\big[\mathbb{I}\big\{\bar{\mathbf{A}} = \bar{\mathbf{a}}\big\}\,\pi(Z;a',\bar{\mathbf{a}})\, f(Z;w,a')\,\mu(Z;a',\bar{\mathbf{a}})\big]\\
&= \mathbb{E}\big[\mathbb{I}\big\{A = a',\ \bar{\mathbf{A}} = \bar{\mathbf{a}}\big\}\, f(Z;w,a')\,\mu(Z;a',\bar{\mathbf{a}})\big]\\
&= \mathbb{E}\big[\mathbb{I}\big\{A = a',\ \bar{\mathbf{A}} = \bar{\mathbf{a}}\big\}\, f(Z;w,a')\, Y\big]\,,
\end{aligned}$$

where the last step uses the definition of μ. Dividing both sides by $\mathbb{E}[\rho_h(W;w)]$ yields the claim.

D.3 Scene distribution given local observed evidence

Let $P_{\bar{\mathbf{a}}}$ be the distribution of Z conditional on $\bar{\mathbf{A}} = \bar{\mathbf{a}}$, and define

$$\begin{aligned}
q_h(z;w) &= r(z;v)\,\eta(z;y,a,\bar{\mathbf{a}})\,\pi(z;a,\bar{\mathbf{a}}),\\
\mathrm{d}Q_h(z \mid w) &= \frac{q_h(z;w)\,\mathrm{d}P_{\bar{\mathbf{a}}}(z)}{\int q_h(u;w)\,\mathrm{d}P_{\bar{\mathbf{a}}}(u)}.
\end{aligned}$$

Here $q_h \geq 0$ and $\int q_h\,\mathrm{d}P_{\bar{\mathbf{a}}} = \mathbb{E}[\rho_h]/\Pr(\bar{\mathbf{A}} = \bar{\mathbf{a}}) \in (0,\infty)$, so Q_h is a probability distribution. The factorisation of q in the IPW proof gives

$$\psi_h(a';w) = \int \mu(z;a',\bar{\mathbf{a}})\,\mathrm{d}Q_h(z \mid w).$$

The evidence thus defines a distribution over nearby scenes, and μ gives the alternative-action response in each scene. Through $\eta(z;y,a,\bar{\mathbf{a}})$, this distribution can change with the observed outcome. Hence the observed scene and the observed outcome jointly determine which scenes are plausible. The scene v restricts the average to a neighbourhood, and the outcome y weights the scenes within it. This holds even though the potential outcomes are independent within each scene. Exact conditioning on $V = v$ leaves a single scene; localization instead gives a target that admits regular estimation at fixed bandwidth (Sec. 4).

E Background on Efficient Estimation via Influence Functions

Efficient estimators have minimum asymptotic variance among all regular asymptotically linear estimators (Bickel et al., 1993), and are favored in causal inference (van der Laan & Rose, 2011;

Chernozhukov et al., 2018). Influence functions are a common tool to construct them. Specifically, given a target estimand $\Psi(\mathbb{P})$, an influence function $\phi(O;\mathbb{P})$ is a function satisfying

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\Psi(\mathbb{P}_\varepsilon)\Big|_{\varepsilon=0} = \mathbb{E}_{\mathbb{P}}[\phi(O;\mathbb{P})\, s(O;\mathbb{P})] \quad \text{and} \quad \mathbb{E}_{\mathbb{P}}[\phi(O;\mathbb{P})] = 0,$$

where $\{\mathbb{P}_\varepsilon : |\varepsilon| < \delta\}$ is a one-dimensional smooth path through $\mathbb{P}$ and s is the score along this path. For a fitted distribution $\widehat{P}$, a first-order expansion gives

$$\Psi(P) - \Psi(\widehat{P}) = \mathbb{E}_P[\phi(O;\widehat{P})] + R_2(P,\widehat{P}),$$

where R_2 is the higher-order remainder. Averaging $\phi(O_i;\widehat{P})$ estimates this leading error and yields the one-step estimator in Eq. (2). Cross-fitting and suitable remainder conditions yield asymptotic linearity; efficiency additionally requires the efficient influence function for the statistical model (Chernozhukov et al., 2018).

F Proof of Prop. 4

Fix the query (w, a'), the representation map g, and $h > 0$. On the unrestricted observed-data model for W, with $Z = g(V)$, define the statistical functional

$$\Psi_h^{\mathrm{OR}}(P) = \frac{\mathbb{E}_P[\rho_h(W;w)\,\mu_P(Z;a',\bar{\mathbf{a}})]}{\mathbb{E}_P[\rho_h(W;w)]}.$$

At the true distribution P, Ass. 1–5 identify this functional with $\psi_h(a';w)$ by Prop. 2. Write $D = \mathbb{E}[\rho_h]$ and $N = \mathbb{E}[\rho_h\mu]$. Assume $D > 0$, and that K_h and f are bounded at this fixed bandwidth; since Y is binary, these bounds ensure $\phi \in L_2(P)$. Consider a dominated regular parametric submodel $\{P_\varepsilon\}$ through $P_0 = P$, with score $s(O) = \partial_\varepsilon \log \mathrm{d}P_\varepsilon(O)|_{\varepsilon=0}$, for which differentiation under the following integrals is valid. It suffices to use bounded mean-zero scores and paths $\mathrm{d}P_\varepsilon = (1+\varepsilon s)\mathrm{d}P$: these scores are dense in $L_2^0(P)$, and the stated bounds justify the differentiations below. The submodel need not satisfy Ass. 5 away from P. In general its transport weight is

$$\frac{q_\varepsilon(z;w)}{\pi_\varepsilon(z;a',\bar{\mathbf{a}})}, \qquad q_\varepsilon(z;w) = \mathbb{E}_\varepsilon[\mathbb{I}\{Y=y, A=a\}K_h(V-v) \mid \bar{\mathbf{A}} = \bar{\mathbf{a}}, Z = z].$$

At $\varepsilon = 0$, Ass. 5 gives $q_0 = r\eta\pi(\cdot;a,\bar{\mathbf{a}})$, so this weight equals f.

Derivative of D. By the score identity, $\partial_\varepsilon D(\varepsilon)\big|_{\varepsilon=0} = \mathbb{E}[\rho_h(W;w)\,s(O)] = \mathbb{E}[\{\rho_h(W;w) - D\}\,s(O)]$.

Derivative of N. Let $\mu_\varepsilon(z;a',\bar{\mathbf{a}}) := \mathbb{E}_\varepsilon\big[Y \mid A = a',\ \bar{\mathbf{A}} = \bar{\mathbf{a}},\ Z = z\big]$ and $N(\varepsilon) = \mathbb{E}_\varepsilon[\rho_h(W;w)\,\mu_\varepsilon(Z;a',\bar{\mathbf{a}})]$. By the product rule,

$$\partial_\varepsilon N(\varepsilon)\big|_{\varepsilon=0} = \mathbb{E}[\rho_h(W;w)\,\mu(Z;a',\bar{\mathbf{a}})\,s(O)] + \mathbb{E}[\rho_h(W;w)\,\dot{\mu}(Z;a',\bar{\mathbf{a}})]\,,$$

where $\dot{\mu}(z;a',\bar{\mathbf{a}}) := \partial_\varepsilon\mu_\varepsilon(z;a',\bar{\mathbf{a}})\big|_{\varepsilon=0}$. For the bounded-score path above, write C_z for conditioning on $A = a', \bar{\mathbf{A}} = \bar{\mathbf{a}}, Z = z$. Its conditional mean satisfies

$$\frac{\mu_\varepsilon(z) - \mu(z)}{\varepsilon} = \frac{\mathbb{E}[(Y-\mu(z))s(O) \mid C_z]}{1 + \varepsilon\mathbb{E}[s(O) \mid C_z]}.$$

For $|\varepsilon|\|s\|_\infty < 1/2$, the denominator is at least $1/2$ and the difference quotient is uniformly bounded. Thus $\dot{\mu}(z) = \mathbb{E}[(Y-\mu(z))s(O) \mid C_z]$, and bounded ρ_h permits dominated convergence in the preceding product rule. Using the tower property and the definitions of π and f,

$$\begin{aligned}\mathbb{E}[\rho_h(W;w)\,\dot{\mu}(Z;a',\bar{\mathbf{a}})] &= \mathbb{E}\big[\mathbb{I}\{\bar{\mathbf{A}} = \bar{\mathbf{a}}\}q_0(Z;w)\dot{\mu}(Z;a',\bar{\mathbf{a}})\big] \\ &= \mathbb{E}\big[\mathbb{I}\big\{A = a',\ \bar{\mathbf{A}} = \bar{\mathbf{a}}\big\}\, f(Z;w,a')\,\dot{\mu}(Z;a',\bar{\mathbf{a}})\big]\,.\end{aligned}$$

Moreover, since $\mu_\varepsilon(z;a',\bar{\mathbf{a}}) = \mathbb{E}_\varepsilon\big[Y \mid A = a',\ \bar{\mathbf{A}} = \bar{\mathbf{a}},\ Z = z\big]$, for any fixed bounded $t(Z)$,

$$\mathbb{E}_\varepsilon\big[\mathbb{I}\big\{A = a',\ \bar{\mathbf{A}} = \bar{\mathbf{a}}\big\}\,\{Y - \mu_\varepsilon(Z;a',\bar{\mathbf{a}})\}\,t(Z)\big] = 0.$$

Differentiating at $\varepsilon = 0$ with $t(Z) = f(Z;w,a')$ held fixed gives

$$\mathbb{E}[\mathbb{I}_{a',\bar{\mathbf{a}}} f\dot{\mu}] = \mathbb{E}[\mathbb{I}_{a',\bar{\mathbf{a}}} f(Y-\mu)s].$$

Substitution into the product rule yields

$$\begin{aligned}\partial_\varepsilon N(\varepsilon)\big|_{\varepsilon=0} &= \mathbb{E}[\{\rho_h(W;w)\,\mu(Z;a',\bar{\mathbf{a}}) - N\}\,s(O)] \\ &\quad + \mathbb{E}\big[\mathbb{I}\big\{A = a',\ \bar{\mathbf{A}} = \bar{\mathbf{a}}\big\}\, f(Z;w,a')\,\{Y - \mu(Z;a',\bar{\mathbf{a}})\}\,s(O)\big]\,.\end{aligned}$$

Derivative of $\Psi_h^{\mathrm{OR}} = N/D$. By the quotient rule and using $N/D = \psi_h(a'; w)$,

$$\partial_\varepsilon \Psi_h^{\mathrm{OR}}(P_\varepsilon)\big|_{\varepsilon=0} = \frac{\dot{N} - \psi_h \dot{D}}{D} = \mathbb{E}[\phi(W, Z; w, a')\, s(O)]\,,$$

where ϕ is given in Eq. (6). Its mean is zero: the prediction term integrates to $N - \psi_h D = 0$, and the residual term has conditional mean zero given $(A, \bar{\mathbf{A}}, Z)$. The unrestricted observed-data model has tangent space $L_2^0(P)$, the space of square-integrable mean-zero functions of the observation. Thus the displayed derivative and $\phi \in L_2^0(P)$ identify its unique canonical gradient. This establishes efficiency for Ψ_h^{OR} in that model; at P, the functional equals the causal local target.

G Proof of Thm. 5

We fix the representation map g, evidence $w = (y, a, \bar{\mathbf{a}}, v)$, alternative action $a' \neq a$, and bandwidth $h > 0$ throughout the sampling limit. For brevity, write $\psi = \psi_h(a'; w)$, $\widehat{\psi} = \widehat{\psi}_h(a'; w)$, $\rho_h(W) = \rho_h(W; w)$, $\mu(Z) = \mu(Z; a', \bar{\mathbf{a}})$, $f(Z) = f(Z; w, a')$, and $\phi(O) = \phi(W, Z; w, a')$.

G.1 Regularity conditions

Thm. 5 requires the following conditions in addition to Ass. 1–5.

(i) The observations $O_1, \ldots, O_n$ are i.i.d. The data are split independently of their values into a fixed number $L \geq 2$ of folds, with $n_\ell = |I_\ell|$ and $n_\ell/n \to \lambda_\ell \in (0, 1)$. The nuisance functions for fold ℓ are fitted using only observations outside I_ℓ.

(ii) The kernel K is bounded and nonnegative, with unit integral. The bandwidth $h > 0$ is fixed, and $D = \mathbb{P}[\rho_h(W)] > 0$.

(iii) Y, μ, and f are bounded. With probability tending to one, all $\widehat{\mu}_\ell$ and $\widehat{f}_\ell$ are bounded by a common finite constant, independent of n.

(iv) For every fold, $\|\widehat{\mu}_\ell - \mu\|_{L_2(P_Z)} = o_p(1)$ and $\|\widehat{f}_\ell - f\|_{L_2(P_Z)} = o_p(1)$, and their product is $o_p(n^{-1/2})$.

(v) The variance $\tau_h^2 = \mathbb{P}[\phi(O)^2]$ is strictly positive.

All bounds in these conditions may depend on the fixed bandwidth. Since ρ_h is bounded, $\widehat{D} = \mathbb{P}_n[\rho_h(W)] = D + O_p(n^{-1/2})$, and τ_h^2 is finite by conditions (ii)–(iii). In particular, $\widehat{D} > 0$ with probability tending to one. On the event $\widehat{D} = 0$, define the estimator and its variance estimate arbitrarily as finite values; this event has probability tending to zero and does not affect the limits below.

G.2 Asymptotic linearization, central limit theorem, and efficiency

Let $J(O) = \mathbb{I}\{A = a', \bar{\mathbf{A}} = \bar{\mathbf{a}}\}$, $\omega_\ell = n_\ell/n$, and $\mathbb{P}_{n,\ell}$ denote the empirical mean over fold I_ℓ. Define the numerator scores, centred at the true target, by

$$S(O) = \rho_h(W)\{\mu(Z) - \psi\} + J(O) f(Z)\{Y - \mu(Z)\},$$
$$\widehat{S}_\ell(O) = \rho_h(W)\{\widehat{\mu}_\ell(Z) - \psi\} + J(O)\widehat{f}_\ell(Z)\{Y - \widehat{\mu}_\ell(Z)\}.$$

Then $\mathbb{P}S = 0$ and $\phi = S/D$. When applied to a fitted function, $\mathbb{P}$ integrates over a fresh observation while holding that function fixed.

Step 1 (foldwise decomposition). On the event $\widehat{D} > 0$, the common denominator in Eq. (7) gives the exact identity

$$\widehat{D}(\widehat{\psi} - \psi) = (\mathbb{P}_n - \mathbb{P})S + \sum_{\ell=1}^{L} \omega_\ell (\mathbb{P}_{n,\ell} - \mathbb{P})(\widehat{S}_\ell - S) + \sum_{\ell=1}^{L} \omega_\ell \mathbb{P}(\widehat{S}_\ell - S).$$

Step 2 (second-order bias). Write $\Delta_{\mu,\ell} = \widehat{\mu}_\ell - \mu$ and $\Delta_{f,\ell} = \widehat{f}_\ell - f$. The transport identity in Sec. D gives $\mathbb{P}[\rho_h \Delta_{\mu,\ell}] = \mathbb{P}[Jf\Delta_{\mu,\ell}]$. Also, $\mathbb{P}[J\Delta_{f,\ell}(Y-\mu)] = 0$ by the definition of μ. Thus

$$\mathbb{P}(\widehat{S}_\ell - S) = -\mathbb{P}[J\Delta_{f,\ell}\Delta_{\mu,\ell}],$$
$$|\mathbb{P}(\widehat{S}_\ell - S)| \le \|\Delta_{f,\ell}\|_{L_2(P_Z)} \|\Delta_{\mu,\ell}\|_{L_2(P_Z)} = o_p(n^{-1/2}).$$

For example, an $o_p(n^{-1/4})$ error for each nuisance estimate suffices. Unequal rates also suffice when both estimates are consistent and their error product is $o_p(n^{-1/2})$.

Step 3 (empirical process control via cross-fitting). The difference can be written as

$$\widehat{S}_\ell - S = (\rho_h - Jf)\Delta_{\mu,\ell} + J\Delta_{f,\ell}(Y - \widehat{\mu}_\ell).$$

Boundedness and condition (iv) imply $\|\widehat{S}_\ell - S\|_{L_2(P)} \le C_h(\|\Delta_{\mu,\ell}\|_{L_2(P_Z)} + \|\Delta_{f,\ell}\|_{L_2(P_Z)}) = o_p(1)$. Let $\mathcal{T}_\ell$ include the training observations outside I_ℓ, the independent fold assignment, and any independent fitting randomness. Conditional on $\mathcal{T}_\ell$, the evaluation observations are independent of $\widehat{S}_\ell$ and are i.i.d. from P. Consequently,

$$\mathbb{E}\left[\left\{\sqrt{n}(\mathbb{P}_{n,\ell} - \mathbb{P})(\widehat{S}_\ell - S)\right\}^2 \mid \mathcal{T}_\ell\right] \le \frac{n}{n_\ell}\|\widehat{S}_\ell - S\|^2_{L_2(P)} = o_p(1).$$

To make the unconditional implication explicit, let $X_{n\ell}$ be the centred fold term inside the square above, and $V_{n\ell} = (n/n_\ell)\|\widehat{S}_\ell - S\|^2_{L_2(P)}$. Conditional Chebyshev's inequality gives, for every $\epsilon, \delta > 0$,

$$\Pr(|X_{n\ell}| > \epsilon) \le \Pr(V_{n\ell} > \delta) + \delta/\epsilon^2.$$

Taking $n \to \infty$ and then $\delta \downarrow 0$ gives $X_{n\ell} = o_p(1)$. Since L is fixed, their weighted sum is also $o_p(1)$; independence between these foldwise terms is not needed.

Step 4 (central limit theorem and efficiency). Steps 1–3 and $\widehat{D} \to_p D > 0$ yield

$$\sqrt{n}(\widehat{\psi} - \psi) = \frac{1}{D\sqrt{n}}\sum_{i=1}^n S(O_i) + o_p(1)$$
$$= \frac{1}{\sqrt{n}}\sum_{i=1}^n \phi(O_i) + o_p(1) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \tau_h^2).$$

The last step uses the i.i.d. central limit theorem, since $\mathbb{P}\phi = 0$ and $0 < \tau_h^2 < \infty$. Replacing $\widehat{D}^{-1}$ by D^{-1} is valid because $n^{-1/2}\sum_i S(O_i) = O_p(1)$ and $\widehat{D}^{-1} - D^{-1} = o_p(1)$. By Prop. 4, this influence function is efficient for the outcome-regression functional in the unrestricted observed-data model specified in Sec. F. For any regular path P_ε with score s and fixed t, contiguity transfers the asymptotic-linear remainder to sampling from $P_{t/\sqrt{n}}$. Le Cam's third lemma shifts the mean of the linear term by $t\mathbb{E}_P[\phi s]$, while Sec. F gives $\sqrt{n}\{\Psi_h^{\mathrm{OR}}(P_{t/\sqrt{n}}) - \Psi_h^{\mathrm{OR}}(P)\} \to t\mathbb{E}_P[\phi s]$. After centring at this local statistical target, the limiting law remains $\mathcal{N}(0, \tau_h^2)$. This establishes regularity, so the estimator attains the efficiency bound in the specified model.

G.3 Variance and confidence interval

The variance in Thm. 5 is

$$\tau_h^2(w, a') = \frac{1}{\{\mathbb{E}[\rho_h(W; w)]\}^2}\Big[\mathbb{E}\big[\rho_h(W; w)^2\{\mu(Z; a', \bar{\mathbf{a}}) - \psi_h(a'; w)\}^2\big]$$
$$+ \mathbb{E}\big[\mathbb{I}\{A = a', \bar{\mathbf{A}} = \bar{\mathbf{a}}\}\, f(Z; w, a')^2\, \mu(Z; a', \bar{\mathbf{a}})\{1 - \mu(Z; a', \bar{\mathbf{a}})\}\big]\Big]. \quad (8)$$

The cross term is zero because $\rho_h J = 0$ when $a' \neq a$; the second term uses the Bernoulli conditional variance of Y. For $i \in I_\ell$ and $\widehat{D} > 0$, define the computable influence value

$$\widehat{\phi}_i = \frac{1}{\widehat{D}}\Big[\rho_h(W_i)\{\widehat{\mu}_\ell(Z_i) - \widehat{\psi}\} + J(O_i)\widehat{f}_\ell(Z_i)\{Y_i - \widehat{\mu}_\ell(Z_i)\}\Big],$$
$$\widehat{\tau}_h^2 = \frac{1}{n}\sum_{i=1}^n \widehat{\phi}_i^2.$$

To prove consistency, conditional Markov's inequality and Step 3 give $\mathbb{P}_{n,\ell}(\widehat{S}_\ell - S)^2 = o_p(1)$ for every fold, since its conditional expectation is $\|\widehat{S}_\ell - S\|^2_{L_2(P)} = o_p(1)$. The conditional-to-unconditional argument is the same as in Step 3, with Markov's bound δ/ϵ in place of δ/ϵ^2. Replacing ψ by $\widehat{\psi}$ changes $\widehat{S}_\ell$ by $-\rho_h(\widehat{\psi} - \psi)$. Its empirical squared norm is $(\widehat{\psi} - \psi)^2\mathbb{P}_n[\rho_h^2] = o_p(1)$. Let $\widetilde{S}_i$ denote this computable numerator, with $\widehat{\psi}$ in place of ψ. Then

$$\widehat{\phi}_i - \phi(O_i) = \frac{\widetilde{S}_i - S(O_i)}{\widehat{D}} + S(O_i)(\widehat{D}^{-1} - D^{-1}).$$

Since $\mathbb{P}_n[S^2] = O_p(1)$ and $\widehat{D} \to_p D > 0$, these facts imply

$$\frac{1}{n}\sum_{i=1}^{n}\{\widehat{\phi}_i - \phi(O_i)\}^2 = o_p(1).$$

Write the empirical squared difference above as $r_n = o_p(1)$. Cauchy–Schwarz gives $|\widehat{\tau}_h^2 - \mathbb{P}_n[\phi^2]| \le r_n + 2\sqrt{r_n\mathbb{P}_n[\phi^2]} = o_p(1)$. The law of large numbers $\mathbb{P}_n[\phi^2] \to_p \tau_h^2$ therefore yields $\widehat{\tau}_h^2 \to_p \tau_h^2$. Thus $\widehat{\psi}_h(a'; w) \pm 1.96\,\widehat{\tau}_h(w, a')/\sqrt{n}$ is an asymptotic 95% confidence interval for the fixed-bandwidth local target $\psi_h(a'; w)$.

G.4 From the kernel localised target to the pointwise target

Let $\psi^\star(v) \coloneqq \mathbb{E}\big[Y_{a'} \mid Y = y,\, A = a,\, \bar{\mathbf{A}} = \bar{\mathbf{a}},\, V = v\big]$ be the pointwise target. Let $F = \{Y = y, A = a, \bar{\mathbf{A}} = \bar{\mathbf{a}}\}$ have positive probability, and let p_F be the density of V given F. Assume that $p_F(v) > 0$, and that p_F and $p_F\psi^\star$ are twice continuously differentiable with bounded second derivatives. Assume also that K is symmetric and nonnegative, with unit integral and finite second moment. The local target can then be written as

$$\psi_h = \frac{\int K(t)\,\psi^\star(v + ht)\,p_F(v + ht)\,dt}{\int K(t)\,p_F(v + ht)\,dt}.$$

For either $H = p_F$ or $H = p_F\psi^\star$, bounded second derivatives give

$$|H(v + ht) - H(v) - h\nabla H(v)^\top t| \le \tfrac{1}{2}M_H h^2\|t\|^2$$

for a finite constant M_H. Symmetry gives $\int tK(t)\,dt = 0$, and the finite second moment controls the integrated remainder. Thus the numerator is $\psi^\star(v)p_F(v) + O(h^2)$ and the denominator is $p_F(v) + O(h^2)$. For all sufficiently small h, the denominator is at least $p_F(v)/2 > 0$; subtracting $\psi^\star(v)$ times the denominator from the numerator and dividing gives $\psi_h - \psi^\star(v) = O(h^2)$. This approximation concerns the target as $h \to 0$; Thm. 5 establishes inference only for fixed h and does not give a confidence interval for $\psi^\star(v)$ along a shrinking bandwidth sequence.

H TongSim-CF Details

H.1 Data generation

We build TongSim-CF in TongSim (Sun et al., 2025; Peng et al., 2024), a simulator that renders household scenes and executes scripted interaction events. For each scenario variant, we randomize the initial scene and sample the exogenous noise of the structural equations. Each episode gives one tuple $\big(Y, A, \bar{\mathbf{A}}, V\big)$, where V is the pre-action frame. All variants share the factorisation

$$\begin{aligned}
&S^d = g_d(U_d)\,, \quad S^c = g_c(U_c)\,, \quad V = g_V\big(S^d, S^c, U_V\big)\,,\\
&\bar{\mathbf{A}} = g_{\bar{\mathbf{A}}}\big(S^d, S^c, U_{\bar{\mathbf{A}}}\big)\,, \quad A = g_A\big(\bar{\mathbf{A}}, S^d, S^c, U_A\big)\,,\\
&Y = g_Y\big(A, \bar{\mathbf{A}}, S^d, S^c, U_{Y,(A,\bar{\mathbf{A}})}\big)\,,
\end{aligned}$$

where S^d is a discrete precondition such as whether power is available, S^c is a continuous scene factor such as layout or camera pose, and all U terms are exogenous noise. The outcome noises for different action configurations are independent of one another and of the action noise. In these equations, Setting 2 adds another action $\bar{\mathbf{A}}$ that is correlated with A and affects Y, and Setting 3 lets S^d, and in some variants S^c, affect both A and Y. Some Setting 3 variants also contain a visual cue, such as an indicator light, that is correlated with the precondition but has no causal effect. Across settings, some variants mask the association of an effective action, and others create an association for an ineffective action. Tab. 4 lists all scenarios.

Table 4: TongSim-CF scenarios. Each scenario has one or more variants that change the mechanism.

Setting	ID	Scenario (actions → outcome)	Variants
1	S-A01	Switch on → table lamp lit	Default
	S-A02	Open → cabinet drawer open	Default
	S-A03	Pick up → cup held	Default
2	S-B01	Left and right switches → ceiling lamp on	Either switch works; only the left switch works
	S-B02	TV remote and front-panel button → TV on	Either control works; remote is ineffective
	S-B03	Place cup and press button → coffee dispensed	Button controls flow; cup is required
	S-B04	Room light switch and PC power button → monitor on	PC power required; both actions required
	S-B05	Power indicator and fan switch → fan spinning	Both actions required; fan switch alone suffices
	S-B06	Unlock switch and open door → door open	Both actions required; opening alone suffices
	S-B07	Unlock switch and open window → window open	Both actions required; unlock/open chain is masked
	S-B08	Close door and start microwave → microwave on	Both actions required; door/start chain is masked
	S-B09	Open and pull up blinds → blinds collapsed	Both actions required; one action suffices
3	S-C01	Toggle lamp switch → lamp lit	Power-conditioned effect; masked continuous confounding
	S-C02	Flick switch and open door → door open	Indicator-gated joint effect; spurious action association
	S-C03	Flick switch and open window → window open	Indicator-gated joint effect; spurious action association
	S-C04	Wake monitor → monitor on	LED-conditioned effect; masked continuous confounding
	S-C05	Flick power switch → device on	Power-indicator conditioning; masked association

H.2 GROUND TRUTH

For a test video with evidence $(Y = y, A = a, \bar{\mathbf{A}} = \bar{\mathbf{a}}, V = v)$, we hold the scene state and the other actions fixed and set $A \leftarrow a'$ in the outcome equation. The ground truth $\psi^{\mathrm{gt}}(a'; y, a, \bar{\mathbf{a}}, v)$ is the resulting success probability. Because the outcome noises are independent, the observed outcome carries no further information once the scene state and other actions are given. Thus, Ass. 4 holds by construction, and the success probability equals the target in Eq. (1). This ground truth is pointwise, so the reported error includes the difference between the local and pointwise targets. The copula study in Sec. J.6 changes the dependence between potential outcomes and recomputes the ground truth accordingly.

I IMPLEMENTATION DETAILS

I.1 OUR ESTIMATOR

Tab. 5 lists the settings of our estimator on both datasets. On TongSim-CF, the discrete attributes Z^{d} are binary preconditions, such as whether power is available or a lock is engaged. We read them from the simulator state. Gemini-3-Flash can also extract them from the first frame. The continuous features are $Z^{\mathrm{c}} = (\alpha \widetilde{z}(V))_{1:d_s}$, where $\widetilde{z}(V) = (z(V) - \mu_{\mathrm{pca}})^{\top} W_{\mathrm{pca}} \oslash \sigma_{\mathrm{pca}}$ standardises the PCA projection of the VA-VAE latent mean $z(V)$. The nuisance models are logistic regressions on action indicators, $(Z^{\mathrm{d}}, Z^{\mathrm{c}})$, and their pairwise interactions. For scene matching, we use a representation-space kernel with exact agreement on Z^{d}. We form a proxy for r by averaging the kernel over Gaussian perturbations of each continuous feature vector. These feature samples do not directly sample $V \mid Z$, so consistency of this proxy for $\mathbb{E}[K_h(V - v) \mid Z]$ is not established here.

Table 5: Settings of our estimator.

	TongSim-CF	Physics 101
Visual encoder	VA-VAE (Yao et al., 2025), first frame	CLIP ViT-B/32 (Radford et al., 2021), object photograph
Discrete attributes Z^{d}	Simulator preconditions	None
Continuous features Z^{c}	PCA, 10 components kept, first 2 to 4 used, scale $\alpha = 0.2$	PCA, dimension 16 or 24
Bandwidth h	0.3	0.2
Selection of d_s and h	Fixed	Nested five-fold cross-validation over objects
Nuisance models	Logistic regression with quadratic features, L_2 penalty, L-BFGS (Liu & Nocedal, 1989), 500 iterations	Same
Matching term r	128 Monte Carlo samples	Same
Post-processing	None	Clipping to $[0, 1]$

Alg. 1 gives the cross-fitted estimator analysed in Thm. 5. In the experiments, we normalise the two terms of Eq. (7) separately and clip the estimates on Physics 101 to $[0, 1]$. The efficiency and inference guarantees concern Eq. (7) under its stated conditions, and do not cover these implementation approximations.

Algorithm 1 Efficient estimation of $\psi_h(a'; y, a, \bar{\mathbf{a}}, v)$.

1: **Input:** data $\{(y_i, a_i, \bar{\mathbf{a}}_i, v_i, z_i)\}_{i=1}^{n}$, evidence $w = (y, a, \bar{\mathbf{a}}, v)$, counterfactual action a', kernel K_h, number of folds L.
2: Split $\{1, \ldots, n\}$ into L disjoint folds $\{I_\ell\}_{\ell=1}^{L}$.
3: **for** $\ell = 1$ to L **do**
4: Fit $\widehat{\mu}_\ell$, $\widehat{\pi}_\ell$, $\widehat{r}_\ell$ on $\{i : i \notin I_\ell\}$, and form $\widehat{\eta}_\ell$ and $\widehat{f}_\ell$.
5: **end for**
6: Compute $\widehat{\psi}_h(a'; w)$ via Eq. (7).
7: **Return:** $\widehat{\psi}_h(a'; w)$.

To measure how broadly the kernel spreads, we use the effective sample size

$$\mathrm{ESS}(v) := \frac{\left(\sum_{j=1}^{n} K_h(V_j - v)\right)^2}{\sum_{j=1}^{n} K_h(V_j - v)^2}.$$

It equals m when the kernel weight is spread evenly over m videos. It ignores the action, outcome, and attribute filters in ρ_h.

I.2 Baselines on TongSim-CF

The VLMs are evaluated on the random subset of 53 tasks in Sec. 5.1, and the other methods on all tasks. On the subset, our estimator has MAE 0.023, 0.038, 0.073 and 0.048 in Settings 1, 2, 3 and overall. Tab. 6 summarises the inputs and training of each baseline.

Table 6: Baselines on TongSim-CF. *Observed* labels are the recorded binary outcomes. The two-fold methods predict each test video with a model fitted on the other fold.

Method	Input	Training
Fire-Zhu (Fire & Zhu, 2016)	Actions	Action and outcome co-occurrence statistics
LLCP (Chen et al., 2024a)	CLIP features of the first frame, actions, outcome	Conditional VAE on frame pairs and observed labels, full pool
Prompted VLMs	First frame, actions, and outcome of the test video and of five nearest videos	None
VLMs with retrieval	Records of all other videos, five images per action value, test frame	None
Qwen2.5-VL-7B (Bai et al., 2025)	First frame, action text	LoRA on observed labels, two folds
V-JEPA 2 (Assran et al., 2025)	Frozen features of the first frame, actions	Predictor head on observed labels, two folds
Counterfactually supervised	CLIP features, actions	Counterfactual probabilities from the simulator, two folds
Ours	Z^{d}, Z^{c}, actions, outcome	Observed labels, full pool

LLCP. We adapt LLCP to scene features, since the original model is object-level. It is a conditional variational autoencoder trained on CLIP features of frame pairs. At test time, it weights latent samples by the likelihood of the observed outcome and predicts the outcome under the alternative action.

VLMs. We query `openai/gpt-5.5` and `google/gemini-3-flash-preview` through OpenRouter. The prompted VLMs receive the test video and its five nearest videos, each with its first frame, actions, and outcome. The retrieval variant receives the records of all other videos and the five most similar frames for each value of the target action, with the template below. We average three replies per task, and Gemini-3-Flash gives no valid reply for one task.

```
Estimate a counterfactual probability for the query episode
in a simulated home.
Action indicators:  1 = performed, 0 = not performed.
Y: 1 = the target state change occurred, 0 = it did not.
Reference episodes:
[reference records on both action arms, with their reference frames]
=== QUERY EPISODE (frame below) ===
QUERY: [observed action indicators] Y=[observed outcome]
[query frame]
For this same query episode, what is the probability that Y
would have been 1 if [target action] had instead been [alternative
value], with [other actions fixed at their observed values / no other actions]?
Return only one number in [0,1].
```

Finetuned foundation models. We finetune Qwen2.5-VL-7B with LoRA and calibrate its Yes/No logit on a held-out part of each training fold. For V-JEPA 2, we encode the repeated first frame with the frozen encoder and train a predictor head. At test time, both models receive the alternative action with the image and other actions unchanged.

Counterfactually supervised model. This model is trained with the outcome probability under the alternative action from the simulator as a soft label. It uses CLIP features of the first frame and all action values, and selects its regressor by inner cross-validation.

Example in Fig. 1. We give GPT-5.5 29 frames of the clip and ask which action caused the door to open. The figure condenses its answer, and this example is not part of the quantitative evaluation.

I.3 PHYSICS 101

Data and protocol. The ramp subset of Physics 101 (Wu et al., 2016) contains 1801 trials, with 3 to 17 trials per object and angle. The action is $A = \mathbb{I}\{20 \text{ degrees}\}$, and there are no other actions. Treating the trials at the other angle as draws of the counterfactual outcome relies on Ass. 4. We hold out each object in turn and fit all methods on the other 100 objects. Our estimator selects the feature dimension and bandwidth by nested cross-validation over objects. The baselines follow their TongSim-CF protocols, with the object photograph as the scene. The VLMs with retrieval receive the sliding counts of all other objects, with the template below.

```
Estimate a counterfactual sliding probability.
Reference records give other objects' sliding counts at ramp
angles of 10 and 20 degrees.  Photographs show the objects
before testing.
Reference objects:
[counts for the other 100 objects, with photographs of the first five references]
Target trial:  the object below was placed at [observed angle]
degrees and [slid / did not slide].
[target photograph]
For this same trial, what is the probability that the object
would have slid if the ramp angle had instead been [alternative
angle] degrees?
Return only one number in [0,1].
```

J ADDITIONAL RESULTS

Unless stated otherwise, the results in this section use the subset of 53 tasks in Tab. 1, on which our estimator has MAE 0.048.

J.1 ESTIMATOR COMPONENTS AND NUISANCE MISSPECIFICATION

We compare three estimators of the local target in Eq. (3). OR plugs the fitted outcome regression into Eq. (4). IPW plugs the fitted weights into Eq. (5). Our estimator combines the two as in Eq. (7). To test the benefit of combining them, we perturb one nuisance function at a time. To misspecify the outcome model $\widehat{\mu}$, we flip 30% of the binary features, add Gaussian noise at 0.5 standard deviations to the continuous features, and drop 50% of the features before fitting. To misspecify the propensity model $\widehat{\pi}$, we remove Z^{c} and its interactions from its features. Tab. 7 shows that, without misspecification, the three estimators have similar error. The benefit of combining them appears under misspecification. OR fails when the outcome model is misspecified, with MAE 0.239, because it uses the corrupted predictions directly. Our estimator corrects these predictions with weighted residuals and keeps its MAE at 0.058. Removing the continuous features from the propensity model changes neither IPW nor our estimator much.

Table 7: MAE of the estimator components under default fitting and nuisance misspecification (mean $\pm$ SEM across variants).

Estimator variant	Setting 1	Setting 2	Setting 3	Overall
OR	$0.015_{\pm 0.007}$	$0.044_{\pm 0.006}$	$0.062_{\pm 0.015}$	$0.047_{\pm 0.006}$
IPW	$0.023_{\pm 0.005}$	$0.036_{\pm 0.005}$	$0.067_{\pm 0.020}$	$0.045_{\pm 0.007}$
Ours	$0.023_{\pm 0.005}$	$0.038_{\pm 0.005}$	$0.073_{\pm 0.023}$	$0.048_{\pm 0.008}$
OR, misspecified μ	$0.168_{\pm 0.025}$	$0.234_{\pm 0.031}$	$0.270_{\pm 0.059}$	$0.239_{\pm 0.026}$
IPW, misspecified π	$0.021_{\pm 0.006}$	$0.036_{\pm 0.006}$	$0.057_{\pm 0.013}$	$0.041_{\pm 0.006}$
Ours, misspecified μ	$0.015_{\pm 0.008}$	$0.061_{\pm 0.007}$	$0.064_{\pm 0.021}$	$0.058_{\pm 0.008}$
Ours, misspecified π	$0.021_{\pm 0.006}$	$0.038_{\pm 0.006}$	$0.066_{\pm 0.021}$	$0.045_{\pm 0.008}$

J.2 Scene representation

Tab. 8 varies the scene representation used by the nuisance models and by the kernel. In every row, the kernel uses the continuous features Z^{c}, so the continuous visual information enters every configuration. Settings 1 and 2 have no discrete attributes. There, the rows differ only in whether the nuisance models also take Z^{c}, and their MAE differs by at most 0.002. In Setting 3, the scene confounds the action, and the discrete attributes are essential. Without them, the kernel mixes scenes with different mechanisms, and MAE on the subset rises from 0.073 to 0.177. On these tasks, nuisance models with Z^{d} alone have lower MAE in Setting 3 than with both features. This gap of 0.016 is within the noise of one test video per variant, and more than half of it comes from a single task. On all tasks, the two features together give the lowest MAE in Setting 3 and overall. With the continuous features alone, our overall MAE is 0.081 on the subset and 0.080 on all tasks, still below that of every baseline in Tab. 1.

Table 8: MAE of our estimator with different scene representations (mean $\pm$ SEM across variants). The nuisance models μ, η and π take the listed features. The kernel always uses Z^{c} and matches Z^{d} exactly when the nuisance models use it.

Tasks	Nuisance features	Setting 1	Setting 2	Setting 3	Overall
Subset	Z^{d}	$0.021_{\pm 0.006}$	$0.036_{\pm 0.006}$	$0.057_{\pm 0.014}$	$0.041_{\pm 0.006}$
	Z^{c}	$0.023_{\pm 0.005}$	$0.038_{\pm 0.005}$	$0.177_{\pm 0.055}$	$0.081_{\pm 0.021}$
	$Z^{\mathrm{d}}+Z^{\mathrm{c}}$	$0.023_{\pm 0.005}$	$0.038_{\pm 0.005}$	$0.073_{\pm 0.023}$	$0.048_{\pm 0.008}$
All tasks	Z^{d}	$0.018_{\pm 0.002}$	$0.040_{\pm 0.005}$	$0.070_{\pm 0.010}$	$0.047_{\pm 0.005}$
	Z^{c}	$0.019_{\pm 0.002}$	$0.041_{\pm 0.005}$	$0.169_{\pm 0.036}$	$0.080_{\pm 0.016}$
	$Z^{\mathrm{d}}+Z^{\mathrm{c}}$	$0.019_{\pm 0.002}$	$0.041_{\pm 0.005}$	$\mathbf{0.061_{\pm 0.009}}$	$\mathbf{0.045_{\pm 0.005}}$

We also add Gaussian noise to the representation. MAE stays near 0.048 down to a signal to noise ratio of 2 and rises to 0.106 at 0.5 (Tab. 9).

Table 9: Overall MAE of our estimator under representation noise at different signal to noise ratios (SNR).

SNR	∞	4	2	1	0.5
Overall MAE	0.048	0.048	0.046	0.057	0.106

To see how much visual information the compressed features keep, we decode them with the fixed VA-VAE decoder. We use 155 held-out frames, five per variant, and invert the PCA with the discarded coordinates set to zero. Tab. 10 shows that reconstruction quality drops as fewer coordinates are kept. These metrics measure pixel retention, not the information relevant to the actions. Fig. 3 shows that nearby features correspond to similar layouts and object configurations.

Table 10: Reconstruction from the full VA-VAE encoding and from compressed coordinates on 155 frames. Pixel values lie in $[0, 1]$.

Visual representation	Dimension	Pixel MSE	SSIM
Full VA-VAE encoding	8192	0.000233	0.950
Retained PCA coordinates	10	0.016999	0.684
Continuous estimator input	2–4	0.024754	0.651

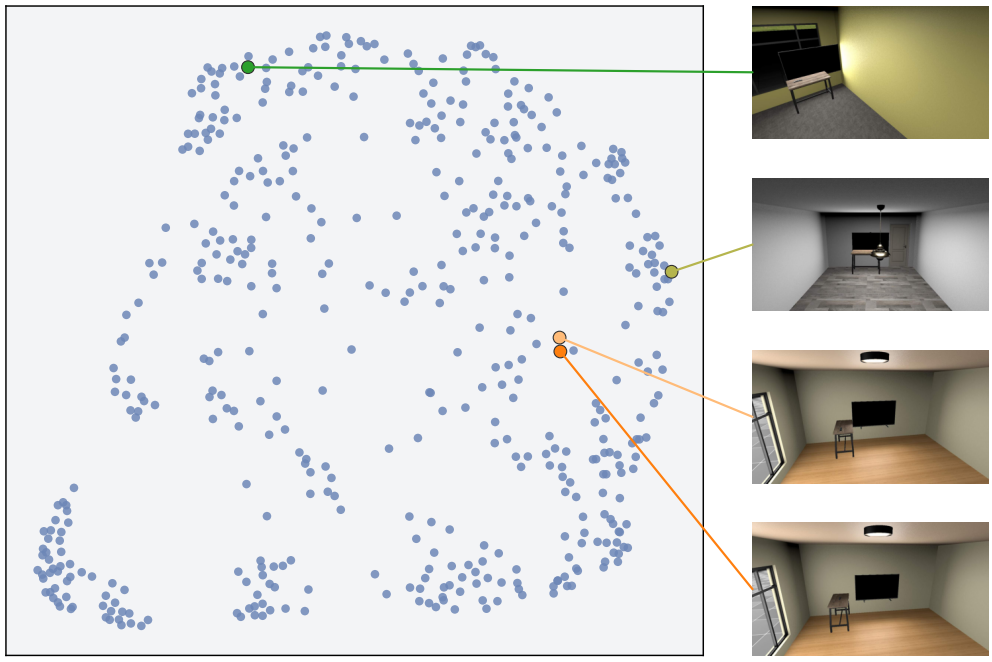

Figure 3: UMAP (McInnes et al., 2018) of the continuous features Z^c with the corresponding first frames.

J.3 BANDWIDTH, NUMBER OF OBSERVATIONS, AND WEIGHT TRUNCATION

Tab. 11 varies the bandwidth. The bandwidth trades the number of matched videos against their similarity to the observed scene. A small bandwidth, $h = 0.1$, keeps fewer videos (ESS 221) and raises the error in every setting. A larger bandwidth raises ESS and slightly lowers error in Setting 1. In Setting 3, it also averages over scenes with different confounders. There, raising h from 0.3 to 0.9 raises ESS from 536 to 590 but also raises MAE from 0.073 to 0.084. The default $h = 0.3$ balances the two effects and gives the lowest error in Setting 3 and overall, in line with Sec. G.4.

Table 11: MAE and effective sample size of our estimator for different bandwidths (mean ± SEM across variants).

	MAE				ESS			
h	S1	S2	S3	Overall	S1	S2	S3	Overall
0.1	$0.029_{\pm 0.008}$	$0.055_{\pm 0.010}$	$0.076_{\pm 0.015}$	$0.059_{\pm 0.008}$	$249_{\pm 43}$	$210_{\pm 26}$	$232_{\pm 41}$	$221_{\pm 20}$
0.3	$0.023_{\pm 0.005}$	$\mathbf{0.038_{\pm 0.005}}$	$\mathbf{0.073_{\pm 0.023}}$	$\mathbf{0.048_{\pm 0.008}}$	$535_{\pm 6}$	$526_{\pm 14}$	$536_{\pm 16}$	$530_{\pm 9}$
0.9	$\mathbf{0.020_{\pm 0.005}}$	$\mathbf{0.038_{\pm 0.005}}$	$0.084_{\pm 0.027}$	$0.051_{\pm 0.010}$	$596_{\pm 2}$	$596_{\pm 1}$	$590_{\pm 6}$	$594_{\pm 2}$
2.0	$\mathbf{0.020_{\pm 0.005}}$	$\mathbf{0.038_{\pm 0.005}}$	$0.085_{\pm 0.027}$	$0.052_{\pm 0.010}$	$598_{\pm 2}$	$597_{\pm 1}$	$592_{\pm 6}$	$596_{\pm 2}$

Tab. 12 keeps the 31 test videos of the subset and subsamples the other videos of each variant to $n \in \{50, 100, 200, 350\}$, with five nested subsets per size. The error of our estimator falls from 0.123 at $n = 50$ to 0.048 with all videos, whereas Fire-Zhu stays near 0.2. More data improves the adjusted estimate but cannot remove the confounding bias of an association.

Tab. 13 keeps only the k largest matching weights. Global truncation selects k neighbours by scene similarity before applying the action and outcome conditions. Truncation within strata selects up to k neighbours after applying the conditions of each term, so it keeps up to $2k$ videos. Truncation within strata stays close to the full result for $k \geq 100$, whereas global truncation loses accuracy for $k \leq 200$. Many global neighbours fail the later conditions and thus contribute nothing.

Table 12: MAE with n videos per variant (mean ± SEM across variants).

n	Ours	Fire-Zhu
50	$0.123_{\pm 0.011}$	$0.216_{\pm 0.024}$
100	$0.089_{\pm 0.010}$	$0.202_{\pm 0.024}$
200	$0.070_{\pm 0.010}$	$0.197_{\pm 0.024}$
350	$0.058_{\pm 0.009}$	$0.194_{\pm 0.025}$
All	$0.048_{\pm 0.008}$	$0.191_{\pm 0.025}$

Table 13: MAE when only the k largest matching weights are kept (mean ± SEM across variants).

k	Global	Within strata
25	$0.153_{\pm 0.023}$	$0.063_{\pm 0.007}$
50	$0.079_{\pm 0.012}$	$0.054_{\pm 0.007}$
100	$0.084_{\pm 0.011}$	$0.047_{\pm 0.008}$
200	$0.063_{\pm 0.009}$	$0.049_{\pm 0.008}$
500	$0.050_{\pm 0.008}$	$0.048_{\pm 0.008}$

J.4 COMPUTATIONAL COST

Tab. 14 profiles one variant with 500 videos and 1000 queries on a single RTX 4090. Visual encoding takes most of the time. The encoded frames and fitted nuisance models are shared across queries. With cached nuisance predictions and weights, each query needs $O(n)$ operations.

Table 14: Wall-clock time for one variant with 500 videos and 1000 queries on a single RTX 4090.

Stage	Component	Time
Representation	VA-VAE encoding (500 frames, batch 16)	7.30 s
Representation	PCA projection	0.01 s
Nuisance	Outcome model $\widehat{\mu}$ (logistic)	0.21 s
Nuisance	Propensity model $\widehat{\pi}$	0.09 s
Estimation	Monte Carlo density ratio (128 samples)	0.10 s
Estimation	Kernel matching and correction	0.73 s
Total		8.44 s

J.5 FOUNDATION MODELS AS OUTCOME PREDICTORS

We use the finetuned Qwen and V-JEPA 2 predictors as the outcome model $\widehat{\mu}$ of our estimator. We compare three uses of the same fitted predictor. *Direct prediction* evaluates it under the alternative action. The *local plug-in* averages its predictions over videos that match the observed evidence. The *complete estimator* adds the weighted residual correction with $h = 0.3$. Each test video is predicted by the model fitted on the other fold. Tab. 15 shows that local averaging lowers the error only modestly, whereas the correction lowers it below 0.05 for both predictors.

Table 15: MAE with the same fitted outcome predictor used directly, in the local plug-in, or in the complete estimator (mean ± SEM across variants).

Outcome predictor	Direct prediction	Local plug-in	Complete estimator
Qwen2.5-VL-7B	$0.169_{\pm 0.025}$	$0.163_{\pm 0.024}$	$\mathbf{0.046_{\pm 0.007}}$
V-JEPA 2	$0.149_{\pm 0.025}$	$0.118_{\pm 0.019}$	$\mathbf{0.049_{\pm 0.007}}$

J.6 SENSITIVITY TO IDENTIFICATION ASSUMPTIONS

This subsection gives the protocols of Sec. 5.4.

Unmeasured confounding. We add an unobserved binary variable U to every variant. It adds γU to the logits of the actions and the outcome, with $\gamma \in \{0, 0.5, 1, 1.5, 2\}$ and three data seeds per strength. We keep the 31 test videos of the subset and regenerate their actions and outcomes. The ground truth averages the alternative-action outcome probability over the posterior of U. Our estimator finds no matching video for 41 of the 795 test predictions, so all methods are scored on the other 754. As γ grows from 0 to 2, the gap between the ground truths for observed outcomes 1 and 0 grows from 0 to 0.122. This growth explains the improvement of LLCP: with uniform weights instead of outcome reweighting, its error at $\gamma = 2$ rises.

Dependence between potential outcomes. We couple Y_a and $Y_{a'}$ with a Gaussian copula with correlation $\rho \in \{-1, -0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75, 0.9, 1\}$. The copula keeps the outcome probability under each action, so we reuse the videos and predictions of the $\gamma = 0$ study and change only the ground truth. The oracle predicts the outcome probability under the alternative action. All methods are scored on the 152 of 159 test predictions that our estimator can form.