From Pixels to Perspectives: Audience-Specific Visual Sentiment in Political Imagery

Olga Gasparyan
Florida State University
ogasparyan@fsu.edu

Elena Sirotkina
New York University
elena.sirotkina@nyu.edu

Abstract

Visual sentiment analysis often assumes that images have a single sentiment label, but political images may be evaluated differently by different audiences. Work on perspectivist annotation in computational linguistics shows that annotator identities can shape data labels, but automated visual analysis still often treats aggregated annotations as ground truth. This paper argues that such aggregation is appropriate when disagreement reflects random error around a shared construct. When disagreement reflects stable audience-specific interpretations, however, aggregation changes the estimand: it replaces group-specific evaluations with a pooled quantity whose meaning depends on the composition of the annotator pool. We formalize this problem as *perspective collapse* and develop a measurement strategy for estimating audience-specific sentiment from image content using pretrained multimodal image embeddings. Using 960 immigration-related images rated by Democratic and Republican respondents, we show that partisan audiences assign systematically different sentiment to the same images. Pooled labels are most misleading for the images on which partisan disagreement is greatest. We also show that aggregated labels can change downstream conclusions about which visual features are associated with positive or negative sentiment. The point is not that aggregation is always inappropriate. Rather, when disagreement follows theoretically meaningful audience divisions, single-label workflows may produce statistically reliable measures of the wrong quantity.

Key words: visual data, sentiment analysis, annotator disagreements, visual data labeling, deep learning.

Introduction

The rapid advance of visual social media platforms like Instagram, TikTok, and X has made images central to political communication where they actively direct public attention (Domke, Perlmutter and Spratt, 2002; Grabe and Bucy, 2014; Schill, 2012), shape attitudes and sentiments (Anastasopoulos et al., 2024; Bossetta and Schmøkel, 2023; Casas and Williams, 2019), and reinforce stereotypes (Carpinella and Bauer, 2021; Domke, Perlmutter and Spratt, 2002). This effect is especially strong in politics, where the way events are visually depicted can influence public reaction, particularly during pivotal moments like elections (Grabe and Bucy, 2014) or periods of heightened issue salience (de Lima-Santos et al., 2023; Casas and Williams, 2019). Consequently, as political scientists increasingly deploy automated computer vision tools to analyze massive image corpora, identifying valid, reliable methods to measure the latent sentiment of political imagery has become a central measurement problem.

Conventional visual sentiment analysis workflows, whether relying on manual human coding or automated architectures, rest on the assumption that annotators are interchangeable units whose disagreements represent random error to be smoothed away through aggregation. While this assumption is defensible for low-stakes or non-political stimuli, it becomes problematic when applied to polarizing political content. When evaluating highly salient visual themes such as mass protests, migrant caravans, or partisan elites, an annotator's decoding of an image is intrinsically bound to their prior attitudinal dispositions (Williams et al., 2026). Here, coder disagreement is not stochastic noise; it is systematic, structurally driven by ideology, and substantively consequential. A similar argument motivates the "perspectivist" paradigm in computational linguistics, where scholars argue that annotator identities systematically shape labels assigned to subjective content and should therefore be modeled rather than averaged away (Cabitza, Campagner and Basile, 2023; Sap et al., 2022; Davani, Díaz and Prabhakaran, 2022). Yet visual communication research continues to rely heavily on aggregate consensus voting, replac-

ing perspective-specific interpretations with an often misleading single point estimate.

The implications of this assumption extend beyond annotation practice and directly affect measurement. Historically, automated content-analytical workflows have treated an image's latent sentiment as a fixed, objective property embedded within its pixels or constituent objects (Borth et al., 2013; You, Jin and Luo, 2017; Song et al., 2018; Ortis, Farinella and Battiato, 2020). These architectures implicitly assume that complex visual scenes transmit a uniform sentiment signal that can be consistently recovered by either human or algorithmic decoders. Yet recent work in political science challenges this premise, demonstrating that identical politically sensitive images frequently fail to evoke consistent attitudinal or affective responses across ideologically diverse audiences (Madrigal and Soroka, 2023; Williams et al., 2026). We argue that a substantial portion of this apparent inconsistency originates not from audience volatility but from the aggregation procedures used to construct sentiment labels. When annotators with opposing political priors assign contradictory sentiment scores to the same image, traditional aggregation metrics collapse these evaluations into a single consensus measure. In doing so, they erase the distinct audience reactions that make political imagery strategically meaningful and behaviorally consequential in the first place.

Increasing the size or diversity of the annotation pool does not solve this measurement problem. Larger panels reduce random individual-level error, but they do not remove disagreement that is systematically structured by annotators' characteristics (Benoit et al., 2016; Williams et al., 2026). When annotators with different political priors evaluate the same image, aggregation estimates a consensus score whose meaning largely depends on the composition of the annotator pool. While recent work has called for greater attention to subgroup variation in annotation processes (Williams et al., 2026), scalable approaches for preserving audience-specific evaluations remain limited. We address this gap by developing a framework that estimates partisan-specific sentiment rather than collapsing these evaluations into a single consensus label.

Existing approaches implicitly assume that disagreement should be reduced. We take the opposite view. If disagreement reflects meaningful differences in how audiences interpret political imagery, then the goal of measurement should not be to eliminate disagreement but to preserve it. Rather than treating sentiment as a single latent property of an image, we conceptualize sentiment as a distribution of evaluations that may vary systematically across politically relevant audiences. The challenge is therefore not merely to predict consensus labels, but to recover and represent heterogeneous evaluations without collapsing them into a single aggregate score.

To address this challenge, we develop a perspectivist machine-learning framework that estimates audience-specific sentiment evaluations from visual content. Using a corpus of immigration images and sentiment annotations from Democratic and Republican respondents, we demonstrate that conventional aggregation procedures systematically compress politically contested images toward the midpoint of the sentiment scale, obscuring meaningful disagreement. Our framework instead estimates partisan-specific sentiment functions from frozen multimodal image representations, allowing identical visual stimuli to receive different evaluations depending on the audience interpreting them. We show that information loss from aggregation is concentrated among the images that generate the greatest political disagreement, while politically meaningful cleavages such as partisanship and immigration attitudes recover substantially more information than arbitrary annotator splits. These findings suggest that visual sentiment is often not a fixed property of an image but a perspective-dependent evaluation that varies systematically across audiences.

Perspective-Dependent Visual Sentiment

Traditional approaches to visual sentiment analysis conceptualize sentiment as a latent property of the image itself — a stable emotional signal embedded within visual con-

tent that can be recovered through either human annotation or automated classification (Borth et al., 2013; Ortis, Farinella and Battiato, 2020; Zhao et al., 2021). Although these approaches differ substantially in methodology, they share a common measurement assumption: that images possess a single underlying sentiment that can be estimated through increasingly accurate observation. Under this view, disagreement among annotators is interpreted primarily as measurement error, while aggregation serves as a mechanism for recovering the image's true sentiment value.

The assumption of a single recoverable sentiment becomes less convincing when applied to political imagery. Appraisal theories of emotion emphasize that affective responses emerge through the interaction between external stimuli and the goals, beliefs, and prior experiences of the observer (Lazarus, 1991; Scherer, 2001). Similarly, research on motivated reasoning demonstrates that political predispositions shape not only how individuals evaluate information but also how they perceive it in the first place (Kunda, 1990; Leong et al., 2019). Recent work in political communication suggests that identical immigration images frequently evoke different reactions across partisan audiences (Madrigal and Soroka, 2023; Williams et al., 2026). In such contexts, variation in sentiment evaluations may reflect systematic differences in interpretation rather than random coding error.

This insight closely parallels recent developments in computational linguistics, where scholars have increasingly questioned the assumption that annotator disagreement should always be eliminated through consensus-building procedures. Perspectivist approaches argue that disagreement is often meaningful because annotators bring distinct social identities, experiences, and political beliefs to the labeling process (Sap et al., 2022; Davani, Díaz and Prabhakaran, 2022; Cabitza, Campagner and Basile, 2023). From this perspective, aggregation can obscure important dimensions of variation by replacing multiple valid interpretations with a single consensus label. Rather than viewing disagreement as a nuisance, perspectivist frameworks treat it as information that should be modeled

directly.

The implications extend beyond annotation practice and directly affect measurement validity. Aggregation is highly effective when disagreement reflects random fluctuations around a shared latent quantity (Benoit et al., 2016). However, when disagreement is structured by stable interpretive cleavages, aggregating may generate indicators that correspond to no theoretically meaningful audience. In the language of measurement validity, the resulting indicator no longer maps cleanly onto the underlying concept it is intended to represent (Adcock and Collier, 2001). Disagreement may itself reveal important features of the phenomenon being measured rather than uncertainty about its true value (Aroyo and Welty, 2015).

Although our empirical setting uses continuous sentiment ratings and therefore illustrates aggregation through averaging, the same measurement problem applies to other consensus procedures such as majority vote or adjudication (more common in the categorical labeling). We argue that majority voting also does not eliminate perspective collapse; it simply resolves disagreement by selecting the interpretation favored by more annotators in the labeling pool. But when disagreement reflects systematic audience differences, a majority-vote label will just reflect the perspective that is more common in the annotator pool, not necessarily the label that is more valid. Our critique is therefore not specific to arithmetic means, but generally to consensus-label workflows that replace structured disagreement with a single target label.

Standard intercoder reliability diagnostics (such as Cohen's κ, Krippendorff's α, and related chance-corrected agreement coefficients) are useful for assessing whether annotators assign labels consistently, but they do not fully resolve this problem. They can reveal that coders disagree, but they cannot exactly distinguish random coding errors from structured disagreement among interpreters. In many annotation workflows, low reliability is treated as evidence that the coding scheme should be clarified, ambiguous cases should be resolved, or a majority label should be imposed (Sabou et al., 2014; Yan

et al., 2014; Äyräväinen, Hinds and Davidson, 2025). These steps may be appropriate when disagreement reflects confusion or noise. But when disagreement reflects systematic audience differences, the same procedures can remove the variation that is theoretically meaningful. Thus, we believe that even though reliability diagnostics are necessary, they might not be sufficient for determining whether aggregation preserves or changes the construct being measured.

As a result, we distinguish between two different sources of annotator disagreement. In the first, disagreement reflects random measurement error around a shared latent sentiment. Under these conditions, aggregation improves measurement by averaging away idiosyncratic noise. In the second, disagreement reflects systematic differences in how audiences interpret political imagery. Here, aggregation no longer recovers a common underlying quantity. Instead, it produces what we term *perspective collapse*: the reduction of heterogeneous audience evaluations into a single consensus score. Under these conditions, politically polarizing images become increasingly likely to appear neutral, not because audiences converge in their interpretations, but because opposing evaluations offset one another.

To formalize this distinction, let Y_{ijg} denote the sentiment rating assigned to an image i by annotator j belonging to audience group g. Under the conventional annotation framework, each observed rating is assumed to be a noisy realization of a single latent image sentiment,

$$Y_{ijg} = \theta_i + \varepsilon_{ijg} \tag{1}$$

where θ_i denotes the image's underlying sentiment and ε_{ijg} is an unobserved, stochastic annotator-level error. Under the standard assumptions that these errors are independent and have mean zero $\mathbb{E}[\varepsilon_{ijg}] = 0$, disagreement reflects random measurement error around a shared latent quantity. Aggregation therefore improves measurement by averaging away idiosyncratic deviations, and as the number of annotations increases, the consensus estimator converges to the underlying image sentiment.

Perspective collapse arises precisely when the assumption of a shared latent sentiment no longer holds. Rather than reflecting random deviations around a common image-level evaluation, systematic disagreement indicates that different audiences assign distinct latent sentiments to the same image. In this case, the observed evaluations should be represented as

$$Y_{ijg} = \theta_{ig} + \varepsilon_{ijg} \tag{2}$$

where θ_{ig} denotes the audience-specific latent sentiment.

Under this model, the conventional consensus estimator no longer converges to a single latent image sentiment. Instead, as the number of annotations increases, it converges to a weighted average of audience-specific evaluations (*aggregated sentiment estimate*),

$$\bar{Y}_i^* = \sum_g \pi_g \theta_{ig} \tag{3}$$

with π_g being the proportion of annotators belonging to audience group g. Consequently, consensus aggregation does not recover the sentiment of any particular audience. Rather, it estimates a mixture of heterogeneous audience evaluations whose composition depends on the distribution of annotators.

The key distinction is that the object of estimation changes rather than the statistical properties of the estimator. Under Equation (1), aggregation improves both the precision and validity of measurement because all annotators provide noisy observations of the same latent sentiment. Under Equation (2), aggregation continues to improve statistical precision, but only with respect to the weighted average in Equation (3). It does not recover the audience-specific latent sentiments that have been combined through aggregation. Perspective collapse therefore reflects a change in the estimand rather than a failure of the estimator.

We demonstrate two implications of this framework using simulation analysis. First, we show that different consensus rules distort structured disagreement in different ways.

Second, increasing the number of annotators improves the precision of the pooled label without making it a valid measure of either of the audience-specific sentiments.

The formal result is easiest to see for aggregation via arithmetic means, but this measurement problem generalizes for other aggregation rules. Many annotation workflows produce categorical labels and rely on majority vote rather than averaging. Majority vote avoids so called neutralization effect, but it does not avoid perspective collapse: when audience groups assign different categories to the same image, by construction majority vote selects the category favored by the larger group in the annotation pool. The estimand therefore remains a property of such an annotation pool rather than of either audience-specific evaluation.

Table 1: Consensus rules under different levels of systematic disagreement.

		Consensus label		
Rule	Gap	Neutral	Matches A	Matches B
Mean	Low (0.5)	100%	100%	100%
	Mid (2.0)	100%	0%	0%
	High (3.5)	99%	1%	0%
Majority vote	Low (0.5)	25%	25%	25%
	Mid (2.0)	1%	98%	1%
	High (3.5)	0%	100%	0%

Note: The table reports a simulation with two audience groups: majority group A (70% of raters) and minority group B (30%). Each image receives 42 ratings from group A and 18 ratings from group B (2,000 images per disagreement level; within-group rating SD $= 1.2$). The disagreement gap is calculated as $\theta_A - \theta_B$. Ratings are collapsed into three sentiment categories: negative (≤ 3), neutral ($(3, 5)$), and positive (≥ 5). Cells report the share of images for which the single consensus label is neutral, matches group A's true category, or matches group B's true category.

Table 1 shows that consensus rules do not eliminate structured disagreement; they collapse it in different ways. When disagreement is low, both audiences occupy the same sentiment category, so a single consensus label can match both. At higher levels of systematic disagreement, however, the two audiences occupy different sentiment categories, so no single label can recover both audience-specific evaluations. Mean aggregation resolves this conflict by moving opposing evaluations toward the neutral category, often matching neither audience. Majority vote avoids this false neutrality, but it does so by

assigning the label to the numerically dominant audience.

The second implication is related to the size of the annotators pool. Equation (1) implies that increasing the size of the annotation pool can average out individual-level noise. Adding annotators reduces sampling error and that way makes the estimated mean more precise. It cannot, however, resolve disagreement that arises because different audiences systematically evaluate the same image differently, as we show in Equation (2). When the pooled label estimates an average of audience-specific sentiments, increasing the number of annotators makes that average increasingly stable, but it does not make the pooled label converge to either audience's sentiment.

We illustrate this distinction with a simulation of 3,000 images whose latent sentiments differ systematically across two audiences. For each annotation-pool size, we compare an audience-specific label built from audience A's ratings with a pooled label derived from ratings from both audiences. We then evaluate both the conventional reliability of the pooled label and the accuracy with which each label recovers audience A's latent sentiment.

Table 2: Annotation pool size and label recovery.

			RMSE for θ_A	
Gap	k	Pooled reliability	Pooled label	Audience label
Large ($\delta = 3$)	5	0.509	1.546	0.523
	30	0.911	1.510	0.259
	90	0.970	1.507	0.210
Moderate ($\delta = 1$)	5	0.808	0.622	0.499
	30	0.964	0.528	0.216
	90	0.988	0.507	0.128

The simulation includes 3,000 images measured on a 1–7 sentiment scale. For each image i, we draw a sentiment midpoint $\mu_i \sim U(2.5, 5.5)$. The two audience-specific latent sentiments are then placed symmetrically around this midpoint, such that $\theta_{Ai} = \mu_i + \delta/2$ and $\theta_{Bi} = \mu_i - \delta/2$. Dividing the gap equally around μ_i ensures that the two audience sentiments differ by exactly δ, while their equally weighted average remains μ_i. The distribution range $U(2.5, 5.5)$ ensures that both latent sentiments remain within the 1–7 response scale even when the simulated audience gap is large ($\delta = 3$). For each image, k independent ratings are generated from each audience around its respective latent sentiment, with within-audience rating SD =1.2. The audience-A label is the mean of its k ratings, whereas the pooled label is the mean of all $2k$ ratings. Pooled reliability is measured using ICC(1,2k), which estimates the reliability of a mean constructed from the $2k$ pooled ratings. The RMSE columns report each label's error in recovering audience A's latent sentiment, θ_{Ai}.

The results in Table 2 demonstrate that as the number of annotators increases, the

pooled label becomes highly reliable: its reliability reaches 0.970 in case of a large audience disagreement ($\delta = 3$) and 0.988 in case of a moderate audience disagreement ($\delta = 1$). Yet its error in recovering audience A's sentiment does not disappear. Under the large gap, pooled-label RMSE remains approximately 1.5 regardless of whether each audience number of annotators increases from 5 to 90. Under the moderate gap, it approaches 0.5. These values correspond to the structural distance, $\delta/2$, between audience A's sentiment and the equally weighted pooled midpoint. By contrast, the audience-specific label becomes increasingly accurate as k grows since additional ratings average out individual random noise around the correct audience-specific target. Thus, increasing annotation pools improve estimation of the pooled sentiment, but they cannot make that pooled sentiment a valid measure of an audience-specific construct.

This framework yields several observable implications which do not depend on a particular computational model or estimation procedure. Instead, they follow directly from the distinction between random annotation error and systematic audience-specific interpretation. These implications allow us to identify conditions under which consensus aggregation provides a valid measure of visual sentiment and those under which it obscures theoretically meaningful variation.

Implication 1. Information loss from aggregation increases with systematic disagreement.

The amount of information loss through aggregation should increase as systematic disagreement between audiences grows. When audience-specific latent sentiments are similar, the aggregated estimate closely approximates each group's evaluation because averaging removes only random annotation error. As audience-specific evaluations diverge, however, the aggregated estimate becomes progressively less representative of any individual audience. Perspective collapse therefore intensifies with the degree of systematic

disagreement, even though the statistical precision of the aggregated estimate continues to improve.

Implication 2. Information loss from aggregation is concentrated among images that elicit systematic disagreement across audiences.

Consensus aggregation does not distort all images equally. Images that evoke similar evaluations across audiences remain well represented by the aggregated estimate because averaging primarily removes random annotation error. By contrast, images that generate systematically different evaluations across audience segments experience substantially greater information loss. Therefore, these images are most susceptible to perspective collapse, as opposing evaluations offset one another and produce estimates that appear more neutral than the underlying audience-specific sentiments.

Implication 3. Theoretically meaningful audience partitions recover more information than arbitrary or consensus aggregation.

Disaggregating annotations should recover information only when audience partitions correspond to the sources of interpretive heterogeneity (meaningful societal cleavages). Partitions based on theoretically relevant characteristics, such as gender, race, ideology and political beliefs, should therefore produce substantially more informative estimates than either a single consensus model or arbitrary partitions of annotators. Recovering audience-specific sentiment thus depends not simply on estimating multiple models, but on identifying dimensions along which systematic differences in interpretation arise.

Empirical Strategy

Empirically testing this framework requires data that preserve systematic differences in sentiment across audiences. We constructed such a dataset using politically salient immigration imagery labeled separately by Democratic and Republican annotators. These audience-specific labels serve as our empirical targets.

We choose the topic of immigration as our empirical setting because it is among the most politically polarizing issues in contemporary American politics (Card et al., 2022; Ollerenshaw and Jardina, 2023). Partisan divisions over immigration have remained both theoretically well established and empirically persistent, shaping citizens' evaluations of immigration policy, migrants, and border enforcement. These stable differences in political predispositions make immigration an especially appropriate setting for examining whether audience-specific beliefs systematically influence visual sentiment.

Many immigration images do not contain a single, unambiguous emotional meaning. Unlike images depicting universally recognizable affective cues, such as smiling faces, natural disasters, or graphic violence, the interpretation of immigration imagery often depends on how viewers understand the broader social and political context. The same image of migrants walking along a road, waiting at a border crossing, or traveling with family members may evoke compassion, concern, or indifference depending on the observer's prior beliefs about immigration. This ambiguity makes immigration imagery particularly well suited for evaluating whether visual sentiment reflects a stable property of the image or instead depends systematically on the audience evaluating it.

We assembled a corpus of immigration-related images from publicly available posts by major U.S. news organizations on X (formerly Twitter) and supplemented this collection with images obtained from Getty Images, a primary source of visual content used in news reporting on immigration. These sources provide a broad sample of the visual material through which Americans routinely encounter immigration in the news. After removing duplicate and near-duplicate images, the final corpus consists of 960 unique im-

ages spanning a diverse range of immigration-related events, settings, and visual frames.

We collected sentiment evaluations from 5,533 unique respondents who self-identified as either Democrats or Republicans. Participants were randomly assigned ten images and rated each image on a seven-point sentiment scale.[1] On average, each image was evaluated by approximately 30 Democratic and 30 Republican respondents, providing sufficient observations to construct reliable audience-specific image-level sentiment labels.

For each image, we compute separate sentiment scores by averaging ratings within each partisan group. These Democratic and Republican image-level means constitute the primary outcomes throughout the analysis. For comparison, we also construct a conventional consensus sentiment score (aggregated sentiment score) by averaging across all annotators irrespective of party. This pooled measure represents the standard annotation practice in visual sentiment analysis and provides the benchmark against which we evaluate perspective collapse.

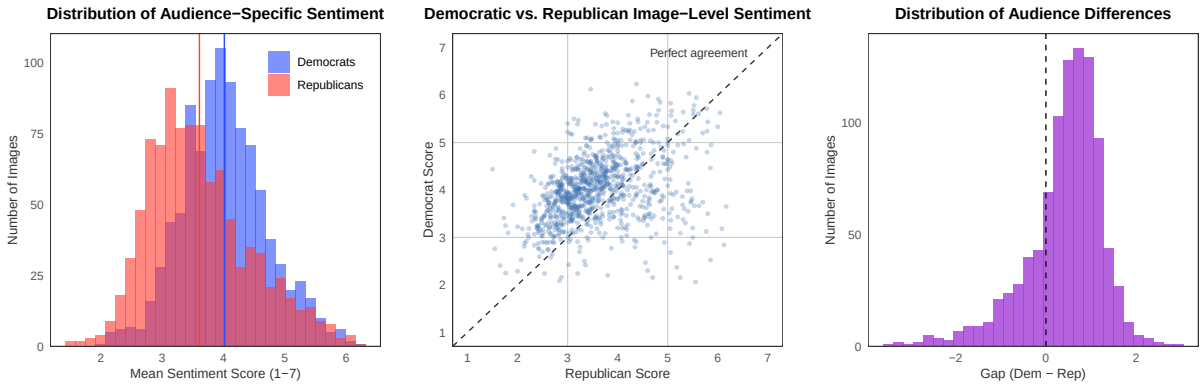
Before turning to predictive models, we first examine the extent to which Democratic and Republican annotators differ in their evaluations of the same images. If visual sentiment is genuinely perspective-dependent, we should observe systematic divergence between the two audience-specific labels and corresponding attenuation in the pooled consensus score.

Figure 1 provides descriptive evidence that evaluations of political imagery differ systematically across partisan audiences. Although the distributions of image-level sentiment overlap substantially (left panel), Democratic respondents consistently assign more positive sentiment than Republican respondents. Across the 960 immigration images, the average Democratic rating exceeds the average Republican rating by 0.41 points on the 7-point sentiment scale.

The middle panel shows considerable heterogeneity in audience-specific evaluations.

[1] We asked respondents the following question: "Would you say that this image portrays the subject(s) or objects(s) in this picture in a positive or negative light? '1' stands for negative, '4' is neutral, and '7' stands for positive."

Figure 1: Audience-Specific Sentiment Evaluations.



Note: $N = 960$ immigration images; each image's score is the mean of individual ratings within an audience (Democratic or Republican) on a 1–7 sentiment scale (1 = very negative, 4 = neutral, 7 = very positive). Each observation corresponds to the same immigration image evaluated independently by Democratic and Republican respondents. The left panel shows the distribution of audience-specific image means, with vertical lines at the Democratic and Republican means; the center panel plots Democratic against Republican image means (one point = one image; dashed 45° line = perfect agreement; faint guides mark the neutral-band edges at 3 and 5); the right panel shows the per-image gap (Democrat minus Republican; dashed line at zero). Mean sentiment is 4.01 ($SD = 0.70$) for Democratic and 3.61 ($SD = 0.84$) for Republican audiences, for a mean Democratic–Republican gap of $+0.41$ (paired $t(959) = 14.5$, $p < .001$; Cohen's $d \approx 0.5$); image-level scores are positively but imperfectly correlated ($r = 0.37$). Although the distributions overlap substantially, Democratic respondents assign consistently higher sentiment than Republican respondents, rating 76% of images (733 of 960) more positively.

While many images lie close to the line of perfect agreement, others generate substantial disagreement between Democratic and Republican respondents. The right panel shows that these differences are directional rather than random: the distribution of Democratic–Republican gaps is shifted to the right, indicating that Democrats assign higher sentiment scores to most images.

The descriptive statistics reinforce this pattern. Democrats rate 76% of images (733 of 960) more positively than Republicans, and two-thirds of images differ by more than half a point on the seven-point scale. The average difference is statistically significant ($t = 14.5$, $p < .001$) and corresponds to a medium effect size ($d \approx 0.5$). At the same time, the moderate correlation between Democratic and Republican image-level sentiment ($r = 0.37$) indicates that the two sets of evaluations are related but not exactly interchangeable. This suggests that disagreement reflects systematic differences in how partisan audiences evaluate political imagery rather than random annotation error.

The descriptive evidence in Figure 1 motivates our measurement strategy. Since evaluations of the same image differ systematically across partisan subgroups, we therefore

estimate audience-specific sentiment labels. The choice of audience partition is guided by theory: rather than modeling every possible source of heterogeneity, we estimate separate labels for groups expected a priori to differ systematically in their evaluations of political imagery. Specifically, we construct separate image-level sentiment labels for Democratic and Republican respondents by averaging ratings only within each audience. Averaging within audiences reduces idiosyncratic respondent-level variation while preserving systematic differences between partisan groups. We then use these audience-specific labels for the predictive and classification models, allowing the same image to receive different sentiment predictions depending on the audience whose evaluations are being modeled.

Because each image is evaluated by multiple respondents within each audience, the resulting labels average over individual-level variation while retaining systematic differences across partisan groups. On average, each image receives approximately 30 independent evaluations from Democratic respondents and 30 from Republican respondents. To assess the reliability of the resulting image-level labels, we compute the intraclass correlation coefficient (ICC), which quantifies the consistency of evaluations within each audience. Table 3 reports ICC estimates for both partisan groups, indicating that the audience-specific labels are measured with high reliability.

Table 3: Reliability of Audience-Specific Image Labels.

Audience	Images	Avg. raters	ICC(1,1) [95% CI]	ICC$(1, k)$ [95% CI]
Democratic	960	31.8	0.102 [0.092, 0.113]	0.783 [0.762, 0.802]
Republican	960	30.2	0.152 [0.138, 0.166]	0.844 [0.829, 0.857]
All (pooled)	960	62.0	0.095 [0.086, 0.104]	0.867 [0.854, 0.879]

Note: This table displays intraclass correlation coefficients for the sentiment ratings of immigration images within each audience. ICC(1,1) is the reliability of a single respondent; ICC$(1, k)$ is the reliability of the image-level mean label (the quantity used in the analysis), corrected for the average number of raters per image. It is estimated using one-way random-effects ICCs. Each image is rated by a different overlapping set of respondents, so a one-way model is used; the average rater count is the bias-corrected effective group size for the unbalanced design. ICC$(1, k)$ is the Spearman–Brown reliability of the audience mean. The pooled single-rater ICC is lower than either audience because partisan disagreement enters the one-way model as within-image noise.

Table 3 indicates that, although individual respondents vary considerably in their evaluations of the same image (as we can see in the low scores of ICC(1,1)), averaging approximately thirty ratings within each partisan audience yields highly reliable image-

level sentiment labels. The resulting ICC$(1, k)$ estimates exceed 0.78 for Democratic and 0.84 for the Republican audiences, indicating that the audience-specific means capture stable patterns of partisan evaluation. As expected, the pooled consensus labels show slightly higher reliability because they average over about twice as many respondents, further reducing sampling variability. This higher reliability reflects greater statistical precision, not necessarily a more appropriate measurement target. Whereas the pooled labels estimate consensus sentiment, the audience-specific labels estimate the distinct partisan evaluations that constitute the theoretical object of interest. Having established that audience-specific sentiment labels capture systematic and reliably measured differences in image evaluations, we next ask whether these differences are predictable from the visual content of the images themselves. To do so, we evaluate how state-of-the-art computer vision models can recover Democratic and Republican sentiment labels directly from images and compare their performance to models trained on conventional pooled consensus labels.

Results

Prediction Setup

We now ask whether the audience-specific evaluations can be recovered directly from image content, and how aggregation affects that recovery. We are careful about which question the prediction exercise can answer. The objective is not to determine which labeling strategy yields the highest predictive accuracy. Averaging over more annotators produces more reliable labels by reducing random measurement error, but greater reliability alone does not imply that a label faithfully represents the relationship between image content and how a particular audience evaluates political imagery. So, the central question is about measurement validity rather than statistical precision: does a given labeling strategy correctly represent how a real audience evaluates political imagery? We

therefore compare models trained on audience-specific and pooled sentiment labels to evaluate our empirical implications. If perspective collapse removes meaningful information through aggregation, pooled labels should be less representative of either partisan audience, with the largest losses occurring for politically contested images. The empirical question is not whether an average differs from its components; mechanically, it must when audience-specific labels diverge. The question is whether this divergence is large, systematically structured by theoretically meaningful audiences, recoverable from visual content, and consequential for subsequent inference.

Task. We treat sentiment recovery as continuous prediction: from an image we predict the image-level mean sentiment score on the original 1–7 scale. We estimate three models that are identical in every respect except the labelling target they are trained on: 1. averaged across only Democratic annotators, 2. averaged across only Republican annotators, and 3. the pooled mean (averaged across all annotators). We evaluate all labeling strategies using a common prediction framework. Holding the prediction pipeline fixed ensures that any observed differences arise from the labeling strategy under the same learning conditions, rather than from differences in model specification.

Architecture. We first benchmark a range of image representations and prediction heads across all three prediction tasks under a common training and evaluation protocol (Table 4). The comparison includes a CNN trained directly on raw image pixels, frozen ImageNet feature extractors (ResNet50V2 and DenseNet169), and frozen CLIP ViT-B/32 embeddings[2] paired with several prediction heads. In Tables C.4 and C.5 of the Appendix we also demonstrate the results with the alternative backbone models. Models trained from scratch perform poorly at the available sample size and fail to outperform even a mean baseline, while pretrained visual representations substantially improve predictive

[2]CLIP (Contrastive Language-Image Pretraining) is a foundation vision model trained on hundreds of millions of image-text pairs. Rather than being fine-tuned for our task, we use its pretrained image embeddings as fixed visual representations (Radford et al., 2021).

performance. The CLIP representation consistently yields the strongest predictive performance across the Democratic, Republican, and pooled sentiment labels. Within the CLIP family, CLIP ViT-B/32 + Ridge performs within one standard deviation of the best-performing alternative while using the simplest prediction head. We therefore use this architecture throughout the remainder of the analysis.

Table 4: Benchmarking prediction architectures across labeling targets.

	Cross-validated R^2				Test
Approach	Dem	Rep	Pooled	Average (SD)	R^2
Mean baseline (no image)	-0.003	-0.006	-0.004	-0.004 (0.003)	-0.003
Basic CNN (from scratch, pixels)	-0.048	0.042	-0.028	-0.012 (0.064)	-0.016
ResNet50V2 (frozen) + Ridge	0.155	0.153	0.168	0.158 (0.056)	0.097
DenseNet169 (frozen) + Ridge	0.144	0.275	0.232	0.217 (0.049)	0.315
CLIP ViT-B/32 (frozen) + Ridge	0.384	0.434	0.435	0.418 (0.036)	0.484
CLIP ViT-B/32 (frozen) + SVR	0.372	0.412	0.429	0.404 (0.030)	0.469
CLIP ViT-B/32 (frozen) + Keras MLP	0.087	0.247	-0.020	0.104 (0.092)	0.274

Note: Candidate architectures are benchmarked separately on the Democratic, Republican, and pooled image-level sentiment labels under a common evaluation protocol. Model development uses 822 images with predefined 5-fold cross-validation for model fitting and hyperparameter selection; final performance is evaluated on a held-out test set of 138 images ($\sim$ 15% of the data) that is not used during model development. Cross-validated R^2 is reported separately for each prediction task; Average reports the mean cross-validated R^2 across the three tasks, with the standard deviation across folds shown in parentheses. Test R^2 is the mean coefficient of determination across the three held-out prediction tasks. Pretrained models are used as frozen feature extractors; only the prediction head is estimated. We retain CLIP ViT-B/32 + Ridge because it performs consistently among the strongest architectures while using the simplest prediction head.

Evaluation. We evaluate all models under a common protocol using 5-fold cross-validation on the training and validation sets and a held-out test set for final evaluation. Performance is reported using R^2, RMSE, MAE, and Pearson correlation. The comparison of interest throughout is the relative performance across Democratic, Republican, and pooled sentiment labels.

Main Prediction Results

We begin by comparing overall predictive performance across the three labeling strategies. Table 5 reports prediction accuracy for the Democratic, Republican, and pooled sentiment labels. The pooled label achieves the highest predictive performance ($R^2 = 0.54$ on

the held-out test set), followed by the Republican ($R^2 = 0.47$) and Democratic ($R^2 = 0.44$) labels. This result is expected because the three labels differ in measurement reliability. The pooled label averages evaluations from approximately twice as many respondents as either audience-specific label and therefore has the highest reliability (ICC$(1, k) = 0.87$ versus 0.84 for Republicans and 0.78 for Democrats; Table 3). A more reliable target contains less measurement noise and is therefore easier to predict from image features. Higher predictive performance therefore reflects greater label reliability rather than evidence that the pooled label provides a more substantively meaningful representation of audience sentiment. The key question is therefore not which label is easiest to predict, but whether each labeling strategy accurately represents how real audiences evaluate political imagery. We therefore next evaluate predictive performance separately for images on which Democratic and Republican respondents agree versus those on which they disagree.

Table 5: Prediction performance across alternative sentiment labels.

	Label	Cross-validation			Test	
Target	SD	R^2 (SD)	RMSE	r	R^2	r
Democratic	0.70	0.384 (0.030)	0.550	0.622	0.443	0.673
Republican	0.83	0.434 (0.050)	0.623	0.660	0.466	0.695
Pooled	0.64	0.435 (0.027)	0.482	0.660	0.542	0.747

Note: All models use frozen CLIP ViT-B/32 image embeddings with ridge regression. Results are based on five-fold cross-validation using 822 images and evaluation on a held-out test set of 138 images. SD denotes the standard deviation of the target label. R^2 and Pearson's r are reported because they are directly comparable across targets with different variances.

Table 6 examines this question by partitioning the held-out images according to whether Democratic and Republican respondents assign them to the same sentiment category. This allows us to compare the performance of audience-specific and pooled labels separately for images on which the two audiences agree and those on which they disagree.

This pattern provides direct support for Implication 2, which predicts that the information lost through aggregation should be concentrated and be largest among politically contested images. Among non-polarizing images, the audience-specific and pooled

approaches perform almost identically. The additional prediction error introduced by pooling is only 0.02 sentiment scale points, and the predicted-observed correlations differ little (0.57 versus 0.54). The pattern changes sharply for politically contested images. Audience-specific models recover each party's evaluations substantially more accurately than the pooled label when used as a proxy for each audience's evaluations. The cost of pooling increases to 0.16 scale points, while the predicted-observed correlation falls from 0.82 for the audience-specific models to 0.64 for the pooled model. The overall predictive advantage of the pooled label therefore masks a systematic loss of information concentrated precisely among politically contested images, where aggregation obscures meaningful differences in audience evaluations.

Table 6: Prediction accuracy by image polarization.

		Mean abs. error			Correlation r	
Image group	n	Aud.-spec.	Pooled	Cost	Aud.-spec.	Pooled
Non-polarizing	87	0.410	0.429	0.019	0.568	0.535
Polarizing	51	0.530	0.685	0.155	0.818	0.639

Note: Images are classified as polarizing when Democratic and Republican image-level labels fall into different sentiment categories (negative, neutral, or positive; cut points at 3 and 5 on the seven-point scale). Audience-specific models predict the corresponding party-specific labels, whereas the pooled model predicts the aggregated label, which is evaluated as a proxy for each audience separately. Prediction architecture is identical to Table 5.

When Does Pooling Become Costly?

The binary comparison above establishes that the benefits of audience-specific labels are concentrated on politically contested images. We next examine the related scaling prediction in Implication 1: whether information loss increases progressively with the degree of partisan disagreement, rather than appearing only once an image crosses a particular polarization threshold. We test this prediction and examine whether the observed pattern could arise mechanically from defining polarizing images using the same labels later used for evaluation.

To test whether information loss increases continuously with partisan disagreement,

Table 7 groups held-out images into terciles according to the absolute difference between Democratic and Republican sentiment labels and compares prediction error across the three groups.

Table 7: Prediction accuracy by degree of partisan disagreement.

Disagreement	Gap range	n	Audience-specific	Pooled	Cost of pooling
Low	0.00-0.48	46	0.469	0.404	-0.07
Medium	0.49-0.98	46	0.382	0.442	0.06
High	0.98-3.33	46	0.512	0.725	0.21

Note: Images are grouped into terciles according to the absolute difference between Democratic and Republican image-level sentiment labels. Entries report mean absolute prediction error averaged across the two audiences. Pooling cost is defined as the difference between pooled and audience-specific prediction error. Positive values indicate that audience-specific labels provide more accurate predictions.

The results support this caling prediction. Among images in the lowest disagreement tercile, pooling imposes essentially no cost and even performs marginally better because the two audiences already assign similar sentiment to the image. As partisan disagreement increases, however, the cost of pooling rises progressively, becoming positive in the middle tercile and reaching 0.21 sentiment scale points in the highest tercile. The information lost through aggregation therefore increases smoothly with systematic disagreement between audiences rather than appearing only once an image crosses an arbitrary threshold.

We next test this relationship directly by regressing each test image's pooling cost (the increase in prediction error from using the pooled rather than the audience-specific model) on the magnitude of partisan disagreement,

$$\text{PoolingCost}_i = \beta_0 + \beta_1 \left|\text{Dem}_i - \text{Rep}_i\right| + \varepsilon_i$$

The estimated relationship is positive and significant. Each one-point increase in partisan disagreement increases the cost of pooling by approximately one quarter of a sentiment scale point ($\hat{\beta} = 0.253$, 95% CI $[0.197, 0.309]$, $p < 0.001$). Partisan disagreement alone explains nearly forty percent of the variation in image-level pooling cost ($R^2 = 0.386$).

Table 8: Image-level regression of pooling cost on partisan disagreement.

	Pooling cost
Absolute partisan disagreement	0.253^{***}
	(0.028)
Constant	-0.133^{***}
	(0.022)
95% CI (disagreement)	$[0.197,\ 0.309]$
R^2	0.386
Observations	138

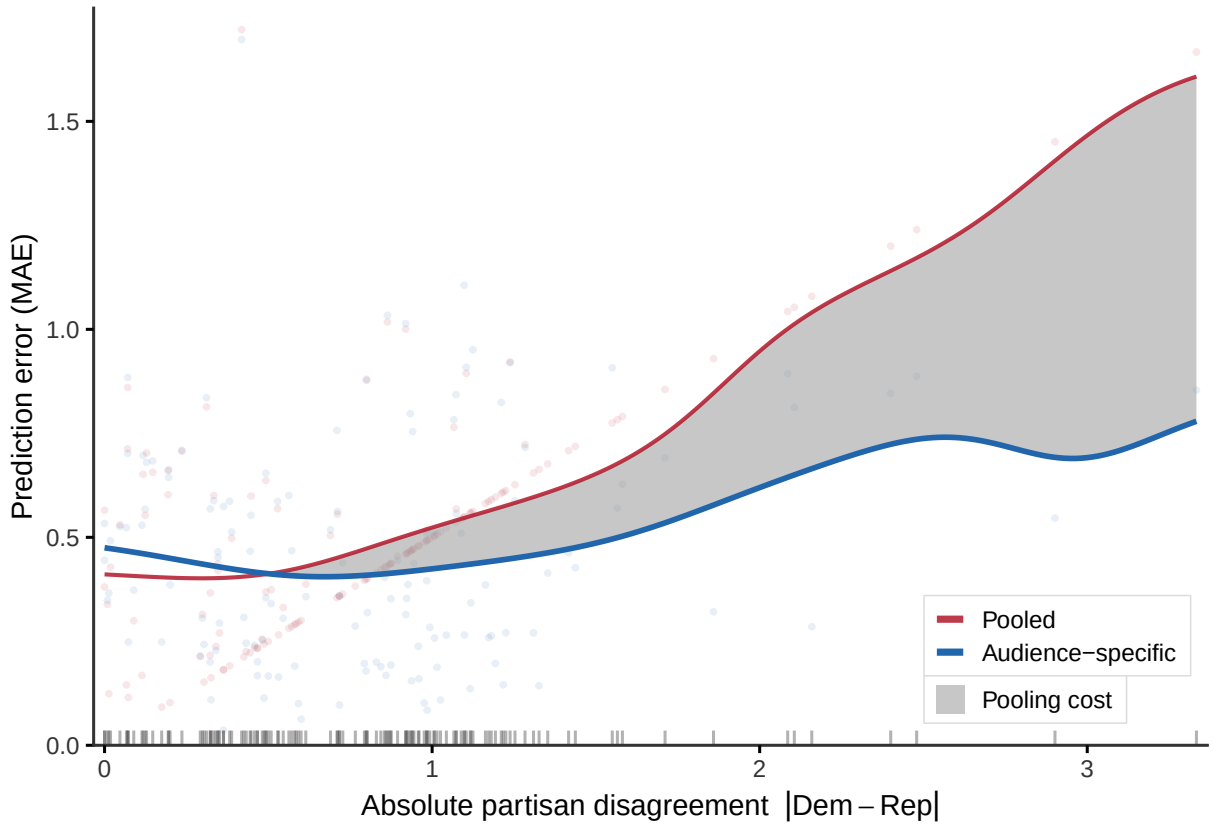
Note: The dependent variable is the image-level increase in prediction error from using the pooled rather than the audience-specific model. HC1 robust standard errors are reported in parentheses.

Replacing the linear specification with a natural cubic spline yields the same conclusion ($F(2, 135) = 43.1$, $p < 0.001$), indicating that the relationship is not an artifact of functional-form assumptions. Figure 2 visualizes this continuous gradient.

When Democratic and Republican evaluations are similar, aggregation removes primarily random variation, so pooling carries little cost. As audience-specific evaluations diverge, however, the pooled label increasingly represents neither audience, producing progressively larger prediction errors for both. The cost of pooling therefore emerges precisely where systematic disagreement is greatest. The same qualitative pattern holds when Democratic and Republican prediction errors are examined separately (Figure B.1 in the Appendix).

One concern is that the relationship between partisan disagreement and pooling cost is mechanically induced because both quantities are computed from the same respondent ratings. To eliminate this possibility, we randomly partition respondents evaluating each image into two disjoint groups. One half is used solely to classify images as polarizing or non-polarizing, while the other half is used exclusively to construct the audience-specific and pooled labels used for prediction and evaluation. Because no respondent contributes to both quantities, any mechanical dependence between the disagreement measure and the evaluation metric is eliminated. Table 9 shows that the cost of pooling remains con-

Figure 2: Pooling cost increases with partisan disagreement.



Note: Each point represents one test image. For each image, prediction error is measured as the mean absolute error (MAE) averaged across Democratic and Republican evaluations. The blue curve shows the Gaussian-kernel local average of the audience-specific prediction error, computed by averaging the Democratic model's error against Democratic labels and the Republican model's error against Republican labels. The red curve shows the corresponding average error of the pooled model, evaluated against both audiences' labels. The shaded region between the curves therefore represents the image-level pooling cost (pooled error minus audience-specific error).

Table 9: Robustness to sample splitting.

Image group (defined by Half A)	n	Audience-specific	Pooled	Cost of pooling
Non-polarizing (Half A)	76	0.544	0.570	0.026
Polarizing (Half A)	57	0.546	0.622	0.076

Note: Respondents evaluating each image are randomly divided into two disjoint groups. Half A is used only to classify images as polarizing or non-polarizing, while Half B is used exclusively to construct the audience-specific and pooled labels used for prediction and evaluation. Prediction error is measured as mean absolute error (MAE) averaged across Democratic and Republican labels. Because the disagreement measure and the prediction evaluation are derived from different respondents, the observed relationship cannot be a mechanical consequence of using the same ratings in both steps.

centrated among the polarizing images. Although the estimates are attenuated because each label is constructed from only half as many raters, the qualitative pattern stays consistent. It indicates that the concentration of pooling costs among contested images is not an artifact of subdividing respondents or conditioning on the outcome, but reflects genuine differences in audience-specific sentiment.

Placebo Partition

If our framework is correct, only partitions that reflect systematic differences in interpretation should recover additional information. Arbitrary partitions should not.

We evaluate this prediction by reconstructing audience-specific labels under several alternative partitions. In addition to party identification, we consider ideology and immigration attitudes, both of which plausibly structure evaluations of immigration imagery, and compare them with a placebo partition obtained by randomly assigning respondents to two groups. The prediction exercise is then repeated under each partition using the same modeling pipeline. If merely dividing respondents were sufficient to recover information, the random partition should perform similarly to the theoretically motivated ones.

Table 10 demonstrates the placebo exercise. When respondents are assigned to groups at random, the resulting labels differ only because of sampling variability (mean divergence 0.41), and fitting separate models provides essentially no improvement over

Table 10: Robustness to meaningful partitions.

Partition	Images	Label divergence	Cost of pooling
Random (placebo)	960	0.407	-0.003
Party (Dem vs Rep)	960	0.803	0.069
Ideology (lib vs con)	960	0.832	0.070
Immigration attitude	517	4.157	1.898

Note: Label divergence is the mean absolute difference between the two audience-specific image labels on the held-out test set. Cost of pooling is the reduction in label-recovery error obtained by using audience-specific rather than pooled prediction models. For each partition, audience-specific labels are reconstructed from the same underlying respondent ratings and evaluated using the identical prediction pipeline. All partitions require at least three raters per image.

a pooled model (pooling cost -0.003). Simply partitioning respondents therefore does not recover additional signal. In contrast, theoretically motivated partitions produce substantially greater disagreement between audience-specific labels and corresponding gains from audience-specific prediction. Democratic and Republican labels diverge nearly twice as much (0.80), with a pooling cost of 0.07, while ideology yields nearly identical results (0.83 and 0.07). The strongest effects arise when respondents are partitioned by immigration attitudes, where disagreement is greatest.

These results illustrate the logic of Implication 3. Theoretically meaningful audience partitions recover information that is lost under consensus aggregation, whereas arbitrary partitions do not. Simply subdividing respondents creates no additional signal; what matters is whether the partition captures systematic differences in how individuals interpret visual stimuli. Partisanship provides one empirically supported partition for immigration imagery, but the broader principle is that audience-specific modeling is useful only when those systematic differences exist.

Choosing Audience Partitions

A natural question raised by our framework is how researchers should choose audience partitions. Although we focus on partisan identity, the proposed framework is not specific to politics or party identification. Rather, it applies whenever sentiment depends systematically on characteristics of the audience evaluating the image. Our argument is

not that every observable characteristic warrants a separate sentiment model. Estimating separate models for every possible audience would quickly become impractical, increasing estimation variance while fragmenting the available data, whereas relying on a single consensus label risks obscuring important heterogeneity. The methodological challenge is therefore not to maximize the number of audience partitions, but to identify those that correspond to meaningful and reproducible differences in how visual stimuli are interpreted.

We suggest three considerations for choosing audience partitions. First, partitions should be theoretically motivated: they should correspond to characteristics expected to shape how individuals interpret visual stimuli. Second, these differences should be empirically identifiable, producing systematic rather than idiosyncratic variation in sentiment. Third, each partition should exhibit sufficient within-group agreement and sample size to permit reliable estimation of audience-specific sentiment labels.

These considerations are intended as principles rather than an algorithm. Future work may develop data-driven procedures for identifying audience partitions, combining theory with statistical evidence on between-group disagreement and within-group reliability. More generally, our framework suggests that audience heterogeneity should be modeled when it is theoretically meaningful and empirically supported, rather than eliminated through indiscriminate aggregation.

Consequences for the Inference Work

Thus far, we showed that pooled sentiment labels measure a different quantity than audience-specific labels whenever evaluations differ systematically across audiences. Now one might ask whether this measurement difference affects the follow-up empirical analyses that rely on such sentiment measures. We show that it does: because aggregation changes the quantity being measured rather than merely introducing random error, its

consequences propagate into subsequent inference. We illustrate three common consequences: misclassification of image valence, attenuation of estimated relationships, and attenuation of treatment effects, using quantities calibrated to the 960 immigration images in our data (Table 11).

Table 11: Consequences of label aggregation for the inference.

		Disagreement tercile			
Audience	Quantity	Low	Mid	High	Overall
Panel A. Label distortion by disagreement tercile					
Democratic	Sentiment sign-flip	7%	22%	42%	24%
	Slope recovered	95%	90%	80%	90%
Republican	Sentiment sign-flip	5%	19%	28%	17%
	Slope recovered	105%	107%	113%	107%
Panel B. Treatment-effect recovery (simulation)					
		High tercile			Overall
Democratic		18% [−4%, 40%]			52% [40%, 64%]
Republican		36% [15%, 58%]			56% [46%, 66%]

Note: All quantities are computed using the 960 immigration images. *Sentiment sign-flip* reports the proportion of images for which the pooled sentiment label and the audience-specific sentiment label fall on opposite sides of the neutral midpoint (4.0 on the 7-point scale). *Slope recovered* reports the percentage of the audience-specific regression coefficient recovered when the pooled label is used in place of the audience-specific label. *Treatment effect* recovered reports the percentage of a true treatment effect of negatively valenced imagery recovered when treatment assignment is based on the pooled rather than the audience-specific label (2,000 simulation draws; 95% Monte Carlo simulation intervals). Columns report estimates overall and by terciles of partisan disagreement.

The most immediate consequence of aggregation is the misclassification of image sentiment. Using a continuous definition, we count a sign-flip whenever the pooled sentiment label and the audience-specific label fall on opposite sides of the neutral midpoint (4 on a 7-point scale). Under this definition, the pooled label disagrees with the Democratic audience label for 24% of images and with the Republican audience label for 17% of images (Table 11). These disagreements though are not distributed uniformly across images. They are concentrated among the images that generate the greatest partisan disagreement, rising from 7% in the lowest disagreement tercile to 42% in the highest. Table D.6 in the Appendix shows that this pattern is robust to increasingly conservative definitions of a sentiment disagreement. Although the overall disagreement rate declines when

the conservative midpoint criterion is relaxed, the remaining disagreements remain concentrated almost entirely among the most polarizing images.

A second consequence arises when researchers use image sentiment as an explanatory variable. Suppose a study regresses an image-level outcome on audience sentiment but, lacking audience-specific labels, it substitutes the pooled label as a proxy. We quantify how much of the audience-specific regression coefficient is recovered under this substitution. The pooled proxy recovers different fractions of the audience-specific coefficient depending on both the audience and the degree of partisan disagreement. For the Democratic audience sentiment, recovery declines from 95% among the least polarizing images to 80% among the most polarizing ones (Table 11), indicating increasing attenuation as disagreement grows. For Republican audience sentiment, recovery modestly exceeds 100%, reflecting a slight amplification rather than attenuation. Thus, aggregation does not uniformly shrink downstream estimates; it systematically distorts them in ways that depend on the relationship between the pooled and audience-specific sentiment labels.

The asymmetry between Democratic and Republican sentiment suggests that the consequences of aggregation may themselves depend on properties of the underlying audience-specific sentiment distributions. Characterizing when aggregation attenuates versus amplifies downstream estimates is an interesting direction for future methodological work.

A related concern is post-treatment conditioning. The simulation should not be read as recommending that researchers classify images using perceptions measured from the same respondents whose outcomes are later analyzed. If perceived sentiment is measured after image exposure and then used to subset observations, define treatment categories, or adjust the outcome model, standard post-treatment-bias concerns apply (Rosenbaum, 1984). Our point is different: when researchers use externally generated sentiment labels to select or classify visual stimuli, pooled labels may misclassify the treatment category relative to the audience-specific sentiment they intend to manipulate. Audience-specific labels therefore do not solve all causal-design problems; they clarify what visual treat-

Table 12: Aggregation Changes Substantive Inference About Image Content.

Image content	Democratic	Republican	Pooled
American flag †	-0.07^{**}	$+0.18^{***}$	$+0.05^{*}$
Children & families ‡	$+0.03$	-0.16^{***}	-0.07^{**}
Protest ‡	$+0.01$	-0.21^{***}	-0.09^{***}
Crowd of migrants ‡	-0.03	-0.32^{***}	-0.17^{***}
Law enforcement	-0.14^{***}	-0.09^{**}	-0.12^{***}
Detention	-0.15^{***}	-0.36^{***}	-0.25^{***}
Border barrier	-0.18^{***}	-0.16^{***}	-0.17^{***}

Note: This table reports coefficients from separate regressions of Democratic, Republican, and pooled image-level sentiment on standardized zero-shot CLIP content scores (one regression per row). Coefficients represent the change in sentiment associated with a one-standard-deviation increase in content presence. † indicates content for which Democratic and Republican responses have opposite signs, attenuating the pooled estimate through cancellation. ‡ indicates content for which the pooled estimate appears significant despite the association being driven primarily by one audience. $N = 960$ images. $p < .05$, $p < .01$, $p < .001$.

ment is being assigned.

When Pooling Changes Scientific Conclusions

We now look at the empirical example where image sentiment labels can be used as a dependent variable, and ask which visual features make immigration images be perceived more positively or negatively? To answer it, we regress image-level sentiment on measures of image content extracted directly from the images using a pretrained CLIP vision-language model, independently of the human sentiment ratings. The resulting conclusions differ depending on whether sentiment is measured using pooled or audience-specific labels. Table 12 shows that aggregation can change the substantive inference in two distinct ways.

The first mechanism happens when the two audiences respond in opposite directions to the same visual concept. American flag imagery provides the clearest example: it is associated with more positive sentiment among Republicans ($+0.18$ points per standard deviation of content presence) but more negative sentiment among Democrats (-0.07). Averaging these opposing responses attenuates the pooled coefficient to just $+0.05$. A researcher relying on the pooled labels would therefore conclude that the American flag has

only a weak association with sentiment, even though it elicits substantial but opposing reactions across the two audiences.

The second distortion occurs when only one audience responds systematically to a visual concept. For example, images depicting crowds of migrants are associated with substantially more negative sentiment among Republicans (-0.32) but not among Democrats (-0.03 and not significant). The pooled analysis though reports a significant negative association (-0.17), making an audience-specific relationship appear to be a general property of the image.

In both cases, the pooled label is not merely a less precise measure of sentiment; it yields different conclusions about which visual features shape political evaluations. Depending on the pattern of audience disagreement, aggregation either masks genuinely polarizing content through cancellation or makes audience-specific associations appear population-wide. Disaggregation helps to recover this structure.

One might argue that such differences could simply be recovered through subgroup analyses or interaction models. However, those approaches address heterogeneity in relationships conditional on the outcome being correctly measured. Our argument concerns an earlier stage of the research process: when audience-specific evaluations have already been averaged into a single consensus label, the outcome itself has changed. No subsequent modeling strategy would be able to recover information that was discarded during label aggregation.

Discussion

The existing work on sentiment analysis, including visual sentiment, starts from a reasonable assumption: people may disagree, but that disagreement is mostly noise around shared sentiment judgment that can be smoothed over at scale. If that is true, averaging labels makes sense because it brings the estimate closer to the sentiment the image is

presumed to have. Our results show that this assumption does not always hold in political settings. Some disagreements are stable and systematic, and are tied to the cleavages rooted in the audiences evaluating the visuals. In those cases, pooling individual evaluations in the aggregated image-level labels does not just remove individual-level noise. It changes what exactly we measure. A pooled label replaces the evaluations of real audiences with a combined score that may not match to how politically meaningful audience groups actually evaluate the image.

We first show this distinction analytically and then examine it with an empirical application using images from the topic of immigration. We emphasize that we do not choose a relevant audience division because particular images are already known to be polarizing. We choose it because partisanship is theoretically likely to shape how immigration imagery is interpreted. The results show where this expectation matters. Democratic and Republican labels are not simply noisier versions of the same image sentiment. They diverge in systematic ways, and the cost of pooling is concentrated on the images where that divergence is largest. Audience-specific labels recover Democratic and Republican sentiment better, while pooled labels are weakest as measures of audience sentiment precisely where the two groups disagree in their evaluations.

We are not arguing that consensus labels are always wrong, or that every project needs separate labels for every possible group, but rather that the choice depends on what the researchers want the label to represent. If disagreement is mostly random variation around a shared construct, aggregation remains useful because it improves precision without changing the target of measurement. If disagreement reflects stable differences in interpretation, the pooled label measures something else: an average across groups rather than the evaluation of any one group. The point is therefore not to replace aggregation as a general rule, but to recognize when aggregation changes the object of measurement.

As political science relies more on automated measures of unstructured data such as textual, visual, and audio content, the challenge is not only to build more accurate pre-

diction models, it is also to make sure that the labels those models learn from actually match the concepts researchers want to measure. Audience differences in the interpretation of political imagery are not surprising. The contribution of this study is to show what follows for computational measurement when those differences enter the annotation process. If sentiment labels are built by averaging across audiences that interpret the same image differently, the resulting label may be reliable in a statistical sense while no longer representing the evaluation of any politically meaningful group. Pooling those evaluations may produce a cleaner label, but it can also change what the label represents and, in turn, what researchers conclude from it in their subsequent inferential and causal work. For visual sentiment analysis, and for computational measurement more broadly, the task is therefore not only to improve prediction, but to ask whose judgment the prediction is meant to recover.

References

Adcock, Robert and David Collier. 2001. "Measurement validity: A shared standard for qualitative and quantitative research." *American political science review* 95(3):529–546.

Anastasopoulos, L Jason, Dhruvil Badani, Shiry Ginosar and Jake Ryland Williams. 2024. "Visible home style." *Electoral Studies* 90:102794.

Aroyo, Lora and Chris Welty. 2015. "Truth is a lie: Crowd truth and the seven myths of human annotation." *AI magazine* 36(1):15–24.

Äyräväinen, Laura EM, Joanne Hinds and Brittany I Davidson. 2025. "Disambiguating sentiment annotation: A mixed methods investigation of annotator experience and impact of instructions on annotator agreement." *Plos one* 20(12):e0336269.

Benoit, Kenneth, Drew Conway, Benjamin E Lauderdale, Michael Laver and Slava Mikhaylov. 2016. "Crowd-sourced text analysis: Reproducible and agile production of political data." *American Political Science Review* 110(2):278–295.

Borth, Damian, Rongrong Ji, Tao Chen, Thomas Breuel and Shih-Fu Chang. 2013. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM international conference on Multimedia*. pp. 223–232.

Bossetta, Michael and Rasmus Schmøkel. 2023. "Cross-platform emotions and audience engagement in social media political campaigning: Comparing candidates' Facebook and Instagram images in the 2020 US election." *Political Communication* 40(1):48–68.

Cabitza, Federico, Andrea Campagner and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37 pp. 6860–6868.

Card, Dallas, Serina Chang, Chris Becker, Julia Mendelsohn, Rob Voigt, Leah Boustan, Ran Abramitzky and Dan Jurafsky. 2022. "Computational analysis of 140 years of US

political speeches reveals more positive but increasingly polarized framing of immigration." *Proceedings of the National Academy of Sciences* 119(31):e2120510119.

Carpinella, Colleen and Nichole M Bauer. 2021. "A visual analysis of gender stereotypes in campaign advertising." *Politics, Groups, and Identities* 9(2):369–386.

Casas, Andreu and Nora Webb Williams. 2019. "Images that matter: Online protests and the mobilizing role of pictures." *Political Research Quarterly* 72(2):360–375.

Davani, Aida Mostafazadeh, Mark Díaz and Vinodkumar Prabhakaran. 2022. "Dealing with disagreements: Looking beyond the majority vote in subjective annotations." *Transactions of the Association for Computational Linguistics* 10:92–110.

de Lima-Santos, Mathias-Felipe, Isabella Gonçalves, Marcos G Quiles, Lucia Mesquita and Wilson Ceron. 2023. "Visual Political Communication in a Polarized Society: A Longitudinal Study of Brazilian Presidential Elections on Instagram." *arXiv preprint arXiv:2310.00349* .

Domke, David, David Perlmutter and Meg Spratt. 2002. "The primes of our times? An examination of the 'power'of visual images." *Journalism* 3(2):131–159.

Grabe, Maria Elizabeth and Erik P Bucy. 2014. Image bite analysis of political visuals: Understanding the visual framing process in election news. In *Sourcebook for Political Communication Research*. Routledge pp. 209–237.

Kunda, Ziva. 1990. "The case for motivated reasoning." *Psychological bulletin* 108(3):480.

Lazarus, Richard S. 1991. *Emotion and adaptation*. Oxford University Press.

Leong, Yuan Chang, Brent L Hughes, Yiyu Wang and Jamil Zaki. 2019. "Neurocomputational mechanisms underlying motivated seeing." *Nature human behaviour* 3(9):962–973.

Madrigal, Guadalupe and Stuart Soroka. 2023. "Migrants, caravans, and the impact of news photos on immigration attitudes." *The International Journal of Press/Politics* 28(1):49–69.

Ollerenshaw, Trent and Ashley Jardina. 2023. "The asymmetric polarization of immigration opinion in the United States." *Public Opinion Quarterly* 87(4):1038–1053.

Ortis, Alessandro, Giovanni Maria Farinella and Sebastiano Battiato. 2020. "Survey on visual sentiment analysis." *IET Image Processing* 14(8):1440–1456.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR pp. 8748–8763.

Rosenbaum, Paul R. 1984. "The consequences of adjustment for a concomitant variable that has been affected by the treatment." *Journal of the Royal Statistical Society Series A: Statistics in Society* 147(5):656–666.

Sabou, Marta, Kalina Bontcheva, Leon Derczynski and Arno Scharl. 2014. Corpus annotation through crowdsourcing: Towards best practice guidelines. In *LREC*. pp. 859–866.

Sap, Maarten, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi and Noah A Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies*. pp. 5884–5906.

Scherer, Klaus R. 2001. "Appraisal considered as a process of multilevel sequential checking." *Appraisal processes in emotion: Theory, methods, research* 92(120):57.

Schill, Dan. 2012. "The visual image and the political image: A review of visual com-

munication research in the field of political communication." *Review of communication* 12(2):118–142.

Song, Kaikai, Ting Yao, Qiang Ling and Tao Mei. 2018. "Boosting image sentiment analysis with visual attention." *Neurocomputing* 312:218–228.

Williams, Nora Webb, Andreu Casas, Kevin Aslett and John Wilkerson. 2026. "When conservatives See red but liberals feel blue: Labeler characteristics and variation in content annotation." *The Journal of Politics* 88(2):000–000.

Yan, Yan, Rómer Rosales, Glenn Fung, Ramanathan Subramanian and Jennifer Dy. 2014. "Learning from multiple annotators with varying expertise." *Machine learning* 95(3):291–327.

You, Quanzeng, Hailin Jin and Jiebo Luo. 2017. Visual sentiment analysis by attending on local image regions. In *Proceedings of the AAAI conference on artificial intelligence*. Vol. 31.

Zhao, Sicheng, Xingxu Yao, Jufeng Yang, Guoli Jia, Guiguang Ding, Tat-Seng Chua, Bjoern W Schuller and Kurt Keutzer. 2021. "Affective image content analysis: Two decades review and new perspectives." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44(10):6729–6751.

Appendix

A Classification Task

After gathering individual evaluations, we averaged them to produce a single score for each image on a 1 to 7 interval scale (average evaluation score - AES). While we use these interval-scale scores for linear predictions in our analysis, we also convert them into conventional sentiment labels. Specifically, we categorize the interval scores into three sentiment groups—negative, neutral, and positive:

$$\text{Categorical Label} = \begin{cases} \text{negative}, & \text{if AES} \leq 3 \\ \text{neutral}, & 3 < \text{if AES} < 5 \\ \text{positive}, & \text{if AES} \geq 5 \end{cases} \tag{4}$$

The classification results match the continuous-prediction results. Audience-specific labels are classified more accurately than the pooled label (Table A.1); the pooled classifier's disadvantage is concentrated on the politically contested images and disappears where the parties agree (Table A.2); and the recovered class structure is specific to theoretically motivated partitions, while a random split of the same raters does not recover any meaningful information (Table A.3).

Table A.1: Overall Classification Results

Target	Macro-F1	Accuracy	$F1_{neg}$	$F1_{neu}$	$F1_{pos}$
Democratic	0.567	0.797	0.364	0.877	0.462
Republican	0.683	0.739	0.611	0.802	0.636
Pooled	0.637	0.841	0.564	0.904	0.444

Note: Negative ≤ 3, Neutral $(3, 5)$, Positive ≥ 5 on the 1–7 scale; macro-F1 weights the three classes equally, so it is not inflated by the dominant neutral class.

Table A.2: Classification task based on image groups.

Image group	n	Audience-specific	Pooled proxy	Gain
Non-polarizing	87	0.833	0.908	-0.075
Polarizing	51	0.657	0.480	0.176
All	138	0.768	0.750	0.018

Note: Accuracy averaged over the Democratic and Republican labels; gain is audience-specific minus pooled-proxy.

Table A.3: Classification task for various audience partitions.

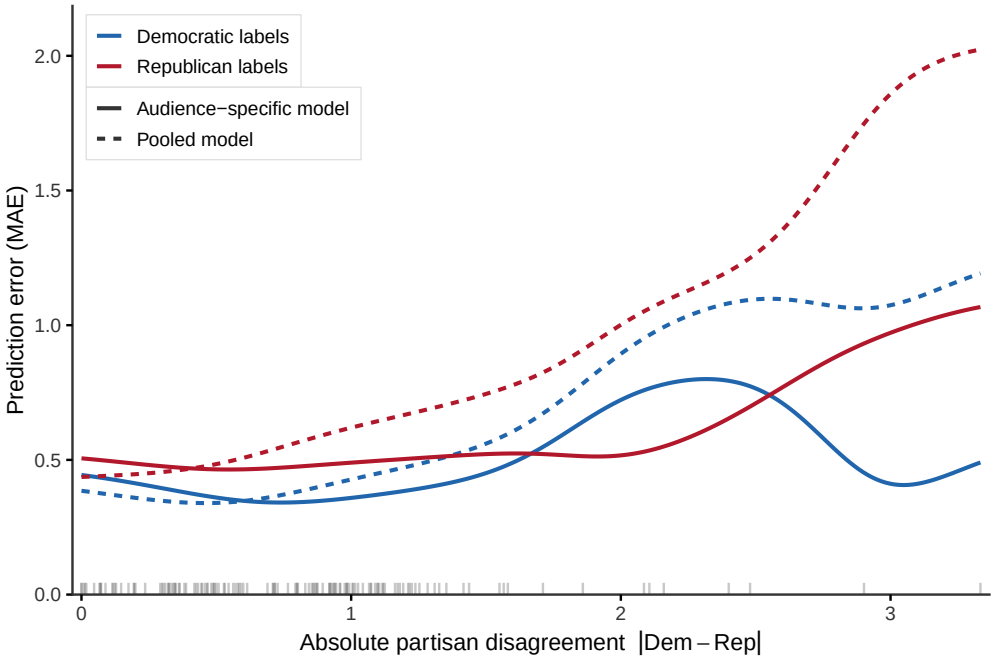
Partition	Images	Category disagreement	Accuracy gain
Random (placebo)	960	0.167	-0.014
Party (Dem vs Rep)	960	0.370	0.018
Ideology (lib vs con)	960	0.391	0.065

Note: Two-group 3-class labels re-built from the raw immigration ratings under each partition; category disagreement is the share of test images on which the two groups' categories differ; gain is the accuracy of group-specific classifiers minus a single pooled classifier.

B Pooling Costs by Partisan Subgroups

Figure B.1 disaggregates prediction error by audience. The increase in pooling error is evident for both Democratic and Republican labels, indicating that the average relationship reported in the main text is not driven by one audience alone. Although the magnitude differs somewhat across audiences with pooling producing larger errors for Republican labels on this dataset, the qualitative pattern is unchanged: pooled predictions become increasingly inaccurate as partisan disagreement grows.

Figure B.1: Pooling costs by partisan subgroups.



Note: Curves show Gaussian-kernel local averages of image-level prediction error (MAE) as a function of absolute partisan disagreement. Solid lines denote the audience-specific models evaluated against their corresponding audience labels, while dashed lines denote the pooled model evaluated separately against Democratic and Republican labels. Gray tick marks indicate the distribution of test images along the disagreement scale. Pooling becomes increasingly costly for both audiences as partisan disagreement grows, indicating that the average relationship reported in Figure 2 is not driven by one audience alone. Although the increase is somewhat larger for Republican labels in this dataset, the qualitative pattern is the same for both audiences.

C Robustness Across Model Specifications

Table C.4: Robustness to other known models.

Frozen backbone	Dim	Cost (non-polarizing)	Cost (polarizing)
CLIP ViT-B/32 (OpenAI)	768	0.019	0.155
CLIP ViT-L/14 (OpenAI)	1024	0.019	0.190
DINOv2 ViT-B/14 (self-supervised)	768	0.012	0.139
SigLIP ViT-B/16 (Google)	768	0.001	0.190
OpenCLIP ViT-B/32 (LAION-2B)	512	0.034	0.158
EVA-02-CLIP ViT-B/16 (merged-2B)	512	0.009	0.161
ResNet50V2 (ImageNet, frozen)	2048	-0.042	0.123
DenseNet169 (ImageNet, frozen)	1664	0.012	0.116

Note: Every backbone was frozen, with no fine-tuning, with the same the same selected prediction head (ridge regression).

Here we use a multi-task continuous prediction model to predict sentiment for two distinct partisan groups, Democrats and Republicans, within a unified framework. We test whether training one model with two separate heads for each partisan predictive label can outperform training two separate audience-specific models. Our results show essentially no difference between these two approaches, so we opt to work with two separate audience-specific models as a main specification.

Table C.5: Individual models vs. multi-task model.

	Democratic R^2		Republican R^2	
Design	CV	Test	CV	Test
Separate (one net per audience)	-0.040	0.124	0.156	0.352
Multi-task (shared trunk, two heads)	0.044	0.236	0.195	0.268

Note: Both designs use a frozen CLIP ViT-B/32 embedding and the same two-block hidden architecture; they differ only in whether the hidden trunk is shared across audiences. Per-audience performance (R^2) is essentially unchanged.

D Subsequent Research

Table D.6: Consequences of label aggregation with conservative cut-offs.

		Disagreement tercile			
Audience	Reversal definition	Low	Mid	High	Overall
Democratic	Continuous midpoint criterion (4.0 threshold)	0.07	0.22	0.42	0.24
	Continuous midpoint criterion (0.25 margin)	0.00	0.00	0.10	0.03
	Polarity criterion (≤ 3 vs ≥ 5)	0.00	0.00	0.00	0.00
Republican	Continuous midpoint criterion (4.0 threshold)	0.05	0.19	0.28	0.17
	Continuous midpoint criterion (0.25 margin)	0.00	0.00	0.09	0.03
	Polarity criterion (≤ 3 vs ≥ 5)	0.00	0.00	0.00	0.00

Note: Entries report the share of images for which the pooled label and the audience-specific label fall on opposite sides of neutral. Disagreement terciles are based on the absolute Democratic–Republican sentiment gap. The midpoint criterion counts any crossing of the neutral midpoint (4.0). The 0.25-margin criterion additionally requires both labels to be at least 0.25 scale points from neutral. The polarity criterion requires one label to be at or below 3 and the other at or above 5.