CLAM: CAUSAL SPATIAL DISAGGREGATION TO INFER LOCAL EFFECTS FROM COARSE DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning fine-grained spatial patterns from coarse-resolution data is challenging, especially in causal settings where high-resolution effects must be inferred from aggregated interventions and outcomes. We introduce CLAM, a method for estimating localized causal effects from coarse observations by exploiting high-resolution contextual covariates that modulate these effects. By jointly learning the causal mechanism and a disaggregation mapping, CLAM captures interactions that are missed when addressing these problems independently. The method supports localized effect estimation, principled outcome disaggregation, and reliably captures spatially varying causal effects across diverse settings. This is particularly relevant for applications such as public health and environmental policy, where decisions are made at broad scales despite substantial local heterogeneity. Code is available at `github.com/dsanonym/clam`.

1 INTRODUCTION

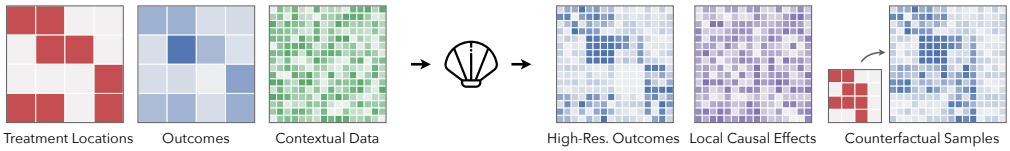


Figure 1: **Motivating example.** CLAM combines regional treatments (e.g., a vaccination awareness campaign) and outcomes (e.g., vaccination levels) with subregional covariates (e.g., demographics) to learn a causal mechanism that can then be used to infer a high-resolution outcome map, estimate local causal effects, and sample counterfactual outcomes.

Going from high-resolution (HR) to low-resolution (LR) data is typically straightforward, as it can be done via aggregation or pooling functions such as sums or means. These operations remove information. In contrast, going from LR to HR data is fundamentally difficult because it requires restoring information that has been lost. This generally necessitates some form of prior knowledge about the HR data. This inverse problem is typically studied under terms such as *spatial disaggregation* (Patil et al., 2024), *downscaling* (Sun et al., 2024), *ecological inference* (Schuessler, 1999), or *super-resolution* (Anwar et al., 2020). It is relevant across many fields including earth science (Kumar et al., 2023), public health (Matisziw et al., 2008), and the social sciences (Ye et al., 2016).

Causality (Pearl, 2009), by contrast, studies the effects of interventions on a system. Its core insight is that statistical associations do not imply causal relationships. In other words, observational patterns may be useful for prediction, but they do not, in general, tell us how outcomes will change under interventions.

Answering causal questions is often attempted using LR data, but this data is frequently insufficient for this (Beckers & Halpern, 2019). For example, interventions, like vaccination campaigns, might be applied across broad regions, yet their impact could vary significantly at the subregional level. Surprisingly little effort has been made to properly combine spatial disaggregation with the quest to answer causal queries.

Our hypothesis is that causal inference and spatial disaggregation are better solved together than apart, and that doing so makes questions answerable that neither can address alone.

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.

Contribution. This work addresses the gap between spatial disaggregation and causal inference by introducing a unified framework grounded in structural causal models (SCMs) and the do-calculus proposed by Pearl (1995). We argue that jointly tackling these tasks is natural in many real-world settings, where interventions are implemented at coarse scales but effects manifest locally. While our primary focus is the formulation of this framework, we illustrate its capabilities and limitations through a series of simulation studies that mimic realistic data-generating processes, as well as a case study based on real-world data on the impact of heat waves on gun violence.

CLAM (*C*aus*a*l dis*a*ggregation *m*ethod) can be summarized as follows: We learn a mechanistic function that maps an intervention indicator and contextual information to high-resolution outcomes. In doing so, we treat HR contextual data (or *covariates*) as an auxiliary variable that implicitly defines a prior over the HR outcome (cf. Figure 1). The model is trained by aggregating the predicted HR outcomes and enforcing consistency with the observed low-resolution measurements. The learned mechanism then supports downstream tasks such as inferring high-resolution outcome maps and estimating local causal effects.

We introduce causal spatial disaggregation as a distinct problem class, with the problem formulation as our primary contribution and the algorithm as a concrete instantiation. We make the target estimand, assumptions, and limitations explicit, and investigate what can be recovered from aggregated observations.

2 PROBLEM SETUP

For notational simplicity, we assume that all regions share the same subregion structure, interventions are binary, and all variables are scalar rather than vector-valued.

We consider N regions, each divided into M subregions. The LR *treatment data* is a vector $\widehat{T} \in \{0,1\}^N$, where each entry $\widehat{t}_i$ indicates whether an intervention occurred in region i. We assume that, when present, the intervention affects all subregions within that region. The LR *outcome data* is a vector $\widehat{Y} \in \mathbb{R}^N$, obtained by applying a row-wise aggregation function (e.g., summation or averaging) to an unobserved HR outcome matrix $Y \in \mathbb{R}^{N \times M}$, where $y_{i,j}$ denotes the outcome value for subregion j within region i. The HR *contextual data* (covariates) is given by a matrix $C \in \mathbb{R}^{N \times M}$, where $c_{i,j}$ denotes the contextual value for subregion j in region i.

Finally, we assume that N and M are perfect squares, allowing us to represent the spatial structure as a $\sqrt{N} \times \sqrt{N}$ grid of regions, each containing a $\sqrt{M} \times \sqrt{M}$ grid of subregions (see Figure 1 for an example with $N = M = 16$ and $\sqrt{N} = \sqrt{M} = 4$).

Goal. Our goal is to learn a function that estimates local causal effects and to disaggregate coarse outcomes into fine-grained predictions. We formalize this objective in the next section.

Generalizations. The setup admits several natural extensions. Variables can be vector-valued, treatments can be continuous instead of binary, and the observed LR treatment vector can be interpreted as an aggregation of an underlying HR treatment matrix that varies within a region. The framework also extends to spatio-temporal settings, where T, Y, and C evolve over time. More generally, additional structure such as hidden confounding, mediation, or non-trivial treatment assignment can be incorporated (see also Section C). Naturally, the spatial layout need not follow a regular grid, which we adopt only for illustration.

Assumptions. Causal spatial disaggregation relies on both causal and recovery assumptions. On the causal side, we assume that the causal graph is known or specified from domain knowledge and that the target effect is identifiable under this graph, for example by ruling out unmodeled confounding where required. On the recovery side, inferring fine-grained structure from aggregated observations requires a causal mechanism that is shared across subregions, sufficient variation in HR context across regions, and sufficient overlap between treated and untreated observations over the relevant context values. The aggregation from HR to LR must also be known or sufficiently constrained. We discuss these assumptions and their implications in more detail in Section 3.4 and Section 5.

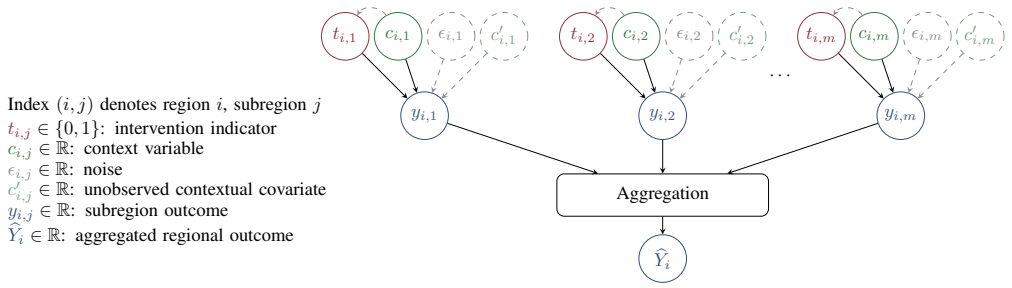


Figure 2: Causal graph for one region. Colors follow the scheme used throughout the paper: treatment in red, context in green, and outcomes in blue. All functional relations are shared among subregions and regions. Variables could also be vector-valued. Unobserved covariates and a causal link from context to treatment are optional. The aggregation is typically a mean or a sum.

3 Our Method: CLAM

Our method is based on two insights. First, structural causal models provide a natural way to formalize causal disaggregation by making the assumptions and unknown components explicit. Second, these unknown components can be learned from LR observations using an aggregate-learning paradigm, where predicted HR outcomes are combined and matched to the observed aggregates.

3.1 Structural Causal Model

Each region–subregion pair (i, j) is associated with a treatment $t_{i,j}$, context $c_{i,j}$, noise $\epsilon_{i,j}$, and an optional hidden confounder (or non-confounding covariate) $c'_{i,j}$ (cf. Figure 2). A shared structural function maps these inputs to a subregional outcome $y_{i,j}$, and then aggregates the subregional outcomes to a regional outcome $\widehat{Y}_i$:

$$y_{i,j} = f_{\boldsymbol{\theta}}(t_{i,j},\, c_{i,j},\, \epsilon_{i,j},\, c'_{i,j}), \quad \widehat{Y}_i = g_{\boldsymbol{\phi}}\big(\{y_{i,j}\}_{j=1}^{M}\big). \tag{1}$$

The shared mechanism $f_{\boldsymbol{\theta}}(\cdot)$ is evaluated separately for each subregion, with uncertainty captured through $\epsilon_{i,j}$, while $g_{\boldsymbol{\phi}}(\cdot)$ aggregates the resulting subregional outcomes in a permutation-invariant way. Although both functions are shared across regions, region-specific behavior can still be modeled through contextual variables, such as region indicators or positional encodings. The difference between $c'_{i,j}$ and $\epsilon_{i,j}$ is that $\epsilon_{i,j}$ represents i.i.d. noise, while $c'_{i,j}$ may have structure, for example spatial autocorrelation or persistence across time.

For simplicity, we typically assume sum aggregation, additive noise, and no unobserved confounding. The model then simplifies to:

$$\widehat{Y}_i = \sum_{j=1}^{M} \big(f_{\boldsymbol{\theta}}(t_{i,j}, c_{i,j}) + \epsilon_{i,j}\big) = \sum_{j=1}^{M} f_{\boldsymbol{\theta}}(t_{i,j}, c_{i,j}) + \sum_{j=1}^{M} \epsilon_{i,j}. \tag{2}$$

While we focus on the more general formulations above, the model can in principle be simplified further by using a linear parameterization of $f_{\boldsymbol{\theta}}(\cdot)$, with $\boldsymbol{\theta} = (\alpha_1, \beta_1, \alpha_2, \beta_2)$. This corresponds to two linear context–outcome relationships, one under treatment and one under control:

$$\widehat{Y}_i = \sum_{j=1}^{M} \big(t_{i,j}\,(\alpha_1 + \beta_1 c_{i,j}) + (1 - t_{i,j})\,(\alpha_2 + \beta_2 c_{i,j}) + \epsilon_{i,j}\big). \tag{3}$$

3.2 Training

We use μ_i to denote the regional (low-resolution) prediction of the model, obtained by aggregating the outputs of the neural network. Under additive zero-mean noise, μ_i estimates the expected re-

gional outcome $\widehat{Y}_i$. The model is trained by matching these predictions to the observed LR outcomes using an MSE loss:

$$\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}} = \frac{1}{N}\sum_{i=1}^{N}\left(\mu_i - \widehat{Y}_i\right)^2.$$

Importantly, the loss is available only at the regional level. The model does not observe individual subregional outcomes and must instead learn a shared mechanism whose aggregated predictions match the observed regional outcomes.

When relaxing the assumptions from Section 2, additional components can be learned jointly. For example, if the aggregation function $g_{\boldsymbol{\phi}}(\cdot)$ is unknown, it can be learned together with the mechanism $f_{\boldsymbol{\theta}}(\cdot)$. In general, $g_{\boldsymbol{\phi}}(\cdot)$ can be implemented using a permutation-invariant set neural network, such as the Deep Sets architecture proposed by Zaheer et al. (2017).

This is also conceptually similar to graph neural networks, where both the local message-passing functions and the aggregation of messages from neighboring nodes can be learned (Gilmer et al., 2017).

Similarly, if treatments are not uniformly applied across subregions, the underlying HR treatment matrix can be inferred jointly during training. Other latent variables, such as unobserved contextual variables, can be handled in the same way. This includes uneven delivery, for example a vaccination campaign reaching subregions to different degrees.

3.3 Downstream Applications

Training yields the predictive mechanism $f_{\boldsymbol{\theta}}(\cdot)$, from which several causal quantities can be derived. The most direct is what we call the *local conditional average treatment effect* (LOCATE), a version of the conditional average treatment effect (CATE) conditioned on subregion-specific factors:

$$E_{i,j} = f_{\boldsymbol{\theta}}(1,\, c_{i,j}) - f_{\boldsymbol{\theta}}(0,\, c_{i,j}). \tag{4}$$

This describes the treatment–control difference for a given subregion. Aggregating these local effects yields ATEs for regions, populations, or environments. For continuous treatments, the same mechanism can be used to compute conditional dose–response curves. LOCATE is noise-marginalized.

Beyond effect estimation, the framework supports outcome *disaggregation*, i.e., mapping coarse regional outcomes back to fine-grained predictions, as well as counterfactual reasoning, e.g., predicting outcomes under hypothetical intervention placements or alternative contexts. In the simplest case, this amounts to evaluating $f_{\boldsymbol{\theta}}(\cdot)$ under new inputs. For counterfactuals, if the noise term is believed to capture realization-specific information, inferred noise values can be reused to produce sharper, sample-specific predictions.

The framework can also recover latent HR treatment locations by representing them as trainable variables (cf. Exp. 2 and Appendix E.2). With time-series data, CLAM can further infer latent variables that modulate causal effects, provided temporal variation supplies enough constraints (cf. Exp. 3 and Appendix E.3).

3.4 Identifiability

Identifiability is central to causal spatial disaggregation because the same aggregate observations can often be explained by different local mechanisms. We do not provide a complete identifiability theory, but Appendix A lays the groundwork for one by disentangling causal identification, statistical identification from aggregate observations, and finite-sample recovery by the optimizer. It also shows why recovering the mechanism and recovering the high-resolution outcome map are distinct problems, and that neither is generally possible without additional assumptions. In short, in a noise-free setting with known sum aggregation, recovery reduces to a linear system with a simple rank condition. Beyond this, the consistency result of Zhang et al. (2020) covers an idealized version of our setting. Extending these foundations to spatial dependence, latent variables, general aggregation functions, and finite-sample optimization remains an important open problem.

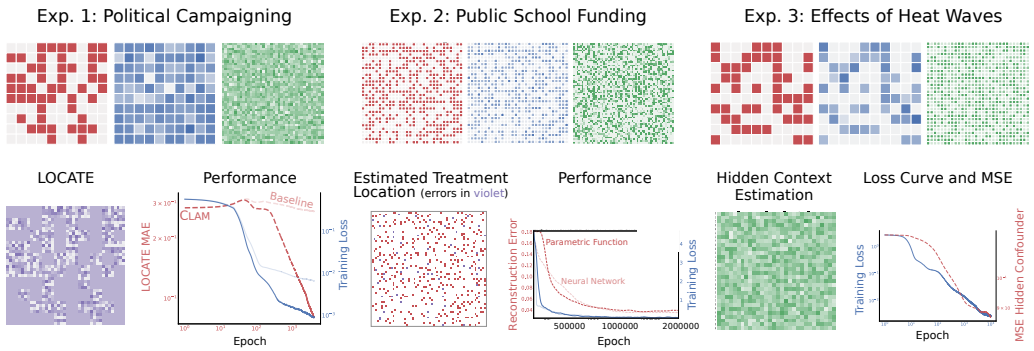


Figure 3: **Results Overview. Top:** Input of LR intervention locations (red), LR outcomes (blue), and observable HR context (green). **Bottom:** Selected inference results.

3.5 Relationship to Causal Deabstraction

We view CLAM as an instance of *causal deabstraction*: recovering a fine-grained causal mechanism from coarse, aggregated observations. While prior work studies when causal structure is preserved under abstraction (Beckers & Halpern, 2019; Zennaro, 2022), we consider the inverse direction: inferring a high-resolution model that is consistent with observed aggregates. Aggregation deletes the pairing between outcomes and subregions. HR covariates and an invariant causal mechanism (Peters et al., 2016; Schölkopf et al., 2021) can partially restore it when covariate compositions vary across regions. Appendix B formalizes this view and clarifies the causal semantics.

3.6 Exploring Motifs in the Causal Graph

So far, we have focused on the simplest causal motif, where treatment and context both affect the outcome, but context does not affect treatment. This keeps the causal assumptions clear and lets us focus on the first principles of spatial causal disaggregation. The same principle extends to richer causal graphs and we discuss possible extensions and their limitations in Appendix C.

As a short outlook, context may instead act as a confounder, mediator, collider, or instrumental variable. A confounder affects both treatment and outcome and therefore has to be accounted for when estimating treatment effects. A mediator lies on the causal path from treatment to outcome, while conditioning on a collider can introduce bias. Moreover, instrumental variables can provide useful treatment variation when hidden confounding is present. Other extensions include spatial spillovers, where neighboring treatments or contexts affect the outcome, and time-varying settings, where treatment and context can influence each other over time.

In all cases, the basic idea remains the same: high-resolution variables leave a structured signal in the regional observations through the causal mechanism, and the causal graph tells us how this information can be exploited.

4 Experiments

We evaluate CLAM on three synthetic studies and one real-world study. In the synthetic studies, the ground-truth mechanism is available, allowing us to isolate distinct capabilities of the framework under controlled conditions: heterogeneous local-effect recovery, latent intervention localization, and reconstruction of unobserved covariates. The real-world study then evaluates the method on U.S. gun-violence data, where we assume that only coarse regional outcomes are observed.

Full experimental details, per-experiment figures, and additional analyses are provided in Appendices E and F. Two additional studies on confounded treatment allocation and an unknown aggregation function are deferred to Appendix G. Our focus is qualitative, with the aim of characterizing the behavior of the method. Figure 3 provides an overview of the inputs and main results. Code that runs off-the-shelf on Google Colab is available on `GitHub`.

4.1 SYNTHETIC STUDIES

Exp. 1: Political Campaigning. We first ask whether CLAM can recover heterogeneous local causal effects from aggregated outcomes when the effect of treatment depends on subregional context. The setup is similar to the basic case in Figure 1. We use 10×10 regions, each containing 4×4 subregions. We model a binary regional intervention (campaign spending) whose impact varies with subregional wealth (z-scored income values). The context is constructed so that regional mean wealth is largely uninformative, while the distribution of wealth across subregions differs. The reported outcome is the candidate's relative improvement, observed only at the regional level. We compare against a naïve **baseline** (Figure 3, with confidence intervals in the Appendix) that uniformly disaggregates the regional outcome and then learns the causal mechanism.

This experiments highlights the limitation of such baselines. Because the regional mean context carries little information, it does not reveal how context, and therefore causal effects, vary across subregions. In the limiting case where regional outcomes are also identical, uniform, bicubic, and ecological-regression baselines (Goodman, 1953) cannot distinguish subregions and can at best assign the regional average effect (see also Appendix E.1.2). Exp. 1 is close to this regime, with only small variation in regional outcomes. CLAM can operate below this error floor because it directly uses the HR context and learns a shared mechanism linking local context to local outcomes across regions. This is the gain hypothesized in the introduction, where treating disaggregation and causal inference separately caps what is achievable.

We use a simple neural network to estimate the causal mechanism. CLAM accurately recovers the underlying subregional causal effects, with the MAE to the ground-truth LOCATE matrix (Eq. (4)) decreasing alongside the training loss. The learned model supports coherent counterfactual reasoning at both regional and subregional resolutions and correctly identifies regions whose outcome differences arise purely from contextual heterogeneity rather than mean effects. Full setup and per-subregion visualizations are given in Appendix E.1 (Figure 5).

Exp. 2: Public School Funding. Next, we test whether CLAM can recover *latent intervention locations* jointly with local causal effects when only aggregated intervention totals and outcomes are observed. Each of the 30×30 regions consists of 4×4 subregions and receives binary funding in either no subregion or exactly one unobserved subregion. Treatment effects vary with subregional socioeconomic status, and the outcome is education-score improvement.

We implement $f_\theta(\cdot)$ using either a parametric function or a small neural network. In both cases, CLAM jointly identifies the treated subregions and accurately recovers the underlying causal-effect function. To sample a treated subregion, we use the common Gumbel-Max trick. Details are provided in Appendix E.2 and Figure 8.

Exp. 3: Spatiotemporal Effects of Heat Waves. Finally, we consider how heat waves affect school performance over time. The effect is mitigated by two subregional factors: parental education, which is observed, and vegetation coverage, which is unobserved. We model 100 regions with 9 subregions each over 48 months. CLAM uses these repeated observations to jointly learn the causal mechanism and reconstruct the fixed, hidden vegetation pattern.

This experiment also highlights a flexibility–identifiability trade-off. With a correctly specified linear mechanism, CLAM recovers the vegetation map almost perfectly. With a flexible MLP, it still recovers meaningful spatial structure, but with higher error because the model can trade off changes in the learned mechanism against changes in the latent vegetation while producing similar regional outcomes. Details are provided in Appendix E.3 (Figure 9).

Takeaways. Across the three main studies, CLAM can recover local causal effects from aggregate supervision and can, in favorable settings, also recover latent spatial structure such as intervention locations or hidden effect modifiers. At the same time, Exp. 3 shows that recovery becomes less reliable as the model becomes more flexible. We also observe that increasing the size of the neural network can lead to overfitting. Two additional experiments in Appendix G consider harder settings with confounded treatment allocation and an unknown aggregation function. Overall, these experiments demonstrate the feasibility and flexibility of the approach, but they do not demonstrate a general recovery guarantee.

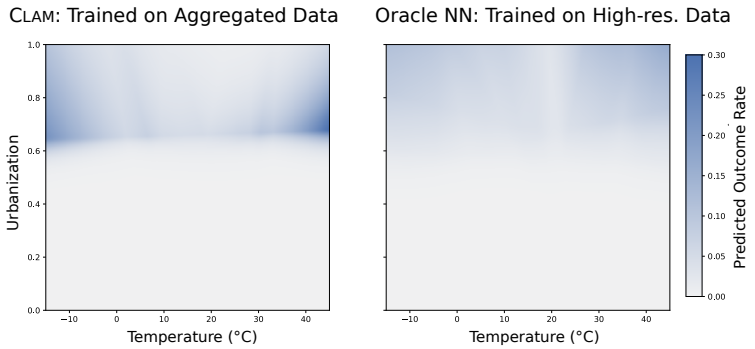


Figure 4: Side-by-side comparison of the learned rate functions $f_{\boldsymbol{\theta}}(\text{temp}, \text{urban})$ for the disaggregation from CLAM (left) and the oracle (a NN with the ground truth subregional data; right). Predictions are generated across a temperature range of $[-15, 45]^{\circ}\text{C}$ and an urbanization scale of $[0, 1]$.

4.2 SEMI-SYNTHETIC CASE STUDY WITH REAL-WORLD DATA: HEAT AND GUN VIOLENCE

We study how changes in temperature affect gun violence across space in the United States using data from the Gun Violence Archive (GVA) for 2022–2023 (Gun Violence Archive, 2015), motivated by prior evidence linking higher ambient temperatures to increased gun violence (Lyons et al., 2022). We treat temperature as a continuous exposure. Urbanization is the high-resolution contextual covariate, and gun-incident counts are the outcome and observed at weekly resolution.

A core difficulty in evaluating disaggregation methods on real data is the absence of ground truth at the subregional scale. We work around this by starting from data that is natively HR (gun-incident counts) and *artificially* coarsening it into LR regional counts, which serve as the inputs to CLAM. This makes the study semi-synthetic. For evaluation, we additionally fit an **oracle baseline** directly to the original HR counts. Unlike CLAM, this model has access to the true subregional outcomes and therefore provides a reference for what could be learned with full HR supervision.

We use $N = 100$ regions with $M = 100$ subregions each, indexed by region i, subregion j, and week w. Because gun incidents are sparse count data, with many subregions recording no incidents in a given week, we model their occurrence through Poisson rates. Let $Y_{i,w}^{\text{LR}}$ be the observed regional count, $c_{i,j}$ the HR urbanization covariate, and $\widetilde{t}_{i,w}$ the standardized regional temperature. We model subregional rates as $\lambda_{i,j,w} = f_{\boldsymbol{\theta}}(\widetilde{t}_{i,w}, c_{i,j})$ and sum them to obtain the regional rate $\mu_{i,w} = \sum_{j=1}^{M} \lambda_{i,j,w}$, with $Y_{i,w}^{\text{LR}} \sim \text{Poisson}(\mu_{i,w})$. Training minimizes the regional Poisson negative log-likelihood.

Figure 4 shows that CLAM recovers a qualitatively similar temperature–urbanization interaction to the oracle. The spatial comparison in Appendix Figure 10 shows the same pattern, with plausible structure but inflated and more diffuse local intensities, reflecting an aggregation-induced identifiability gap.

5 LIMITATIONS

CLAM relies on assumptions that are necessary given the nature of inferring fine-grained causal effects from aggregated data, but are nonetheless restrictive. We discuss the main limitations here.

Do-Calculus Identifiability. CLAM is not a substitute for causal identification. The method should only be applied in settings where do-calculus (applied to a known or expert-specified causal graph) establishes that the causal effect is identifiable in principle. If the causal graph is unknown or the required identification conditions fail (e.g., the backdoor criterion is violated by unmodeled confounders), CLAM will fit a model to the data that does not carry a causal interpretation.

No Identification Guarantee from Optimization Alone. Even when do-calculus certifies identifiability in principle, we do not provide a formal guarantee that our optimization procedure recovers

the true causal effect. The training objective is non-convex, and the best we can reasonably claim is that, under good initialization and sufficient covariate diversity, gradient descent finds a local minimum whose value is close to the true causal effect. It is easy to come up with pathological cases in which different fine-grained disaggregations are equally consistent with the observed regional outcomes. How to detect or avoid them in practice, remains an open question.

Assumptions and Their Scope. Successful recovery of the local mechanism generally relies on (i) an invariant causal mechanism across subregions, (ii) sufficient diversity in subregional covariate distributions across regions, (iii) no hidden confounding, or its explicit modeling, and (iv) a known or sufficiently constrained aggregation function.

The invariance assumption (i) is less restrictive than it may appear: although $f_{\boldsymbol{\theta}}(\cdot)$ is shared across regions and subregions, region-specific dynamics can still be captured through contextual covariates such as region indicators or positional encodings. Apparent heterogeneity across regions is absorbed by the context, yielding a *conditionally* invariant model. The stronger requirement is that the contextual covariates provide sufficient variation to constrain the shared mechanism and capture sources of heterogeneity that are correlated with the treatment. Unobserved factors that are uncorrelated with the treatment can effectively act as additional noise: they make recovery harder but need not bias the estimated effect. Unobserved confounders, by contrast, will generally lead to incorrect causal estimates unless they are explicitly modeled. The aggregation function (iv) need not be perfectly known in advance: as demonstrated in Exp. 4 (Appendix G), it can be jointly learned alongside the local mechanism when it is sufficiently parameterized and the available variation allows the two components to be distinguished.

Identifiability Gap and Latent Variable Recovery. Even when all assumptions hold, aggregation induces an intrinsic identifiability gap: matching aggregated predictions to aggregated observations may admit multiple equally valid solutions. Our real-world study illustrates this, with CLAM systematically inflating local effect magnitudes relative to the cell-supervised oracle. The problem is amplified when latent variables are jointly inferred (e.g., intervention locations in Exp. 2, unobserved covariates in Exp. 3), since different inferred values can each correspond to a mechanism that explains the data equally well. We discuss this further in Appendix A and regard a rigorous treatment, together with practical diagnostics, as the most important direction for future work.

Limited Real-World Validation. Our evaluation is primarily qualitative and relies substantially on synthetic studies, where the data-generating assumptions are known and can be matched to the model. The real-world study uses a subregion-supervised oracle as a reference rather than a true causal benchmark. However, the identified causal effect of temperature on gun-violence intensity is also supported by prior literature (Lyons et al., 2022). Robustness to noise, measurement error, and treatment–context dependence is explored in the additional experiments, but a more systematic benchmarking remains an important direction for future work.

6 Related Work

CLAM sits at the intersection of multiple research strands. We provide an extended discussion of related work in Appendix D.

Spatial Disaggregation and Downscaling. A long line of work in statistical downscaling and spatial disaggregation infers high-resolution estimates from coarse data using interpolation, regression, or learning-based techniques (Patil et al., 2024; Sun et al., 2024; Maraun, 2019; Chau et al., 2021). These methods are predominantly associational, recent extensions integrate causally relevant predictors (Vázquez-Patiño et al., 2022; Dutta & Maity, 2020) but still rely on relatively fine-grained information and do not target interventional or counterfactual queries.

Causal Inference under Aggregation and Abstraction. Reasoning causally with aggregated data is well known to be hazardous: results on the ecological fallacy show that aggregate statistics may obscure or invert relationships at the unit level (Freedman, 1999; Rogers et al., 1991; Weinstein & Blei, 2026; Li et al., 2026). Work on causal abstraction formalizes when mappings between models at different granularities preserve interventional or counterfactual semantics (Beckers &

Halpern, 2019; Beckers et al., 2020; Zennaro, 2022). Spatiotemporal causal inference (Christiansen et al., 2022; Zhou et al., 2024; Mukaigawara et al., 2025; Song & Papadogeorgou, 2024; Oprescu et al., 2026; Ninad et al., 2025; Balkus et al., 2024) addresses related challenges with different data regimes. None of these directly target the joint problem of disaggregating outcomes *and* estimating local causal effects from coarse interventions.

Invariant Mechanisms, GNNs, and Hierarchical Models. Conceptually, CLAM relies on the principle of *invariant causal mechanisms* (Peters et al., 2016; Schölkopf et al., 2021; Heinze-Deml et al., 2018; Rojas-Carulla et al., 2018): the local mechanism $f_{\theta}(\cdot)$ is shared across subregions while contextual covariates drive heterogeneity. This turns aggregation into a constraint and connects our architecture to message-passing GNNs (Gilmer et al., 2017) and Deep Sets (Zaheer et al., 2017), as well as to hierarchical Bayesian models that pool information across groups (Feller & Gelman, 2015; Dias et al., 2013). It also places CLAM in the setting of learning from aggregate observations (Zhang et al., 2020), whose training loss coincides with ours under mean aggregation (cf. Appendix A).

7 Conclusions and Future Work

The primary contribution of this work is not a single algorithm but the articulation of a *problem class*: estimating high-resolution causal effects from low-resolution interventions and outcomes by exploiting high-resolution context. We frame this as *causal deabstraction*, inverting the standard notion of causal abstraction and recasting spatial aggregation as a constraint satisfaction problem. We thereby bridge structural causal models and spatial statistics. In many real-world settings such as public health, environmental policy, and education, spatial disaggregation and causal estimation are inseparable, since interventions are implemented at coarse scales but their effects manifest locally. Treating them jointly within a structural causal model makes assumptions explicit, clarifies what is and is not identifiable, and turns causal inference into a tractable optimization problem.

CLAM is one concrete instantiation of this framework. It explicitly adjusts for subregional contextual effects using only coarse regional supervision and jointly learns latent high-resolution variables such as exact intervention locations and latent spatial effect modifiers alongside the causal mechanism. Across synthetic and real-world studies, we recover meaningful spatial and functional structure from aggregated data, while exposing the inherent limits of the setting. We find that relative structure and causal dependencies are recoverable, but absolute local magnitudes may be systematically overestimated without fine-grained outcome supervision.

We also provide first identifiability results for this problem class. In the general setting, we give negative results showing that neither the local mechanism nor the high-resolution outcome map is identifiable without additional restrictions. For more structured settings, we obtain positive results through a rank-based characterization and by connecting CLAM to existing aggregate-regression theory, which establishes recovery of the shared local function under stronger assumptions. Together, these results begin to map the boundary between identifiable and non-identifiable settings. Because this boundary is not yet fully understood, a central open problem is to characterize which combinations of aggregation, contextual variation, spatial dependence, latent structure, and model restrictions still permit recovery, and which do not. More broadly, principled baselines and theory remain scarce for this problem class, and we hope that making the problem explicit helps motivate both.

Although this paper primarily proposes a research framework, misuse in high-stakes domains such as epidemiology could cause harm if poorly identified local effects guided interventions or resource allocation. Likewise, outputs should be accompanied by uncertainty estimates, sensitivity analyses, and domain expertise. Disaggregation itself may also raise privacy concerns, since inferred high-resolution patterns can reveal sensitive local information.

Methodologically, principled regularization, hierarchical priors, or Bayesian formulations could control aggregation-induced bias and provide uncertainty quantification. Relaxing the shared-mechanism assumption, integrating causal discovery over contextual variables, incorporating instrumental-variable settings, and extending to richer spatiotemporal aggregation are natural next steps. We hope that framing CLAM as a first concrete entry in a broader problem class makes these extensions easier to pursue.

AI USE STATEMENT

We used generative AI tools to improve wording and readability, to cross-check our literature search for potentially overlooked relevant work, and to support the implementation in collaboration with an AI coding agent. All suggestions and generated code were reviewed by the authors, who take responsibility for the final content and implementation.

ETHICS STATEMENT

Causal spatial disaggregation can support decisions in domains such as public health, education, and environmental policy, but poorly identified local effects may also mislead interventions or resource allocation. Practical use should therefore include uncertainty estimates, sensitivity analysis, and domain expertise. Inferring high-resolution patterns from coarse data may also create privacy risks.

REFERENCES

Saeed Anwar, Salman Khan, and Nick Barnes. A deep journey into super-resolution: A survey. *ACM computing surveys (CSUR)*, 53(3):1–34, 2020.

Salvador V Balkus, Scott W Delaney, and Nima S Hejazi. The causal effects of modified treatment policies under network interference. *arXiv preprint arXiv:2412.02105*, 2024.

Sander Beckers and Joseph Y Halpern. Abstracting causal models. In *Proceedings of the aaai conference on artificial intelligence*, volume 33, pp. 2678–2685, 2019.

Sander Beckers, Frederick Eberhardt, and Joseph Y Halpern. Approximate causal abstractions. In *Uncertainty in artificial intelligence*, pp. 606–615. PMLR, 2020.

Siu Lun Chau, Shahine Bouabid, and Dino Sejdinovic. Deconditional downscaling with gaussian processes. *Advances in Neural Information Processing Systems*, 34:17813–17825, 2021.

Rune Christiansen, Matthias Baumann, Tobias Kuemmerle, Miguel D Mahecha, and Jonas Peters. Toward causal inference for spatio-temporal data: Conflict and forest loss in colombia. *Journal of the American Statistical Association*, 117(538):591–601, 2022.

Ronaldo Dias, Nancy L Garcia, and Alexandra M Schmidt. A hierarchical model for aggregated functional data. *Technometrics*, 55(3):321–334, 2013.

Riya Dutta and Rajib Maity. Identification of potential causal variables for statistical downscaling models: effectiveness of graphical modeling approach. *Theoretical and Applied Climatology*, 142 (3):1255–1269, 2020.

Robert E. Fay and Roger A. Herriot. Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74(366):269–277, 1979.

Avi Feller and Andrew Gelman. Hierarchical models for causal effects. *Emerging trends in the social and behavioral sciences*, 1:16, 2015.

David A Freedman. Ecological inference and the ecological fallacy. *International Encyclopedia of the social & Behavioral sciences*, 6(4027-4030):1–7, 1999.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pp. 1263–1272. Pmlr, 2017.

Leo A Goodman. Ecological regressions and behavior of individuals. *American sociological review*, 18(6):663, 1953.

Carol A. Gotway and Linda J. Young. Combining incompatible spatial data. *Journal of the American Statistical Association*, 97(458):632–648, 2002.

Gun Violence Archive. Gun violence archive gva. Web Archive, 2015. URL https://www.loc.gov/item/lcwaN0016293/. United States. Retrieved from Library of Congress: https://www.loc.gov/item/lcwaN0016293/.

Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2):20170016, 2018.

Gary King. *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton University Press, 1997.

Bipin Kumar, Kaustubh Atey, Bhupendra Bahadur Singh, Rajib Chattopadhyay, Nachiketa Acharya, Manmeet Singh, Ravi S Nanjundiah, and Suryachandra A Rao. On the modern deep learning approaches for precipitation downscaling. *Earth Science Informatics*, 16(2):1459–1472, 2023.

Ho Chung Leon Law, Dino Sejdinovic, Ewan Cameron, Tim C.D. Lucas, Seth Flaxman, Katherine Battle, and Kenji Fukumizu. Variational learning on aggregate outputs with Gaussian processes. In *Advances in Neural Information Processing Systems*, 2018.

Xintong Li, Haoran Zhang, and Xiao Zhou. Spatio-temporal hierarchical causal models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 23230–23238, 2026.

James Liley, Samuel Emerson, Bilal Mateen, Catalina Vallejos, Louis Aslett, and Sebastian Vollmer. Model updating after interventions paradoxically introduces bias. In *International Conference on Artificial Intelligence and Statistics*, pp. 3916–3924. PMLR, 2021.

Vivian H Lyons, Emma L Gause, Keith R Spangler, Gregory A Wellenius, and Jonathan Jay. Analysis of daily ambient temperature and firearm violence in 100 us cities. *JAMA network open*, 5 (12):e2247207–e2247207, 2022.

Douglas Maraun. Statistical downscaling for climate science. In *Oxford Research Encyclopedia of Climate Science*. 2019.

Timothy C Matisziw, Tony H Grubesic, and Hu Wei. Downscaling spatial structure for the analysis of epidemiological data. *Computers, Environment and Urban Systems*, 32(1):81–93, 2008.

Mitsuru Mukaigawara, Kosuke Imai, Jason Lyall, and Georgia Papadogeorgou. Spatiotemporal causal inference with arbitrary spillover and carryover effects. *arXiv preprint arXiv:2504.03464*, 2025.

Urmi Ninad, Jonas Wahl, Andreas Gerhardus, and Jakob Runge. Causal discovery on vector-valued variables and consistency-guided aggregation. *arXiv preprint arXiv:2505.10476*, 2025.

Miruna Oprescu, David Park, Xihaier Luo, Shinjae Yoo, and Nathan Kallus. Gst-unet: A neural framework for spatiotemporal causal inference with time-varying confounding. *Advances in Neural Information Processing Systems*, 38:17295–17322, 2026.

Shruthi Patil, Noah Pflugradt, Jann M Weinand, Detlef Stolten, and Jürgen Kropp. A systematic review of spatial disaggregation methods for climate action planning. *Energy and AI*, 17:100386, 2024.

Judea Pearl. From bayesian networks to causal networks. In *Mathematical models for handling partial knowledge in artificial intelligence*, pp. 157–182. Springer, 1995.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Jian Peng, Alexander Loew, Olivier Merlin, and Niko EC Verhoest. A review of spatial downscaling of satellite remotely sensed soil moisture. *Reviews of Geophysics*, 55(2):341–366, 2017.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):947–1012, 2016.

William S. Robinson. Ecological correlations and the behavior of individuals. *American Sociological Review*, 15(3):351–357, 1950.

David F Rogers, Robert D Plante, Richard T Wong, and James R Evans. Aggregation and disaggregation techniques and methodology in optimization. *Operations research*, 39(4):553–582, 1991.

Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19(36):1–34, 2018.

Jakob Runge, Sebastian Bathiany, Erik Bollt, Gustau Camps-Valls, Dim Coumou, Ethan Deyle, Clark Glymour, Marlene Kretschmer, Miguel D Mahecha, Jordi Muñoz-Marí, et al. Inferring causation from time series in earth system sciences. *Nature communications*, 10(1):2553, 2019.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Alexander A Schuessler. Ecological inference. *Proceedings of the National Academy of Sciences*, 96(19):10578–10581, 1999.

Zhaoyan Song and Georgia Papadogeorgou. Bipartite causal inference with interference, time series data, and a random network. *arXiv preprint arXiv:2404.04775*, 2024.

Yongjian Sun, Kefeng Deng, Kaijun Ren, Jia Liu, Chongjiu Deng, and Yongjun Jin. Deep learning in statistical downscaling for deriving high spatial resolution gridded meteorological data: A systematic review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 208:14–38, 2024.

Angel Vázquez-Patiño, Esteban Samaniego, Lenin Campozano, and Alex Avilés. Effectiveness of causality-based predictor selection for statistical downscaling: a case study of rainfall in an ecuadorian andes basin. *Theoretical and Applied Climatology*, 150(3):987–1013, 2022.

Eli N Weinstein and David M Blei. Hierarchical causal models. *Journal of Machine Learning Research*, 27(37):1–73, 2026.

Xinyue Ye, Qunying Huang, and Wenwen Li. Integrating big social data, computing and modeling for spatial social science, 2016.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017.

Fabio Massimo Zennaro. Abstraction between structural causal models: A review of definitions and properties. *arXiv preprint arXiv:2207.08603*, 2022.

Yivan Zhang, Nontawat Charoenphakdee, Zhenguo Wu, and Masashi Sugiyama. Learning from aggregate observations. *Advances in Neural Information Processing Systems*, 33:7993–8005, 2020.

Lingxiao Zhou, Kosuke Imai, Jason Lyall, and Georgia Papadogeorgou. Estimating heterogeneous treatment effects for spatio-temporal causal inference: How economic assistance moderates the effects of airstrikes on insurgent violence. *arXiv preprint arXiv:2412.15128*, 2024.

A IDENTIFIABILITY

Identification in CLAM depends on the target of interest and on the model assumptions. We distinguish causal identification under an assumed graph, statistical identification from aggregate observations, and finite-sample recovery by the optimizer. The last is an estimation problem rather than an identification question.

However, a complete theory for statistical identification from aggregates is beyond the scope of this work. Below, we characterize several tractable special cases and discuss the main remaining open problems.

A.1 IDENTIFIABILITY OF WHAT

First, *causal identification* asks whether a quantity such as LOCATE can be expressed in terms of the observed distribution under the assumed causal graph. We treat the graph as given and do not consider causal discovery. This is separate from whether the required fine-grained mechanism can be recovered from aggregate observations.

Second, *statistical identification* asks whether the local mechanism $f_{\boldsymbol{\theta}}(\cdot)$ is uniquely determined by the population distribution of the aggregate data. We are generally interested in the function itself rather than the parameter vector $\boldsymbol{\theta}$, since different parameterizations may represent the same function.

Third, one can ask whether the high-resolution outcome map $\{y_{i,j}\}$ is identified. Even with a known mechanism, exact local outcomes may remain unknown because the individual noise realizations $\epsilon_{i,j}$ and possibly latent inputs $c'_{i,j}$ are unobserved. Conversely, aggregates can sometimes determine particular local outcomes without identifying the full mechanism.

A.2 GENERAL NON-IDENTIFIABILITY

Observed setup

	context C	
$A,\ \widehat{Y}_A = 1$	0	1
$B,\ \widehat{Y}_B = 1$	0.5	0.5

Explanation 1: $f(0, c) = c$, $g = \text{sum}$

A	0	1
B	0.5	0.5

Explanation 2: $f(0, c) \equiv 1$, $g = \text{mean}$

A	1	1
B	1	1

Explanation 3: $f(0, c) \equiv 0.5$, $g = \text{sum}$

A	0.5	0.5
B	0.5	0.5

Without restrictions, neither the mechanism nor the high-resolution outcome map is identified from aggregate observations.

Consider a noise-free setting similar to Equation (2) in which no region receives treatment. Take two regions A and B, each with two subregions, with context values $(0, 1)$ and $(0.5, 0.5)$, respectively, and suppose that both observed regional aggregates equal 1. If both the local mechanism $f_{\boldsymbol{\theta}}(\cdot)$ and the aggregation function $g_{\boldsymbol{\phi}}(\cdot)$ are unknown, infinitely many pairs are compatible with these observations.

For example, $f_{\boldsymbol{\theta}}(0, c) = c$ yields subregional outcomes $(0, 1)$ for region A and $(0.5, 0.5)$ for region B. With sum aggregation, both regions produce the observed aggregate of 1 (Explanation 1). Alternatively, the constant mechanism $f_{\boldsymbol{\theta}}(0, c) \equiv 1$ yields $(1, 1)$ in both regions, and mean aggregation again gives an aggregate of 1 (Explanation 2). Thus, both explanations fit the observed data while implying different high-resolution outcomes.

The ambiguity remains even when the aggregation function is known. Suppose $g_{\boldsymbol{\phi}}(\cdot)$ is fixed to be a sum. Then $f_{\boldsymbol{\theta}}(0, c) = c$ is compatible with the data, but so is $f_{\boldsymbol{\theta}}(0, c) \equiv 0.5$ (Explanation 3). The latter yields $(0.5, 0.5)$ in both regions, which again sums to 1.

A.3 IDENTIFIABILITY AS A RANK PROBLEM

A useful limiting case is to think of $f_{\boldsymbol{\theta}}(\cdot)$ as a dictionary lookup. Each distinct observed input $(t_{i,j}, c_{i,j})$ gets its own independent unknown value. Let v_k denote the outcome assigned to the kth distinct input. With known sum aggregation and no noise, Equation (2) can then be written as

$$\widehat{Y}_i = \sum_k A_{ik} v_k,$$

or, after stacking all regions,

$$\widehat{\mathbf{Y}} = A\mathbf{v}.$$

Here, each column of A corresponds to one entry in the dictionary, and each row corresponds to one region. The value A_{ik} records how often the kth input occurs among the subregions of region i. Finding $\mathbf{v}$ therefore means finding one outcome value for each distinct treatment–context combination such that, when these values are added within every region, they reproduce all observed regional outcomes. If the regions contain sufficiently different combinations of inputs, these values

can be separated. Formally, they are uniquely identified when A has full column rank. Otherwise, different dictionary values can produce exactly the same aggregates.

A neural network moves away from this saturated case by tying the dictionary entries together through the shared function $f_{\boldsymbol{\theta}}(\cdot)$. For the linear model in Equation (3), the same idea reduces to ordinary linear regression: after aggregation, the regional outcomes are linear in the model parameters. These parameters are uniquely identified exactly when the corresponding aggregate design matrix has full column rank.

A.4 Aggregate Regression

The saturated view above is deliberately pessimistic because it treats every distinct input as unrelated. With a shared mechanism and many varying aggregates, substantially more can be identified. This connects directly to learning from aggregate observations as studied by Zhang et al. (2020).

Our additive model in Equation (2) is closely related to their regression-from-mean setting. Consider known sum aggregation, the same number M of subregions in every region, additive homoscedastic Gaussian noise, no latent c', and the sampling assumptions of Zhang et al. (2020). Since sum and mean aggregation then differ only by the constant factor M, minimizing the aggregate MSE is equivalent up to rescaling.

Following the notation of Zhang et al. (2020), let W_0 denote an optimal parameterization of the regression function. They use W for the model parameters, corresponding to $\boldsymbol{\theta}$ in our notation. Let $[W_0]$ denote all parameter values that are observationally equivalent to W_0, meaning that they induce the same population distribution of aggregate observations. Proposition 7 characterizes this equivalence class as

$$[W_0] = \{W \in \mathcal{W} \mid f(X; W) = f(X; W_0) \text{ almost everywhere}\}. \tag{5}$$

The key implication is that, under their assumptions, aggregate supervision is sufficient to identify the underlying individual-level regression function in the population limit, even though individual outcomes are never observed. The parameters themselves may not be unique, but observationally equivalent parameters must represent the same function almost everywhere. This provides an important base case for CLAM. Specifically, sufficiently varied aggregates can, in fact, identify a shared local mechanism.

The main gap is that the assumptions required for this result are considerably stronger than those of our general spatial setting. In particular, Zhang et al. (2020) assume that individual outcomes within an aggregate are conditionally independent given their inputs. Spatial subregions, however, are often dependent because nearby locations share unobserved factors and spatial structure. The authors explicitly identify spatial aggregation as a setting where this assumption may fail. Moreover, their result concerns the population limit, assumes a known aggregation function and a correctly specified observation model. Causal interpretation and counterfactual effects therefore still require the causal and overlap assumptions discussed above.

Note that this correspondence does not contradict the counterexample above. There, only two fixed aggregates are observed. Under the stronger assumptions of Zhang et al. (2020), many varying aggregates provide enough information to identify the shared function in the population limit.

In summary, their theory thus provides a useful foundation, but not an identifiability result for the full CLAM setting. Given the negative result above, the open problem is to characterize the boundary of identifiability: which forms of spatial dependence, latent structure, aggregation, and finite-sample variation still permit recovery, and which do not.

A.5 Open Problems

The results above give a general negative result and positive results for restricted settings. Several parts of the full problem remain open.

Spatio-Temporal Extensions and Latent Structure. Repeated observations over time can help recover persistent latent structure. A latent variable c' that remains stable across time affects many aggregate observations, while observation-specific noise $\epsilon_{i,j}$ affects only one. Exp. 3 illustrates this:

the same hidden vegetation field influences outcomes across many time points, allowing information about it to accumulate. A general theory for when such latent structure is identifiable remains open.

Unknown Aggregation. If $g_{\boldsymbol{\phi}}(\cdot)$ is unknown and learned jointly with $f_{\boldsymbol{\theta}}(\cdot)$, different combinations of the two may explain the same aggregate observations. Identification therefore requires restrictions on one or both functions, sufficient variation in the data, or prior knowledge about the aggregation process. This makes known sum or mean aggregation substantially easier than the general model in Equation (1).

Nonlinear Aggregation and Noise. With linear aggregation and additive zero-mean noise, aggregation commutes with expectation. For nonlinear aggregation, this generally does not hold:

$$g_{\boldsymbol{\phi}}\left(\{\mathbb{E}[y_{i,j} \mid T, C]\}_j\right) \neq \mathbb{E}\left[g_{\boldsymbol{\phi}}\left(\{y_{i,j}\}_j\right) \mid T, C\right]. \tag{6}$$

Applying a nonlinear aggregator directly to mean predictions can therefore target the wrong quantity. A suitable objective must propagate the noise through the aggregation, for example through a likelihood or Monte Carlo approximation. This fixes the objective mismatch but does not guarantee identification.

Causal Structure. The statistical results above concern recovery of a function from aggregate data, not whether that function has a causal interpretation. We assume a known causal graph. Under the baseline motif, $f(t, c)$ can represent the conditional interventional mean used for LOCATE. Other cases are discussed in Appendix C.

Identification Versus Optimization. Population identification does not guarantee recovery by a finite-sample non-convex optimizer. Neural networks, latent variables, and learned aggregation functions can introduce local optima and flat solutions. Conversely, optimization failure does not imply non-identifiability. We therefore treat optimization stability as a separate empirical question.

A.6 Practical Diagnostics

No finite-sample diagnostic can prove structural identification, but several checks can reveal weakly constrained solutions. We recommend comparing LOCATE estimates across random restarts, architectures, regularization choices, and plausible aggregation or noise models. Large disagreement suggests an underdetermined problem, while agreement is reassuring but not conclusive.

Whenever possible, estimates should also be checked against higher-resolution observations and through sensitivity analyses under plausible violations of the assumptions. Bootstrap intervals capture sampling uncertainty, but should not be interpreted as evidence of structural identification.

A.7 Summary of Claims

The discussion can be summarized in three levels.

1. For the general model in Equation (1), identification is not guaranteed. Our counterexample shows that different local mechanisms, high-resolution outcome maps, and aggregation functions can produce exactly the same aggregate observations.
2. For the additive model in Equation (2), stronger assumptions give positive results. In particular, the i.i.d. fixed-size aggregation setting connects to Zhang et al. (2020), whose results imply identification of the shared regression function almost everywhere on the support of the observed inputs.
3. For the linear model in Equation (3), identification becomes a standard rank problem. Aggregating the linear local mechanism produces a linear regression in the model parameters, which are uniquely identified when the resulting aggregate design matrix has full column rank.

Causal identification is separate: a statistically identified local mechanism has the desired causal interpretation only under the assumptions encoded by the causal graph.

B CAUSAL DEABSTRACTION FRAMEWORK

In this section, we situate our work within the broader causal inference literature. Specifically, we frame it as a novel instance of *causal deabstraction* (Beckers & Halpern, 2019; Zennaro, 2022). Our methodology is informed by the framework of SCMs and guided by the principle of invariant causal mechanisms (ICM).

B.1 STRUCTURAL CAUSAL FRAMEWORK

The structural causal model introduced in Section 3.1 can be formalized in generative form as follows. Each spatial unit (region) $i \in \{1, \dots, N\}$ is composed of M subregions indexed by $j \in \{1, \dots, M\}$. For each subregion (i, j), we posit the following SCM:

$$
\begin{aligned}
& c_{i,j} \text{ observed (exogenous)} \\
& \epsilon_{i,j} \sim P_\epsilon \\
& t_{i,j} = T_i \in \{0, 1\} \\
& y_{i,j} = f_{\boldsymbol{\theta}}(t_{i,j}, c_{i,j}, \epsilon_{i,j}) \\
& \widehat{Y}_i = g_{\boldsymbol{\phi}}(\{y_{i,j}\}_{j=1}^{M}).
\end{aligned}
$$

Here, $c_{i,j}$ denotes a high-resolution contextual covariate for subregion j within region i, which is observed and exogenous, meaning it is not caused by T or Y, and $\epsilon_{i,j}$ represents exogenous noise drawn from P_ϵ. The binary intervention $t_{i,j} = T_i \in \{0, 1\}$ is assigned at the regional level and shared across all subregions. The function $f_{\boldsymbol{\theta}}(\cdot)$ encodes the local causal mechanism, shared across all subregions, that maps the intervention, context, and noise to the subregional outcome $y_{i,j}$. The aggregation function $g_{\boldsymbol{\phi}}(\cdot)$ maps the collection of subregional outcomes $\{y_{i,j}\}_{j=1}^{M}$ to the regional outcome $\widehat{Y}_i$.

In the main section, we considered simplified cases, such as sum aggregation, additive noise, and the absence of hidden confounders, to provide intuition. The formulation here generalizes that view by expressing the structural causal model in explicit stochastic form, specifying how interventions, contextual variables, and noise enter the model, while making clear that both $f_{\boldsymbol{\theta}}(\cdot)$ and $g_{\boldsymbol{\phi}}(\cdot)$ are shared across regions.

B.2 INVARIANT CAUSAL MECHANISM ASSUMPTION

A central assumption in CLAM is that the causal function $f_{\boldsymbol{\theta}}(\cdot)$ is *invariant* across all subregions and regions. That is, while the covariates $\{c_{i,j}\}$ may vary across regions due to differences in local composition, the functional form $f_{\boldsymbol{\theta}}(t, c)$ does not change. This assumption reflects the principle of *invariant causal mechanisms* (Schölkopf et al., 2021; Peters et al., 2016; Heinze-Deml et al., 2018), which posits that causal mechanisms are stable across environments unless directly altered.

We treat each region i as an *environment* characterized by its observed covariate composition $\{c_{i,j}\}_{j=1}^{M}$, but governed by a shared mechanism $f_{\boldsymbol{\theta}}(\cdot)$. Variation in covariate composition across regions provides information for estimating $f_{\boldsymbol{\theta}}(\cdot)$ from the observed aggregated outcomes $\widehat{Y}_i$.

B.3 DEABSTRACTION VIA INVARIANCE

We propose to view the task of estimating $f_{\boldsymbol{\theta}}(\cdot)$ from aggregated outcomes as an instance of *causal deabstraction*, that is, inferring a latent, high-resolution causal model that is consistent with a lower-resolution aggregated model. This perspective is dual to recent work on causal abstraction (Beckers & Halpern, 2019; Zennaro, 2022), which studies when a coarse model preserves the counterfactual semantics of a fine-grained one. Here, we reverse the direction: we seek to recover a fine-grained causal mechanism that is *compatible* with the observed coarse-level outcomes.

This is enabled by two key properties:

1. The aggregation function $g_{\boldsymbol{\phi}}(\cdot)$ is consistent and applied uniformly across regions.
2. The observed covariate compositions $\{c_{i,j}\}_{j=1}^{M}$ differ across regions, providing variation in the aggregates $\widehat{Y}_i$ under the shared mechanism $f_{\boldsymbol{\theta}}(\cdot)$.

Given sufficient diversity and the assumption of causal invariance, the aggregated outputs can provide a supervisory signal for recovering the latent causal function, although, as discussed above, this recovery is not guaranteed in the general setting.

B.4 Causal Semantics via Do-Calculus

Under the assumption of no hidden confounding, the causal effect of the intervention T_i on a subregional outcome $y_{i,j}$ is identified via the backdoor criterion:

$$\mathbb{E}[y_{i,j} \mid do(T_i = t),\, c_{i,j}] = \int f_{\boldsymbol{\theta}}(t, c_{i,j}, \epsilon_{i,j})\, p(\epsilon_{i,j})\, d\epsilon_{i,j}.$$

In Exps. 1–4, we assume $T_i \perp\!\!\!\perp \epsilon_{i,j} \mid c_{i,j}$ and that T_i is either randomized or independent of $C_{i,j}$. Under this setting, the expectation can be estimated from observational data. However, because only the aggregate outcome $\widehat{Y}_i$ is observed, we rely on the consistency constraint

$$\widehat{Y}_i \approx \sum_{j=1}^{M} f_{\boldsymbol{\theta}}(T_i, c_{i,j}).$$

The training procedure thus finds a function $f_{\boldsymbol{\theta}}(\cdot)$ that jointly explains the observed aggregates under this constraint, while enforcing that $f_{\boldsymbol{\theta}}(\cdot)$ is invariant across regions. In this sense, estimating $f_{\boldsymbol{\theta}}(\cdot)$ becomes a constraint satisfaction problem, where causal invariance and aggregation consistency define the feasible solution space. In Exp. 5, this conditional independence assumption is deliberately violated by construction, and valid estimation requires explicitly modeling the treatment allocation mechanism.

B.5 Implications

The assumption of an invariant subregional mechanism plays a central role in enabling causal inference in our disaggregated setting. By treating coarse observations as aggregates of fine-grained outcomes and leveraging variation in covariate compositions, CLAM aims to recover a shared causal mechanism rather than merely statistical associations. Overall, we provide a formal and algorithmic instantiation of *causal deabstraction* from low-resolution outcomes and high-resolution covariates.

C Exploiting Context Across Motifs in the Causal Graph

Throughout the paper, we focus on the simplest causal motif: a treatment T acting on an outcome Y, with a contextual covariate C that modulates the effect. Real-world settings can exhibit a much richer variety of causal structures, and the role of high-resolution context, C, shifts with the motif. Sometimes it must be conditioned on, sometimes ignored, sometimes leveraged for new identification strategies. Do-calculus tells us *whether* an effect is identifiable given the graph, this section asks *how* the LR/HR distinction interacts with that question. Here, we provide an outlook in which we sketch how HR context can be exploited in common causal graph motifs.

Baseline: C as a Pure Effect Modifier: $T \to Y$, $C \to Y$. The default setting throughout this paper is the simplest motif: T and C are independent inputs, both pointing into Y, with no edge between them. C is not a confounder, only a covariate that modulates the local effect of T, and the backdoor criterion holds trivially and gives rise to the LOCATE in Equation (4). Thus, C provides multiple "views" of the same regional outcome and thereby constrains the shared mechanism $f_{\boldsymbol{\theta}}(\cdot)$.

Confounder: $C \to T$, $C \to Y$, $T \to Y$. When the context C also drives whether (or by how much) a region is treated, it becomes a confounder. The backdoor criterion then requires conditioning on C. Treating T as exogenous yields biased estimates. Modeling the treatment-assignment mechanism at HR is what makes the adjustment correct. Exp. 5 (Appendix G) instantiates this motif and shows that ignoring the dependence inflates the MSE on local effects. A concrete real-world example to illustrate this problem is a wildfire-prevention program: forested counties with dry summers are more likely to receive funding, and those same counties also have higher baseline fire incidence, so the program's effect is conflated with the underlying risk.

Collider: $T \to C, Y \to C$. A collider is a variable influenced by both the treatment and the outcome. Conditioning on such a variable opens a spurious path between T and Y, so naïvely passing C as an input to $f_{\boldsymbol{\theta}}(t, c)$ would bias the estimated effect. Nevertheless, HR collider information may still be useful: since C is partly determined by the HR data $y_{i,j}$, it can provide fine-grained information for the reconstruction of Y or serve as a diagnostic check. For example, consider a regional vaccination campaign (T), where the outcome of interest is infection rates (Y), and the HR variable C is neighborhood-level pharmacy purchases of rapid tests. The campaign may increase public attention and test purchases through its messaging, while actual local infections also increase demand for rapid tests.

A possible way to exploit this structure is to invert the usual roles: instead of using C as a covariate of the outcome, we treat C itself as the HR target and learn a local mechanism $f^{\text{inv}}(t, y) = c$ that maps treatment and outcome to the collider. Under this setup, CLAM's aggregation-consistency principle still applies, but it is used to jointly learn the mechanism and the disaggregation of T and Y against the observed HR values of C.

Mediator: $T \to C, C \to Y, T \to Y$. A mediator C lies on the causal pathway from T to Y, so part of the treatment effect is transmitted through C rather than acting directly. Naïve estimation that conditions on C cannot separate the direct effect of T from the indirect effect through C. When the high-resolution mediator is observed, one can instead jointly learn the treatment-to-mediator, mediator-to-outcome, and direct treatment-to-outcome mechanisms. Only the two outcome-side mechanisms rely on the aggregated regional observations, since the treatment-to-mediator mechanism is supervised directly by the observed $c_{i,j}$. This enables a decomposition of the learned effect into a direct component $T \to Y$ and an indirect component $T \to C \to Y$. A concrete example is a regional advertising campaign observed at low resolution, whose effect on regional sales, also observed at low resolution, is partly direct and partly mediated by word-of-mouth dynamics reflected in local online discourse observed at high resolution.

Instrumental variable: $C \to T, T \to Y, U \to T, U \to Y$. If a hidden confounder links treatment and outcome, but the HR covariate acts on the outcome only through the treatment, that covariate may play the role of an instrument. An HR instrument is unusual but plausible. Consider subregional weather (C, HR), which influences whether a vaccination campaign is deployed (T, observed at LR), but does not affect the infection rate (Y, observed at LR) in any other way. One possible extension of CLAM would then be to disaggregate treatment and outcome jointly, as $y_{i,j} = f_2(f_1(c_{i,j}))$, where $f_1(\cdot)$ maps weather to local deployment and is trained against the LR treatment totals, and $f_2(\cdot)$ maps deployment to local infections and is trained against the LR outcome totals. In such a construction, $f_2(\cdot)$ would only see the deployment predicted from the weather, which may help isolate the component of treatment variation attributable to the instrument rather than the hidden confounder.

Spill-over. A subregional outcome may depend on neighboring treatments and contexts: interventions can spill over across space. The local mechanism is then no longer purely local, and aggregated observations alone cannot distinguish a strong local effect from a weaker effect that diffuses outward. HR context can help by exposing spatial structure. Technically, one can replace the pointwise $f_{\boldsymbol{\theta}}(\cdot)$ with a grid-based model. In principle, the model can learn both direct local effects and spill-over patterns from aggregated outcomes. A concrete example is a driving ban in one district that displaces traffic into adjacent ones, so regional air quality depends on both the ban location and surrounding road and vegetation patterns.

Time-Varying Confounding. In spatiotemporal settings, confounders may evolve over time and be shaped by past treatments. Present context can then no longer be treated as a fixed pre-treatment covariate, because it may already encode earlier interventions. This makes estimation harder: naïve adjustment can block part of the causal pathway or introduce bias (Liley et al., 2021). CLAM can be extended by indexing $f_{\boldsymbol{\theta}}(\cdot)$ over time and by explicitly modeling how past treatments, outcomes, and HR covariates affect the current state. Exp. 3 already uses temporal variation to recover a stable latent spatial factor; a natural next experiment would make this factor dynamic and treatment-dependent. For example, neighborhoods with high prior infection rates may receive additional public-health resources, which then change later vulnerability and influence future allocation decisions.

Summary. Across all motifs, the role of HR data can be summarized as follows: it provides *variation within regions* that translates into observable differences at the aggregated level. This variation can be exploited to constrain and identify fine-grained causal mechanisms, but only when the causal graph permits it.

More concretely: (i) for effect modifiers and confounders, HR structure is highly informative and can be directly leveraged for identification and adjustment; (ii) for mediators, it remains informative but requires explicitly modeling the underlying causal pathways; (iii) for colliders, it can be actively misleading if the framework is not changed accordingly.

CLAM does not replace causal identification via do-calculus: the causal graph determines whether a target effect is identifiable in principle, while CLAM estimates the corresponding fine-grained structure from aggregated data. Identifiability in principle, however, does not guarantee recovery in practice, which further requires sufficient variability in the HR covariates, a well-behaved optimization problem, and effective training of the chosen function class. See Appendix A and Section 5.

D Extended Related Work

Our work builds on and extends three main strands of research: spatial disaggregation, causal inference under aggregation, and causal abstraction.

Statistical Downscaling and Disaggregation. Statistical downscaling methods aim to infer high-resolution estimates from coarse data, using spatial interpolation, small area estimation (Fay & Herriot, 1979), or learning-based techniques. These have been widely applied in climate and environmental sciences (Maraun, 2019; Peng et al., 2017; Gotway & Young, 2002), including recent work on probabilistic downscaling using Gaussian processes (Chau et al., 2021; Law et al., 2018). However, such methods typically model associational patterns and do not support interventional or counterfactual reasoning.

To address this, some recent approaches integrate causal reasoning into downscaling pipelines, e.g., by identifying causally relevant predictors (Vázquez-Patiño et al., 2022; Dutta & Maity, 2020). Yet, these still rely on observable variables and assume access to relatively fine-grained information. In contrast, our approach estimates high-resolution *causal effects* using only aggregated outcome data and high-resolution covariates.

The challenges of reasoning with aggregate data are well-documented. Classical results on the ecological fallacy show that aggregated statistics may obscure or even invert causal relationships (Robinson, 1950; Freedman, 1999; King, 1997). Earlier work on aggregation and disaggregation in optimization similarly highlights the complexity introduced when latent heterogeneity is present but unobserved (Rogers et al., 1991). CLAM explicitly addresses these challenges by modeling the aggregation process and learning disentangled causal effects at the subregional level.

Causal Deabstraction. Structurally, our approach is inspired by recent work on causal abstraction, which studies mappings between causal models at different levels of granularity (Beckers & Halpern, 2019; Beckers et al., 2020; Zennaro, 2022). These works formalize when such mappings preserve counterfactual or interventional semantics, but do not address learning disaggregated effects from data. Finally, our work relates to spatiotemporal causal discovery (Runge et al., 2019), but differs by operating in a setting where temporal data is limited and only aggregate observations are available.

Invariant Causal Mechanisms. Our work also relates to the principle of *invariant causal mechanisms*, which posits that the functional relationships between variables remain stable across different environments or interventions (Peters et al., 2016; Schölkopf et al., 2021). This idea has been used for causal discovery (Heinze-Deml et al., 2018) and domain generalisation (Rojas-Carulla et al., 2018), and is particularly relevant for spatial disaggregation where intervention effects may vary locally but share stable dependencies on contextual covariates. CLAM incorporates this principle by enforcing that the learned local causal mechanism $f_{\boldsymbol{\theta}}(\cdot)$ is shared across all subregions while allowing spatial heterogeneity to emerge through contextual variables.

Graph Neural Networks. Conceptually, our architecture is related to message-passing graph neural networks (Gilmer et al., 2017). In a GNN, pairwise messages are computed and aggregated within

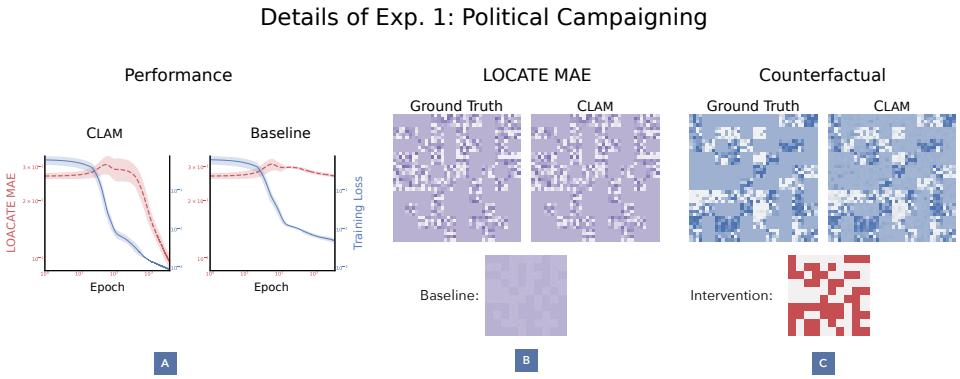


Figure 5: Evaluation of the learned outcomes, causal effect estimates, counterfactual predictions, and training behavior.

a single update step. Similarly, our model learns an update function $f_{\boldsymbol{\theta}}(\cdot)$ that operates on ordered inputs, and optionally an aggregation function $g_{\boldsymbol{\phi}}(\cdot)$ that is permutation-invariant, akin to a Deep Sets model (Zaheer et al., 2017). As in GNNs, we share the parameters of the structural function across nodes or regions.

Hierarchical Bayesian Model. The design also parallels hierarchical Bayesian models (Feller & Gelman, 2015; Dias et al., 2013), where group-specific effects are drawn from a shared prior distribution and information is partially pooled across groups. In CLAM, the local mechanism $f_{\boldsymbol{\theta}}(\cdot)$ plays the role of the shared prior, and contextual covariates capture structured variation across subregions, allowing data-sparse subregions to benefit from patterns learned in others.

Spatiotemporal Causality. Spatiotemporal causal inference has received growing attention as researchers seek to understand how interventions and their effects evolve across both space and time. A recent work presents a causal framework for studying the interplay between conflict and forest loss in Colombia, emphasizing the need to model spatial and temporal dependencies jointly (Christiansen et al., 2022). Extending this perspective, methods proposed by Zhou et al. (2024) estimate heterogeneous treatment effects in settings where economic assistance moderates the impact of airstrikes on insurgent violence, while approaches introduced by Mukaigawara et al. (2025) accommodate arbitrary spill-over and carryover effects across space and time. Network-based settings with temporal dynamics have also been explored. Song & Papadogeorgou (2024) study bipartite causal inference with interference in time series data, and Balkus et al. (2024) study cases of modified treatment policies. Recent methodological innovations include GST-UNet (Oprescu et al., 2026), a deep learning architecture for spatiotemporal causal inference under time-varying confounding, and consistency-guided aggregation for causal discovery in multivariate spatiotemporal data introduced by Ninad et al. (2025). Together, these works highlight the diversity of approaches to spatiotemporal causality, spanning structural models, interference-aware estimation, and machine learning-driven inference.

E Details of Main Experiments

E.1 Exp. 1: Political Campaigning

Context. We consider a simple political campaigning setting in which a candidate either spends campaign resources in a region or does not. The treatment is binary at the regional level and is inherited by all subregions. The outcome is the candidate's relative improvement, but it is observed only as a regional aggregate. The effect of campaigning depends on subregional context, here interpreted as standardized wealth. Thus, two regions can have similar aggregate context and similar aggregate outcomes while still containing different subregional causal effects.

Setup. We use $N = 100$ regions arranged on a 10×10 grid. Each region contains $M = 16$ subregions arranged on a 4×4 grid. The treatment matrix is $T \in \{0, 1\}^{N \times M}$. Exactly half of

the regions are treated, and within a treated region all subregions have treatment value 1; within a control region all subregions have treatment value 0.

The context matrix $C \in \mathbb{R}^{N \times M}$ contains one continuous contextual value per subregion. We sample the entries from a standard Gaussian distribution and then subtract the mean within each region, so that each row of C has mean zero. This makes the regional average context uninformative by construction. We then randomly swap entries across the matrix to introduce additional variation while preserving the overall structure. The observed outcome is generated at the subregional level and then aggregated to the region level by averaging.

The ground-truth outcome function is

$$f(t, c) = \begin{cases} 1, & t = 0, \\ \mathrm{interp}(c), & t = 1, \end{cases}$$

where $\mathrm{interp}(c)$ is the piecewise linear interpolation through the anchor points $\{(-2, 2), (0, 1), (1, 0)\}$.

Thus, untreated subregions have outcome 1, while treated subregions have an outcome that depends nonlinearly on the local context c. The noisy subregional outcome is

$$y_{i,j} = f(t_{i,j}, c_{i,j}) + \epsilon_{i,j}, \qquad \epsilon_{i,j} \sim \mathcal{N}(0, 0.1^2),$$

and the observed regional outcome is

$$\widehat{Y}_i = \frac{1}{M} \sum_{j=1}^{M} y_{i,j}.$$

The ground-truth local causal effect, which we use only for evaluation, is

$$E_{i,j} = f(t_{i,j}, c_{i,j}) - f(0, c_{i,j}).$$

Training. We train a small neural network $f_{\boldsymbol{\theta}}(t, c)$ with two inputs, treatment and context, and one scalar output. The model has two hidden layers with 32 hidden units each, ReLU activations, dropout with probability 0.05, and is trained with AdamW using learning rate 10^{-3} and weight decay 10^{-3} for 4000 epochs. Importantly, CLAM is trained only from the regional aggregate signal. For each region, we average the model's subregional predictions and minimize

$$\mathcal{L}_{\text{CLAM}} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{1}{M} \sum_{j=1}^{M} f_{\boldsymbol{\theta}}(t_{i,j}, c_{i,j}) - \widehat{Y}_i \right)^2 .$$

Baseline. We compare against a uniform disaggregation baseline. Since the true subregional outcomes are not observed, this baseline assigns the same regional outcome to every subregion,

$$\widetilde{y}_{i,j} = \widehat{Y}_i.$$

It then trains the same neural network architecture using an element-wise loss,

$$\mathcal{L}_{\text{base}} = \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} \left(f_{\boldsymbol{\theta}}(t_{i,j}, c_{i,j}) - \widetilde{y}_{i,j} \right)^2 .$$

This gives the baseline a dense subregional loss signal, but the signal is based on a uniformity assumption that removes the within-region heterogeneity needed to recover local causal effects.

Evaluation. We evaluate whether the learned model can recover the fine-grained causal structure from aggregate observations (Figure 5). We report the training loss and LOCATE MAE (cf. Eq. equation 4) over training (Figure 5a). Confidence intervals (95%) are bootstrapped and based on 20 runs. As expected, both decrease for CLAM, while the baseline cannot learn the causal mechanism. We also report the full LOCATE matrix, comparing the ground truth with the one inferred by CLAM, in Figure 5b. The recovered causal effect is almost indistinguishable from the ground truth, while the baseline, as expected, cannot reconstruct the underlying disaggregation. Lastly, we construct a counterfactual treatment matrix (Figure 5c) by inverting the observed treatment assignment, $T^{\text{cf}} = 1 - T$, and compare the ground-truth counterfactual outcome $f(T^{\text{cf}}, C)$ with the neural-network estimate $f_{\boldsymbol{\theta}}(T^{\text{cf}}, C)$. This tests whether the learned mechanism supports coherent counterfactual prediction beyond the observed treatment assignment.

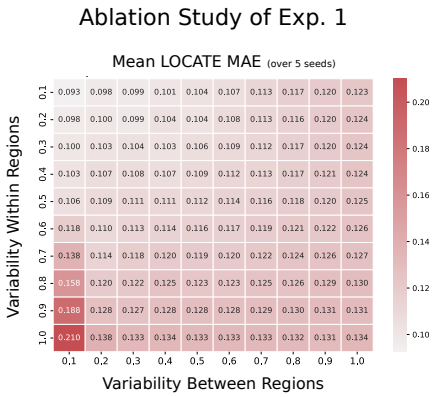


Figure 6: We evaluate the effect of varying the variance of regional context means and, for each fixed context mean, the inter-regional variance. Results are averaged over 5 random seeds for each parameter combination, and uncertainty bounds represent the corresponding variability.

Ablation. To test when this recovery is possible, we repeat the experiment while varying two properties of the context matrix. The first is the standard deviation of the regional context means, denoted σ_{reg}. The second is the standard deviation of context values within each region, denoted σ_{subreg}. For each pair $(\sigma_{\text{reg}}, \sigma_{\text{subreg}})$, we run the experiment over multiple random seeds and average the final LOCATE MAE. This ablation shows that high variability within regions is associated with a smaller error.

E.1.1 Ablation of Exp. 1: Political Campaigning

Our setting targets low-variability aggregate regimes, where regional outcomes are nearly constant and the inverse problem is underdetermined: many fine-grained causal explanations can yield indistinguishable regional observations. The key idea is that high-dimensional subregional covariates can restore information by exposing the shared causal mechanism $f_{\boldsymbol{\theta}}(\cdot)$ to diverse contextual compositions within each aggregate unit. Exp. 1 and the corresponding ablation support this mechanism: as subregional heterogeneity increases, recovery of the true causal effect improves, even when aggregate outcomes remain similarly uninformative. By contrast, varying regional-level heterogeneity has a weaker effect, indicating that the gains do not come primarily from more variable aggregate outcomes, but from richer within-region contextual structure. See Figure 6 for details. Thus, causal deabstraction becomes possible in otherwise weakly identified settings when subregional covariate variation, together with an invariant causal mechanism, provides enough compositional diversity to distinguish the underlying fine-grained causal function.

E.1.2 Ecological Disaggregation Baselines

Standard disaggregation methods are not direct competitors of CLAM, because each of them solves only one half of the problem. This section makes that precise. The setting is the one of Exp. 1: we observe the regional treatment $\widehat{t}_i \in \{0, 1\}$, which is inherited by all subregions of region i, the HR context $c_{i,j}$, and the regional outcome $\widehat{Y}_i$. The subregional outcomes $y_{i,j}$ are never observed. The quantity of interest is the local effect $E_{i,j}$ from Equation (4), and we score every method by the mean absolute error to the ground-truth LOCATE matrix.

Both baselines split the task into two separate steps.

Bicubic Interpolation. This baseline first disaggregates and then learns. The regional outcomes are arranged as a coarse 10×10 image and interpolated up to the 40×40 grid of subregions. This invents an outcome value for every subregion,

$$\widetilde{y}_{i,j} = \text{bicubic}\big(\widehat{Y}\big)_{i,j}, \tag{7}$$

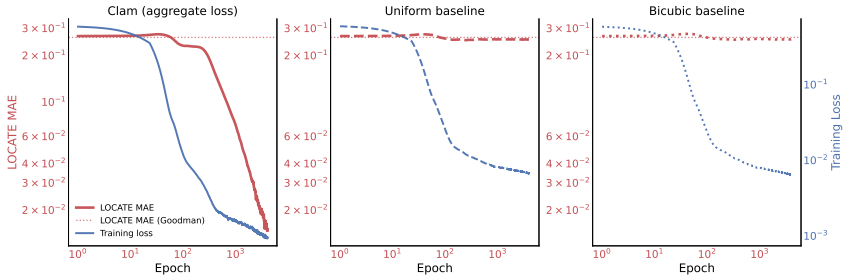


Figure 7: Ecological disaggregation baselines on Exp. 1, one panel per method, on log–log axes. In every panel, red is the LOCATE MAE against the ground truth, which is never used for training, and blue is the training loss that the method actually minimizes. The thin red horizontal line marks the error of Goodman regression, which is computed in closed form and therefore constant. All three trained runs use the same network, initialization, and optimizer, and differ only in the loss. For CLAM, both curves fall together, so fitting the aggregated observations does recover the local effects. For the two baselines, the training loss keeps falling while the LOCATE MAE flattens out at the level that Goodman also reaches.

which is then treated as if it had been measured. The same network as in CLAM is fitted to these values with an element-wise loss,

$$\mathcal{L}_{\text{bicubic}} = \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} \big(f_{\boldsymbol{\theta}}(t_{i,j}, c_{i,j}) - \widetilde{y}_{i,j} \big)^2, \tag{8}$$

and the local effect is read off as $\widehat{E}_{i,j} = f_{\boldsymbol{\theta}}(1, c_{i,j}) - f_{\boldsymbol{\theta}}(0, c_{i,j})$. The uniform baseline above is the special case in which no interpolation is applied, that is, $\widetilde{y}_{i,j} = \widehat{Y}_i$.

Goodman Regression. This baseline (Goodman, 1953) first learns and then disaggregates. The causal analysis is run entirely at the regional level, by regressing the regional outcome on the regional treatment and the regional mean context $\overline{c}_i = \frac{1}{M} \sum_{j=1}^{M} c_{i,j}$ using least squares,

$$\widehat{Y}_i \;\approx\; \beta_0 + \beta_T \, \widehat{t}_i + \beta_c \, \overline{c}_i. \tag{9}$$

The estimated treatment coefficient is then simply handed down to every treated subregion,

$$\widehat{E}_{i,j} = \beta_T \, t_{i,j}. \tag{10}$$

Why Both Hit the Same Error Floor. Neither method uses $c_{i,j}$ to explain variation *within* a region. Bicubic interpolation only knows *where* a subregion lies, but in Exp. 1 the effect varies with context, and the context is spatially unstructured by construction. The invented outcome $\widetilde{y}_{i,j}$ therefore carries almost no information about $c_{i,j}$, and fitting $f_{\boldsymbol{\theta}}(\cdot)$ to it yields an almost constant effect estimate. Goodman regression is constant by construction, since one regional coefficient is copied to all treated subregions. In both cases the remaining error is the spread of the true local effect. The regional coefficient β_T still captures the average effect among the treated reasonably well; it is only the heterogeneity that is lost. Figure 7 shows both failure modes.

E.2 EXP. 2: PUBLIC SCHOOL FUNDING VS IMPROVEMENT IN OUTCOMES

Context. We examine how public spending on schools affects educational outcomes when the exact locations of the expenditures are *unknown*. Each administrative region contains multiple subregions, interpreted here as school districts, but only one district in a treated region receives a funding intervention. At the regional level, we only observe an aggregated treatment indicator and the aggregated outcome, not the specific district that received the intervention. The goal is to infer the high-resolution intervention locations from aggregated data.

The subregional context consists of high-resolution demographic and economic data, represented here by a normally distributed *wealth score*. We hypothesize that the effect of school funding depends on this wealth score through a global shift and scale.

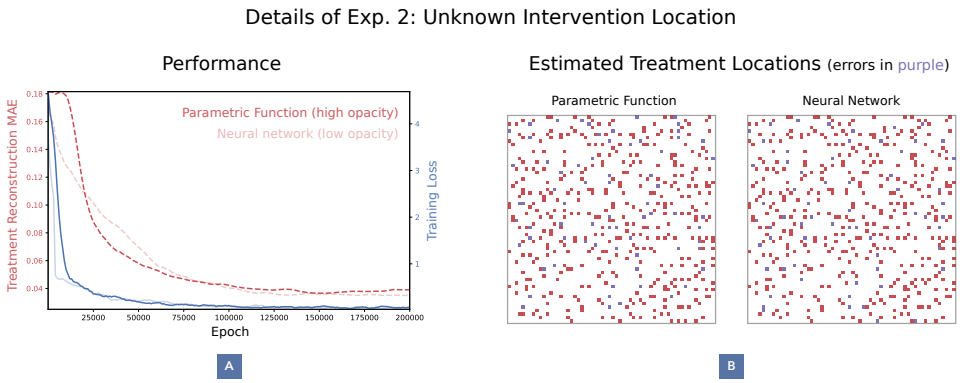


Figure 8: Performance of the training loss and MSE of the local-effect estimation, alongside ground-truth causal effects, estimated causal effects, and the corresponding estimation errors across subregions.

Setup. We assume $N = 30 \times 30 = 900$ regions, each with $M = 2 \times 2 = 4$ subregions.

We simulate:

- A high-resolution intervention matrix $T \in \mathbb{R}^{N \times M}$, where $t_{i,j}$ indicates whether subregion j of region i receives the funding intervention.
- A regional intervention vector $\widehat{T} \in \mathbb{R}^N$, where $\widehat{T}_i = \sum_{j=1}^{M} t_{i,j}$ is the absolute intervention intensity in region i.
- A high-resolution context matrix $C \in \mathbb{R}^{N \times M}$ encoding the socioeconomic status of each subregion, where smaller $c_{i,j}$ values represent lower socioeconomic status and larger $c_{i,j}$ values represent higher socioeconomic status.
- A high-resolution noise matrix $\epsilon \in \mathbb{R}^{N \times M}$ with i.i.d. entries $\epsilon_{i,j} \sim \mathcal{N}(0, \sigma^2)$, where $\sigma^2 = 0.005$.

The intervention matrix T is generated by:

1. Initializing all entries to zero.
2. Randomly selecting 50% of the regions for treatment.
3. For each treated region i, choosing exactly one subregion $j \in \{1, \ldots, M\}$ uniformly at random and assigning $t_{i,j} = 1$.

The context matrix C is sampled i.i.d. from a standard normal distribution:

$$c_{i,j} \sim \mathcal{N}(0, 1).$$

The outcome variable is given by the ground-truth functional relationship

$$y_{i,j} = f_{\boldsymbol{\theta}}(t_{i,j}, c_{i,j}) + \epsilon_{i,j} = (\theta_{\text{shift}} - c_{i,j}) \cdot \theta_{\text{scale}} \cdot t_{i,j} + \epsilon_{i,j},$$

with ground-truth parameters

$$\theta_{\text{shift}} = 4.2, \qquad \theta_{\text{scale}} = 1.23.$$

The observed regional outcome is aggregated by summation:

$$\widehat{Y}_i = \sum_{j=1}^{M} y_{i,j}.$$

As in Exp. 1, we assume no hidden confounding, no time-series effects, treatment assignment independent of context, and aggregation-based regional observations.

Training. We train a model to estimate both the latent subregional intervention assignments and the parameters of the causal effect function.

The model has:

- $N \times M = 900 \times 4$ latent treatment logits, one vector of length M for each region.
- Two scalar parameters $\{\theta_{\text{shift}}, \theta_{\text{scale}}\}$ specifying the parametric form of $f_{\boldsymbol{\theta}}(\cdot)$.

The learned causal effect function is parametrized as

$$f_{\boldsymbol{\theta}}(t_{i,j}, c_{i,j}) = (\theta_{\text{shift}} - c_{i,j}) \cdot \theta_{\text{scale}} \cdot t_{i,j}.$$

To ensure a positive scale parameter, θ_{scale} is represented through a softplus transformation.

An auxiliary preprocessing function is applied to the latent treatment logits using the Gumbel-Softmax relaxation. This produces a soft row-wise assignment over the M subregions. The resulting assignment is processed so that untreated rows remain all-zero and treated rows have approximately one active subregion. During training, the Gumbel-Softmax temperature is exponentially annealed from 2.0 to 0.1.

In the forward pass:

1. Sample a mini-batch of region indices.
2. Process the latent treatment logits using the Gumbel-Softmax relaxation to obtain an estimate of the high-resolution treatment matrix T for the mini-batch.
3. Compute predicted subregional outcomes:
$$\mu_{i,j} = f_{\boldsymbol{\theta}}\big(\widehat{t}_{i,j}, c_{i,j}\big)\,.$$
4. Aggregate to regional predictions by summation:
$$\mu_i = \sum_{j=1}^{M} \mu_{i,j}.$$
5. Compare the regional prediction with the observed regional outcome $\widehat{Y}_i$ using the MSE loss.

We optimize all parameters jointly using AdamW with learning rate 10^{-4}. Training is performed for $200{,}000$ epochs with a fixed random seed.

Neural Network. Additionally, we train a neural-network variant following the same treatment-assignment procedure. In this version, the causal effect function $f_{\boldsymbol{\theta}}(\cdot)$ is implemented as a feed-forward neural network taking $(t_{i,j}, c_{i,j})$ as input.

The network has three hidden layers with hidden dimension 16 and $\tanh$ activations:

$$(t_{i,j}, c_{i,j}) \mapsto \text{MLP}_{\boldsymbol{\theta}}(t_{i,j}, c_{i,j}).$$

To stabilize the local-effect interpretation, the model uses the centered form

$$f_{\boldsymbol{\theta}}(t_{i,j}, c_{i,j}) = \text{MLP}_{\boldsymbol{\theta}}(t_{i,j}, c_{i,j}) - \text{MLP}_{\boldsymbol{\theta}}(0, c_{i,j}),$$

which enforces $f_{\boldsymbol{\theta}}(0, c_{i,j}) \approx 0$. The latent treatment indicators are processed in the same way as in the parametric model, using Gumbel-Softmax with temperature annealing from 2.0 to 0.1.

The neural-network model is also trained for $200{,}000$ epochs using AdamW with learning rate 10^{-4} and mini-batches containing 20% of the regions.

Results. Figure 8a shows the evolution of the mini-batch training loss and the quality of the treatment-location reconstruction (measured as MAE). Both the neural network and the parametric function recover the locations reasonably well. We however find that this depends on the noise level and on proper temperature annealing. The final reconstructed locations are shown in Figure 8b, where incorrect predictions are shown in purple.

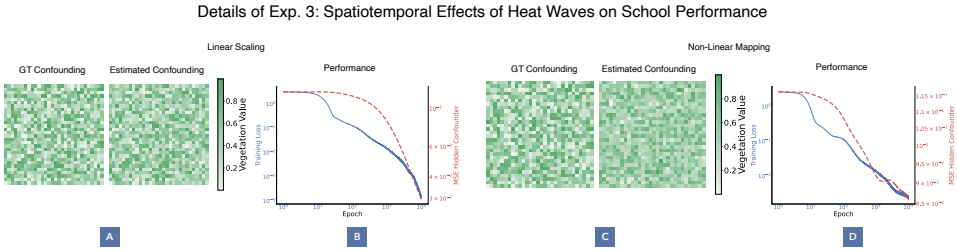


Figure 9: Ground-truth vegetation index, the corresponding estimated latent effect-modifier fields, and joint plots of training loss and vegetation MSE. Results are shown for two settings: linear scaling (left) and MLP learning (right).

E.3 Exp. 3: Spatiotemporal Effects of Heat Waves on School Performance

Context. In this experiment, we examine the impact of extreme heat events ("heat waves") on educational outcomes across both space and time. We model a spatial grid of $N = 100$ regions, each containing $M = 9$ subregions, observed over $W = 48$ consecutive months. The intervention represents the occurrence of a heat wave in a given subregion and month. Heat waves have a detrimental effect on school performance, but the strength of this effect depends on the educational background of parents in the subregion and on an unobserved environmental factor, vegetation coverage, which mitigates the impact of heat.

The observed context for each subregion is a categorical indicator of parents' education level, taking one of three values (*low* = 1, *medium* = 2, *high* = 3), which evolves slowly over time to reflect gradual demographic changes. The unobserved context is a fixed vegetation index between 0 and 1 for each subregion, representing the proportion of green cover and its protective effect against heat. The intervention process is binary at the subregional level, with most regions remaining unaffected in a given month, but some experiencing heat waves simultaneously in all their subregions, corresponding to large-scale weather patterns.

The outcome, measured as a monthly change in school performance for each subregion, is driven by the interaction between the intervention, parents' education level, and the vegetation coverage. Subregions with lower parental education are more strongly affected, while vegetation dampens the negative effect. The challenge in this experiment is to recover both the mapping from context, intervention, and unobserved effect modifier to outcomes, and the vegetation values themselves, given only aggregated regional-level outcomes over time.

Setup. We model $N = 100$ regions, each with $M = 9$ subregions, observed over $W = 48$ discrete time points (months).

For each month $w \in \{1, \dots, W\}$, we define:

- A high-resolution binary treatment matrix $T^{(w)} \in \{0, 1\}^{N \times M}$, where $t_{i,j}^{(w)} = 1$ indicates that subregion j of region i is experiencing a heat wave in month w. In our data generation, each row of $T^{(w)}$ is either all zeros (probability 0.7) or all ones (probability 0.3), corresponding to large-scale heat events affecting entire regions.
- An observed context matrix $C^{(w)} \in \{1, 2, 3\}^{N \times M}$ encoding the categorical education level of parents in each subregion. At $w = 1$, these values are sampled uniformly. For $w > 1$, $C^{(w)}$ is obtained from $C^{(w-1)}$ by flipping each entry to a random category with probability 0.05, capturing slow demographic change.
- An unobserved covariate context matrix $U \in [0, 1]^{N \times M}$ containing fixed vegetation index values for each subregion, drawn uniformly from $[0, 1]$ and constant across all months.
- A noise matrix $\epsilon^{(w)} \in \mathbb{R}^{N \times M}$ with i.i.d. entries $\epsilon_{i,j}^{(w)} \sim \mathcal{N}(0, \sigma^2)$, with $\sigma^2 = 0.02$.

The subregional outcome is generated as:

$$y_{i,j}^{(w)} = f_{\boldsymbol{\theta}}\left(t_{i,j}^{(w)}, c_{i,j}^{(w)}, u_{i,j}\right) + \epsilon_{i,j}^{(w)},$$

where $c_{i,j}^{(w)}$ is the parents' education category, $u_{i,j}$ is the vegetation index, and:

$$f_{\boldsymbol{\theta}}\Big(t_{i,j}^{(w)}, c_{i,j}^{(w)}, u_{i,j}\Big) = \begin{cases} 0 & \text{if } t_{i,j}^{(w)} = 0, \\ \big(10 \cdot \mathbb{1}[c_{i,j}^{(w)} = 1] + 5 \cdot \mathbb{1}[c_{i,j}^{(w)} = 2] \\ \quad + \ \mathbb{1}[c_{i,j}^{(w)} = 3]\big) \cdot (1 - u_{i,j}) & \text{if } t_{i,j}^{(w)} = 1. \end{cases}$$

This function specifies that the subregional outcome depends on whether a heat wave occurs and, if so, how vulnerable the subregion is based on the education level of parents and the vegetation coverage. If $t_{i,j}^{(w)} = 0$, the effect is zero. If $t_{i,j}^{(w)} = 1$, the impact is determined by the categorical education level with coefficients $(10, 5, 1)$ corresponding to low, medium, and high parental education, respectively, and is scaled by $(1 - u_{i,j})$, which reduces the effect in proportion to the amount of vegetation present. The indicator $\mathbb{1}[\cdot]$ selects the appropriate coefficient for each subregion.

The regional outcome at month w is obtained via mean aggregation. This setup induces a spatiotemporal causal problem with heterogeneous treatment effects driven by both observed and unobserved context, where the latter must be recovered from aggregated outcomes across multiple time points.

Training. We train a model to jointly estimate:

1. The mapping $f_{\boldsymbol{\theta}}(\cdot)$ from observed context, intervention, and vegetation index to subregional outcomes,
2. The static vegetation index values $U \in [0, 1]^{N \times M}$ for all subregions.

The function $f_{\boldsymbol{\theta}}(\cdot)$ is implemented as a 4 layer multilayer perceptron (MLP) with hidden dimension 16 and a dropout rate of 0.1 after the second hidden layer. Its inputs for each subregion are:

$$z_{i,j}^{(w)} = \big(\mathbb{1}[c_{i,j}^{(w)} = 1],\ \mathbb{1}[c_{i,j}^{(w)} = 2],\ \mathbb{1}[c_{i,j}^{(w)} = 3],\ t_{i,j}^{(w)},\ u_{i,j}\big),$$

where $u_{i,j}$ is the vegetation index. The output is the predicted outcome $y_{i,j}^{(w)}$ for that subregion and month.

In the forward pass for a given month w:

1. The model takes as input $T^{(w)}$, $C^{(w)}$, and U,
2. The MLP evaluates $f_{\boldsymbol{\theta}}(\cdot)$ at each (i, j) to produce predicted subregional outcomes $y_{i,j}^{(w)}$,
3. These are aggregated to the regional level via mean aggregation.
4. The primary loss is the MSE between predicted and observed regional outcomes.

Training proceeds over all months in random order at each epoch. The total parameter set consists of the MLP weights $\boldsymbol{\theta}$ and the vegetation matrix U. We optimize using Adam with a learning rate of 0.001 for $10{,}000$ epochs, with fixed random seeds for reproducibility. Loss curves for $\mathcal{L}_{\text{region}}$ and the MSE of $\mathcal{L}_{\text{veg}}$ are recorded throughout training.

Results.

1. Figure 9a shows the loss curves and vegetation MSE under the linear scaling parameterization. Both training loss and MSE decrease rapidly, and the final error is very small, indicating near-perfect recovery of the unobserved covariate.
2. Figure 9b compares the ground-truth vegetation field to the estimated vegetation under linear scaling. The two maps are almost indistinguishable, confirming that the restricted functional form matches the data-generating process and allows highly accurate recovery.
3. Figures 9c and d present the same results for the non-linear MLP parameterization. While the model still recovers meaningful structure in the vegetation, the MSE remains orders of magnitude higher than in the linear case, and deviations from the ground truth are clearly visible in the estimated field. This is notable since the true vegetation values are not necessarily identifiable; the MLP could instead store a transformed version and map them back during the forward process.

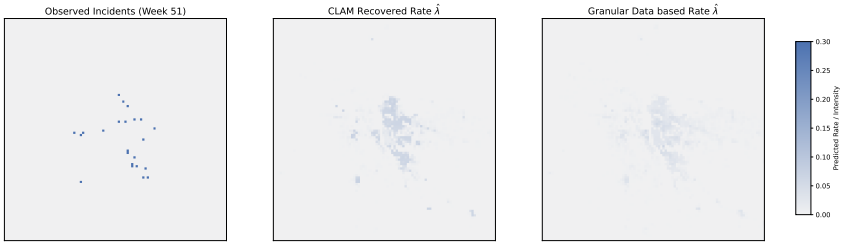


Figure 10: Comparison of learned spatial intensities for week 51. Both models are rendered on a shared color scale across the $\sqrt{N \cdot M} \times \sqrt{N \cdot M}$ grid, alongside the observed events for that week.

4. Overall, these results demonstrate that the model can reconstruct regional outcomes while also uncovering hidden drivers of heterogeneity. The comparison between linear and non-linear parameterizations highlights the trade-off: structural restrictions yield more accurate recovery when they match the true mechanism, while flexible approximators such as MLPs still learn the vegetation but with reduced precision.

F Real-World Case Study Details: Effect of Heat on Gun Violence

This section provides the full details of the real-world study summarized in Section 4.2. The notation follows Section 2: regions are indexed by $i \in \{1, \ldots, N\}$, subregions or cells by $j \in \{1, \ldots, M\}$, and weeks by $w \in \{1, \ldots, W\}$. Temperature is used as a continuous, time-varying treatment-like exposure; urbanization is the HR static contextual covariate; and gun-incident counts are the observed LR outcomes.

Problem Setting. The target variable is observed at a coarse spatial resolution, while covariates are available at a finer resolution. The study area is discretized into a fine grid of $\sqrt{N \cdot M} \times \sqrt{N \cdot M}$ cells, grouped into $\sqrt{N} \times \sqrt{N}$ coarse regions. With $N = 100$ and $M = 100$, this gives a 100×100 cell grid grouped into a 10×10 region grid. For each week w and region i, we observe a nonnegative count outcome

$$Y_{i,w}^{\mathrm{LR}} \in \mathbb{Z}_{\geq 0}, \tag{11}$$

representing the number of gun-incident events occurring in region i during week w.

Data Representation and Covariates. Let $c_{i,j,w}$ denote the observed urbanization measurement for cell (i, j) and week w. Since urbanization is treated as time-invariant in this study, we use its temporal mean,

$$c_{i,j} = \frac{1}{W} \sum_{w=1}^{W} c_{i,j,w}. \tag{12}$$

Temperature is modeled at the region level. Let $t_{i,j,w}^{\mathrm{HR}}$ denote the cell-level temperature for cell (i, j) and week w. We compute the region-average temperature exposure as

$$t_{i,w}^{\mathrm{LR}} = \frac{1}{M} \sum_{j=1}^{M} t_{i,j,w}^{\mathrm{HR}}, \tag{13}$$

and standardize it using the global mean and standard deviation across all region–week pairs:

$$\widetilde{t}_{i,w} = \frac{t_{i,w}^{\mathrm{LR}} - \mu_t}{\sigma_t + \varepsilon}, \tag{14}$$

where $\varepsilon > 0$ is a small constant for numerical stability. Missing temperature values are imputed using the global median temperature, and missing outcome counts are set to zero.

Disaggregation Model. We model the latent HR intensity, or rate, $\lambda_{i,j,w} > 0$ as a function of the standardized regional temperature exposure and the HR urbanization covariate:

$$\lambda_{i,j,w} = f_{\boldsymbol{\theta}}\big(\widetilde{t}_{i,w},\, c_{i,j}\big)\,, \tag{15}$$

where $f_{\boldsymbol{\theta}}(\cdot)$ is a feed-forward neural network with parameters $\boldsymbol{\theta}$. To enforce nonnegativity, the network output is passed through a softplus transformation.

The model-implied LR expected count is obtained by summing the HR rates over all cells in region i:

$$\mu_{i,w} = \sum_{j=1}^{M} \lambda_{i,j,w}. \tag{16}$$

Thus, $\mu_{i,w}$ denotes the predicted regional count intensity, while $Y_{i,w}^{\mathrm{LR}}$ denotes the observed regional count. This additive aggregation is consistent with the SCM in Section 3.1.

Training Objective. We assume a Poisson observation model at the region–week level,

$$Y_{i,w}^{\mathrm{LR}} \sim \mathrm{Poisson}(\mu_{i,w})\,, \tag{17}$$

and train $f_{\boldsymbol{\theta}}$ by minimizing the negative Poisson log-likelihood, up to additive constants:

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{|\mathcal{W}_{\mathrm{train}}| \cdot N} \sum_{w \in \mathcal{W}_{\mathrm{train}}} \sum_{i=1}^{N} \big(\mu_{i,w} - Y_{i,w}^{\mathrm{LR}} \log(\mu_{i,w} + \varepsilon)\big)\,, \tag{18}$$

where $\mathcal{W}_{\mathrm{train}}$ is the set of training weeks and $\varepsilon > 0$ is a small constant for numerical stability.

Spatial Intensity Comparison. Figure 10 visualizes observed incidents for a representative week, week 51, alongside spatial intensity estimates learned under different supervision regimes. When trained only on region-level counts, CLAM produces spatially structured intensity maps that concentrate mass around incident clusters but exhibit noticeably higher peak intensities and broader spatial support compared to the cell-supervised estimate. This overestimation reflects a fundamental ambiguity induced by aggregation: without access to fine-grained outcomes, the model must allocate regional mass across plausible subregional locations, leading to inflated local rates in areas consistent with the observed context and covariates. The response-surface comparison in the main paper, Figure 4, shows the same pattern in covariate space rather than physical space.

Learning. The training procedure optimizes the function parameters $\boldsymbol{\theta}$ (or any other missing components of the SCM). The structural function $f_{\boldsymbol{\theta}}(\cdot)$ can be represented by a neural network or any other parameterized function. Since the fine-grained outcomes $y_{i,j}$ are unobserved, training relies solely on their aggregated counterparts.

We optimize with the goal that the inferred aggregated predictions (μ_i) match the observed aggregate: $\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}} = \sum_{i=1}^{N} \big(\widehat{Y}_i - \mu_i\big)^2$, where $\widehat{Y}_i$ denotes the observed aggregate of region i. If the noise terms are not i.i.d. and normally distributed, they can also be explicitly learned within the optimization loop.

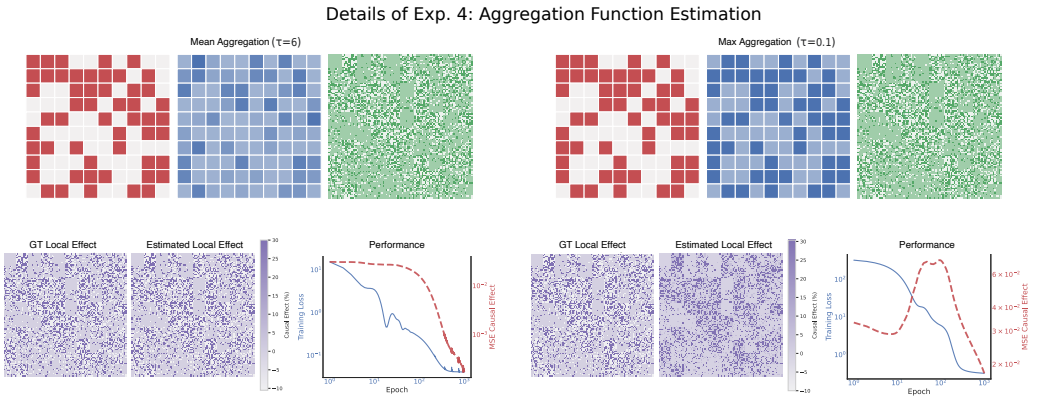


Figure 11: **Results Overview. Top:** Input of LR intervention locations (red), LR outcomes (blue), and HR context (green) for mean aggregation and max aggregation cases. **Bottom:** We observe the estimated local effects and the performance of our model over multiple epochs for both cases.

G Additional Experiments

Exp. 4 considers driving bans with an unknown aggregation function between subregional and regional outcomes. Exp. 5 introduces treatment–context confounding, where the probability of subregional treatment depends on contextual variables, affecting both allocation and outcomes

G.1 Exp. 4: Unknown Aggregation Functions

Context. We study the causal effect of driving bans on air quality under the assumption that the reported region-level air quality values are an unknown aggregation of finer-scale (subregional) measurements. Specifically, we model the aggregation as a learned combination of the subregional mean and maximum values, with the combination weights parameterized via a logit transform.

A realistic example is a city-wide driving ban where air quality is officially reported as a single regional index (e.g., $PM_{2.5}$ concentration), while the underlying measurement network collects data from multiple monitoring stations. Depending on reporting policies, the published index might resemble an average across stations, a worst-case station reading, or something in between.

The high-resolution context for each subregion is a single continuous variable: a vegetation index, capturing the amount of green coverage in the area. Vegetation can mitigate pollution and thus may modulate the effect of a driving ban. We hypothesize that the causal effect of driving bans varies with vegetation and that the unknown aggregation mechanism must be learned alongside the effect parameters to estimate subregional impacts correctly.

Setup. We assume $N = 100$ regions, each with $M = 100$ subregions.

We are given:

- A binary treatment vector $T \in \{0, 1\}^N$, where $T_i = 1$ indicates that a driving ban was implemented in region i (and thus all its subregions).
- A high-resolution context matrix $C \in [0, 1]^{N \times M}$ encoding the vegetation index of each subregion, where $c_{i,j} = 0$ represents minimal vegetation cover and $c_{i,j} = 1$ represents dense vegetation.
- A high-resolution noise matrix $\epsilon \in \mathbb{R}^{N \times M}$ with i.i.d. entries $\epsilon_{i,j} \sim \mathcal{N}(0, \sigma^2)$, where $\sigma^2 = 0.02$.

The treatment vector T is generated using a Bernoulli distribution with probability 0.5, assigning each region to either the intervention or control group at random. The vegetation index values in C are drawn i.i.d. from $\mathrm{Uniform}(0, 1)$.

The subregional outcome variable is given by the ground-truth functional relationship:

$$y_{i,j} = f_{\boldsymbol{\theta}}(t_{i,j}, c_{i,j}) + \epsilon_{i,j},$$

where:

- $t_{i,j} = T_i$ is the treatment status of subregion j in region i (inherited from the region),
- $c_{i,j}$ is the vegetation index,
- $\epsilon_{i,j}$ is Gaussian noise.

The causal effect function is parameterized as:

$$f_{\boldsymbol{\theta}}(t_{i,j}, c_{i,j}) = \begin{cases} 0.0 & \text{if } t_{i,j} = 0, \\ \theta_{\text{base}} + \theta_{\text{veg}} \cdot c_{i,j} & \text{if } t_{i,j} = 1, \end{cases}$$

where θ_{base} captures the baseline effect of a driving ban and θ_{veg} captures how this effect changes with vegetation cover.

Unlike Exp. 1, the regional outcome x_i is not a simple mean over subregions. Instead, we model the aggregation function as:

$$y_i = \sum_{j=1}^{M} p_{i,j}(\tau) \cdot y_{i,j},$$

where:

$$p_{i,j}(\tau) = \frac{\exp(y_{i,j}/\tau)}{\sum_{k=1}^{M} \exp(y_{i,k}/\tau)}$$

is the softmax weight assigned to subregion j with temperature parameter $\tau > 0$. When $\tau \to \infty$, $p_{i,j}$ approaches a uniform distribution (mean aggregation); when $\tau \to 0$, $p_{i,j}$ concentrates on the subregion with the highest outcome (max aggregation).

Training. We train a model to estimate:

1. The parameters $\boldsymbol{\theta} = \{\theta_{\text{base}}, \theta_{\text{veg}}\}$ of the causal effect function $f_{\boldsymbol{\theta}}(\cdot)$,
2. The aggregation temperature parameter $\tau > 0$ governing the softmax-based aggregation from subregional to regional outcomes.

The learned function takes the form:

$$f_{\boldsymbol{\theta}}(t_{i,j}, c_{i,j}) = \begin{cases} 0.0 & \text{if } t_{i,j} = 0, \\ \theta_{\text{base}} + \theta_{\text{veg}} \cdot c_{i,j} & \text{if } t_{i,j} = 1. \end{cases}$$

In the forward pass:

1. Compute predicted subregional outcomes: $y_{i,j} = f_{\boldsymbol{\theta}}(t_{i,j}, c_{i,j})$,
2. Compute subregion-level aggregation weights via the softmax: $p_{i,j}(\tau) = \frac{\exp(y_{i,j}/\tau)}{\sum_{k=1}^{M} \exp(y_{i,k}/\tau)}$,
3. Compute predicted regional outcomes as the expectation under $p_{i,j}(\tau)$: $y_i = \sum_{j=1}^{M} p_{i,j}(\tau) \cdot y_{i,j}$,
4. Compute the loss as the mean squared error with the observed Y_i.

All parameters $\{\theta_{\text{base}}, \theta_{\text{veg}}, \tau\}$ are optimized jointly using Adam with a learning rate of 0.001 for 1000 epochs. We initialize τ to a moderate value (e.g., $\tau = 1.0$) and enforce $\tau > 0$ during training by optimizing its logarithm.

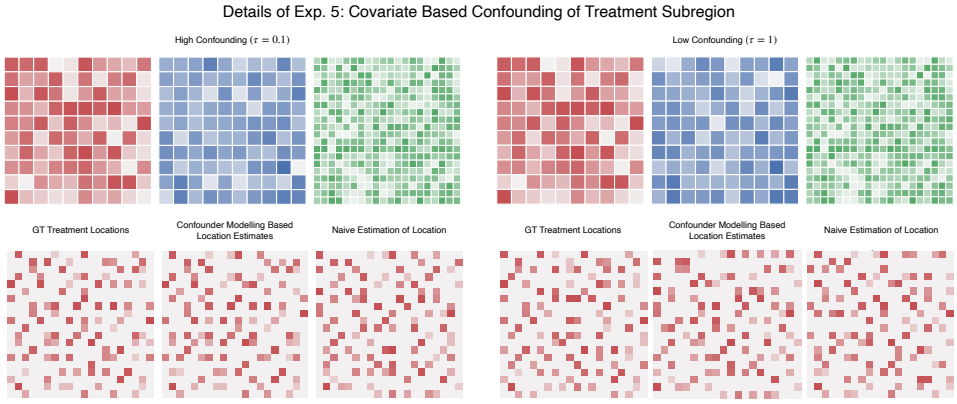


Figure 12: **Results Overview. Top:** Input of LR intervention locations (red), LR outcomes (blue), and HR context (green). **Bottom:** Ground truth treatment locations, estimation of treatment locations based on modeling of location covariate-based confounding, naïve estimation of treatment locations with all free parameters.

Results.

1. Visualizations of the aggregated causal effects show that mean aggregation preserves more variation across regions, whereas max aggregation suppresses smaller effects and produces sharper contrasts.
2. Although the subregional causal effect loss is never used directly for training, it decreases steadily over epochs, indicating that the model recovers fine scale effects from only regional supervision. Training loss is substantially higher under max aggregation, as weaker subregional signals are obscured.
3. Comparing estimated and ground truth local effects confirms this bias: areas with small true effects tend to be overestimated when max aggregation dominates.
4. The aggregation function, parameterized by the temperature τ, is recovered with high accuracy. In the mean aggregation case, τ is estimated within 0.01 of the ground truth, while in the max aggregation case, the estimate is slightly underestimated (5.6), consistent with the plateauing behavior of the softmax function at low temperatures.
5. Overall, the experiment demonstrates that our approach can successfully recover both the aggregation mechanism (within its parametric form) and the underlying causal effects, despite only observing region-level outcomes.

G.2 Exp. 5: Covariate-Based Confounding of Treatment Subregion

Context. In this experiment, we study how confounding between treatment allocation and contextual variables affects causal effect estimation. We take the example of school districts receiving additional public funding and the corresponding changes in educational outcomes.

In reality, funding allocation is rarely random: wealthier or poorer districts may have systematically different chances of receiving funding due to political priorities, lobbying power, or policy constraints. This creates a dependency between the probability of treatment and district level context variables, violating the standard assumption of independent treatment assignment.

We model this confounding explicitly by making the probability of treatment allocation a deterministic function of the contextual variable, passed through a logistic transform to obtain valid probabilities for each region. The treatment effect itself is also a function of the same contextual variable, representing heterogeneous impacts across different socioeconomic conditions.

Setup. We assume $N = 100$ regions, each with $M = 100$ subregions (school districts).

We are given:

- A high-resolution treatment matrix $T \in \{0,1\}^{N\times M}$ with *exactly one* treated subregion per region, i.e., $\sum_{j=1}^{M} t_{i,j} = 1$ for all i.
- A context matrix $C \in [0,1]^{N\times M}$ encoding a socioeconomic/need index for each subregion, where a higher $c_{i,j}$ indicates greater need.
- A noise matrix $\epsilon \in \mathbb{R}^{N\times M}$ with i.i.d. entries $\epsilon_{i,j} \sim \mathcal{N}(0, \sigma^2)$ (default $\sigma^2 = 0.02$).

Confounded Treatment Allocation. For each region i, we form logits from the subregional context,

$$\ell_{i,j} = \alpha_0 + \alpha_1\, c_{i,j},$$

and convert them to a categorical distribution over subregions via a row-wise softmax,

$$p_{i,j} \;=\; \frac{\exp(\ell_{i,j})}{\sum_{k=1}^{M}\exp(\ell_{i,k})}.$$

We then draw exactly one treated subregion $j^\star \sim \mathrm{Categorical}(p_{i,1:M})$ and set

$$t_{i,j} = \mathbb{1}\{j = j^\star\}, \qquad \sum_{j=1}^{M} t_{i,j} = 1.$$

The parameters (α_0, α_1) control the strength and direction of confounding between context and treatment assignment (default $\alpha_0 = 0$, $\alpha_1 > 0$ so higher-need subregions are more likely to be treated).

Outcome Model With Context-Dependent Effects. Subregional outcomes follow the ground-truth functional relationship

$$y_{i,j} \;=\; f_{\boldsymbol{\theta}}(t_{i,j}, c_{i,j}) + \epsilon_{i,j},$$

with the causal effect function

$$f_{\boldsymbol{\theta}}(t_{i,j}, c_{i,j}) \;=\; \begin{cases} 0 & \text{if } t_{i,j} = 0, \\ \theta_{\text{base}} + \theta_{\text{soc}}\, c_{i,j} & \text{if } t_{i,j} = 1, \end{cases}$$

where $\boldsymbol{\theta} = \{\theta_{\text{base}},\, \theta_{\text{soc}}\}$ captures the baseline funding effect and its modulation by socioeconomic need.

Regional Aggregation. Region-level outcomes are the sum over subregions (as in Exp. 1).

This construction induces confounding because (i) treatment allocation depends on $c_{i,j}$ through the softmax logits, and (ii) treatment effects also vary with $c_{i,j}$ via $f_{\boldsymbol{\theta}}(\cdot)$.

Training. We compare two training regimes for estimating the causal effect parameters $\boldsymbol{\theta} = \{\theta_{\text{base}}, \theta_{\text{soc}}\}$ and the high resolution treatment assignments $T \in \mathbb{R}^{N\times M}$ in the presence of treatment context confounding.

Regime 1: Naïve Training Under Independence Assumption. This setup directly applies the training procedure of Exp. 2 (Section E) to the confounded data, it:

1. Treats the treatment allocation as if independent of the context,
2. Learns T as free parameters, constrained to be one hot per row via a differentiable argmax with temperature,
3. Learns $\boldsymbol{\theta}$ by minimizing the MSE between observed and predicted regional outcomes.

This ignores the fact that the treatment assignment mechanism is context dependent.

Regime 2: Confounding Aware Training. Here, we explicitly model the context-dependent treatment allocation mechanism. For each region i, we predict logits

$$\ell_{i,j} = \beta_0 + \beta_1 \, c_{i,j},$$

where (β_0, β_1) are learned parameters, and convert them to a categorical distribution via a row-wise softmax with a learnable temperature $\tau_{\text{conf}} > 0$:

$$p_{i,j}(\tau_{\text{conf}}) = \frac{\exp(\ell_{i,j}/\tau_{\text{conf}})}{\sum_{k=1}^{M} \exp(\ell_{i,k}/\tau_{\text{conf}})}.$$

We then use $p_{i,j}(\tau_{\text{conf}})$ as the (differentiable) treatment assignment in the forward pass.

In both regimes, predicted subregional outcomes are given by:

$$y_{i,j} = f_{\boldsymbol{\theta}}(t_{i,j}, c_{i,j}) = \begin{cases} 0 & \text{if } t_{i,j} = 0, \\ \theta_{\text{base}} + \theta_{\text{soc}} \cdot c_{i,j} & \text{if } t_{i,j} = 1, \end{cases}$$

where in Regime 1, $t_{i,j}$ comes from the naïvely learned $\widehat{T}$, and in Regime 2, $t_{i,j} = p_{i,j}(\tau_{\text{conf}})$ from the confounding model.

Regional outcomes are summed and aggregated. and the training loss is the mean squared error, All parameters are optimized jointly using Adam with a learning rate of 0.001 for 1000 epochs. Regime 1 learns $\{\widehat{T}, \theta_{\text{base}}, \theta_{\text{soc}}\}$, while Regime 2 learns $\{\beta_0, \beta_1, \tau_{\text{conf}}, \theta_{\text{base}}, \theta_{\text{soc}}\}$.

Results.

1. We compare two cases of confounding in subregional treatment allocation, holding region-level treatments and contextual factors fixed. The aggregated causal effects differ markedly depending on the level of confounding, showing that the treatment context dependency directly shapes the observed regional outcomes.
2. We visualize the ground truth intervention locations alongside the estimates obtained with and without explicitly modeling confounding. Under strong confounding, the confounding-aware approach recovers the true intervention locations with high fidelity, while the naïve method from Exp. 2 retrieves only a small fraction correctly.
3. In the low confounding setting, both methods perform similarly, and location probabilities are estimated with comparable accuracy. This confirms that adjusting for confounding does not reduce performance when confounding is weak.
4. For causal effect estimation, the final MSE of subregional effects under low confounding was 2.04 (naïve) versus 1.60 (confounding aware). Under high confounding, the respective errors were 0.74 versus 0.32. These results show that accounting for confounding improves both intervention location recovery and the accuracy of estimated causal effects by conditioning on the correct contextual information.

Overall, this experiment demonstrates that explicitly modeling treatment context confounding is essential as it enables accurate recovery of intervention locations and yields more reliable causal effect estimates, particularly when treatment assignment is strongly biased by contextual factors.