Partisan Portraits: Measuring Visual Polarization in News Media*

Mohammad Mosaffa
Cornell University

Omid Rafieian
Cornell University

Hema Yoganarasimhan
University of Washington

Abstract

News outlets choose not only what readers read but also what they see. Studies of polarization in news media have focused almost entirely on verbal content, in part because images are high-dimensional and the visual attributes an editor selects are chosen jointly, which makes it hard to attribute differences across outlets to any one of them. We develop a measure of polarization in visual content that combines economic theory with generative models and multi-modal deep learning and uses the full richness of image data rather than a handful of extracted features. Applying this framework to a decade-long dataset featuring 30 prominent politicians across 20 major news outlets, we identify significant polarization in visual content, with notable variation across outlets, across politicians, and over time. We document which outlets exhibit higher conservative visual slant, which politicians receive the most polarized visual portrayal, and how visual polarization has evolved over the past decade. Finally, we conduct a series of validation tests demonstrating the consistency of our proposed measures with external measures of media slant that rely on non-image-based sources, and the shortcomings of reduced-form approaches that rely on feature extraction.

JEL Codes: L82, D72, C55, C81
Keywords: Media slant, political polarization, news media, generative models, counterfactual image generation, measurement

*We would like to thank the participants of the 2024 PhD student workshop at Cornell University, 2024 MarkTech Conference, 2025 AI and Managment (AIM) conference at USC, 2025 Workshop on Platform Analytics (WoPA), Boston University, 2025 ACM Conference on Economics and Computation, 2025 Northeastern Marketing Consortium (NEMC), 2025 AI & Marketing: New Methods and New Risks Conference at Stanford GSB, and Carnegie Melon University for their feedback. We also thank Sachin Gupta, David Holtz, Jura Liaukonyte, Simha Mummalaneni, Mor Naaman, Hoori Rafieian, Navdeep Sahni, Jesse Shapiro, Zikun Ye, and Dennis Zhang for their detailed comments, which have significantly improved the paper. Please address all correspondence to: mm3322@cornell.edu, or83@cornell.edu, and hemay@uw.edu.

1 Introduction

News outlets choose not only what readers read but also what they see. Economists have studied the first choice at length. Theory explains slant in news content as a response to readers' priors and to reputational incentives (Mullainathan and Shleifer, 2005; Gentzkow and Shapiro, 2006), and a large empirical literature measures it and traces its consequences for political behavior (Groseclose and Milyo, 2005; Gentzkow and Shapiro, 2010; DellaVigna and Kaplan, 2007; Martin and Yurukoglu, 2017). Most empirical measures in this literature are constructed from textual data, such as the think tanks an outlet cites, the phrases it shares with one party's legislators, the topics it covers and omits.

The choice of images has received considerably less attention, although little about it suggests it is less consequential. Photographs occupy the most prominent position on a front page or a homepage. They are processed faster and remembered longer than text (Sullivan and Masters, 1988; Townsend and Kahn, 2014), and facial appearance alone predicts electoral outcomes (Todorov et al., 2005; Hoegg and Lewis, 2011). Political actors themselves also appear to treat visual portrayal as carrying weight beyond the text. Political campaigns maintain libraries of approved photographs, press offices negotiate camera positions, and the occasional public dispute reveals how the tradeoff is understood by its subjects. When *Time* put Donald Trump on its cover in October 2025 above a story praising his role in brokering a Gaza ceasefire, his response was not to the article:

> "Time Magazine wrote a relatively good story about me, but the picture may be the Worst of All Time. They 'disappeared' my hair, and then had something floating on top of my head that looked like a floating crown, but an extremely small one. Really weird!"[1]

The reaction is informative because it isolates the visual dimension of the coverage: the accompanying text was favorable, while the objection concerned the image. It illustrates that visual portrayal can convey information or impressions that are distinct from those conveyed by words. If images are both salient to audiences and consequential enough for the subjects of coverage to contest publicly, then understanding whether news outlets systematically differ in their visual portrayal of political figures is an important part of measuring media slant and polarization.

Figure 1 illustrates the high-level concept behind visual polarization. It shows coverage of Donald Trump and Kamala Harris–the two major-party presidential candidates in 2024 election–by *CNN* and *Fox News*. In the examples shown, Trump appears mid-shout at a rally in one image and composed in a suit in the other; Harris appears smiling on a convention stage in one and mid-gesture in the other. Such differences in visual portrayal are immediately salient, but measuring them systematically is difficult. Images are high-dimensional objects, and the attributes that shape a viewer's impression—facial expression, pose, camera angle, lighting, background, cropping, and setting—are chosen jointly.

To see the measurement challenge, consider the natural reduced-form approach: running photographs through an expression classifier and comparing the resulting scores for a positive expression like smile across outlets. Suppose, for example, that we compare images of Donald Trump published by *CNN* and

[1] Posted on Truth Social at 1:30 a.m. ET on October 14, 2025. The cover photograph was taken from a low angle against backlighting.

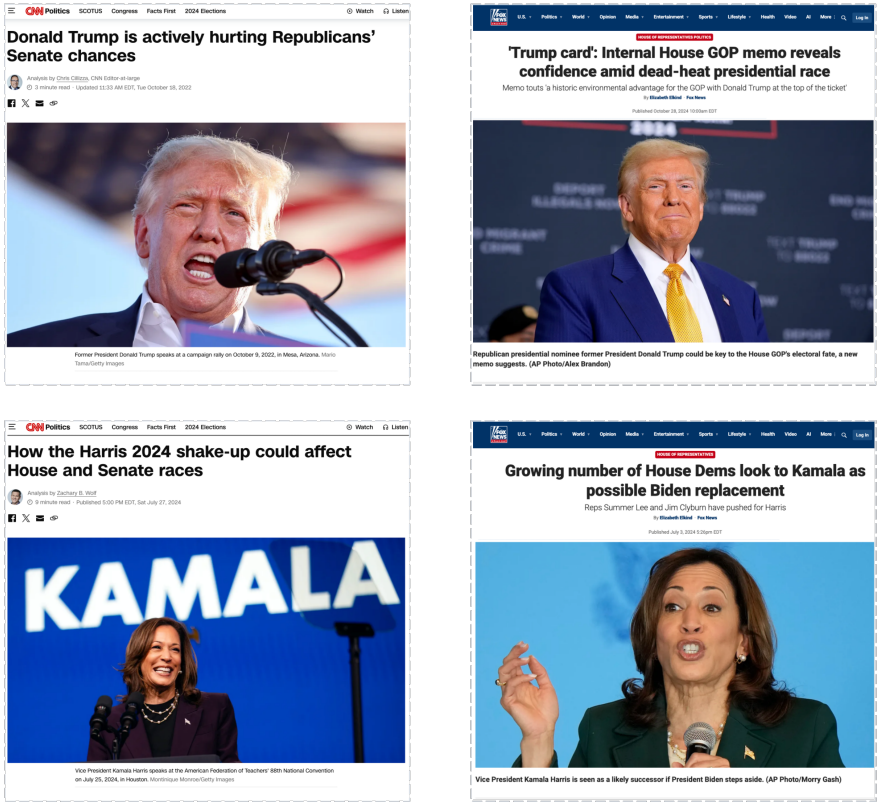


Figure 1: Examples of 2024 presidential candidates' portrayal by *CNN* and *Fox News*.

Fox News, and that the smile labels are measured perfectly. Now consider another feature of the image: whether an American flag appears in the background. Suppose flags appear more often in images selected by *Fox News*, consistent with its audience having stronger nationalist preferences. At the same time, it is easy to conceive that politicians are also more likely to smile in the kinds of formal or campaign settings in which flags are present. We could then observe more smiling images of Trump on *Fox News* even if *Fox News* had no greater propensity to select a smiling photograph once the rest of the image were held fixed. The same problem arises with rallies versus hearings, cheering crowds, lighting, camera angle, and other aspects of the photograph. In naturally occurring images, these attributes move together. A comparison across realized images therefore compares two *bundles* of visual attributes selected jointly by the outlets, and therefore does not isolate whether one outlet is more likely than another to select a smiling portrayal, all else equal. Enlarging the set of extracted features captures more of this bundle but does not eliminate the problem, because any finite representation specifies ex ante which dimensions of the image are represented and leaves the remainder outside the comparison.

In this paper, we instead define visual polarization through a counterfactual comparison rather than a comparison of published images. Consider a politician, an article, and two photographs constructed to differ in exactly one visual attribute, which in our application is whether the politician is smiling. We ask how much more inclined each outlet would be to run the article with the favorable rather than the unfavorable version, and we define *visual polarization* between two outlets as the difference between their inclinations.

Visual slant is the same quantity measured relative to a reference outlet. Because the two photographs are constructed to differ only in the focal attribute, this comparison isolates that attribute from the many attributes that accompany it in naturally occurring images.

Our measure is an object in a model of editorial image choice. We model an editor at each outlet as selecting among the photographs available to her so as to maximize a payoff that depends on the image, the article, the politician, and the outlet's editorial position. Under a logit structure, the log odds with which an outlet would choose the favorable photograph over the unfavorable one equals the difference in the utility it derives from the two. Visual polarization is therefore a difference in valuations recovered from editorial choices, giving the measure an interpretation in terms of preferences rather than merely observed differences in published content.

The empirical challenge is that we do not observe these choices. Our data contain the photograph each outlet published, but not the photographs it declined, and not the collection of images from which it chose. What the data do identify is the reverse conditional probability: how likely it is that an article with a given photograph and context came from a given outlet. Inverting the model of editorial choice shows that the relative log odds of two outlets at a fixed photograph equals the difference in their valuations of that photograph plus three nuisance terms. These are the relative availability of the photograph to the two outlets, the relative inclusive value of the collections from which they choose, and their relative propensity to cover the story at all.

Two of these three terms cancel when we difference across the two photographs. The inclusive value is a property of an outlet's entire collection and the coverage propensity is a property of its news agenda, so neither depends on which photograph the outlet published. Relative availability does not cancel, because it asks whether a particular frame reached a particular newsroom, and it is therefore where our identifying assumption is required. However, given our specific counterfactual choice design, the assumption that both images have the same relative availability at each outlet is plausible, because the two counterfactual versions are designed to represent alternative frames of the same photographic moment, naturally being seconds apart in a realistic setting. Under this restriction and standard overlap conditions, the third term also cancels, identifying the structural parameter.

We then characterize the behavior of the estimator when the publication probability is replaced by an estimate. Because the article-level measure is a linear contrast in four estimated log probabilities, first-stage error passes through it without being amplified. Whether this is sufficient for an aggregate measure depends on how the aggregate is constructed. Our outlet-level index is additive in the article-level parameter, so consistency transfers directly to it. Our politician-level index is not additive, because it is built from a dispersion across outlets, and estimation error inflates a dispersion rather than averaging out of it. We show that the plug-in estimator of the variance underlying this index is biased upward by an amount that does not shrink as articles accumulate, and we adopt a cross-fitted estimator that removes this bias from the variance.

Implementing the measure requires two inputs. We use a conditional generative model to add a smile to neutral photographs of each politician while leaving the rest of the frame intact, and we find no systematic changes in the other image characteristics we measure. We then estimate a multi-modal deep learning model that predicts the outlet from the photograph together with the article title, the politician's identity and party

affiliation, and the publication date. The role of prediction in our framework is instrumental. We use it to hold the remainder of the image and the article fixed when we make the counterfactual comparison, rather than to assume that content away.

We apply the measure to 63,188 photographs of 30 prominent Democratic and Republican politicians published by 20 major news outlets between 2011 and 2021, each verified to contain a single face belonging to the politician named in the article. The outlet prediction model classifies correctly in 44.5 percent of held-out cases in a twenty-way problem, against 5 percent under uniform guessing, while a specification that uses the politician, party, date, and article text but no image reaches 16.7 percent. This gap is informative because it indicates that photographs carry substantial information about which outlet published an article beyond who is covered and how the story is written.

Our central result is that the response to a smile has a systematic partisan structure. Measured relative to *Reuters* as our reference outlet, a smile raises an outlet's relative likelihood of running an article when the outlet and the politician are politically aligned, and lowers it when they are opposed. For Republican politicians, average visual slant is 0.45 in the three most Republican-leaning outlets and -0.69 in the three most Democratic-leaning ones. For Democratic politicians, the corresponding figures are -0.57 and 0.82. The sign of the response therefore reverses with outlet ideology and reverses again with the politician's party.

Aggregating to the outlet level measures *Conservative Visual Slant* (CVS), the difference between an outlet's smile response for Republican and for Democratic politicians. *Fox News*, *Daily Mail*, and *Newsmax* rank highest on this measure, while *The Washington Post*, *The New York Times*, and *CNN* rank lowest. Three exercises support interpreting CVS as the partisan component of visual portrayal. First, it correlates 0.72, 0.54, and 0.77 with slant measures constructed from Twitter sharing, browsing panels, and editorial ratings (Faris et al., 2017; Flaxman et al., 2016; AllSides, 2024), none of which uses image data. Second, these correlations collapse toward zero when we randomly permute politicians' party labels, providing a placebo test of the partisan structure captured by the measure. Third, a feature-extraction measure in the style of Boxell (2021), constructed on our own photographs, correlates only weakly and insignificantly with the same benchmarks, consistent with our argument that reducing images to extracted features can discard information relevant to visual slant.

Because we estimate slant article by article, we can also ask which politicians outlets disagree about most. *Overall Visual Polarization* (OVP) measures the dispersion across outlets in their preference for a favorable portrayal of a given politician. Donald Trump, Barack Obama, Bernie Sanders, and Joe Biden rank among the most polarized politicians in our sample, and Joe Manchin, Susan Collins, and Liz Cheney among the least. OVP correlates with an independent measure of how politicians' campaign-finance-based ideology scores established in (Bonica, 2014). Because this measure is constructed independently of our image data, the correlation provides an external benchmark for the variation captured by OVP.

Finally, the ten-year span of our data allows us to document when visual divergence emerged. The average slant across all outlets for both parties is close to zero through 2014 and then separates sharply from 2015 onward, reaching roughly 0.18 to 0.20 for Democratic politicians and -0.17 for Republican politicians in the period between 2018 and 2020. This timing coincides with Donald Trump's entry into presidential

politics, and his coverage accounts for much of the decline on the Republican side. Some of the partisan gap nevertheless persists when we exclude his images entirely, which suggests a broader shift in partisan visual framing, indicating that the divergence is not driven solely by Trump's entry into national politics.

The paper makes three contributions. First, it introduces a new economic object for the study of visual media slant. Existing work measures differences in the visual content outlets choose; we measure how an isolated visual manipulation changes outlets' relative revealed affinity for an image. This distinction separates a focal attribute from the many correlated characteristics of naturally occurring images and yields an estimand with a direct editorial-choice interpretation. Second, the paper develops an empirical strategy for recovering this object from high-dimensional visual data. Counterfactual image generation creates controlled within-image variation, while the multimodal prediction model conditions on a rich learned representation of the remaining image and article context rather than reducing the image ex ante to a small set of researcher-selected variables. Third, the method produces new longitudinal and granular evidence on media slant: it measures outlet-specific direction, politician-specific polarization, and changes over time, and it shows that the choice of measurement approach has materials effects on the resulting outlet rankings and their relationship to established measures of ideology.

1.1 Related Literature

Our starting point is the literature that measures ideological slant in news content. Groseclose and Milyo (2005) map outlets' citations of think tanks onto the ideological positions of legislators. Gentzkow and Shapiro (2006) show how slant toward consumers' priors can arise from reputational incentives even when consumers value accuracy, and Mullainathan and Shleifer (2005) derive slant from readers' taste for confirmatory news. Gentzkow and Shapiro (2010) construct a text-based measure of newspaper slant from partisan differences in congressional language and attribute a substantial share of observed slant to consumer demand. Gentzkow et al. (2019) generalize the underlying logic to high-dimensional choice data, defining partisanship by how readily party can be inferred from a speaker's language and correcting the finite-sample bias of plug-in estimators with a leave-out procedure. We borrow their bias correction and share their premise that predictability reveals systematic group differences. Our target, however, is not the level of predictability in observed content but the change in relative predictability induced by a controlled intervention on one component of it. Prat and Strömberg (2013) survey the broader political economy of mass media, and DellaVigna and Kaplan (2007) and Martin and Yurukoglu (2017) establish that exposure to slanted news shifts voting behavior, which is what makes measurement on the supply side consequential.

The closest work to ours studies visual rather than verbal slant. Peng (2018) uses computer-vision measures of facial expression and related attributes to document differential portrayals of the 2016 presidential candidates across fifteen news websites. Boxell (2021) analyzes a far larger set of front-page images, extracts facial emotions, and shows that outlets with more Republican audiences portray Republican politicians more favorably in relative terms. These papers establish that partisan visual portrayal is present in published images, and their measures are functions of machine-extracted attributes of those images. Caprini (2023) broadens the object from facial favorability to visual language, constructing an interpretable vocabulary of subjects, facial and positional characteristics, and contextual elements, and adapting the estimator of Gentzkow et al. (2019) to quantify how well images discriminate between outlets of different ideologies;

she also shows experimentally that partisan lead images shift readers' opinions with text held fixed. Ash et al. (2021) study visual content in televised news along related lines.

Our question is complementary to this work. A dictionary-based measure characterizes partisanship in the dimensions that its vocabulary represents, and as Caprini (2023) notes, the possible tokenizations of an image are effectively unlimited. More fundamentally for our purpose, even a rich dictionary applied to published images measures joint differences in the bundle of attributes that outlets select. We instead target the response to one specified attribute under a within-image counterfactual. This yields an interpretation in terms of the isolated contribution of that attribute. Our study advances previous research in three significant ways. First, we enhance the traditional two-stage approach by addressing its inherent biases and issues with omitted variables, introducing an estimation procedure that uses the rich information in the image content to measure the polarization in visual content. Second, our method delivers results that are both precise and available at a lower level of granularity, enabling us to quantify polarization at both the news outlet and politician levels and to track their evolution over the past decade. Third, we leverage a long term dataset spanning from 2011 to 2021, which allows us to capture long-term trends in political coverage that earlier studies may have missed.

The paper also relates to a methodological literature on inference with machine-learned variables. Wei and Malik (2022) characterize the biases that arise when econometric models use variables extracted from unstructured data, and Singh and Zheng (2023) and Xu et al. (2024) propose estimators that incorporate high-dimensional unstructured data into causal analysis directly. Our treatment of extraction bias and omitted-variable bias in §7.1 follows this line. Finally, a growing body of work uses generative models to construct controlled visual variation. Athey et al. (2022) manipulate smiles in profile images on a micro-lending platform, Ludwig and Mullainathan (2024) alter facial features to study judicial decisions, Luo and Toubia (2024) generate faces to study discrimination in online marketplaces, and Li et al. (2024) integrate consumer preferences into generative product design. These studies use manipulated images to estimate downstream responses or to generate hypotheses about behavior. We use the same technology to construct a measure of heterogeneity in editorial choice.

The remainder of the paper proceeds as follows. Section 2 describes the data and sample construction. Section 3 develops the intuition and defines visual slant, visual polarization, and their aggregate measures. Section 4 establishes identification and studies how estimation error affects these measures. Section 5 explains our choice of visual feature, presents the estimation procedure, and evaluates its components. Section 6 reports differences across outlets and politicians and changes over time. Section 7 compares our approach with feature-extraction methods and external benchmarks. Section 8 concludes.

2 Data

Our analysis uses photographs of U.S. politicians published by major news outlets, together with information about the articles in which they appeared. We describe how we collect these images and article information, then explain how we identify photographs containing a single face belonging to the intended politician.

2.1 Data Collection

We collect photographs of 30 politicians from 20 news outlets over the period 2011–2021. The set of politicians consists of those who ran for important public offices and/or held important national roles during the ten-year span of 2011–2021 and were commonly searched on Google (based on their popularity on Google searches using Google Trends data). The news outlets used in our study consist of the list of popular outlets that have been used in earlier studies on media polarization; see Flaxman et al. (2016) for details. We refer readers to Web Appendix §A.1 for a full list of politicians and outlets.

The data collection is done using the SerpAPI application (SerpAPI, 2023). SerpAPI facilitates efficient large-scale image scraping by leveraging Google's image search to extract relevant images and metadata. The process involves generating search queries based on specific criteria, such as targeted news outlets and date ranges, to extract images, article links, titles, and dates. For each politician–outlet pair, we use two-year search windows that advance one year at a time, issuing ten queries. This approach was chosen because it provides a structured, efficient, and reliable interface for large-scale data collection, automating the process while ensuring compliance with Google's data access policies.



Figure 2: Search results for "Joe Biden" site: "https://www.cnbc.com" before:2022 after:2020.

Figure 2 presents an example of a query we use for data collection, where we use "site:" to restrict searches to specific news websites, "before:" and "after:" to filter results by publication date, and exact phrase searches to capture relevant content precisely. For each query, we aim to collect 80 news articles for each politician from each news outlet for the specified period.[2] This information for each image in the query is compiled into a structured data frame with the following fields:

- Image: A URL pointing to the image.
- Alt: A short description or alternate text corresponding to the image.
- Href: A hyperlink where the related news article can be found.
- Title: The title of the news article or caption associated with the image.
- Publication Date: The publication date of the associated news article.

[2]Some queries retrieve fewer articles/images for certain politicians-outlet combinations because the outlet may not have published 80 images for that specific politician.

- Query Parameter: The search query string utilized to retrieve the image and associated news, emphasizing the political figure and the source website along with a defined temporal span.

Overall, our data collection strategy gives us a comprehensive data set of 287,275 images for 30 politicians from 20 news outlets.

2.2 Data Cleaning

We clean the search results to retain images containing a single face belonging to the intended politician. This restriction ensures that each news outlet's selection reflects its portrayal of the intended politician, avoiding confounding effects from multiple individuals in the same image. Three types of images must be excluded:

- Images showing relevant scenes or events without the intended politician.
- Images showing another politician despite being retrieved under the intended politician's name.
- Images containing multiple faces.

To address these challenges, we employ a two-phase computer vision framework. We provide a brief overview of this framework here and refer readers to Web Appendix §A.2 for the technical details. In the first phase, we use a series of computer vision models to keep only images with one face presented in them. In the second phase, we build and deploy a face-verification tool to ensure that the face in an image belongs to the intended politician. First, we manually select 20 high-quality, single-face images for each of the 30 politicians. These selected images serve as true labels identifying the correct politician, ensuring a reliable foundation for training the face verification model. Next, we use the trained verification model for each instance in our one-face sample obtained from the first phase to verify that the predicted label for the face is the same as the intended politician.

After applying the above data cleaning procedure, we are left with a set of 63,188 images, where each image shows a single face that belongs to the intended politician. Web Appendix §A.2 provides further details on our two-step verification and includes a comprehensive table (Figure A.2) listing the number of images for each politician-outlet combination after cleaning. As expected, the sample contains more images of presidents, presidential candidates, and other prominent political figures than of less prominent politicians, reflecting differences in the news coverage they receive (Web Appendix Figure A.2).

3 A Measure of Visual Polarization

Our objective is to develop measures of visual slant and visual polarization that can be estimated from news articles and their accompanying images. These measures should capture differences across outlets in how they prefer to portray a politician, while accounting for the many attributes that vary jointly across photographs. We begin with the intuition underlying our proposed measures in Section 3.1, then define visual polarization and visual slant using a model of editorial image choice in Section 3.2. Section 3.3 aggregates the article-level measures to characterize outlets' partisan visual slant and disagreement across outlets in their portrayal of individual politicians.

3.1 The Measurement Problem and Core Idea

A central challenge is to specify what constitutes a more favorable visual portrayal. A photograph combines many attributes, including facial expression, pose, lighting, and background, and these attributes can convey different impressions of the politician. Rather than assign an overall judgment of favorability to the entire image, we define our measures with respect to a chosen visual feature. We ask how outlets differ in their preference for that feature when the article and all other image attributes are held fixed.

To illustrate, take whether the politician is smiling as the focal feature. Consider an article about Joe Biden entering the White House in January 2021. Suppose that *Fox News* and *CNN* each choose between two versions of the same photograph, one showing Biden smiling and the other showing him with a neutral expression, with all other visual content unchanged. Figure 3 illustrates this hypothetical choice. Each outlet has some probability of selecting the smiling version. If both outlets are equally likely to choose the smiling version, they have the same preference for smiling in this example. This holds whether both favor the smiling version or both favor the neutral version. We seek to measure how much their preferences differ.

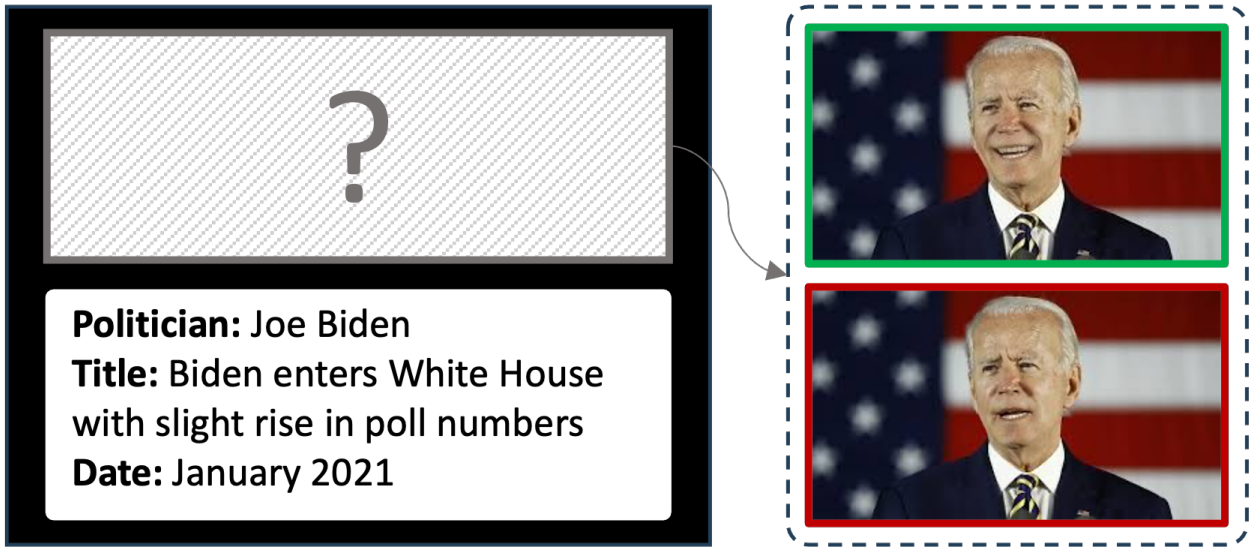


Figure 3: Editorial Choice between Smiling and Neutral Portrayals

To measure polarization with respect to a specific visual feature, we need to distinguish outlets' preferences for that feature from their preferences for other attributes of the photograph. For smiling, this means separating an outlet's preference for a smile from its preferences for the setting, lighting, or background. A natural approach is to use an image-analysis model to measure the feature in each photograph and compare its prevalence across outlets. Extracting a smile indicator discards the rest of the image. Yet the discarded attributes may be related both to whether the politician is smiling and to which outlet publishes the photograph. A comparison based only on the smile indicator can therefore reflect outlets' preferences for those accompanying attributes as well as for smiling itself.

We address this problem by constructing two versions of the same photograph that differ only in the focal feature. Generative AI allows us to create these counterfactual image pairs while holding the remaining visual content fixed. We use a model of editorial image choice to define polarization through outlets'

preferences between these two versions.

The key comparison is therefore between outlets' preferences for two versions of the same photograph that differ only in a chosen visual feature. In our main application, that feature is smiling: we compare a smiling portrayal with a neutral portrayal while holding the article and the remaining visual content fixed. Section 5.1 explains this empirical choice. We next formalize this comparison for a general visual feature through a model of editorial image choice and define the corresponding measures of visual polarization and slant.

3.2 The Structural Parameter

Consider article i about politician p, with textual and contextual characteristics x_i, and let $\mathscr{Y}$ denote the outlets in our analysis. Fixing the article and outlet, we model the editor as choosing the photograph z that maximizes

$$U_{iy}(z) = u(z, x_i, p, y) + \xi_{izy}, \tag{1}$$

where $u(\cdot)$ captures the outlet's systematic valuation of an image in the article's context and ξ_{izy} is an idiosyncratic shock, distributed i.i.d. Type I Extreme Value across photographs, articles, and outlets.

Let T denote a binary visual feature. Fix a pair (Z_p^1, Z_p^0) of versions of the same photograph of politician p: the feature is present in Z_p^1 and absent in Z_p^0, while all remaining visual content is held fixed. For smiling, these are the smiling and neutral versions. Consider a hypothetical choice in which outlet y selects between only these two versions for article i. Under the utility model, the probability of selecting the version with the feature is

$$q\big(Z_p^1 \mid y, x_i, p\big) = \frac{e^{u(Z_p^1, x_i, p, y)}}{e^{u(Z_p^1, x_i, p, y)} + e^{u(Z_p^0, x_i, p, y)}}. \tag{2}$$

The probability of choosing the version without the feature is $q(Z_p^0 \mid y, x_i, p) = 1 - q(Z_p^1 \mid y, x_i, p)$. The log odds of choosing the version with the feature equal the difference in its systematic valuation:

$$\log \frac{q(Z_p^1 \mid y, x_i, p)}{q(Z_p^0 \mid y, x_i, p)} = u(Z_p^1, x_i, p, y) - u(Z_p^0, x_i, p, y). \tag{3}$$

This difference measures how much outlet y values the presence of the focal feature in this image and article context. We define polarization by comparing this incremental valuation across two outlets.

Definition 1 (Visual Polarization)**.** *For article i about politician p and outlets y_1 and y_2, visual polarization with respect to feature T is*

$$\rho_i^T(y_1, y_2) = \big[u(Z_p^1, x_i, p, y_1) - u(Z_p^0, x_i, p, y_1)\big] - \big[u(Z_p^1, x_i, p, y_2) - u(Z_p^0, x_i, p, y_2)\big]. \tag{4}$$

Equivalently, the log odds interpretation of the measure is:

$$\rho_i^T(y_1, y_2) = \log \frac{q\big(Z_p^1 \mid y_1, x_i, p\big)}{q\big(Z_p^0 \mid y_1, x_i, p\big)} - \log \frac{q\big(Z_p^1 \mid y_2, x_i, p\big)}{q\big(Z_p^0 \mid y_2, x_i, p\big)}. \tag{5}$$

A positive value of $\rho_i^T(y_1, y_2)$ means that outlet y_1 values the presence of the focal feature more than outlet

y_2 does; a negative value indicates the reverse. Its absolute magnitude measures the extent of disagreement. A zero value means that the outlets have the same incremental valuation of the feature for this image pair and article. Exponentiating $\rho_i^T(y_1, y_2)$ gives the ratio of their odds of choosing the version with the feature.

Visual polarization compares two outlets but does not locate either outlet relative to a common benchmark. To define visual slant for an individual outlet, we designate a reference outlet.

Definition 2 (Visual Slant)**.** *For article i about politician p, the visual slant of outlet y with respect to feature T is*

$$\tau_i^T(y) = \rho_i^T(y, y_n), \tag{6}$$

where y_n denotes the reference outlet.

We use *Reuters* as our reference. This normalization defines slant relative to *Reuters*; it does not require that *Reuters* be perfectly neutral. The reference outlet has zero slant by construction: $\tau_i^T(y_n) = 0$. The reference affects outlet-specific slant but not pairwise polarization, since

$$\rho_i^T(y_1, y_2) = \tau_i^T(y_1) - \tau_i^T(y_2) \tag{7}$$

for any common reference outlet.

3.3 Aggregate Measures

We aggregate the article-level measures to answer two questions about the focal visual feature. First, how does an outlet's preference for the feature differ between Republican and Democratic politicians, relative to the reference outlet? Second, how much do outlets disagree in their valuation of the feature for a given politician? We address these questions using Conservative Visual Slant (CVS) and Overall Visual Polarization (OVP), respectively.

Let P_i denote the politician featured in article i, let $\mathscr{Y}$ denote the full set of 20 outlets evaluated for every article context, and let $\mathscr{P}_G$, for $G \in \{R, D\}$, denote the sets of Republican and Democratic politicians. Define average article-level slant for outlet y and politician group G as

$$\bar{\tau}^T(y \mid \mathscr{P}_G) = \mathbb{E}\left[\tau_i^T(y) \,\middle|\, P_i \in \mathscr{P}_G\right]. \tag{8}$$

Definition 3 (Conservative Visual Slant)**.** *For outlet y,*

$$\mathrm{CVS}(y) = \bar{\tau}^T(y \mid \mathscr{P}_R) - \bar{\tau}^T(y \mid \mathscr{P}_D). \tag{9}$$

To construct CVS, we first average an outlet's article-level slant separately for Republican and Democratic politicians, then take the difference between these two averages. The averaging is over articles within each politician group, so politicians represented by more articles receive greater weight within their group. In our smile application, a positive CVS means that, relative to *Reuters*, the outlet has a stronger average preference for smiling portrayals of Republican politicians than of Democratic politicians. A negative value indicates the reverse. The measure compares these preferences across the two groups; it does not require that an outlet favor smiling portrayals of one group and disfavor them for the other. Differencing removes any component

of the outlet's relative preference for smiling images that is common across the two groups. We use CVS as our main outlet-level measure and compare it with external measures of media slant.

The second aggregate, OVP, measures disagreement across outlets in their valuation of the focal feature for a given politician:

Definition 4 (Overall Visual Polarization)**.** *For politician p,*

$$\text{OVP}(p) = \mathbb{E}\left[\sqrt{\text{Var}_{y\in\mathscr{Y}}\left(\tau_i^T(y)\right)} \;\middle|\; P_i = p\right], \tag{10}$$

where the variance is taken across all 20 outlets in $\mathscr{Y}$ for each article context i, and the expectation is taken across articles featuring politician p.

To construct OVP, we first calculate the standard deviation of article-level slant across outlets for each article about the politician. We then average these standard deviations across the articles featuring that politician. Overall Visual Polarization (OVP) measures how dispersed outlets' valuations of the focal feature are for a given politician. Higher OVP indicates that a politician is more visually polarizing with respect to the focal feature: outlets disagree more strongly in their valuation of that feature when portraying the politician. In the smile application, this means greater disagreement over smiling versus neutral portrayals. OVP is zero when all compared outlets have the same incremental valuation of smiling in each article context, whether they all favor the smiling version or all favor the neutral version.

These measures describe differences in editorial valuations, but the underlying image choices are unobserved. The next section establishes how the measures can be recovered from the distribution of published articles and photographs.

4 Theoretical Analysis

This section establishes identification and examines the statistical properties of the proposed estimator. Section 4.1 connects the structural polarization measure to observable outlet probabilities and states the identifying assumptions. Section 4.2 studies estimation error and the construction of the aggregate measures.

4.1 Identification

The measures in Section 3 depend on how editors would choose between two versions of the same photograph. Our data record the photographs outlets publish, but not the alternatives available to their editors or their choices between the hypothetical versions. Identification therefore requires connecting the structural valuation differences to the joint distribution of observed articles, images, and outlets.

4.1.1 Linking Editorial Choice to Observed Data

For an observed image z and article context (x, p), the data identify the conditional probability that the article comes from outlet y. Denote this probability by

$$g(y \mid z, x, p) = \Pr(Y = y \mid Z = z, X = x, P = p), \tag{11}$$

where the probability refers to the distribution of articles in our analysis sample. We now derive its relationship to editorial valuations. The structural measure differences utility across two margins, image version

and outlet. Reversing the order of these differences leaves the measure unchanged:

$$\begin{aligned}\rho_i^T(y_1,y_2) &= \big[u(Z_p^1,x_i,p,y_1) - u(Z_p^0,x_i,p,y_1)\big] - \big[u(Z_p^1,x_i,p,y_2) - u(Z_p^0,x_i,p,y_2)\big] \\ &= \big[u(Z_p^1,x_i,p,y_1) - u(Z_p^1,x_i,p,y_2)\big] - \big[u(Z_p^0,x_i,p,y_1) - u(Z_p^0,x_i,p,y_2)\big].\end{aligned} \tag{12}$$

The first line compares the incremental valuation of the focal feature across the two outlets. The second first compares the outlets' valuations of a fixed image, then differences that comparison across the two image versions. This suggests recovering polarization from relative outlet valuations at each image. To connect these valuations to g, we must account for both story coverage and photograph availability.

Let $\pi(y \mid x,p)$ denote the probability that outlet y covers a story with characteristics (x,p), and let $\phi_y(z \mid x,p)$ denote the probability that photograph z is available to that outlet for the story. Conditional on covering the story, the probability that outlet y publishes photograph z is

$$h(z \mid y,x,p) = \frac{\phi_y(z \mid x,p)\, e^{u(z,x,p,y)}}{\mathscr{I}_y(x,p)}, \qquad \mathscr{I}_y(x,p) = \sum_{z' \in \mathscr{Z}_p} \phi_y(z' \mid x,p)\, e^{u(z',x,p,y)}. \tag{13}$$

This formulation is an availability-weighted conditional logit over the space $\mathscr{Z}_p$ of photographs of p that could accompany the article, and allows us to distinguish the actual publication probability from the hypothetical binary-choice probability q of Section §3.2. Availability enters multiplicatively, as a supply shifter separate from preferences: a frame an outlet rarely receives is one it rarely runs, however much it would like to. The denominator $\mathscr{I}_y(x,p)$ is the expected *inclusive value* of outlet y's choice set, one number summarizing how attractive the whole collection on its desk is.

This specification draws on the literature that models choice when consideration sets or product availability are unobserved (Goeree, 2008; Manzini and Mariotti, 2014; Abaluck and Adams-Prassl, 2021; Barseghyan et al., 2021; Conlon and Mortimer, 2013). Web Appendix §B.1 derives the exact publication probability and relates it to Equation (13). The availability-weighted logit provides an approximation when collection attractiveness concentrates around its expected value and no individual photograph materially changes it.

Applying Bayes' rule, we can connect the observable outlet probability g to coverage and publication probabilities as follows:

$$g(y \mid z,x,p) = \frac{\pi(y \mid x,p)\, h(z \mid y,x,p)}{\sum_{y' \in \mathscr{Y}} \pi(y' \mid x,p)\, h(z \mid y',x,p)}, \tag{14}$$

whose denominator is common to all outlets and so drops out of the odds ratio between two outlets. Substituting Equation (13) and taking logs, we can write:

$$\begin{aligned}\log \frac{g(y_1 \mid z,x,p)}{g(y_2 \mid z,x,p)} = \underbrace{\big[u(z,x,p,y_1) - u(z,x,p,y_2)\big]}_{\text{relative valuation of } z} + \underbrace{\log \frac{\phi_{y_1}(z \mid x,p)}{\phi_{y_2}(z \mid x,p)}}_{\text{relative availability of } z} \\ - \underbrace{\log \frac{\mathscr{I}_{y_1}(x,p)}{\mathscr{I}_{y_2}(x,p)}}_{\text{relative inclusive value}} + \underbrace{\log \frac{\pi(y_1 \mid x,p)}{\pi(y_2 \mid x,p)}}_{\text{relative coverage}}.\end{aligned} \tag{15}$$

Relative outlet log odds therefore reflect differences in editorial valuations, but also differences in photograph availability, the attractiveness of the available image collections, and the propensity to cover the story. Recovering visual polarization requires separating the valuation differences from these other influences.

4.1.2 An Observable Counterpart

Our data record the photographs outlets publish, rather than the alternatives their editors considered. From these observations, we can estimate the probability g that an image and article context are associated with each outlet. Comparing these probabilities across matched image versions provides an estimable counterpart to our theoretical polarization measure. We define this contrast as the difference in outlet-prediction log odds across the two versions, holding the politician and article context fixed.

Definition 5 (Outlet-Prediction Log-Odds Contrast)**.** *For article i about politician p, matched image versions Z_p^1 and Z_p^0 that differ in feature T, and outlets y_1 and y_2, define*

$$\tilde{\rho}_i^T(y_1,y_2) \equiv \log \frac{g(y_1 \mid Z_p^1, x_i, p)}{g(y_2 \mid Z_p^1, x_i, p)} - \log \frac{g(y_1 \mid Z_p^0, x_i, p)}{g(y_2 \mid Z_p^0, x_i, p)}. \tag{16}$$

To relate this contrast to structural polarization, take the difference across image versions in Equation (15). For any component $F(z)$, write $\Delta_i F = F(Z_p^1) - F(Z_p^0)$, holding (x_i, p) fixed. Then

$$\begin{aligned} \tilde{\rho}_i^T(y_1,y_2) = \rho_i^T(y_1,y_2) &+ \underbrace{\Delta_i \log \frac{\phi_{y_1}(z \mid x_i, p)}{\phi_{y_2}(z \mid x_i, p)}}_{\text{availability}} \\ &\underbrace{- \Delta_i \log \frac{\mathscr{I}_{y_1}(x_i, p)}{\mathscr{I}_{y_2}(x_i, p)}}_{\text{inclusive value}} + \underbrace{\Delta_i \log \frac{\pi(y_1 \mid x_i, p)}{\pi(y_2 \mid x_i, p)}}_{\text{coverage}}. \end{aligned} \tag{17}$$

Equation (17) shows that the observable contrast $\tilde{\rho}_i^T$ can differ from structural polarization ρ_i^T because outlet probabilities reflect photograph availability, the attractiveness of the available collection, and story coverage as well as editorial valuations. To recover structural polarization, these other influences must cancel when we compare the two image versions. This leads to the identifying conditions below.

4.1.3 Assumptions and Identification

Throughout the identification analysis, we compare matched image versions that differ only in the focal feature, following the definition of our measure. We then impose three assumptions to recover this comparison from observed data.

The first restricts the relative availability of the matched image versions; the second ensures positive probabilities and data support at the evaluation points; and the third separates story coverage from photograph selection. Together with the maintained publication model, these conditions connect the estimable contrast in Section 4.1.2 to the theoretical polarization measure.

Assumption 1 (Symmetric Availability of Matched Frames)**.** *For every article i in the evaluation sample,*

about politician p, and every outlet $y \in \mathscr{Y}$,

$$\phi_y\left(Z_p^1 \mid x_i, p\right) = \phi_y\left(Z_p^0 \mid x_i, p\right). \tag{18}$$

Equal availability across the two versions within each outlet makes the availability ratio between outlets the same for both versions, so the availability term in Equation (17) cancels when we difference across versions.

Assumption 1 requires the two matched versions to be equally likely to reach a given outlet, while allowing access to photographs to differ across outlets. News agencies and image providers, including Reuters, the Associated Press, and Getty Images, distribute photographs of political events to news outlets. Competing outlets may therefore draw on overlapping collections of photographs of the same politician at the same event while selecting different expressions or poses to accompany their articles.

Our matched image versions depict the same politician in the same setting and differ only in expression. We therefore interpret them as alternative frames from the same photographic moment. Access to photographs from that occasion depends on factors such as wire subscriptions, press credentials, or a photographer's presence. These occasion-level sources of access motivate equal availability of the two versions within each outlet.[3] Web Appendix §B.2 provides further discussion of this availability condition.

Assumption 2 (Overlap)**.** *There exist $\underline{\phi} > 0$ and $\underline{\pi} > 0$ such that, for every article i in the evaluation sample, every outlet $y \in \mathscr{Y}$, and each $t \in \{0,1\}$:*

(i) $\phi_y\left(Z_p^t \mid x_i, p\right) \geq \underline{\phi}$ *and* $\pi(y \mid x_i, p) \geq \underline{\pi}$*; and*

(ii) $\left(Z_p^t, x_i, p\right)$ *lies in a region the data inform: there is a neighborhood of the evaluation point, in the representation on which the estimate of g conditions, that receives positive mass in the training distribution.*

Part (i) requires each outlet to have a positive probability of covering the story and receiving either photograph. This gives us positive publication probabilities and lets us calculate the outlet log odds. The condition fits comparisons among outlets that cover the same politicians and draw on photographs of their public appearances.

Part (ii) requires the training data to contain images and article contexts near the points where we evaluate g. We construct each pair from an ordinary photograph of the politician and change only the focal expression, preserving the surrounding visual context. This construction helps keep the comparison close to images the model encounters in training. Section 5.2.4 examines contextually similar photographs across outlets and how the model responds when we add a smile. These exercises illustrate the support condition; Section 4.2 states the separate convergence condition we use for estimation. Web Appendix §B.2 discusses the two parts of overlap in detail.

Assumption 3 (Coverage Precedes Selection)**.** *The coverage propensity $\pi(y \mid x, p)$ does not depend on the photograph selected.*

[3] Web Appendix §B.3 shows that identification requires only the weaker condition that the two outlets have the same ratio of availability probabilities across versions.

Under Assumption 3, an outlet first decides to report on a politician or event and then selects a photograph to accompany the article. This sequence lets us compare how the outlet portrays the politician while holding its decision to cover the story fixed. Because the coverage term depends only on (x, p), differencing across image versions removes it. Web Appendix §B.2 provides further discussion of this timing restriction.

We can now apply these conditions to Equation (17). Symmetric availability removes the availability term, and the timing of coverage and image selection removes the coverage term. Our publication model assigns the same inclusive value to both versions within each outlet, so differencing also removes that term. Part (i) of Assumption 2 lets us form the log odds, while part (ii) supplies data support for estimation.

Thus the observable log-odds contrast equals the structural valuation difference. The following proposition formalizes this identification result.

Proposition 1 (Identification)**.** *Under Assumptions 1 and 3, and part (i) of Assumption 2,*

$$\tilde{\rho}_i^T(y_1, y_2) = \rho_i^T(y_1, y_2)$$

for every article i in the evaluation sample and every pair of outlets in $\mathscr{Y}$.

Proof. See Web Appendix §B.3.

Proposition 1 lets us recover the structural valuation difference using the outlet-probability contrast in Equation (16). We can therefore identify visual polarization from the joint distribution of (Y, Z, X, P) without observing editors' collections, their inclusive values, photograph availability, or the utility function. A positive value means that outlet y_1 has a greater incremental valuation of the feature than outlet y_2 for this image pair and article context; a negative value indicates the reverse.

Setting $y_2 = y_n$ gives the corresponding observable slant contrast relative to the reference outlet, which we take to be *Reuters*. Under the same conditions, this contrast recovers structural slant. Applying the aggregations in Section 3.3 then recovers the corresponding CVS and OVP measures. Section 5 explains how we estimate the outlet probabilities and evaluate these contrasts empirically.

In summary, the difference-in-differences structure of our estimand, together with the use of two counterfactual images that differ only in one feature, allows the additional terms to cancel under plausible assumptions. The latter property also highlights an important restriction: the parameter is defined for a specific pair of photographs and can be sensitive to the choice of photographs. In Section 5, we construct several pairs per politician and average the estimates across them to ensure that our estimates are less dependent on the specific images.

4.2 Statistical Properties

Section 4.1 shows how outlet probabilities identify visual slant and polarization. In practice, we estimate these probabilities from data. We now examine how estimation error affects the resulting measures. We establish two results: accurate outlet-probability estimates yield accurate article-level measures, while measuring dispersion across outlets requires an additional correction for estimation noise.

4.2.1 Accuracy of the Article-Level Measures

Let $\hat{\rho}_i^T$ denote the contrast in Equation (16) with estimated probabilities $\hat{g}$ in place of g, and let $\hat{\tau}_i^T(y) =$

$\hat{\rho}_i^T(y, y_n)$ denote the corresponding slant estimate. Write $\|\hat{g} - g\|_{\infty,i}$ for the largest probability-estimation error across outlets at the two image versions for article i.

Assumption 2(ii) supplies data support near these evaluation points. To study estimation accuracy, we also consider convergence of the fitted probabilities to their population values at these points. Because our measures use logarithms, small probability errors can have large effects near zero. We therefore bound both the true and fitted probabilities away from zero. Let $\mathscr{S}_i = \{(Z_p^t, x_i, p) : t \in \{0,1\}\}$ denote the two evaluation points for article i.

Assumption 4 (Bounded Probabilities)**.** *There exists $\varepsilon \in (0,1)$ such that, for every article i in the evaluation sample and every $y \in \mathscr{Y}$, $g(y \mid \cdot) \geq \varepsilon$ at every point in $\mathscr{S}_i$, and $\hat{g}(y \mid \cdot) \geq \varepsilon$ at every such point with probability one.*

This condition provides a bound on the sensitivity of the logarithm to probability errors. The following proposition combines it with the identification conditions to bound estimation error in the structural measure.

Proposition 2 (Error Bound and Consistency)**.** *Suppose the conditions of Proposition 1 and Assumption 4 hold. For any article i and outlets y_1, y_2,*

$$\left|\hat{\rho}_i^T(y_1, y_2) - \rho_i^T(y_1, y_2)\right| \leq \frac{4}{\varepsilon} \|\hat{g} - g\|_{\infty,i}.$$

Consequently, if the estimated outlet probabilities converge to their true values at the evaluation points, then $\hat{\rho}_i^T(y_1, y_2)$ converges in probability to $\rho_i^T(y_1, y_2)$.

Proof. See Web Appendix §B.4.

Proposition 2 links the accuracy of estimated outlet probabilities to the accuracy of estimated polarization. The article-level polarization estimate, $\hat{\rho}_i^T(y_1, y_2)$, combines four estimated log probabilities, so its error is at most $4/\varepsilon$ times the largest probability-estimation error at the evaluation points. The same result applies to slant by setting $y_2 = y_n$. As the fitted probabilities converge to their true values, the article-level measures converge to their structural targets.

4.2.2 Aggregation and Estimation Noise

CVS averages article-level slant differences, so the error bound extends directly to this aggregate (Corollary 1). OVP instead averages the cross-outlet standard deviation of slant within each article context. Estimation noise can increase the variance underlying each standard deviation, making outlets appear to disagree more than they do. A shared classifier can also introduce the same outlet-specific errors across many articles, so adding articles need not eliminate this source of dispersion.

We address this problem through cross-fitting, following Gentzkow et al. (2019). We split the training data into two disjoint samples, fit a classifier on each, and use both classifiers to estimate slant for the same article contexts and outlets. We then calculate the covariance between the two sets of slant estimates. Under the error conditions below, this covariance recovers the underlying cross-outlet variance because independent estimation errors do not contribute to its expectation.

Let $e_i(y)$ denote the error in estimated slant, so that $\hat{\tau}_i^T(y) = \tau_i^T(y) + e_i(y)$, and let Var_y and Cov_y denote moments across the full outlet set $\mathscr{Y}$ for article context i. Proposition 3 compares the variance calculated from one estimated mapping with the covariance calculated from two independently estimated mappings.

Proposition 3 (Bias of Dispersion Estimators)**.** *Suppose the mean error $\mathbb{E}[e_i(y)]$ does not vary across outlets. Then the plug-in estimator is biased upward:*

$$\mathbb{E}\big[\text{Var}_y\big(\hat{\tau}_i^T(y)\big)\big] = \text{Var}_y\big(\tau_i^T(y)\big) + \mathbb{E}\big[\text{Var}_y\big(e_i(y)\big)\big] \;\geq\; \text{Var}_y\big(\tau_i^T(y)\big),$$

with strict inequality unless the error is constant across outlets. Suppose instead that $\hat{\tau}_i^{T,(1)}$ and $\hat{\tau}_i^{T,(2)}$ are constructed using two classifiers whose errors $e_i^{(1)}$ and $e_i^{(2)}$ are independent of one another and each have mean constant across outlets. Then

$$\mathbb{E}\big[\text{Cov}_y\big(\hat{\tau}_i^{T,(1)}(y),\, \hat{\tau}_i^{T,(2)}(y)\big)\big] = \text{Var}_y\big(\tau_i^T(y)\big).$$

Proof. See Web Appendix §B.4.

Proposition 3 isolates the estimation-noise contribution to dispersion: the variance from one mapping includes this contribution, whereas the covariance between two mappings removes it under the stated error conditions. The unbiasedness result concerns the variance underlying OVP. The truncation and square-root transformation used to report OVP can retain finite-sample bias.

These results guide our implementation: we estimate outlet probabilities to recover article-level slant and polarization, and use cross-fitted covariance to correct the variance component of OVP for estimation noise. Section 5 brings these steps together. We first explain the choice of smiling as the focal feature, then describe image generation, outlet prediction, and the evaluation of the resulting comparisons.

5 Empirical Implementation and Evaluation

In this section, we present the details of our estimation procedure and algorithm in Section 5.1. Next, in Section 5.2, we empirically evaluate different components of our algorithm using data.

5.1 Estimation Procedure

Our framework quantifies visual slant and polarization with respect to a chosen visual feature by comparing outlets' preferences for matched image versions that differ in that feature. In our main application, we focus on whether the politician is smiling. Smiling offers an interpretable dimension of favorable portrayal because it generally conveys a positive expression (Sülflow and Maurer, 2019). The attention politicians' smiles receive in campaigns and public commentary also illustrates their relevance to political imagery.[4] Our resulting slant measures describe outlets' preferences for smiling rather than neutral portrayals, and our polarization measures capture disagreement across outlets over that preference. We apply the same framework to brightness and frowning in Web Appendix §E.7.

Our estimation procedure has two main components: counterfactual image generation and news outlet prediction. Section 5.1.1 describes how we construct matched image pairs, and Section 5.1.2 explains how

[4]For example, Zohran Mamdani's smile featured prominently in his 2025 mayoral campaign and became the subject of parody in a Saturday Night Live sketch.

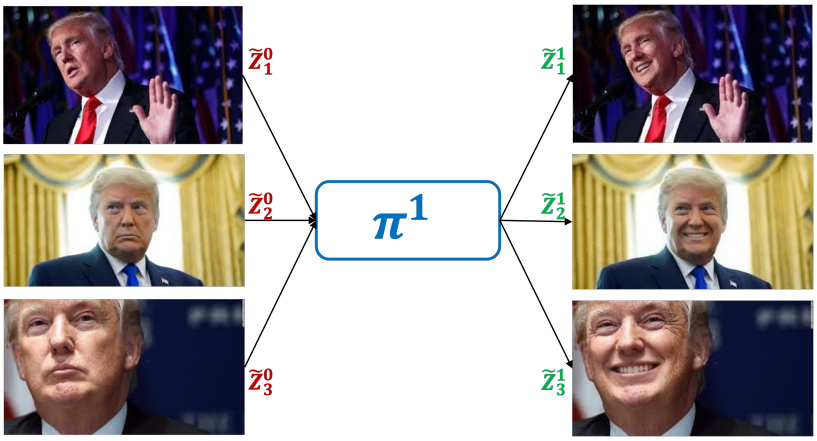


Figure 4: Counterfactual image generation for three neutral images of Donald Trump.

we estimate outlet probabilities. Section 5.1.3 brings these components together in the complete algorithm for estimating visual slant and polarization.

5.1.1 Counterfactual Image Generation

A crucial step in our algorithm is counterfactual image generation. We use recent advances in generative image models to modify a specific visual feature while holding the remaining image content fixed. Generative models have been employed in several recent studies on image analysis (Athey et al., 2022; Ludwig and Mullainathan, 2024; Luo and Toubia, 2024).

For each politician p, we select three neutral photographs, $\tilde{Z}^0_{p1}$, $\tilde{Z}^0_{p2}$, and $\tilde{Z}^0_{p3}$, from outside the training sample. Using three starting photographs reduces dependence on any particular portrayal while allowing us to inspect each generated pair manually. We choose photographs typical of those in the training data to keep the counterfactual comparisons close to image contexts the prediction model has encountered, consistent with the support requirement in Assumption 2(ii). Let π^1 denote the operation that adds the focal feature. In our smile application, π^1 adds a smile to each neutral photograph, while π^0 leaves the neutral photograph unchanged.

To generate the smiling version of each photograph, we employ AILabTools as our π^1 operator; this tool uses conditional Generative Adversarial Networks (cGANs) (Mirza and Osindero, 2014) and produces high-quality, realistic outputs. We generate the smiling counterparts $\tilde{Z}^1_{ps} = \pi^1(\tilde{Z}^0_{ps})$ for $s = 1,2,3$. Each pair consists of the original neutral photograph and its smiling version, designed to preserve the politician's identity and the rest of the visual content. We repeat this process for all 30 politicians, obtaining three matched pairs per politician. Figure 4 illustrates the construction for Donald Trump. To assess whether adding a smile preserves the rest of the photograph, we examine changes in other measured image characteristics in Section 5.2.1.

5.1.2 News Outlet Prediction via a Multi-Modal Deep Learning

We estimate a news outlet prediction model $g(y_i \mid Z_i, X_i, P_i; \theta)$ that maps article content and images into outlet likelihoods. Because the inputs are inherently multi-modal, including high-dimensional images, textual information, and contextual variables, we require a flexible prediction model that can learn complex

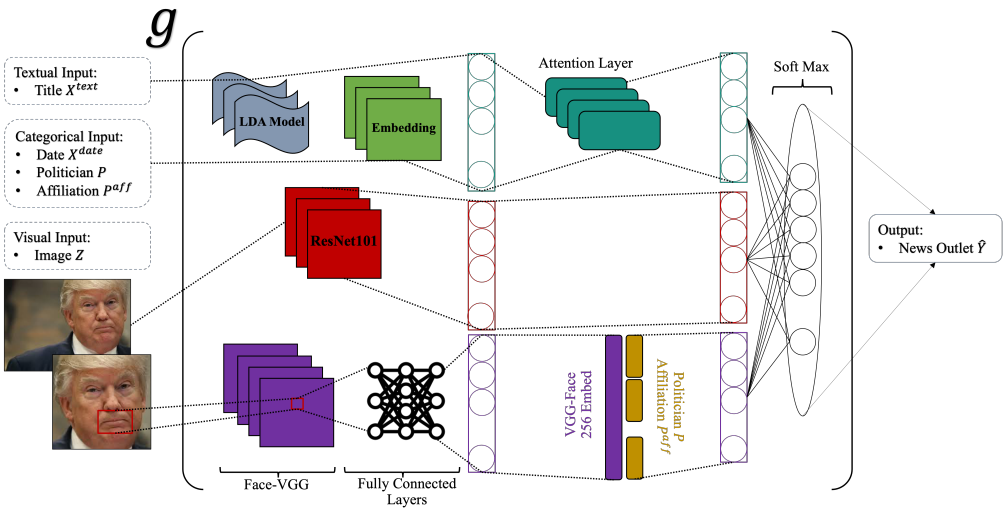


Figure 5: The multi-modal deep learning model for the news outlet prediction task

associations between visual content and outlet identity. In particular, we require the model to correctly attribute variation in outlet likelihoods to specific visual features, such as the presence of a smile, rather than to correlated visual or textual attributes.

We impose structure on the prediction architecture to ensure that the model learns the relationship between outlet identity and visual features in a way that supports counterfactual evaluation. Specifically, we design the model to jointly process visual and non-visual inputs while preserving a clear mapping from changes in visual content to changes in predicted outlet likelihoods. Figure 5 illustrates the architecture we use to implement this approach. The multi-modal deep learning model has the following architecture:

- First, to leverage the variation in similar events, we use the contextual information in news articles. This includes textual data (X^{text}), publication dates (X^{date}), politicians' names (P), political affiliations (P^{aff}), and image data (Z). Textual data is processed using Latent Dirichlet Allocation (LDA), which encodes each news title as a 40-dimensional topic vector. Second, we capture metadata such as dates, names, and affiliations as categorical variables by embedding them into dense vectors. Then, a *Attention Mechanism* processes the structured data (X^{text}, X^{date}, P, P^{aff}), and learns to prioritize the most relevant text and metadata features for the classification task (Vaswani, 2017).
- Second, *ResNet-101* processes the entire image Z by leveraging its capabilities from general image recognition tasks and extracts hierarchical and context-rich features from the broader visual content and scene information (He et al., 2016). Intuitively, given similarities in the event/news covered, these two parts of the model architecture effectively capture a given outlet's stylistic preference for certain styles of text and image features.
- Finally, to accurately capture smiles and link them to news outlets, we use *MTCNN* performs face detection and crops the face alone from the rest of the image (Zhang et al., 2016). We pass the detected face through a pretrained *VGG-Face* network to extract facial features (Parkhi et al., 2015), and fine-tune the network for our outlet-prediction task. Finally, we apply *Chunk attention* to VGG-Face embeddings,

and combine them with categorical data (politicians' names P and affiliations P^{aff}) to capture correlations between facial features and structured metadata related to the image. This step of separately capturing facial features and emotions ensures that the model accurately captures the relationship between smiles and news outlets, mitigating the risk of misattribution to correlated features.

The final classification layer combines all modalities. The model is trained using the AdamW optimizer (Loshchilov and Hutter, 2017), which ensures efficient optimization and robust regularization. To train the model, we maximize the following entropy function over the training data:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \sum_{i \in \mathscr{D}_{\text{train}}} \sum_{y \in \mathscr{Y}} w_y \mathbf{1}\{Y_i = y\} \log g(Y_i = y \mid Z_i, X_i, P_i; \theta), \tag{19}$$

The resulting model $\hat{g}$ is then used to estimate the polarization parameter using Equation (A.32). Please see Web Appendix §D.1 for additional details on implementation and model training.

5.1.3 Algorithmic Procedure

We present the complete estimation procedure in Algorithm 1. We split our data $\mathscr{D}$ into training and test sets, denoted by $\mathscr{D}_{\text{train}}$ and $\mathscr{D}_{\text{test}}$, respectively. We use the training data to train the deep learning model, and use the estimated model to calculate the polarization measure on the test data.

This training/test split lets us evaluate the measures on articles we did not use to fit the prediction model. For OVP, we additionally divide the training sample into two disjoint folds and fit a separate classifier on each. For each test article, we repeat Step 3 with each classifier to obtain two slant vectors across all 20 outlets. We then calculate their cross-outlet covariance, rather than the variance of a single vector. This is the cross-fitting correction discussed in Section 4.2; both mappings evaluate the same article contexts and matched image pairs. Under the error conditions in Proposition 3, this covariance is unbiased for the cross-outlet variance underlying OVP.

The reference outlet enters through y_n in Steps 3 and 4. To change the reference, we subtract the new reference outlet's estimated slant from every outlet's slant within each article context. This adjustment is common across outlets for that context, but can vary across articles. We can therefore change the reference without refitting the prediction model.

5.2 Empirical Evaluation of Algorithm's Components

We now present the results to validate different components of our algorithm. First, in §5.2.1, we provide evidence demonstrating that the counterfactual images generated in the Step 1 of our algorithm are consistent with the original images, preserving all features except the added smile. Next, in §5.2.2, §5.2.3, and §5.2.4, we present results on the performance of our news outlet prediction model from Step 2 of our algorithm.

5.2.1 Validating Counterfactual Image Consistency

This exercise examines the feature-isolation requirement underlying the structural comparison in Section 3.2 and the identification argument in Section 4.1. It is important to ensure that adding a smile does not systematically change other contextual features of the image. Ensuring this consistency is essential for isolating the effect of the smile without introducing biases from other image characteristics, such as brightness or colorfulness (see §B.5.2). To that end, we measure a few other image characteristics (such as brightness and

Algorithm 1 Estimating Visual Slant and Polarization

Input:
Data $\mathscr{D} = \{(Z_i, X_i, P_i, Y_i)\}_{i \in \mathscr{N}}$, partitioned into $\mathscr{D}_{\text{train}}$ and $\mathscr{D}_{\text{test}}$
Politicians $\mathscr{P}$; outlets $\mathscr{Y} = \{y_n, y^1, \ldots, y^K\}$ with reference outlet y_n
For each $p \in \mathscr{P}$, neutral photographs $\{\tilde{Z}_{ps}\}_{s=1}^{S}$ with $\tilde{Z}_{ps} \notin \mathscr{D}_{\text{train}}$, satisfying Assumption 2
Transformation π^1; class weights $\{w_y\}_{y \in \mathscr{Y}}$
Output: Visual slant $\hat{\tau}^T(p, y^k)$ for every $p \in \mathscr{P}$, $y^k \in \mathscr{Y}$

Step 1. Construct counterfactual pairs (§5.1.1)
for each $p \in \mathscr{P}$ and $s = 1, \ldots, S$ **do**
  $\tilde{Z}^1_{ps} \leftarrow \pi^1(\tilde{Z}_{ps})$ ▷ feature activated
  $\tilde{Z}^0_{ps} \leftarrow \tilde{Z}_{ps}$ ▷ feature absent
end for
Step 2. Estimate outlet probabilities on $\mathscr{D}_{\text{train}}$ (§5.1.2)
$\hat{\theta} \leftarrow \underset{\theta}{\operatorname{argmax}} \sum_{i \in \mathscr{D}_{\text{train}}} \sum_{y \in \mathscr{Y}} w_y \mathbf{1}\{Y_i = y\} \log g(y \mid Z_i, X_i, P_i; \theta)$
Write $\hat{g}(\cdot) \equiv g(\cdot; \hat{\theta})$
Step 2′. *(For dispersion measures only.)* Partition $\mathscr{D}_{\text{train}}$ into disjoint folds $\mathscr{D}^{(1)}, \mathscr{D}^{(2)}$; repeat Step 2 on each to obtain $\hat{g}^{(1)}, \hat{g}^{(2)}$
Step 3. Evaluate article-level slant on $\mathscr{D}_{\text{test}}$
for each $p \in \mathscr{P}$, $y^k \in \mathscr{Y}$, and article $j \in \mathscr{D}_{\text{test}}$ with $P_j = p$ **do**
  $$\hat{\tau}^T_j(y^k) \leftarrow \frac{1}{S} \sum_{s=1}^{S} \left[\log \frac{\hat{g}\left(y^k \mid \tilde{Z}^1_{ps}, x_j, p\right)}{\hat{g}\left(y_n \mid \tilde{Z}^1_{ps}, x_j, p\right)} - \log \frac{\hat{g}\left(y^k \mid \tilde{Z}^0_{ps}, x_j, p\right)}{\hat{g}\left(y_n \mid \tilde{Z}^0_{ps}, x_j, p\right)} \right]$$
end for
Step 3′. For OVP, evaluate both mappings. Repeat Step 3 with $\hat{g}^{(1)}$ and $\hat{g}^{(2)}$ to obtain $\hat{\tau}^{T,(1)}_j(y)$ and $\hat{\tau}^{T,(2)}_j(y)$ for every $y \in \mathscr{Y}$.
For each test article j, compute $\widehat{C}_j \leftarrow \operatorname{Cov}_{y \in \mathscr{Y}}\left(\hat{\tau}^{T,(1)}_j(y), \hat{\tau}^{T,(2)}_j(y)\right)$.
Step 4. Aggregate within politician
$$\hat{\tau}^T(p, y^k) \leftarrow \frac{1}{N^{\text{test}}_p} \sum_{j \in \mathscr{D}_{\text{test}}, P_j = p} \hat{\tau}^T_j(y^k), \quad N^{\text{test}}_p = |\{j \in \mathscr{D}_{\text{test}} : P_j = p\}|$$

Downstream aggregations. Polarization between any pair of outlets via (7); outlet-level CVS via (9); politician-level OVP by averaging $\sqrt{\max\{\widehat{C}_j, 0\}}$ over test articles about each politician, as in Equation (23)

colorfulness) before and after applying the smile operator π^1 and use the Kolmogorov-Smirnov (K-S) test to compare the distributions of these characteristics for the original and smiley images. Our tests suggest that there are no statistically significant differences between the original and smiley images for other contextual features. These results support preservation of the measured characteristics; isolation of all nonfocal visual content remains a requirement of the counterfactual design. Detailed analysis of the results is shown in the Web Appendix §D.3.

5.2.2 Predictive Performance of the News Outlet Prediction Model

In this section, we present the performance of our multi-modal multi-class classification problem for news outlet prediction. As mentioned earlier, we use an 85%-15% split between training and test data. We consider four measures to evaluate the model's predictive performance – (1) Accuracy, (2) Precision, (3) Recall, and (4) Weighted Cross-Entropy (WCE). Detailed explanations of these metrics are provided in Web Appendix §E.1.

Features Used	Accuracy (%)	Precision (%)	Recall (%)	WCE Loss
Politician, Party, Date	11.39	10.99	15.18	2.76
Politician, Party, Date, Text	16.72	15.01	17.08	2.66
Politician, Party, Date, Text, Image	44.04	43.82	40.90	2.28

Table 1: Performance metrics for the 20-class classification task on the test data, by input set.

We use the above performance measures to evaluate different versions of our deep learning model on the test data and present the results in Table 1. Each row shows a more complex model that progressively adds more explanatory variables/features. The first row considers a model that only uses categorical inputs such as politician identity, party affiliation, and date. We then add textual information in the second row and add image features in the final row. Two key points emerge from Table 1. First, our final outlet prediction model achieves a remarkably good performance of approximately 45%. It is worth emphasizing that a random benchmark only achieves a 5% accuracy (given that there are 20 news labels). Second, we notice that the image information greatly contributes to the predictive accuracy of the model: while the non-image models achieve a maximum accuracy of 16.72%, the multi-modal model achieves an accuracy of 44.52%. This aligns with prior research that highlights the importance of visual data in predictive tasks (Dzyabura et al., 2023).

We present additional details on the model's performance measures across outlets in Web Appendix E.1. In the rest of this section, we provide some additional interpretation and intuition for the results from the black-box predictive model. In §5.2.3, we discuss how our model captures the information in a smile, and in §5.2.4, we examine how outlet predictions shift when we add a smile to an image.

5.2.3 How Does the Model Account for Smiles?

Understanding how a multi-modal model learns and utilizes visual features is essential for interpreting predictions, especially in complex tasks like news outlet classification. However, this can be challenging since visual features are high-dimensional. In particular, it is important to examine whether the model is really utilizing information from facial features and emotions or whether it is simply utilizing the broader contextual information in the image for prediction (such as background color, text, etc.).

We employ a recently developed, popular tool – Grad-CAM – to visualize the contribution of specific visual features to the model's predictions by helping us understand the attention patterns of different components in the model (Selvaraju et al., 2017). Grad-CAM highlights the regions of an image that drive the decision-making process by examining the gradient of the model's prediction function $g(\cdot)$ with respect to its inputs. This technique is particularly helpful in our architecture, where two different CNN branches, Face-VGG and ResNet101, process facial features and global image context, respectively. We can therefore use Grad-CAM to analyze attention patterns in both CNN branches separately.

As an illustrative example, consider the last image of Donald Trump in Figure 6, from *Washington Post*. Figure 6 shows visualizations for two versions of this image: the original image (top row) and a smile-added version (bottom row) generated using the transformation function π^1. The Grad-CAM visualizations illustrate two distinct patterns of model attention. First, the modified ResNet101 heatmaps predominantly focus on broader contextual features, such as the lectern, the logo, and background elements, including the event

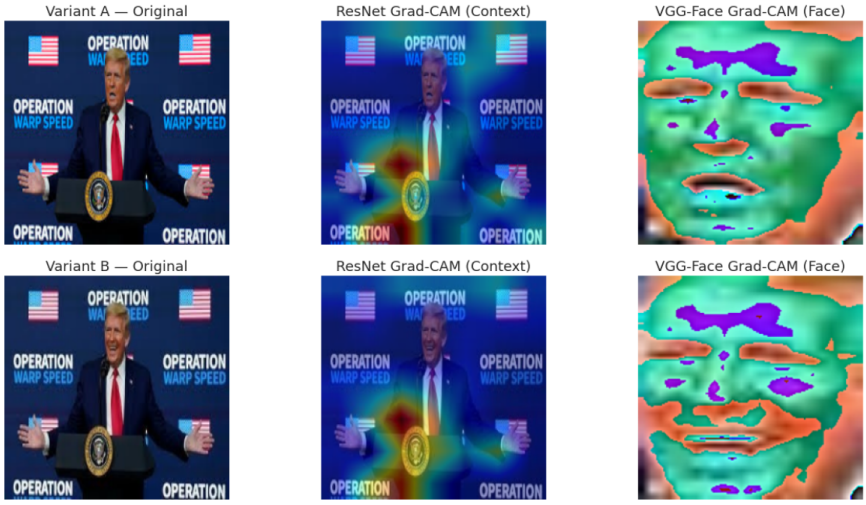


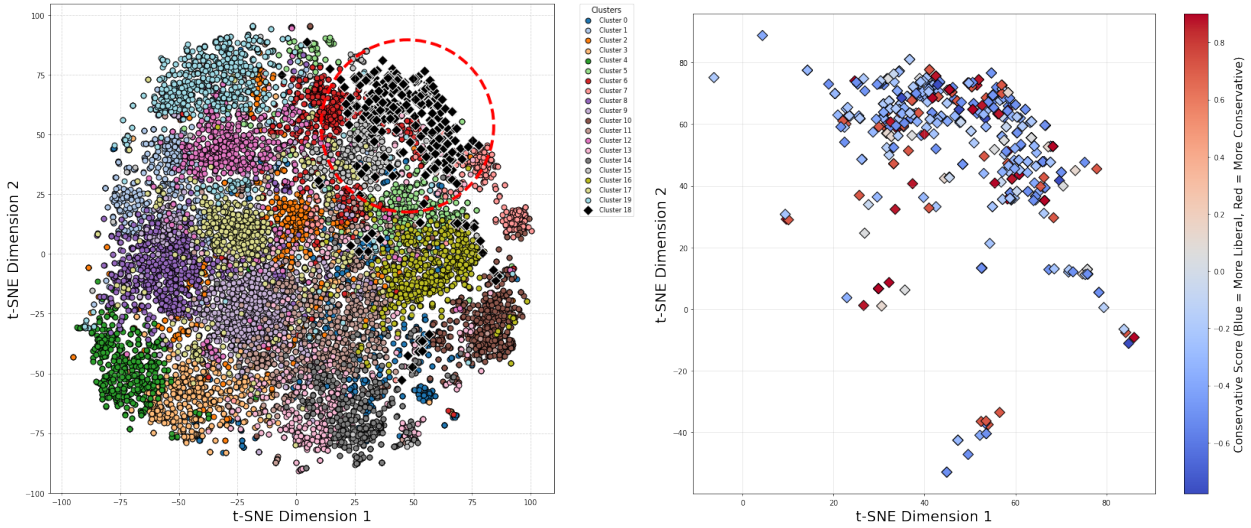
Figure 6: Grad-CAM visualizations for theoriginal and smile-added versions of a *Washington Post* image. The ResNet101 heatmaps (right column) capture information on the global image context, while the Face-VGG heatmaps (middle column) highlight the mouth region and facial expression changes.

name ("Operation Warp Speed"). The gradients of $g(\cdot)$ with respect to the global image embeddings emphasize that ResNet101 captures cues related to the overall spatial and contextual layout of the image, such as positioning and visual saliency. Second, the modified Face-VGG heatmaps display heightened sensitivity to facial features, with particular emphasis on the mouth region. For the original image, the gradient of $g(\cdot)$ with respect to the facial embeddings shows moderate activation around the mouth, corresponding to a neutral expression. When a smile is introduced, the activation in the mouth region intensifies and extends to other facial areas associated with smile dynamics. This pattern suggests that Face-VGG is able to successfully detect smile-related features and subtle changes in facial expressions, which provides strong evidence in support of including Face-VGG in our network architecture.

5.2.4 How Does Adding a Smile Change Predictions?

We now examine how adding a smile changes outlet predictions within sets of contextually similar photographs, using images from Operation Warp Speed as an illustration. This exercise relates to the support requirement in part (ii) of Assumption 2: counterfactual evaluation requires the fitted model to be informed by comparable images in the data. We examine the presence of similar photographs across outlets and the location of a smile-modified image relative to those photographs in a visual embedding. The clustering visualization and the prediction changes provide descriptive diagnostics; they do not establish overlap in the full input space or test the symmetric-availability and coverage-timing assumptions.

To do so, we turn to unsupervised learning techniques. First, we extract embeddings for all images using ResNet101. Each image is represented as a high-dimensional vector $\mathbf{e}_i \in \mathbb{R}^d$, where d denotes the dimensionality of the embedding space. These embeddings capture general visual features of the images, such as texture, color, and background features, without being tied to specific news outlets. To make the embeddings interpretable and suitable for visualization, we reduce their dimensionality to two dimensions using Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE),



(a) Clusters for Donald Trump: Images grouped into 20 clusters based on t-SNE projections.

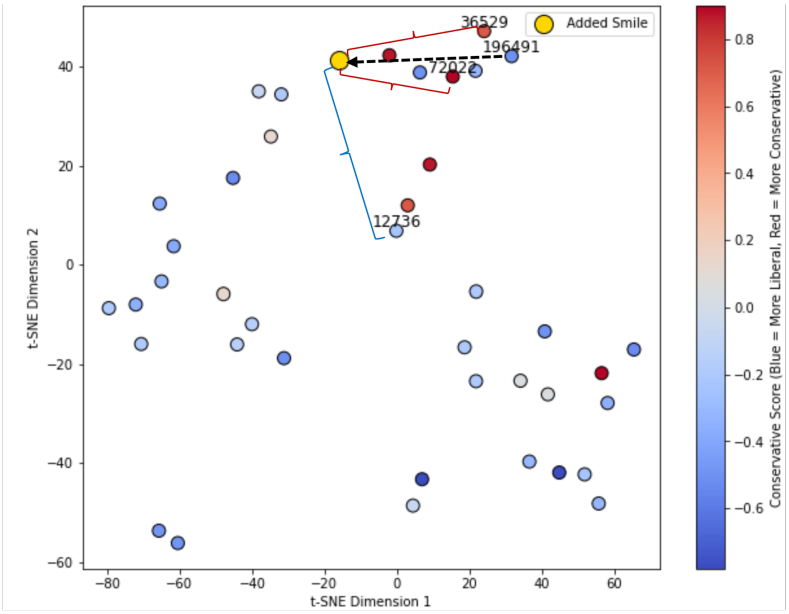
(b) t-SNE plot for Cluster 18: Highlighting conservative scores for images in this cluster.

Figure 7: Image clustering based on t-SNE.

followed by clustering to identify patterns. We present the details of PCA and t-SNE in our context in Web Appendix §D.2.

Figure 7a shows the t-SNE projection for all images of Donald Trump, grouped into 20 clusters based on their visual similarity. Each cluster captures images with shared contextual or visual features and provides a conceptualization of "similar events". For instance, consider Cluster 18, which is highlighted in Figure 7b. Here, each point represents an image, and the color indicates the conservative score of the news outlet that selected the image, ranging from liberal (blue) to conservative (red). The diversity in news outlets within the cluster suggests that multiple news outlets show images that are similar in structure/background (likely covering similar events). This variation is critical since it allows the outlet prediction model to capture how variation in facial features influences the likelihood of an image being chosen by a given outlet, conditional on the other image features (e.g., background, text, texture) being similar.

Figure 8a illustrates examples from Cluster 18, showcasing images used by different news outlets, while Figure 8b displays the corresponding images for the highlighted IDs. Within this cluster, consider the image corresponding to ID 196491 in our dataset by the *Washington Post*, which is the same image of Donald Trump used in the previous section. In this image, Trump appears unhappy, leaning slightly towards an angry expression. As shown in Figure 8a, this image is initially positioned on the top right-hand side of the plot. Next, we modify the original image (ID 196491) by applying the transformation function π^1, which adds a smile to Donald Trump while preserving the contextual elements of the image. After computing the embedding for this modified image, its coordinates in the t-SNE space are updated, as illustrated by the gold dot in the scatter plot. Interestingly, the smile-added version of the original image shifts to the left. Now, the updated position results in a shorter distance to images such as ID 36529 from the *Daily Mail* and ID 72022 from *Newsmax*, compared to its distance to the original image ID 196491 by the *Washington Post* and ID 12736 by *ABC News*. This demonstrates how the addition of a smile alters the embedding position to align



(a) Embedding positions of the original image (ID: 196491), similar images, and the image with a smile added (highlighted in yellow).



(b) Images corresponding to highlighted points in the t-SNE plot.

Figure 8: t-SNE plot of the Cluster 18 of Donald Trump's images illustrating the effect of adding a smile.

News Outlet	Original Image (%)	Smile-Added Image (%)
Daily Mail	3.78	20.60
Newsmax	1.90	5.54
ABC News	9.86	3.91
Washington Post	32.74	22.30

Table 2: Outlet predictions for the original and smile-added versions of the Donald Trump image.

with images that share similar facial expressions, even though the overall context remains unchanged.

Although the t-SNE coordinates provide interpretable insights into how our model works, they do not capture the full complexity of the image features that our model $\hat{g}$ captures. Therefore, in Table 2, we present our model's outlet predictions for the *Washington Post* image of Donald Trump in Figure 6 (the one that was originally shown in Washington Post) for two cases – the original image and smile-added image version. We see substantial differences in the outlet predictions between the original image and the counterfactual version with the smile. Notably, we find that adding a smile substantially increases the predicted probability for *Daily Mail* (from 3.78% to 20.60%) and *Newsmax* (from 1.90% to 5.54%), slightly decreases the predicted probability for *ABC News* (from 9.86% to 3.91%), and substantially decreases the predicted probability for *Washington Post* (from 32.74% to 22.30%).

In summary, we see that adding a smile systematically shifts the outlet prediction probabilities, consistent with the finding in the prior work in the advertising domain that shows facial expressions and emotions can shape audience engagement by capturing attention and influencing perception (Teixeira et al., 2012). Specifically, we see an increase (decrease) in the predicted probabilities for right-leaning (left-leaning) news outlets when we add a smile to an image of Donald Trump. Nevertheless, this is just a single instance in the

data that we use to illustrate the intuition behind how our algorithm works. In the next section, we present more systematic measures to quantify visual polarization.

6 Main Results

We present the substantive findings in three parts: visual slant across outlets in §6.1, visual polarization across politicians in §6.2, and changes over time in §6.3. Section 7 then examines external validity and compares CVS with a feature-extraction benchmark.

6.1 Visual Slant Across News Outlets

We now present our main results on visual slant. All results are direct applications of Algorithm 1 in §5.1.3. Before presenting the results, we review a few important considerations in applying our algorithm. First, recall that our visual slant parameter requires a neutral outlet denoted by y_n. We use *Reuters* as the baseline outlet y_n to measure visual slant, as *Reuters* is a largely neutral and fact-based outlet.[5] Second, we use a train-test split, and all the polarization and slant measures are shown for the test data $\mathscr{D}_{test}$ using the estimates obtained from the training data $\mathscr{D}_{train}$. Third, for each politician, we start with three neutral images, generate counterfactual smiling versions of these three images, and use these six images in our polarization and slant measurement for each politician.

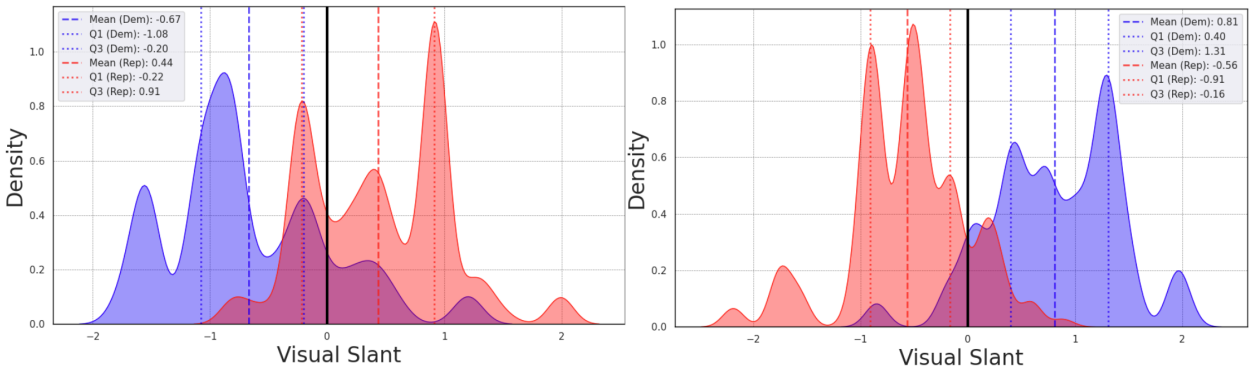
This section is organized as follows. In §6.1.1, we present the overall distribution of visual slant across all outlets and politicians. We then explore the extent of heterogeneity across news outlets in §6.1.2 to see which outlets exhibit higher levels of visual slant. In §6.2, we document the heterogeneity in visual slant at the politician level. Finally, in §6.3, we analyze how visual polarization evolves over time.

6.1.1 Distribution of Visual Slant

An important feature of our algorithm is its ability to produce an individual-level measure of visual slant $\hat{\tau}_i^T(y^k)$. As such, for each article i featuring politician p, we can estimate 19 visual slant measures corresponding to all the outlets other than Reuters. Based on the scores from Faris et al. (2017), AllSides (2024), and Flaxman et al. (2016), we categorize the three most Republican-leaning news outlets as: *Fox News*, *Newsmax*, and *Daily Mail*, and the three most Democratic-leaning news outlets as: *Washington Post*, *CNN*, and *The New York Times*. Figure 9 shows the distributions of visual slant for Democratic and Republican politicians as featured in Democratic- and Republican-leaning news outlets. If there were no visual slant or polarization, all distributions would be tightly centered around zero. However, we observe that for both Republican- and Democratic-leaning outlets, the individual-level visual slant measures deviate from zero and exhibit significant dispersion. This indicates the presence of a visual slant in these outlets. Additionally, we notice a clear divide in the distributions of Democratic- and Republican-leaning outlets for both Republican and Democratic politicians, pointing to evidence of visual polarization.

Next, we examine if there is a systematic difference between Republican- and Democratic-learning outlets by examining their averages in each panel. We denote the set of Republican and Democratic politicians by $\mathscr{P}_R$ and $\mathscr{P}_D$, respectively. Similarly, we denote the sets of Republican- and Democratic-leaning outlets defined above are denoted by $\mathscr{Y}_R$ and $\mathscr{Y}_D$, respectively. For Republican politicians, in Figure 9a, the

[5] It is worth emphasizing that one could easily change the choice of base. Although the visual slant measure will change with a different choice, the overall visual polarization between two outlets remains unchanged.



(a) Republican politicians.

(b) Democratic politicians.

Figure 9: Histograms of *visual slant* for Democratic and Republican Politicians in Democratic- and Republican-leaning News Outlets.

distributions of $\hat{\tau}_i^T(y^k)$ show distinct differences between left- and Republican-leaning news outlets. The visual slant measurement for left-leaning news outlets is *-0.67*, implying that, on average, smiling images of Republican politicians decrease the utility of a news outlet classified as Democratic-leaning compared to Reuters. In contrast, the visual slant measurement for Republican-leaning news outlets is *0.44*, suggesting that, on average, smiling images of Republican politicians significantly increase their utility compared to Reuters. Therefore, for Republican politicians, we can conclude that:

$$\bar{\tau}^T\left(y^k \in \mathscr{Y}_D \mid \mathscr{P}_R\right) < 0 < \bar{\tau}^T\left(y^k \in \mathscr{Y}_R \mid \mathscr{P}_R\right) \tag{20}$$

For Democratic politicians, in Figure 9b, the visual slant distribution for Democratic and Republican news outlets also shows clear differences. The average visual slant for Democratic news outlets is *0.81*, indicating that, on average, images of smiling Democratic politicians increase the utility of a Democratic-leaning outlet compared to Reuters (neutral baseline). Conversely, the average visual slant for Republican-leaning new outlets is *-0.56*, suggesting that, on average, smiling images of democratic politicians decrease their utility compared to Reuters. Therefore, for Democratic politicians, we can conclude that:

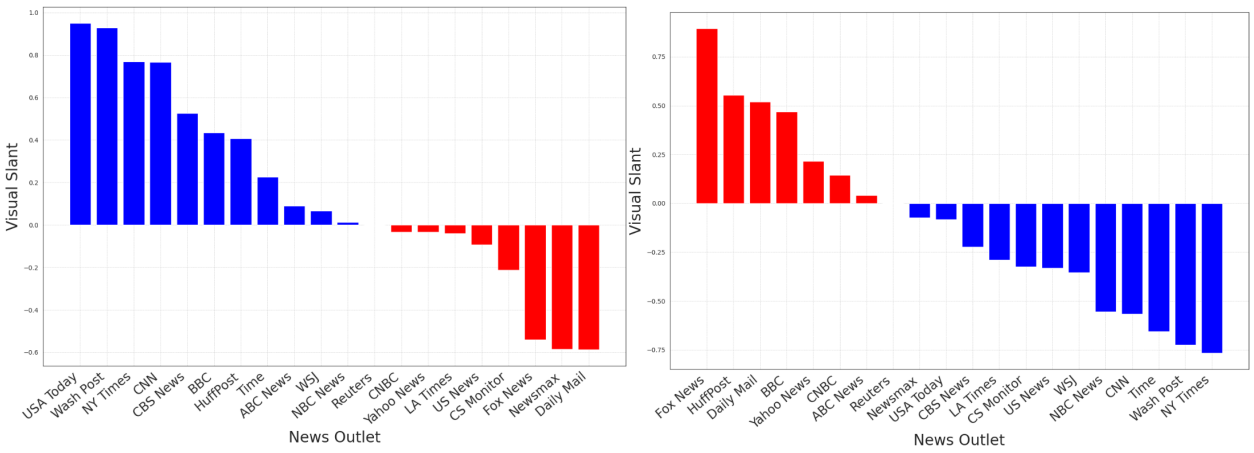
$$\bar{\tau}^T\left(y^k \in \mathscr{Y}_R \mid \mathscr{P}_D\right) < 0 < \bar{\tau}^T\left(y^k \in \mathscr{Y}_D \mid \mathscr{P}_D\right) \tag{21}$$

In Web Appendix §E.2, we employ two statistical tests to analyze these differences: the Kolmogorov-Smirnov (K-S) test and one-sample t-tests. Both confirm that there are significant differences between the distributions and means of Republican- and Democratic-leaning outlets for both Democratic and Republican politicians.

6.1.2 Conservative Visual Slant Across News Outlets

We now document the extent to which visual slant varies across outlets. Figure 10 shows our visual slant measure for each news outlet ($y^k \in \mathscr{Y}$) for Republican politicians as $\bar{\tau}^T(y^k \mid \mathscr{P}_R)$, and for Democratic politicians as $\bar{\tau}^T(y^k \mid \mathscr{P}_D)$, ranked in bar charts. In Figure 10a, we see that Democratic news outlets have a positive visual slant measurement for Democratic politicians and a negative visual slant for Republican

politicians. Conversely, in Figure 10b, we see that Republican news outlets exhibit a positive visual slant measurement for Republican politicians and a negative visual slant measurement for Democratic politicians.



(a) Democratic politicians.

(b) Republican politicians.

Figure 10: Overall visual slant measures for Democratic and Republican politicians by news outlet.

Further, outlets such as *Fox News* and *Daily Mail* display a more positive visual slant measurement for Republican politicians, indicating a more favorable portrayal of these politicians. Simultaneously, these outlets show a more negative visual slant measure of Democratic politicians, displaying a less favorable portrayal of these politicians. On the other hand, outlets like *CNN*, *Washington Post*, and *The New York Times* exhibit a more positive visual slant measurement for Democratic politicians and a more negative visual slant for Republican politicians. Interestingly, some outlets, such as the *BBC* and *HuffPost*, demonstrate a positive visual slant for both Democratic and Republican politicians. This suggests that their editorial approach may balance portrayals of politicians from both parties, reflecting a more centrist ideological positioning rather than strong partisan alignment.

To establish a single outlet-specific visual slant measurement, we compute the difference between the visual slant measures for Republican and Democratic politicians within each outlet. A larger difference indicates a stronger conservative visual slant. Accordingly, we define our unified measure as the *Conservative Visual Slant (CVS)* as follows:

$$\widehat{\text{CVS}}(y^k) = \bar{\tau}^T\left(y^k \mid \mathscr{P}_R\right) - \bar{\tau}^T\left(y^k \mid \mathscr{P}_D\right), \quad \text{where } y^k \in \mathscr{Y}. \tag{22}$$

Intuitively, this metric captures the degree to which an outlet portrays Republican politicians more favorably compared to Democratic politicians. For example, if a Republican-leaning outlet scores high on this measure, it implies that it emphasizes showing Republican politicians with a smile while depicting Democratic politicians without a smile. Figure 11 illustrates the conservative visual slant scores for all outlets in our dataset, sorted in increasing order. The results reveal that outlets such as *Daily Mail* and *Fox News* demonstrate high conservative visual slant, strongly favoring Republican politicians. On the other hand, outlets like *Washington Post*, *NY Times*, and *CNN* exhibit a pronounced liberal slant, favoring Democratic politicians. Meanwhile, outlets such as *BBC*, *CS Monitor*, and *HuffPost* exhibit relatively neutral or low

levels of slant. Overall, there is significant heterogeneity in visual slant across outlets, providing insights into which outlets are the most and least polarized visually. We refer readers to Web Appendix §E.4 for a comprehensive outlet-level analysis, including additional statistical tests and robustness checks.

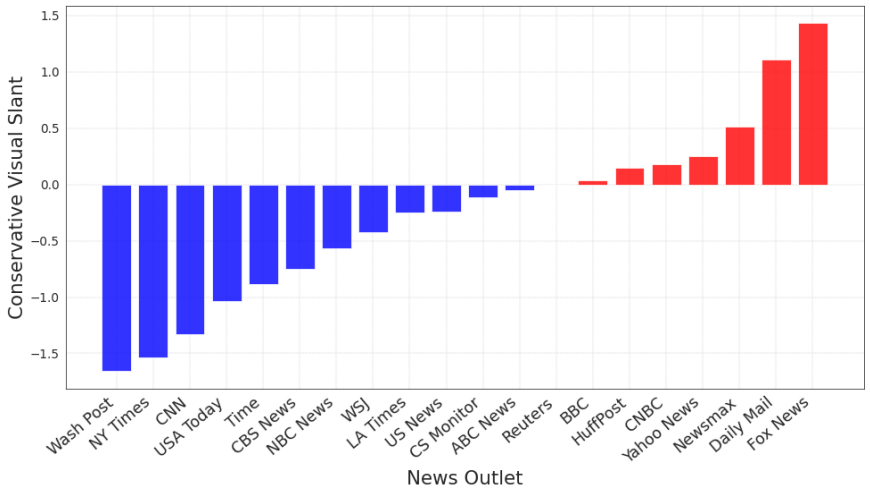


Figure 11: *Conservative Visual Slant (CVS)* of news outlets.

Section 7 examines the external validity of these outlet-level patterns and compares CVS with a feature-extraction benchmark.

6.2 Overall Visual Polarization Across Politicians

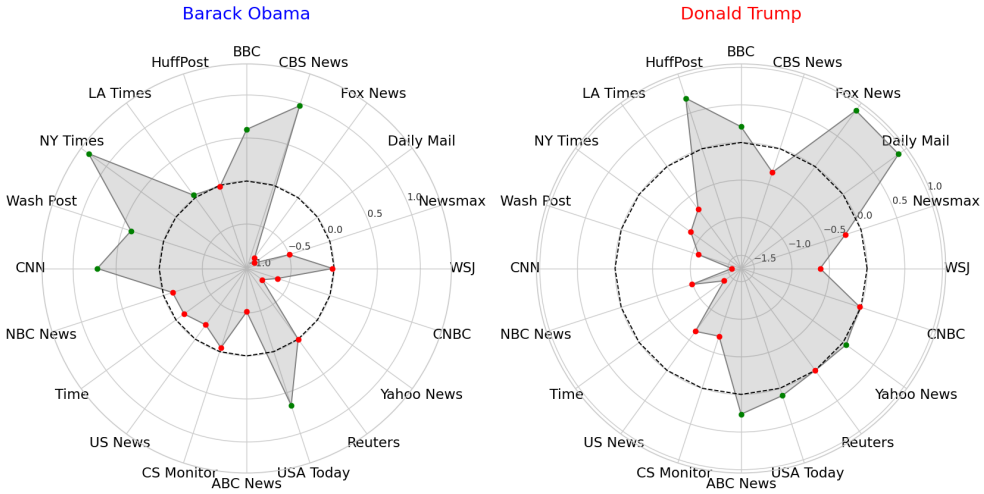


Figure 12: Radar plots of visual slant in each news outlet for Barack Obama and Donald Trump.

We now examine how different politicians are portrayed across news outlets and identify the politicians with the most polarizing depictions in the media. Figure 12 presents the visual slant measures for two prominent figures: Barack Obama and Donald Trump. Consistent with our earlier findings, we observe

a clear divide in visual slant scores – Obama receives more favorable portrayals in Democratic-leaning outlets, while Trump is depicted more positively in Republican-leaning outlets. We extend this analysis to other politicians, with detailed results provided in Web Appendix §E.8.

Next, we aim to document visual polarization across politicians. The core intuition is that visual polarization appears as greater dispersion in how the same politician is visually portrayed across news outlets. At the article level, this dispersion can be naturally summarized by the standard deviation of outlet-specific visual slant. We therefore define Overall Visual Polarization (OVP) by first measuring, for each article about a politician, the cross-outlet standard deviation in visual slant, and then averaging this article-level dispersion across all articles covering that politician.

Although a standard-deviation–based measure is intuitive, implementing it empirically raises challenges. As Gentzkow et al. (2019) document, in high-dimensional settings, a naive plug-in estimator of dispersion suffers from finite-sample bias because the same estimated slant mapping enters both the construction of outlet-specific slant scores and the aggregation of their dispersion, creating mechanical dependence across components. To address this concern, we adopt their cross-fitting approach. We partition the training data into two disjoint folds and estimate two independent slant mappings, $\hat{\tau}^{T,(1)}(\cdot)$ and $\hat{\tau}^{T,(2)}(\cdot)$. We then use these mappings to evaluate dispersion on the test data. Rather than computing a conventional within-article standard deviation, we construct dispersion using the cross-product of deviations from the two independently estimated mappings. This approach eliminates mechanical bias while preserving the interpretation of dispersion as a standard deviation. Web Appendix §E.3 reports the predictive performance of the two estimated mappings. Equation (23) formalizes this construction by defining OVP for politician p as the article-level average of the square root of the nonnegative cross-outlet covariance in visual slant:

$$\text{OVP}(p) = \frac{1}{|\mathscr{A}_p|} \sum_{i \in \mathscr{A}_p} \sqrt{\max\left\{ \text{Cov}_{y \in \mathscr{Y}} \left(\hat{\tau}_i^{T,(1)}(y), \hat{\tau}_i^{T,(2)}(y) \right), 0 \right\}}. \tag{23}$$

where $\mathscr{A}_p$ denotes the set of articles featuring politician p, and the covariance is taken across outlets using predictions from the two independently estimated slant mappings. The truncation at zero ensures that the measure is well-defined in finite samples.

Figure 13 ranks the politicians from most polarized to least polarized based on Equation 23. We see that Donald Trump, Barack Obama, Bernie Sanders, and Joe Biden are the top politicians with the most visually polarizing portrayal across all outlets. Given that all four were presidents or presidential candidates for significant chunks of our observation period and at the forefront of multiple polarizing discussions and events, this is understandable. Further, several prominent Republican politicians, such as Marco Rubio, also show high levels of polarization, likely attributed to their roles in policy debates and media prominence during key political events. On the Democratic side, figures such as Hillary Clinton, display significant polarization, reflecting their leadership roles and ideological positions within the party. Further, on the right end of Figure 13, we observe politicians with lower *OVP* scores. Notably, Liz Cheney, Joe Manchin, and Susan Collins rank among the least polarized, aligning with their reputations for bipartisanship and moderation (Killough, 2018). These findings have important implications for political strategy, particularly in how politicians shape their election campaigns when deciding whether to mobilize their base or appeal to

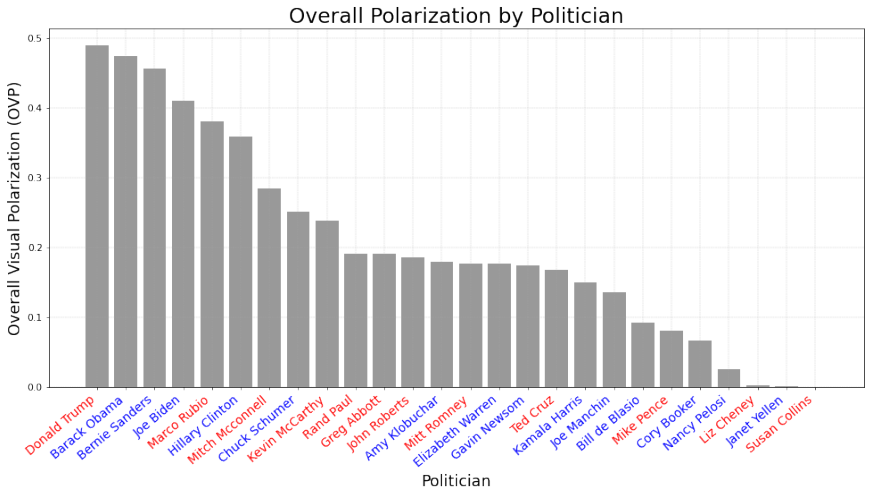


Figure 13: *Overall Visual Polarization (OVP)* of politicians, ranked from most to least polarized.

a broader, centrist electorate.

Finally, we assess whether our politician-specific *OVP* measure aligns with an external, politician-level indicator of polarization. Because we do not observe an established politician-level measure of polarization, we follow Bonica (2014) and use campaign-finance-based ideology scores (CFscores) as a proxy for a politician's underlying polarization. CFscores place politicians on a common left–right scale based on who contributes to their campaigns, and therefore cover not only members of Congress but also presidents, governors, and other officials who have raised campaign funds. The intuition is that politicians whose ideological positions lie farther from the center should be more polarizing; if *OVP* captures meaningful polarization in visual portrayals, these politicians should also exhibit more polarized depictions across news outlets. We therefore average each politician's CFscore over the 2012–2022 election cycles, take its absolute value as a measure of ideological extremity, and then examine its relationship with *OVP*. Figure 14 presents the results. We observe a positive and statistically significant correlation between the two measures, providing additional support for the interpretability of *OVP*.

This finding supports the idea that politicians with more ideologically extreme positions tend to receive more polarized visual portrayals across news outlets. For example, figures such as Donald Trump and Bernie Sanders have relatively extreme CFscores and also display high *OVP* values, consistent with more polarized visual representations in the media. In contrast, politicians such as Joe Manchin and Susan Collins, whose CFscores are closer to the center, exhibit substantially lower *OVP* values, reflecting more uniform visual portrayals across outlets. Overall, the positive relationship between ideological extremity and *OVP* suggests that visual polarization aligns with ideological polarization as revealed by campaign contributions.

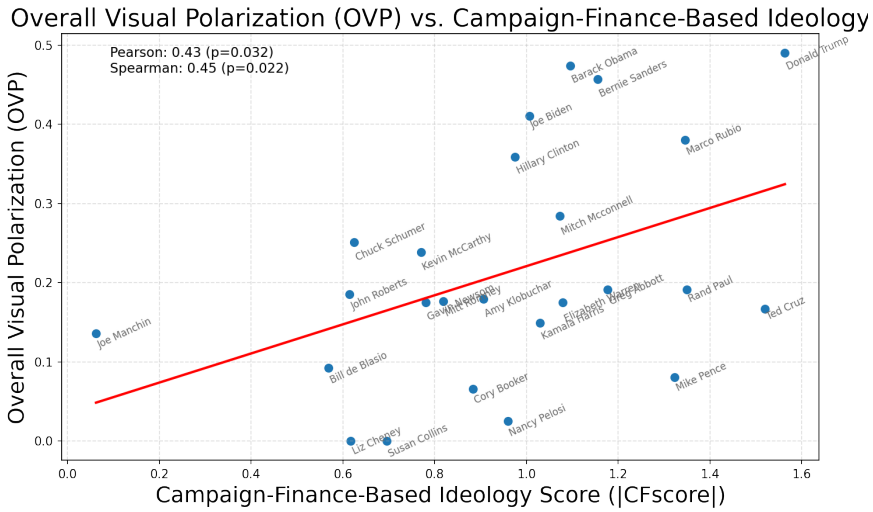


Figure 14: Relationship between a politician's *Overall Visual Polarization (OVP)* and their ideological extremity, measured by the absolute DIME CFscore.

6.3 Slant and Polarization Over Time

Now we turn to the time–series behavior of visual slant in the news media. For each calendar year $t \in \mathscr{T}$ and party $g \in \{R, D\}$, we compute an average visual slant

$$\bar{\tau}^T_{t,g} = \frac{1}{|I_{t,g}|} \sum_{(p,y)\in I_{t,g}} \bar{\tau}^T(y \mid p,t)\,,$$

where $I_{t,g}$ is the set of politician–outlet pairs (p,y) in year t with politician $p \in \mathscr{P}_g$ and outlet $y \in \mathscr{Y}$. Figure 15 plots $\bar{\tau}^T_{t,R}$ for Republican politicians (solid red), the same quantity excluding Donald Trump (dashed red), and $\bar{\tau}^T_{t,D}$ for Democratic politicians (blue).

Figure 15 traces the evolution of average visual slant over time for Democratic and Republican politicians. Two descriptive patterns emerge. First, visual slant for both parties remains close to zero from the early 2011 through 2014, indicating little systematic partisan differentiation in visual portrayal during this period. Second, beginning around 2015, the series diverge sharply across parties. Average visual slant for Democratic politicians increases steadily, reflecting a higher relative prevalence of visually favorable portrayals, such as smiling images. In contrast, average visual slant for Republican politicians declines, indicating a relative shift toward less favorable visual presentation. In the period between 2018 to 2020, the gap becomes pronounced: Democratic visual slant reaches approximately 0.18–0.20, whereas Republican visual slant falls to around -0.17 at its lowest point.

The timing of this divergence coincides with the emergence of Donald Trump as a presidential candidate in 2015, which motivates a separate analysis of Republican visual slant excluding his coverage. The dashed series in Figure 15 reports Republican slant computed without Trump. Excluding Trump substantially attenuates the decline in Republican visual slant. In the absence of Trump, Republican slant remains close to zero or slightly positive rather than strongly negative. However, the gap between Democratic and Repub-

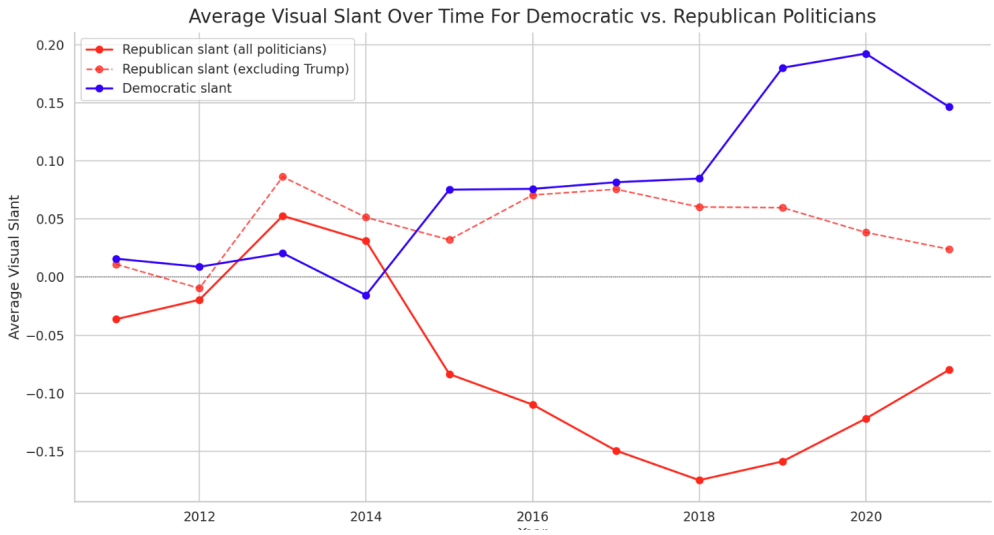


Figure 15: Average visual slant over time for Democratic vs. Republican politicians. Positive values indicate more positive visual treatment relative to Reuters; negative values indicate more negative visual treatment.

lican slant persists, indicating that Trump's coverage amplifies the divergence but does not solely generate the emerging partisan separation in visual portrayal.

Taken together, these time-series patterns document a clear increase in visual polarization in news media over time. Beginning in the mid-2015, outlets exhibit a sustained divergence in the visual portrayal of Democratic and Republican politicians. The emergence of Donald Trump coincides with a sharp widening of this gap and contributes to its magnitude. However, the persistence of partisan separation after excluding his coverage indicates that the rise in visual polarization reflects a broader evolution in partisan visual framing rather than differences driven by a single political figure.

7 Validation and Comparison with Alternative Measures

We assess the external validity of our outlet-level measure and compare it with a conventional feature-extraction approach. We first explain the benchmark and its relationship to the counterfactual measure, then compare both measures with independent indicators of outlet ideology. These indicators measure related dimensions of media slant and provide external benchmarks rather than ground truth for visual polarization.

7.1 Feature-Extraction Benchmark

The conventional approach to studying visual content follows a two-step procedure: extract a feature from each photograph using a pretrained model, then relate the extracted feature to outlet identity and article characteristics through a regression (Davenport, 2017; Boxell, 2021; Peng, 2018). The first step produces

$$\hat{T}_i = f_1(Z_i). \tag{24}$$

For a binary feature such as smiling, a natural second step is a logistic regression. Consider photographs of

a fixed politician p published by two outlets y_1 and y_2, and let $D_i = \mathbf{1}\{Y_i = y_1\}$. The specification is

$$\log \frac{\Pr\left(\hat{T}_i = 1 \mid D_i, X_i, P_i = p\right)}{\Pr\left(\hat{T}_i = 0 \mid D_i, X_i, P_i = p\right)} = \alpha_0 + \alpha_1' X_i + \beta D_i. \tag{25}$$

The coefficient β describes the difference between outlets in the log odds that a published image contains the extracted feature, conditional on the included article characteristics. It provides a reduced-form comparison with our structural measure, but its interpretation as a difference in preferences requires further conditions. Web Appendix §B.5.1 develops the two-step specification in detail.

An important distinction arises even before considering errors in feature extraction. The regression uses the images outlets actually publish. As Equation (15) shows, associations between image content and outlet identity reflect photograph availability and the attractiveness of the available collections as well as editorial valuations. Our counterfactual comparison removes these additional terms under the conditions in Section 4.1; a comparison of extracted features across published images does not automatically do so. Beyond this distinction, two sources of error complicate the interpretation of the regression coefficient.

The first is *extraction error*. The first step produces an estimated feature $\hat{T}_i$ rather than the true feature T_i. A model trained to detect facial expressions in natural photographs may also learn characteristics correlated with those expressions, so its predictions can reflect both the focal attribute and other visual content (Wei and Malik, 2022). Define the extraction error as $e_i^{\text{ext}} = \hat{T}_i - T_i$. If this error is systematically associated with outlet identity after accounting for the included controls, the second-step comparison can reflect differences in extraction errors as well as differences in expressions. Random measurement error and systematic extraction error have different implications; the relevant distinction is whether the error is related to the regressors, not simply whether the feature is estimated.

The second is *omitted-variable bias*, which can arise even when the feature is measured perfectly. Write a generic second-step relationship for the true feature as

$$T_i = f_2(Y_i, X_i) + \varepsilon_{0i} + \varepsilon_{\text{ov},i}, \tag{26}$$

where ε_{0i} is the remaining mean-zero disturbance conditional on the included regressors, and $\varepsilon_{\text{ov},i}$ represents the contribution of omitted visual characteristics that can be related to outlet identity and the controls. The flag-background example in Section 1 illustrates the problem: an observed difference in smiling photographs may reflect differences in the settings outlets select, rather than a stronger preference for smiling when other image attributes are held fixed. Adding extracted controls can mitigate this problem, but any finite representation may leave relevant visual dimensions outside the comparison.

Combining extraction error and omitted visual content gives the schematic decomposition

$$\hat{T}_i = f_2(Y_i, X_i) + \varepsilon_{0i} + \varepsilon_{\text{ov},i} + e_i^{\text{ext}}. \tag{27}$$

This decomposition makes the two sources distinct: $\varepsilon_{\text{ov},i}$ can affect the second step even with true feature labels, whereas e_i^{ext} arises from replacing those labels with model predictions. Equations (26)–(27) describe the generic regression problem; the logistic specification in Equation (25) is an example of the fitted model

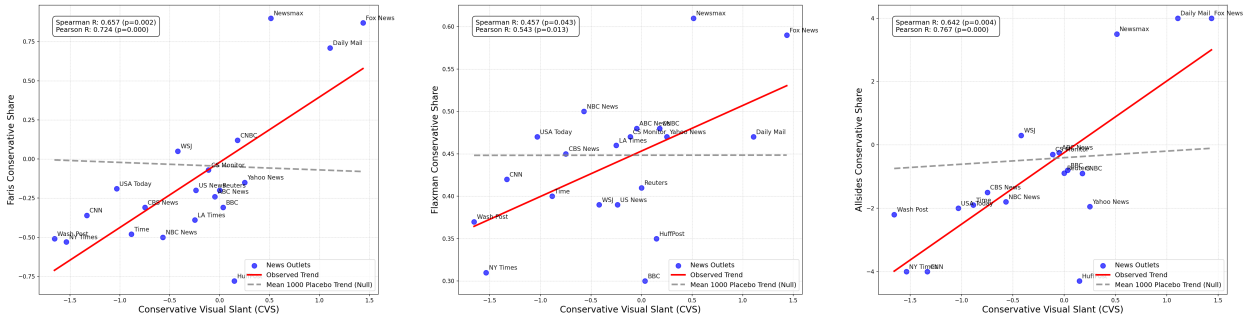
used for a binary extracted feature. Web Appendix §B.5.2 gives the detailed decomposition, and §B.5.3 and §B.5.4 provide the formal extraction-bias and omitted-variable analyses.

Our approach uses a different measurement strategy. We construct image pairs designed to differ only in the focal feature and evaluate them using a model that retains the remaining visual content. This holds the other image attributes approximately fixed in the comparison and avoids relying on a separate extracted-feature score. The interpretation still depends on the identification conditions, the quality of the image transformation, and prediction performance at the evaluation points. The empirical exercise below examines whether these methodological differences lead to different relationships with independent ideology benchmarks.

For that exercise, we construct an outlet-level benchmark following Boxell (2021) on the same photographs used for our analysis. This benchmark compares the expected relative favorability of published images of Republican and Democratic politicians within each outlet. Relative favorability is measured as happiness minus the sum of negative emotions inferred from facial-expression analysis. It is a feature-extraction benchmark distinct from the binary-smile regression used above to illustrate the general measurement issues. Web Appendix §C.2 gives its construction and additional specifications.

7.2 Comparison with External Benchmarks

We compare CVS for our 20 outlets with three measures constructed from independent, non-visual sources: the conservative sharing score of Faris et al. (2017), the browsing-based measure of Flaxman et al. (2016), and the ideology ratings of AllSides (2024). These sources are not used to train our prediction model. For example, Faris et al. (2017) measure the proportion of an outlet's stories shared on Twitter by users who predominantly retweet conservative-leaning sources.



(a) *CVS* vs. Conservative Share (Faris et al., 2017)

(b) *CVS* vs. News Diet Slant (Flaxman et al., 2016)

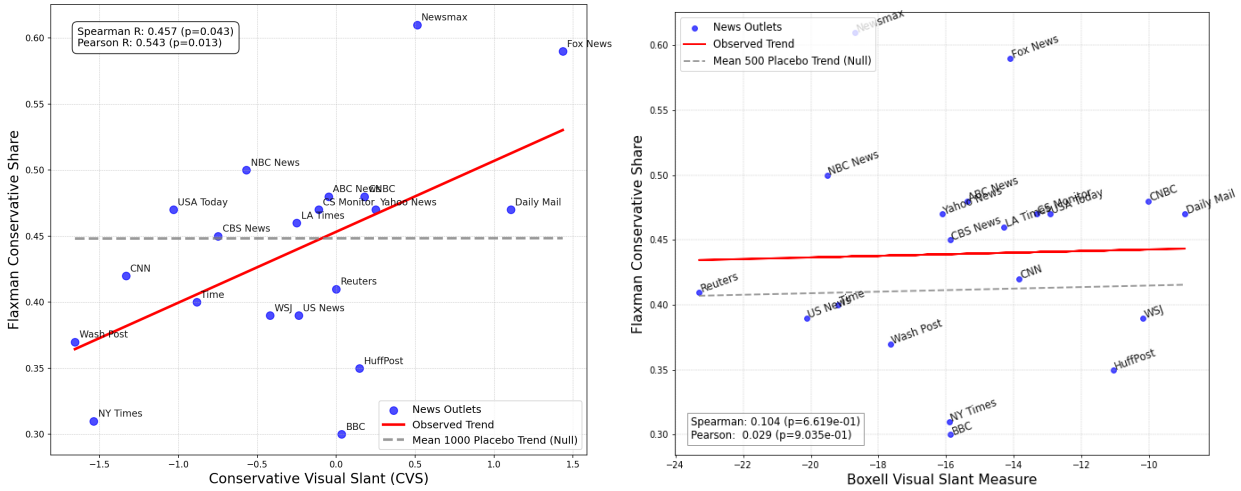
(c) *CVS* vs. AllSides Ideology Ratings (AllSides, 2024)

Figure 16: Correlation between *Conservative Visual Slant (CVS)* and external, non-visual measures of media slant constructed from independent data sources.

Figure 16 shows significant positive correlations of 0.72, 0.54, and 0.76 between CVS and the three benchmarks, respectively. The agreement across independent sources supports interpreting the outlet-level pattern as partisan visual slant. To assess whether this alignment depends on politicians' actual party affiliations, we randomly reassign party labels 1,000 times and recompute CVS. The correlations collapse toward zero under these placebo assignments, indicating that the external alignment depends on the partisan

grouping used in the measure.

We next compare CVS and the feature-extraction benchmark against the same external ideology measure. Figure 17 uses the conservative share score of Flaxman et al. (2016); we assess alignment using Pearson and Spearman correlations. The visual measures are constructed from the same image sample, so differences in their external alignment reflect differences in the measurement procedures applied to those photographs.



(a) *Conservative Visual Slant (CVS)* from our estimation procedure

(b) Visual Slant Measure from Boxell (2021).

Figure 17: Comparison of *CVS* and visual slant measure by Boxell (2021) against Conservative Share Score from Flaxman et al. (2016).

CVS has a strong, statistically significant association with the Flaxman et al. benchmark, while the feature-extraction measure has a weak, statistically insignificant association. Web Appendix §E.6 reports the comparisons with the other two external benchmarks. These findings support the external validity of CVS and show that the choice of measurement strategy matters for its alignment with independent measures of ideology. They do not establish that the external scores are the true values of visual slant or identify the contribution of each potential source of error in the feature-extraction approach.

Web Appendix §E.7 also applies the counterfactual framework to brightness and frowning and reports their relationships with the same external benchmarks.

8 Conclusion

News outlets choose what readers see as deliberately as they choose what readers read, and this paper develops a measure of the resulting divergence. The obstacle to measuring it is that photographs are high-dimensional objects whose attributes are chosen jointly, so that comparing the images outlets publish compares bundles of attributes rather than any single one of them. We therefore define visual slant as the response to a counterfactual change in one visual attribute, holding the remainder of the image fixed by construction. We show that this object is identified from observed coverage under three conditions: an overlap condition on the photographs at which the measure is evaluated, a condition requiring that the two versions of a photo-

graph be equally available to each outlet, and a timing condition under which an outlet's coverage decision precedes its choice of photograph. We also characterize the behavior of the estimator that recovers it, and show that the natural plug-in estimator of our dispersion measure is biased while the cross-fitted estimator is not.

Applying the measure to 63,188 photographs spanning a decade yields three findings. First, visual slant has a systematic partisan structure, with the response to a smile reversing sign with outlet ideology and reversing again with the politician's party. Second, our outlet-level measure agrees closely with slant measures constructed from sharing behavior, browsing panels, and editorial ratings, none of which uses image data. This agreement suggests that visual and verbal slant express a single editorial position rather than two independent ones. Third, visual divergence is a recent development rather than a constant feature of the period we study. It is negligible through 2014 and widens sharply thereafter, and Donald Trump's coverage amplifies the change without generating it.

Our comparison with feature-extraction measures carries a lesson that extends beyond this application. Applied work increasingly derives variables from unstructured data using pretrained models and then treats those variables as data. Our results indicate that this step is not innocuous. The same photographs, analyzed by extraction and by counterfactual comparison, order outlets differently, and only one of the two orderings accords with independent evidence about these outlets. Where the object of interest is the contribution of a specific attribute, constructing the counterfactual is an alternative to the extraction approach rather than a refinement of it.

Three limitations mark the boundaries of what we have shown and suggest directions for further work. First, our estimates concern the supply side. They establish that outlets differ in the valuation they place on visual attributes, but they do not establish whether this reflects editors' own ideological preferences or their reading of audience demand. Gentzkow and Shapiro (2010) draw this distinction for text using variation in audience composition that our data do not contain. Second, our measure is specific to the attribute we manipulate. We estimate the response to a smile and report results for brightness and for a frown in the Web Appendix, but a photograph can flatter or diminish along dimensions that no single manipulation captures, and extending the framework to several attributes jointly is a natural next step as generative models improve. Third, we measure slant without measuring its consequences. Caprini (2023) shows experimentally that partisan lead images move readers' opinions with text held fixed, and linking outlet-level visual slant to political behavior, as DellaVigna and Kaplan (2007) and Martin and Yurukoglu (2017) do for cable news, would establish whether the divergence we document has the effects that the literature on textual slant has found.

Finally, our framework is not specific to politics. Any setting in which agents select images from a common pool presents the same measurement problem, including product photographs, job market profiles, and advertising creative, and the same approach applies. One constructs the counterfactual, conditions flexibly on the remainder of the image, and reads valuations off the resulting change in choice probabilities.

Competing Interests Declaration

Author(s) have no competing interests to declare.

References

Jason Abaluck and Abi Adams-Prassl. What do consumers consider before they choose? identification from asymmetric demand responses. *The Quarterly Journal of Economics*, 136(3):1611–1663, 2021.

AllSides. Media bias rating methods, 2024. URL https://www.allsides.com/media-bias/media-bias-rating-methods.

Elliott Ash, Ruben Durante, Maria Grebenschikova, and Carlo Schwarz. Visual representation and stereotypes in news media. 2021.

Susan Athey, Dean Karlan, Emil Palikot, and Yuan Yuan. Smiles in profiles: Improving fairness and efficiency using estimates of user preferences in online marketplaces. Technical report, National Bureau of Economic Research, 2022.

Levon Barseghyan, Maura Coughlin, Francesca Molinari, and Joshua C. Teitelbaum. Heterogeneous choice sets and preferences. *Econometrica*, 89(5):2015–2048, 2021.

Adam Bonica. Mapping the ideological marketplace. *American Journal of Political Science*, 58(2):367–386, 2014.

Levi Boxell. Slanted images: Measuring nonverbal media bias during the 2016 election. *Available at SSRN 3837521*, 2021.

Giulia Caprini. Visual bias. 2023.

Christopher T. Conlon and Julie Holland Mortimer. Demand estimation under incomplete product availability. *American Economic Journal: Microeconomics*, 5(4):1–30, 2013.

Thomas H Davenport. How analytics has changed in the last 10 years (and how it's stayed the same). *Harvard Business Review*, 28(08):2017, 2017.

Stefano DellaVigna and Ethan Kaplan. The fox news effect: Media bias and voting. *The Quarterly Journal of Economics*, 122(3):1187–1234, 2007.

Daria Dzyabura, Siham El Kihal, John R Hauser, and Marat Ibragimov. Leveraging the power of images in managing product return rates. *Marketing Science*, 42(6):1125–1142, 2023.

Robert Faris, Hal Roberts, Bruce Etling, Nikki Bourassa, Ethan Zuckerman, and Yochai Benkler. Partisanship, propaganda, and disinformation: Online media and the 2016 us presidential election. *Berkman Klein Center Research Publication*, 6, 2017.

Seth Flaxman, Sharad Goel, and Justin M Rao. Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly*, 80(S1):298–320, 2016.

Matthew Gentzkow and Jesse M Shapiro. Media bias and reputation. *Journal of political Economy*, 114(2): 280–316, 2006.

Matthew Gentzkow and Jesse M Shapiro. What drives media slant? evidence from us daily newspapers. *Econometrica*, 78(1):35–71, 2010.

Matthew Gentzkow, Jesse M Shapiro, and Matt Taddy. Measuring group differences in high-dimensional choices: method and application to congressional speech. *Econometrica*, 87(4):1307–1340, 2019.

Michelle Sovinsky Goeree. Limited information and advertising in the US personal computer industry. *Econometrica*, 76(5):1017–1074, 2008.

Tim Groseclose and Jeffrey Milyo. A measure of media bias. *The quarterly journal of economics*, 120(4): 1191–1237, 2005.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

Joandrea Hoegg and Michael V Lewis. The impact of candidate appearance and advertising strategies on election results. *Journal of Marketing Research*, 48(5):895–909, 2011.

Ashley Killough. Moderate senators feel boost as shutdown ends. *CNN*, 2018. URL https://www.cnn.com/2018/01/22/politics/moderate-senators-shutdown-end-bipartisan-group-senate/index.html.

Hui Li, Jian Ni, and Fangzhu Yang. Product design using generative adversarial network: Incorporating consumer preference and external data. *arXiv preprint arXiv:2405.15929*, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Jens Ludwig and Sendhil Mullainathan. Machine learning as a tool for hypothesis generation. *The Quarterly Journal of Economics*, 139(2):751–827, 2024.

Lan E Luo and Olivier Toubia. Using ai for controllable stimuli generation: An application to gender discrimination with faces. *Available at SSRN 4865798*, 2024.

Paola Manzini and Marco Mariotti. Stochastic choice and consideration sets. *Econometrica*, 82(3):1153–1176, 2014.

Gregory J. Martin and Ali Yurukoglu. Bias in cable news: Persuasion and polarization. *American Economic Review*, 107(9):2565–2599, 2017.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Sendhil Mullainathan and Andrei Shleifer. The market for news. *American Economic Review*, 95(4):1031–1053, 2005.

Omkar Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *BMVC 2015-Proceedings of the British Machine Vision Conference 2015*. British Machine Vision Association, 2015.

Yilang Peng. Same candidates, different faces: Uncovering media bias in visual portrayals of presidential candidates with computer vision. *Journal of Communication*, 68(5):920–941, 2018.

Andrea Prat and David Strömberg. The political economy of mass media. In Daron Acemoglu, Manuel Arellano, and Eddie Dekel, editors, *Advances in Economics and Econometrics: Tenth World Congress*, volume 2, pages 135–187. Cambridge University Press, 2013.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

SerpAPI. Serpapi - real-time search engine results api, 2023. URL https://serpapi.com/. Accessed on January, 2023.

Amandeep Singh and Bolong Zheng. Causal regressions for unstructured data. In *Causal Representation Learning Workshop at NeurIPS 2023*, 2023.

Michael Sülflow and Marcus Maurer. The power of smiling. how politicians' displays of happiness affect viewers' gaze behavior and political judgments. *Visual political communication*, pages 207–224, 2019.

Denis G Sullivan and Roger D Masters. " happy warriors": Leaders' facial displays, viewers' emotions, and political support. *American Journal of Political Science*, pages 345–368, 1988.

Thales Teixeira, Michel Wedel, and Rik Pieters. Emotion-induced engagement in internet video advertisements. *Journal of marketing research*, 49(2):144–159, 2012.

Alexander Todorov, Anesu N. Mandisodza, Amir Goren, and Crystal C. Hall. Inferences of competence from faces predict election outcomes. *Science*, 308(5728):1623–1626, 2005.

Claudia Townsend and Barbara E. Kahn. The visual preference heuristic: Effects of visual vs. verbal depiction on assessment of others, choice of hedonic products, and voting. *Journal of Consumer Research*, 41 (2):392–411, 2014.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Yanhao Wei and Nikhil Malik. Unstructured data, econometric models, and estimation bias. SSRN, 2022.

Sikun Xu, Dennis J. Zhang, Zhenling Jiang, and Zhengling Qi. Causal inference when controlling for unstructured data. *Working Paper*, 2024. Olin Business School, Washington University in St. Louis. Accessed: 2025-01-21.

Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016.

Web Appendix

A Details of Data Collection and Cleaning

A.1 Details of Politicians and News Outlets

Two important details about our data collection relate to the set of politicians and news outlets. Below we present these two lists:

- **Politicians:** To select the sample of 30 politicians, we first created a large set of individuals who have run for public offices from either Democratic or Republican party. We then sampled a set of top 30 politicians based on their search volume on Google Trends. As such, the higher the search volume for a politician, the more likely it is to select that politician for our main sample. The resulting sets of Democratic and Republican politicians is presented below:
  - *Democratic Politicians*: This list contains (1) Barack Obama, (2) Joe Biden, (3) Hilary Clinton, (4) Bernie Sanders, (5) Nancy Pelosi, (6) Kamala Harris, (7) Chuck Schumer, (8) Corey Booker, (9) Amy Klobuchar, (10) Elizabeth Warren, (11) Gavin Newsom, (12) Bill de Blasio, (13) James Clyburn, (14) Janet Yellen, (15) Joe Manchin, and (16) Steny Hoyer.
  - *Republican Politicians*: This list contains (1) Donald Trump, (2) Ted Cruz, (3) Marco Rubio, (4) Mitt Romney, (5) Mitch McConnell, (6) Greg Abbott, (7) Rand Paul, (8) Mike Pence, (9) Kevin McCarthy, (10) Susan Collins, (11) Liz Chenney, (12) John Roberts, (13) Hal Rogers, (14) Andy Biggs.
- **News Outlets:** We selected 20 news outlets following Flaxman et al. (2016): ABC News, BBC, CBS News, CNBC, CNN, The Christian Science Monitor, Daily Mail, Fox News, HuffPost, Los Angeles Times, NBC News, Newsmax, The New York Times, Reuters, Time, USA Today, U.S. News & World Report, The Wall Street Journal, The Washington Post, and Yahoo News.

A.2 Details of Data Cleaning

In this section, we present details about data cleaning. We first present the manual verification procedure we used to ensure that SerpAPI does not miss important data. To address the limitations of the automated method, we employ an additional manual process for quality control. This involves saving HTML files of search results locally and using Beautiful Soup to extract images and metadata. This method ensures the accuracy and completeness of our dataset.

We now present the two-phase data cleaning procedure in Figure A.1 in greater detail. The first phase, designed for accurate face detection, consists of two main steps. Initially, we perform pyramid rescaling, adjusting each image to three scales (1.1, 0.8, and 0.6) to enhance focus, similar to adjusting a camera lens (Lin et al., 2017). Following this, Mask R-CNN, which uses the ResNet50 architecture as its backbone, identifies and outlines humans within these images (He et al., 2017), referring to an object detection and instance segmentation task. Next, we use multitask cascaded Convolutional Neural Networks (MTCNN) for face detection within the regions identified by Mask R-CNN. The detected faces are subsequently processed by FaceNet for face recognition (Zhang et al., 2016). This dual approach not only increases the efficiency of our face detection process but also significantly improves its accuracy.

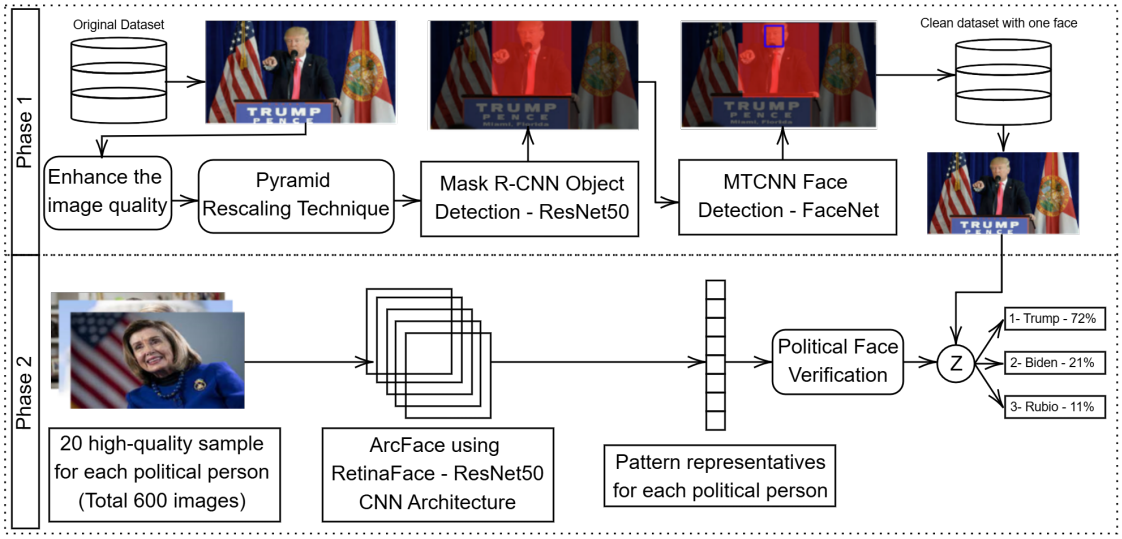


Figure A.1: Two-phase computer vision framework for political person verification.

The second phase is the face verification stage of our computer vision framework. We initiate the process by feeding ArcFace with twenty samples for each of the thirty politicians to create a unique facial pattern representative for each political person (Deng et al., 2019). Then, we take the images from the first phase, which are confirmed to have just one face, and run them through ArcFace. This system compares the detected face with our database of political figures and ranks the top three matches, providing a similarity percentage for each.

In total, we retain 63,188 images with exactly one face, that have been verified by our dual framework. These images represent a clean and reliable dataset for our study. The distribution of the final data is shown in Figure A.2. This dataset allows us to accurately analyze media representation of political figures, ensuring that each image is correctly attributed to the intended individual.

B Theoretical Analysis: Details and Proofs

This appendix develops the argument summarized in Section 4. §B.1 specifies the editor's selection problem when the collection of available photographs is unobserved, states the publication probability (13) used in the main text, and derives the exact random-consideration-set representation it approximates. §B.2 discusses the identifying assumptions stated in Section 4.1. §B.3 proves Proposition 1 and records what the argument does and does not require. §B.4 proves Propositions 2 and 3 and the corollary governing additive aggregates.

B.1 Image Selection with an Unobserved Menu

An editor does not choose between two photographs. She chooses from the photographs that have reached her desk: the frames her outlet's wire subscriptions delivered, the frames its own staff photographers filed, etc. We do not observe which photos are presented to each editor, how they vary across outlets, or how they vary across stories within an outlet. Treating this menu as fixed and common would attribute to editorial taste whatever differences arise from differences in holdings, and treating it as fixed but outlet-specific would still impose that two stories about the same politician in the same week were prepared from the same

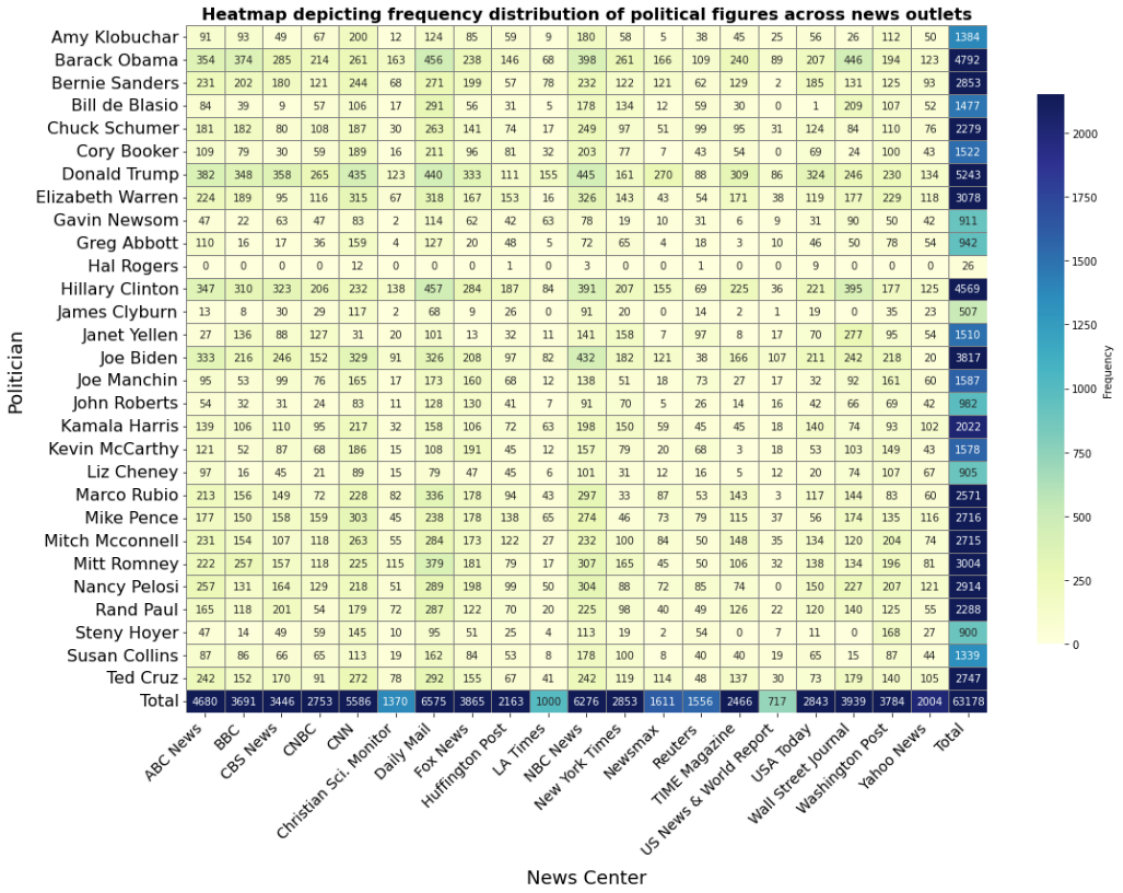


Figure A.2: Frequency distribution of politicians across outlets.

photographs, which differs from how the newsroom operates.

Let $\mathscr{Z}_p$ denote the space of photographs of p that could accompany an article about him, understood to contain not only the frames that were in fact transmitted to newsrooms but every frame a camera present at one of p's public appearances could have recorded.[6] The collection $\mathscr{C}_y(x_i, p) \subseteq \mathscr{Z}_p$ from which the editor at outlet y chooses the image for article i is random, and $\phi_y(z \mid x_i, p) = \Pr(z \in \mathscr{C}_y(x_i, p))$ gives the marginal probability that any given frame belongs to it. Availability is a property of the outlet's supply relationships rather than of its editorial preferences, and the two enter as distinct primitives: ϕ_y governs what reaches the editor's desk and u governs how she ranks what she finds there.

Specification (13) writes the publication probability as an availability-weighted conditional logit. Equivalently, it is the conditional logit an editor would generate if she faced all of $\mathscr{Z}_p$ with payoffs shifted by a supply term, $\tilde{u}_y(z) = u(z,x,p,y) + \log \phi_y(z \mid x, p)$. It is a proper distribution over $\mathscr{Z}_p$ by construction, since $\mathscr{I}_y(x,p) = \sum_{z'} \phi_y(z' \mid x, p) e^{u(z',x,p,y)}$ normalizes it, and $\mathscr{I}_y(x,p) = \mathbb{E}\left[\sum_{z' \in \mathscr{C}_y(x,p)} e^{u(z',x,p,y)}\right]$ is the expected inclusive value of the collection the outlet draws. This is the form used in the literature that estimates demand when the menu is unobserved: Goeree (2008) estimates demand for personal computers with consideration probabilities that depend on advertising exposure; Manzini and Mariotti (2014) axiomatize stochastic

[6] A digital photograph is a finite array, so $\mathscr{Z}_p$ is a finite set and the sums here are literal. The continuous representation over which we estimate is a device for smoothing across photographs in estimation, not a different population object.

consideration and characterize the choice behavior it generates; Abaluck and Adams-Prassl (2021) identify consideration separately from preferences in insurance choice; and Barseghyan et al. (2021) study identification when choice sets are heterogeneous and unobserved. Conlon and Mortimer (2013) make the closely related point that treating incomplete availability as though it were full availability biases estimated preferences. The problem here is of the same kind, and if anything more acute: the number of photographs of a prominent politician is very large, no outlet holds more than a small fraction of them, and the fraction each holds is a consequence of contracts and logistics that we observe only indirectly.

The exact publication probability implied by a random menu is not quite (13), and the following lemma shows what it is.

Write

$$\mathscr{J}_y \;=\; \sum_{z' \in \mathscr{C}_y(x,p)} e^{u(z',x,p,y)}$$

for the *realized* inclusive value of the collection, a random variable, as against the expected inclusive value $\mathscr{I}_y(x,p) = \mathbb{E}[\mathscr{J}_y]$ that appears in (13). That the two are related in this way,

$$\mathscr{I}_y(x,p) \;=\; \sum_{z' \in \mathscr{Z}_p} \phi_y(z' \mid x,p)\, e^{u(z',x,p,y)} \;=\; \mathbb{E}\left[\sum_{z' \in \mathscr{C}_y(x,p)} e^{u(z',x,p,y)}\right] \;=\; \mathbb{E}\big[\mathscr{J}_y\big],$$

is linearity of expectation, and requires no restriction on how availability is correlated across photographs.

Lemma 1 (Exact choice representation)**.** *Suppose $\mathscr{C}_y(x,p) \neq \emptyset$ with probability one and that the taste shocks $\{\xi_{izy}\}$ in (1) are independent of the realization of $\mathscr{C}_y(x,p)$. Then, for any z with $\phi_y(z \mid x,p) > 0$, the probability that outlet y publishes z is*

$$h^\star(z \mid y,x,p) = \phi_y(z \mid x,p) \cdot e^{u(z,x,p,y)} \cdot \Lambda_y(z \mid x,p)\,, \tag{A.28}$$

where

$$\Lambda_y(z \mid x,p) = \mathbb{E}\big[\,\mathscr{J}_y^{-1} \,\big|\, z \in \mathscr{C}_y(x,p)\big]\,. \tag{A.29}$$

The representation holds for an arbitrary joint distribution of availability across photographs.

Proof. Condition on the event $\{z \in \mathscr{C}_y(x,p)\}$, which has probability $\phi_y(z \mid x,p)$; on its complement the editor cannot publish z. On the event, the Type I Extreme Value assumption in (1) gives the conditional logit probability $e^{u(z,x,p,y)} / \mathscr{J}_y$, which is well defined because $\mathscr{J}_y$ is bounded below by $e^{u(z,x,p,y)} > 0$. Taking expectations over the distribution of $\mathscr{C}_y(x,p)$ conditional on that event and applying the law of total probability gives (A.28). Non-emptiness of $\mathscr{C}_y(x,p)$ is what makes $h^\star(\cdot \mid y,x,p)$ a proper distribution over $\mathscr{Z}_p$. □

B.2 Identifying Assumptions

This appendix first discusses the isolation requirement underlying the matched-image comparison, then explains Assumptions 1–3 in the order stated in Section 4.1.3.

Matched-image construction: isolation

The comparison in Section 3.2 holds all visual content other than the focal feature fixed. This isolation

requirement determines whether Equation (4) measures polarization with respect to T. If the transformation that produces Z_p^1 from Z_p^0 also changes brightness, framing, or background, the comparison captures preferences over a bundle of features and reintroduces the attribution problem discussed in Section 3.1. Section 5.2.1 examines preservation of selected nonfocal image characteristics.

The matched-image construction also determines the scope of the estimand. Definition 1 concerns these two photographs, and if outlets' valuation of the feature interacts with other visual content, a different pair would give a different number. Section 5 describes how we select the pairs and, as a guard against this sensitivity, averages the estimate over several pairs for each politician.

Assumption 1: symmetric availability

Assumption 1 of the main text requires that the two frames of a matched pair be equally likely to reach any given outlet. It places no restriction across outlets: some newsrooms are much better equipped than others, and the assumption allows ϕ_{y_1} and ϕ_{y_2} to differ arbitrarily in level. It permits smiling photographs of a politician to be more plentiful than neutral ones in the world at large, or less so, since it is stated for a matched pair from one occasion rather than for the two types of photograph in general.

The construction of the evaluation pair is what makes this credible. The two photographs depict the same politician in the same setting and the same pose, differ only in the focal expression, and stand for two frames of a single burst, taken seconds apart at one photographic occasion. A wire service that transmits an occasion transmits a selection of frames from it, a staff photographer who covers an event files from that event, and an outlet that holds one frame of a sequence therefore holds, or could have held, its neighbors. As §4.1 discusses, the conditions that determine whether a photographic occasion reaches a newsroom operate at the level of the occasion rather than at the level of the expression within it. The assumption would be considerably less comfortable if the two images came from different occasions, since one outlet might then hold a flattering portrait from a campaign event while another held only an unflattering frame from a hearing.

Assumption 2: overlap

The two parts restrict different objects and are used by different results. Part (i) is a statement about the model, and it is what Proposition 1 uses: every outlet under comparison must have a non-negligible chance of covering the story and of holding each evaluation photograph, so that the publication probabilities are strictly positive and the log odds in (15) are well defined. An evaluation photograph that a given outlet could never obtain is one for which that outlet's valuation of the feature is not behaviorally meaningful. Part (ii) is a data-support requirement for estimation. It is not used in the algebraic error-bound proof and does not establish first-stage consistency, which is separately assumed for the convergence conclusion of Proposition 2. It is deliberately not a requirement that the evaluation photographs appear in the data. They do not: Z_p^1 is constructed, and $\mathscr{Z}_p$ contains frames that were never distributed. What part (ii) requires is that the pair lie among photographs like those the model has seen, which is why we build it from an ordinary photograph of the politician drawn from outside the training sample rather than from an unusual or synthetic starting point: an image far from anything in the data would return a fitted value that is an extrapolation rather than an estimate.

Assumption 3: coverage precedes selection

Assumption 3 holds if an outlet decides whether to run a story before choosing its photograph, which is the ordinary sequence of newsroom production; it fails if a striking image induces coverage that would not otherwise have occurred. Note that it is what makes the coverage term in (15) a function of (x, p) alone, and hence what allows the second difference to remove it.

B.3 Identification

Proof of Proposition 1. Fix (x, p). The observations are outlet–article pairs: an article with characteristics (x, p) is covered by outlet y with probability $\pi(y \mid x, p)$, which by Assumption 3 does not depend on the photograph, and conditional on covering it, the outlet publishes photograph z with probability $h(z \mid y, x, p)$, a proper distribution by (13). Among outlet–article pairs with characteristics (x, p), the share attributable to outlet y publishing z is therefore $\pi(y \mid x, p)\, h(z \mid y, x, p)$ divided by $\sum_{y' \in \mathscr{Y}} \pi(y' \mid x, p)$, and conditioning further on $Z = z$ gives (14), the common factor $\sum_{y'} \pi(y' \mid x, p)$ having cancelled. Note that this requires no assumption that π sum to one across outlets: several outlets cover the same story, and the normalization is absorbed in the denominator.

The denominator of (14) does not depend on y, so it cancels in any ratio across outlets, and substituting (13) yields (15). Assumption 2(i) bounds $\phi_y(Z_p^t \mid x_i, p)$ and $\pi(y \mid x_i, p)$ away from zero, and $\mathscr{I}_y(x, p) \in (0, \infty)$, so h and hence g are strictly positive at both evaluation points and every logarithm in (15) exists.

Of the three nuisance terms in (15), the inclusive-value term $\log[\mathscr{I}_{y_1}(x, p)/\mathscr{I}_{y_2}(x, p)]$ and the coverage term $\log[\pi(y_1 \mid x, p)/\pi(y_2 \mid x, p)]$ are functions of (y_1, y_2, x, p) alone and do not depend on z. Evaluating (15) at $z = Z_p^1$ and at $z = Z_p^0$ and differencing therefore annihilates both, leaving

$$\tilde{\rho}_i^T(y_1, y_2) = \underbrace{\Big[u(Z_p^1, \cdot, y_1) - u(Z_p^1, \cdot, y_2)\Big] - \Big[u(Z_p^0, \cdot, y_1) - u(Z_p^0, \cdot, y_2)\Big]}_{= \rho_i^T(y_1, y_2) \text{ by (12)}} + \log \frac{\phi_{y_1}(Z_p^1 \mid x_i, p)\, \phi_{y_2}(Z_p^0 \mid x_i, p)}{\phi_{y_1}(Z_p^0 \mid x_i, p)\, \phi_{y_2}(Z_p^1 \mid x_i, p)}, \tag{A.30}$$

Assumption 1 sets $\phi_y(Z_p^1 \mid x_i, p) = \phi_y(Z_p^0 \mid x_i, p)$ for each y, so the second term is zero. Identification follows because g is a functional of the joint distribution of the observed data and $\tilde{\rho}_i^T$ is a known functional of g evaluated at two points at which it is well defined. □

Remark 1 (What the second difference removes)**.** *The proof turns on a single structural feature of the estimand: it differences log odds across two photographs, and so annihilates any term in* (15) *that is constant in z. Three unobserved quantities are removed in this way. The first is the relative propensity of the two outlets to cover the story, which would otherwise confound editorial taste with differences in news agendas. The second is the inclusive value of each outlet's collection, the object that makes an unobserved menu problematic in the first place: a newsroom with a rich archive publishes any particular frame less often than one with a thin archive, holding preferences fixed, and that difference is common to the two versions. The third is any component of an outlet's valuation shared by the two photographs, including the newsworthiness of the event and the technical quality of the image. None of these quantities is identified, and the argument*

does not require that any of them be. What the second difference does not remove is anything that varies with z, which is why availability requires Assumption 1 rather than being implied. That assumption is what the matched-pair construction buys: the pair is chosen so that the one object that fails to cancel is equal.

Remark 2 (Identification requires only balance across outlets)**.** *Equation* (A.30) *shows that availability enters as a ratio of ratios. It vanishes under the weaker condition*

$$\frac{\phi_{y_1}\left(Z_p^1 \mid x_i, p\right)}{\phi_{y_1}\left(Z_p^0 \mid x_i, p\right)} = \frac{\phi_{y_2}\left(Z_p^1 \mid x_i, p\right)}{\phi_{y_2}\left(Z_p^0 \mid x_i, p\right)}, \tag{A.31}$$

which permits smiling frames to be systematically easier or harder to obtain than neutral ones, provided the asymmetry is the same for the outlets being compared. In particular, (A.31) *holds for every pair of outlets whenever log availability is additively separable in outlet and version,* $\log \phi_y(Z_p^t \mid x_i, p) = \alpha_y + \beta_t$*. Proposition 1 goes through verbatim with Assumption 1 replaced by* (A.31)*. We state the assumption in the stronger form because the matched-pair construction justifies it at that level, and because the failure that would concern a reader, an outlet with privileged access to flattering frames of particular politicians, violates the weak and the strong form alike.*

B.4 Statistical Properties

Let

$$\hat{\rho}_i^T(y_1, y_2) = \log \frac{\hat{g}\left(y_1 \mid Z_p^1, x_i, p\right)}{\hat{g}\left(y_2 \mid Z_p^1, x_i, p\right)} - \log \frac{\hat{g}\left(y_1 \mid Z_p^0, x_i, p\right)}{\hat{g}\left(y_2 \mid Z_p^0, x_i, p\right)}, \tag{A.32}$$

let $\mathscr{S}_i = \left\{\left(Z_p^t, x_i, p\right) : t \in \{0,1\}\right\}$ denote the points at which $\hat{g}$ is evaluated, and let

$$\left\|\hat{g} - g\right\|_{\infty,i} \equiv \max_{\mathscr{S}_i} \max_{y \in \mathscr{Y}} \left|\hat{g}(y \mid \cdot) - g(y \mid \cdot)\right|.$$

Note that (A.32) is a linear contrast in $\log \hat{g}$, with coefficients $(+1,-1,-1,+1)$ on the four evaluated log-probabilities. This structure delivers the two results below.

Assumption 4 in Section 4.2 states the bounded-probability condition used below. Assumption 2(i) delivers the qualitative content of this condition for the population object, since it makes g strictly positive at the evaluation points. Assumption 4 strengthens it to a bound uniform across outlets and articles and extends it to the estimate, both of which the arguments below require.

Proof of Proposition 2. By Proposition 1, $\rho_i^T = \tilde{\rho}_i^T$, and $\tilde{\rho}_i^T$ and $\hat{\rho}_i^T$ are the same contrast in $\log g$ and $\log \hat{g}$ respectively. By Assumption 4 and the mean value theorem applied to $\log(\cdot)$ on $[\varepsilon, 1]$, along which the derivative is at most ε^{-1}, we have $|\log \hat{s} - \log s| \leq \varepsilon^{-1}|\hat{s} - s|$ for any $s, \hat{s} \in [\varepsilon, 1]$. The two expressions differ in four logarithms, each entering with coefficient of modulus one, so the triangle inequality gives the bound. If $\|\hat{g} - g\|_{\infty,i} \xrightarrow{p} 0$, the bound also implies $\hat{\rho}_i^T \xrightarrow{p} \rho_i^T$. □

We do not characterize the rate at which $\|\hat{g} - g\|_{\infty,i}$ vanishes, since no such result is available for the class of models we estimate. What the bound establishes is that first-stage error is transmitted linearly, with constant $4/\varepsilon$.

Remark 3 (Partial cancellation of first-stage bias)**.** *Because the logarithm is nonlinear, an unbiased $\hat{g}$ does not deliver an unbiased* $\log \hat{g}$*. A second-order expansion around g gives*

$$\mathbb{E}\big[\log \hat{g}(y \mid z)\big] \;\approx\; \log g(y \mid z) + \frac{b(y,z)}{g(y \mid z)} - \frac{m(y,z)}{2\,g(y \mid z)^2},$$

where $b(y,z) = \mathbb{E}[\hat{g} - g]$ is the bias of the first stage at the evaluation point and $m(y,z) = \mathbb{E}[(\hat{g} - g)^2] = \mathrm{Var}(\hat{g}) + b(y,z)^2$ is its mean squared error. These corrections enter (A.32) with signs $(+,-,-,+)$. The two photographs are close in image space and differ only in the focal feature, so b and m are of similar magnitude at the two points for a given outlet, and the corresponding corrections largely offset. The leading distortion is therefore governed by the difference *of the first-stage errors across the two versions rather than by their level. The second-difference structure that removes the unobserved menu in Proposition 1 thus also attenuates the first-stage bias.*

Additive aggregates

Corollary 1 (Additive measures)**.** *Let $M = \sum_{k \in \mathscr{K}} w_k \rho_k$ be a linear combination of article-level parameters over a finite index set $\mathscr{K}$, and let $\hat{M} = \sum_{k \in \mathscr{K}} w_k \hat{\rho}_k$. Then*

$$\big|\hat{M} - M\big| \;\le\; \Big(\textstyle\sum_{k \in \mathscr{K}} |w_k|\Big) \cdot \max_{k \in \mathscr{K}} \big|\hat{\rho}_k - \rho_k\big|. \tag{A.33}$$

Proof. Immediate from the triangle inequality. □

Let $\mathscr{A}_R$ and $\mathscr{A}_D$ denote the evaluation-sample articles about Republican and Democratic politicians. The sample analogue of $\mathrm{CVS}(y)$ is of the form covered by Corollary 1, with weights $1/|\mathscr{A}_R|$ on articles about Republicans and $-1/|\mathscr{A}_D|$ on articles about Democrats, summing in absolute value to two. Two qualifications attach. First, because $|\mathscr{K}|$ grows with the evaluation sample, transferring consistency from the article level to the aggregate through the bound of Corollary 1 requires the article-level errors to vanish uniformly, that is $\sup_i \|\hat{g} - g\|_{\infty,i} \xrightarrow{p} 0$, which is stronger than the pointwise premise of Proposition 2; convergence of the average of the article-level errors would suffice for the conclusion, but is not what the max-based bound delivers. Second, replacing the population expectations in (9) by sample averages introduces sampling error. Assumption 4 makes the article-level parameters bounded, and hence of finite variance: each of the two log odds ratios in (16) lies in $[-\log(1/\varepsilon), \log(1/\varepsilon)]$, so $|\tilde{\rho}_i^T| \le 2\log(1/\varepsilon)$, and the same bound holds for $|\rho_i^T|$ by Proposition 1. Sampling error then vanishes at rate $O_p(n^{-1/2})$ provided the summands are independent across units. Articles about the same politician are not independent of one another, since they share the evaluation pair and the politician's photographic record, so we take the politician as the unit of independence and n as the number of politicians. Because the weights in (9) are assigned per article rather than per politician, the rate additionally requires that no politician account for a dominant share of the articles in $\mathscr{A}_R$ or $\mathscr{A}_D$, that is $\max_p n_p/|\mathscr{A}_G| = O(n^{-1})$.

Independence of the evaluation articles from the training sample is a separate matter, and it is what the train–test split delivers. Absent it, $\hat{g}$ would have been fitted in part to the idiosyncratic features of the articles used to evaluate it, and the resulting log odds would overstate systematic differences between outlets, most severely for politicians with the fewest articles.

Dispersion aggregates

Let $\mathscr{A}_p$ denote the evaluation-sample articles about politician p, with $n_p = |\mathscr{A}_p|$, and write

$$\hat{\tau}_i^T(y) = \tau_i^T(y) + e_i(y), \tag{A.34}$$

so that $e_i(y)$ is the first-stage error carried into the article-level slant estimate. Let Var_y and Cov_y denote finite-population moments across $\mathscr{Y}$, and let

$$\widehat{\mathrm{OVP}}(p) = \frac{1}{n_p} \sum_{i \in \mathscr{A}_p} \sqrt{\max\left\{ \mathrm{Cov}_y\left(\hat{\tau}_i^{T,(1)}(y),\, \hat{\tau}_i^{T,(2)}(y) \right), 0 \right\}} \tag{A.35}$$

be the cross-fitted estimator of (10), where $\hat{g}^{(1)}$ and $\hat{g}^{(2)}$ are classifiers fitted on disjoint folds of the training data, $\hat{\tau}_i^{T,(k)}(y)$ is the article-level slant estimate computed from $\hat{g}^{(k)}$, and $e_i^{(k)}(y)$ is its error. Replacing the covariance by $\mathrm{Var}_y(\hat{\tau}_i^T(y))$ gives the plug-in estimator.

Proof of Proposition 3. Fix an article i. All expectations are conditional on the evaluation sample, so $\tau_i^T(\cdot)$ and $\mathscr{Y}$ are fixed. Write $\mu(y) = \mathbb{E}[e_i(y)]$ and, for the cross-fitted case, $\mu^{(k)}(y) = \mathbb{E}\left[e_i^{(k)}(y)\right]$ for $k \in \{1,2\}$; by hypothesis each is constant across outlets, with values μ and $\mu^{(k)}$ respectively.

For the plug-in estimator, $\mathrm{Var}_y(\tau_i^T + e_i) = \mathrm{Var}_y(\tau_i^T) + \mathrm{Var}_y(e_i) + 2\,\mathrm{Cov}_y(\tau_i^T, e_i)$. Taking expectations and using linearity of Cov_y in its second argument, $\mathbb{E}[\mathrm{Cov}_y(\tau_i^T, e_i)] = \mathrm{Cov}_y(\tau_i^T(y), \mu(y)) = 0$ because $\mu(y)$ is constant across outlets. The stated identity follows, and the inequality since $\mathrm{Var}_y(e_i) \geq 0$, with equality only if $e_i(y)$ is constant across outlets almost surely.

For the cross-fitted estimator, bilinearity of the finite-population covariance gives

$$\mathrm{Cov}_y\left(\tau_i^T + e_i^{(1)},\, \tau_i^T + e_i^{(2)}\right) = \mathrm{Var}_y(\tau_i^T) + \mathrm{Cov}_y\left(\tau_i^T, e_i^{(2)}\right) + \mathrm{Cov}_y\left(e_i^{(1)}, \tau_i^T\right) + \mathrm{Cov}_y\left(e_i^{(1)}, e_i^{(2)}\right).$$

The two middle terms have expectation zero by the argument just given. For the final term, write

$$\mathrm{Cov}_y(a,b) \;=\; |\mathscr{Y}|^{-1} \sum_{y \in \mathscr{Y}} a(y)\, b(y) \;-\; \bar{a}\,\bar{b}, \qquad \bar{a} = |\mathscr{Y}|^{-1} \sum_{y \in \mathscr{Y}} a(y).$$

Independence of the two error vectors gives $\mathbb{E}\left[e_i^{(1)}(y)\, e_i^{(2)}(y')\right] = \mu^{(1)}\mu^{(2)}$ for every pair (y, y'), so the first term has expectation $|\mathscr{Y}|^{-1} \sum_y \mu^{(1)}\mu^{(2)} = \mu^{(1)}\mu^{(2)}$ and the second has expectation $|\mathscr{Y}|^{-2} \sum_y \sum_{y'} \mu^{(1)}\mu^{(2)} = \mu^{(1)}\mu^{(2)}$. The difference is zero. Dependence of $e_i^{(k)}(y)$ across outlets within a fold is left unrestricted. □

Four comments on the result. First, the hypothesis is weaker than requiring the first stage to be unbiased. A mean error shared by all outlets shifts every slant estimate for politician p equally and leaves the dispersion across outlets untouched; what a dispersion measure is sensitive to is the part of the error specific to an outlet. This is also why Remark 3 does not substitute for Proposition 3: that remark concerns cancellation across the two *versions* within an outlet, and it does not deliver constancy across outlets.

Second, the bias does not vanish as coverage of a politician grows. Since $\mathrm{OVP}(p)$ averages article-level dispersions, averaging over $\mathscr{A}_p$ averages the bias term $\mathbb{E}[\mathrm{Var}_y(e_i)]$ rather than removing it, and that term does

not shrink with n_p: every article about p is scored by the same $\hat{g}$, so whatever that model gets systematically wrong about an outlet recurs in each article's dispersion. The bias therefore shrinks only as the first stage improves, and it is largest for politicians with the thinnest coverage, whose slant estimates are least precise. That is the wrong direction for a measure meant to rank politicians by how polarizing they are.

Third, it is important to distinguish what cross-fitting does and does not deliver. Fitting the two classifiers on disjoint folds of the training data makes the *estimation* error of one statistically unrelated to that of the other, which is the independence the second part of Proposition 3 requires, and it is the component of e_i that inflates measured dispersion in the first part. It does not make the two errors independent in every respect: both classifiers are evaluated at the same generated pair (Z_p^1, Z_p^0) and on the same evaluation articles, so any error that originates in the pair itself, such as an artifact of the generative transformation, is common to both folds and survives the cross-product. Cross-fitting is therefore a remedy for model-specific noise, not for a flawed counterfactual, which is why we validate the pair construction separately in §5.2.1.

Fourth, the cross-fitted estimator we report replaces each article's variance by the covariance of the two fold-specific slant vectors and takes the square root. Unbiasedness holds for the quantity inside the radical; the square root is concave, so the reported measure is downward biased by Jensen's inequality, and the finite-sample truncation of negative covariances at zero works in the opposite direction. Each fold-specific classifier is also fitted on half the training data, so it is less accurate than one fitted on all of it. Web Appendix §E.3 reports the predictive performance of the two fold-specific models.

B.5 Standard Reduced-Form Approach and its Limitations

In this part, we discuss the standard reduced-form approach used to measure polarization in visual content. In §B.5.1, we describe the two-step reduced-form approach and connect the estimated parameter under this approach to the measure of polarization in visual content as defined in Definition 1. Next, in §B.5.2, we provide a theoretical and conceptual discussion of the drawbacks of this approach.

B.5.1 Two-step Model

Recall that our goal is to measure *visual polarization*, $\rho^T(p, y_1, y_2)$, with respect to a focal feature T, such as whether the politician in the image is smiling or not. However, this is inherently challenging when dealing with unstructured image data, because, unlike the standard causal inference literature, where treatment is observed (Imbens and Rubin, 2015), here we do not directly observe T_i, i.e., the presence or absence of the treatment (smile in this case) in a given article i. As a result, a common approach in social science settings is to use a *Two-step Approach*. This approach addresses the high dimensionality and complexity of unstructured data by first extracting meaningful features and then using these features in a structured econometric model. See Davenport (2017) for a general discussion of this approach in the broader social sciences literature, and Boxell (2021) and Peng (2018) for applications of this approach to the context of image polarization in media.[7] Methodologically, the two-step approach can be outlined as follows:

[7]Beyond the political polarization context, this two-step approach is now commonly used in marketing research involving images in other settings as well. For instance, unstructured data such as video and audio streams (Z) are analyzed to extract features like facial expressions (T), which are then used to study their impact on user engagement (Y) (Lu et al., 2021). In the online labor market, profile pictures (Z) are examined to identify features such as perceived race or attire (T) to assess job offer likelihood (Y), highlighting visual biases in employment opportunities (Troncoso and Luo, 2022). Additionally, research on social media posts about e-cigarettes (Z) extracts demographic features (T) to study tax policy compliance (Y), revealing demographic responses to

1. *Feature extraction using a machine learning model:* The first step involves using an off-the-shelf machine learning model, denoted as f_1, to extract relevant feature(s) T from the unstructured data Z. This model $f_1 : Z \to T$ is used to derive the feature(s) T:

$$\hat{T} = f_1(Z) \tag{A.36}$$

2. *Econometric analysis:* After extracting features $\hat{T}$, the second step involves analyzing these features within a structured econometric model. There are two potential ways to relate $\hat{T}$ and Y:

$$Y = f_2^{\text{T-independent}}(\hat{T}, \boldsymbol{X}) + \varepsilon, \quad \text{or,} \tag{A.37a}$$
$$\hat{T} = f_2^{\text{T-dependent}}(Y, \boldsymbol{X}) + \varepsilon, \tag{A.37b}$$

where $f_2^{\text{T-independent}}$ is the second-stage econometric model where extracted feature $\hat{T}$ is used as an independent variable, and $f_2^{\text{T-dependent}}$ is the second-stage model where $\hat{T}$ is used as dependent variable. Social scientists typically prefer the second form because $\hat{T}$ is an estimated variable that could include measurement error, and therefore using it as the independent variable is not appropriate (Bollen and Davis, 2009).

We now describe how this approach can be used to quantify polarization in visual content with respect to feature T, defined as $\rho^T(p, y_1, y_2)$ in Definition 1. Consider a simple example where news outlets from Democratic-leaning outlets like the *The New York Times*, *CNN*, and *BBC* (y_1) and Republican-leaning outlets like *Fox News*, *Newsmax*, and *Daily Mail* (y_2) choose images for a politician, say Hillary Clinton (p). In this context, the first step involves estimating the binary indicator $\hat{T}$, which denotes whether the focal politician (Hillary Clinton) is smiling in the image ($\hat{T} = 1$) or not ($\hat{T} = 0$). This can be obtained from standard off-the-shelf emotion recognition models such as Face++ (Face++, 2023), Google Vision (Vision, 2023)[8], or through in-house model training where we first label a sample of the images using human subjects and then train a model to predict the label given the image.

For the second step, we can use a logistic regression where we regress $\hat{T}_i$ on characteristics X_i and Y_i. For notational simplicity, consider the case where Y is a binary variable (e.g., $Y = 1$ for Democratic-leaning outlets, and $Y = 0$ for Republican-leaning outlets). We can write the log odds ratio for the resulting logistic regression as follows:

$$\log\left(\frac{\Pr(\hat{T} = 1 \mid Y, \boldsymbol{X})}{\Pr(\hat{T} = 0 \mid Y, \boldsymbol{X})}\right) = \alpha_0 + \alpha_1 \boldsymbol{X} + \beta Y \tag{A.38}$$

If the model in Equation (A.38) is well specified, the log-odds coefficient β captures how outlet identity shifts the likelihood that an image exhibits the visual feature, conditional on article characteristics. This object provides a reduced-form approximation to our visual polarization measure in Definition 1, which is defined as the change in relative outlet log odds induced by an isolated change in visual content. In the

legislative changes (Anand and Kadiyali, 2024).

[8] Azure (2023)'s Face API, a popular model previously used in studies such as Boxell (2021); Caprini (2023), has been unavailable for use since June 2022 due to updated responsible AI policies (Microsoft, 2022).

binary-outlet case, collapsing counterfactual outlet likelihoods into observed variation in $\hat{T}$ yields

$$\begin{aligned}\rho^T(p, Y=1, Y=0) &= (\alpha_0 + \alpha_1 \boldsymbol{X} + \beta) - (\alpha_0 + \alpha_1 \boldsymbol{X}) \\ &= \beta.\end{aligned} \tag{A.39}$$

Thus, the coefficient β in the logistic regression model serves as an empirical estimate of the polarization measure $\rho^T(p, Y=1, Y=0)$. If $\beta > 0$, it indicates that Democratic-leaning outlets ($Y=1$) assign a higher probability to using a smiling image of Hillary Clinton than Republican-leaning outlets, conditional on article characteristics, implying a larger change in relative outlet likelihoods induced by the visual feature for Democratic-leaning outlets. Conversely, if $\beta < 0$, Republican-leaning outlets ($Y=0$) exhibit a larger shift in relative likelihoods than Democratic-leaning outlets. Therefore, $\rho^T(p, Y=1, Y=0)$ can be approximated by β, providing a quantitative measure of ideological polarization across news outlets.

B.5.2 Drawbacks of the Two-Step Approach

While the two-step model described in §B.5.1 is easy to interpret and apply, it relies on the assumption that the regression model is well-specified. However, this assumption can fail due to two natural reasons and lead to incorrect inference: (1) *extraction bias*, and (2) *omitted variable bias*. In this section, we discuss these biases and explain how they can manifest in our setting.

Extraction bias occurs due to imperfections in the feature extraction process, or the first step of the two-step process (Wei and Malik, 2022). That is, the true label T may differ from the extracted label $\hat{T} = f_1(Z)$ as follows:

$$T = f_1(Z) + \varepsilon_1 = \hat{T} + \varepsilon_1 \tag{A.40}$$

The error in Equation (A.40) can be viewed as a measurement error. It is well-known that systematic dependence of this error on other relevant factors in the model can introduce biases or inconsistencies in the estimates; see Chapter 4 of Wooldridge (2010). As such, we decompose this error into two parts, such that $\varepsilon_1 = \varepsilon_r + \varepsilon_e$, where ε_r is the random i.i.d. noise in the measurement process, and ε_e is the extraction error that can be correlated with the other relevant image features contained in $Z^{(-T)}$. For instance, a machine learning model designed to detect the facial expressions (e.g., whether a person is smiling) may also inadvertently capture brightness levels in the image, as brighter images are often associated with happier expressions. We can rewrite the relationship between T and $\hat{T}$ as:

$$T = \hat{T} + \varepsilon_r + \varepsilon_e. \tag{A.41}$$

Then, the predicted feature $\hat{T}$ is a combination of the true feature of interest T (e.g., smile) and a correlated nuisance feature ε_e (e.g., brightness). Thus, when we perform the second stage estimation, the estimated effect (β) captures the effect of the outlet on the biased measurement $\hat{T}$ rather than the true feature T.

Next, *omitted variable bias* arises from relevant factors omitted from the second step. This issue can arise even if we have the true labels T. To illustrate this point, consider the second-step relationship between the true label and covariates:

$$T = f_2(Y, X) + \varepsilon_2, \tag{A.42}$$

where ε_2 consists of unobserved variables that are not captured by the observed covariates Y and X through the semi-parametric function f_2. If this error term is independent of the covariates, we can identify function f_2 correctly. However, this is often a very strong assumption given the amount of information in $Z^{(-T)}$ that is omitted from the model. For example, consider a collection of images by Donald Trump in articles by *CNN* and *Fox News*, where the smile labels are accurate, i.e., $\hat{T} = T$. Now, consider an image feature, such as the presence of the US flag. It is likely that the US flag is present more often in images from *Fox News*, given their higher share of nationalist viewers. At the same time, it is more likely for a politician to smile in formal settings with flags present in the background. In this scenario, estimating the parameters of Equation (A.42) will lead to function f_2 picking up the association between the presence of the flag and the smile in the image because the presence of the flag is omitted from the model. Given the high-dimensional and unstructured nature of the images, numerous other image features can result in omitted variable bias in our estimates. To characterize the endogenous part of the error term ε_2, we rewrite Equation (A.42) as follows:

$$T = f_2(Y,X) + \varepsilon_0 + \varepsilon_{ov}, \tag{A.43}$$

where ε_0 is the part that is independent of covariates, and ε_{ov} is the part that is correlated with Y or X, which can lead to omitted variable bias. It is worth noting that image features that lead to extraction bias can also lead to omitted variable bias. For example, omitting image features like brightness can lead to omitted variable bias if brightness is correlated with both the actual presence of a smile in the image (not the extracted one) and the news outlet.

Together, if we want to estimate the second step of the two-step model in §B.5.1 by using $\hat{T}$ as the outcome, the errors in Equation (A.41) also appear in the second step equation as follows:

$$\hat{T} = f_2(Y,X) + \varepsilon_0 + \varepsilon_{ov} + \varepsilon_r + \varepsilon_e, \tag{A.44}$$

where ε_{ov} leads to omitted variable bias, ε_e leads to extraction bias due to the use of a machine learning model in the first step, and ε_r contributes to higher uncertainty in model estimates. In summary, the information loss resulting from extracting a single feature from an image can introduce both extraction bias and omitted variable bias. We develop a formal characterization of these biases and their sources in Web Appendix §B.5.3 and B.5.4.

B.5.3 Theoretical Analysis of Extraction Bias for the Two-step Model

As discussed in §B.5.2, *extraction bias* arises from the use of the predicted variable $\hat{T}$ in the analysis because the true variable T is not observable. This issue is primarily related to the machine learning model f_1, where transfer learning is employed to predict the feature of interest. We provide the theoretical econometric framework demonstrating how extraction bias can lead to biased estimation of β.

Recall that T is not directly observed; rather, it is extracted from the high-dimensional unstructured image data Z using a first-stage machine learning model f_1. This leads to two types of errors in the estimate $\hat{T}$: Measurement Error (ε_r) and Extraction Error (ε_e).

Remark 4. *Measurement Error (ε_r) arises from inaccuracies in predictions or measurements that are not*

systematically correlated with the true value of T (Hansen, 2022). This can be formally defined as:

$$T = \hat{T} + \varepsilon_r \quad where \quad Cov(T, \varepsilon_r) = 0$$

Remark 5. *Extraction Error (ε_e) occurs when the estimation method includes additional, indirectly related features (Wei and Malik, 2022). This error implies that the error term is correlated with the true value T:*

$$T = \hat{T} + \varepsilon_e \quad where \quad Cov(T, \varepsilon_e) \neq 0$$

Measurement error can be viewed as random noise added to the estimate $\hat{T}$, resulting in a variable that is not systematically biased. In contrast, extraction errors occur when the machine learning model, trained to predict the specific feature of interest T, inadvertently captures additional non-focal features that are correlated with T. For instance, a machine learning model designed to detect the facial expression of faces whether a person is happy may also inadvertently capture brightness levels in the image, as brighter images are often associated with happier expressions. This can introduce an extraction error ε_e into the model's predictions, potentially biasing the results.

In our setting, suppose that images used by *CNN* are generally brighter. In this case, the model may also capture brightness rather than true happiness. As a result, the predicted variable $\hat{T}$ reflects not only the intended feature (happiness) but also the unintended influence of brightness. This issue complicates the analysis by conflating the natural effect of emotional expression with confounding factors, making it difficult to separate their impacts. To assess this challenge theoretically, we can divide it into *Well-specified* and *Partially-specified* models.

Consider scenarios where T represents the actual value of the feature. Achieving this oracle estimate can be done by using human judgment to classify photos as Happy ($T = 1$) or Unhappy ($T = 0$), though this method is expensive. Now, assume a linear model where Y is a binary variable indicating the news outlet (1 for *CNN*, 0 for *Fox news*). In this case, the parameter of interest, β, can be estimated without bias because of the following Proposition.

Proposition 4. ***Well-specified:*** *Let (Y, X) denote observed regressors, where Y is the variable of interest and $X \in \mathbb{R}^k$ captures other explanatory variables. Consider the model:*

$$\begin{aligned} &\hat{T}_{oracle} = T + \varepsilon_r, \quad where \quad \hat{T}_{oracle} = Human, \\ &T = \alpha + \beta Y + \gamma^\top X + e, \end{aligned} \tag{A.45}$$

where $\mathbb{E}[e \mid Y, X] = 0$ and $\mathbb{E}[\varepsilon_r \mid Y, X] = 0$. Suppose $\hat{\beta}$ is the OLS estimator of the parameter of interest (coefficient of Y). Then $\hat{\beta}$ is an unbiased estimator of β. As $n \to \infty$:

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}\left(0, \Sigma_\beta = \mathbf{e}_2^\top \left(A_0^{-1} \Sigma A_0^{-1}\right) \mathbf{e}_2\right),$$

where $\mathbf{Z} = (1, Y, X_1, \ldots, X_k)^\top$, $A_0 = \mathbb{E}[\mathbf{Z}\mathbf{Z}^\top]$, $\Sigma = \mathbb{E}[\mathbf{Z}\theta\theta\mathbf{Z}^\top]$, $\theta = e + \varepsilon_r$, and $\mathbf{e}_2 = (0, 1, 0, \ldots, 0)^\top$.

Proof. In scenarios where T represents an oracle estimation of the population, it is possible to leverage

human judgment to categorize photos into Happy ($T = 1$) or Unhappy ($T = 0$). Assume a linear model where Y is a binary variable representing the news outlet (1 for *Fox News*, 0 for *CNN*).

Let $\hat{T}_{oracle} = \alpha + \beta Y + \gamma^\top X + \theta$, where $\theta = e + \varepsilon_r$. Define $\mathbf{Z} = (1, Y, X_1, \ldots, X_k)^\top$ and stack the observations:

$$\hat{\mathbf{T}}_{oracle} = \mathbf{Z}\boldsymbol{\beta}_{\text{true}} + \boldsymbol{\theta}, \quad \boldsymbol{\beta}_{\text{true}} = (\alpha, \beta, \gamma_1, \ldots, \gamma_k)^\top.$$

Here, the OLS estimator is:

$$\hat{\boldsymbol{\beta}} = (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \hat{\mathbf{T}}_{oracle} = \boldsymbol{\beta}_{\text{true}} + (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \boldsymbol{\theta}.$$

Since $\mathbb{E}[e \mid Y, X] = \mathbb{E}[\varepsilon_r \mid Y, X] = 0$, we have $\mathbb{E}[\boldsymbol{\theta} \mid \mathbf{Z}] = \mathbf{0}$. Therefore:

$$\mathbb{E}[\hat{\boldsymbol{\beta}} \mid \mathbf{Z}] = \boldsymbol{\beta}_{\text{true}} + (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbb{E}[\boldsymbol{\theta} \mid \mathbf{Z}] = \boldsymbol{\beta}_{\text{true}},$$

so $\hat{\boldsymbol{\beta}}$ is unbiased. In particular, $\hat{\beta}$ is an unbiased estimator of β. By the Law of Large Numbers (LLN) and Central Limit Theorem (CLT):

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{\text{true}}) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, A_0^{-1} \Sigma A_0^{-1}\right),$$

where:

$$A_0 = \mathbb{E}[\mathbf{Z}\mathbf{Z}^\top], \quad \Sigma = \mathbb{E}[\mathbf{Z}\theta\theta\mathbf{Z}^\top].$$

For $\hat{\beta}$, focusing on the second component:

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}\left(0, \mathbf{e}_2^\top A_0^{-1} \Sigma A_0^{-1} \mathbf{e}_2\right),$$

where $\mathbf{e}_2 = (0, 1, 0, \ldots, 0)^\top$. This confirms that in a well-specified model, the estimator $\hat{\beta}$ remains unbiased, and only the variance is affected by the inclusion of the error terms. □

The Proposition above confirms the reliability of using such models in well-specified scenarios.

Proposition 5. ***Partial-specified:*** *Let (Y, X) again denote observed regressors, where Y is the variable of interest and $X \in \mathbb{R}^k$ captures other explanatory variables, ensuring no omitted variable bias. Consider the model:*

$$\begin{aligned} &\hat{T}_{ml} = T + \varepsilon_r + \varepsilon_e, \quad \textit{where} \quad \hat{T}_{ml} = f_1(Z), \\ &T = \alpha + \beta Y + \gamma^\top X + e, \end{aligned} \tag{A.46}$$

where $\mathbb{E}[e \mid Y, X] = \mathbb{E}[\varepsilon_r \mid Y, X] = 0$, and ε_e is an extraction error correlated with outlet identity after partialling out X: $\mathbb{E}[\tilde{Y}\varepsilon_e] \neq 0$, where $\tilde{Y}$ is the residual from the population projection of Y on $(1, X)$. Suppose $\hat{\beta}$ is the OLS estimator of the parameter of interest (coefficient of Y). Then $\hat{\beta}$ is inconsistent for β. As $n \to \infty$:

$$\operatorname{plim} \hat{\beta} = \beta + \mathbf{e}_2^\top A_0^{-1} \mathbb{E}[\mathbf{Z}\varepsilon_e],$$

and, under standard regularity conditions,

$$\sqrt{n}\big(\hat{\beta} - \operatorname{plim}\hat{\beta}\big) \xrightarrow{d} \mathcal{N}\big(0, \Sigma_{\beta}^{*}\big),$$

where $\mathbf{Z} = (1, Y, X_1, \ldots, X_k)^{\top}$, $A_0 = \mathbb{E}[\mathbf{Z}\mathbf{Z}^{\top}]$, $\mathbf{e}_2 = (0,1,0,\ldots,0)^{\top}$, *and* Σ_{β}^{*} *is the sandwich variance computed from residuals around* $\operatorname{plim}\hat{\beta}$.

Proof. In real-world scenarios, $\hat{T}_{ml}$ often includes extraction errors, leading to biases.

Let $\hat{T}_{ml} = \alpha + \beta Y + \gamma^{\top} X + \theta$, where $\theta = e + \varepsilon_r + \varepsilon_e$. Define $\mathbf{Z} = (1, Y, X_1, \ldots, X_k)^{\top}$ and stack the observations:

$$\hat{\mathbf{T}}_{ml} = \mathbf{Z}\beta_{\text{true}} + \boldsymbol{\theta}, \quad \beta_{\text{true}} = (\alpha, \beta, \gamma_1, \ldots, \gamma_k)^{\top}.$$

Here, the OLS estimator is:

$$\hat{\boldsymbol{\beta}} = (\mathbf{Z}^{\top}\mathbf{Z})^{-1}\mathbf{Z}^{\top}\hat{\mathbf{T}}_{ml} = \beta_{\text{true}} + (\mathbf{Z}^{\top}\mathbf{Z})^{-1}\mathbf{Z}^{\top}\boldsymbol{\theta}.$$

Since $\mathbb{E}[e \mid Y, X] = \mathbb{E}[\varepsilon_r \mid Y, X] = 0$, we have $\mathbb{E}[\mathbf{Z}e] = \mathbb{E}[\mathbf{Z}\varepsilon_r] = \mathbf{0}$, and therefore

$$b_0 \equiv \mathbb{E}[\mathbf{Z}\boldsymbol{\theta}] = \mathbb{E}[\mathbf{Z}\varepsilon_e],$$

which is nonzero because the extraction error is correlated with outlet identity. Let $A_0 = \mathbb{E}[\mathbf{Z}\mathbf{Z}^{\top}]$, and let a_{ij} denote the (i, j) entry of A_0^{-1}. By the LLN, $n^{-1}\mathbf{Z}^{\top}\mathbf{Z} \xrightarrow{p} A_0$ and $n^{-1}\mathbf{Z}^{\top}\boldsymbol{\theta} \xrightarrow{p} b_0$, so by the continuous mapping theorem

$$\operatorname{plim}\hat{\boldsymbol{\beta}} = \beta_{\text{true}} + A_0^{-1}b_0 \equiv \boldsymbol{\beta}^{*} \neq \beta_{\text{true}}.$$

To characterize the sampling distribution, define the residual around the pseudo-true value, $u = \boldsymbol{\theta} - \mathbf{Z}^{\top}A_0^{-1}b_0$, which satisfies $\mathbb{E}[\mathbf{Z}u] = b_0 - A_0A_0^{-1}b_0 = \mathbf{0}$. Then

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{*} = (\mathbf{Z}^{\top}\mathbf{Z})^{-1}\mathbf{Z}^{\top}\mathbf{u},$$

and, under finite fourth moments, by the LLN and CLT:

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{*}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, A_0^{-1}\Sigma^{*}A_0^{-1}),$$

where $\Sigma^{*} = \mathbb{E}[\mathbf{Z}\mathbf{Z}^{\top}u^2]$. Note that

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \beta_{\text{true}}) = \sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{*}) + \sqrt{n}A_0^{-1}b_0$$

diverges whenever $A_0^{-1}b_0 \neq \mathbf{0}$, so the limiting distribution must be centered at $\boldsymbol{\beta}^{*}$ rather than at β_{true}.

To isolate the second component (β) in $\hat{\boldsymbol{\beta}}$, let $\mathbf{e}_2 = (0,1,0,\ldots,0)^{\top}$ and $\mu = \mathbf{e}_2^{\top}A_0^{-1}b_0$. Then:

$$\operatorname{plim}\hat{\beta} = \beta + \mu, \qquad \sqrt{n}(\hat{\beta} - \beta - \mu) \xrightarrow{d} \mathcal{N}\Big(0, \quad \Sigma_{\beta}^{*} = \mathbf{e}_2^{\top}A_0^{-1}\Sigma^{*}A_0^{-1}\mathbf{e}_2\Big),$$

where:

$$A_0^{-1} b_0 = \begin{pmatrix} a_{11}\,\mathbb{E}[\varepsilon_e] + a_{12}\,\mathbb{E}[Y\,\varepsilon_e] + \sum_{j=1}^{k} a_{1,(j+2)}\,\mathbb{E}[X_j\,\varepsilon_e] \\ a_{21}\,\mathbb{E}[\varepsilon_e] + a_{22}\,\mathbb{E}[Y\,\varepsilon_e] + \sum_{j=1}^{k} a_{2,(j+2)}\,\mathbb{E}[X_j\,\varepsilon_e] \\ \vdots \\ a_{(2+k)1}\,\mathbb{E}[\varepsilon_e] + a_{(2+k)2}\,\mathbb{E}[Y\,\varepsilon_e] + \sum_{j=1}^{k} a_{(2+k),(j+2)}\,\mathbb{E}[X_j\,\varepsilon_e] \end{pmatrix}.$$

Since $\mathbf{e}_2^\top = (0,\ 1,\ 0, \ldots, 0)$, it follows that

$$\mu = \mathbf{e}_2^\top \left(A_0^{-1} b_0\right) = a_{21}\,\mathbb{E}[\varepsilon_e] + a_{22}\,\mathbb{E}[Y\,\varepsilon_e] + \sum_{j=1}^{k} a_{2,(j+2)}\,\mathbb{E}[X_j\,\varepsilon_e].$$

By the Frisch–Waugh–Lovell theorem, the second row of A_0^{-1} applied to b_0 equals the coefficient on Y in the population projection of ε_e on $(1, Y, X)$. Hence

$$\mu = \frac{\mathbb{E}[\tilde{Y}\,\varepsilon_e]}{\mathbb{E}[\tilde{Y}^2]},$$

where $\tilde{Y}$ is the residual from the population projection of Y on $(1, X)$. By assumption, $\mathbb{E}[\tilde{Y}\,\varepsilon_e] \neq 0$, and therefore

$$\mu \neq 0.$$

Note that $\mathrm{Cov}(T, \varepsilon_e) \neq 0$ alone would not suffice for this conclusion, since

$$\mathrm{Cov}(T, \varepsilon_e) = \beta\,\mathrm{Cov}(Y, \varepsilon_e) + \gamma^\top\,\mathrm{Cov}(X, \varepsilon_e) + \mathrm{Cov}(e, \varepsilon_e),$$

and the last term can be nonzero even when ε_e is uncorrelated with (Y, X).

Thus, $\hat{\beta}$ is asymptotically biased: it converges in probability to $\beta + \mu$ rather than to β. □

In summary, in a partially-specified model where $\hat{T}_{ml}$ includes extraction errors correlated with outlet identity after partialling out X, the OLS estimator $\hat{\beta}$ is inconsistent. This bias stems from the correlation between the residualized regressor $\tilde{Y}$ and the error term ε_e. Consequently, $\hat{\beta}$ converges in probability to

$$\beta + \frac{\mathbb{E}[\tilde{Y}\,\varepsilon_e]}{\mathbb{E}[\tilde{Y}^2]}$$

rather than to β, and its limiting distribution is centered at this biased limit.

B.5.4 Theoretical Analysis of Omitted Variable Bias for the Two-step Model

This bias can occur when the variable of interest Y is correlated with the information not captured in T. Recall that $Z^{(-T)}$ represents the components of Z that are not captured by the extracted feature T. To quantify this bias, we first start by characterizing the true relationship between the variable of interest Y,

the extracted feature T, and the omitted components $Z^{(-T)}$ as $T = f_2(Y, k(Z^{(-T)}); \beta) + \varepsilon_0$, where $k(Z^{(-T)})$ is a non-parametric function that captures the potentially complex effects of the omitted variables $Z^{(-T)}$. In contrast, the econometric model that researchers estimate (as described earlier) is given by $T = f_2(Y; \beta) + \varepsilon$. The key challenge is that the error term ε is composed of both the original noise ε_0 and the bias due to the exclusion of $k(Z^{(-T)})$.

Remark 6. *Omitted Variable Bias (ε_{ov}) occurs when the error term includes the effect of omitted variables that are correlated with the independent variable Y. Formally, the error term can be expressed as:*

$$\varepsilon = \varepsilon_0 + \varepsilon_{ov} \quad where \quad Cov(Y, \varepsilon_{ov})) \neq 0$$

Proposition 6. ***Omitted Variable Bias:*** *Consider the linear model for T:*

$$\begin{aligned} T &= \beta Y + \delta k(Z^{(-T)}) + \varepsilon_0 \quad \textit{(True Model)}, \\ T &= \tilde{\beta} Y + \varepsilon \quad \textit{(Observed Model)}, \end{aligned} \tag{A.47}$$

where $\varepsilon = \varepsilon_0 + \varepsilon_{ov}$ and $\varepsilon_{ov} = \delta k(Z^{(-T)})$ represents the impact of the omitted variable. Then, the asymptotic bias of the OLS estimator $\hat{\tilde{\beta}}$ is given by:

$$\operatorname{plim} \hat{\tilde{\beta}} - \beta = \lambda \delta, \tag{A.48}$$

where $\lambda = Cov(Y, k(Z^{(-T)}))/Var(Y)$ is the coefficient from the projection of $k(Z^{(-T)})$ on Y (Hansen, 2022).

Proof. We start by considering the true model:

$$T = \beta Y + \delta k(Z^{(-T)}) + \varepsilon_0$$

and the estimated model, which omits the variable $k(Z^{(-T)})$:

$$T = \tilde{\beta} Y + \varepsilon$$

where $\varepsilon = \varepsilon_0 + \varepsilon_{ov}$ and $\varepsilon_{ov} = \delta k(Z^{(-T)})$. The OLS estimator for $\tilde{\beta}$ converges in probability to:

$$\operatorname{plim} \hat{\tilde{\beta}} = \frac{\operatorname{Cov}(Y, T)}{\operatorname{Var}(Y)}$$

Substituting the true model into this equation, we have:

$$\operatorname{Cov}(Y, T) = \operatorname{Cov}\left(Y, \beta Y + \delta k(Z^{(-T)}) + \varepsilon_0\right)$$

Expanding the covariance:

$$\operatorname{Cov}(Y, T) = \beta \operatorname{Cov}(Y, Y) + \delta \operatorname{Cov}(Y, k(Z^{(-T)})) + \operatorname{Cov}(Y, \varepsilon_0)$$

Assuming $\text{Cov}(Y, \varepsilon_0) = 0$, we get:

$$\text{Cov}(Y,T) = \beta \text{Var}(Y) + \delta \text{Cov}(Y, k(Z^{(-T)}))$$

Thus, the probability limit of the OLS estimator becomes:

$$\text{plim}\, \hat{\tilde{\beta}} = \frac{\beta \text{Var}(Y) + \delta \text{Cov}(Y, k(Z^{(-T)}))}{\text{Var}(Y)}$$

Rearranging:

$$\text{plim}\, \hat{\tilde{\beta}} = \beta + \frac{\delta \text{Cov}(Y, k(Z^{(-T)}))}{\text{Var}(Y)}$$

This simplifies to:

$$\text{plim}\, \hat{\tilde{\beta}} - \beta = \frac{\delta \text{Cov}(Y, k(Z^{(-T)}))}{\text{Var}(Y)}$$

where $\frac{\text{Cov}(Y, k(Z^{(-T)}))}{\text{Var}(Y)} = \lambda$ is the coefficient from the projection of $k(Z^{(-T)})$ on Y.

□

C Empirical Evidence on Bias in Reduced-Form Visual Slant

This section provides empirical evidence on the sources of bias that arise in reduced-form approaches to measuring visual media slant. We proceed in two steps. First, in § C.1, we investigate how extraction bias and omitted-variable bias arise in a two-step reduced-form model using a focused politician-level illustration based on images of Hillary Clinton. Second, in § C.2, we examine reduced-form outlet-level visual slant measures following Boxell (2021), assess their alignment with external ideology benchmarks, and show how adding politician and outlet controls improves, though does not fully resolve, these biases.

C.1 Bias in Two-Step Reduced-Form Models: A Politician-Level Illustration

In the previous section, we theoretically characterized how the two-step approach based on feature extraction results in information loss that can appear in forms of extraction bias and omitted variable bias. In this section, we empirically examine if the sources of bias discussed in §B.5.2 can bias the estimates of the two-step reduced form model outlined in §B.5.1. We focus on visual representations of Hillary Clinton as the focal politician p, where y_1 corresponds to images published by Democratic-leaning news outlets such as *CNN*, *The New York Times*, and the *BBC*, and y_2 corresponds to those from Republican-leaning outlets like *Fox News*, *Newsmax*, and *The Daily Mail*. Our goal is to estimate the parameter β as characterized in Equation (A.38). For illustrative purposes, we focus on the subset of our data containing images of Hillary Clinton (comprising 1,021 articles).

Following the two-step approach, we first employ the DeepFace model developed by Meta as the machine learning algorithm f_1 (Taigman et al., 2014). We extract the presence ($Z(\hat{T} = 1)$) or lack of a smile ($Z(\hat{T} = 0)$) in the images of Hilary Clinton using DeepFace.[9] In the second step, we apply a logistic regres-

[9] DeepFace utilizes a deep neural network with a nine-layer architecture and over 120 million parameters, including convolu-

sion model f_2 to quantify the relationship between the estimated facial expression $\hat{T}$ and the independent variable Y as:

$$\log\left(\frac{\Pr(\hat{T}=1 \mid Y)}{\Pr(\hat{T}=0 \mid Y)}\right) = \alpha + \beta \cdot Y,$$

where $Y = 1$ represents Democratic outlets (*CNN*, *The New York Times*, or *BBC*) and $Y = 0$ corresponds to Republican outlets (*Fox News*, *Newsmax*, or *The Daily Mail*). The coefficients α (intercept) and β (slope) are estimated from the data. For simplicity, this example excludes article-level features X.

However, the above specification ignores all the non-smile-related information in the image (represented by $Z^{(-T)}$). As discussed earlier, doing so can lead to biased estimates of β. To that end, we now consider a modified specification that accounts for a few other image characteristics. For illustrative purposes, we select three arbitrary image features: edge density, brightness, and contrast, which capture a few relevant aspects of the visual context. Thus, the extended specification of the logistic regression model becomes:

$$\log\left(\frac{\Pr(\hat{T}=1 \mid Y, Z^{(-T)})}{\Pr(\hat{T}=0 \mid Y, Z^{(-T)})}\right) = \alpha + \beta \cdot Y + \gamma_1 \cdot \text{EdgeDensity} + \gamma_2 \cdot \text{Brightness} + \gamma_3 \cdot \text{Contrast},$$

where $Z^{(-T)}$ includes the additional image characteristics, and γ_1, γ_2, and γ_3 are the coefficients that measure the effects of edge density, brightness, and contrast, respectively.

The results from both specifications are shown in Table A1. Column (1) presents the estimates for the model without other image characteristics. Here, the *Democratic Outlets* variable is statistically significant ($p < 0.01$), with a positive coefficient of 0.369, indicating that Democratic news outlets are more likely to show Hillary Clinton smiling as a democratic politician. However, when we include other image characteristics in Column (2), the coefficient for *Democratic Outlets* changes to 0.278 and loses statistical significance ($p < 0.10$). This reduction in both magnitude and significance suggests that part of the association initially attributed to *Democratic Outlets* in Column (1) was actually capturing differences in how images were presented across outlets. In Column (3), we show that the change in the coefficient for Democratic Outlets comes from the fact that the variable Democratic Outlets is correlated with some of the three image features excluded from the model in Column (1). It is important to emphasize that the correlation structure between the Democratic Outlet dummy and three arbitrary images only highlight the potential for the bias in the estimates from the reduced form model, and the correct estimate that can be obtained with including all relevant controls can have any directional relationship with the estimates in Column (1).

In sum, the empirical evidence confirms the existence of extraction and/or omitted variable bias and highlights the crucial role it plays in inference when social scientists use reduced-form models. Even in this simple illustrative example, we see that failing to account for correlated image characteristics can lead to biased estimates and incorrect conclusions about media bias and political polarization. Finally, it is crucial to emphasize that edge density, brightness, and contrast are just three arbitrary image characteristics selected from the much broader set of potential contextual features $Z^{(-T)}$. The visual content of an image encompasses a wide range of information, and the choice of these particular features is meant to demonstrate

tional layers and a 3D alignment step. This model has been adapted for tasks such as facial expression and gender recognition in a variety of social science contexts (Dvorkin et al., 2021; Luca et al., 2022).

Variable	*Dependent Variable: $\hat{T}$ (Happiness)*		*Dependent Variable: Y*
	(1)	(2)	(3)
Intercept	-1.170***	-1.809***	-1.604***
	(0.104)	(0.257)	(0.235)
Democratic Outlets (Y)	0.369**	0.278†	
	(0.141)	(0.145)	
Brightness Level		0.356	0.325
		(0.561)	(0.512)
Edge Density		0.878†	2.312***
		(0.477)	(0.456)
Contrast		0.537	1.450**
		(0.561)	(0.513)
No. of Obs.	1021	1021	1021
Pseudo-R^2	0.005	0.013	0.040
Log-Likelihood	-595.4	-591.3	-679.2
Likelihood Ratio Test	6.87**	14.97**	56.94***
Note:		†p<0.10; *p<0.05; **p<0.01; ***p<0.001	

Table A1: Results for logistic regressions. Columns 1–2 show results where T is the binary dependent variable (happiness), while column 3 shows results where Y (news outlet side) is the dependent variable where Republican outlets (*Fox News*, *Newsmax*, *Daily Mail*) are coded as $Y = 0$ and Democratic outlets (*The New York Times*, *BBC*, *CNN*) are coded as $Y = 1$. Image characteristics include brightness level, edge density, and contrast.

that even a limited selection of characteristics can significantly impact the results.

C.2 Bias in Reduced-Form Outlet-Level Visual Slant

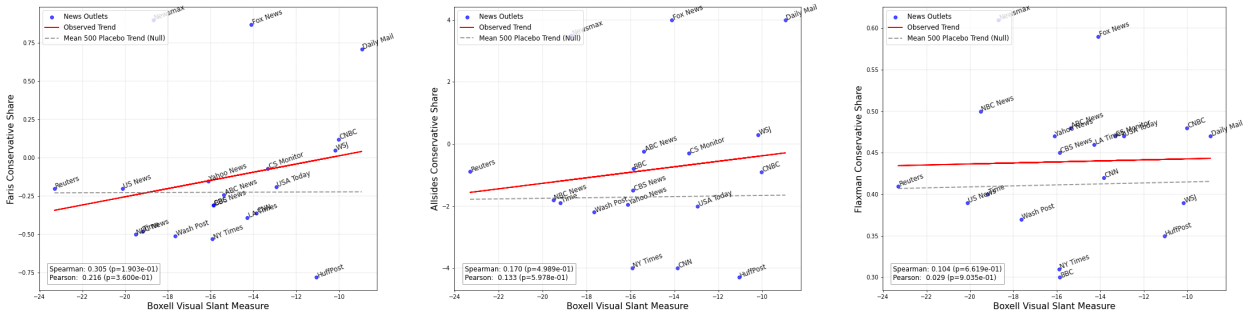
In this section, we examine visual slant using a reduced-form approach based on observed images, following Boxell (2021). The objective is to construct an outlet-level measure of visual slant that is directly comparable to existing benchmark measures, while highlighting the limitations of the reduced-form approach relative to our proposed estimation procedure.

Boxell (2021) measures outlet-level visual slant by comparing the expected values of a visual feature of interest across images used to portray Republican versus Democratic politicians. Let $\hat{T}_{pyi}$ denote the estimated value of feature T for image i of politician p on outlet y, where T is obtained by an off-the-shelf machine-learning model applied to the observed image. Boxell (2021) defines T corresponds to *relative favorability*, defined as happiness minus the sum of negative emotions inferred from facial-expression analysis. The outlet-level slant reported in Table 4 of Boxell (2021) is based on the difference $\mathbb{E}[\hat{T}_{pyi} \mid y, p \in \mathscr{P}_R] - \mathbb{E}[\hat{T}_{pyi} \mid y, p \in \mathscr{P}_D]$, with these party-specific expectations normalized relative to their cross-outlet averages to account for systematic baseline differences in politicians' visual presentation. This construction is purely reduced form: it relies on variation in the estimated feature of interest $\hat{T}$ as it appears in selected images and therefore overlooks the full richness of visual information embedded in those images, rather than isolating or holding fixed specific visual attributes.

The same reduced-form outlet-level slant measure can be recovered from a single image-level regression. Specifically, we can rewrite the construction as

$$\hat{T}_{pyi} = \sum_{y} \delta_y \mathbf{1}(y)\,\mathbf{1}(p \in \mathscr{P}_R) + \varepsilon_{pyi},$$

where $\mathbf{1}(\cdot)$ denotes an indicator function. The coefficient δ_y captures the Republican–Democrat gap in the expected value of the estimated feature $\hat{T}$ for outlet y. We apply this reduced-form specification to our data and evaluate how the resulting outlet-level visual slant correlates with established external measures of media ideology.



(a) Faris Conservative Share (b) AllSides Conservative Share (c) Flaxman Conservative Share

Figure A.3: Relationship between Boxell-style reduced-form visual slant and external conservative share benchmarks. Each panel plots outlet-level visual slant against a benchmark measure, with the red line indicating the observed linear trend.

Figure A.3 reports the results. Across all three benchmarks, Faris, AllSides, and Flaxman, we do not observe a statistically meaningful relationship between the reduced-form visual slant measure and external conservative share scores. The estimated correlations are small in magnitude and statistically insignificant, indicating that this reduced-form approach yields little alignment with established measures of media slant.

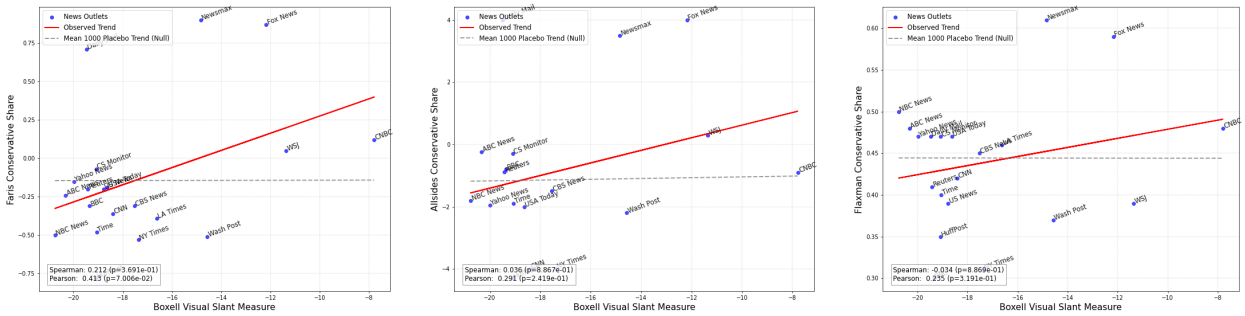
Moreover, the observed relationships are quantitatively similar to those obtained from a placebo exercise in which we randomly permute each politician's party affiliation between Republican and Democratic and re-estimate the same reduced-form specification. The resulting placebo correlations are centered close to zero and closely track the observed trends shown in Figure A.3. This similarity suggests that, absent feature isolation, reduced-form visual slant measures based on observed images capture little systematic partisan structure beyond what would arise by chance.

Given the drawbacks of the reduced-form approach discussed earlier, in particular, extraction bias from relying on a single estimated feature $\hat{T}$ and omitted-variable bias due to unobserved features and outlet-level heterogeneity, we extend the baseline specification by introducing additional controls. Specifically, we estimate

$$\hat{T}_{pyi} = \alpha_p + \gamma_y + \sum_{y} \delta_y \mathbf{1}(y)\,\mathbf{1}(p \in \mathscr{P}_R) + \varepsilon_{pyi},$$

which augments the reduced-form interaction with politician fixed effects α_p and outlet fixed effects γ_y.

Figure A.4 shows that adding politician and outlet fixed effects leads to stronger correlations between the



(a) Faris Conservative Share (b) AllSides Conservative Share (c) Flaxman Conservative Share

Figure A.4: Relationship between reduced-form visual slant with politician and outlet fixed effects and external conservative share benchmarks. Each panel plots outlet-level visual slant against the corresponding benchmark, with the red line indicating the observed linear trend and the dashed line denoting the mean placebo trend.

reduced-form visual slant measure and external ideology benchmarks relative to Figure A.3. Although the correlations with Faris, AllSides, and Flaxman remain statistically insignificant, their magnitudes increase and deviate more clearly from the placebo trend compared to the baseline reduced-form specification.

This improvement indicates that omitted politician-level heterogeneity and outlet-specific baseline differences contribute to bias in the simpler reduced-form model. Introducing these controls improves alignment with external measures, though substantial noise remains in the absence of feature isolation.

D Appendix: Implementation Details

In this appendix, we present implementation and validation details underlying our empirical analysis. In §D.1, we describe the implementation of the multi-modal machine learning architecture used to predict news outlet identity from textual, visual, and structured inputs. In §D.2, we describe the dimensionality reduction and clustering procedures used to organize image embeddings. In §D.3, we validate the generative model used to produce counterfactual images and assess whether the generated images isolate the intended visual changes.

D.1 Multi-Modal ML Implementation

First, to fully exploit the presence of similar events, we use the contextual information in articles. The contextual information comprises textual data X^{text} from article titles, publication date X^{date}, politicians' names P, and political affiliations P^{aff}, and image data Z. We want our model architecture to fully capture the clustering structure of the contextual information for two reasons. First, identifying clusters with diverse news outlets allows for identification of visual factors related to polarization. Second, identifying clusters with a single or only a few news outlets helps identify stylistic preferences of outlets, preventing those factors to play a role in determining the polarization parameter.

Figure 5 in the main text of the paper offers a visual overview of the model. To extract meaningful patterns in textual data X^{text}, we utilize Latent Dirichlet Allocation (LDA), which models the text data by identifying underlying topics across the articles. Although methods like BERT (Devlin et al., 2018) are known for their semantic understanding, LDA outperformed these models in our specific prediction task, particularly in identifying and quantifying topic distributions. Each article is thus represented as a 40-dimensional vector

F_{LDA}, encapsulating the relevance of different topics within the text. The LDA processing steps, including pre-processing and model training, are detailed in Web Appendix §D.1.1. Categorical data, including the publication date X^{date}, politicians' names P, and political affiliations P^{aff}, provides additional context that enhances the model's ability to discern patterns relevant to news outlet classification. These features are represented using *embedding layers*, where each categorical value is mapped to a dense vector representation. Specifically, we use an embedding size of 4 for the publication dates X^{date}, 8 for the politicians' names P, and 2 for political affiliations P^{aff}. For the image input, *ResNet-101* processes the entire image, leveraging its exceptional performance in general image recognition tasks to extract hierarchical and context-rich features from the broader visual content and scene information (He et al., 2016). To adapt these architectures to the specific requirements of our task, the last 10 layers of ResNet-101 and the last 5 layers of VGG-Face are fine-tuned, ensuring that both contextual and facial embeddings are optimized for our application.

The second challenge is related to correctly estimating the link between smile and news outlet prediction. We use *MTCNN* architecture for face detection due to its ability to perform joint face detection and alignment with high accuracy, ensuring precise focus on facial regions (Zhang et al., 2016). We pass the detected face through a pretrained *VGG-Face* network to extract facial features (Parkhi et al., 2015), and fine-tune the network for our outlet-prediction task. We design this part of the architecture to ensure that the predictive model accounts for the information in the politicians' faces. Later in results, we show how adding this element to the structure allows the model to correctly identify the differences between the counterfactual image versions.

In summary, the integration of modalities occurs through specialized attention mechanisms:

- *Chunk attention* is applied to the VGG-Face embeddings, combining them with categorical data (politicians' names P and affiliations P^{aff}) to capture correlations between facial features and structured metadata related to the image. This fusion ensures the model can link specific facial attributes to political or identity-related information (Liang et al., 2024) (see Web Appendix §D.1.2 for details).
- *Attention Mechanism* processes the structured data (categorical features, LDA topics), learning to prioritize the most relevant metadata features for the classification task (Vaswani, 2017) (see Web Appendix §D.1.3 for details).
- The embeddings from ResNet-101 that are passed through fully connected layers, enriched with contextual image information, are directly incorporated into the final representation (see Web Appendix §D.1.4 for details).

The outputs from these attention mechanisms and embeddings are used together in the final classification layer, which predicts the target news outlet. This design ensures that facial details, image context, and structured data interact effectively, enhancing the model's performance in discerning patterns across modalities.

After concatenating the feature vectors, the output from this layer is then passed through a final softmax layer to produce the classification prediction. The complete model details are explained in Web Appendix §D.1.5. The model is trained using the AdamW optimizer (Loshchilov and Hutter, 2017), which is particularly suited for large-scale data due to its adaptive learning rate and regularization through weight decay. The loss function used is weighted cross-entropy, which measures the discrepancy between the predicted probabilities and the true labels, optimizing the model to improve classification accuracy. The training pro-

cess's specifics, including the dataset's division, hyperparameter tuning, and regularization techniques, are discussed in Web Appendix §D.1.6.

With the model architecture defined, we choose parameters to maximize the weighted log-likelihood of the observed outlet over the training data, equivalently minimizing the weighted cross-entropy loss:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{\mathscr{N}} \sum_{y \in \mathscr{Y}} w_y \mathbf{1}\{Y_i = y\} \log g(Y_i = y \mid Z_i, X_i, P_i; \theta), \tag{A.49}$$

where $\mathbf{1}\{\cdot\}$ is the indicator function and the class weights w_y correct for the unequal number of articles across outlets. The resulting model is what we use to estimate the polarization parameter using Equation (12).

D.1.1 Textual Data Processing using LDA

Textual data X^{text} is pivotal as it provides contextual and content information from the articles. We evaluated several approaches for modeling this data, including BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) and LDA (Latent Dirichlet Allocation) (Blei et al., 2003). Despite BERT's advanced capabilities in understanding context and semantics, LDA solely demonstrated superior performance in our prediction task, particularly in identifying topics and their distributions across the articles.

LDA is a generative statistical model that explains sets of observations by unobserved groups, revealing why certain parts of the data are similar. The model assumes that each document is a mixture of a small number of topics and that each word in the document is attributable to one of the document's topics. The hyperparameters α and β are set to 'auto' to allow the model to learn these parameters during training. Mathematically, LDA posits the following generative process for a corpus D consisting of M documents, each containing N words:

Algorithm 2 LDA Generative Process (Blei et al., 2003)

for each topic k in $\{1, \ldots, K\}$ **do**
    Draw a distribution over words $\phi_k \sim \text{Dir}(\beta)$
end for
for each document d in $\{1, \ldots, M\}$ **do**
    Draw a distribution over topics $\theta_d \sim \text{Dir}(\alpha)$
    for each word n in $\{1, \ldots, N_d\}$ **do**
        Draw a topic $z_{dn} \sim \text{Multinomial}(\theta_d)$
        Draw a word $w_{dn} \sim \text{Multinomial}(\phi_{z_{dn}})$
    end for
end for

Here, α and β are hyperparameters of the Dirichlet distributions, θ_d is the topic distribution for document d, ϕ_k is the word distribution for topic k, z_{dn} is the topic assignment for the n-th word in document d, and w_{dn} is the n-th word in document d. In our model, we set the number of topics K to 40, and the model

is trained with 40 passes over the corpus to ensure topic extraction.

The preprocessing steps for the textual data include cleaning and tokenizing the news titles, followed by lemmatization and stopword removal using NLTK's WordNetLemmatizer and stopwords list (Bird et al., 2009). The processed tokens are then used to create bigrams and trigrams using Gensim's Phrases model (Řehůřek and Sojka, 2010). These n-grams help capture contextual relationships between words, making the textual data more informative for the LDA model.

The trained LDA model generates topic distributions for each document. For each document d, LDA provides the distribution θ_d over topics, which can be interpreted as the document's composition in terms of latent topics. These distributions are then scaled and normalized for input into the neural network. This transformation converts textual data into a structured format encapsulating underlying topics and their relevance to each document. As a result, we have a 40-dimensional vector representation for each news article title.

D.1.2 Face Information Branch: Framework and Operations

The Face Analysis branch of the proposed architecture is designed to extract, refine, and integrate facial information using a multi-step process that combines advanced detection, feature extraction, dimensionality reduction, and attention mechanisms. This section details each operation, starting from face detection via the MTCNN, progressing to feature extraction using Face-VGG, and culminating with chunked attention for integrating structured metadata. Given an input image $\mathbf{I} \in \mathbb{R}^{H\times W\times C}$, the MTCNN algorithm detects faces and aligns them for further processing. The detection process is divided into three stages. In the first stage, a proposal network is used to generate candidate bounding boxes:

$$\mathbf{R}_1 = f_1(\mathbf{I};\boldsymbol{\theta}_1),$$

where $\mathbf{R}_1$ denotes the set of candidate bounding boxes, and $\boldsymbol{\theta}_1$ are the parameters of the proposal network. These candidate regions are refined in the second stage using a refine network:

$$\mathbf{R}_2 = f_2(\mathbf{R}_1;\boldsymbol{\theta}_2),$$

where $\mathbf{R}_2$ represents the refined bounding boxes. Finally, the output network predicts the bounding boxes and facial landmarks:

$$(\mathbf{R}_{\text{final}}, \mathbf{L}) = f_3(\mathbf{R}_2;\boldsymbol{\theta}_3),$$

where $\mathbf{L} \in \mathbb{R}^{5\times 2}$ are the coordinates of the facial landmarks. Using $\mathbf{R}_{\text{final}}$ and $\mathbf{L}$, the aligned face crop $\mathbf{F} \in \mathbb{R}^{160\times 160\times 3}$ is extracted from the original image. Each aligned face crop $\mathbf{F}$ is passed through a pre-trained Face-VGG network to extract a 512-dimensional feature vector. Let $g_{\text{VGG}}(\cdot;\boldsymbol{\Theta}_{\text{VGG}})$ denote the Face-VGG model, where $\boldsymbol{\Theta}_{\text{VGG}}$ represents its parameters. The extracted feature vector $\mathbf{v} \in \mathbb{R}^{512}$ is computed as:

$$\mathbf{v} = g_{\text{VGG}}(\mathbf{F};\boldsymbol{\Theta}_{\text{VGG}}).$$

The last five layers of the Face-VGG model are fine-tuned to adapt to the specific task. The extracted feature vector $\mathbf{v}$ is projected into a lower-dimensional space using a fully connected layer. The reduced feature vector $\mathbf{v}_{\text{red}} \in \mathbb{R}^{256}$ is given by:

$$\mathbf{v}_{\text{red}} = \mathbf{W}_{\text{red}}\mathbf{v} + \mathbf{b}_{\text{red}},$$

where $\mathbf{W}_{\text{red}} \in \mathbb{R}^{256\times 512}$ and $\mathbf{b}_{\text{red}} \in \mathbb{R}^{256}$ are learnable parameters of the layer. To improve generalization, batch normalization and dropout are applied to $\mathbf{v}_{\text{red}}$. The normalized and regularized feature vector $\mathbf{v}_{\text{norm}}$ is computed as:

$$\mathbf{v}_{\text{norm}} = \text{BatchNorm}(\mathbf{v}_{\text{red}}),$$

and the final output after dropout is:

$$\mathbf{v}_{\text{drop}} = \text{Dropout}(\mathbf{v}_{\text{norm}}, p = 0.3).$$

The reduced feature vector $\mathbf{v}_{\text{drop}} \in \mathbb{R}^{256}$ is divided into $k = 8$ equally sized chunks, each of dimension 32:

$$\mathbf{v}_{\text{chunk}} = \{\mathbf{v}_i \in \mathbb{R}^{32} \mid i = 1,\ldots,k\}, \quad \mathbf{v}_i = \mathbf{v}_{\text{drop}}[32(i-1) : 32i].$$

This chunking operation allows the model to analyze localized regions of the image feature vector (Fu et al., 2019). To incorporate structured metadata, such as person and party embeddings, the chunked features $\mathbf{v}_{\text{chunk}}$ interact with a metadata embedding $\mathbf{m} \in \mathbb{R}^{32}$. The metadata embedding is computed as:

$$\mathbf{m} = \mathbf{W}_{\text{meta}}\mathbf{x} + \mathbf{b}_{\text{meta}},$$

where $\mathbf{x}$ represents the concatenated embeddings of person and party, $\mathbf{W}_{\text{meta}} \in \mathbb{R}^{32\times d_{\text{meta}}}$, and $\mathbf{b}_{\text{meta}} \in \mathbb{R}^{32}$ are learnable parameters. An attention mechanism assigns weights α_i to each chunk $\mathbf{v}_i$ based on its similarity to the metadata embedding $\mathbf{m}$:

$$\alpha_i = \frac{\exp\left(\mathbf{v}_i^\top \mathbf{m}\right)}{\sum_{j=1}^{k} \exp\left(\mathbf{v}_j^\top \mathbf{m}\right)}.$$

The attention-weighted feature representation $\mathbf{v}_{\text{att}} \in \mathbb{R}^{256}$ is computed as:

$$\mathbf{v}_{\text{att}} = \sum_{i=1}^{k} \alpha_i \mathbf{v}_i.$$

The output of the Face Analysis branch $\mathbf{v}_{\text{att}}$ is concatenated with features from the ResNet branch $\mathbf{f}_{\text{ResNet}}$ and the metadata branch $\mathbf{f}_{\text{meta}}$ to form the final aggregated feature vector:

$$\mathbf{z} = [\mathbf{v}_{\text{att}}; \mathbf{f}_{\text{ResNet}}; \mathbf{f}_{\text{meta}}].$$

This feature vector $\mathbf{z} \in \mathbb{R}^{d_{\text{final}}}$ is passed through subsequent layers for classification.

D.1.3 Structured Metadata Branch: Framework and Operations

The Structured Data Branch in the proposed architecture processes tabular metadata, including embeddings for categorical features (e.g., person, party, and date) and topic distributions derived from LDA. The key component of this branch is the multi-head attention mechanism, inspired by the Transformer model (Vaswani, 2017), which dynamically highlights relevant interactions among structured features. This section provides the mathematical foundations and operations in a coherent manner.

Let the structured input be represented as $\mathbf{x} \in \mathbb{R}^{d_{\text{struct}}}$, where $d_{\text{struct}} = 54$, obtained by concatenating the embeddings of categorical features and topic distributions. The embeddings for person, party, and date are denoted as $P \in \mathbb{R}^{d_{\text{person}}}$, $P^{\text{aff}} \in \mathbb{R}^{d_{\text{party}}}$, and $X^{\text{date}} \in \mathbb{R}^{d_{\text{date}}}$, respectively, while the topic vector is $X^{\text{text}} \in \mathbb{R}^{d_{\text{topics}}}$, where $d_{\text{text}} = 40$. The structured input vector $\mathbf{x}$ is thus formed as:

$$\mathbf{x} = [P; P^{\text{aff}}; X^{\text{date}}; X^{\text{text}}],$$

where $[\cdot]$ denotes vector concatenation. This input vector is then projected into a latent representation space of dimension $d_{\text{latent}} = 64$ via a linear transformation:

$$\mathbf{h} = \mathbf{W}_{\text{linear}}\mathbf{x} + \mathbf{b}_{\text{linear}},$$

where $\mathbf{W}_{\text{linear}} \in \mathbb{R}^{d_{\text{latent}} \times d_{\text{struct}}}$ and $\mathbf{b}_{\text{linear}} \in \mathbb{R}^{d_{\text{latent}}}$ are learnable weights and biases. The resulting vector $\mathbf{h} \in \mathbb{R}^{d_{\text{latent}}}$ serves as the input to the multi-head attention mechanism. The multi-head attention mechanism begins by projecting $\mathbf{h}$ into three distinct subspaces corresponding to the query ($\mathbf{Q}$), key ($\mathbf{K}$), and value ($\mathbf{V}$) representations (Vaswani, 2017). These projections are computed as:

$$\mathbf{Q} = \mathbf{W}_Q\mathbf{h}, \quad \mathbf{K} = \mathbf{W}_K\mathbf{h}, \quad \mathbf{V} = \mathbf{W}_V\mathbf{h},$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d_{\text{head}} \times d_{\text{latent}}}$ are learnable weight matrices, and d_{head} represents the dimensionality of each attention head. The scaled dot-product attention mechanism computes the compatibility between queries and keys, resulting in an attention score matrix:

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_{\text{head}}}}\right),$$

where $\text{softmax}(\cdot)$ normalizes the scores across all keys for each query, and the scaling factor $\sqrt{d_{\text{head}}}$ stabilizes gradients during training by mitigating the effect of large dot-product magnitudes. The attention scores $\mathbf{A} \in \mathbb{R}^{d_{\text{head}} \times d_{\text{head}}}$ modulate the value vectors $\mathbf{V}$, yielding the attention output:

$$\mathbf{O}_{\text{single}} = \mathbf{A}\mathbf{V}.$$

To capture diverse interactions among the structured features, multiple attention heads are employed. For each attention head i, the process is repeated independently, producing:

$$\mathbf{O}_i = \mathbf{A}_i\mathbf{V}_i = \text{softmax}\left(\frac{\mathbf{Q}_i\mathbf{K}_i^\top}{\sqrt{d_{\text{head}}}}\right)\mathbf{V}_i,$$

where $i = 1, \ldots, h$ and h is the number of heads. The outputs from all attention heads are concatenated and projected back into the latent representation space using a linear transformation:

$$\mathbf{O}_{\text{multi}} = \mathbf{W}_{\text{out}} \left[\mathbf{O}_1; \mathbf{O}_2; \ldots; \mathbf{O}_h\right],$$

where $\mathbf{W}_{\text{out}} \in \mathbb{R}^{d_{\text{latent}} \times (h \cdot d_{\text{head}})}$ is a learnable projection matrix. The multi-head attention output $\mathbf{O}_{\text{multi}} \in \mathbb{R}^{d_{\text{latent}}}$ encodes a rich and dynamic representation of the structured data. To ensure stability and prevent overfitting, the output $\mathbf{O}_{\text{multi}}$ is normalized using batch normalization and regularized with dropout. These operations are defined as:

$$\mathbf{h}_{\text{norm}} = \text{BatchNorm}(\mathbf{O}_{\text{multi}}), \quad \mathbf{h}_{\text{drop}} = \text{Dropout}(\mathbf{h}_{\text{norm}}, p),$$

where p is the dropout probability. The final representation of the structured data branch is $\mathbf{h}_{\text{struct}} = \mathbf{h}_{\text{drop}}$, a 64-dimensional vector ready to be fused with the outputs from other branches in the model. The multi-head attention mechanism allows the model to dynamically prioritize different aspects of the structured data by assigning relevance scores based on the compatibility of queries and keys. The queries ($\mathbf{Q}$) represent the model's current focus, the keys ($\mathbf{K}$) encode contextual information about all features, and the values ($\mathbf{V}$) provide the corresponding information content. By using multiple attention heads, the model captures diverse patterns and relationships, resulting in a robust and task-specific representation of the structured metadata.

D.1.4 Image Contextual Information Branch: Framework and Operations

The ResNet101 branch is responsible for extracting high-dimensional feature representations from the full raw image data. Let $\mathbf{I} \in \mathbb{R}^{H \times W \times C}$ represent an input image, where $H = 244$ and $W = 244$ are the height and width, and C is the number of color channels. The input image passes through a pre-trained ResNet101 network, with the last 10 layers trainable. Denote the ResNet101 transformation as $g_{\text{ResNet}}(\cdot; \boldsymbol{\Theta}_{\text{ResNet}})$, where $\boldsymbol{\Theta}_{\text{ResNet}}$ are the trainable weights of the final layers. The output feature map $\mathbf{F}_{\text{ResNet}} \in \mathbb{R}^{d_{\text{ResNet}}}$ is given by:

$$\mathbf{F}_{\text{ResNet}} = g_{\text{ResNet}}(\mathbf{I}; \boldsymbol{\Theta}_{\text{ResNet}}),$$

where $d_{\text{ResNet}} = 2048$ is the dimensionality of the extracted feature vector. This high-dimensional feature vector is flattened and passed through a linear layer for dimensionality reduction:

$$\mathbf{f}_{\text{ResNet}} = \mathbf{W}_{\text{reduce}} \mathbf{F}_{\text{ResNet}} + \mathbf{b}_{\text{reduce}},$$

where $\mathbf{W}_{\text{reduce}} \in \mathbb{R}^{d_{\text{reduce}} \times d_{\text{ResNet}}}$, $\mathbf{b}_{\text{reduce}} \in \mathbb{R}^{d_{\text{reduce}}}$, and $d_{\text{reduce}} = 512$. Batch normalization and dropout are applied to enhance generalization:

$$\mathbf{f}_{\text{norm}} = \text{BatchNorm}(\mathbf{f}_{\text{ResNet}}), \quad \mathbf{f}_{\text{drop}} = \text{Dropout}(\mathbf{f}_{\text{norm}}, p),$$

where $p = 0.6$ is the dropout probability. The final output of the ResNet101 branch is $\mathbf{f}_{\text{drop}} \in \mathbb{R}^{512}$, which is fused with outputs from other branches for classification.

Layer (type)	Output Shape	Param #
Branch 1: Face Information		
MTCNN (Face Detection)	[-1, 160, 160, 3]	0
Face-VGG (Last 5 layers trainable)	[-1, 512]	23,560,896
Linear (Face-VGG to 256)	[-1, 256]	131,328
BatchNorm1d (Face-VGG normalization)	[-1, 256]	512
Dropout (Face-VGG, 0.3)	[-1, 256]	0
Chunked VGG Features (8 chunks of 32)	[-1, 8, 32]	0
Linear (person + party to 32)	[-1, 1, 32]	384
Chunked Attention (Chunked VGG attention with person + party)	[-1, 256]	65,792
Branch 2: Image Contextual Information		
ResNet101 (Last 10 layers trainable)	[-1, 2048]	44,549,160
Flatten (ResNet feature vector)	[-1, 2048]	0
Linear (ResNet to 512)	[-1, 512]	1,049,088
BatchNorm1d (ResNet normalization)	[-1, 512]	1,024
Dropout (ResNet, 0.6)	[-1, 512]	0
Branch 3: Structured Metadata		
Embedding (person: 8, party: 2, date: 4)	[-1, 14]	0
Latent Dirichlet Allocation (LDA topics: 40)	[-1, 40]	0
Concatenation (person, party, date, topics)	[-1, 54]	0
Linear (Structured data to 64)	[-1, 64]	3,520
MultiHeadAttention (Structured data attention)	[-1, 64]	16,512
BatchNorm1d (Structured data normalization)	[-1, 64]	128
Dropout (Structured data, 0.4)	[-1, 64]	0
Final: Concatenation		
Concatenation (VGG + Structured + ResNet)	[-1, 832]	0
Dropout (Final dropout, 0.5)	[-1, 832]	0
MLP classifier (Linear 832→256, ReLU, Linear 256→20)	[-1, 20]	218,388
Total params		69,608,628
Trainable params		10,790,888
Non-trainable params		58,817,740

Table A2: Architecture of the Multi-Modal Machine Learning Model

D.1.5 Final Model Architecture

The final architecture of the multi-modal model is summarized in Table A2. This table provides an overview of each layer in the model.

The outputs from the three branches—Face Analysis ($\mathbf{f}_{\text{face}} \in \mathbb{R}^{d_{\text{face}}}$), Structured Data ($\mathbf{h}_{\text{struct}} \in \mathbb{R}^{d_{\text{struct}}}$), and ResNet101 ($\mathbf{f}_{\text{drop}} \in \mathbb{R}^{d_{\text{reduce}}}$)—are concatenated to form a unified feature vector:

$$\mathbf{z} = [\mathbf{f}_{\text{face}}; \mathbf{h}_{\text{struct}}; \mathbf{f}_{\text{drop}}] \in \mathbb{R}^{d_{\text{final}}},$$

where $d_{\text{final}} = d_{\text{face}} + d_{\text{struct}} + d_{\text{reduce}}$. In this case, $d_{\text{face}} = 256$, $d_{\text{struct}} = 64$, and $d_{\text{reduce}} = 512$, resulting in $d_{\text{final}} = 832$. To enhance generalization and mitigate overfitting, dropout is applied to the concatenated representation:

$$\mathbf{z}_{\text{drop}} = \text{Dropout}(\mathbf{z}, p),$$

where $p = 0.5$ is the dropout rate. The resulting vector $\mathbf{z}_{\text{drop}} \in \mathbb{R}^{d_{\text{final}}}$ is passed through a fully connected classification layer to compute the logits for the n_{class} target classes:

$$\mathbf{y} = \mathbf{W}_{\text{class}} \, \text{ReLU}\big(\mathbf{W}_{\text{fuse}} \, \mathbf{z}_{\text{drop}} + \mathbf{b}_{\text{fuse}}\big) + \mathbf{b}_{\text{class}},$$

where $\mathbf{W}_{\text{fuse}} \in \mathbb{R}^{256 \times d_{\text{final}}}$ and $\mathbf{b}_{\text{fuse}} \in \mathbb{R}^{256}$ are the weights and biases of the hidden fusion layer, and $\mathbf{W}_{\text{class}} \in \mathbb{R}^{n_{\text{class}} \times 256}$ and $\mathbf{b}_{\text{class}} \in \mathbb{R}^{n_{\text{class}}}$ are the weights and biases of the classification layer. The predicted class probabilities are obtained by applying the softmax function to the logits:

$$\hat{\mathbf{y}}_i = \frac{\exp(y_i)}{\sum_{j=1}^{n_{\text{class}}} \exp(y_j)}, \quad i = 1, \ldots, n_{\text{class}},$$

where y_i is the i-th component of $\mathbf{y}$, and $\hat{\mathbf{y}}_i$ represents the probability of the i-th class. The final output $\hat{\mathbf{y}} \in \mathbb{R}^{n_{\text{class}}}$ is a normalized probability distribution over the target classes, enabling robust multi-class predictions.

D.1.6 Model Training and Fine-Tuning

To ensure robust model evaluation, the dataset is split into training (85%) and test (15%) sets. The splitting process preserves the distribution of key features by stratifying based on both news center and date, ensuring balanced records across all classes. The training set is used to optimize the model parameters, while the test data set is used for assessing model performance and polarization measurement to avoid overfitting bias.

The training process employs the AdamW optimizer (Loshchilov and Hutter, 2017), configured with a learning rate $\eta = 0.0001$ and a weight decay coefficient $\lambda = 0.01$. This optimizer is particularly well-suited for large-scale data due to its adaptive learning rates and decoupled weight decay, effectively balancing convergence speed and regularization. The loss function used for training is a weighted cross-entropy loss, adjusted to account for class imbalance. Let $\mathscr{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$ represent a batch of N training samples, where $\mathbf{y}_i \in \{0,1\}^C$ is the one-hot encoded true label for the i-th sample, and $C = 20$ denotes the number of classes. The predicted class probabilities are given by $\hat{\mathbf{y}}_i = \text{softmax}(\mathbf{z}_i)$, where $\mathbf{z}_i \in \mathbb{R}^C$ is the logit vector for the i-th sample. The weighted cross-entropy loss with label smoothing is defined as:

$$\mathscr{L} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} w_c \, y_{i,c} \log \hat{y}_{i,c},$$

where w_c is the weight for class c, computed as the inverse of its relative frequency in the training set to mitigate the impact of class imbalance:

$$w_c = \frac{1}{\text{freq}(c)} \cdot \frac{\sum_{j=1}^{C} \text{freq}(j)}{C}.$$

The optimizer updates the model parameters $\boldsymbol{\theta}$ at each iteration by minimizing the total loss $\mathscr{L}$. The parameter updates follow the AdamW rule:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \frac{\hat{\mathbf{m}}_t}{\sqrt{\hat{\mathbf{v}}_t} + \varepsilon} - \lambda \boldsymbol{\theta}_t,$$

where $\hat{\mathbf{m}}_t$ and $\hat{\mathbf{v}}_t$ are the bias-corrected first and second moments of the gradient, respectively, and ε is a

small constant to ensure numerical stability. The algorithm for parameter updates is formally presented below:

Algorithm 3 AdamW Optimizer

Input: Learning rate η, decay rates β_1, β_2, weight decay coefficient λ, small constant ε

Initialize: $\mathbf{m}_0 = 0$, $\mathbf{v}_0 = 0$, $t = 0$

for each iteration **do**

$t = t + 1$

Compute gradients of the loss: $\mathbf{g}_t = \nabla_{\boldsymbol{\theta}} \mathscr{L}(\boldsymbol{\theta}_{t-1})$

Update biased first moment estimate:

$$\mathbf{m}_t = \beta_1 \mathbf{m}_{t-1} + (1 - \beta_1)\mathbf{g}_t$$

Update biased second moment estimate:

$$\mathbf{v}_t = \beta_2 \mathbf{v}_{t-1} + (1 - \beta_2)\mathbf{g}_t^{\odot 2}$$

Compute bias-corrected first moment estimate:

$$\hat{\mathbf{m}}_t = \frac{\mathbf{m}_t}{1 - \beta_1^t}$$

Compute bias-corrected second moment estimate:

$$\hat{\mathbf{v}}_t = \frac{\mathbf{v}_t}{1 - \beta_2^t}$$

Update parameters:

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \eta \frac{\hat{\mathbf{m}}_t}{\sqrt{\hat{\mathbf{v}}_t} + \varepsilon} - \lambda \boldsymbol{\theta}_{t-1}$$

end for

The optimization process continues over 30 epochs, where each epoch consists of processing mini-batches of data of size 64. At each iteration, the model updates its parameters using the computed gradients and the AdamW update rule. The weight decay term $\lambda \boldsymbol{\theta}_{t-1}$ ensures regularization by penalizing large weights, helping to improve generalization. Metrics such as validation loss and classification accuracy are monitored at the end of each epoch to assess the model's performance and ensure generalization to unseen data.

D.2 Details of PCA and t-SNE for Reducing Image Dimensionality to Two

PCA is used to reduce the dimensionality of the embeddings from d to $d' \ll d$ by projecting them onto a lower-dimensional subspace that retains most of the variance in the data. Mathematically, PCA identifies a set of orthogonal components $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{d'} \in \mathbb{R}^d$, where each component maximizes the variance of the

projected data. The reduced embedding for an image is given by:

$$\mathbf{e}_i^{\text{PCA}} = \mathbf{W}^\top \mathbf{e}_i, \quad \mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{d'}],$$

where $\mathbf{W}$ is the matrix of the top d' eigenvectors of the covariance matrix of the embeddings, sorted by their corresponding eigenvalues. This transformation ensures that the majority of the information in the original embeddings is preserved in the reduced representation, making subsequent computations more efficient.

After reducing the embeddings to d' dimensions using PCA, we project them into a two-dimensional space, $\mathbf{e}_i^{\text{t-SNE}} \in \mathbb{R}^2$, using t-SNE for visualization. t-SNE, or t-Distributed Stochastic Neighbor Embedding, is a dimensionality reduction technique designed specifically for high-dimensional data visualization (Belkina et al., 2019). It aims to preserve the local structure of the data by modeling the similarity between points i and j in the original high-dimensional space as probabilities $P = \{p_{ij}\}$, and then finding a 2D embedding where the similarities, represented by $Q = \{q_{ij}\}$, approximate P. In this 2D space, the similarity between points i and j is modeled using a Student-t distribution, where the pairwise similarity is defined as:

$$q_{ij} = \frac{\left(1 + \|\mathbf{e}_i^{\text{t-SNE}} - \mathbf{e}_j^{\text{t-SNE}}\|^2\right)^{-1}}{\sum_{k \neq l} \left(1 + \|\mathbf{e}_k^{\text{t-SNE}} - \mathbf{e}_l^{\text{t-SNE}}\|^2\right)^{-1}},$$

where $\mathbf{e}_i^{\text{t-SNE}}$ and $\mathbf{e}_j^{\text{t-SNE}}$ represent the 2D coordinates of points i and j, $\|\mathbf{e}_i^{\text{t-SNE}} - \mathbf{e}_j^{\text{t-SNE}}\|$ is the Euclidean distance between points i and j in the 2D space, and k and l iterate over all points in the dataset to compute the denominator of the similarity measure for proper normalization. t-SNE attempts to arrange the 2D points such that their similarities, denoted as $Q = \{q_{ij}\}$, approximate the similarities in the high-dimensional space, denoted as $P = \{p_{ij}\}$. Here, $P = \{p_{ij}\}$ represents the pairwise similarities computed from the PCA-reduced high-dimensional embeddings, and $Q = \{q_{ij}\}$ represents the pairwise similarities in the 2D space computed using the formula above. The arrangement of points in 2D is achieved by minimizing the Kullback-Leibler (KL) divergence:

$$\text{KL}(P\|Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}},$$

where p_{ij} and q_{ij} denote the similarity between points i and j in the high-dimensional and 2D spaces, respectively. By minimizing this divergence, t-SNE ensures that points with high similarity in the high-dimensional space (large p_{ij}) remain close in the 2D space (large q_{ij}), while dissimilar points (small p_{ij}) are positioned farther apart. The resulting 2D embeddings, $\mathbf{e}_i^{\text{t-SNE}}$, enable visualization of images' space,

To further analyze the patterns in the embedding space for a specific politician p, we perform clustering on the reduced embeddings associated with their images. For each politician p, let the set of embeddings corresponding to their images be denoted as $\mathscr{E}^p = \{\mathbf{e}_i^{\text{PCA}} \mid i \in \mathscr{I}_p\}$, where $\mathscr{I}_p$ is the set of image indices for politician p. These embeddings are grouped into $k = 20$ clusters using K-Means clustering. The clustering objective for politician p is to minimize the within-cluster sum of squared distances:

$$\mathscr{C}_i^p = \arg\min_{\mathscr{C}} \sum_{j=1}^{k} \sum_{\mathbf{e}_i^{\text{PCA}} \in \mathscr{C}_j^p} \|\mathbf{e}_i^{\text{PCA}} - \mu_j^p\|^2,$$

where $\mathscr{C}_j^p$ is the set of embeddings assigned to cluster j for politician p, and μ_j^p is the centroid of cluster j for p. Intuitively, each cluster represents a group of visually similar images, which may correspond to specific events or contextual features, such as similar camera angles and settings. For instance, images of politician p during a particular speech or event are likely to form a distinct cluster, capturing the shared visual context of those images. Now, for a politician p and a given cluster k, we can investigate how different news outlets select images.

D.3 Validating Generated Counterfactual Images

In this part, we explain generating counterfactual images. First, for each politician p, we select three images $\tilde{Z}_{p1}^0, \tilde{Z}_{p2}^0, \tilde{Z}_{p3}^0$ that are not used to train our multi-modal ML model, sum up to 84 images. To ensure unbiasedness, these images are chosen to represent the politician in a neutral manner, avoiding any inherent slant. Consequently, this selected image set inherently represents the neutral version z^0. Additionally, selecting three images helps mitigate any specific biases that might arise from using a single image and ensures that our analysis is robust across different visual contexts.

Using the approach proposed by Mirza and Osindero (2014), we generate a smiley version of each image while keeping other factors constant. In other words, we apply the transformation π^1 on images $\tilde{Z}_{p1}^0, \tilde{Z}_{p2}^0, \tilde{Z}_{p3}^0$ to obtain $\pi^1(\tilde{Z}_{p1}^0) = \tilde{Z}_{p1}^1, \pi^1(\tilde{Z}_{p2}^0) = \tilde{Z}_{p2}^1, \pi^1(\tilde{Z}_{p3}^0) = \tilde{Z}_{p3}^1$. Each $\tilde{Z}^1$ now embodies the slant notion, specifically a smile, as shown in Figure 4 in the main text of the paper.

Since the isolated change in smile and facial expression is crucial, we further assess the generated images to see if other image characteristics have changed. The generated photos should have no other significant differences. As discussed, one of the challenges with the two-stage approach is that image characteristics such as brightness or colorfulness can cause bias in the model (see §B.5.2). Therefore, we measure the brightness and colorfulness of images before and after applying the smile operator π^1. The histograms of these characteristics are shown in Figure A.5:

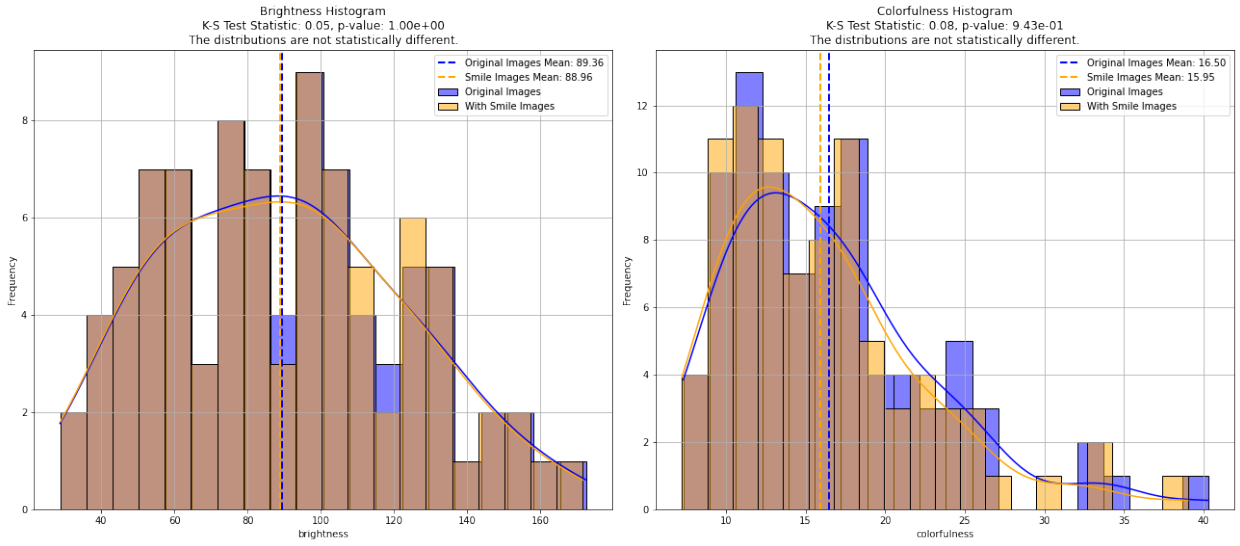


Figure A.5: Histograms of brightness and colorfulness for original and smiley images.

For brightness, the mean value for the original images is 89.36, while for the smiley images, it is 88.96. The Kolmogorov-Smirnov (K-S) test statistic is 0.05 with a p-value of $\approx$1.00, indicating that the distributions are not statistically different. This suggests that the introduction of a smile does not significantly alter the brightness of the images. Similarly, for colorfulness, the mean value for the original images is 16.50, compared to 15.95 for the smiley images. The K-S test statistic here is 0.08 with a p-value of 0.943, again indicating no statistical difference between the distributions.

This analysis confirms that the generated smiley images do not exhibit significant changes in brightness and colorfulness compared to the original images. This supports the validity of using these counterfactual images to isolate the effect of the smile on the parameter of interest, without introducing additional biases from other image characteristics.

E Supplementary Results

In this section, we report supplementary empirical results that support and extend the analysis in the main text. In §E.1, we document the predictive performance of the multi-modal news outlet prediction model. In §E.2, we present hypothesis tests for the distribution of visual polarization. In §E.3, we report results from sample splitting and assess relative prediction performance. In §E.4, we present individual news outlet results. In §E.5, we provide additional correlation analyses using existing partisanship scores. In §E.6, we compare our estimation procedure with the two-stage approach. In §E.7, we extend the analysis to alternative visual features. Finally, in §E.8, we present individual politician-level results.

E.1 Detailed Multi-Modal ML Performance

In this section, we detail the performance of our multi-modal multi-class classification problem for news outlet prediction. As mentioned earlier, we use an 85%-15% split between training and test data. We consider four measures to evaluate the model's predictive performance – (1) Accuracy, (2) Precision, (3) Recall, and (4) Weighted Cross-Entropy (WCE). Accuracy is simply measured as:

$$\text{Accuracy} = \frac{\sum_{i=1}^{N} \mathbf{1}(Y_i = \text{argmax}_c \hat{Y}_{ic})}{N},$$

where $\hat{Y}_{ic}$ is our predicted probability for news outlet c producing article i.

Precision evaluates the reliability of the model's predictions by measuring the proportion of correctly predicted articles for a given outlet among all articles predicted for that outlet. Formally, it is defined as:

$$\text{Precision} = \frac{\sum_{i=1}^{N} \mathbf{1}(Y_i = c \wedge \text{argmax}_c \hat{Y}_{ic} = c)}{\sum_{i=1}^{N} \mathbf{1}(\text{argmax}_c \hat{Y}_{ic} = c)},$$

where $\mathbf{1}(Y_i = c \wedge \text{argmax}_c \hat{Y}_{ic} = c)$ counts the true positives (correct predictions for outlet c), and $\mathbf{1}(\text{argmax}_c \hat{Y}_{ic} = c)$ counts all predictions made for outlet c.

Recall, on the other hand, measures the model's ability to identify all true articles for a given outlet. It is defined as:

$$\text{Recall} = \frac{\sum_{i=1}^{N} \mathbf{1}(Y_i = c \wedge \text{argmax}_c \hat{Y}_{ic} = c)}{\sum_{i=1}^{N} \mathbf{1}(Y_i = c)},$$

where $\mathbf{1}(Y_i = c)$ counts all true articles for outlet c. While Precision emphasizes minimizing false positives, Recall ensures that false negatives are minimized, making them complementary metrics to assess the model's performance comprehensively.

Finally, WCE Loss is defined as:

$$-\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} w_c Y_{ic} \log\left(\hat{Y}_{ic}\right),$$

where C is the number of classes and w_c is the weight assigned to class c to handle class imbalance.

Source	Accuracy	Precision	Recall	F1-Score	No. of observations
Daily Mail	0.654	0.613	0.654	0.633	987
Newsmax	0.609	0.511	0.609	0.556	241
LA Times	0.576	0.629	0.576	0.601	151
Fox News	0.493	0.509	0.493	0.501	577
ABC News	0.489	0.314	0.489	0.381	703
CNN	0.470	0.417	0.470	0.441	832
CNBC	0.444	0.373	0.444	0.405	410
The New York Times	0.440	0.380	0.440	0.408	426
Wall Street Journal	0.433	0.454	0.433	0.444	593
CS Monitor	0.429	0.668	0.429	0.522	203
Time	0.416	0.418	0.416	0.417	369
BBC	0.403	0.636	0.403	0.492	554
NBC News	0.394	0.448	0.394	0.419	937
Washington Post	0.373	0.316	0.373	0.342	567
CBS News	0.297	0.463	0.297	0.360	517
Reuters	0.295	0.382	0.295	0.332	233
HuffPost	0.285	0.316	0.285	0.300	321
US News	0.249	0.411	0.249	0.309	109
USA Today	0.234	0.303	0.234	0.264	425
Yahoo News	0.196	0.202	0.196	0.199	296

Table A3: Predictive accuracy results on the test set for different news outlets (sorted by accuracy)

Table A3 provides information on our multi-modal model performance at each class level (news outlet) and presents the test performance of the model across various news sources, sorted by accuracy. The model performs best with *Daily Mail* and shows strong, consistent results across other sources, such as *Newsmax* and *CNN*. The model effectively addresses class imbalances by employing a weighted cross-entropy loss, ensuring fair representation for both frequent and rare classes. This is reflected in the balanced precision, recall, and F1 scores across sources with different sample sizes. The weighted loss function helps prevent overfitting to dominant classes, resulting in reliable performance across various news outlets.

E.2 Hypothesis Tests for the Distribution of Visual Polarization

We employ two statistical tests to analyze these differences statistically: the Kolmogorov-Smirnov (K-S) test and one-sample t-tests. The K-S test determines if two samples come from the same distribution. The null hypothesis H_0 states that the two samples are drawn from the same distribution. On the other hand, the one-sample t-test determines if the mean of a sample is different from a known value (zero in this case). The

null hypothesis H_0 states that the sample mean is equal to the known value. A low p-value indicates that we can reject the null hypothesis. The results of these statistical tests are summarized in Table A4.

Politicians	Test	Statistic	P-Value	n
Democratic Politicians	K-S Test	0.7929	0.0000	31728
	Democratic Outlets Mean Test	176.5033	0.0000	15864
	Republican Outlets Mean Test	-130.7972	0.0000	15864
Republican Politicians	K-S Test	0.6859	0.0000	24978
	Democratic Outlets Mean Test	-111.9999	0.0000	12489
	Republican Outlets Mean Test	83.9318	0.0000	12489

Table A4: Summary of Statistical Test Results

The results presented in Table A4 provide strong statistical evidence of visual slant in the portrayal of both Democratic and Republican politicians. For Democratic politicians, the statistical tests confirm a significant and positive visual slant for Democratic-leaning outlets, and a significant and negative visual slant for Republican-leaning outlets. These findings reflect a clear partisan divide, as shown by the significant test results.

For Republican politicians, the table demonstrates an opposing trend. The visual slant is significant and positive for Republican-leaning outlets, whereas it is significantly lower than zero for Democratic-leaning outlets. The statistical significance across all tests underscores the robustness of these patterns. Table A4 highlights how visual elements, such as smiles, are interpreted differently depending on the political context and outlet alignment, illustrating the media's role in amplifying partisan biases.

E.3 Sample Splitting and Relative Prediction Performance

To construct dispersion-based measures of visual polarization discussed in §6.2 and §6.3, we adopt a sample-splitting strategy and divide the training data into two disjoint subsamples, which we use to train two independent outlet-prediction models. This approach avoids mechanically propagating estimation error into dispersion statistics that depend on cross-outlet variation in predicted relative likelihoods (Gentzkow et al., 2019).

Table A5 compares the predictive performance of these two half-sample models to the baseline model trained on the full training sample. As expected, reducing the training sample size leads to a modest decline in predictive accuracy and F1 scores. Relative to the full-sample model, overall accuracy declines from approximately 0.44 to 0.33 and 0.29, respectively, while weighted F1 scores fall from about 0.41 to 0.33 and 0.29.

Despite this loss in performance, both half-sample models continue to perform well above chance and preserve substantial discriminatory power across outlets. High-frequency outlets remain reliably predicted, and cross-outlet differences in predictive strength are broadly consistent with the full-sample model. These results indicate that sample splitting imposes a moderate efficiency cost but leaves the outlet-prediction model sufficiently informative for use in dispersion-based analyses.

Model	Accuracy	Weighted F1	Macro F1
Full-sample model	0.44	0.43	0.42
Half-sample model A	0.33	0.33	0.32
Half-sample model B	0.29	0.29	0.28

Table A5: Comparison of predictive performance across full-sample and half-sample outlet prediction models

E.3.1 Robustness of the Smile Effect under Sample Splitting

Although sample splitting reduces overall predictive accuracy, it does not compromise the model's ability to recover systematic visual effects. In particular, both half-sample models continue to identify a meaningful smile-based visual slant that aligns closely with established non-visual measures of outlet ideology.

To assess this, we compare outlet-level smile-based visual slant under each model to three widely used non-visual slant measures. Table A6 reports both Spearman rank and Pearson correlations. Across all benchmarks, correlations remain sizable and statistically significant under sample splitting, indicating that the smile-based visual signal is stable and not driven by full-sample estimation or overfitting.

	Flexman		Faris		AllSides	
Model	Spearman	Pearson	Spearman	Pearson	Spearman	Pearson
Full-sample model	0.457 (0.043)	0.540 (0.013)	0.657 (0.002)	0.720 (0.000)	0.642 (0.004)	0.760 (0.000)
Half-sample model A	0.520 (0.017)	0.490 (0.020)	0.600 (0.004)	0.610 (0.004)	0.760 (0.002)	0.740 (0.002)
Half-sample model B	0.550 (0.012)	0.540 (0.012)	0.470 (0.036)	0.570 (0.008)	0.430 (0.070)	0.440 (0.060)

Table A6: Correlation between outlet-level smile-based visual slant and external non-visual slant measures. Spearman rank and Pearson correlations are reported with p-values in parentheses.

Taken together, these results show that both half-sample models recover the same systematic smile-based visual pattern identified by the full-sample model. Although trained on reduced samples, each model captures the cross-outlet structure in visual slant that underlies our analysis. This evidence supports the use of predictions from the split-sample models in the dispersion metrics defined in §6.2 and §6.3, where sample splitting is required to avoid mechanically inducing bias. The persistence of the smile-based signal ensures that the dispersion measures reflect genuine cross-outlet variation rather than artifacts of estimation error.

E.4 Individual News Outlet Results

We report outlet-level distributions of the smile-based visual slant and polarization measures defined in §3.2. We compute, for each news outlet, the mean visual polarization separately for Democratic and Republican politicians relative to *Reuters*, which we use as the neutral reference outlet and normalize to zero by construction. Figure A.6 visualizes these outlet-level means.

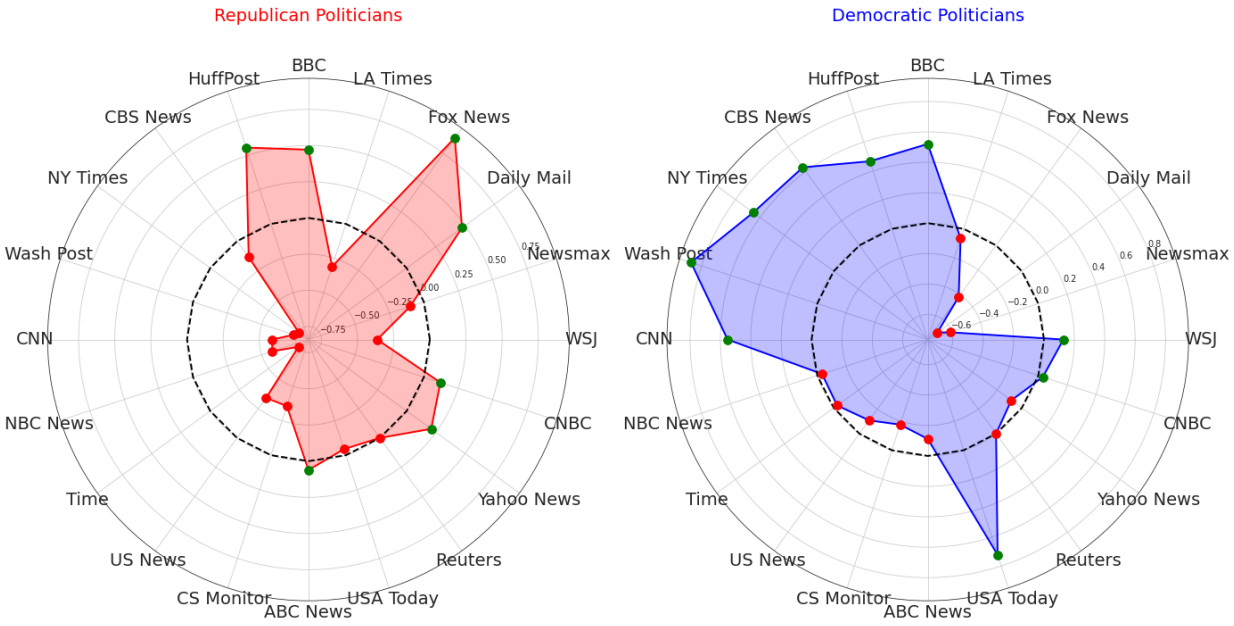


Figure A.6: Radar plots showing the mean polarization measurement for Republican and Democratic politicians across various news outlets. *Reuters* polarization is zero since it serves as the baseline.

We plot the full empirical distribution of polarization estimates for each outlet and party affiliation to characterize dispersion across politicians and article contexts. Figure A.7 shows these distributions and complements the mean patterns in Figure A.6.

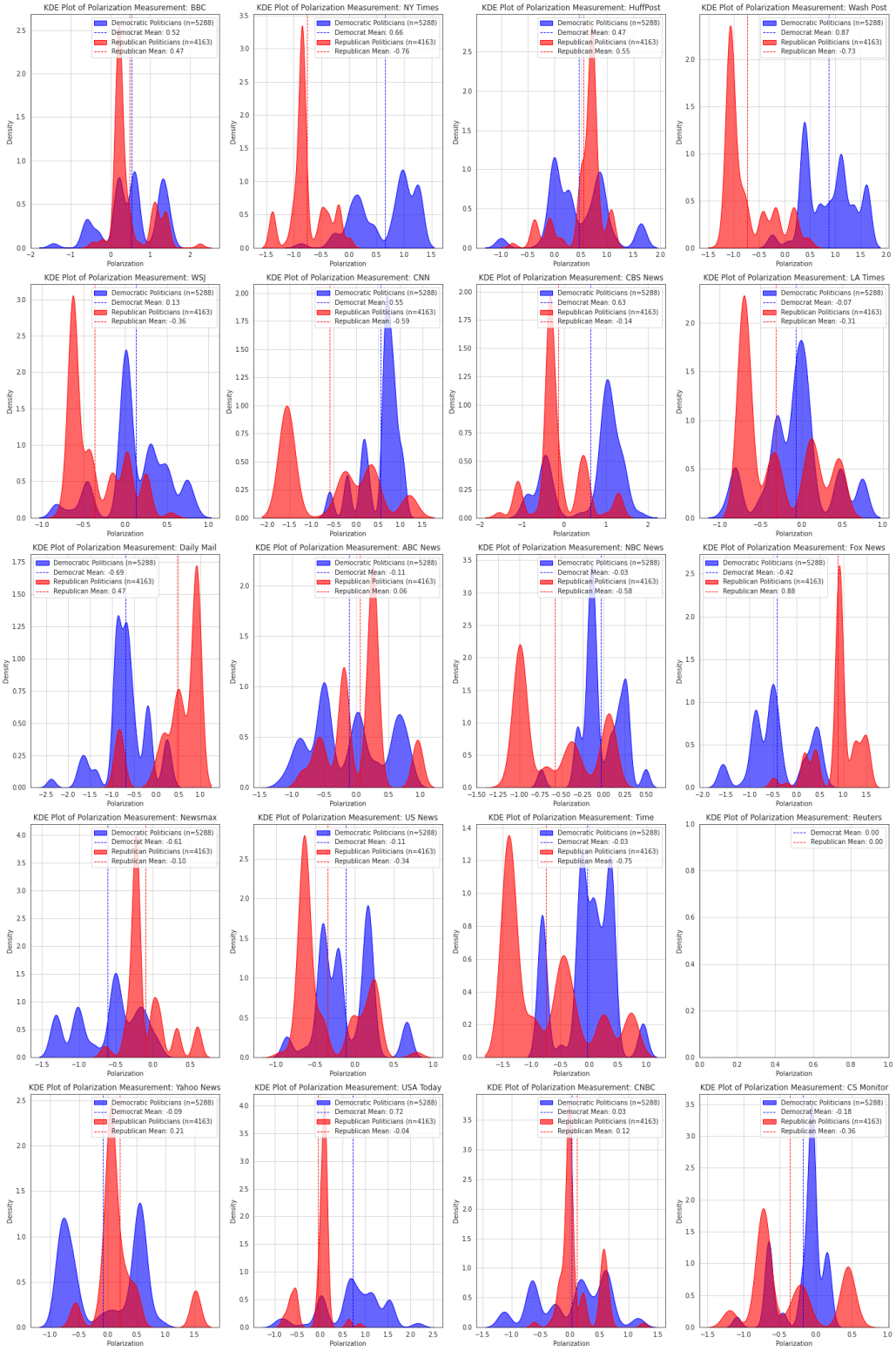


Figure A.7: Histograms of polarization for Democratic and Republican Politicians in each News Outlet

Finally, we report numerical summaries and formal one-sample t-tests against zero for each outlet–party

pair in Table A7. Except for the benchmark outlet *Reuters*, we find statistically significant deviations from zero for nearly all outlet–party combinations. We use these outlet-level results to decompose the aggregate patterns reported in the main text.

News Center	Pol. Side	Mean	Q1	Q3	T-Test Stat	T-Test P-Val	Significant?	Support (n)
BBC	Dem	0.43	-0.62	1.46	56.64	0.00	True	5288
BBC	Rep	0.47	0.21	1.38	65.39	0.00	True	4163
NY Times	Dem	0.77	0.03	1.31	93.60	0.00	True	5288
NY Times	Rep	-0.77	-1.00	-0.18	-116.81	0.00	True	4163
HuffPost	Dem	0.41	-0.01	1.57	54.10	0.00	True	5288
HuffPost	Rep	0.55	-0.09	0.88	96.85	0.00	True	4163
Wash Post	Dem	0.93	0.36	1.96	109.63	0.00	True	5288
Wash Post	Rep	-0.73	-1.08	0.19	-102.14	0.00	True	4163
WSJ	Dem	0.07	-0.36	0.54	14.52	7.39×10^{-47}	True	5288
WSJ	Rep	-0.35	-0.63	0.25	-70.40	0.00	True	4163
CNN	Dem	0.77	0.05	1.50	105.88	0.00	True	5288
CNN	Rep	-0.57	-1.55	0.59	-36.60	1.72×10^{-254}	True	4163
CBS News	Dem	0.53	-0.15	1.41	50.71	0.00	True	5288
CBS News	Rep	-0.22	-1.10	0.40	-27.29	5.18×10^{-151}	True	4163
LA Times	Dem	-0.04	-0.81	0.49	-6.38	1.92×10^{-10}	True	5288
LA Times	Rep	-0.29	-0.71	0.44	-39.95	9.08×10^{-296}	True	4163
Daily Mail	Dem	-0.59	-0.91	0.17	-75.79	0.00	True	5288
Daily Mail	Rep	0.52	-0.11	0.91	69.64	0.00	True	4163
ABC News	Dem	0.09	-0.87	1.92	7.27	4.02×10^{-13}	True	5288
ABC News	Rep	0.04	-0.52	0.27	6.92	5.27×10^{-12}	True	4163
NBC News	Dem	0.01	-0.31	0.39	3.84	1.23×10^{-4}	True	5288
NBC News	Rep	-0.56	-0.99	0.07	-79.57	0.00	True	4163
Fox News	Dem	-0.54	-0.98	0.33	-69.80	0.00	True	5288
Fox News	Rep	0.89	0.17	1.38	110.91	0.00	True	4163
Newsmax	Dem	-0.59	-1.75	-0.09	-81.71	0.00	True	5288
Newsmax	Rep	-0.07	-0.27	0.32	-16.67	1.95×10^{-60}	True	4163
US News	Dem	-0.09	-0.42	0.28	-16.88	2.46×10^{-62}	True	5288
US News	Rep	-0.33	-0.64	0.25	-50.58	0.00	True	4163
Time	Dem	0.23	-0.12	0.82	42.89	0.00	True	5288
Time	Rep	-0.66	-1.40	0.33	-55.10	0.00	True	4163
Reuters	Dem	0.00	0.00	0.00	NaN	NaN	False	5288
Reuters	Rep	0.00	0.00	0.00	NaN	NaN	False	4163
Yahoo News	Dem	-0.03	-0.80	1.40	-3.17	1.54×10^{-3}	True	5288
Yahoo News	Rep	0.22	-0.02	0.50	30.03	1.66×10^{-179}	True	4163
USA Today	Dem	0.95	0.51	1.79	133.66	0.00	True	5288
USA Today	Rep	-0.08	-0.56	0.10	-16.58	7.90×10^{-60}	True	4163
CNBC	Dem	-0.03	-0.88	1.28	-3.06	2.22×10^{-3}	True	5288
CNBC	Rep	0.15	-0.17	0.57	30.50	1.41×10^{-184}	True	4163
CS Monitor	Dem	-0.21	-1.05	0.15	-40.94	0.00	True	5288
CS Monitor	Rep	-0.33	-0.73	0.45	-43.22	0.00	True	4163

Table A7: Statistical analysis of polarization by news outlet and politician affiliation

E.5 Correlation Analysis for Existing Partisanship Scores

We extend our analysis to all 20 outlets, $y^k \in \mathscr{Y}$, to evaluate whether polarization patterns observed in §6.1.1 hold more broadly. Specifically, we calculate outlet-level *CVS* measurement for all outlets following Equation (9). To validate our measure, we compare *CVS* with existing benchmarks, focusing on the *conservative share score* by Faris et al. (2017), AllSides (2024), Flaxman et al. (2016). Figures A.8, A.9, and A.10 visualize these relationships alongside correlation analysis using Pearson and Spearman.

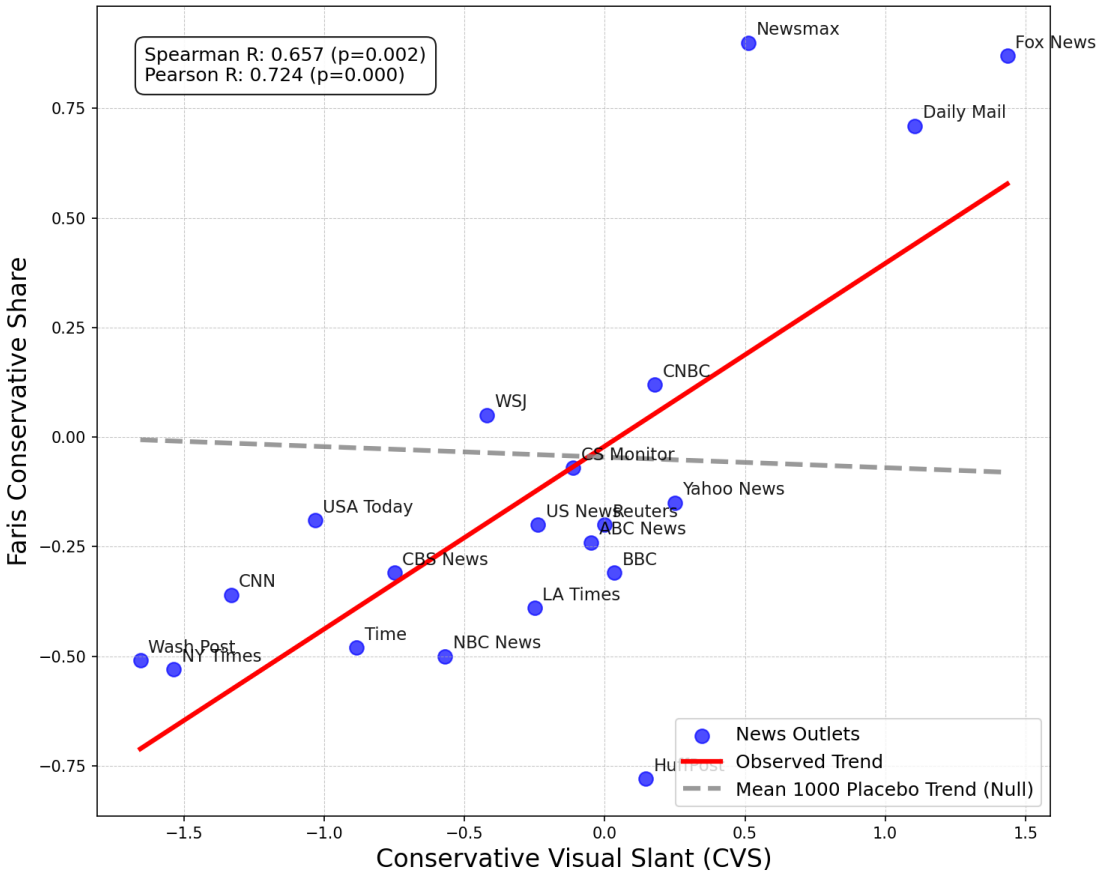


Figure A.8: Correlation between *CVS* and the conservative share score of news outlets (based on Faris et al. (2017)'s Conservative Score).

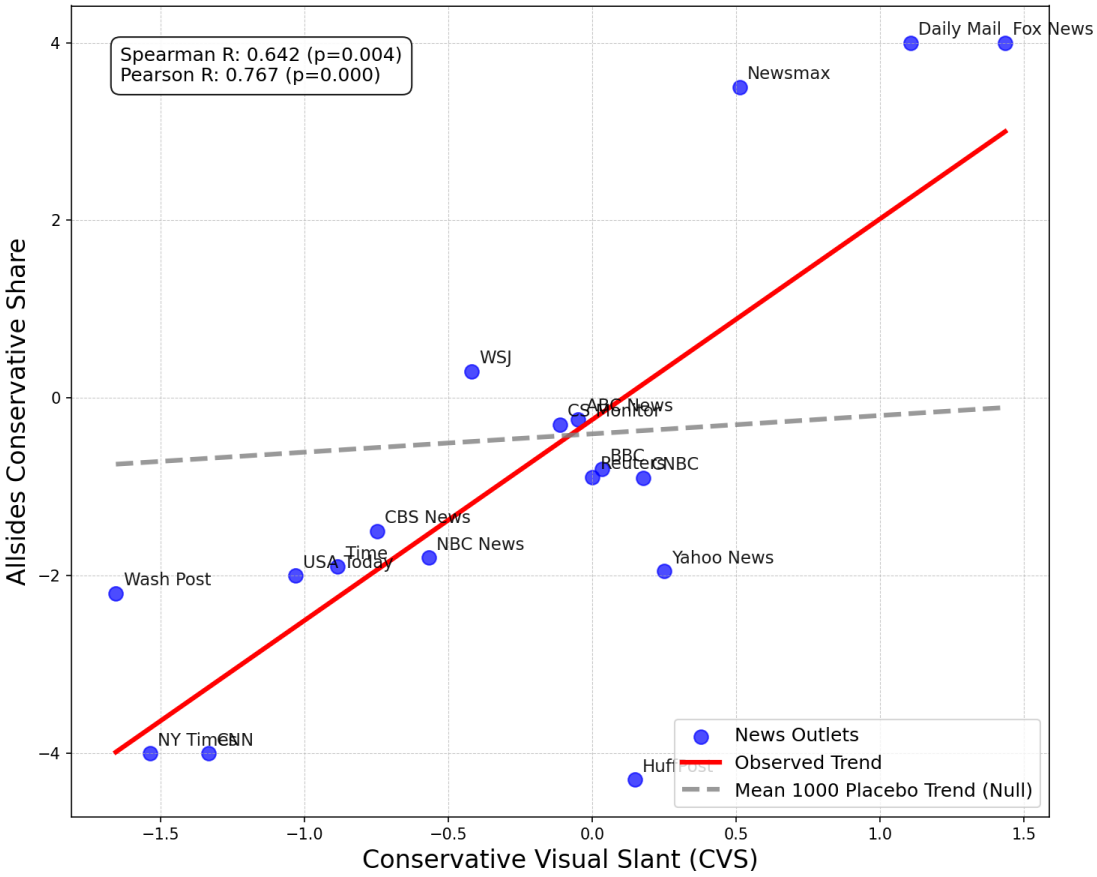


Figure A.9: Correlation between *CVS* and the conservative share score of news outlets (based on AllSides (2024) Conservative Score).

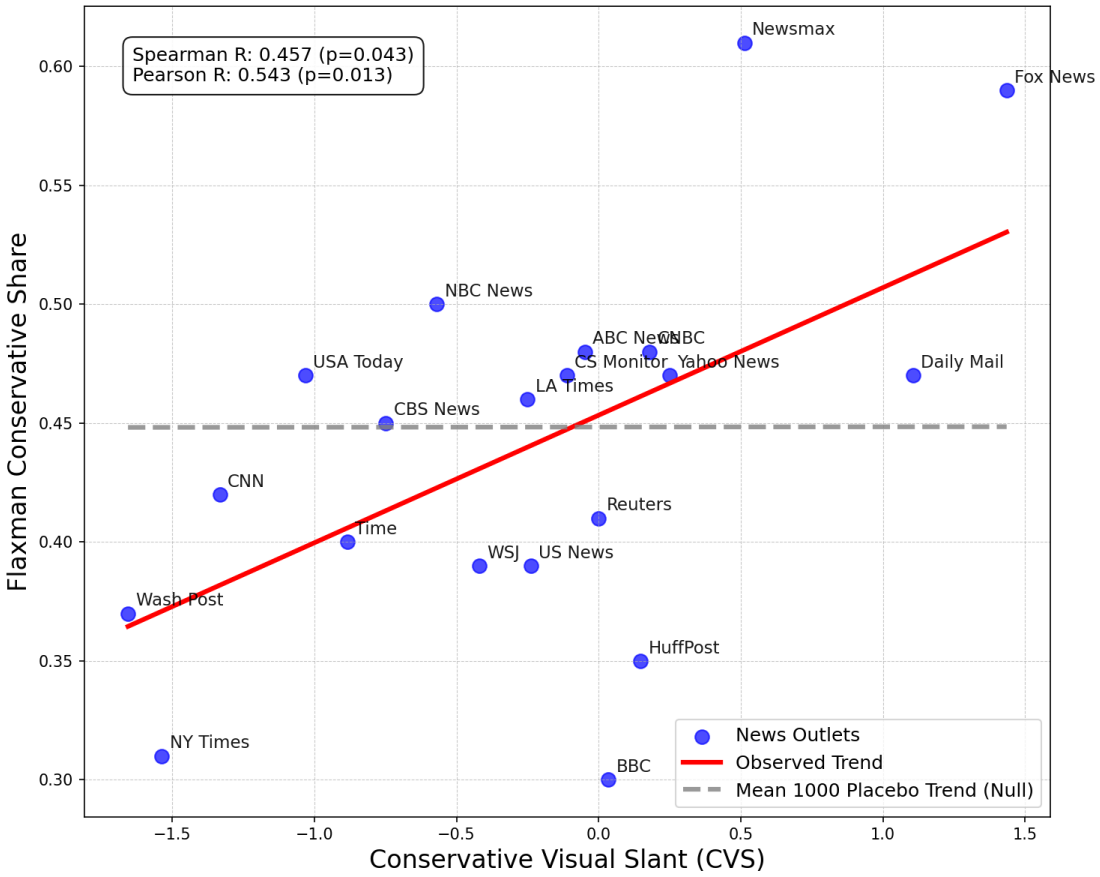


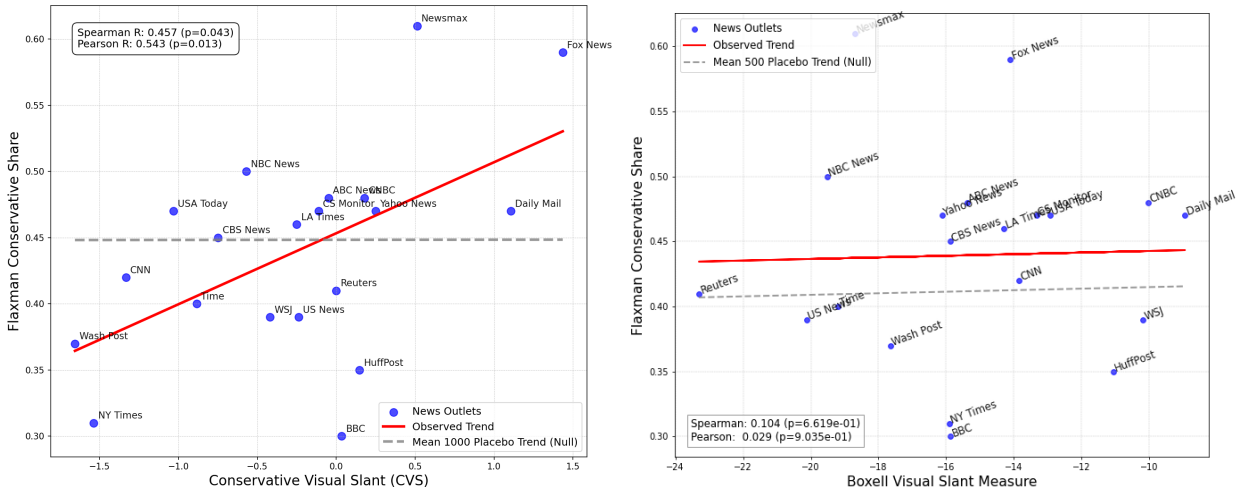
Figure A.10: Correlation between *CVS* and the conservative share score of news outlets. (Based on Flaxman et al. (2016) Conservative Score)

Across all three benchmarks, we observe a significant positive correlation between the conservative share scores and our *CVS* measure. The strength of correlation varies across sources, with AllSides (2024) showing the highest Pearson correlation ($r = 0.760$, $p < 0.001$; Spearman $\rho = 0.642$), followed by Faris et al. (2017) ($r = 0.720$, $\rho = 0.657$), and Flaxman et al. (2016) showing the weakest but still statistically significant association ($r = 0.540$, $\rho = 0.457$). These results support the validity of *CVS* as an independent measure of conservative visual slant, aligning well with established textual and audience-based conservative share scores.

E.6 Comparison with the Two-Stage Approach

In this section, we present a comparison of our estimation procedure with the established two-stage approach for analyzing slant in visual content. For this purpose, we collect the overall visual slant scores from Boxell (2021), which are calculated as the difference between the relative favorability towards Republicans minus Democrats (similar to *CVS* we present in Equation (9), but using the two-stage approach as discussed in §C.2). To evaluate the performance, we use both Pearson correlation analysis and Spearman rank correlation analysis. Each of these analyses is conducted by comparing our estimation procedure and the feature-extraction benchmark against three existing conservative share measures, providing an assess-

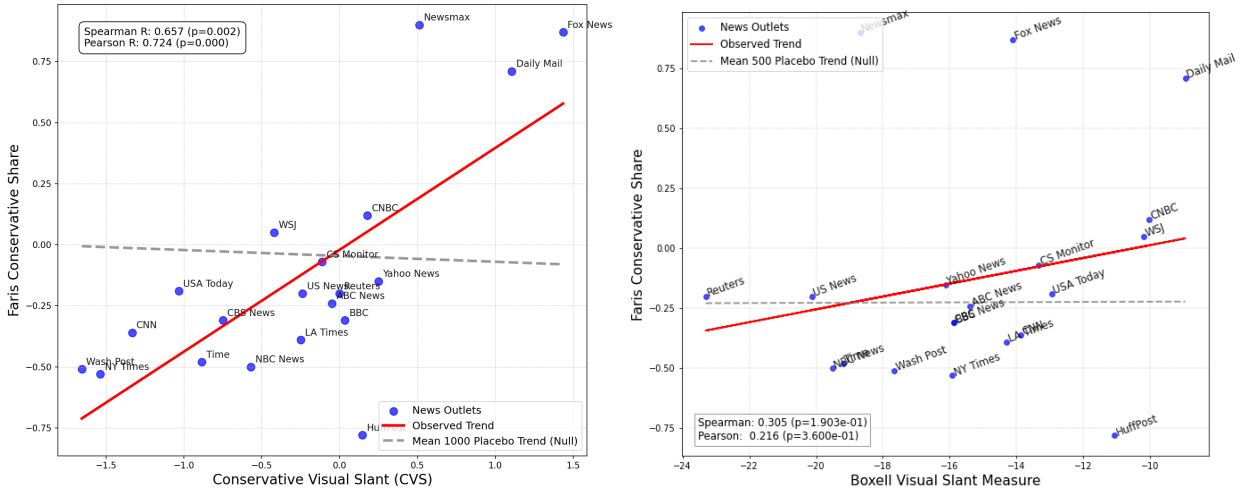
ment of the alignment and ranking accuracy across approaches. Also, to assess whether these correlations arise mechanically, we also report placebo results obtained by randomly reassigning politician party labels and recomputing visual slant. Figures A.11, A.12, and A.13 summarize our findings, providing statistical insights into the relative effectiveness of each method.



(a) *Conservative Visual Slant (CVS)* from our estimation procedure

(b) Visual Slant Measure from Boxell (2021).

Figure A.11: Comparison of *CVS* and visual slant measure by Boxell (2021) against Conservative Share Score from Flaxman et al. (2016).

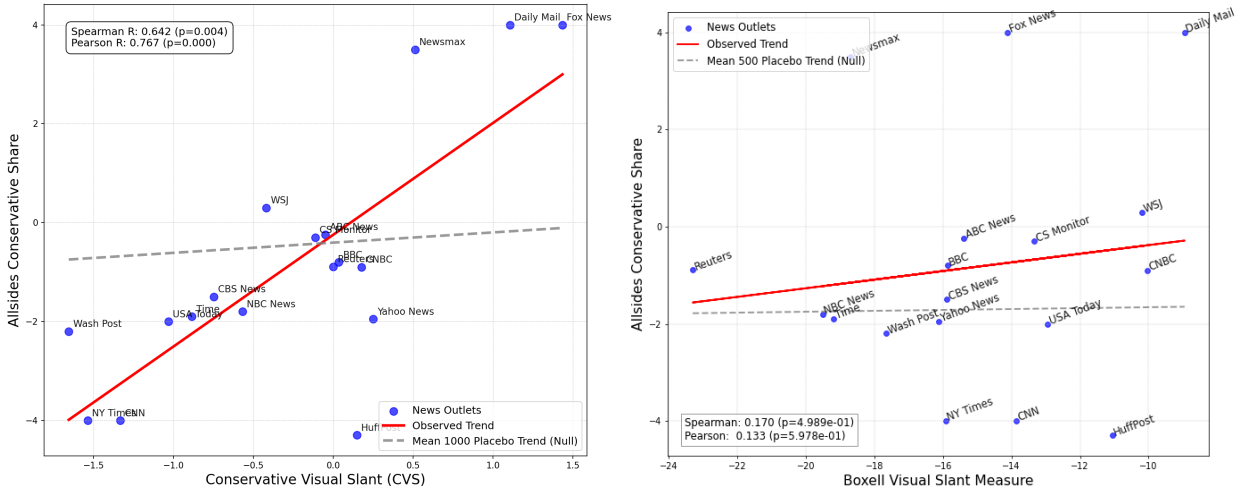


(a) *Conservative Visual Slant (CVS)* from our estimation procedure

(b) Visual Slant Measure from Boxell (2021).

Figure A.12: Comparison of *CVS* and visual slant measure by Boxell (2021) against Conservative Share Score from Faris et al. (2017).

The results clearly show that our estimation procedure outperforms Boxell (2021) two-stage approach in measuring visual slant. Across all three benchmarks, our measure achieves higher Pearson and Spearman



(a) *Conservative Visual Slant (CVS)* from our estimation procedure

(b) Visual Slant Measure from Boxell (2021).

Figure A.13: Comparison of *CVS* and visual slant measure by Boxell (2021) against Conservative Share Score from AllSides (2024).

correlations, demonstrating stronger alignment with established conservative share measures. Additionally, the correlations for our measure are statistically significant, with lower p-values, whereas Boxell (2021) approach often produces weaker and less reliable correlations. our measure also excels in ranking accuracy, as reflected in its consistently higher Spearman rank correlations, confirming that it orders news outlets' visual slant more accurately.

E.7 Visual Slant for Brightness and Sadness

We extend our framework to two additional visual features T, brightness and sadness (frown), to assess whether the visual slant patterns documented for facial expression (smile) generalize to other dimensions of visual presentation. Figure A.14 illustrates how we generate these counterfactual images using our image-transformation procedure. Starting from an original image, we construct counterfactual versions that vary only along a single visual dimension, either inducing a frown to capture sadness or increasing luminance to capture brightness, while holding all other visual attributes fixed.

Using these feature-specific counterfactual images, we compute outlet-level visual slant measures separately for Democratic and Republican politicians, relative to *Reuters* as the reference outlet. Analogous to the main analysis, we then construct a unified outlet-level measure, the Conservative Visual Slant (CVS), by taking the difference between Republican and Democratic slant within each outlet, feature by feature.

Figures A.15 and A.16 relate the resulting CVS measures for brightness and sadness, respectively, to three widely used external benchmarks of media slant: the conservative share scores from Faris et al. (2017), Flaxman et al. (2016), and AllSides (2024). Each panel plots outlet-level CVS against the corresponding conservative share score, with red trend lines summarizing the fitted linear relationship.

For brightness, we do not find statistically significant correlations between CVS and any of the three conservative share benchmarks. While the estimated slopes are generally positive, consistent with brightness

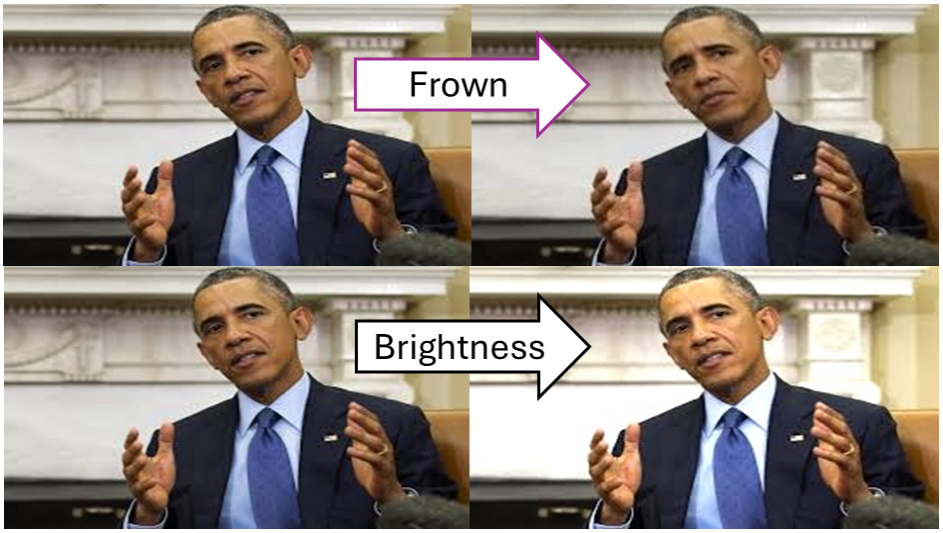


Figure A.14: Illustration of visual features for sadness (frown) and brightness using an example image of Barack Obama.

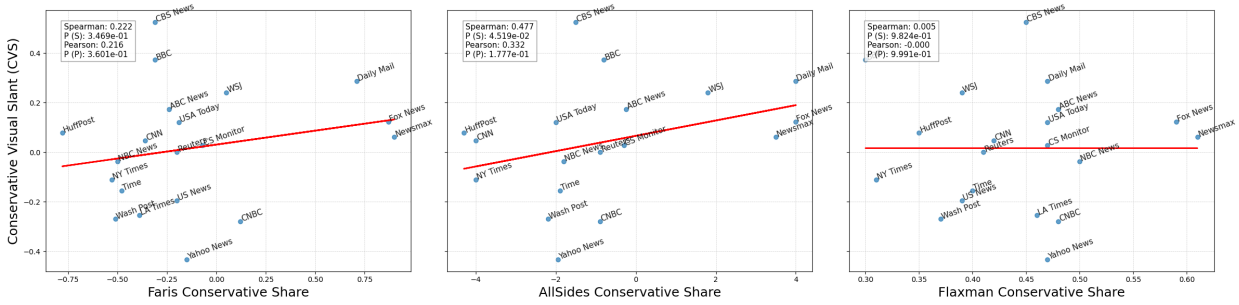


Figure A.15: Conservative Visual Slant (CVS) based on brightness plotted against established conservative share measures.

being interpreted as a favorable visual attribute, the relationships are weak and noisy. This pattern suggests that, unlike smiling, brightness is not used in a consistently partisan manner across outlets. In other words, although brightness may signal positive representation, it does not appear to be a uniformly slanted visual channel through which outlets differentiate between Republican and Democratic politicians.

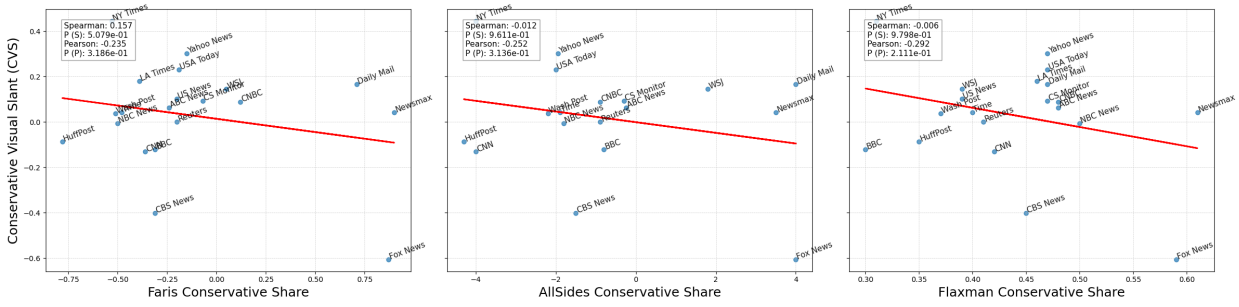


Figure A.16: Conservative Visual Slant (CVS) based on sadness (frown) plotted against established conservative share measures.

For sadness, we observe negative correlations between CVS and the conservative share benchmarks, reflecting the fact that sadness (frown) is a negatively valenced visual feature. However, as with brightness, these correlations are not statistically significant. This indicates that while outlets may differ in their use of negative visual cues, sadness is not systematically deployed in a way that aligns tightly with established measures of ideological slant.

Taken together, these results reinforce the central role of facial expression (smile) as the most salient and consistently slanted visual feature in political imagery. Brightness and sadness exhibit intuitive directional patterns, positive and negative, respectively, but lack the statistical strength and consistency observed for smile.

E.8 Individual Politicians Results

We report additional politician-level results for the smile-based visual polarization measure. We compute outlet-specific polarization for each politician relative to *Reuters* and plot these estimates to illustrate cross-outlet variation in visual portrayal at the individual politician level.

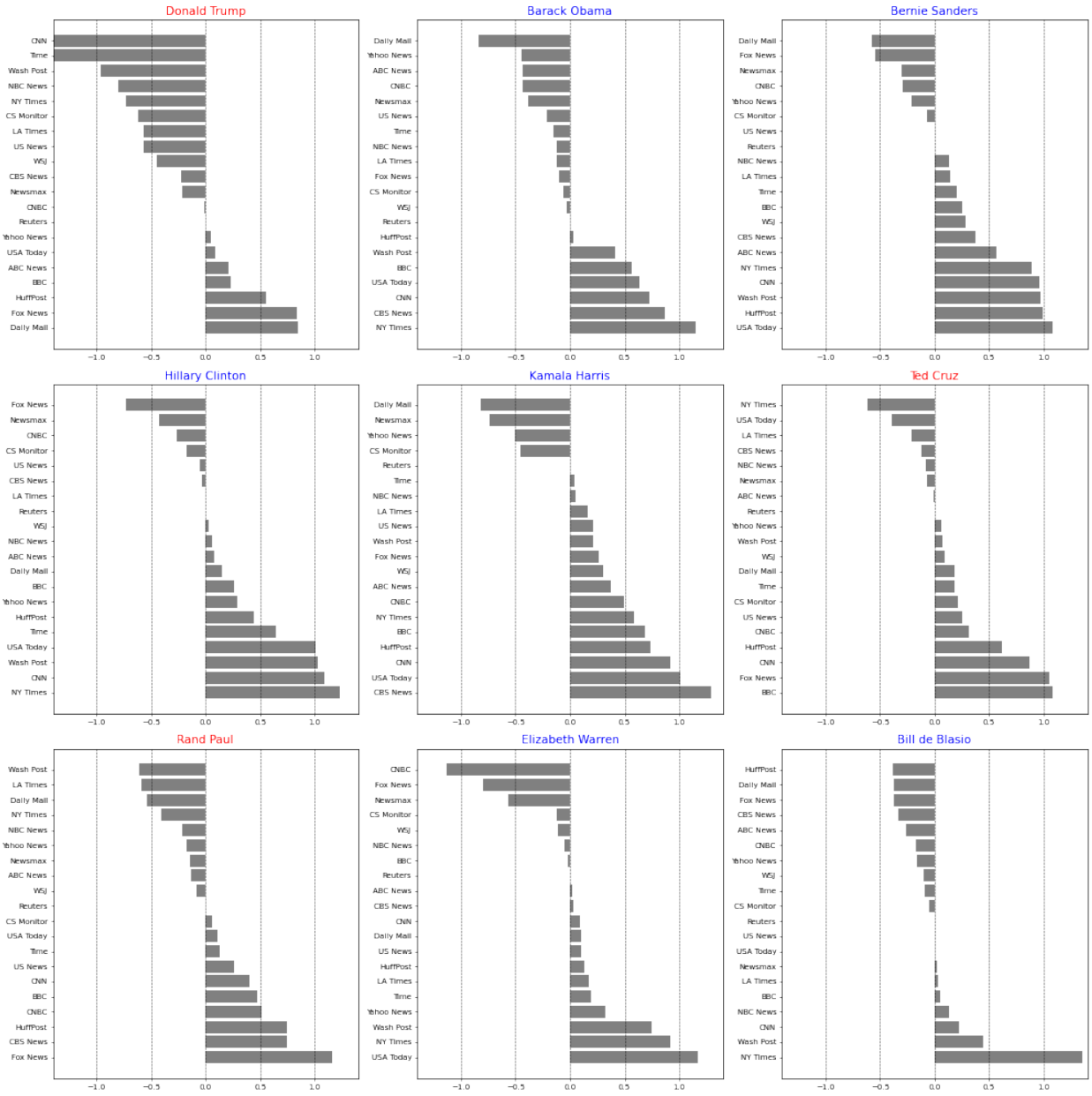


Figure A.17: Bar plot showcasing the polarization across various Politicians, segmented by different News Outlets, Part I

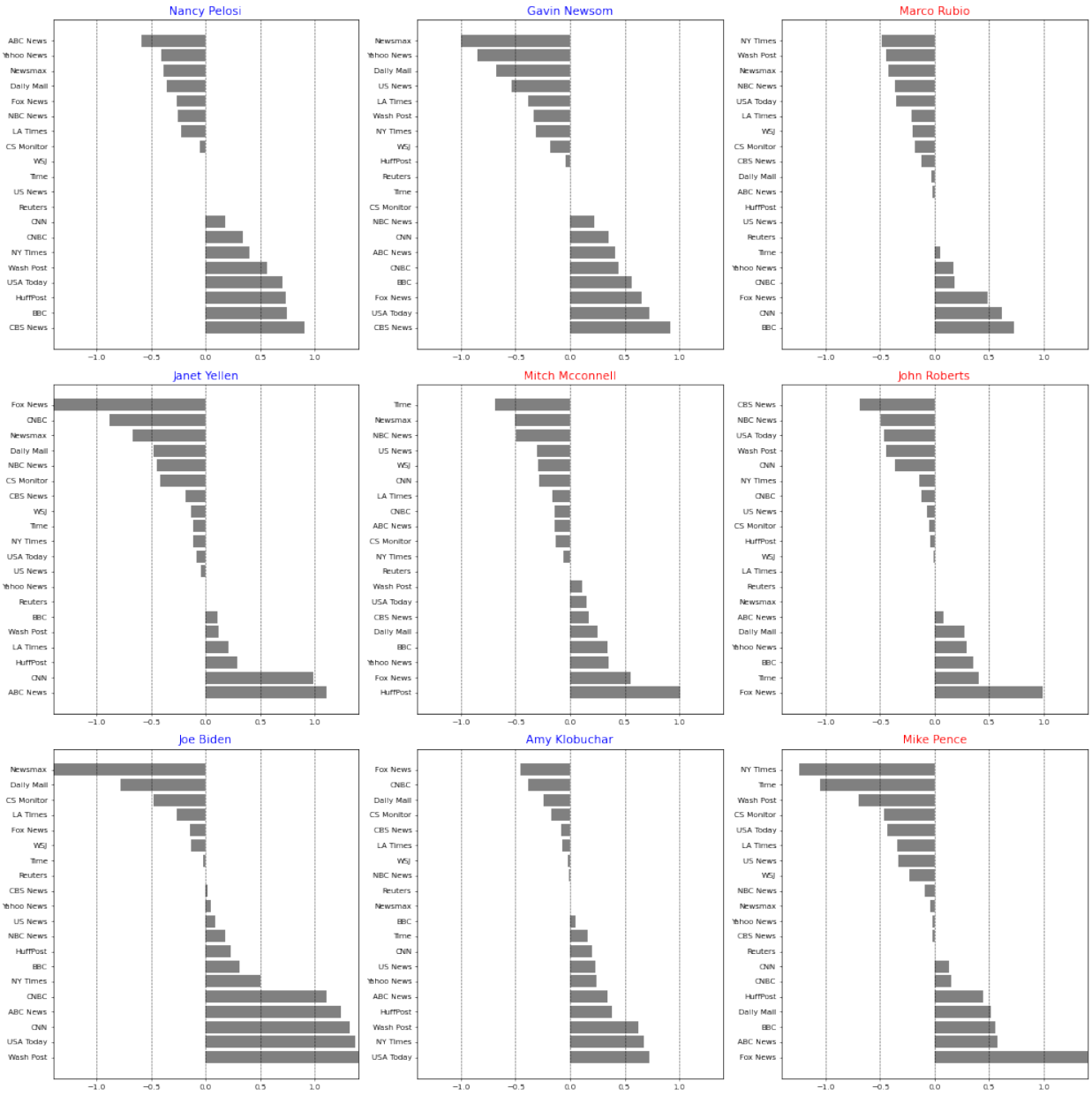


Figure A.18: Bar plot showcasing the polarization across various Politicians, segmented by different News Outlets, Part II

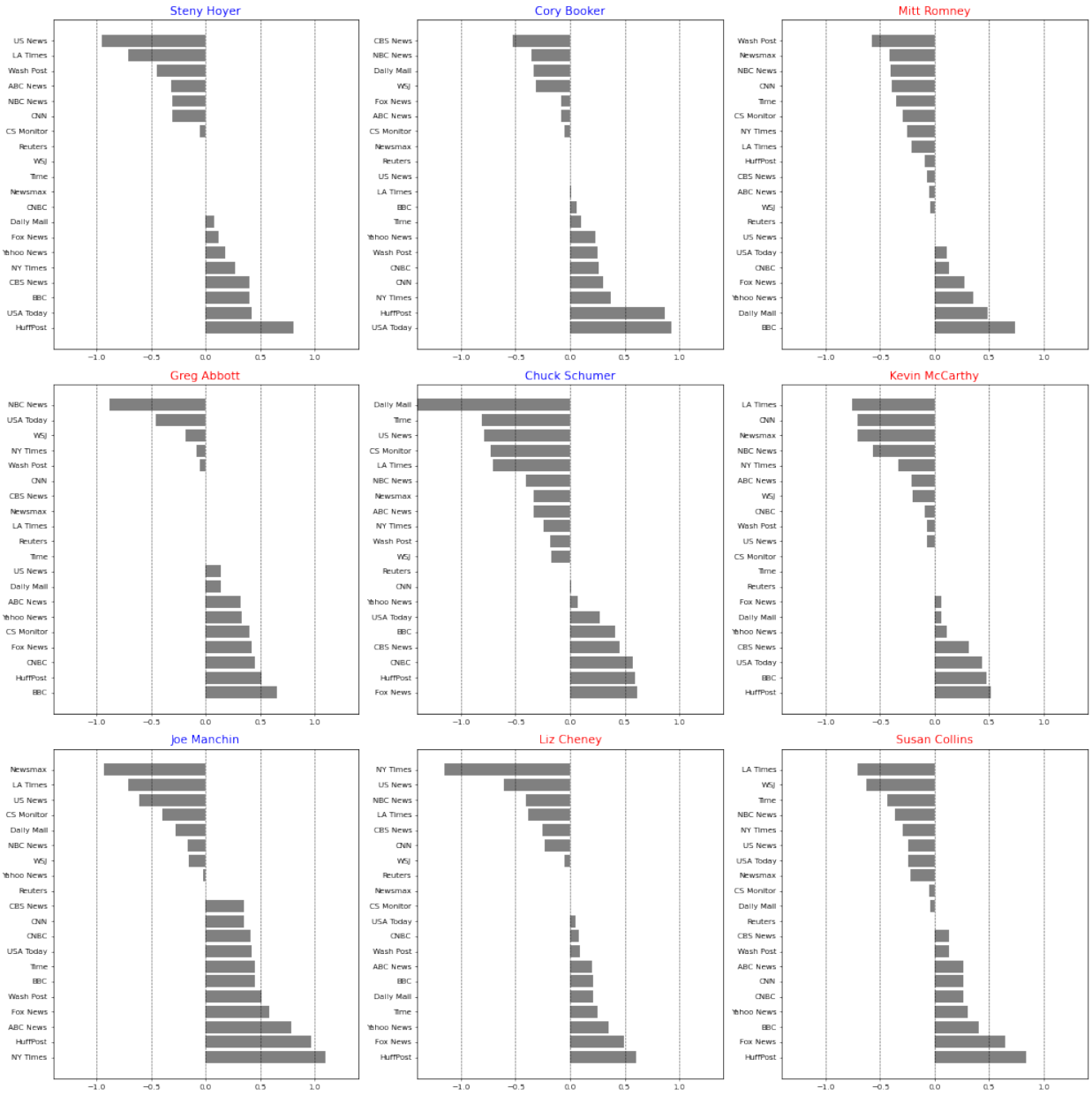


Figure A.19: Bar plot showcasing the polarization across various Politicians, segmented by different News Outlets, Part III

Web Appendix References

Jason Abaluck and Abi Adams-Prassl. What do consumers consider before they choose? identification from asymmetric demand responses. *The Quarterly Journal of Economics*, 136(3):1611–1663, 2021.

AllSides. Media bias rating methods, 2024. URL https://www.allsides.com/media-bias/media-bias-rating-methods.

Piyush Anand and Vrinda Kadiyali. Frontiers: Smoke and mirrors: Impact of e-cigarette taxes on underage social media posting. *Marketing Science*, 2024.

Microsoft Azure. Face api - cognitive services, 2023. URL https://azure.microsoft.com/en-us/services/cognitive-services/face/. Accessed on September, 2023.

Levon Barseghyan, Maura Coughlin, Francesca Molinari, and Joshua C. Teitelbaum. Heterogeneous choice sets and preferences. *Econometrica*, 89(5):2015–2048, 2021.

Anna C Belkina, Christopher O Ciccolella, Rina Anno, Richard Halpert, Josef Spidlen, and Jennifer E Snyder-Cappione. Automated optimized parameters for t-distributed stochastic neighbor embedding improve visualization and analysis of large datasets. *Nature communications*, 10(1):5415, 2019.

Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. " O'Reilly Media, Inc.", 2009.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

Kenneth A Bollen and Walter R Davis. Causal indicator models: Identification, estimation, and testing. *Structural Equation Modeling: A Multidisciplinary Journal*, 16(3):498–522, 2009.

Levi Boxell. Slanted images: Measuring nonverbal media bias during the 2016 election. *Available at SSRN 3837521*, 2021.

Giulia Caprini. Visual bias. 2023.

Christopher T. Conlon and Julie Holland Mortimer. Demand estimation under incomplete product availability. *American Economic Journal: Microeconomics*, 5(4):1–30, 2013.

Thomas H Davenport. How analytics has changed in the last 10 years (and how it's stayed the same). *Harvard Business Review*, 28(08):2017, 2017.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Maximiliano Dvorkin, Juan M Sánchez, Horacio Sapriza, and Emircan Yurdagul. Sovereign debt restructurings. *American Economic Journal: Macroeconomics*, 13(2):26–77, 2021.

Face++. Emotion recognition, 2023. URL https://www.faceplusplus.com/emotion-recognition/. Accessed on September, 2023.

Robert Faris, Hal Roberts, Bruce Etling, Nikki Bourassa, Ethan Zuckerman, and Yochai Benkler. Partisanship, propaganda, and disinformation: Online media and the 2016 us presidential election. *Berkman Klein Center Research Publication*, 6, 2017.

Seth Flaxman, Sharad Goel, and Justin M Rao. Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly*, 80(S1):298–320, 2016.

Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3146–3154, 2019.

Matthew Gentzkow, Jesse M Shapiro, and Matt Taddy. Measuring group differences in high-dimensional choices: method and application to congressional speech. *Econometrica*, 87(4):1307–1340, 2019.

Michelle Sovinsky Goeree. Limited information and advertising in the US personal computer industry. *Econometrica*, 76(5):1017–1074, 2008.

Bruce Hansen. *Econometrics*. Princeton University Press, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press, 2015.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Computing Surveys*, 56(10):1–42, 2024.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Shijie Lu, Dai Yao, Xingyu Chen, and Rajdeep Grewal. Do larger audiences generate greater revenues under pay what you want? evidence from a live streaming platform. *Marketing Science*, 40(5):964–984, 2021.

Michael Luca, Elizaveta Pronkina, and Michelangelo Rossi. Scapegoating and discrimination in times of crisis: Evidence from airbnb. Technical report, National Bureau of Economic Research, 2022.

Paola Manzini and Marco Mariotti. Stochastic choice and consideration sets. *Econometrica*, 82(3):1153–1176, 2014.

Microsoft. Responsible ai investments and safeguards for facial recognition, 2022. URL https://azure.microsoft.com/en-us/blog/responsible-ai-investments-and-safeguards-for-facial-recognition.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Omkar Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *BMVC 2015-Proceedings of the British Machine Vision Conference 2015*. British Machine Vision Association, 2015.

Yilang Peng. Same candidates, different faces: Uncovering media bias in visual portrayals of presidential candidates with computer vision. *Journal of Communication*, 68(5):920–941, 2018.

Radim Řehůřek and Petr Sojka. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.

Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.

Isamar Troncoso and Lan Luo. Look the part? the role of profile pictures in online labor markets. *Marketing Science*, 2022.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Google Vision. Ml kit: Face detection api, 2023. URL https://developers.google.com/ml-kit/vision/face-detection. Accessed on September, 2023.

Yanhao Wei and Nikhil Malik. Unstructured data, econometric models, and estimation bias. SSRN, 2022.

Jeffrey M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge, MA, 2nd edition, 2010.

Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016.