CLAM: Causal Spatial Disaggregation to Infer Local Effects From Coarse Data

Gerrit Großmann[*]
DFKI[†]
Kaiserslautern, Germany
Gerrit.Grossmann@dfki.de

Sumantrak Mukherjee[*]
DFKI[†]
Kaiserslautern, Germany
Sumantrak.Mukherjee@dfki.de

Sebastian Vollmer
DFKI[†], RPTU
Kaiserslautern, Germany
Sebastian.Vollmer@dfki.de

Abstract

Learning spatially fine-grained patterns from coarse-resolution data is inherently difficult; doing so in a causal setting—estimating high-resolution effects from coarse interventional data—adds an extra layer of complexity.

We introduce CLAM, a method for estimating fine-grained causal effects when only coarse-resolution data on interventions and outcomes is available. We assume high-resolution contextual covariates exist that modulate these effects and can be exploited to infer localized causal effects, support counterfactual reasoning, and enable disaggregation of the outcome. Through simulation studies, we demonstrate that CLAM can recover spatially varying causal impacts under diverse conditions.

This has important implications for domains such as public health and environmental policy, where decisions are made at broad scales but causal pathways vary locally. Code is available here.

1 Introduction

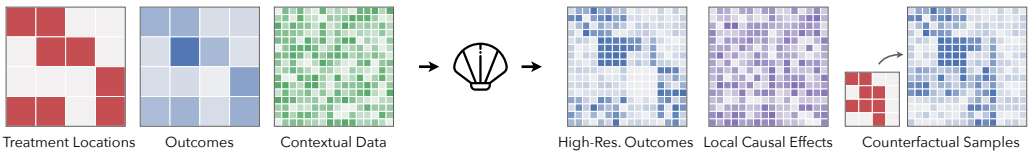

Figure 1: **Motivating example.** CLAM takes regional treatments (a vaccination awareness campaign) and outcomes (vaccination levels), along with subregional covariates (demographics), to learn treatment–outcome relationships and estimate local causal effects or sample counterfactuals.

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 typically 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* [1], *downscaling* [2], *ecological inference* [3], or *super-resolution* [4]. It is relevant

[*]Equal contribution.

[†]DFKI: Deutsche Forschungszentrum für Künstliche Intelligenz GmbH (German Research Center for Artificial Intelligence).

Accepted to the NeurIPS 2025 Workshop on CauScien: Uncovering Causality in Science.

across many fields involving spatio-temporal statistics, including earth science [5], public health [6], and the social sciences [7].

Causality [8], on the other hand, deals with the effect of interventions on outcome variables. Its core insight is that statistical associations do not imply causal relationships. In other words, observing a system is not the same as knowing how intervening on it will change its dynamics.

Answering causal questions is often attempted using LR data, but this data is frequently insufficient for this [9]. 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 questions.

Contribution. This work aims to close this gap by offering a unifying framework for disaggregation grounded in the formalism of structural causal models (SCM) and Pearl's do-calculus [8] that is evaluated through multiple simulation studies that mimic real-world datasets and dynamics.

CLAM (*C*aus*al* dis*a*ggregation *m*ethod) can be summarized as follows: we treat HR contextual data (or *covariances*) as an auxiliary variable that implicitly defines a prior over the HR outcome (cf. Figure 1). Using this, we learn a mechanistic function that maps an intervention indicator and contextual information to the HR outcomes. The model is trained by aggregating the predicted HR outcomes and enforcing consistency with the observed LR measurements. This allows us to perform causal effect estimation, counterfactual reasoning, and computing HR estimates for the LR input data—in some conditions, even when the aggregation function is unknown or in the presence of hidden confounding.

See Appendix A for related work and Appendix B for the connection to causal literature.

2 Problem Setup

For notational simplicity and w.l.o.g., we assume that each region has the same subregion structure, interventions are binary, and all variable are scalar instead of vector-valued.

We assume that we have N regions, each divided into M subregions. The low-resolution *treatment data* is a vector $T \in \{0,1\}^N$, where each element t_i indicates whether a treatment/intervention occurred in region i. Typically, we assume that all subregions of that region are affected. The LR *outcome data* is a vector $X \in \mathbb{R}^N$ that was generated by an unobserved HR outcome matrix via aggregation. The high-resolution *contextual data* (covariate) is a matrix $C \in \mathbb{R}^{N \times M}$, where $c_{i,j}$ corresponds to a contextual value or region $1 \leq i \leq N$, subregion $1 \leq j \leq M$.

We furthermore assume N and M are perfect squares, allowing us to model a spatial grid of $\sqrt{N} \times \sqrt{N}$ regions, each containing a grid of $\sqrt{M} \times \sqrt{M}$ subregions.

Causal Assumptions. Depending on the experimental setting, different assumptions may hold or be relaxed. We may assume a known causal graph, which constrains the model and reduces the risk of spurious explanations, particularly important when contextual variables are high-dimensional and correlated, but only a subset is causally relevant. We typically assume no hidden confounding, meaning all relevant variables affecting the outcome are observed, allowing interventions to be interpreted causally up to reasonable noise. Causal effect identification also requires variation in interventions across regions. If the aggregation function from subregional to regional outcomes is known (e.g., a mean or sum), disaggregation becomes more tractable. Moreover, we typically assume that the treatment does not depend on the context values. Note that we are implicitly using the backdoor criterion (cf. Appendix B.4) to estimate causal effects.

Downstream Applications. Our framework supports several downstream tasks. These include estimating causal effects such as average treatment effects (ATE), localized effects at the subregional level, or heterogeneous effects conditioned on contextual variables. We can also disaggregate coarse outcome measurements into fine-grained outcomes. The framework allows for counterfactual reasoning, such as predicting outcomes under hypothetical intervention placements or alternative contextual conditions. Additionally, we can jointly infer latent variables that modulate causal effects.

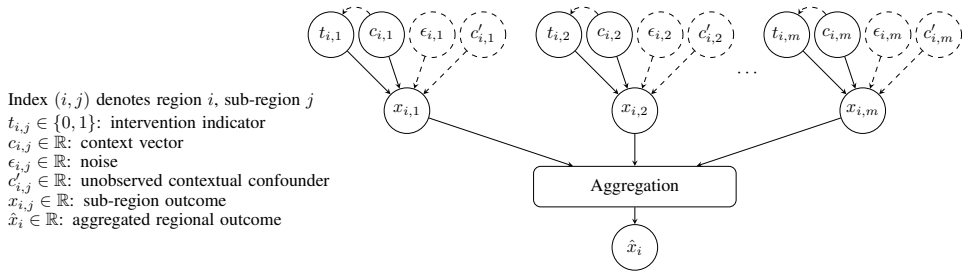

Figure 2: Causal graph for one region. All functional relations are shared among (sub)regions. All scalar values can also be vector-valued. Unobserved confounders and a causal link from the context to the treatment location are optional. The aggregation is typically a mean or a sum.

3 Our Method: CLAM

Our method combines two insights. First, we represent the causal disaggregation problem using a structural causal model formalism. This formalism encodes assumptions and specifies the missing pieces (functions, parameters). Second, we train an ML model to infer the missing pieces where we use the LR data as a loss signal to estimate the most likely HR data and functional relationships.

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 $c'_{i,j}$ (cf. Figure 2). A shared function deterministically maps these inputs to a subregional outcome ($x_{i,j}$ for subregion j of region i), and then aggregates the set of subregional outcomes to a regional outcome ($\hat{x}_i$ for region i):

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

Each subregion is modeled by a local functional graph with aleatoric uncertainty captured via $\epsilon_{i,j}$. Although the functions $f_{\boldsymbol{\theta}}(\cdot)$ and $g_{\boldsymbol{\phi}}(\cdot)$ are shared across regions, region-specific dynamics can still be modeled through context variables (e.g., region indicators or positional encodings).

In general, $g_{\boldsymbol{\phi}}(\cdot)$ can be any function parameterized by $\boldsymbol{\phi}$. For simplicity, we typically assume the aggregation is a sum, the noise is additive, and there are no hidden confounders. Then the model simplifies to:

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

In this case, we define the local causal effect (akin to the ATE definition) $E_{i,j} = f_{\boldsymbol{\theta}}(1,\, c_{i,j}) - f_{\boldsymbol{\theta}}(0,\, c_{i,j})$ as the outcome difference between control and treatment.

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 $x_{i,j}$ are unobserved, training relies solely on their aggregated counterparts.

We optimize with the goal that the inferred aggregated predictions ($\hat{x}_i$) match the observed aggregate: $\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}} = \sum_{i=1}^{N} \big(\hat{x}_i - x_i\big)^2$, where x_i denotes the ground truth 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.

4 Experiments

We present **three simulation studies** demonstrating the capabilities of CLAM, along with **two additional studies** in Appendix C. The additional studies address special cases with confounding in treatment allocation and with an unknown aggregation function.

The appendix contains experiments related to max-based aggregation functions, an unknown aggregation function, internal confounding, and spill-over effects, as well as an ablation study. Generally, we focus on qualitative aspects of the experiments to understand the behavior of our method.

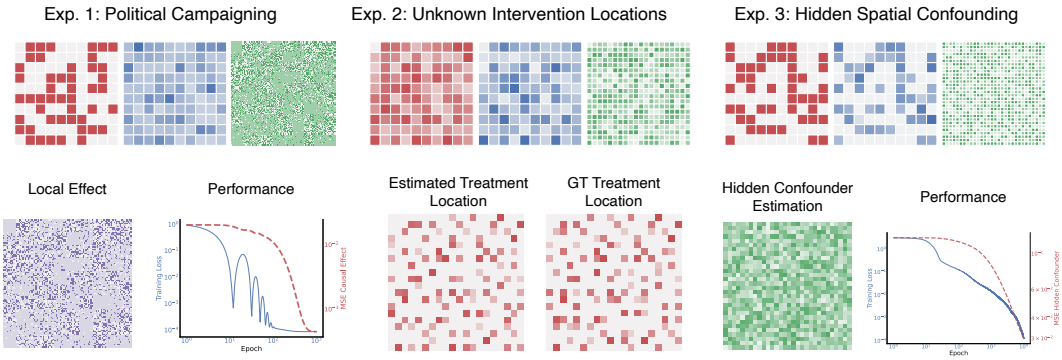

Figure 3: **Results Overview. Top:** Input of LR intervention locations (red), LR outcomes (blue), and HR context (green). **Bottom:** Selected inference results.

Overview results are given in Figure 3. The first row contains the input data. Appendix C contains details of the experiments. We provide code that runs off-the-shelf on Google Colab.

Exp. 1 Political Campaigning. We model how political campaign spending (intervention) affects a candidates relative improvement (outcome), given subregion wealth levels (context). We report the estimated local causal effect and track how the MSE to the ground-truth local causal effect (red) decreases during training, alongside the training loss based on comparing estimated and observed regional outcomes (blue).

Exp. 2 Unknown Intervention Locations. We examine how public school spending (real-valued intervention) affects education scores (outcome) across regions with varying wealth (context), when the exact expenditure location within each region is unknown. Each region contains 4×4 subregions, and the task is to identify the single treated subregion. We report both the ground-truth and estimated treatment locations.

Exp. 3 Hidden Spatial Confounding. We examine time-series data of heat waves (intervention) and their impact on education scores (outcome) across regions with varying vegetation (observed, static) and wealth (unobserved, time-varying). In Figure 3, a single sample from this time series is shown as input. We report the reconstructed hidden confounder and show how the MSE to the unknown ground-truth confounder (red) decreases alongside the training loss (blue) under a linearity assumption. Notably, we are able to reconstruct the hidden confounder reasonably well.

As an overall takeaway from our experiments, we conclude that our method generally performs well in the synthetic setting, particularly for estimating causal effects. The recovery of latent variables (such as intervention locations or hidden confounders) depends on the specific experiment. Overfitting may occur, but only when large amounts of data are available as in Exp. 3. Two main challenges remain: (i) a technical one, where the neural network sometimes fails to find a good optimum, and (ii) a conceptual one, where latent variables are not always identifiable.

5 Conclusions and Future Work

We present CLAM, a method for estimating high-resolution causal effects using low-resolution data on intervention locations and outcomes, by leveraging high-resolution contextual information. The structural causal model formulation makes it easy to formalize assumptions and incorporate prior knowledge. Ultimately, the problem of causal spatial disaggregation reduces to an optimization task, where the structural function and other latent variables are learned from data. For future work, it is necessary to study identifiability more rigorously and bridge the gap between theoretical identifiability and the practical question of when the optimization procedure recovers the most likely solution. A proper Bayesian formulation of the problem is also possible. Moreover, we could perform proper causal discovery on the set of contextual variables to relax our assumptions about which variables are relevant. Finally, while our conceptual framework and thorough investigation demonstrate many promising directions, the most obvious next step is to rigorously evaluate the method on real-world data.

6 Acknowledgement

This work was funded by the Bundesministerium für Bildung und Forschung - (BMFTR) under Grant 01IW23005. We thank the anonymous reviewers of the "NeurIPS 2025 Workshop on CauScien: Uncovering Causality in Science" for their constructive feedback.

References

[1] 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.

[2] 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.

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

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

[5] 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.

[6] 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.

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

[8] Judea Pearl. *Causality*. Cambridge university press, 2009.

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

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

[11] 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.

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

[13] 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.

[14] 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.

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

[16] 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.

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

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

[19] 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.

[20] 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*, pages 1263–1272. Pmlr, 2017.

[21] 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.

[22] 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.

[23] 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.

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

[25] 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.

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

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

[28] 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.

[29] 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.

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

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

[32] 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.

[33] Miruna Oprescu, David K Park, Xihaier Luo, Shinjae Yoo, and Nathan Kallus. Gst-unet: Spatiotemporal causal inference with time-varying confounders. *arXiv preprint arXiv:2502.05295*, 2025.

[34] 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.

A 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 or learning-based techniques. These have been widely applied in climate and environmental sciences [10, 11], including recent work on probabilistic downscaling using Gaussian processes [12]. 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 [13, 14]. 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 [15]. Earlier work on aggregation and disaggregation in optimization similarly highlights the complexity introduced when latent heterogeneity is present but unobserved [16]. CLAM explicitly addresses these challenges by modeling the aggregation process and learning disentangled causal effects at the subregional level.

Causal (De)abstraction. Structurally, our approach is inspired by recent work on causal abstraction, which studies mappings between causal models at different levels of granularity [9, 17, 18]. 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 [19], but differs by operating in a setting where temporal data is limited and only aggregate observations are available.

Graph Neural Networks. Conceptually, our architecture is related to message-passing graph neural networks (GNNs) [20]. In a GNN, pairwise messages are computed and aggregated within 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 $g_{\phi}(\cdot)$ that is permutation-invariant, akin to a Deep Sets model [21]. As in GNNs, we share the parameters of the structural function across nodes or regions.

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 [22, 23]. This idea has been used for causal discovery [24] and domain generalisation [25], 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.

Hierarchical Bayesian Model. The design also parallels hierarchical Bayesian models [26, 27], 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 [28]. Extending this perspective, methods proposed in [29] estimate heterogeneous treatment effects in settings where economic assistance moderates the impact of airstrikes on insurgent violence, while approaches introduced in [30] accommodate arbitrary spillover and carryover effects across space and time. Network-based settings with temporal dynamics have also been explored. In [31] the

authors study bipartite causal inference with interference in time series data, and in [32] methods are developed to analyse the causal effects of modified treatment policies under network interference. Recent methodological innovations include GST-UNet [33], a deep learning architecture for spatiotemporal causal inference under time-varying confounding, and consistency-guided aggregation for causal discovery in multivariate spatiotemporal data introduced in [34]. Together, these works highlight the diversity of approaches to spatiotemporal causality, spanning structural models, interference-aware estimation, and machine learning-driven inference.

B Causal Deabstraction Framework

In this section, we situate our work within the broader causal inference literature. Specifically, we frame this as a novel instance of *causal deabstraction* [9, 18]. Our methodology is informed by the framework of SCMs and is guided by the principle of ICM.

B.1 Structural Causal Framework

The structural causal model introduced in the main text (Section 3) can be formalized in generative form as follows. Each spatial unit (region) $i \in \{1, \ldots, N\}$ is composed of M subregions indexed by $j \in \{1, \ldots, M\}$. For each subregion (i, j), we posit the following SCM:

$$
\begin{aligned}
c_{i,j} &\sim P_C && \text{(high-resolution contextual covariate)} \\
\epsilon_{i,j} &\sim P_\epsilon && \text{(exogenous noise)} \\
t_{i,j} &= T_i \in \{0, 1\} && \text{(binary intervention, shared across subregions)} \\
x_{i,j} &= f_{\boldsymbol{\theta}}(t_{i,j}, c_{i,j}, \epsilon_{i,j}) && \text{(subregional outcome)} \\
\hat{x}_i &= g_\phi\left(\{x_{i,j}\}_{j=1}^{M}\right) && \text{(aggregated regional outcome)}
\end{aligned}
$$

Here, $f_{\boldsymbol{\theta}}(\cdot)$ encodes the local causal mechanism, shared across all subregions, while $g_{\boldsymbol{\phi}}(\cdot)$ aggregates subregional outcomes to the regional level. In the main section we considered simplified casessuch as sum aggregation, additive noise, and the absence of hidden confoundersto provide intuition. The formulation here generalizes that view by writing the SCM in explicit stochastic form, specifying how interventions, contexts, and noise are drawn, while making clear that $f_{\boldsymbol{\theta}}(\cdot)$ 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 distribution of 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* [23, 22, 24], which posits that causal relations are stable across environments unless directly intervened upon.

We treat each region i as an *environment* defined by its distribution over covariates $\{c_{i,j}\}$, but governed by a shared mechanism $f_{\boldsymbol{\theta}}(\cdot)$. The variability in covariate composition across regions allows us to disentangle $f_{\boldsymbol{\theta}}(\cdot)$ from the observed aggregated outcomes x_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 [9, 18], 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 effects.

This is possible due to two key properties:

1. The aggregation function $g_\phi(\cdot)$ is consistent and applied uniformly across regions.

2. The diversity in the covariate compositions $\{c_{i,j}\}_{j=1}^{M}$ across regions induces identifiable variation in the aggregates x_i under the shared mechanism $f_{\boldsymbol{\theta}}(\cdot)$.

Given sufficient diversity and the assumption of causal invariance, the aggregated outputs provide a supervisory signal to recover the latent causal function.

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 $x_{i,j}$ is identified via the backdoor criterion:

$$\mathbb{E}[x_{i,j} \mid do(T_i = t)] = \int \int f_{\boldsymbol{\theta}}(t, c, \epsilon)\, p(c)\, p(\epsilon)\, dc\, d\epsilon$$

In Exps 14, we assume $T_i \perp\!\!\!\perp \epsilon_{i,j} \mid c_{i,j}$, and that T_i is either randomized or conditionally independent of $C_{i,j}$. Under this setting, the expectation can be approximated via observational data. However, because only the aggregate outcome x_i is observed, we must rely on the consistency constraint:

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

The training procedure thus finds the function $f_{\boldsymbol{\theta}}(\cdot)$ that jointly explains all observed aggregates under this constraint, while enforcing that $f_{\boldsymbol{\theta}}(\cdot)$ is invariant across regions. In this sense, the estimation of $f_{\boldsymbol{\theta}}(\cdot)$ 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 the coarse observations as aggregates of fine-grained causal effects and leveraging variation in covariate compositions, CLAM performs inference not merely on statistical associations but on causal mechanisms that are meaningful and robust to changes in population structure. All in all, we contribute a formal and algorithmic instantiation of *causal deabstraction* from low-resolution outcomes and high-resolution covariates.

C Details of Experiments

We present five simulation studies to illustrate different aspects of our framework. Exp. 1 models political campaigning, where interventions interact with socioeconomic context in a heterogeneous manner. Exp. 2 examines public school funding when intervention locations are unknown and must be inferred from aggregated data. Exp. 3 studies the spatiotemporal effects of heat waves, where repeated observations enable recovery of a hidden confounder. Exp. 4 considers driving bans with an unknown aggregation function between subregional and regional outcomes. Exp. 5 introduces treatmentcontext confounding, where the probability of subregional treatment depends on contextual variables, affecting both allocation and outcomes.

C.1 Exp. 1: Political Campaigning

Context. We assume that politicians spend money on political campaigning, but only in certain states (or regions). This is modeled as a binary decision: either money is spent on a region or not. The outcome we measure is the **relative improvement** in the politicians performance compared to the previous year.

We also have contextual data mapping each subregion to a wealth level categorized as *poor*, *middle class*, or *rich*. The rationale is that a campaign may have varying effectiveness depending on the socioeconomic background of the population.

Setup. We assume a 10×10 grid of regions, where each region contains 10×10 subregions. That is, $N = 100$ regions, each with $M = 100$ subregions.

We are given:

- A binary treatment vector $T \in \{0, 1\}^N$, where $T_i = 1$ indicates that region i was treated (i.e., the politician spent money there). If a region is treated, all of its subregions are considered treated.
- A context matrix $C \in \{0, 1, 2\}^{N \times M}$ encoding the wealth level of each subregion, where 1 = poor,2 = middle class, 3 = rich.
- An outcome vector $X \in \mathbb{R}^N$ representing the observed improvement per region.

We generate the treatment matrix using a Bernoulli distribution with probability 0.5, assigning each subregion to either the intervention or control group at random. The subregional wealth data is generated randomly, but with the constraint that each region (roughly) has the same average wealth. This ensures that the model cannot learn a mapping from regional wealth to the outcome.

We also construct two extreme cases: one region where all subregions are middle-class, and another region where 50% of subregions are rich and 50% are poor. Both regions have the same average regional wealth of 1.0.

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

$$\hat{x}_i = \sum_{j=1}^{M} \left(f_{\boldsymbol{\theta}}(t_{i,j}, c_{i,j}) + \epsilon_{i,j} \right),$$

where:

- $t_{i,j} = T_i$ is the treatment status of subregion j in region i (inherited from the region),
- $\epsilon_{i,j} \sim \mathcal{N}(0, \sigma^2)$ is i.i.d. Gaussian noise. The default noise variance σ^2 is set to 0.02.

The function $f_{\boldsymbol{\theta}}(\cdot)$ encodes the treatment effect:

$$f_{\boldsymbol{\theta}}(t_{i,j}, c_{i,j}) = \begin{cases} 0.0 & \text{if } t_{i,j} = 0, \\ -0.1 & \text{if } t_{i,j} = 1 \text{ and } c_{i,j} = 1 \text{ (poor)}, \\ 0.0 & \text{if } t_{i,j} = 1 \text{ and } c_{i,j} = 2 \text{ (middle class)}, \\ 0.3 & \text{if } t_{i,j} = 1 \text{ and } c_{i,j} = 3 \text{ (rich)}. \end{cases}$$

As we see, the campaign has:

- A slightly negative effect in poor neighborhoods,
- No effect in middle-class neighborhoods, and
- A large positive effect in rich neighborhoods.

This means that the higher the variance in each region becomes, the higher the effect strength.

As in most experiments, we assume no hidden confounding, no time series data, treatment assignment independent of context, and mean-based aggregation.

Training. We train a simple model with 6 trainable parameters ($\boldsymbol{\theta}$):

$$f_{\boldsymbol{\theta}}(t_{i,j}, c_{i,j}) = \begin{cases} \theta_1 & \text{if } t_{i,j} = 0 \text{ and } c_{i,j} = 1 \text{ (poor)}, \\ \theta_2 & \text{if } t_{i,j} = 0 \text{ and } c_{i,j} = 2 \text{ (middle class)}, \\ \theta_3 & \text{if } t_{i,j} = 0 \text{ and } c_{i,j} = 3 \text{ (rich)}, \\ \theta_4 & \text{if } t_{i,j} = 1 \text{ and } c_{i,j} = 1 \text{ (poor)}, \\ \theta_5 & \text{if } t_{i,j} = 1 \text{ and } c_{i,j} = 2 \text{ (middle class)}, \\ \theta_6 & \text{if } t_{i,j} = 1 \text{ and } c_{i,j} = 3 \text{ (rich)}. \end{cases}$$

We use a learning rate of 0.001 and train for 1000 epochs.

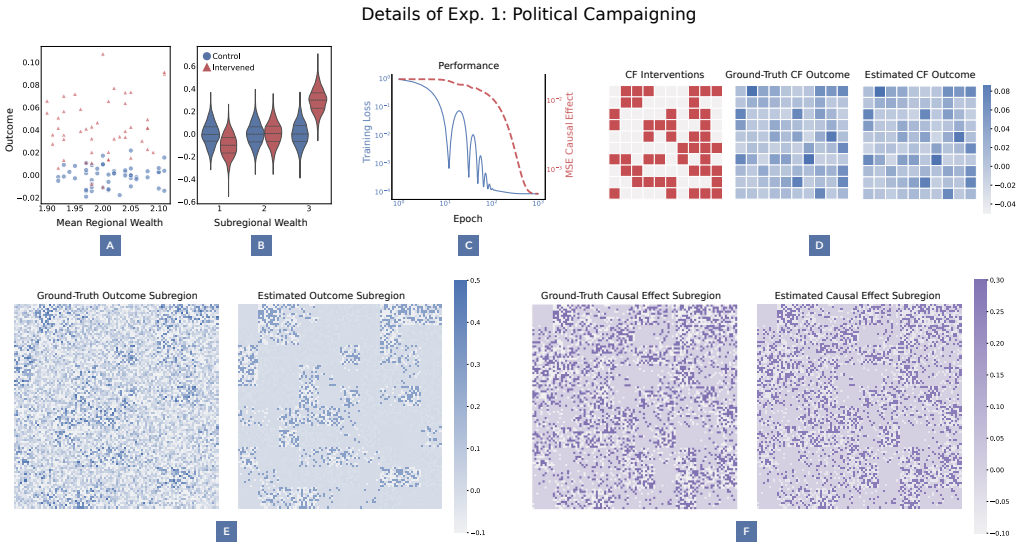

Figure 4: Evaluation of the learned parameters, causal effect estimates, and outcome predictions.

Results.

1. We show the correlation between regional wealth (Figure 4a) and outcome, and between subregional wealth (Figure 4b) and outcome under the two treatment conditions. This shows that the regional wealth value alone is not sufficient to predict the outcome. Two outliers (regions with maximal and minimal variability), denoted with black arrows, are clearly visible.
2. We also show how the MSE of the estimated causal effect matrix evolves with the training loss (Figure 4c).
3. We use the learned parameters and a hypothetical intervention mask to generate both regional and subregional counterfactual outcomes (Figure 4d). Note that subregional counterfactual outcomes can be obtained by sampling from the noise posterior, subject to preserving the correct overall sum, provided that the noise distribution is known.
4. We use the inferred function $f_{\hat{\boldsymbol{\theta}}}(\cdot)$ to estimate the HR outcome matrix $\hat{X}$ (Figure 4e) and the HR causal effect matrix $\hat{E}$ (Figure 4f), and compare them against their respective ground truths. Note that the actual HR outcome matrix contains the noise form the data generation, which is stripped away in the estimated version and is not relevant for the causal effect estimation. If we were to generate multiple HR outcome samples, the mean would be closer to the estimated matrix than the one sample we consider.
5. We test whether the inferred parameters $\hat{\boldsymbol{\theta}}$ are close to the ground truth $\boldsymbol{\theta}$. We find that they can consistently be recovered within a small margin of error (about 0.1). On the one hand, this is not surprising because there are only six parameters to fit; on the other hand, the relatively large subregions with relatively high noise weaken the loss signal. We also observe that the MSE is quite robust against Gaussian noise.

Ablation Study Interpretation. Our setting is motivated by cases where region-level outcomes exhibit low variability across regions, resulting in an underdetermined inverse problem. In such scenarios, standard aggregate-level causal inference fails to distinguish between competing fine-grained causal explanations, as multiple subregional configurations may produce indistinguishable regional outcomes.

To address this, our framework leverages high-dimensional contextual covariates observed at the subregional level. These covariates, when sufficiently heterogeneous across and within regions, condition the shared causal mechanism $f_{\boldsymbol{\theta}}(\cdot)$ in such a way that each aggregate outcome can be

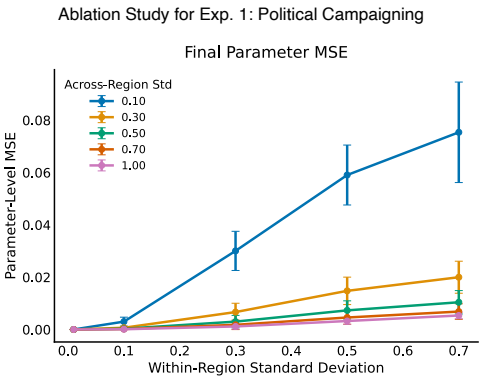

Figure 5: We evaluate the effect of varying the variance of regional context means and, for each fixed context mean, the inter-regional variance. Results are averaged over 5 random seeds for each parameter combination, and uncertainty bounds represent the corresponding variability. The experimental settings and estimation technique are identical to those of Exp 1.

viewed as a distinct mixture over contextual configurations. This compositional diversity effectively provides multiple "views" on the same regional-level signal, enabling the model to disentangle the underlying causal function from the aggregated outcomes.

Exp. 1 illustrates this behavior: although the regional outcomes remain nearly constant across units, increasing the variance of subregional covariates leads to improved recovery of the true causal mechanism. This is further supported by our ablation study, where we systematically vary subregional and regional variance. We observe that increasing subregional variance while keeping region-level outcome variability fixed improves parameter recovery, as measured by the final MSE between estimated and true causal parameters. In contrast, increasing regional variance has comparatively little effect on parameter estimation, indicating that the identifiability gains stem not from more diverse outcomes, but from the richer structure encoded in the subregional context.

These results validate the central insight of our approach: in low-variability regimes where traditional identification strategies break down, it is the variation in subregional covariate compositions combined with the assumption of an invariant causal mechanism that enables accurate causal deabstraction.

C.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 (school districts), but only a single district per region receives a funding intervention. At the regional level, we only observe the *total* spending increase, not the specific district that received it. The goal is to infer the high-resolution intervention locations from aggregated data.

The subregional context consists of high-resolution demographic and economic data here represented as a normalized *wealth score* between 0 (low socioeconomic status) and 1 (high socioeconomic status). We hypothesize that the effect of school funding depends on this wealth score via a global shift and scale.

Setup. We assume $N = 100$ regions, each with $M = 4$ subregions (school districts).

We simulate:

- A high-resolution intervention matrix $T \in \mathbb{R}^{N \times M}$, where $t_{i,j}$ is the spending increase in subregion j of region i.
- A regional intervention vector $T^{\text{reg}} \in \mathbb{R}^N$, where $T_i^{\text{reg}} = \sum_{j=1}^{M} t_{i,j}$ is the total spending increase in region i.

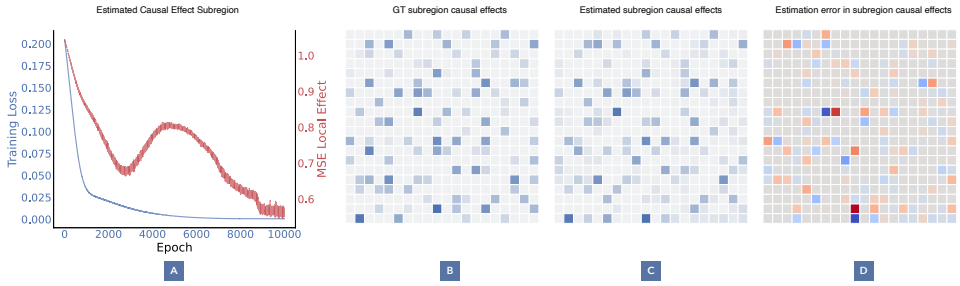

Figure 6: Here we report additional details such as the performance of training loss and MSE of the local effect estimation, alongside that we visualize the ground truth causal effects, the estimated causal effects, and the estimation errors for the different subregions.

- A high-resolution context matrix $C \in [0,1]^{N \times M}$ encoding the socioeconomic status of each subregion, where $c_{i,j} = 0$ represents low socioeconomic status and $c_{i,j} = 1$ represents high 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 $\sigma^2 = 0.02$.

The intervention matrix T is generated by:

1. Initializing all entries to zero,
2. For each $i \in \{1, \ldots, N\}$, choosing exactly one $j \in \{1, \ldots, M\}$ uniformly at random and assigning $t_{i,j} \sim \mathrm{Uniform}(0, 1)$,
3. For $i = 0$, leaving all entries $t_{0,j} = 0$.

The context matrix C is sampled i.i.d. from $\mathrm{Uniform}(0, 1)$.

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

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

where:

- $t_{i,j}$ is the observed subregional intervention,
- $c_{i,j}$ is the socioeconomic score,
- $\epsilon_{i,j}$ is Gaussian noise.

The causal effect function is parameterized as:

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

with ground-truth parameters $\theta_{\text{shift}} = 4.2$ and $\theta_{\text{scale}} = 1.23$.

The regional outcome vector $X \in \mathbb{R}^N$ is obtained by taking a mean over subregional outcomes.

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

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

The model has:

- $N \times M = 400$ free parameters $\{\hat{t}_{i,j}\}$ representing the estimated subregional intervention intensities,
- Two scalar parameters $\{\theta_{\text{shift}}, \theta_{\text{scale}}\}$ specifying the parameterization of $f_{\boldsymbol{\theta}}(\cdot)$.

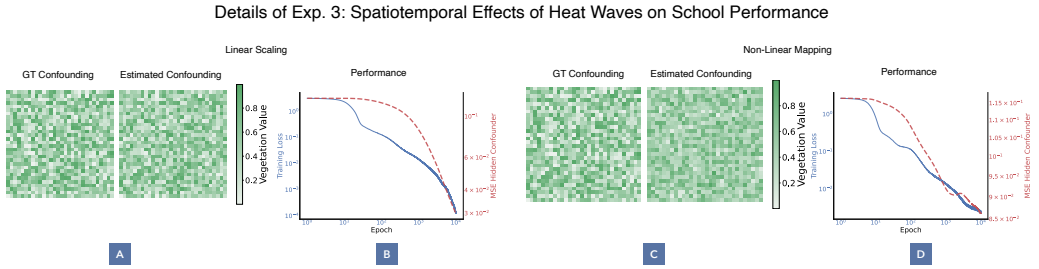

Figure 7: Shown are the groundtruth vegetation index, the corresponding estimated confounder fields, and the joint plots of training loss and confounder MSE. Results are presented for two settings: linear scaling (left) and MLP learning (right).

The learned function takes the form:

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

An auxiliary preprocessing function $h(\cdot)$ is applied to the free parameters $\hat{t}_{i,j}$ to ensure they represent valid interventions:

1. Square each $\hat{t}_{i,j}$ to enforce non-negativity,
2. Apply a differentiable row-wise argmax (temperature-controlled softmax) so that one subregion per region has the maximal value,
3. Row-wise normalization so that $\sum_{j=1}^{M} t_{i,j} = T_i^{\text{reg}}$ matches the observed total regional intervention.

In the forward pass:

1. Process the learned $\hat{T}$ via $h(\cdot)$ to obtain T,
2. Compute predicted subregional outcomes: $\hat{x}_{i,j} = f_{\boldsymbol{\theta}}(t_{i,j}, c_{i,j})$,
3. Aggregate to regional predictions by averaging.
4. Compare with observed x_i from $X \in \mathbb{R}^N$ using the MSE loss.

We optimize all parameters jointly using Adam with a learning rate of 0.001, training for $100{,}000$ epochs with a fixed random seed.

Results.

1. Figure 6a shows the evolution of the training loss and the subregional MSE. While the loss decreases monotonically, the estimation error is more variable but exhibits a consistent downward trend, indicating convergence toward the true causal mechanism.
2. The ground truth subregional causal effects are visualised in Figure 6b, providing the reference against which predictions are evaluated. The predicted effects in Figure 6c closely match the ground truth, capturing both magnitude and spatial structure of the interventions.
3. The error map in Figure 6d confirms this alignment: estimation error is concentrated outside treated regions, while within treated regions it is nearly zero, demonstrating accurate recovery of both intervention locations and local effect sizes.

C.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 confounding 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 largescale 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 confounder to outcomes, and the vegetation values themselves, given only aggregated regionallevel 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 highresolution 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 largescale 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 confounding 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:

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

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

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

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 $\hat{U} \in [0,1]^{N \times M}$ 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 $\hat{x}_{i,j}^{(w)}$ for that subregion and week.

In the forward pass for a given week w:

1. The model takes as input $T^{(w)}$, $C^{(w)}$, and $\hat{U}$,
2. The MLP evaluates $f_{\boldsymbol{\theta}}(\cdot)$ at each (i,j) to produce predicted subregional outcomes $\hat{x}_{i,j}^{(w)}$,
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 weeks in random order at each epoch. The total parameter set consists of the MLP weights $\boldsymbol{\theta}$ and the vegetation tensor $\hat{U}$. 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 7a shows the loss curves and confounder MSE under the linear scaling parameterization. Both training loss and MSE decrease rapidly, and the final error is very small, indicating nearperfect recovery of the hidden confounder.
2. Figure 7b compares the groundtruth vegetation field to the estimated confounder under linear scaling. The two maps are almost indistinguishable, confirming that the restricted functional form matches the data-generating process and allows highly accurate confounder recovery.
3. Figures 7c and d present the same results for the nonlinear MLP parameterization. While the model still recovers meaningful structure in the hidden confounder, the confounder 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.
4. Overall, these results demonstrate that the model can reconstruct regional outcomes while also uncovering hidden drivers of heterogeneity. The comparison between linear and nonlinear parameterizations highlights the tradeoff: structural restrictions yield more accurate recovery when they match the true mechanism, while flexible approximators such as MLPs still learn the hidden confounder but with reduced precision.

C.4 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

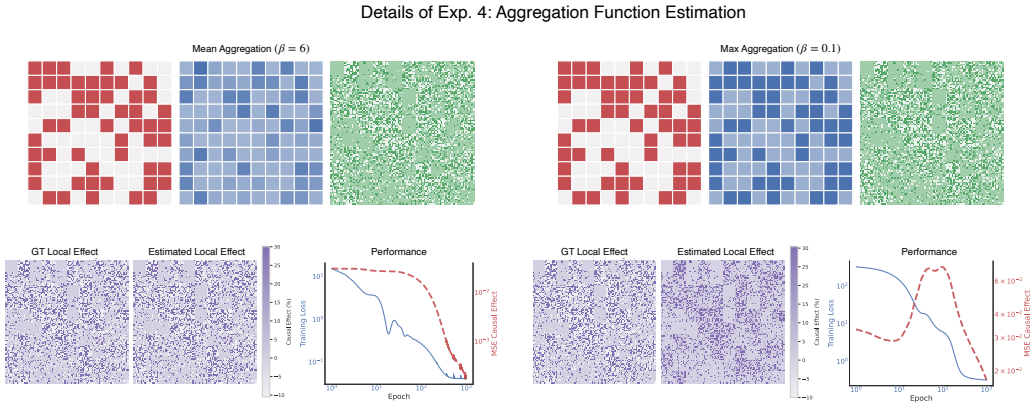

Figure 8: **Results Overview. Top:** Input of LR intervention locations (red), LR outcomes (blue), and HR context (green) for mean aggregation and max aggregation cases. **Bottom:** We observe the estimated local effects and the performance of our model over multiple epochs for both cases.

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:

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

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:

$$x_i = \sum_{j=1}^{M} p_{i,j}(\tau) \cdot \hat{x}_{i,j},$$

where:

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

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

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.

C.5 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.

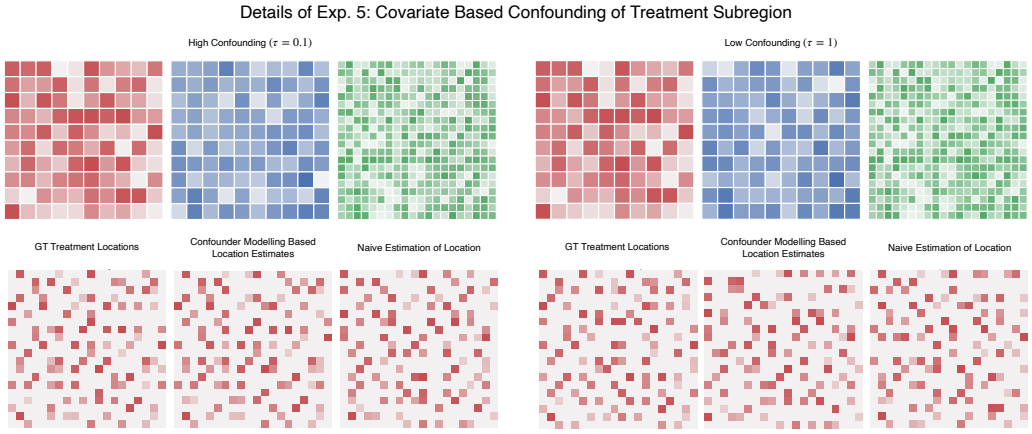

Figure 9: **Results Overview. Top:** Input of LR intervention locations (red), LR outcomes (blue), and HR context (green). **Bottom:** Ground truth treatment locations, estimation of treatment locations based on modeling of location covariate-based confounding, naive estimation of treatment locations with all free parameters.

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

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

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 $\hat{T} \in \mathbb{R}^{N \times M}$ 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 C) to the confounded data, it:

1. Treats the treatment allocation as if independent of the context,
2. Learns $\hat{T}$ 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:

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

where in Regime 1, $t_{i,j}$ comes from the naïvely learned $\hat{T}$, 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 $\{\hat{T}, \theta_{\text{base}}, \theta_{\text{soc}}\}$, 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.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We have accurately described the scope and contribution of our work in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We mention the limitations of our approach and its applicability in specific sections.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: We do not make any theoretical claims in our paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide our code which can be verified indepdently and also include thorough details of our setup in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We have provided the code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so No is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Hyperparameters used and optimizers are reported.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: We focus on qualitative results, more on quantitative results, although we report statistical results in an ablation study.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Our code can be reproduced on online Google Colab instances or any reasonably modern computer.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers, CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Based on our reading of the ethics guidelines our contribution does not violate of any of the rules.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: Our work introduces a new research direction; however, it is not developed enough to cause broader impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our code/data cannot be utilised for high risk misuse so there was no need for safeguard protocols.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We use standard tools such as PyTorch in our code; however, the data models etc, are developed by us.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We do not release any new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We do not involve any human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our contribution did not require or qualify for institutional review board.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLMs were used for grammar checks and rephrasing of some parts of the manuscript, and were also used as coding assistants for submodules of our experiments, which were later verified by the authors of the paper.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.