Appendix

A.1.1. Supplementary Information for the Image-Type Probabilistic Models

A.1.1.1. VISUALIZING THE PROBABILISTIC MODEL STRUCTURE

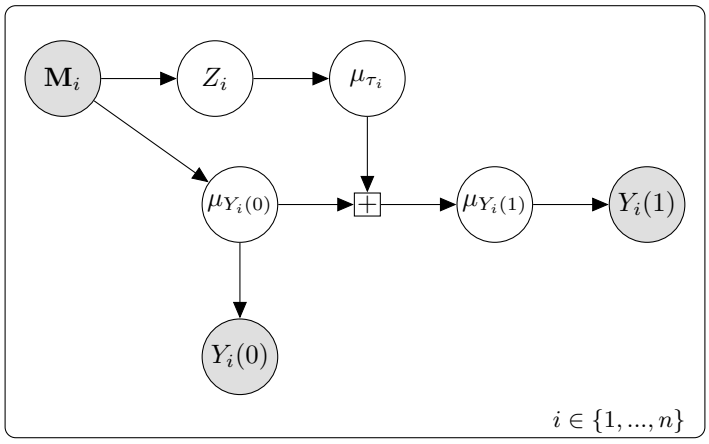

Figure A.1: This figure, which is not a DAG, is a stylized schematic depiction of the probabilistic treatment heterogeneity model for images. The gray circles denote observed random variables; the white circles denote latent variables. The square node denotes deterministic transformations. Z_i denotes the image type generating a distribution over treatment effects. Arrows denote statistical dependency in the probabilistic model, not causal dependencies.

A.1.1.2. DERIVING THE CONDITIONAL DISTRIBUTION, $\{\tau_i = Y_i(1) - Y_i(0) | Z_i = z\}$

Using the model outlined in §3.1, conditioning on τ_i, and exploiting Normality,

$$\{Y_i(1) - Y_i(0) \mid Z_i = z, \mu_{\tau_i}\} \sim \mathcal{N}(\mu_{\tau_i},\ \sigma^2_{0,z} + \sigma^2_{1,z})$$

Integrating out μ_{τ_i}:

$$\begin{aligned}
p(Y_i(1) - Y_i(0) = \tau_i \mid Z_i = z) &= \int_{-\infty}^{\infty} p(Y_i(1) - Y_i(0) = \tau_i \mid Z_i = z,\ \mu_{\tau_i})\, p(\mu_{\tau_i} \mid Z_i = z)\, \mathrm{d}\mu_{\tau_i} \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi(\sigma^2_{0,z} + \sigma^2_{1,z})}} \exp\left\{\frac{-(\tau_i - \mu_{\tau_i})^2}{2(\sigma^2_{0,z} + \sigma^2_{1,z})}\right\} \\
&\qquad \times \frac{1}{\sqrt{2\pi\sigma^2_{\tau,z}}} \exp\left\{\frac{-(\mu_{\tau_i} - \mu_{\tau,z})^2}{2\sigma^2_{\tau,z}}\right\} \mathrm{d}\mu_{\tau_i} \\
&= \frac{1}{\sqrt{2\pi([\sigma^2_{0,z} + \sigma^2_{1,z}] + \sigma^2_{\tau,z})}} \exp\left\{\frac{-(\tau_i - \mu_{\tau,z})^2}{2([\sigma^2_{0,z} + \sigma^2_{1,z}] + \sigma^2_{\tau,z})}\right\}.
\end{aligned}$$

Therefore,

$$\{\tau_i = Y_i(1) - Y_i(0) \mid Z_i = z\} \ \sim\ \mathcal{N}\left(\mu_{\tau,z},\ \sigma^2_{0,z} + \sigma^2_{1,z} + \sigma^2_{\tau,z}\right).$$

A.1.2. Simulation Details

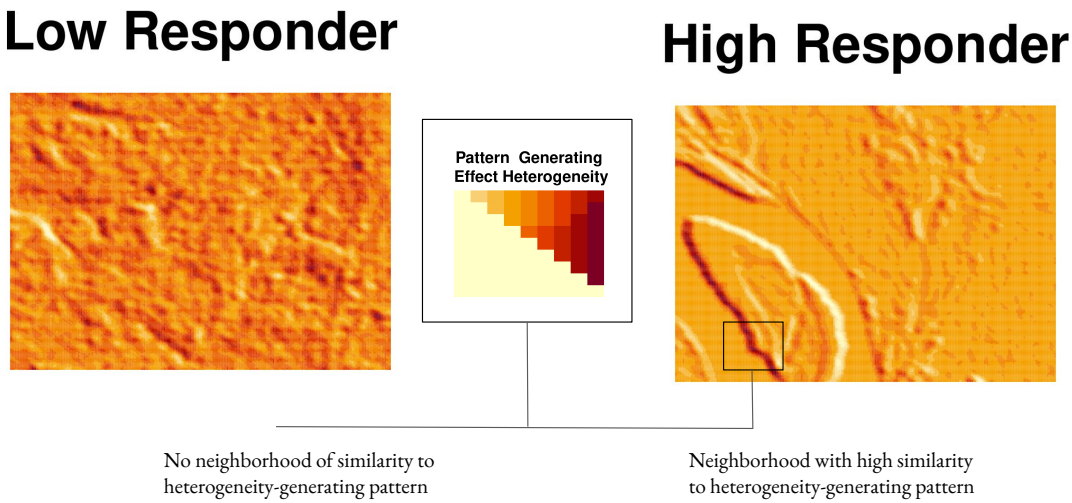

Figure A.2: Simulation design illustration. *Center:* The image pattern used in generating the heterogeneity response in the simulation design of §4. Left: An image having no regions of strong similarity to the heterogeneity-generating pattern (leading to a low treatment effect). Right: An image with many regions of strong similarity to the heterogeneity-generating pattern (leading to a high treatment effect).

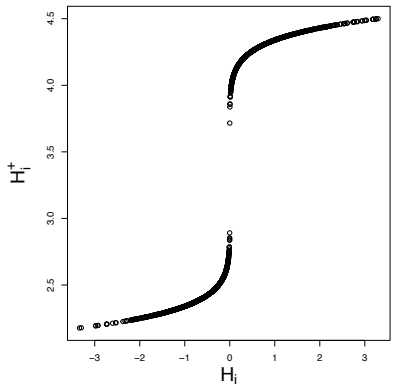

Figure A.3: Illustration of the non-linear transformation used in the simulation in generating H_i^+ from H_i.

A.1.3. Supplementary Analyses for the Application

A.1.3.1. ADDITIONAL DATA DESCRIPTION

We obtain satellite data for the neighborhood around each experimental unit in the following way. First, the place of residence for each unit was geo-referenced using OpenStreetMap. When geo-referencing failed, we use the geometric center for the layer associated with the geographic unit as our focal point for the given unit. Satellite information was then obtained for a cube around focal points with side lengths of 5000 meters. For the skilled work outcome, we take the scaled sum of the log hours worked by experimental units in the last 7 days in skilled or highly skilled trades.

A.1.3.2. ADDITIONAL ANALYSES

Table A.1: Correlation of estimated image cluster 1 probabilities with key tabular covariates.

	Correlation
Urban	0.18
Longitude	-0.01
Latitude	0.27
Female indicator	0.04
Human capital score	-0.11

Figure A.4: Images with most uncertainty in cluster probabilities from the main empirical analysis.

A.1.4. Empirical Analysis with Orthogonalized Potential Outcomes

We orthogonalize outcomes by, in line with the original experimental analysis, fitting a regression model predicting the outcome using main treatment effects and interactions between treatment and gender, treatment and baseline human capital, and treatment and baseline business capital (as well as the main effect terms for the associated interaction). We find a 0.85 correlation between the cluster probabilities using the orthogonalized and raw outcomes.

A.1.5. Model Implementation Details

In the implementation of our models using Bayesian CNN arms, we leave the number of hidden layers, filter size, and so forth as parameters that can be set by investigators.

The unconstrained components of the uncertainties are drawn from Gaussians with mean and variance scaled indexed to z; the non-negativity of the variance is enforced through the softplus

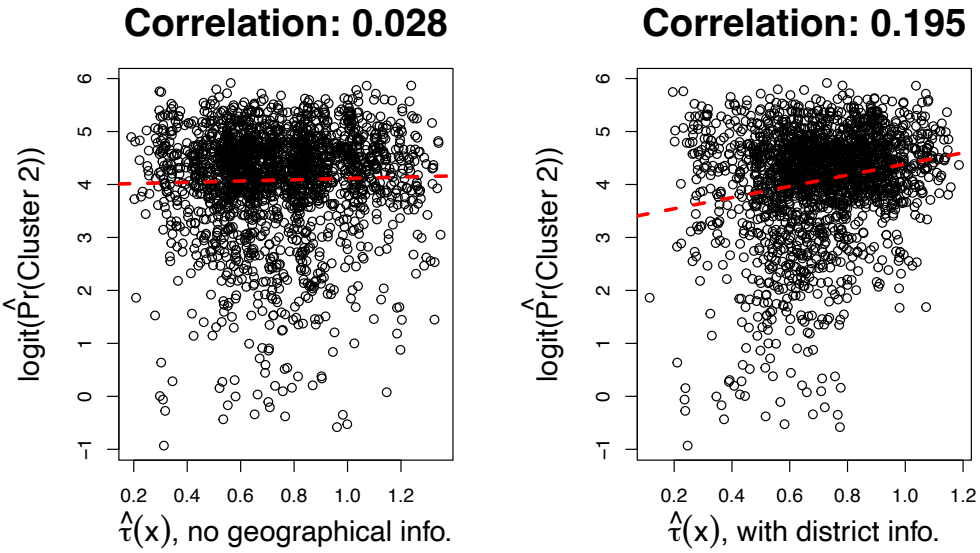

Figure A.5: *Left*. Correlation between estimated treatment effects using a causal forest with individual-level tabular covariates and the posterior mean cluster 2 probabilities from the image heterogeneity model. Individual-level covariates include gender, education, parental education, and indicators for whether a unit's mother and father were alive at the start of the experiment. *Right*. Correlation between estimated treatment effects using a causal forest with individual-level tabular covariates along with district-level indicators and the posterior mean cluster 2 probabilities. Unsurprisingly, the correlation increases, but there is still considerable information present in the estimated clusters not reducible to district indicators alone.

transformation (where softplus(x) = $\log(1 + \exp(x))$). Neural network parameters receive priors using the Empirical Bayes' approach described in Krishnan et al. (2020).

In our application, we use four convolutional layers (filter dimension 5×5), separated by max-pooling layers (2×2). Each convolutional layer applies 32 filters. Bottleneck projection layers are used after each convolutional layer, projecting the 32 dimensions down to 3 to keep the number of parameters reasonably low. Batch normalization layers are used across the feature dimension after each non-linearity (batch normalization momentum across each update step is $= 0.90$). The swish activation is used. We apply the Gumbel-Softmax to approximate the random categorical sampling with the inverse temperature parameter set to 0.5. With this model structure, each batch sample of 20 units takes about one second on a single Apple M1 GPU using Metal-optimized tensorflow 2.10. The full simulation suite takes about 12 hours on local hardware.

A.1.5.1. Encouraging ATE Model Equivalence via Causal Regularization

The modeling process just described involves selecting several parameters, such as the number of clusters and even the kind of images used. Given the numerous possibilities involved, we can consider the addition of a causal regularization term to encourage all models to be equivalent in their implied marginal effect (for discussion of causal regularization in a different context, see (Oberst

et al., 2021)). By the Law of Total Expectation, the ATE $\overline{\tau}$ satisfies

$$\overline{\tau} = \mathbb{E}[Y_i(1) - Y_i(0)] = \sum_{z=1}^{K} \mathbb{E}\big[Y_i(1) - Y_i(0) \mid Z_i = z\big] \Pr(Z_i = z) = \sum_{z=1}^{K} \tau(z) \Pr(Z_i = z),$$

a fact that gives rise to a natural estimator using the sample analogs of the theoretical quantities: $\widehat{\overline{\tau}}_{\text{Model}} = \sum_{z=1}^{K} \widehat{\tau}(z)\widehat{\Pr}(Z_i = z)$, where the $\widehat{\tau}(z)$'s are taken from the mixture components and the $\widehat{\Pr}(Z_i = z)$ term is estimated from the marginal cluster probabilities.

However, the ATE can also be estimated using the non-parametric difference-in-means estimator

$$\widehat{\overline{\tau}}_{\text{Non-parametric}} = n_1^{-1} \sum_{i=1}^{n} Y_i \cdot T_i - n_0^{-1} \sum_{i=1}^{n} Y_i \cdot (1 - T_i),$$

where n_t denotes the number of units in treatment group, $t \in \{0, 1\}$. This non-parametric estimator is under minimal assumptions, consistent (Imbens and Rubin, 2015). Thus, if the proposed heterogeneity model is consistent as well,

$$\{\widehat{\overline{\tau}}_{\text{Model}} \overset{n\to\infty}{\to} \overline{\tau}\},\ \{\widehat{\overline{\tau}}_{\text{Non-parametric}} \overset{n\to\infty}{\to} \overline{\tau}\} \Rightarrow (\widehat{\overline{\tau}}_{\text{Model}} - \widehat{\overline{\tau}}_{\text{Non-parametric}})^2 \overset{n\to\infty}{\to} 0 \tag{4}$$

If the implied ATE from the model diverges too far from the non-parametric estimator, the credibility of the proposed model would be thereby reduced (see Figure A.6 for an illustration). Under additional modeling assumptions, we can in fact re-parameterize the parametric model exactly so that $\widehat{\overline{\tau}}_{\text{Model}} = \widehat{\overline{\tau}}_{\text{Non-parametric}}$ exactly (see §A.1.6). However, the exact re-parameterization forcing $\widehat{\overline{\tau}}_{\text{Model}} = \widehat{\overline{\tau}}_{\text{Non-parametric}}$ involves similar problems as found in the compositional statistics literature (e.g., ordering of clusters can affect the results (Greenacre, 2021)), we instead incorporate a soft penalty that is invariant to the ordering of clusters:

$$\textit{Model ATE Equivalence Regularization Term:}\quad \lambda \left(\sum_{z=1}^{K} \widehat{\tau}(z)\widehat{\Pr}(Z_i = z) - \widehat{\overline{\tau}}_{\text{Non-parametric}} \right)^2$$

We can add this to the variational objective to encourage marginal effects to be equivalent regardless of the parameterization of the model, using a non-parametric estimator for the ATE as a baseline.

A.1.6. Causal Regularization Details

In a simplified model where the distribution of each potential outcome, $Y_i(0)$ and $Y_i(1)$, is characterized by a Gaussian mixture with means $\mu_{t,z}$ for $z \in \{1, 2, ..., K\}$, Equation 4 can be made to hold exactly through parameterization. In particular, we would like to solve:

$$\textstyle\sum_{z=1}^{K} \widehat{\tau}(z)\widehat{\Pr}(Z(\mathbf{M}) = z) - \widehat{\overline{\tau}} = 0$$

Under this simplified model, $\widehat{\tau}(z) = \widehat{\mu}_{1,z} - \widehat{\mu}_{0,z}$, so

$$\textstyle\sum_{z=1}^{K} (\widehat{\mu}_{1,z} - \widehat{\mu}_{0,z})\widehat{\Pr}(Z(\mathbf{M}) = z) - \widehat{\overline{\tau}} = 0$$
$$(\widehat{\mu}_{1,z=1} - \widehat{\mu}_{0,z=1})\widehat{\Pr}(Z(\mathbf{M}) = 1) + \textstyle\sum_{z=2}^{K} (\widehat{\mu}_{1,z} - \widehat{\mu}_{0,z})\widehat{\Pr}(Z(\mathbf{M}) = z) - \widehat{\overline{\tau}} = 0$$

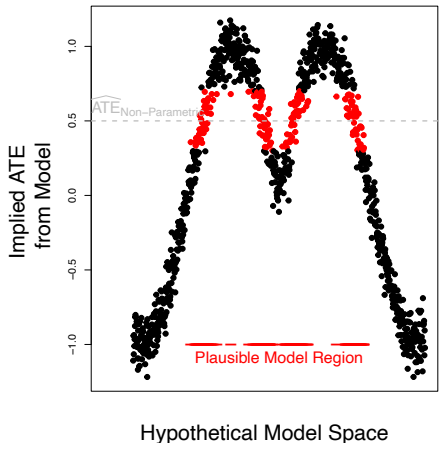

Figure A.6: Visualizing the plausible region for treatment effect heterogeneity models. The most plausible models are those where the implied ATE is close to the non-parametric estimate.

which implies

$$\widehat{\mu}_{0,z=1} = \widehat{\mu}_{1,z=1} \; - \; \left(\widehat{\overline{\tau}} - \sum_{z=2}^{K} (\widehat{\mu}_{1,z} - \widehat{\mu}_{0,z}) \widehat{\Pr}(Z(\mathbf{M}) = z) \right) \widehat{\Pr}(Z(\mathbf{M}) = 1)^{-1}.$$

Thus, in some modeling contexts, the exact non-parametric ATE can be recovered in the clustering model by parameterization.