Economic Value of Visual Product Characteristics

Ankit Sisodia
Purdue University, asisodia@purdue.edu

Vineet Kumar
Yale School of Management, vineet.kumar@yale.edu

February 2026

Key words: visual analytics, deep learning, demand models

1. Introduction

2. Literature Review

Our work draws on four research streams: disentangled representation learning in machine learning, structural demand estimation for differentiated products using the BLP class of random coefficient logit models, visual product design in marketing, and a growing body of work on incorporating unstructured data in consumer choice models. We draw on these streams to motivate a framework that learns interpretable visual product characteristics from images and incorporates them as additional product attributes in a BLP demand model.

2.1. Disentangled Representation Learning

Representation learning studies how high-dimensional observations can be encoded using lower-dimensional factors that retain information useful for downstream tasks (Bengio et al. 2013). This paper focuses on a branch called disentangled representation learning, which aims to isolate meaningful, independent factors of variation in data (Bengio et al. 2013). For example, the dSprites dataset (Higgins et al. 2017) contains 2D images generated by varying shape, size, color, and position. Disentanglement seeks to separate these factors into distinct latent dimensions. If successful, it can recover meaningful factors of variation even when the latent space is high-dimensional.

A large class of disentanglement approaches is based on variational autoencoders (VAEs), including β-VAE (Higgins et al. 2017, Burgess et al. 2017), FactorVAE (Kim and Mnih 2018), and β-TCVAE (Chen et al. 2018). A key challenge, however, is that in purely unsupervised settings there is no general theoretical guarantee that a learned representation is uniquely disentangled (Locatello et al. 2019). Learning semantically interpretable and stable latent dimensions typically requires some form of inductive bias or supervision. Locatello et al. (2020) showed that even limited, potentially imperfect supervision can be sufficient for model selection when learning disentangled representations.

Building on this insight, Sisodia et al. (2025) demonstrated that structured product characteristics available in standard datasets (e.g., brand, price, physical dimensions) can provide weak supervisory signals that help learn and select disentangled representations for product images without manually labeling visual attributes or pre-specifying the visual factors to recover. The learned disentangled representations can then be validated post hoc. Whether such machine-learned visual characteristics can be incorporated into structural

models of demand and competition, where they must interact with equilibrium pricing, endogeneity, and heterogeneous consumer preferences, remains an open question that the present paper addresses.

2.2. Demand Models

The random coefficient logit model described in Berry et al. (1995) is a foundational framework for estimating demand for differentiated products in economics and marketing. The framework addresses price endogeneity using instrumental variables in the spirit of Berry (1994), recognizing that consumer utility depends on observed product characteristics available to the econometrician as well as unobserved product characteristics observed by market participants but not by the econometrician. These unobservables can be correlated with price, motivating the instrumental-variable approach. By allowing heterogeneous tastes over product characteristics, the framework yields flexible substitution patterns and is well-suited to settings with aggregate market-level data and many differentiated products.

Subsequent work has extended the BLP framework along several dimensions. Nevo (2000) incorporated interactions between observed consumer demographics and product characteristics and showed how to include product fixed effects. Petrin (2002) and Berry et al. (2004) integrated micro data with market-level data, yielding richer substitution patterns and reducing instrument requirements. These extensions have improved how the econometrician models consumer heterogeneity and controls for unobservables. However, the set of product characteristics entering the demand model has remained limited to variables that can be recorded as structured data.

2.3. Visual Product Design

A longstanding literature in marketing argues that product form influences consumer preference and choice beyond purely functional attributes. Bloch (1995) proposes that product form generates cognitive and affective responses that affect approach and avoidance behavior, while Creusen and Schoormans (2005) highlights multiple roles of appearance in consumer choice, including aesthetic value, symbolic meaning, and communication of functional information. In the automotive context, Kang et al. (2019) shows that consumers make explicit trade-offs between visual form and functional attributes, and related work on design typicality and novelty suggests that relative visual positioning shapes consumer evaluation (Talke et al. 2009).

Despite this evidence, incorporating visual design directly into structural demand models has been challenging because visual design is inherently unstructured and difficult to quantify using standard product databases. Moreover, for structural estimation and interpretation, any measured visual characteristic must be interpretable in substantive terms, i.e., it should be possible to understand what a unit change represents. As a result, visual design is often absorbed into the unobserved component of utility, which can matter for demand estimation and inference when correlated with endogenous variables such as price.

This motivates approaches that transform product images into structured and interpretable product characteristics that can enter the utility specification and be valued within a structural model of demand.

2.4. Unstructured Data in Consumer Choice Models

A growing literature in marketing and economics leverages unstructured data, such as product images, titles, descriptions, and customer reviews, to enrich discrete choice models. A common approach is to transform unstructured inputs into low-dimensional representations (embeddings) using pre-trained deep learning models, reduce dimensionality using methods such as principal components analysis, and incorporate the resulting components as additional product characteristics in demand models. This work emphasizes that unstructured data can capture otherwise hard-to-quantify dimensions of differentiation and, in turn, improve the model's ability to recover substitution patterns relevant for counterfactual analyses.

Within this literature, several papers use unstructured data primarily to predict mean utilities or product intercepts. For example, Quan and Williams (2019) incorporate product image embeddings from pre-trained convolutional neural networks as shifters of mean utility intercepts, showing that visual information helps explain cross-sectional variation in product demand. Similarly, **?** uses large language models to predict utility intercepts for new products based on their textual descriptions. These approaches are valuable for measuring aspects of product positioning that are otherwise unobserved in structured datasets, but they are not designed to directly recover the covariance structure of utilities that governs substitution and competitive interactions in random coefficient models. Recent work goes beyond mean-utility shifters by incorporating text- and image-based representations directly into random coefficient demand models to capture richer substitution patterns from unstructured data (e.g., Compiani et al. (2025)). While their approach is scalable

across categories and does not require pre-specified attributes, the principal components that enter the model are less directly interpretable.

A complementary set of approaches uses survey-based approaches to measure perceived similarity and incorporate it into demand estimation. For instance, Dotson et al. (2016) elicit ratings of product images and use rating correlations to shift utility correlations, while Magnolfi et al. (2025) elicit relative similarity judgments (e.g., triplets) to construct embeddings that enter a random coefficient logit model. These methods can provide direct information about perceived similarity but typically require category-specific data collection.

Our paper contributes to this literature by proposing an approach that is both interpretable and structural. Rather than using opaque embeddings or principal components as product characteristics, we extract interpretable visual dimensions from product images using disentangled representation learning and incorporate them as additional observed product attributes in a BLP demand model. Our approach complements embedding-based structural methods by recovering a small set of interpretable visual dimensions that enter utility directly as product attributes. In doing so, it shifts systematic measured visual variation from the unobserved component ξ_{jt} into the utility specification, making it available for preference estimation and counterfactual analysis.

3. Methodology

We propose a two-stage approach to incorporate interpretable visual product characteristics into a structural model of demand and competition. In the first stage, we learn a low-dimensional, disentangled representation of product images using a variational autoencoder (VAE) with weak supervision from structured product characteristics (Sisodia et al. 2025). In the second stage, we estimate a random-coefficients logit demand model in the BLP class (Berry et al. 1995) and a supply-side pricing model, treating the learned visual characteristics as additional observed product characteristics. Endogeneity of price is addressed using instrumental variables. Table 1 summarizes the key notation.

3.1. Learning Visual Characteristics from Images

Our approach follows Sisodia et al. (2025) and builds on disentangled representation learning with VAEs (Kingma and Welling 2014, Higgins et al. 2017, Burgess et al. 2017, Kim and Mnih 2018, Chen et al. 2018). Let $\mathcal{I}$ denote a product image. A VAE consists of (i) an

encoder $q_\phi(\mathbf{h} \mid \mathcal{I})$ that maps an image to a latent vector $\mathbf{h} \in \mathbb{R}^D$ and (ii) a decoder $p_\theta(\mathcal{I} \mid \mathbf{h})$ that reconstructs the image from $\mathbf{h}$. We assume a standard Gaussian prior $p(\mathbf{h}) = \mathcal{N}(0, I)$, and a Gaussian variational posterior $q_\phi(\mathbf{h} \mid \mathcal{I}) = \mathcal{N}(\mu_\phi(\mathcal{I}), \mathrm{diag}(\sigma^2_\phi(\mathcal{I})))$.

Why weak supervision is needed. A core challenge is that in purely unsupervised settings there is no general theoretical guarantee that the learned representation is uniquely disentangled (Locatello et al. 2019). In practice, learning stable, semantically interpretable factors typically requires inductive biases and/or some form of supervision (Locatello et al. 2020). In our setting, the ground-truth visual attributes are not observed. We therefore use structured product characteristics as weak supervisory signals, following Sisodia et al. (2025). Intuitively, structured characteristics (e.g., brand, price, and other observed product attributes) may be correlated with underlying design elements and can guide the latent space toward economically meaningful variation.

Objective function. We use the β-TCVAE framework (Chen et al. 2018) augmented with a supervised loss. Let $\mathbf{y}$ denote a vector of supervisory signals for image $\mathcal{I}$ and let $\widehat{\mathbf{y}}(\mathbf{h})$ be predictions from a supervised head that takes $\mathbf{h}$ as input. The learning objective is:

$$\underbrace{\mathcal{L}(\theta, \phi; \mathcal{I}, \mathbf{h})}_{\text{Disentanglement Loss}} = \underbrace{-\mathbf{E}_{q_\phi(\mathbf{h}|\mathcal{I})}[\log p_\theta(\mathcal{I} \mid \mathbf{h})]}_{\substack{\text{Reconstruction} \\ \text{Loss}}} + \underbrace{I_q(\mathbf{h}, \mathcal{I})}_{\substack{\text{Mutual} \\ \text{Information} \\ \text{Loss}}} + \beta \underbrace{KL\left[q(\mathbf{h}) \,\middle\|\, \prod_{d=1}^{D} q(h_d)\right]}_{\substack{\text{Total Correlation} \\ \text{Loss}}}$$
$$+ \underbrace{\sum_{d=1}^{D} KL[q(h_d)||p(h_d)]}_{\substack{\text{Dimension-Wise} \\ \text{KL Divergence Loss}}} + \delta \underbrace{P(\widehat{\mathbf{y}(\mathbf{h})}, \mathbf{y})}_{\substack{\text{Supervised} \\ \text{Loss}}} \tag{1}$$

Here $q(\mathbf{h})$ and $q(h_d)$ denote the aggregate posterior and its marginals, and β controls the strength of the disentanglement penalty while δ controls the strength of weak supervision.

Hyperparameter and model selection. We tune (β, δ) using a grid search and select the best pair for each candidate supervisory-signal vector using cross-validated supervised loss, following Locatello et al. (2020). We then compare representations learned under different supervisory-signal vectors using Unsupervised Disentanglement Ranking (UDR) (Duan et al. 2020), which evaluates robustness of the learned representation across random initializations without requiring ground-truth labels. We select the supervisory-signal vector that yields the highest UDR score. We consider a pre-specified menu $\mathcal{Y}$ of candidate supervisory-signal vectors $\mathbf{y}_{jt}$ constructed from structured product characteristics

(defined in Appendix **??**). For each $\mathbf{y} \in \mathcal{Y}$ we tune (β, δ) by cross-validated supervised loss (Locatello et al. 2020), and then select $\mathbf{y}^\star$ as the supervisory-signal specification that yields the highest UDR score (Duan et al. 2020).

Constructing v_{jt}. The trained VAE produces a latent vector $\mathbf{h}_{jt}$ for each product image. We interpret a subset (or transformation) of these latent dimensions as learned visual product characteristics and denote them by v_{jt}. These v_{jt} are then treated as additional observed product characteristics in the structural demand model below. The VAE yields a latent vector $\mathbf{h}_{jt}$ for each image. Following standard practice in disentanglement work, we retain as *informative* the subset of latent dimensions whose KL contribution exceeds a fixed threshold (Appendix **??**) and define v_{jt} as the raw (unstandardized) scores on these informative dimensions. In the results we interpret the informative dimensions via latent traversals and focus on the subset corresponding to design-related factors (e.g., body shape, boxiness, grille height, grille width); we exclude a non-design dimension related to image appearance (e.g., lighting/contrast) from the structural demand model.

3.2. Demand and Supply Model

We estimate a model of market equilibrium using a random-coefficients logit demand system in the BLP class (Berry et al. 1995) with the specification in Berry et al. (1999). In each market $t = 1, \ldots, T$, consumers choose among J_t differentiated products and an outside option $j = 0$.

Utility. Let x_{jt} denote structured (non-price) product characteristics and let v_{jt} denote the learned visual characteristics from Section 4.1. Consumer i's indirect utility from product j in market t is

$$U_{ijt} = x_{jt}\,\overline{\beta_1} + v_{jt}\,\overline{\beta_2} - \alpha\,\frac{p_{jt}}{y_{it}} + \sum_{k_1} \sigma_{\beta_1}^{k_1} x_{jt}^{k_1} \nu_{it}^{k_1} + \sum_{k_2} \sigma_{\beta_2}^{k_2} v_{jt}^{k_2} \nu_{it}^{k_2} + \xi_{jt} + \epsilon_{ijt}, \tag{2}$$

where p_{jt} is price, y_{it} is consumer income, ν_{it} are consumer-specific taste shocks (assumed mean zero), ϵ_{ijt} is an i.i.d. Type-I extreme value error, and ξ_{jt} is the remaining product-market unobservable after controlling for (x_{jt}, v_{jt}). Price is endogenous because firms observe ξ_{jt} when setting p_{jt}.

We set $\sigma_{\text{price}} = 0$ so that heterogeneity in price sensitivity operates through observable income variation via $-\alpha p_{jt}/y_{it}$, following the standard approach in which demographics are the primary source of price-sensitivity heterogeneity (Berry et al. 1995).

Shares. Let δ_{jt} denote mean utility and μ_{ijt} the individual deviation:

$$\begin{aligned}\delta_{jt} &= x_{jt}\,\overline{\beta_1} + v_{jt}\,\overline{\beta_2} + \xi_{jt},\\ \mu_{ijt} &= -\alpha\,\frac{p_{jt}}{y_{it}} + \sum_{k_1} \sigma^{k_1}_{\beta_1} x^{k_1}_{jt}\,\nu^{k_1}_{it} + \sum_{k_2} \sigma^{k_2}_{\beta_2} v^{k_2}_{jt}\,\nu^{k_2}_{it}.\end{aligned} \tag{3}$$

Choice probabilities and market shares are

$$s_{ijt} = \frac{\exp(\delta_{jt} + \mu_{ijt})}{\sum_{\ell \in \{0,1,\ldots,J_t\}} \exp(\delta_{\ell t} + \mu_{i\ell t})}, \qquad s_{jt} = \int s_{ijt}\, dF(i), \tag{4}$$

and we approximate the integral by simulation (we use 1000 scrambled Halton draws per market).

Supply. Firms play a static, full-information, simultaneous-move pricing game each period (Nash–Bertrand pricing). Let $f(j)$ denote the firm that owns product j, and let $\mathcal{J}_{ft} = \{k \in J_t : f(k) = f\}$ denote firm f's product set in market t. Firm f chooses prices for all $j \in \mathcal{J}_{ft}$ to maximize profits. The multi-product first-order condition for each $j \in \mathcal{J}_{ft}$ is:

$$s_{jt} + \sum_{k \in \mathcal{J}_{f(j)t}} (p_{kt} - c_{kt}) \frac{\partial s_{kt}}{\partial p_{jt}} = 0. \tag{5}$$

We parameterize log marginal costs as

$$\ln(c_{jt}) = x_{jt}\gamma_1 + w_{jt}\gamma_2 + \omega_{jt}, \tag{6}$$

where w_{jt} are observed cost shifters and ω_{jt} is an unobserved cost shock. We do not include visual characteristics directly in the cost equation; thus any cost-side variation correlated with visual design is absorbed into ω_{jt}. The vector w_{jt} contains observed cost shifters used in the marginal cost equation and is defined in Section **??** (Table **??**).

3.3. Instruments and Estimation

Moment conditions. Price is endogenous because p_{jt} may be correlated with ξ_{jt}. Let Z^D_{jt} denote demand instruments and Z^S_{jt} denote supply instruments. We use moment conditions:

$$\mathbb{E}[\xi_{jt} Z^D_{jt}] = 0, \qquad \mathbb{E}[\omega_{jt} Z^S_{jt}] = 0. \tag{7}$$

BLP-style instruments. We construct BLP instruments using sums of observed (non-price) characteristics of other products in the same market, distinguishing products made by the same firm from those made by rival firms (Berry et al. 1995). Let $\tilde{x}_{jt}$ denote the observed characteristics used in instrument construction (excluding price). Then the standard BLP-style sums are:

$$Z_{jt}^{\text{BLP}} = \Big\{ \tilde{x}_{jt},\, w_{jt},\, \sum_{\substack{k \in J_t \\ k \neq j,\, f(k)=f(j)}} \tilde{x}_{kt},\, \sum_{\substack{k \in J_t \\ f(k) \neq f(j)}} \tilde{x}_{kt},\, \sum_{\substack{k \in J_t \\ k \neq j,\, f(k)=f(j)}} 1,\, \sum_{\substack{k \in J_t \\ f(k) \neq f(j)}} 1 \Big\}. \tag{8}$$

These instruments leverage the idea that the competitive environment (as summarized by rivals' characteristics) shifts equilibrium prices, while the timing of characteristic choice relative to pricing supports orthogonality with ξ_{jt}. In constructing BLP-style instruments we use observed structured (non-price) characteristics x_{jt}, learned visual characteristics v_{jt}, and cost shifters w_{jt}, forming the usual sums of characteristics over own-firm and rival products within each market. We treat v_{jt} as observed product characteristics for instrument construction; alternative instrument sets that exclude v_{jt} are reported as robustness checks (Appendix **??**).

GMM estimation. Let ζ collect all parameters in the demand and supply system. Define the sample moments

$$g(\zeta) = \begin{bmatrix} \frac{1}{N} \sum_{j,t} \xi_{jt}(\zeta)\, Z_{jt}^{D} \\ \frac{1}{N} \sum_{j,t} \omega_{jt}(\zeta)\, Z_{jt}^{S} \end{bmatrix}, \tag{9}$$

and estimate ζ by minimizing

$$\widehat{\zeta} = \arg\min_{\zeta}\; g(\zeta)^{\top} W\, g(\zeta), \tag{10}$$

where W is a weighting matrix. We implement a two-step procedure: we first obtain consistent parameter estimates with an initial W, then update W using the estimated moment covariance to obtain an efficient GMM estimator. Following Conlon and Gortmaker (2020), we also implement a two-step procedure with approximate optimal instruments: we (i) estimate the model using the constructed instruments to obtain consistent estimates and (ii) use these estimates to compute approximate optimal instruments and re-estimate the model for improved efficiency.

4. Methodology

Our method to obtain visual product characteristics that are independent and human-interpretable employs a disentanglement-based approach using Variational Autoencoders (VAEs). However, disentanglement methods typically require ground truth visual characteristics as supervisory signals, which are not available in our case since we aim to discover these characteristics. To address this challenge, we use structured product characteristics as supervisory signals. Next, the demand model identifies how those visual characteristics (alongside other more conventional structured product characteristics) affect a demand system comprising a competitive market of firms and heterogeneous consumers. Section 4.1 describes the disentanglement-based deep learning model that identifies visual characteristics from product images. Section 4.2 describes the demand model, including its supply-side and demand-side assumptions, as well as parameter estimation. We provide a table of notation in Table 1.

4.1. Disentanglement with Variational Autoencoder

Our approach to discovering visual product characteristics follows the approach used by (Sisodia et al. 2025). It employs the use of Variational Autoencoders (VAEs) (Kingma and Welling 2014), a class of deep generative models that learn to encode input data into a latent space, and simultaneously enable the generation of new data samples from this latent space. In line with previous research (Higgins et al. 2017, Burgess et al. 2017, Chen et al. 2018, Kim and Mnih 2018), we utilize these VAEs specifically for disentangled representation learning (Bengio et al. 2013), allowing us to identify statistically independent and semantically meaningful visual factors that vary across products in our dataset.

We consider a dataset of product images, each of which we assume is generated by an underlying distribution parameterized by visual characteristics. Our goal is to learn a low-dimensional representation of these visual characteristics that captures the most salient factors of variation in the product images. To achieve this, we utilize a VAE, consisting of an encoder network $q_\phi(\mathbf{z}|\mathbf{x})$, which compresses each high-dimensional product image $\mathbf{x}$ into a lower-dimensional latent space of visual characteristics $\mathbf{z}$, and a decoder network $p_\theta(\mathbf{x}|\mathbf{z})$, which reconstructs images from these latent representations. Both the encoder and decoder are deep neural networks, parameterized by ϕ and θ, respectively.

The VAE framework assumes a generative model where the latent visual characteristics $\mathbf{z}$ are first sampled from a prior distribution $p(\mathbf{z})$, set to an isotropic unit Gaussian $\mathcal{N}(0, \mathbf{I})$.

Subsequently, product images are generated through the decoder distribution $p_\theta(\mathbf{x}|\mathbf{z})$. During training, the encoder network approximates the true posterior distribution $p(\mathbf{z}|\mathbf{x})$ with a variational distribution $q_\phi(\mathbf{z}|\mathbf{x})$, typically modeled as a multivariate Gaussian with diagonal covariance, i.e., $\log q_\phi(\mathbf{z}|\mathbf{x}) = \log \mathcal{N}(\mathbf{z}; \boldsymbol{\mu_d}, \boldsymbol{\sigma_d}^2\mathbf{I})$, where $\boldsymbol{\mu_d}$ and $\boldsymbol{\sigma_d}$ are the mean and standard deviation outputs of the encoder network.

While various disentanglement methods built on the VAE framework—such as β-VAE (Higgins et al. 2017, Burgess et al. 2017), FactorVAE (Kim and Mnih 2018), and β-TCVAE (Chen et al. 2018)—have shown promising results, achieving true disentanglement in these representations faces fundamental theoretical challenges. Specifically, Locatello's theorem (Locatello et al. 2019) showed that unsupervised disentanglement is fundamentally limited without additional structure or assumptions—commonly referred to as inductive biases. These biases can be implicit or explicit and are essential for learning meaningful and interpretable representations from limited data. Some examples of inductive biases include architectural choices (e.g., convolutional neural networks for image data), prior distributions for latent variables, regularization techniques, and data augmentation. Consequently, recent efforts in the deep learning literature have focused on improving disentanglement methods by utilizing benchmark datasets with known ground truth labels corresponding to each visual characteristic (Locatello et al. 2020). However, the very visual characteristics we aim to discover are precisely these ground truth labels.

Instead, we follow Sisodia et al. (2025) in using structured product characteristic (such as brand or price) as supervisory signals to address Locatello's theorem (Locatello et al. 2019). Structured product characteristics, such as brand, material, performance attributes, and price, can be informative for guiding the learning of disentangled representations due to their potential correlation with visual characteristics. For example, a product's brand can significantly influence its visual appearance. A luxury brand like Louis Vuitton may be associated with specific visual design elements that set them apart from other brands. Similarly, the price point of a product can often be reflected in its visual design, with higher-priced items frequently exhibiting more refined or elaborate visual features compared to lower-priced alternatives. The supervised loss term directly quantifies the discrepancy between predicted structured characteristics derived from the latent space and their actual observed values. The machine learning literature has shown that even weak

supervision with a set of signals that has some correlation to the ground truth will help obtain a disentangled representation in practice (Locatello et al. 2020).

To promote disentanglement of the learned representations, we incorporate additional regularization terms following the β-TCVAE approach (Chen et al. 2018). We minimize the objective function in Equation 11 to learn a disentangled representation of the visual characteristics present in the product image data.

$$\underbrace{\mathcal{L}(\theta,\phi;\mathbf{x},\mathbf{z})}_{\text{Disentanglement Loss}} = \underbrace{-\mathbf{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right]}_{\substack{\text{Reconstruction}\\ \text{Loss}}} + \underbrace{I_q(\mathbf{z},\mathbf{x})}_{\substack{\text{Mutual}\\ \text{Information}\\ \text{Loss}}} + \beta \underbrace{KL\left[q(\mathbf{z})||\prod_{j=1}^{J} q(z_j)\right]}_{\substack{\text{Total Correlation}\\ \text{Loss}}}$$
$$+ \underbrace{\sum_{j=1}^{J} KL\left[q(z_j)||p(z_j)\right]}_{\substack{\text{Dimension-Wise}\\ \text{KL Divergence Loss}}} + \delta \underbrace{P(\widehat{\mathbf{y}(\mathbf{z})},\mathbf{y})}_{\substack{\text{Supervised}\\ \text{Loss}}} \tag{11}$$

The objective function in Equation 11 includes five terms.

The first term, the reconstruction loss (-$\mathbf{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right]$), evaluates how well the model can regenerate the original image $\mathbf{x}$ from its latent representation or visual characteristics $\mathbf{z}$. By minimizing this term, we ensure that the latent space retains the key visual elements needed to accurately reproduce each product image. This acts as a fidelity constraint, grounding the learning process in the observable input data.

The second term is the mutual information loss ($I_q(\mathbf{z},\mathbf{x})$). It is a measure of how much information the latent visual characteristic $\mathbf{z}$ contain about the input data $\mathbf{x}$. In the context of disentanglement $I_q(\mathbf{z},\mathbf{x}) = D_{KL}(q(\mathbf{z},\mathbf{x})||q(\mathbf{z})p(\mathbf{x}))$. Higher mutual information means the latent variables capture more information about the inputs. This helps ensure the representation is meaningful and useful. If this term is completely minimized, then the latent visual characteristics would be independent of the input image $\mathbf{x}$. At the same time, if this term is maximized without constraint, then the model might simply memorize the training data without learning useful structure.

Next, the total correlation loss ($KL\left[q(\mathbf{z})||\prod_{j=1}^{J} q(z_j)\right]$) measures the dependence between the individual dimensions of the latent representation $\mathbf{z}$. By penalizing the KL divergence between the joint distribution $q(\mathbf{z})$ and the product of its marginals $\prod_{j=1}^{J} q(z_j)$, this loss term promotes statistical independence among the learned visual characteristics. This is

critical for disentanglement: we want each latent variable to reflect a distinct and independent factor of variation. By encouraging the joint posterior distribution to approximate a factorized distribution, the model learns a more disentangled latent structure. The hyperparameter β controls the strength of this penalty.

A fourth term, the dimension-wise KL divergence ($\sum_{j=1}^{J} KL\left[q(z_j)||p(z_j)\right]$), ensures that each individual latent dimension remains close to its prior distribution (typically a standard Gaussian). This regularization serves multiple purposes: it helps smooth the latent space, discourages redundant encodings, and implicitly controls model complexity by allowing the model to "turn off" unnecessary latent dimensions.

Finally, we include a supervised loss term ($P(\widehat{\mathbf{y}(\mathbf{z})}, \mathbf{y})$) that measures the discrepancy between the predicted supervisory signals $\widehat{\mathbf{y}(\mathbf{z})}$ and the actual observed signals $\mathbf{y}$. This allows us to steer the latent space using weak but informative signals found in structured product characteristics that could be potentially correlated with underlying visual characteristics. The hyperparameter δ balances the importance of this supervised objective with the other unsupervised loss terms.

Our model architecture, detailed in Appendix A, is a modified version of the one used by Burgess et al. (2017). We adapt the architecture to work with 128×128 pixel images and incorporate the supervisory signals from our model of market equilibrium and structured product characteristics. The architecture consists of an encoder neural network, a decoder neural network, and a supervised neural network, which work together to learn disentangled visual representations and predict the supervisory signals.

In addition to the model parameters that are learned during training, we also make modeling choices in the form of hyperparameters. These hyperparameters impact the estimation process but are not directly estimated with the model. The separation between parameters and hyperparameters is common in machine learning, as hyperparameters are typically set before training and control the learning process, while parameters are learned from data during training (Goodfellow et al. 2016, Murphy 2012, Bishop 2006). Examples of hyperparameters include the learning rate, batch size, and regularization strengths like β and δ in our model. The underlying logic is that hyperparameters define the model's capacity, regularization, and optimization settings, which need to be tuned separately from the model parameters to achieve the best performance and generalization. While there are techniques for automatically searching for optimal hyperparameter values, such as grid

search, random search, and Bayesian optimization (Bergstra and Bengio 2012, Snoek et al. 2012), hyperparameters are typically not learned directly during the main model training process.

In our model, we have two key hyperparameters: β and δ. The hyperparameter β controls the weight of the total correlation loss within the disentanglement loss (Chen et al. 2018). This term encourages the model to learn statistically independent latent factors. The hyperparameter δ represents the weight of the supervised loss when incorporated into the overall loss function. A higher value of δ prioritizes the model's ability to predict the vector of supervisory signals, relative to other loss terms like mutual information or reconstruction loss. However, placing too much emphasis on the supervised loss could potentially reduce the quality of disentanglement. On the other hand, setting δ to zero implies that we are not addressing the impossibility theorem (Locatello et al. 2019) and thus have no theoretical guarantees for discovering disentangled visual characteristics.

To select the optimal values for β and δ, we follow the approach proposed by Locatello et al. (2020). We perform a grid search over a range of values for these hyperparameters and select the combination that yields the lowest 10-fold cross-validated supervised loss for each vector of supervisory signals. This approach allows us to find the best balance between the disentanglement and supervised objectives, tailored to each specific set of supervisory signals. To compare the quality of disentangled representations produced by different vectors of supervisory signals, we use the Unsupervised Disentanglement Ranking (UDR) metric proposed by Duan et al. (2020). UDR is an automated method that assesses the robustness of disentangled representations to variance at different starting points without requiring access to the ground truth data generative process. It relies on the assumption that for a particular dataset, a disentangling VAE will converge on the same disentangled representation up to certain isomorphic transformations. We select the hyperparameters β and δ based on the lowest supervised loss for each vector of supervisory signals and then choose the supervisory signals that yield the highest UDR score. The details of the UDR algorithm are provided in Appendix B.

4.2. Model of Market Equilibrium

We use the BLP demand model (Berry et al. 1995) with the specification laid out in Berry et al. (1999) to estimate a model of market equilibrium. Table 1 presents the notation used in the demand model.

Table 1 Table of Notation in Demand Model

Symbol	Meaning
j	Products
t	Markets
i	Consumers
f	Firms
T	Number of Markets
J_t	Number of products in market t
I_t	Number of consumers in market t
F_t	Number of firms in market t
ζ	Model Parameters
ζ_1	Linear demand-side parameters
ζ_2	Non-linear common parameters
ζ_3	Linear supply-side parameters
p_{jt}	Price
c_{jt}	Marginal Cost
x_{jt}	Observed product characteristic
v_{jt}	Visual product characteristic
U_{ujt}	Indirect utility
δ_{jt}	Mean utility
μ_{ijt}	Heterogeneous utility
ϵ_{ijt}	Idiosyncratic taste shock
d_{ijt}	Choice indicator
s_{ijt}	Choice probability
s_{jt}	Market share
ξ_{jt}	Demand-side structural error
ω_{jt}	Supply-side structural error
Z^D	Demand Instruments
Z^S	Supply Instruments
W	Weighting matrix
g	Sample Moments

Consumers: In each market $t = 1, \ldots, T$, there are J_t differentiated goods and I_t consumers. For each market, we observe average quantities, prices and product characteristics for all J_t products.

Consistent with the standard BLP model, the indirect utility of consumer i from purchasing product j in market t is a function of observed product characteristics $\mathbf{x_{jt}}$, unobserved product-market characteristics ξ_{jt}, price p_{jt}, consumer characteristics $\nu_{\mathbf{it}}$ and unknown parameters, ζ. The total number of observed characteristics for the product is K. We use the specification written in Equation (12). Here, price p_{jt} is endogeneous, since it could be based on the unobserved product-market characteristics ξ_{jt}, and hence correlated with it. The indirect utility is specified as:

$$U_{ijt} = \mathbf{x_{jt}}\overline{\beta_1} - \alpha p_{jt}/y_{it} + \xi_{jt} + \sum_{k_1}(\sigma_{\beta_1}^{k_1} x_{jt}^{k_1} \nu_{it}^{k_1}) + \epsilon_{ijt} \tag{12}$$

where $\mathbf{x_{jt}}$ or observed product characteristics only includes structured product characteristics, p_{jt} is the price of product j in market t, y_{it} is the income of the consumer i in

market t, ξ_{jt} is the unobserved product-market characteristic, $\nu_{it}^{k_1}$ represents consumers i's taste for characteristic k_1 in market t, and finally, ϵ_{ijt} denotes a mean-zero idiosyncratic taste shock. The unobserved product-market characteristics can reflect hard to quantify aspects of the product such as quality or style. The unobserved product characteristics can be decomposed into visual product characteristics $\mathbf{v_{jt}}$ and rest of the unobserved product-market characteristics $\tilde{\xi}_{jt}$. This decomposition is written in Equation 13. In a demand model without visual characteristics, this would be same as the typical BLP structural error ξ_{jt}.

Note that, even after we account for visual characteristics, the remaining unobserved product-market characteristics $\tilde{\xi}_{jt}$ may still contain some unobserved visual characteristic, as well as any other aspects of unobservable quality.

$$\xi_{jt} = \mathbf{v_{jt}}\overline{\beta_2} + \sum_{k_2}(\sigma_{\beta_2}^{k_2} v_{jt}^{k_2} \nu_{it}^{k_2}) + \tilde{\xi}_{jt} \tag{13}$$

Each consumer i in market t has unit demand. Consumers choose among $J_t = \{0, 1, \ldots, J_t\}$ discrete choices including the outside good, denoted by $j = 0$. The utility of the outside good represents the choice of not purchasing any product in the market and is given by $U_{i0t} = \epsilon_{i0t}$. Consumers select the alternative (including outside good) with the highest utility:

$$d_{ijt} = \begin{cases} 1 \text{ if } U_{ijt} > U_{ikt} \text{ for all } k \neq j \\ 0 \qquad\qquad \text{otherwise} \end{cases} \tag{14}$$

Note that as in BLP, we can decompose the indirect utility in Equation (12) into a mean utility, $\delta_{\mathbf{jt}}$ and a deviation from that mean, μ_{ijt}, in Equation (15).

$$\begin{aligned} \delta_{jt}(\mathbf{x_{jt}}, p_{jt}, \xi_{jt}; \zeta_{\mathbf{1}}) &= \mathbf{x_{jt}}\overline{\beta_1} + \mathbf{v_{jt}}\overline{\beta_2} + \tilde{\xi}_{jt} \\ \mu_{ijt}(\mathbf{x_{jt}}, p_{jt}, \nu_{\mathbf{ijt}}, y_i; \zeta_{\mathbf{2}}) &= -\alpha p_{jt}/y_{it} + \sum_{k_1}(\sigma_{\beta_1}^{k_1} x_{jt}^{k_1} \nu_{it}^{k_1}) + \sum_{k_2}(\sigma_{\beta_2}^{k_2} v_{jt}^{k_2} \nu_{it}^{k_2}) + \epsilon_{ijt} \end{aligned} \tag{15}$$

where $\mathbf{x_{jt}}$ or observed product characteristics only includes structured product characteristics, $\mathbf{v_{jt}}$ is the visual product characteristics, p_{jt} is the price of product j in market t, y_{it} is the income of the consumer i in market t, $\tilde{\xi}_{jt}$ is the unobserved product-market characteristic, $\nu_{it}^{k_1}$ and $\nu_{it}^{k_2}$ represents consumers i's taste for structured product characteristics k_1

and visual characteristics k_2 in market t, and finally, ϵ_{ijt} denotes a mean-zero idiosyncratic taste shock.

We denote $\zeta = (\zeta_{\mathbf{1}}, \zeta_{\mathbf{2}})$, a vector of all the parameters in the model. The vector $\zeta_{\mathbf{1}}$ contain the linear parameters or the mean preference on $\mathbf{x_{jt}}$, i.e. $\overline{\beta_{\mathbf{1}}}$ and on $\mathbf{v_{jt}}$, i.e. $\overline{\beta_{\mathbf{2}}}$. These preferences are common across all consumers. The vector $\zeta_{\mathbf{2}}$ contain the nonlinear parameters or the standard deviation from mean preference i.e. $\sigma_{\beta_{\mathbf{1}}}$ and $\sigma_{\beta_{\mathbf{2}}}$ as well as the term on the price α. These nonlinear parameters introduce heterogeneity in preferences over structured product characteristics. We fix $\sigma_{\text{price}} = 0$, so that all heterogeneity in price sensitivity is driven by observable income variation through the $-\alpha p_{jt}/y_{it}$ term rather than an unobserved random coefficient. This follows the standard BLP approach in which demographics are the primary source of price sensitivity heterogeneity (Berry et al. 1995).

Using the standard assumption that ϵ_{ijt} are i.i.d. with the Type I extreme value distribution, the probability s_{ijt} that consumer i chooses product j in market t and aggregate product market shares are given by equation (16) below. The integral in Equation (16) does not have a closed-form solution and is approximated by simulation. We use 1000 scrambled Halton draws per market to compute simulated market shares.

$$s_{ijt} = \frac{\exp(\delta_{jt} + \mu_{ijt})}{\Sigma_{l \in J_t} \exp(\delta_{lt} + \mu_{ilt})} \qquad \text{and} \qquad s_{jt} = \int \frac{\exp(\delta_{jt} + \mu_{ijt})}{\Sigma_{l \in J_t} \exp(\delta_{lt} + \mu_{ilt})} dFi \tag{16}$$

Firms: We assume that automobile firms, indexed by f and part of a set F_t, play a static, full information, simultaneous move pricing game each period. Firms choose the price levels of all their models (products) with the objective of maximizing overall profit. We specify a constant marginal cost c_{jt} for a product j in market t. The pricing first order condition for vehicle j is given by Equation (17).

$$s_{jt} + \Sigma_{j \in J_t} (p_{jt} - c_{jt}) \frac{\partial s_{jt}}{\partial p_{jt}} = 0 \tag{17}$$

We parameterize log marginal costs as written below in Equation (18).

$$\ln(c_{jt}) = \mathbf{x_{jt}} \gamma_{\mathbf{1}} + \mathbf{w_{jt}} \gamma_{\mathbf{2}} + \omega_{jt} \tag{18}$$

where $\mathbf{x_{jt}}$ are product characteristics, $\mathbf{w_{jt}}$ are observable cost-shifters and ω_{jt} are unobserved cost-shifters. We can estimate the marginal costs for each product when we solve the supply model jointly with the demand model. We do not explicitly include visual characteristics in the supply-side and so they are assumed to be part of the unobservables.

Instruments: In this demand model, we assume that a consumer's utility depends up on the observed product characteristics as well as unobserved (to the researcher) product characteristics. Firms observe these unobserved product characteristics and set then set prices, which implies that price is endogenous and necessitates the use of instruments. There are multiple options for instruments. First, we could use *exogenous cost shifters*. These are valid if we assume that firms respond to cost shifts by changing prices, and not by changing product characteristics. Second, we could use *observed product characteristics* other than price. This would be valid if we make a timing assumption that firms first set observed product characteristics, then observing the "unobserved" product characteristics (structural error), and then set prices. This assumption would be supported by the observation that firms change prices frequently, whereas product characteristics are altered less frequently. Third, we could use Hausman instruments, a common example of which includes prices in other markets, if we have multiple markets that have the same product. Finally, we could use observed product characteristics of other products. One example of these instruments is referred to as *BLP instruments* in which we take sums of characteristics of other products made by the same firm and sums of characteristics of all other firms. We present them in Equation (19). Note that the instruments used in Sudhir (2001), who calculated average of characteristics only within the same segment and not the entire market, so as to reflect the localized nature of competition, are similar in spirit to differentiation IVs (Gandhi and Houde 2019).

$$Z_{BLP} = \{1, x_{jt}, w_{jt}, \Sigma_{j \in J_t \ \{j\}} 1, \Sigma_{j \notin J_t} 1, \Sigma_{j \in J_t \ \{j\}} x_{jt}, \Sigma_{j \notin J_t} x_{jt}\} \tag{19}$$

With the addition of demand instruments Z_{jt}^D, we construct demand-side moment conditions of the form $E[\tilde{\xi}_{jt} Z_{jt}^D] = 0$. Similarly, we also construct supply-side moment conditions of the form $E[\omega_{jt} Z_{jt}^S] = 0$ using supply instruments Z_{jt}^S.

GMM Estimation: We construct a GMM estimator using both supply-side and demand-side moment conditions.

$$g(\theta) = \begin{bmatrix} \frac{1}{N} \Sigma_{jt} E[\xi_{jt} Z_{jt}^D] \\ \frac{1}{N} \Sigma_{jt} E[\omega_{jt} Z_{jt}^S] \end{bmatrix} \tag{20}$$

We construct a nonlinear GMM estimator for ζ with some weighting matrix W in Equation (21). We solve this problem twice. First, we obtain a consistent estimate of W and then an efficient GMM estimator.

$$\widehat{\theta} = \min_{\theta} \ g(\zeta)' W g(\zeta) \tag{21}$$

Following Conlon and Gortmaker (2020), we implement a two-step procedure with approximate optimal instruments. We first estimate the model using the constructed instruments described above to obtain consistent parameter estimates. We then use these estimates to compute approximate optimal instruments and re-estimate the model, which yields more efficient estimates.

5. Empirical Setting and Results

We focus on the automobile industry in the United Kingdom (UK) from 2008 to 2017, which is well-suited for this analysis given the importance of visual design in consumer purchase decisions and the competitive nature of the market. We begin by describing our dataset, which combines information on automobile characteristics, market shares, and images. Next, we discuss the visual characteristics discovered through our disentangled representation learning approach, comparing the performance of supervised and unsupervised methods. To validate the interpretability and quantification of these visual characteristics, we present the results of human subject surveys. Finally, we incorporate the learned visual characteristics into our BLP demand model and present estimation results, including how design factors affect consumer choices.

5.1. Data

We compiled a data set covering 2008 through 2017 consisting of automobile characteristics, market shares and their images from the United Kingdom (UK). We obtain information on sales and images of the automobiles from DVM-CAR (Huang et al. 2021). Market research studies have shown that up to 70% of consumers identify and judge automobiles by the appearance of headlights and grille located on the face of the automobile.[1] So we only select the images of the front face of the automobiles and ignore other views. Since our sales data comes at the make-model level, we choose the median product characteristic across trims. Our estimation sample includes passenger cars with at least 10 registered sales per model-year. The final sample comprises approximately 2,125 model-year observations covering 326 distinct models from 44 makes across 10 years, representing roughly 96% of sales in the UK over the sample period.[2][3]

[1] URL: https://www.wsj.com/articles/SB114195150869994250

[2] We do not have images for 10% of make-model-years that contribute to 4% of sales in the 2008-2017 period.

[3] We also exclude battery electric vehicles. EVs represented a negligible share of UK sales in our sample.

We collected manufacturer suggested retail prices (MSRP), and characteristics of all automobiles sold in the UK from 2008-2017 from Parker's. Our product characteristics include horsepower, weight, length, width, and carbon dioxide emissions. Prices in all years are deflated to 2015 UK using the consumer price index. We supplemented the Parker's information with additional information, including vehicle country of production and company ownership information. We also supplemented additional information from the Office of National Statics (ONS), UK. We gathered the price of ultra low sulphur petrol per litre and ultra low sulphur diesel per litre as well as the number of households in the UK. Following Berry et al. (1995), we construct a fuel cost–adjusted efficiency measure, miles per pound sterling (MP£). We infer fuel consumption from vehicle-level carbon dioxide emissions and prevailing fuel prices, and compute MP£ as the inverse of fuel cost per mile.

We define a market as the UK new automobile market in a given year, spanning 2008–2017. The potential market size is the total number of UK households in a year, obtained from the ONS. Each household is assumed to purchase at most one new vehicle per year. Also, the random-coefficients specification requires a distribution of consumer demographics. We use annual mean and median real equivalised household disposable income for the UK, published by the ONS, expressed in tens of housands of pounds. For each market-year, we fit a lognormal distribution by setting $\mu_t = \ln(\text{median}_t)$ and $\sigma_t = \sqrt{2(\ln(\text{mean}_t) - \mu_t)}$, and draw 1,000 simulated consumers with equal weights (Nevo 2001).

In Table 2, we display sales-weighted summary statistics for the products at the make-model-year level. The variables include quantity, price (in £), horsepower, weight (in kgs), space (measured as length times width in m^2), grams of carbon dioxide emitted per kilometer, and miles that can be driven per pound sterling (MP£). We report sales-weighted means for each variable. We see that automobiles have improved in terms of both power and fuel efficiency over these ten years.

Table 2 Summary Statistics by Market

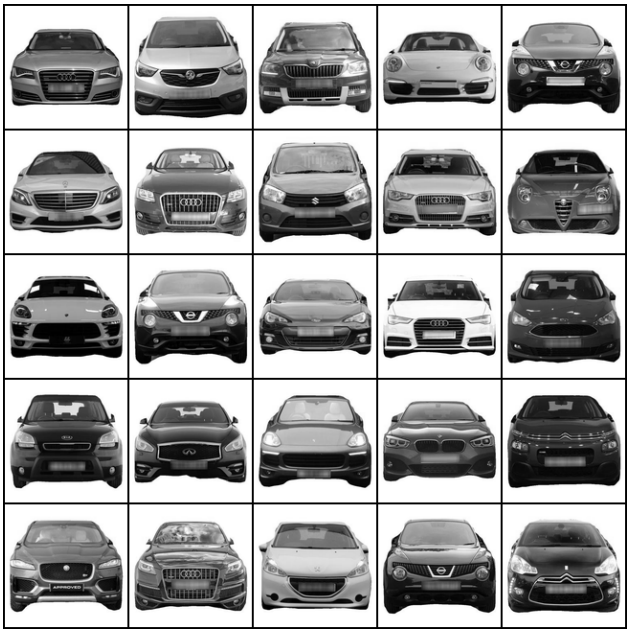
Market	Models	Quantity (units)	Price (£)	HP (HP)	Weight (kg)	Space (m^2)	CO_2 (g/km)	MP£ (miles/£)
2008	189	7411.98	21356.97	119.56	1319.01	8.03	156.80	7.51
2009	204	7151.62	21089.88	114.85	1291.79	7.98	148.05	8.88
2010	209	7529.94	21397.50	117.21	1298.75	8.05	142.44	8.23
2011	203	7227.45	21892.95	124.17	1329.86	8.11	138.48	7.67
2012	217	7436.30	21842.88	124.42	1331.91	8.19	131.64	8.17
2013	212	8657.40	21678.64	123.31	1319.19	8.22	126.66	8.77
2014	221	9107.77	22003.75	124.22	1315.50	8.28	123.65	9.52
2015	222	9518.27	23151.88	127.40	1323.26	8.37	121.30	11.15
2016	231	9712.38	24087.19	130.93	1332.52	8.46	122.26	11.43
2017	217	9070.10	24659.27	134.61	1347.75	8.48	124.16	10.66
All	2125	8323.49	22462.54	124.76	1322.08	8.24	131.83	9.40

Price is endogenous in the demand system and requires instruments. Following Grieco et al. (2021), we construct a cost-side instrument using the price level of consumer expenditure (pl_con) from the Penn World Table, version 9.1 (Feenstra et al. 2015), for the country where each vehicle is primarily produced, lagged by one year. This variable captures the real exchange rate (RXR), defined as the purchasing power parity exchange rate divided by the nominal exchange rate, relative to the United States. We normalize this measure relative to the UK by dividing each country's RXR by the UK's RXR in the same year, so that UK-produced vehicles receive an instrument value of one. The instrument varies across products within a market because vehicles sold in the UK are manufactured in multiple countries. Our sample includes vehicles produced in twelve countries: the Czech Republic, France, Germany, Italy, Japan, South Korea, Malaysia, Romania, Spain, Sweden, the United Kingdom, and the United States. The RXR provides plausibly exogenous variation in prices through two channels. First, increases in local wages or input costs in the country of manufacture raise production costs and are reflected in a higher RXR through the PPP component. Second, movements in nominal exchange rates affect the cost of imported vehicles in the UK market. Both channels shift marginal costs without directly affecting UK consumer preferences for vehicle characteristics, satisfying the exclusion restriction. The real exchange rate enters both the demand and supply instrument sets, as it satisfies the exclusion restriction for both: it is uncorrelated with unobserved demand quality and with unobserved cost shocks. In addition, we employ two sets of characteristic-based instruments. First, following Berry et al. (1995), we compute the sum of characteristics of own-firm and rival-firm products within each market. Second, following Sudhir (2001), we

construct local instruments that restrict these sums to products within the same market segment, and separately within the same production region. Specifically, for each product and each characteristic, we compute the sum of that characteristic across other own-firm products in the same segment (and region) and the corresponding sum across rival-firm products in the same segment (and region), along with own-firm and rival-firm product counts within each group. The segment-level instruments provide additional variation from the competitive positioning of nearby products in characteristic space, while the region-level instruments capture competitive effects among products with similar cost structures.

In Figure 1, we display images of 25 automobiles present in our dataset. Note that, we converted color images of size 128×128 to grayscale for our study (sales are also not available separately by color). Moreover, our goal is to extract visual characteristics that are related to the shape of the automobile and not related to the color. For each image, we have its associated make, model, year, structured product characteristics and price.

Figure 1 Sample of Automobile Images

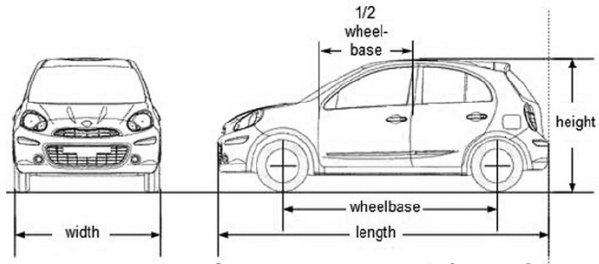


5.2. Discovered Visual Characteristics

We learn the visual characteristics of each make-model sold in the UK between 2008 and 2017 using disentanglement representation learning. We compare the unsupervised approach to learn visual characteristics with supervised approaches. In the supervised

approach, we train the learned visual characteristics to predict the supervisory signal associated with each make-model.[4] Figure 2 shows the dimensions of automobiles that serve as a subset of supervisory signals for disentanglement learning.

Figure 2 Dimensions of Automobiles



Note: This figure is sourced from **?**.

We follow the hyperparameter selection approach and the UDR metric described in the Methodology section (Section 4.1). From Table 3, we find that the visual characteristics learned from supervising on the combination of wheelbase, width, and height achieve the best disentanglement in terms of UDR.

We show the discovered visual characteristics in Figure 3 corresponding to the model with $\beta = 50$, $\delta = 10$ and the combination of 'Wheelbase, Width, and Height' as supervisory signals. Each row in the image corresponds to a visual characteristic. In each row, we change the value of one visual characteristic while fixing the value of all the other characteristics. Note that since we use a generative deep learning-based method, we can change the underlying learned visual characteristics and generate counterfactual images. The ability to generate counterfactual images allows us to interpret each visual characteristic as it isolates the effect of change in one visual characteristic while keeping the other characteristics fixed. We find five informative visual characteristics of an automobile's front

[4] We use the following supervisory signals whose summary statistics are provied in Appendix **??**:
1. 'Make' of the make-model
2. 'Country of Origin' of the make
3. 'Segment' of the make-model
4. 'Price' of the make-model
5. 'Length' of the make-model
6. 'Width' of the make-model
7. 'Height' of the make-model
8. 'Wheelbase' of the make-model
9. Combination of structured product characteristics of the make-model ('HP/Weight', 'MPG', 'Space')
10. Combination of structured product characteristics of the make-model ('Length', 'Width', 'Height')
11. Combination of structured product characteristics of the make-model ('Wheelbase', 'Width', 'Height')

Table 3 Comparison of Different Supervisory Approaches

Number of Signals	Supervisory Signals	β	δ	UDR
3	Wheelbase, Width, Height	50	10	0.739
3	HP/Weight, MPG, Space	50	30	0.710
1	Price	50	30	0.708
1	Weight	50	40	0.708
1	Wheelbase	50	30	0.690
1	Width	50	5	0.689
3	Length, Width, Height	50	40	0.678
1	Length	50	40	0.666
0	Unsupervised β-TCVAE	50	0	0.658
1	Height	30	20	0.378
1	Country of Origin	10	10	0.139
1	Segment	10	10	0.134
1	Unsupervised VAE	1	0	0.073
1	Unsupervised AE	0	0	0.074
1	Make	1	1	0.072

$\beta \in [1, 5, 10, 20, 30, 40, 50]$ and $\delta \in [0, 1, 5, 10, 20, 30, 40, 50]$.

view, while the rest were uninformative (i.e., changing the visual characteristic produces no change in the image). These informative characteristics are:

1. Body Shape: Automobiles scoring high on this characteristic have a narrower, more angular, and less rounded shape, resembling a sedan. Those scoring low have a wider, less angular, and more rounded shape, resembling a hatchback.
2. Color: Automobiles scoring low on this characteristic are lighter, while those scoring high are darker.
3. Grille Height: As the score of this visual characteristic increases, the grille becomes more prominent, larger, and more defined, with the top and bottom parts beginning to merge.
4. Boxiness: Automobiles scoring low on this characteristic have a high degree of boxiness, characterized by a taller, more upright, and narrower shape. Those scoring high have a lower degree of boxiness, with a lower, flatter, and wider appearance.
5. Grille Width: Automobiles scoring low on this characteristic have a narrower, less pronounced grille, while those scoring high have a wider, more prominent grille.

Relationship to Physical Dimensions To better understand our discovered visual dimensions, we examine their correlation with basic physical measurements such as wheelbase, weight, length, height, and width in Table 4. For instance, *Body Shape* (hatchback-like vs. sedan-like) has notable positive correlations with wheelbase (ρ=0.30), length (ρ=0.39), and weight (ρ=-0.33), consistent with hatchback-like profiles tending to have smaller wheelbase, lower length, and lighter. Likewise, its negative correlation with height-to-width

ratio (ρ=-0.42) implies that cars with lower *Body Shape* scores appear taller and narrower—characteristics commonly associated with hatchbacks or smaller crossover-style cars.

Boxiness is strongly negatively correlated with vehicle height (ρ=-0.59) and height-to-width ratio (ρ=-0.49), indicating that cars with higher degree of boxiness (lower score on the visual characteristic of boxiness) tend to look taller and more upright from the front. In contrast, vehicles with lower degree of boxiness appear flatter and sleeker. Notably, *Boxiness* does not align closely with length or wheelbase, suggesting it captures more of a cabin "uprightness" or silhouette shape rather than overall vehicle size.

Meanwhile, *Grille Height* exhibits near-zero correlations with physical measures, implying that it reflects primarily stylistic design choices (e.g., tall vs. short front grilles) rather than purely functional or size-related factors. Finally, *Grille Width* is weakly but consistently positively correlated with physical measurements like wheelbase (ρ=0.12), length (ρ=0.12), and width (ρ=0.15). While this might hint at a slight tendency for grille width to increase with vehicle size, these correlations are too weak to establish a meaningful relationship between grille width and vehicle dimensions.

Make the high correlations stand out. Make it like a heatmap.

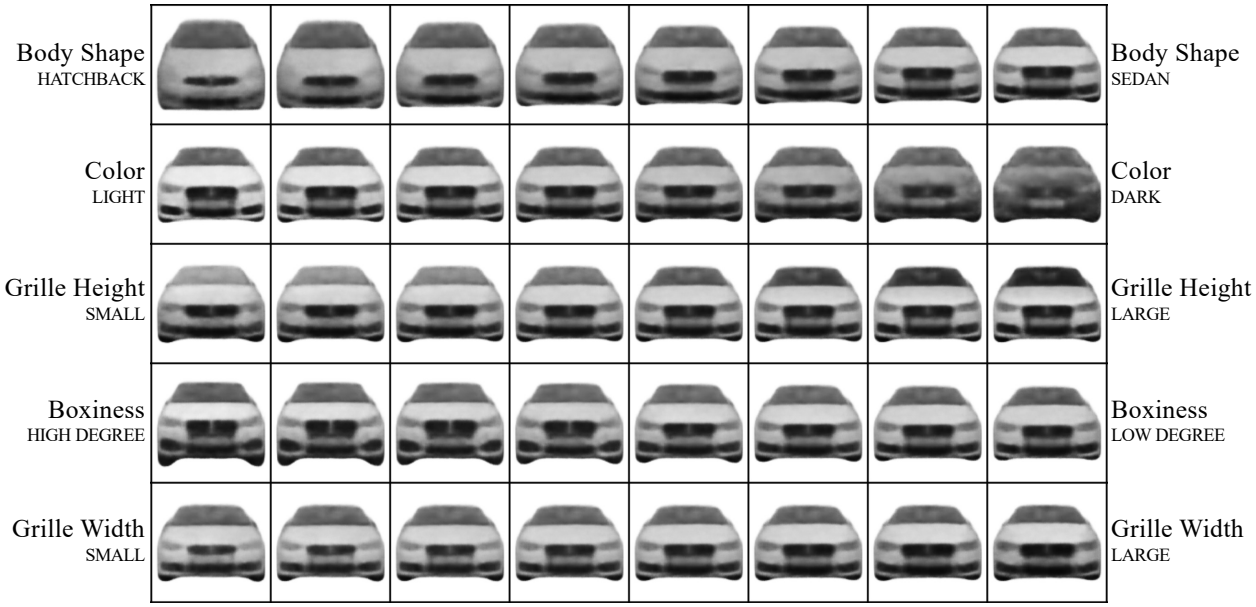
Table 4 Correlation Between Discovered Visual Characteristics and Vehicle Dimensions

	Wheelbase	Weight	Length	Height	Width	Height/Width Ratio
Body Shape	0.30	0.33	0.39	-0.28	0.25	-0.42
Boxiness	0.05	-0.07	0.14	-0.59	0.02	-0.49
Grille Height	0.04	0.02	0.05	-0.04	0.03	-0.05
Grille Width	0.12	0.08	0.12	0.03	0.15	-0.09

Interpretability To validate the interpretability and quantification of the discovered visual characteristics, we conducted surveys with human respondents. The results show that the majority of respondents agreed with each other on the interpretation and with the algorithm on the quantification of the visual characteristics. Detailed information about these surveys can be found in Appendix C.

Correlation between Functional & Form Characteristics We analyzed the correlations between structured and visual product characteristics and found that they are weakly correlated with each other, as shown in Table 5. The visual product characteristics are largely uncorrelated with each other, with the exception of a weak correlation between

Figure 3 **Discovered Visual Characteristics**



Left to Right: Vary one visual characteristic, keeping all others fixed

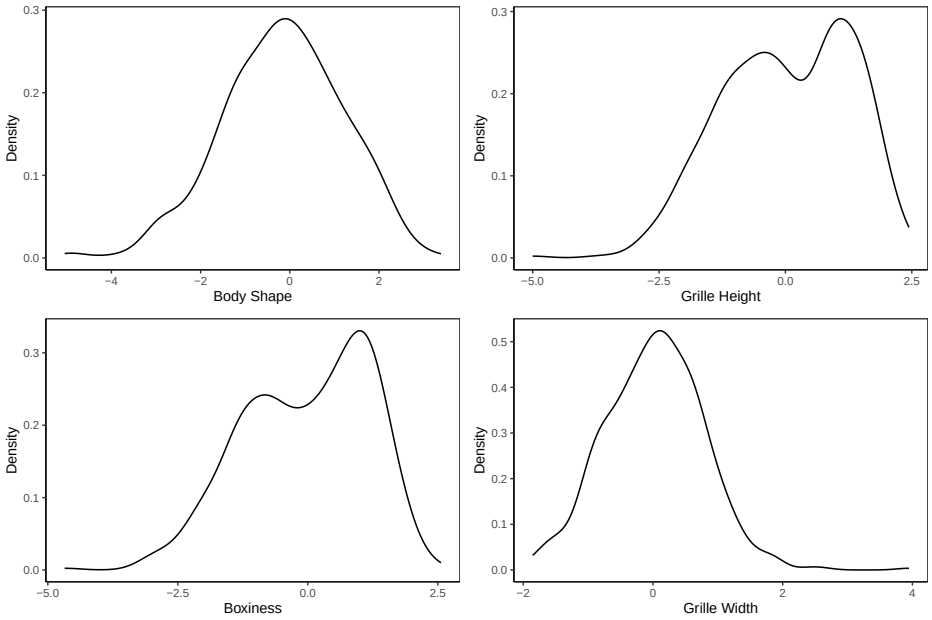
boxiness and body shape. In contrast, the structured product characteristics exhibit correlations with each other. Notably, the structured product characteristics and visual product characteristics are only weakly correlated. This suggests that visual characteristics provide additional information not captured by structured characteristics, highlighting the potential value of incorporating visual information in market structure analysis. This empirical pattern aligns with experimental findings from Kang et al. (2019), who show that consumers make explicit trade-offs between visual form and functional attributes—supporting the idea that visual design holds independent value in consumer decision-making.

One Table: Visual Characteristic, Form, Dimensions - in one table.. Replace 1.00 with -.

Table 5 **Correlation Matrix**

	Structured Characteristics				Visual Characteristics			
	Price	MPG	HP/Weight	Space	Boxiness	Body Shape	Grille Height	Grille Width
Price	1.00							
MPG	-0.60	1.00						
HP/Weight	0.74	-0.48	1.00					
Space	0.67	-0.47	0.36	1.00				
Boxiness	0.06	0.04	0.29	0.09	1.00			
Body Shape	0.50	-0.25	0.54	0.36	0.13	1.00		
Grille Height	0.11	0.03	0.12	0.05	0.04	-0.02	1.00	
Grille Width	0.07	-0.05	0.04	0.15	0.01	-0.12	-0.05	1.00

Figure 4 Discovered Visual Characteristics - Density Plot



5.3. Incorporating Visual Characteristics in Model of Market Equilibrium

We now map the general demand and supply framework of Section 4.2 to the specific variables in our data. Table 6 summarizes the specification. The structured product characteristics entering mean utility are horsepower-to-weight ratio (hpwt), miles per pound sterling (MP£), and vehicle space measured as length times width (space). The visual characteristics are the four disentangled dimensions extracted by our representation learning approach (Section 5.2): body shape, grille height, boxiness, and grille width.[5]

Table 6 Model Specification

Component	**Specification**
Demand: Mean Utility (δ_{jt})	
Structured characteristics ($\mathbf{x_{jt}}$)	hpwt, MP£, space
Visual characteristics ($\mathbf{v_{jt}}$)	body shape, grille height, boxiness, grille width
Demand: Heterogeneous Preferences (μ_{ijt})	
Price	$-\alpha p_{jt}/y_{it}$ (income demographic; $\sigma_{\text{price}} = 0$)
Random coefficients (σ)	hpwt, MP£, space, body shape, grille height, boxiness, grille width
Supply: Log Marginal Cost ($\ln c_{jt}$)	
Product characteristics	log(hpwt), log(space), log(CO_2)
Cost shifters	trend

[5] The visual characteristics also enter the BLP and local instrument construction.

On the demand side, all seven product characteristics—three structured and four visual—enter the mean utility δ_{jt} with linear coefficients $\overline{\beta_1}$ and $\overline{\beta_2}$ (Equation 15). We also allow for unobserved heterogeneity in preferences over all seven characteristics by estimating random coefficients σ_{β_1} and σ_{β_2}. As described in Section 4.2, we fix $\sigma_{\text{price}} = 0$ so that all heterogeneity in price sensitivity operates through the income demographic interaction $-\alpha p_{jt}/y_{it}$.

On the supply side, we parameterize log marginal costs as a function of log(hpwt), log(space), log(CO_2), and a linear time trend. Visual characteristics do not enter the supply side directly and are absorbed into the unobserved cost shock ω_{jt}.

We estimate the model jointly using both demand- and supply-side moment conditions as described in Section 4.2, with the two-step procedure using approximate optimal instruments following Conlon and Gortmaker (2020).

5.4. Estimation Results

Table 7 presents the parameter estimates. Panel A reports the demand-side parameters: the mean preference coefficients ($\overline{\beta}$), the standard deviations of the random coefficients (σ), and the demographic interaction (π) of price with the inverse of income. Panel B reports the supply-side parameters from the log marginal cost specification.

Demand estimates. The mean preference coefficients on all three structured characteristics carry the expected positive signs. Consumers value performance: the coefficient on horsepower-to-weight ratio is $\bar{\beta}_{\text{hpwt}} = 2.657$ ($t = 4.65$), and the coefficient on fuel efficiency is $\bar{\beta}_{\text{MP£}} = 0.249$ ($t = 3.95$). Vehicle space also enters positively at $\bar{\beta}_{\text{space}} = 0.926$ ($t = 2.19$). These estimates are consistent in sign and magnitude with prior work on automobile demand (Berry et al. 1995, Grieco et al. 2021).

Turning to the visual characteristics, three of the four enter mean utility with statistically significant coefficients. Body shape carries the largest positive effect ($\bar{\beta} = 0.329$, $t = 3.50$), indicating that consumers on average prefer a more sedan-like profile over a hatchback-like shape. Grille height is also positive and significant ($\bar{\beta} = 0.216$, $t = 3.13$): a more prominent, well-defined front grille is valued by the average consumer. Boxiness enters negatively ($\bar{\beta} = -0.143$, $t = 1.81$), suggesting that, all else equal, consumers prefer a lower, flatter, wider appearance (low degree of boxiness) over a taller, more upright shape. The coefficient on grille width is positive but not statistically distinguishable from zero ($\bar{\beta} = 0.099$, $t = 0.88$).

Table 7 Parameter Estimates

	$\bar{\beta}$ (Mean)	σ (Std. Dev.)	π $(1/y)$
Panel A: Demand Parameters			
Constant	-13.972	—	—
	(1.416)		
HP/Weight	2.657	0.000	0.000
	(0.572)		
Miles/£	0.249	0.009	0.000
	(0.063)	(0.425)	
Space	0.926	0.388	0.000
	(0.423)	(0.236)	
Body Shape	0.329	0.000	0.000
	(0.094)		
Grille Height	0.216	0.278	0.000
	(0.069)	(0.165)	
Boxiness	-0.143	0.472	0.000
	(0.079)	(0.200)	
Grille Width	0.099	0.308	0.000
	(0.112)	(0.976)	
Price/y	—	0.000	-179.667
			(46.897)
Panel B: Supply Parameters ($\ln c_{jt}$)			
	γ		
Constant	-3.296		
	(0.169)		
log(HP/Weight)	0.644		
	(0.045)		
log(Space)	1.774		
	(0.073)		
log(CO_2)	0.416		
	(0.074)		
Trend	0.013		
	(0.003)		

The price–income interaction coefficient is $\hat{\pi} = -179.7$ ($t = 3.83$), implying that higher-income consumers are less price-sensitive, consistent with the standard BLP formulation in which $\alpha p_{jt}/y_{it}$ generates demand patterns where wealthier households are more willing to pay for premium vehicles.

The random coefficient estimates reveal meaningful consumer heterogeneity in preferences over visual design. The estimated standard deviation for boxiness is $\hat{\sigma} = 0.472$ ($t = 2.36$), statistically significant at the 5% level. This implies that while the average consumer prefers a sleeker, less boxy appearance, a nontrivial fraction of consumers actually prefer boxier, more upright designs—consistent with the popularity of SUVs and crossovers alongside sedans and coupes. The standard deviation for grille height ($\hat{\sigma} = 0.278$, $t = 1.68$) indicates that consumers disagree about the appeal of prominent front grilles. For the

structured characteristics, the random coefficient on space is $\hat{\sigma} = 0.388$ $(t = 1.64)$, consistent with heterogeneous valuations of vehicle size. These random coefficients on visual characteristics are economically important: they generate richer substitution patterns in which consumers with similar visual tastes are more likely to substitute toward products with comparable design features, not just comparable performance attributes.

Supply estimates. The supply-side parameters in Panel B are precisely estimated and economically sensible. All cost shifters enter positively: $\hat{\gamma}_{\log(\text{hpwt})} = 0.644$ $(t = 14.3)$, $\hat{\gamma}_{\log(\text{space})} = 1.774$ $(t = 24.3)$, and $\hat{\gamma}_{\log(CO_2)} = 0.416$ $(t = 5.6)$. These imply that more powerful, larger, and higher-emission vehicles are more costly to produce. The elasticity of marginal cost with respect to space exceeds unity, reflecting the disproportionate cost of engineering larger vehicle platforms. The time trend is positive ($\hat{\gamma}_{\text{trend}} = 0.013$, $t = 4.3$), indicating that real production costs have risen by approximately 1.3% per year over the sample period. This is consistent with increasingly stringent EU emissions regulations and the adoption of advanced safety and infotainment technologies that have raised manufacturing costs in the UK automobile market.

Own-Price Elasticities. Table 8 reports sales-weighted mean own-price elasticities by year. Figure 6 displays the distribution of own-price elasticities across all product-years. The average own-price elasticity across our sample is -6.10, which is comparable to estimates in the automobile demand literature (Berry et al. 1995, Grieco et al. 2021).

Table 8 Own-Price Elasticities by Market Year

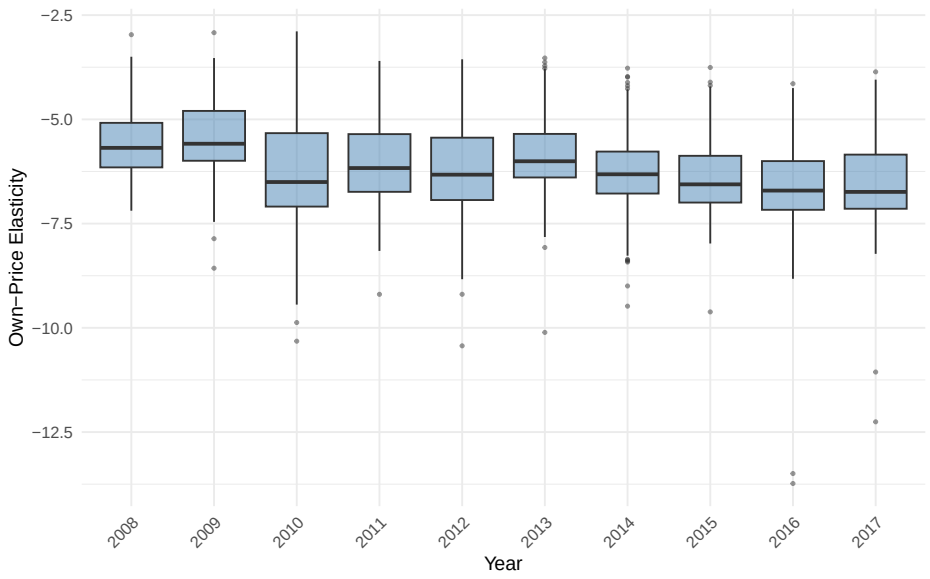
Market	Models	Mean	Median	SD
2008	189	-5.517	-5.683	0.835
2009	204	-5.411	-5.585	0.898
2010	209	-6.249	-6.503	1.204
2011	203	-6.032	-6.168	1.066
2012	217	-6.174	-6.326	1.131
2013	212	-5.838	-6.005	0.916
2014	221	-6.220	-6.317	0.973
2015	222	-6.324	-6.559	0.902
2016	231	-6.611	-6.708	1.134
2017	217	-6.470	-6.739	1.033

The elasticities display two important patterns. First, there is a modest increasing trend over time: mean elasticities range from -5.41 in 2009 to -6.61 in 2016 (Table 8), suggesting that the UK automobile market has become more competitive over the sample period, potentially driven by the entry of new models and brands. Second, there is substantial

Table 9 Own-Price Elasticities by Segment

Segment	Models	Mean	Median	SD
1	189	-4.295	-4.299	0.535
2	260	-4.996	-5.019	0.562
3	375	-6.117	-6.148	0.571
4	285	-6.779	-6.829	0.551
5	153	-7.226	-7.246	0.682
7	580	-6.521	-6.453	0.965
8	283	-6.149	-6.230	0.765

Figure 5 Own Price Elasticities by Year



variation across market segments (Table 9). Segment 1 (mini/subcompact) exhibits the least elastic demand (−4.30), reflecting stronger brand differentiation and loyal customer bases in the small car segment, while Segment 5 (executive/luxury) is the most elastic (−7.23), consistent with the wider array of close substitutes available in the premium segment.

Markups. We infer marginal costs for each product from the supply-side first-order conditions (Equation 17) evaluated at the estimated demand parameters. Table 10 reports markups by year, and Figure 7 presents their distribution.

The mean markup (price minus marginal cost as a fraction of price) is approximately 18% across the full sample, declining from 20% in 2008–2009 to 17% in 2016–2017, mirroring the trend toward more elastic demand. Across segments, markups are highest in Segment 1 (25%), where less elastic demand allows firms to charge larger margins, and lowest in

Figure 6 Own Price Elasticities Distribution

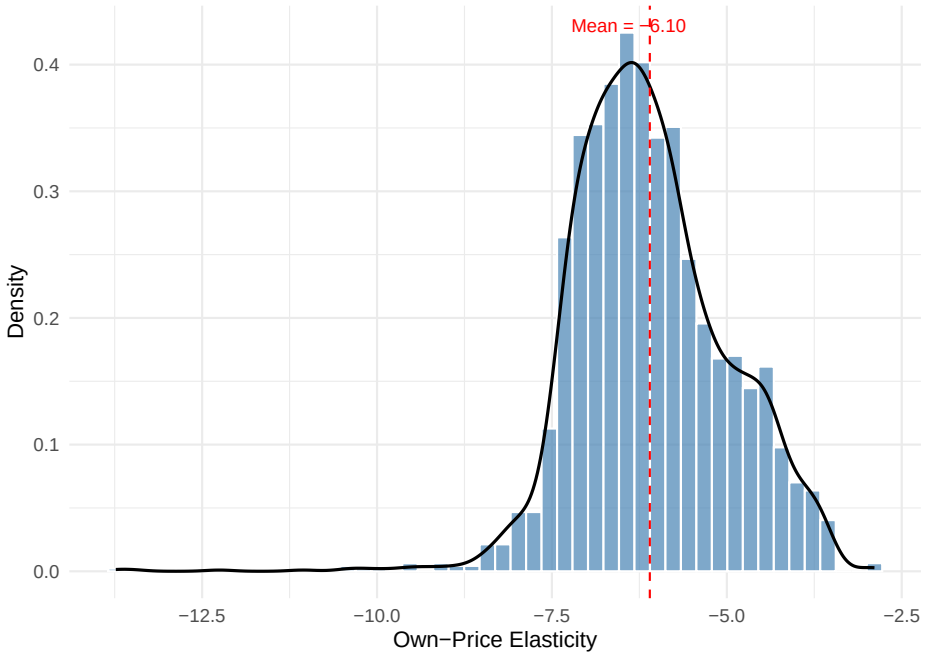


Table 10 Markups by Market

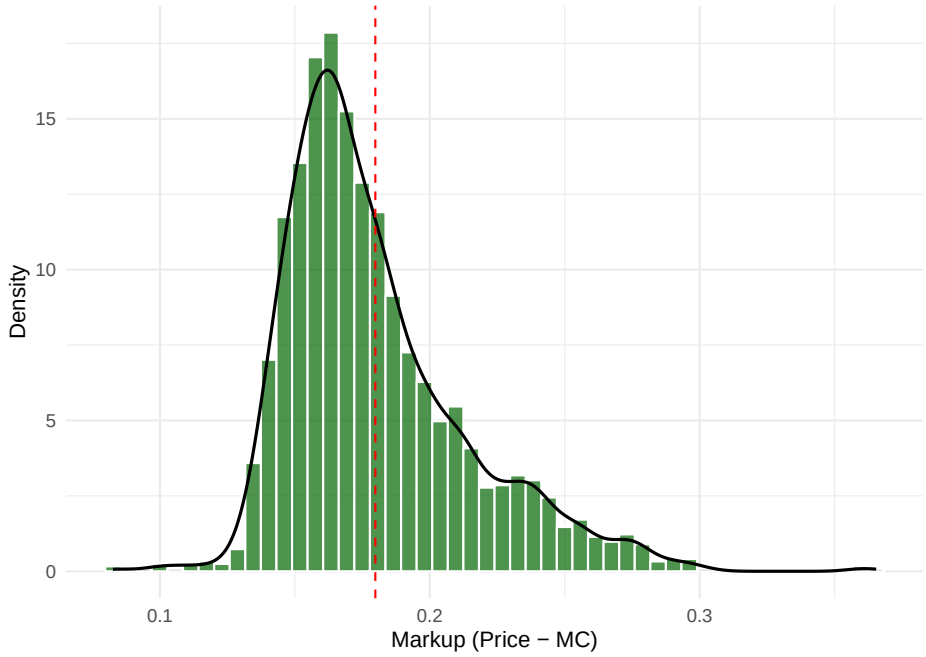
Market	Models	Mean	Median
2008	189	0.20	0.18
2009	204	0.20	0.19
2010	209	0.18	0.16
2011	203	0.18	0.17
2012	217	0.18	0.17
2013	212	0.19	0.18
2014	221	0.18	0.17
2015	222	0.17	0.17
2016	231	0.17	0.16
2017	217	0.17	0.16

Table 11 Markups by Segment

Segment	Models	Mean	Median
1	189	0.25	0.24
2	260	0.21	0.21
3	375	0.18	0.17
4	285	0.16	0.16
5	153	0.15	0.15
7	580	0.16	0.16
8	283	0.18	0.17

Segment 5 (15%), where intense competition among luxury brands compresses margins.

Figure 7 Distribution of Markups



The markup distribution in Figure 7 is right-skewed, with most products concentrated between 10% and 25%, and a thin right tail extending to approximately 35%.

5.5. Counterfactual Analysis: Functional vs. Visual Competition

To quantify the competitive role of visual design relative to functional characteristics, we conduct a neighbor-removal counterfactual exercise in the spirit of product removal analyses common in the merger simulation literature. For each product j in market $t = 2017$, we identify its nearest functional neighbor (based on Euclidean distance in the structured characteristic space of hpwt, MP£, and space) and its nearest visual neighbor (based on distance in the disentangled visual characteristic space).

We then compute three equilibria for each focal product: (i) the baseline equilibrium with all products present, (ii) a counterfactual equilibrium with the functional neighbor removed, and (iii) a counterfactual equilibrium with the visual neighbor removed. In each counterfactual, we remove the neighbor from the market and re-solve for Nash-Bertrand equilibrium prices holding marginal costs fixed, following the standard approach in Nevo (2001). We then record the change in the focal product's price, market share, and profit relative to the baseline.

If visual design is an important dimension of product differentiation and competition, removing a visually similar competitor should increase the focal product's market power—raising its price and profit—in a manner comparable to removing a functionally similar competitor.

Table 12 Counterfactual: Effect of Removing Functional vs. Visual Neighbors

Type	Models	Mean Price	SD Price	Mean Share	SD Share	Mean Profit	SD Profit
Functional	190	-0.0667	0.0284	0.003935	0.010521	-0.0002	0.0003
Visual	217	-0.0656	0.0292	0.003609	0.009742	-0.0002	0.0003

Table 13 Neighbor Removal Effects by Segment

Segment	N	Mean_dFunc_Price	Mean_dVis_Price	Mean_dFunc_Share	Mean_dVis_Share
1	16	-0.1215	-0.1189	0.028880	0.022857
2	25	-0.1001	-0.0996	0.011298	0.010579
3	40	-0.0688	-0.0674	0.002065	0.001990
4	29	-0.0448	-0.0457	0.000300	0.000257
5	15	-0.0408	-0.0369	0.000071	0.000057
7	69	-0.0567	-0.0557	0.000720	0.000744
8	23	-0.0648	-0.0624	0.000629	0.000593

Figure 8 Effect of Removing Nearest Functional vs. Visual Neighbor

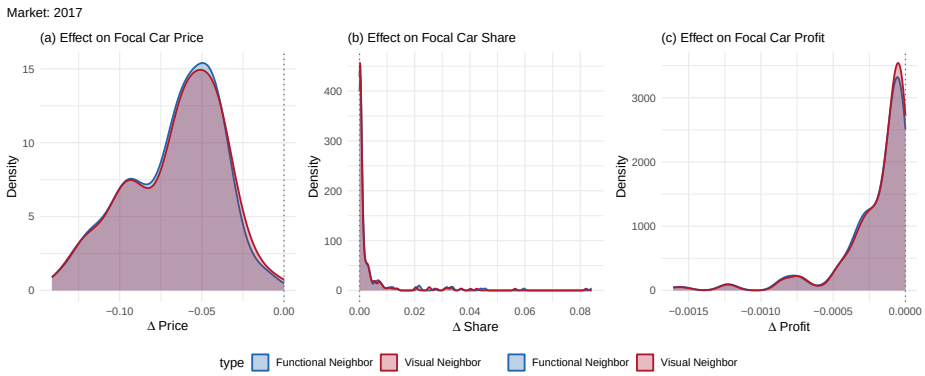


Table 12 summarizes the average effects, and Figure 8 displays the distribution of price, share, and profit changes under functional and visual neighbor removal. The central finding is that the competitive effects of removing a visual neighbor are remarkably similar in magnitude to those of removing a functional neighbor. The mean equilibrium price adjustment under functional neighbor removal is −6.67%, while visual neighbor removal

produces a mean adjustment of −6.56%—the visual effect is approximately 98% of the functional effect. Market share changes are also comparable: focal products gain an average of 0.39 percentage points of market share when a functional neighbor exits, compared with 0.36 percentage points under visual neighbor removal (a ratio of 92%). The profit effects are essentially identical across both scenarios. The distributions of these effects, shown in Figure 8, overlap almost entirely, confirming that the similarity holds not only on average but across the full distribution of products.

Table 13 disaggregates the counterfactual effects by market segment. The competitive effects of neighbor removal are largest in the lower segments. In Segment 1 (mini/subcompact), removing a functional neighbor produces a mean price adjustment of −12.2% and a share gain of 2.89 percentage points, while visual neighbor removal generates effects of −11.9% and 2.29 percentage points, respectively. These larger effects in the small car segment reflect the greater degree of product overlap: small cars tend to be more similar to one another in both functional and visual dimensions, so removing a close competitor has a proportionally larger impact. In contrast, the effects are more modest in Segments 4 and 5 (upper-medium and executive), where products are more differentiated and the loss of a single competitor has less influence on any focal product's competitive position.

Across all seven segments, the ratio of visual-to-functional competitive effects on price is consistently close to one, ranging from 0.90 in Segment 5 to 1.02 in Segment 4. This uniformity indicates that the competitive relevance of visual design is not confined to a single market niche; visual similarity constrains pricing power throughout the product spectrum.

These results carry two implications. First, they provide direct evidence that visual design is an economically meaningful dimension of product differentiation: the market power a firm derives from the absence of a visually similar competitor is on par with the market power derived from the absence of a functionally similar one. Second, they suggest that firms' design choices—the shape of the body, the prominence of the grille, the degree of boxiness—have strategic consequences comparable to decisions about horsepower, fuel efficiency, and vehicle size. From a competitive strategy perspective, a firm that differentiates its design language from rivals can soften price competition in much the same way as a firm that differentiates on engineering attributes.

6. Dummy

References

Bengio Y, Courville A, Vincent P (2013) Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence* 35(8):1798–1828.

Bergstra J, Bengio Y (2012) Random search for hyper-parameter optimization. *Journal of machine learning research* 13(2).

Berry S, Levinsohn J, Pakes A (1995) Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society* 841–890.

Berry S, Levinsohn J, Pakes A (1999) Voluntary export restraints on automobiles: Evaluating a trade policy. *American Economic Review* 89(3):400–430.

Berry S, Levinsohn J, Pakes A (2004) Differentiated products demand systems from a combination of micro and macro data: The new car market. *Journal of political Economy* 112(1):68–105.

Berry ST (1994) Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics* 242–262.

Bishop CM (2006) Pattern recognition and machine learning. *Springer google schola* 2:1122–1128.

Bloch PH (1995) Seeking the ideal form: Product design and consumer response. *Journal of marketing* 59(3):16–29.

Burgess C, Higgins I, Pal A, Matthey L, Watters N, Desjardins G, Lerchner A (2017) Understanding disentangling in β-vae. *Workshop on Learning Disentangled Representations at the 31st Conference on Neural Information Processing Systems*.

Chen RTQ, Li X, Grosse RB, Duvenaud DK (2018) Isolating sources of disentanglement in variational autoencoders. *Advances in Neural Information Processing Systems*, 2615–2625.

Compiani G, Morozov I, Seiler S (2025) Demand estimation with text and image data. *arXiv preprint arXiv:2503.20711* .

Conlon C, Gortmaker J (2020) Best practices for differentiated products demand estimation with pyblp. *The RAND Journal of Economics* 51(4):1108–1161.

Creusen ME, Schoormans JP (2005) The different roles of product appearance in consumer choice. *Journal of product innovation management* 22(1):63–81.

Dotson JP, Beltramo MA, Feit EM, Smith RC (2016) Modeling the effect of images on product choices .

Duan S, Matthey L, Saraiva A, Watters N, Burgess C, Lerchner A, Higgins I (2020) Unsupervised model selection for variational disentangled representation learning. *International Conference on Learning Representations*.

Dumoulin V, Visin F (2016) A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285* .

Feenstra RC, Inklaar R, Timmer MP (2015) The next generation of the penn world table. *American economic review* 105(10):3150–3182.

Gandhi A, Houde JF (2019) Measuring substitution patterns in differentiated-products industries. Technical report, National Bureau of Economic Research.

Goodfellow I, Bengio Y, Courville A (2016) *Deep learning* (MIT press).

Grieco PL, Murry C, Yurukoglu A (2021) The evolution of market power in the us auto industry. Technical report, National Bureau of Economic Research.

Higgins I, Matthey L, Pal A, Burgess C, Glorot X, Botvinick M, Mohamed S, Lerchner A (2017) beta-vae: Learning basic visual concepts with a constrained variational framework. *International Conference on Learning Representations*.

Huang J, Chen B, Luo L, Yue S, Ounis I (2021) Dvm-car: A large-scale automotive dataset for visual marketing research and applications. *arXiv preprint arXiv:2109.00881* .

Kang N, Ren Y, Feinberg F, Papalambros P (2019) Form + function: Optimizing aesthetic product design via adaptive, geometrized preference elicitation. *arXiv preprint arXiv:1912.05047* .

Kim H, Mnih A (2018) Disentangling by factorising. *ICML*, 2649–2658.

Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Kingma DP, Welling M (2014) Auto-encoding variational bayes. *International Conference on Learning Representations*.

Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25.

LeCun Y, Bengio Y, et al. (1995) Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks* 3361(10):1995.

Locatello F, Bauer S, Lučić M, Rätsch G, Gelly S, Schölkopf B, Bachem OF (2019) Challenging common assumptions in the unsupervised learning of disentangled representations. *International Conference on Machine Learning*, 4114–4124.

Locatello F, Tschannen M, Bauer S, Rätsch G, Schölkopf B, Bachem O (2020) Disentangling factors of variations using few labels. *International Conference on Learning Representations*.

Maas AL, Hannun AY, Ng AY, et al. (2013) Rectifier nonlinearities improve neural network acoustic models. *Proc. icml*, volume 30, 3 (Atlanta, GA).

Magnolfi L, McClure J, Sorensen A (2025) Triplet embeddings for demand estimation. *American Economic Journal: Microeconomics* 17(1):282–307.

Murphy KP (2012) *Machine learning: a probabilistic perspective* (MIT press).

Nevo A (2000) Mergers with differentiated products: The case of the ready-to-eat cereal industry. *The RAND Journal of Economics* 395–421.

Nevo A (2001) Measuring market power in the ready-to-eat cereal industry. *Econometrica* 69(2):307–342.

Petrin A (2002) Quantifying the benefits of new products: The case of the minivan. *Journal of political Economy* 110(4):705–729.

Quan TW, Williams KR (2019) Extracting characteristics from product images and its application to demand estimation. *University of Georgia, Department of Economics* .

Sisodia A, Burnap A, Kumar V (2025) Generative interpretable visual design: Using disentanglement for visual conjoint analysis. *Journal of Marketing Research* 62(3):405–428.

Snoek J, Larochelle H, Adams RP (2012) Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems* 25.

Sudhir K (2001) Competitive pricing behavior in the auto market: A structural analysis. *Marketing Science* 20(1):42–60.

Talke K, Salomo S, Wieringa JE, Lutz A (2009) What about design newness? investigating the relevance of a neglected dimension of product innovativeness. *Journal of product innovation management* 26(6):601–615.

Electronic Companion Supplement

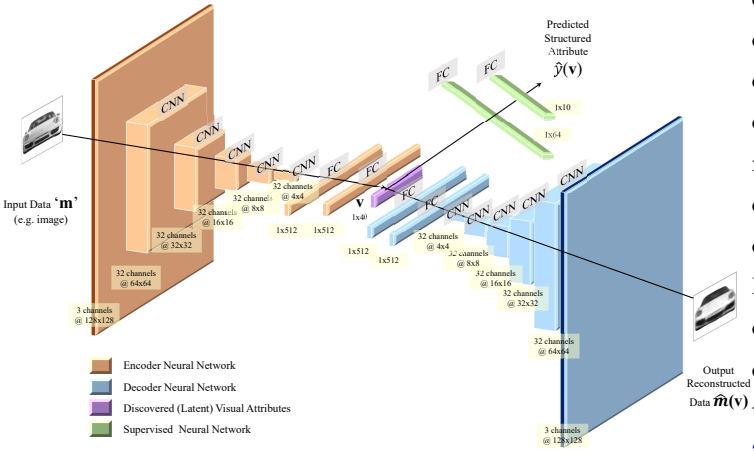
Appendix A: Neural Net Architecture

Figure EC.1 shows the detailed neural net architecture of our model. We modify the architecture proposed by Burgess et al. (2017) to accommodate 128×128 pixel images and incorporate the supervisory signals from our model of market equilibrium.

The encoder neural network uses a sequence of convolutional neural network (CNN) layers to learn high-level representations of the input images. CNNs are well-suited for working with image data, as they can effectively capture spatial hierarchies and learn translation-invariant features (LeCun et al. 1995, Krizhevsky et al. 2012). We stack multiple CNN layers in the encoder to progressively learn more complex and abstract visual concepts. The output of the final CNN layer is then flattened and passed through two fully-connected (FC) layers. The first FC layer reduces the dimensionality of the flattened representation, while the second FC layer further compresses the information into a compact set of latent visual characteristics, with a maximum of J dimensions.

The decoder neural network is designed to reconstruct the original image from the latent visual characteristics. Its architecture is essentially the transpose of the encoder network, consisting of FC layers followed by a sequence of transposed convolutional layers (Dumoulin and Visin 2016). The decoder takes the J-dimensional latent visual characteristics as input and gradually upsamples and expands the representation until it reaches the original image size of 128×128 pixels. Finally, the supervised neural network takes the discovered visual characteristics as input and predicts the vector of supervisory signals, which serve as labels for training the model. The supervised network allows the model to learn visual characteristics that are predictive of the supervisory signals, guiding the disentanglement process.

Figure EC.1 Model Architecture



Notes: The encoder neural net for the VAEs consisted of 5 convolutional layers, each with 32 channels, 4×4 kernels, and a stride of 2. This was followed by 2 fully connected layers, each of 512 units. The latent distribution consisted of one fully connected layer of 40 units parameterizing the mean and log standard deviation of 20 Gaussian random variables. The decoder neural net architecture was the transpose of the encoder neural net but with the output parameterizing Bernoulli distributions over the pixels. Leaky ReLU activations were used throughout, which help alleviate the vanishing gradient problem and improve the model's ability to learn complex representations (Maas et al. 2013). We used the Adam optimizer (Kingma and Ba 2014) with the learning rate 5e-4 and parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We set batch size equal to 64. We train for 200 epochs to ensure convergence.

Appendix B: UDR Algorithm

The Unsupervised Disentanglement Ranking (UDR) metric, proposed by Duan et al. (2020), assesses the similarity of disentangled representations learned by different models. The key idea behind UDR is that if two models have learned similar disentangled representations, their informative latent dimensions should exhibit strong correlations.

To compute UDR for a pair of models i and j, we first calculate the correlation matrix R, where each entry $R(a,b)$ represents the correlation between latent dimensions a and b from models i and j, respectively (Equation EC.1). We then identify the most correlated latent dimension in model j for each dimension a in model i, denoted as r_a (Equation EC.2).

$$R(a,b) = cor(v_i(a), v_j(b)) \tag{EC.1}$$

$$r_a = \max_{b \in V(j)} corV(a,b) \tag{EC.2}$$

The UDR score for the pair of models, UDR_{ij}, is computed using Equation EC.3. This equation consists of two symmetric terms, each focusing on one model. The first term considers each informative latent dimension b in model j and calculates the ratio of the squared correlation of its most similar dimension in model i (r_b^2) to the sum of correlations between b and all dimensions in model i. This ratio is then multiplied by

an indicator function $I_{KL}(b)$, which equals 1 if dimension b is informative (determined by its KL divergence from the prior distribution) and 0 otherwise. The second term follows the same process, but with the roles of models i and j reversed.

$$UDR_{ij} = \frac{1}{d_i + d_j}\left[\sum_{b\in Z(j)} \frac{r_b^2}{\sum_{a\in Z(i)} R(a,b)} I_{KL}(b) + \sum_{a\in Z(i)} \frac{r_a^2}{\sum_{b\in Z(j)} R(a,b)} I_{KL}(a)\right] \tag{EC.3}$$

The final UDR score is normalized by the total number of informative dimensions in both models $(d_i + d_j)$ to ensure that having more informative dimensions does not automatically lead to a higher score. A perfect one-to-one correspondence between the informative dimensions of the two models would result in a UDR score of 1.

Appendix C: Validation Surveys for Interpretability and Quantification

To validate the interpretability of the discovered visual characteristics, we conducted a survey with 93 respondents after removing those who failed attention checks. Respondents were shown a sequence of five images for each visual characteristic, where the characteristic varied from left to right while keeping all other characteristics fixed. They were asked to describe how the car changed the most across the sequence of images. Figure EC.2 presents an example of the open-ended question posed to the respondents.

We then used language models (ChatGPT4 and Claude 3 Opus) to summarize the main themes that respondents agreed upon for each visual characteristic.[6]

Both language models agreed on the summaries, leading us to label the visual characteristics as follows:

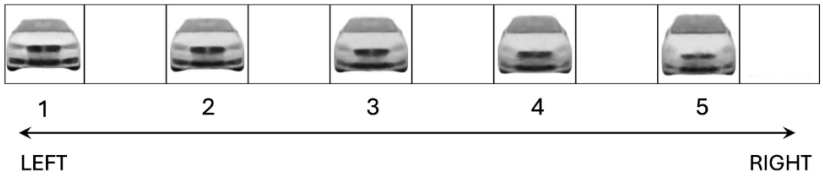
1. Body Shape: The LLM summary of survey respondents states that the "car appears to change in shape, particularly becoming narrower, less angular, and more rounded with each successive image."
2. Color: Automobiles scoring low on this characteristic are darker and vice-versa.
3. Grille Height: The LLM summary of survey respondents indicates that "grilles are become larger, darker, and more defined." Although this summary also mentions the windscreen becoming darker, which is entangled with the grille height, it captures the essence of the characteristic. We interpret this as automobiles scoring low on this characteristic have less prominent grilles, while those scoring high have more prominent, and larger grilles.
4. Boxiness: The LLM summary of survey respondents describes that the "car becomes lower, flatter, and wider as the sequence progresses." We interpret this as automobiles scoring low on this characteristic have a high degree of boxiness, characterized by a taller, more upright, and narrower shape. In contrast, those scoring high on this characteristic have a lower degree of boxiness, with a lower, flatter, and wider appearance.
5. Grille Width: The LLM summary of survey respondents states that "grille width become smaller, narrower, and less pronounced as the sequence progresses." Based on this, we interpret that automobiles

[6] LLM Prompt: Summarize the below responses and share the biggest theme that most respondents agree upon? https://chat.openai.com/c/95ee3363-61a5-4604-8e15-4cd68ac1bae3 or https://chatgpt.com/c/95ee3363-61a5-4604-8e15-4cd68ac1bae3 or https://yalesurvey.ca1.qualtrics.com/jfe/form/SV_ai6c1wUD39kZP8O

Figure EC.2 Interpretability Validation Survey Question

Q1/4: Look at the below image to see the various parts of a car.

Now, carefully examine each car image below from 1 to 5, going from left to right.
Note: Images are low-quality on purpose. Be sure to see all the images 1 to 5.



How does the car change the most as you go from image 1 to 5? Go through each part of the car one by one before deciding your response. Write it in a few words.

scoring low on this characteristic have a wider, more prominent grille, while those scoring high have a narrower, less pronounced grille.

To further validate the quantification of the visual characteristics determined by our method, we conducted a second survey (Figure EC.3). In this survey, we presented respondents with several pairs of automobile images that differed only along one visual characteristic. Respondents were asked to select the pair of automobiles that they perceived as more similar. We then compared the responses to our algorithm's quantification to assess consistency with human interpretation. For the characteristic we labeled as "Body Shape," 97% of the 104 respondents agreed with the algorithm's quantification scale. Similarly, for "Grille Height," 98% of the 107 respondents were in agreement. The characteristic we termed "Boxiness" had a 95% agreement rate among 103 respondents, while "Grille Width" saw 93% of the 104 respondents concurring with the algorithm's quantification. Overall, a strong majority of respondents (averaging 96% across four visual characteristics) agreed with the algorithm's quantification scale for the visual characteristics, demonstrating that our method's quantification aligns well with human perception.

Figure EC.3 Quantification Validation Survey Question

Which pair of cars in your judgment are visually more similar? Carefully check both large and small visual aspects. Do not consider any non-visual features like brand or price.

Left Pair ○ Right Pair ○

Appendix D: Dummy