Gender Disparities in Judicial Attention:
Audiovisual Evidence from Appellate Oral Arguments

Elliott Ash & Aniket Kesari*

March 27, 2026

Abstract

We study procedural disparities in appellate courts by examining how judges allocate attention during oral argument. Using a large collection of video-recorded oral arguments from the U.S. Court of Appeals for the Ninth Circuit between 2010 and 2025, we construct multimodal measures of judicial engagement that combine video, audio, and automated transcripts. We extract and validate measures of judicial demeanor—(dis)agreement and (in)attentiveness—and attorney confidence for descriptive empirical analysis. Female judges display higher rates of visible agreement, while male judges exhibit higher rates of inattention, particularly toward female attorneys. The gender-engagement pattern is concentrated among private-firm attorneys and disappears for government advocates. Judicial interaction styles have immediate behavioral consequences: hostile or cold interventions reduce attorney confidence, with larger effects for female advocates. These findings highlight interactional processes as a channel for inequality in the judiciary, even when textual records of proceedings appear unbiased.

JEL Codes: J16, K40, C55, D83
Keywords: judicial behavior, gender, appellate courts, multimodal measurement, oral argument

*Elliott Ash, ETH Zurich, ashe@ethz.ch; Aniket Kesari (corresponding author), Fordham University, akesari@fordham.edu. We thank Wei Pin for helpful research assistance.

1 Introduction

Oral argument is one of the few moments in appellate adjudication where judges and advocates interact face-to-face. In this interaction, judges allocate attention, signal evaluation, and shape the direction of the exchange through questions, demeanor, and engagement (Johnson et al., 2006; Jacobi and Sag, 2019). These interactional choices matter. They influence how attorneys present arguments, how they perform in real time, and how judicial authority is exercised in a public, professional setting. Yet most empirical research on judicial behavior relies on written opinions or text transcripts, which capture what is said but not how it is conveyed or how interaction unfolds in real time (Dietrich et al., 2018; Ash et al., 2024).

This paper asks whether judicial engagement during appellate oral argument differs systematically across advocates, and in particular whether judges interact differently with male and female attorneys. We study interactional inequality in appellate courts by measuring how judicial attention and evaluation are expressed through observable behavior, including nonverbal cues and conversational dynamics, and by examining how attorneys respond in the moment. By focusing on interaction rather than votes or outcomes, we examine a channel through which inequality may arise even when formal legal rules are neutral.

To answer this question, we analyze a large collection of oral arguments before the U.S. Court of Appeals for the Ninth Circuit from 2010 through 2025. We construct a scalable multimodal measurement pipeline that integrates video, audio, and automated transcripts to recover who is speaking, how speech is delivered, and how interaction unfolds over time. The pipeline attributes speech and nonverbal behavior to individual judges and attorneys, segments proceedings into speaker turns with time-aligned transcripts, and classifies communicative intent using both verbal and nonverbal cues. It further identifies interruptions and related interactional events, coding the structure, tone, and immediate behavioral responses associated with judicial interventions. These components together produce a structured, time-resolved record of courtroom interaction that captures attention, agreement, skepticism, and attorney confidence, and that can be aggregated for systematic empirical analysis.

We validate the key measurements using ground truth labels generated by AI-generated perturbations of judge characteristics and courtroom behaviors. In synthetic courtroom clips, we hold the underlying interaction fixed while varying participant demographics, verbal content, atomic actions, and intended communicative states. This design allows us to test whether the pipeline responds to observable behavior rather than to demographic artifacts. The validation shows high accuracy in recovering participant gender and race, and strong

performance in classifying communicative intent for both judges and attorneys. Taken together, these results provide evidence that the measurement system is capturing the intended behavioral constructs and that the main classifications are driven by observed conduct rather than identity-based stereotypes, while still leaving room for residual measurement error and model-based bias in real-world footage.

We estimate how judicial engagement varies with judge and attorney gender using within-case comparisons that hold constant the legal and institutional context of each proceeding. The baseline specification includes indicators for judge gender, attorney gender, and their interaction, along with a control for which side is arguing and case fixed effects that absorb all time-invariant features of the dispute and panel. This design separates differences in baseline judicial style from differences in how judges respond to male and female advocates. A complementary specification replaces judge gender with judge fixed effects, isolating within-judge variation and directly testing whether the same judge behaves differently across attorneys. Standard errors are clustered at the judge level and adjusted for multiple testing, and we interpret all estimates as conditional associations.

We find that female judges display systematically higher levels of visible agreement and lower levels of inattentiveness during oral argument. These differences are present across attorney types, but are most pronounced when the attorney is also female. The scope of this pattern depends on the institutional context of the advocate. When we distinguish between private practitioners and government attorneys, the gender–engagement relationship is concentrated entirely among private-firm advocates and is absent for government representatives. This suggests that institutional role shapes how attorneys are received. Government attorneys, whose identity is tied to their office, appear to attenuate or override the gender-related patterns in engagement that are otherwise present. The result is consistent with a compositional interpretation: differences in judicial style interact with the salience of individual versus institutional identity.

Judicial interaction style also has immediate behavioral consequences for attorneys. Hostile or cold interventions are associated with clear declines in attorney confidence, while more supportive or neutral exchanges are associated with stability or improvement. There is suggestive evidence that these effects are larger for female attorneys, indicating that the same form of judicial pressure may have uneven behavioral impacts. These interactional effects operate in real time, shaping how attorneys respond during argument, even though they do not appear to translate into differences in case outcomes.

The paper contributes to several literatures on judicial behavior and political economy.

A large body of work studies judges through the lens of the strategic and attitudinal models, emphasizing how judicial decisions reflect institutional constraints and policy preferences (Epstein and Knight, 1997; Segal, 2002; Stephenson, 2009). Empirical research in this tradition has relied primarily on voting patterns, coalition formation, and opinion assignment to study judicial behavior (Martin and Quinn, 2002; Sunstein et al., 2007; Posner, 2010; Epstein et al., 2013). While this work has generated important insights, it focuses almost entirely on outcomes and written products of judging.

A related literature studies discrimination and bias in the legal system, documenting disparities in judicial decisions across race and gender (Fagan and Ash, 2017; Harris and Sen, 2019; Ash et al., 2022). For example, Glynn and Sen (2015) show that judges with more daughters are more likely to vote in favor of women's rights, while Ash et al. (2023) show that judges who use more gender-stereotyped language in their written opinions tend to rule against women's rights and also treat their female colleagues less favorably, citing them less and reversing them more often. This literature has largely inferred bias from decisions or written language. It has not examined behavior during hearings, which are more immediate and less deliberative, and therefore plausibly more prone to latent bias. Outside the courtroom, evidence from legislatures and elite professional settings shows that interactional treatment differs systematically by gender, including in patterns of attention and conversational control (Miller and Sutherland, 2023; Ash et al., 2025; Dupas et al., 2025).

The paper also contributes to the literature on oral argument and transparency in courts. Prior work, largely focused on the U.S. Supreme Court, has emphasized the informational role of oral arguments (Epstein and Kobylka, 1992; Spriggs and Wahlbeck, 1997; Johnson, 2001; Johnson et al., 2006; Samaha et al., 2020), the role of oral argument in judicial deliberation and panel dynamics (Kastellec, 2011; Jacobi and Sag, 2019; Ostberg and Wetstein, 2023), and gender dynamics in questioning and interruption (Jacobi and Sag, 2023). A small number of studies exploit audio recordings to study vocal pitch and emotional intensity, showing that how judges speak predicts how they vote (Dietrich et al., 2019). These perspectives also connect to broader debates about transparency and legitimacy, in which oral argument is viewed as an information-gathering device, a forum for judicial deliberation, and a mechanism of public reason (Rawls, 1973; Segall, 2016; Marder, 2016). We extend this literature by analyzing audiovisual data from a large appellate court and by measuring nonverbal and paralinguistic cues that are central to interaction but absent from text-only data. By providing validated measures of judicial engagement that respond to observable behavior rather than demographic artifacts (Harris and Sen, 2019), we offer new evidence

on how authority and evaluation are expressed in appellate courts and provides tools that can be applied to other deliberative institutions.

Methodologically, the paper builds on a growing literature using audio and video to study political institutions and elite communication. Prior work has examined emotional arousal in legislative speeches (Dietrich et al., 2019), physical movement and cross-party interaction in legislatures (Dietrich, 2021), and treatment of women speakers in economics seminars (Dupas et al., 2025). Work on courts, by contrast, has largely been limited to text analysis with one paper analyzing an emotional arousal dimension in Supreme Court oral argument audio (Dietrich et al., 2019). We integrate visual, acoustic, and textual signals within a single interpretable framework and provide a scalable method for automated behavioral analysis of courtroom interactions.

The rest of the paper proceeds as follows. Section 2 describes the institutional setting and data. Section 3 details the measurement approach. Section 4 presents the validation exercises. Section 5 develops the empirical framework. Section 6 reports results. Section **??** concludes.

2 Institutional Setting and Data

2.1 The Ninth Circuit and Its Video Archive

The United States Court of Appeals for the Ninth Circuit is the largest federal appellate court by geography and caseload, hearing appeals from nine western states and two territories and disposing of roughly 14,000 cases per year. About three to four percent of cases are selected for oral argument.

Oral arguments follow a standard format. Each side typically receives approximately fifteen to twenty minutes. The appellant argues first, then the appellee; appellants may reserve time for rebuttal. A three-judge panel presides, and judges routinely interrupt to pose hypotheticals, press inconsistencies, and signal their reactions through questions and nonverbal cues. Because oral argument is a highly structured but interaction-rich setting in which judges and attorneys engage face-to-face, these videos provide a rare opportunity to study how judicial attention and demeanor are allocated across advocates in real time.

Beginning around 2010, the Ninth Circuit started posting video recordings of these proceedings on its YouTube channel. The archive now contains approximately 15,000 recordings—the most comprehensive public audiovisual record of any federal appellate court. Unlike Supreme Court audio, which has attracted substantial scholarly attention (Dietrich et al.,

2019; Jacobi and Sag, 2019), this video archive has remained largely unexamined. The reason is straightforward: until recently, no scalable method existed for extracting structured behavioral data from courtroom video. The emergence of multimodal AI changes this.

2.2 Sample Construction

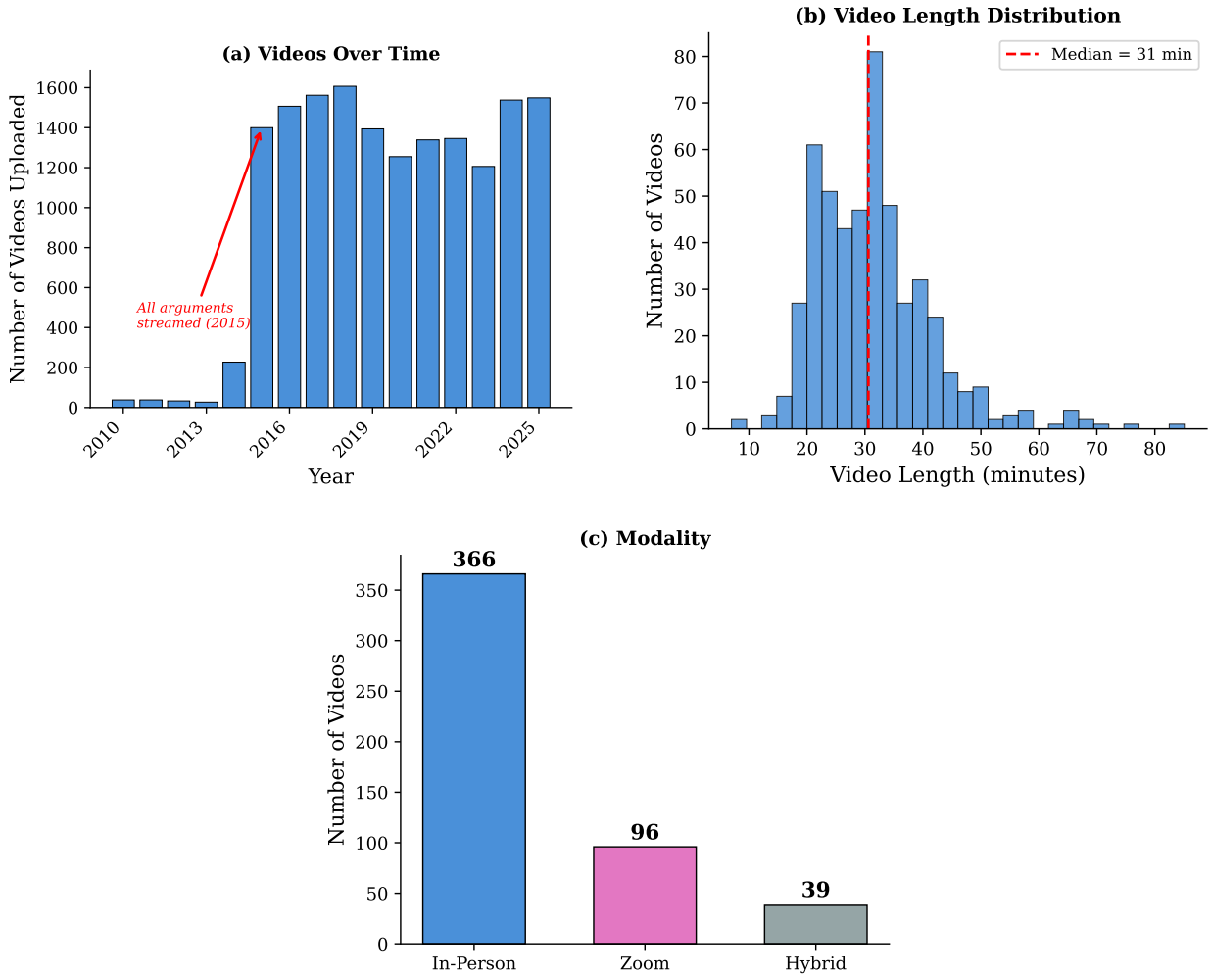
Our sample covers 501 oral argument videos downloaded from the Ninth Circuit's YouTube channel, spanning docket years 2010 through 2025. The majority are from post-2015 arguments, when the Ninth Circuit began livestreaming all oral arguments directly to Youtube.

Of the 501 videos, 366 (73.1%) are in-person arguments, 96 (19.2%) are remote Zoom arguments conducted during and after the COVID-19 pandemic, and 39 (7.8%) are hybrid. The shift to remote argument introduced variation in camera angle, audio quality, and the observability of nonverbal behavior. Figure 1 summarizes the distribution of videos over time, by duration, and by modality.

These videos appear in three principal formats (see Figure 2): (1) split-screen or multi-camera arrangements with fixed views of the panel and podium, common in earlier years; (2) dynamic, operator-controlled camera feeds with shifting focus among participants; and (3) remote Zoom arguments with individual participant windows labeled by name, primarily during the COVID-19 pandemic. While these formats differ visually, all preserve continuous audio and clear speaker attribution cues, enabling a consistent measurement pipeline across time.

Unit of observation. The raw data contain 48,509 actor-by-ten-second-interval observations across the 501 videos: 33,856 at the judge level and 14,653 at the attorney level. We observe 277 unique judge names after cleaning, of which approximately 160 appear in two or more proceedings. The remainder are visiting judges from other circuits. For the main regressions, we aggregate judge-level observations to the judge-by-argument-segment-by-video cell, yielding 2,646 observations in the estimation sample (after excluding observations with missing attorney gender) across 447 videos and 266 judge clusters. The 266 figure refers to the number of unique judges appearing in the estimation sample; standard errors are clustered at this level.

Figure 1: Sample Characteristics of Courtroom Videos



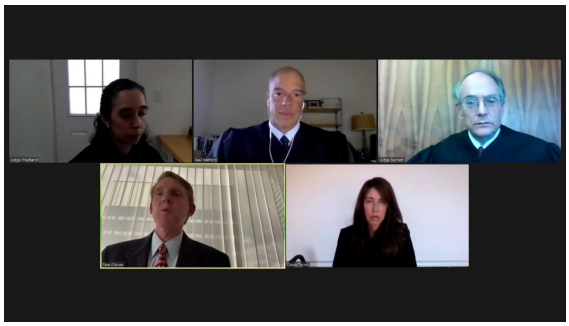
Notes: Summary tabulations on courtroom video dataset: (a) number of videos by year, (b) number of videos by duration (in minutes), (c) number of videos by modality (in-person, Zoom, or hybrid).

Figure 2: Typical Camera Views

(a) Stationary Camera



(b) Zoom Proceeding



Notes. Illustration of the camera positioning and screen segmenting in the Ninth Circuit video data.

2.3 Summary Statistics

Table 1 reports summary statistics for the estimation sample. Female judges account for 37.9% of judge-role observations; female attorneys account for 31.0%. The behavioral probability distributions are heavily concentrated at neutral: mean neutral is 79.7 out of 100, with agreement averaging 3.3, disagreement 11.3, and inattentiveness 5.7. Attorney confidence averages 83.4. The concentration at neutral is expected—judges spend most of oral argument listening without strong affective signals—and means that the regressions are identifying variation in the relatively thin tails of the behavioral distribution.

Table 2 compares observable case characteristics across cases with female and male attorneys. Argument year, modality, panel composition, fraction of Democrat-appointed judges, and mean judge age are all balanced, with no characteristic significantly different at the 10% level. This does not rule out selection on unobservables—attorney quality, argument style, and case strength may still vary with gender—but it does establish that the observable case environment facing female and male attorneys is similar.

3 Measurement

Our objective is to measure interactional judicial behavior during appellate oral argument, with a focus on how judges engage with attorneys in real time. Each proceeding encodes who is speaking, what is said, how it is delivered, and how interaction unfolds over time. Recovering these dimensions requires integrating visual, audio, and textual signals within a single analytical framework.

Table 1: Summary Statistics

Variable	**N**	**Mean**	**SD**	**Min**	**Max**
Panel A: Judge Observations (unit = judge-segment window)					
Female judge (=1)	33,856	0.379	0.485	0	1
Democrat-appointed (=1)	32,614	0.587	0.492	0	1
Minority judge (=1)	33,856	0.294	0.455	0	1
Age (years)	33,749	63.5	12.2	40.0	94.0
P(neutral) (0–100)	33,856	79.70	23.06	0.00	100.00
P(agreement) (0–100)	33,856	3.26	9.36	0.00	100.00
P(disagreement) (0–100)	33,856	11.29	23.63	0.00	100.00
P(inattentiveness) (0–100)	33,856	5.70	10.67	0.00	100.00
Polarity score (agree−disagree)	33,856	-8.03	26.68	-100.00	100.00
Behavior entropy	33,856	0.507	0.331	-0.000	1.522
Panel B: Attorney Observations (unit = attorney-segment window)					
Female attorney (=1)	14,653	0.310	0.462	0	1
Age (years)	664	49.5	10.2	35.0	65.0
P(confidence) (0–100)	14,653	83.39	10.12	10.00	100.00
P(nervousness) (0–100)	14,653	16.61	10.12	0.00	90.00
Polarity score (conf−nerv)	14,653	66.78	20.24	-80.00	100.00
Behavior entropy	14,653	0.601	0.183	-0.000	1.000
Panel C: Case (Video) Level					
Year of argument	480	2022.7	2.4	2011.0	2026.0
In-person modality (=1)	501	0.731	0.444	0	1
Zoom modality (=1)	501	0.192	0.394	0	1
Hybrid modality (=1)	501	0.078	0.268	0	1
Immigration case (=1)	501	0.066	0.248	0	1
Criminal case (=1)	501	0.156	0.363	0	1
Appellant wins (=1)	251	0.183	0.388	0	1

Notes: Panels A and B summarize individual actor-segment observations as extracted by the Gemini multimodal pipeline. Behavioral probabilities are reported on the 0–100 scale returned by the model; they sum to 100 within a given observation by construction. Polarity score equals $P(\text{agreement}) - P(\text{disagreement})$ for judges and $P(\text{confidence}) - P(\text{nervousness})$ for attorneys. Behavior entropy is the Shannon entropy of the probability vector (higher = more ambiguous classification). Panel C reports video-level (case) characteristics. "Appellant wins" is coded from the published circuit court opinion; observations with ambiguous outcomes are excluded. The full sample contains 501 Ninth Circuit oral arguments (2015–2026).

Table 2: Balance Table: Female- vs. Male-Attorney Cases

	Female-attorney cases	**Male-attorney cases**	**Full sample**	***p*-value**
N videos	132	369	501	
Year of argument	2022.750 (2.413)	2022.700 (2.438)	2022.712 (2.422)	0.924
Zoom modality (=1)	0.159 (0.367)	0.203 (0.403)	0.192 (0.394)	0.249
In-person modality (=1)	0.765 (0.426)	0.718 (0.451)	0.731 (0.444)	0.285
Hybrid modality (=1)	0.076 (0.266)	0.079 (0.269)	0.078 (0.268)	0.917
Pct. female judges on panel	0.402 (0.293)	0.391 (0.279)	0.394 (0.282)	0.707
All-male panel (=1)	0.237 (0.427)	0.229 (0.421)	0.231 (0.422)	0.865
Has $\geq$1 female judge (=1)	0.763 (0.427)	0.771 (0.421)	0.769 (0.422)	0.865
N judges on panel	3.000 (0.215)	2.994 (0.074)	2.996 (0.127)	0.774
Pct. Democrat-appointed judges	0.571 (0.300)	0.577 (0.302)	0.576 (0.301)	0.839
Mean judge age (years)	63.777 (6.629)	64.011 (6.628)	63.949 (6.623)	0.731

Notes: Unit of observation is the video (case). A case is classified as a "female-attorney case" if more than 50% of observed attorney turns in that video are by female attorneys. Cells report mean (standard deviation). *p*-values from two-sample Welch *t*-tests comparing female- to male-attorney cases. All characteristics are pre-determined (measured before or independently of the argument), supporting identification. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

We organize the measurement approach in four stages. First, we identify judges and attorneys and attribute observed speech and nonverbal behavior to specific participants. Second, we segment the audio stream into speaker turns and generate verbatim transcripts with speaker attribution, which are used to characterize tone and content. Third, we classify nonverbal communication—body language, facial expressions, vocal tone—into theoretically motivated categories of communicative intent. Fourth, we identify interruptions and related interactional events and classify both judicial interventions and attorney responses. Together, these steps yield a structured record of judicial–attorney interaction suitable for systematic empirical analysis.

The core measurement tool is Google's Gemini multimodal language model.[1] The intuition is similar to what a trained human coder would do while watching a courtroom recording: identify who is speaking, note how the judges are reacting, and characterize the tone of any interruptions. The advantage of AI coding is scale—we process hundreds of hours of video that would be impractical for a human team. Appendix A provides the complete technical specification.

[1]The production pipeline uses Gemini 3.1 Pro Preview.

3.1 Visual Identification of Judges and Attorneys

The first step is participant attribution: linking observed speech and nonverbal behavior to individual judges and attorneys. Accurate attribution is essential, because all downstream measures of attention, demeanor, and interaction are meaningful only when tied to specific actors. Appellate courtroom videos provide strong institutional cues that facilitate this task. Attorneys typically introduce themselves and the parties they represent at the outset of argument, while judges' names appear in on-screen overlays or, in remote proceedings, in Zoom identifiers.

We combine these textual cues with visual and acoustic information to construct a roster of participants for each proceeding. Each individual is labeled by role (judge or attorney) and position (e.g., presiding judge, petitioner's counsel), and judges are linked to external metadata such as appointing president and gender. Identification draws on both institutional structure and stable patterns in speech and appearance across video frames and audio segments.

The attribution procedure integrates rule-based supervision with unsupervised inference. Institutional conventions provide strong priors. The first extended speaker is almost always the petitioner's counsel, the respondent follows, and judges intervene intermittently with shorter utterances. Judges also occupy fixed physical seats in in-person proceedings or stable windows in remote hearings. At the same time, speakers are clustered using distinctive acoustic features and matched visually across frames, which allows reliable attribution even during rapid exchanges or overlapping speech.

This approach does not require case-specific training data or manual annotation. Instead, identification relies on contextual information embedded in the video together with widely available public information about federal appellate judges, whose names, appearances, and speech patterns are documented in court websites and official recordings. When judges' names appear in on-screen panels or Zoom labels, these textual cues anchor attribution and are reinforced by visual facial features. Attorney identification follows the same logic. Attorneys routinely state their names and roles at the beginning of argument, and the structured sequence of appellate argument further allows us to infer party alignment.

3.2 Audio Segmentation and Transcription

The next step is to segment each proceeding's audio into speaker turns and to generate verbatim transcripts with speaker attribution. This task involves two components: identi-

fying speaker changes (diarization) and assigning each speech segment to a specific judge or attorney (classification). Both are necessary to recover the structure of judicial questioning and interaction during oral argument.

We integrate diarization and classification using Gemini in a single pipeline. Prompts instruct the model to detect speaker changes, apply institutional sequencing rules, and associate speech segments with visual information from the video. Because the model processes audio and video jointly, it can recognize individual judges directly using publicly available contextual information, without case-specific training data or manual annotation. The result is a timestamped transcript in which each utterance is attributed to a named judge or attorney. Precise attribution of speaker turns is essential for identifying interruptions and rapid judicial interventions, which are central to our analysis of interactional asymmetries.

For each proceeding, Gemini returns a transcript with precise timestamps and explicit speaker labels. The model is instructed to rely on visual identifiers, institutional sequencing, and conversational dynamics such as uninterrupted speech length and questioning behavior. This procedure preserves the interactional structure of oral argument and produces transcripts suitable for downstream quantitative analysis.

3.3 Measuring Nonverbal Communication

The third task is to capture information that is not conveyed in words but is central to judicial behavior during oral argument. Much of the existing law and psychology literature on judicial behavior in courtroom settings has focused on trial court judges, largely motivated by concerns about bias against vulnerable defendants, with comparatively little attention to interactional dynamics in federal appellate courts, where the primary interlocutors are attorneys (Blanck, 1991; Searcy et al., 2005). Yet this literature makes clear that judges routinely signal skepticism, agreement, or disengagement through posture, gaze, gestures, facial expressions, and tone of voice. Transcripts alone obscure these signals and risk mischaracterizing the nature of an exchange. For example, a short question delivered with a skeptical tone and furrowed brow conveys something quite different from the same words spoken with a neutral affect.

These nonverbal cues are particularly important for understanding interactional inequality in appellate courts. The degree of behavioral alignment between speakers, often described as interactional synchrony, is a key indicator of rapport and affiliation (Nadler, 2012; Miles et al., 2009). Systematic differences in attention, affect, or engagement can shape how attorneys are received and how they respond in real time, even when the spoken content of

questioning is similar. Measuring these cues therefore allows us to analyze not only what judges ask, but how judicial authority and evaluation are expressed during oral argument, which is central to understanding judicial influence in this setting.

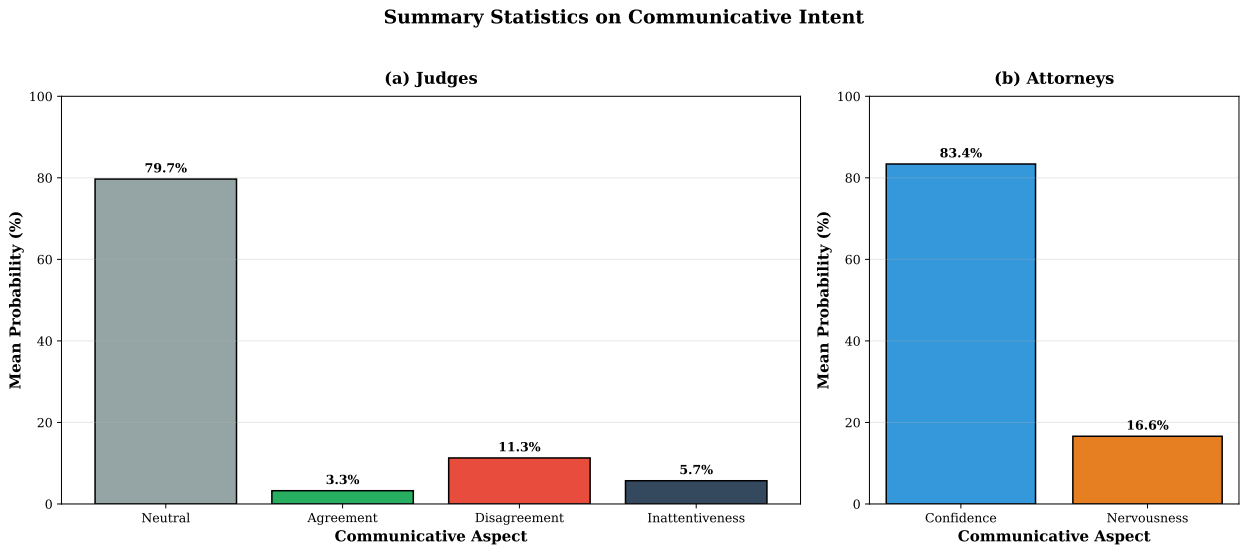
This task presents methodological challenges. Nonverbal signals are fleeting and context-dependent. A nod can signify agreement in one setting or merely attentiveness in another. Traditional approaches require separate supervised classifiers for facial expressions, gesture recognition, or prosody analysis, often trained on narrow datasets that do not generalize well to courtrooms. Gemini's multimodal design offers an alternative. By processing video frame-by-frame in conjunction with the synchronized audio and transcript, the model can align visual, auditory, and textual cues to generate structured annotations.

Prompts instruct the model to provide both objective descriptions (e.g., "leaning forward, arms crossed, low steady tone") and interpretive assessments (e.g., "attentive and probing"). This hybrid approach balances unsupervised observation with guided interpretation. Each annotation is linked to the corresponding transcript timestamp, producing a temporally precise map of judicial demeanor that can be aggregated into measures of engagement, skepticism, or agreement across cases.

One potential issue deserves mention. While Gemini is trained on multimodal data and therefore has captured a wide array of human interactions, it is not necessarily the case that its labels are well-tuned to domain-specific definitions of body language and synchronicity. Courtrooms are unique institutions, and federal appellate courtrooms even more so. An interaction that might be cold or rude in normal conversation could be standard professionalism in a courtroom setting.

Complex behaviors are broken into "atomic behaviors" (e.g., "shuffling papers") to infer communicative intent. Each behavior is then mapped to predefined intent categories. For *judges*, four categories are defined: (i) **neutral engagement**—attentive but noncommittal listening; (ii) **agreement or positive demeanor**—nodding, leaning forward, verbal affirmation; (iii) **disagreement or negative demeanor**—head-shaking, furrowing, skeptical expression; and (iv) **inattentiveness**—gaze averted, writing, looking at materials. For *attorneys*, two categories are defined: **confidence** and **nervousness**, based on posture, voice steadiness, and eye contact. The compositional constraint—categories sum to 100—means the four judge outcomes are mechanically related: a positive coefficient on one outcome must be offset by negative coefficients on the others. We estimate separate OLS regressions for each outcome, note this constraint throughout, and verify key findings with alternative approaches.

Figure 3: Summary Statistics on Communicative Intent



Notes: Distribution of communicative intent categories for judges (four categories: neutral, agreement, disagreement, inattentiveness) and attorneys (two categories: confidence, nervousness). Each observation is a ten-second video window. Probabilities sum to 100 within each observation by construction.

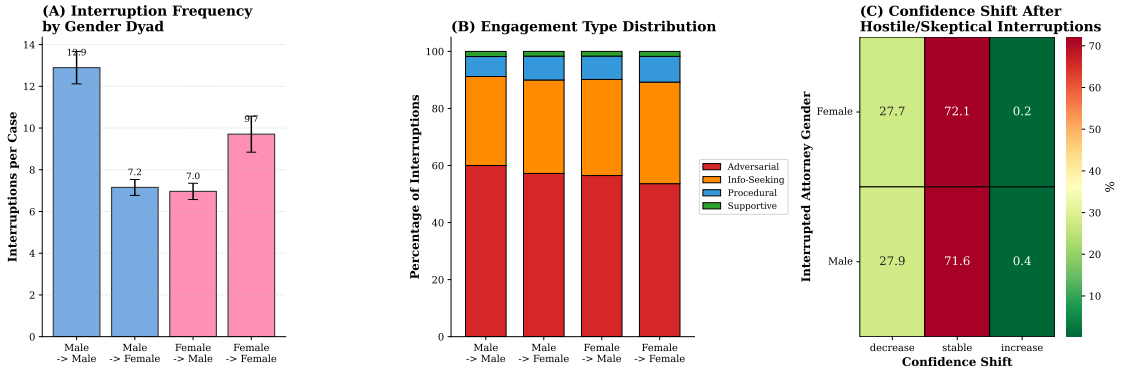
Summary statistics on the communicative intent behaviors are shown in Figure 3. For judges, the most common behavioral intent is neutral, followed by disagreement, then agreement, then inattentiveness. For attorneys, the behaviors are mostly confident with a sizable share of nervousness.

3.4 Identifying and Classifying Interruptions

We next identify judicial interruptions, which are salient interactional events in institutional settings such as courts and legislatures (Dietrich et al., 2019; Miller and Sutherland, 2023; Dupas et al., 2025). We begin by locating all speaker transitions in the transcripts. For each transition, we extract a short video window spanning 15 seconds before and after the change in speaker, which provides the immediate conversational and nonverbal context.

Using these clips, we prompt Gemini to classify each speaker transition along four dimensions: turn mechanics (cutoff, overlap, pause entry, clean yield), engagement type (challenge, press, hypothetical, clarification, redirect, assist, procedural, acknowledgment), affective tone (hostile, skeptical, neutral, supportive, warm, patronizing), and attorney confidence shift (increased, stable, decreased). The interruption data contain 23,668 events. Summary statistics are reported in Figure 4. Table 3 provides examples of how the automated interruption anal-

Figure 4: Summary Statistics on Interruptions



Notes: Distribution of 23,668 interruption events across 357 oral arguments, by turn type, engagement type, and affective tone. Judge-to-attorney interruptions are the dominant type; neutral and skeptical tones predominate.

ysis detects and classifies interruption dynamics, with links to the corresponding YouTube videos showing the behavior described by Gemini.

As a validation exercise, we replicate a well-documented empirical regularity from the U.S. Supreme Court: the party that ultimately loses the case is questioned more frequently during oral argument (Johnson et al., 2006). Applying our interruption and questioning measures to Ninth Circuit proceedings, we find the same pattern. As shown in Figure 5, the losing party receives more questions in approximately 75 percent of cases.

3.5 Demographic Attributes

Gender, race, and (for judges) party of appointing president are recorded for each actor. For judges, these attributes are estimated with a combination of Gemini's analysis and its knowledge of public figures such as judges. For attorneys, all demographics are Gemini-estimated. This creates a measurement concern we take seriously: the conceptual variable of interest is the gender that observing judges perceive. But the reliance on a single model for both treatment and outcome measurement raises the possibility of correlated error, which we address in the validation exercise and the empirical framework.

4 Validation of Visual LLM Outputs

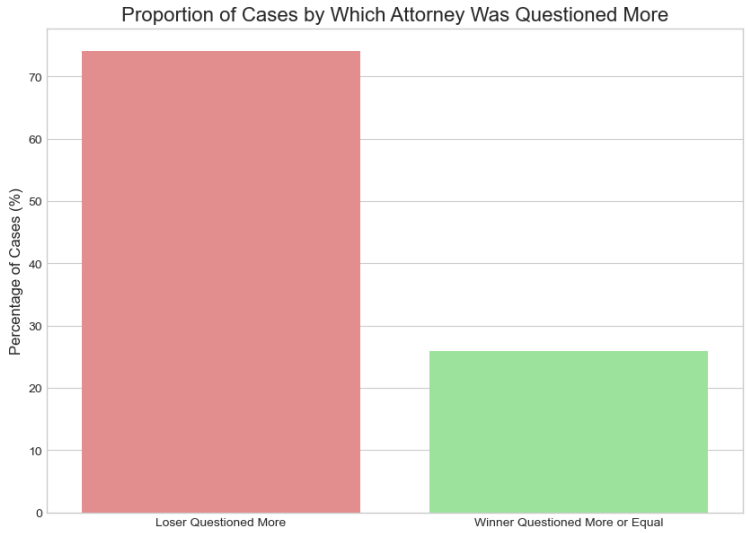
One concern with any AI-based pipeline for data extraction is that the model may be biased both in how it detects different demographic features and in the inferences it makes about

Table 3: Interruption Examples Showing Multimodal Classification

Case	Type	Tone	From	To	Reaction	Gemini Explanation	Link
Sammy Page	Cutoff	Hostile	J(M)	A(F)	Stops	The judge interrupts the attorney mid-sentence while she is attempting to explain, cutting her off to demand a "Simple question" and subsequently a "Yes or no" answer. This is a directive interruption intended to force a specific format of response rather than allowing the speaker to finish her thought.	3:30–4:00
Daniel Crowe	–	Supportive	J(M)	A(M)	Continues	The judge's interjection ("Absolutely") occurs at a natural syntactic boundary after the attorney completes his sentence. It functions as a supportive backchannel agreement rather than a floor-taking interruption, and the attorney immediately continues his argument without disruption.	8:45–9:30
Matta-Lopez	Comp. Overlap	Skeptical	J(F)	A(F)	Defers	The attorney begins her response with "I understand that," clearly intending to continue with a counter-argument. The judge immediately re-takes the floor to insert a probing question, preventing the attorney from finishing her thought. The attorney yields the floor instantly to listen to the new question.	4:20–4:40

Notes: J = Judge, A = Attorney, M = Male, F = Female. Comp. = Competitive. Tone and interruption type determined from multimodal analysis (visual + audio + text). N = 23,668 interruptions from 357 videos.

Figure 5: More Interruptions Are Associated with Higher Lose Rate



Notes: Fraction of cases in which the party receiving more judicial questions ultimately lost. Based on cases with binary outcome data and identified winner/loser.

members of particular groups. For example, Salinas et al. (2025) find that GPT-4 gives worse advice to simulated users with Black female-sounding names. Other work has shown that automated transcription models such as Whisper have higher word error rates for Black speakers relative to white speakers and can hallucinate disproportionately when transcribing audio from speakers with aphasia (Koenecke et al., 2024). These findings motivate careful validation of whether our measurement pipeline responds to observable behavior rather than to demographic artifacts.

To start to address these potential issues, we validate the accuracy of the measurement pipeline and assess potential demographic bias using a synthetic data design. The core idea is to hold the underlying interaction fixed while exogenously varying participant identity and communicative intent. This approach allows us to evaluate whether the model's classifications respond to observable behavior rather than to spurious correlations with demographic attributes.

We generate short courtroom video clips using the generative video model *Veo*. Each synthetic clip lasts eight seconds. We randomize what is said, prompting the model to have the attorney either say "The remaining dispute in this case is very narrow, but has extraordinary legal consequences..." or complete gibberish. We also randomize the atomic action each attorney and judge does, as well as the communicative intent (agreement, disagreement, inattentive, neutral for judges and confidence, nervousness for attorneys). We separate actions

for communicative intent to test whether Gemini can distinguish between the same action delivered with different meanings (e.g. a supportive head nod versus a skeptical head nod). Finally, we randomly perturb participant characteristics, including age, race and gender.

We first assess whether the pipeline correctly identifies participant demographics. For each clip, we compare predicted labels to the known ground truth used in generation. Accuracy is high across both dimensions. Figure 6 reports confusion matrices for gender (Panel A) and race (Panel B), separately for judges and attorneys. Classification errors are rare and do not vary systematically by depicted gender or race, providing evidence against demographic bias in the measurement tool.

We then evaluate the classification of nonverbal communicative intent. Holding participant identity and actions fixed, we randomize the intended affective state of the interaction, such as agreement, disagreement, or inattention. We compare the model's predicted intent labels to the known synthetic ground truth. Confusion matrices for these predictions are shown in Figure 7. Across 32 actor-interval observations, Gemini correctly classified the ground-truth behavioral category in 26 cases (81.2% accuracy). Accuracy is high across categories, with few misclassifications.

Together, these validation exercises provide evidence that the pipeline accurately recovers both demographic attributes and nonverbal intent, and that its classifications are driven by observable behavior rather than by identity-based artifacts.

5 Empirical Framework

The primary regression dataset aggregates the 33,856 judge-level ten-second observations to the judge × argument-segment × video cell. Let i index judges, j index argument segments (appellant or appellee), and c index cases. The outcome $\bar{y}_{ijc}$ is the mean behavioral probability over all observations of judge i during segment j of case c. This aggregation yields an estimation sample of 2,646 cells across 447 videos and 266 judge clusters (after excluding cells with missing attorney gender).

The main estimating equation is:

$$\begin{aligned}\bar{y}_{ijc} = \alpha_c &+ \beta_1 \,\text{FemaleJudge}_i + \beta_2 \,\text{FemaleAttorney}_{jc} \\ &+ \beta_3 \,(\text{FemaleJudge}_i \times \text{FemaleAttorney}_{jc}) + \gamma \,\text{SegmentTwo}_{jc} + \varepsilon_{ijc}\end{aligned} \tag{1}$$

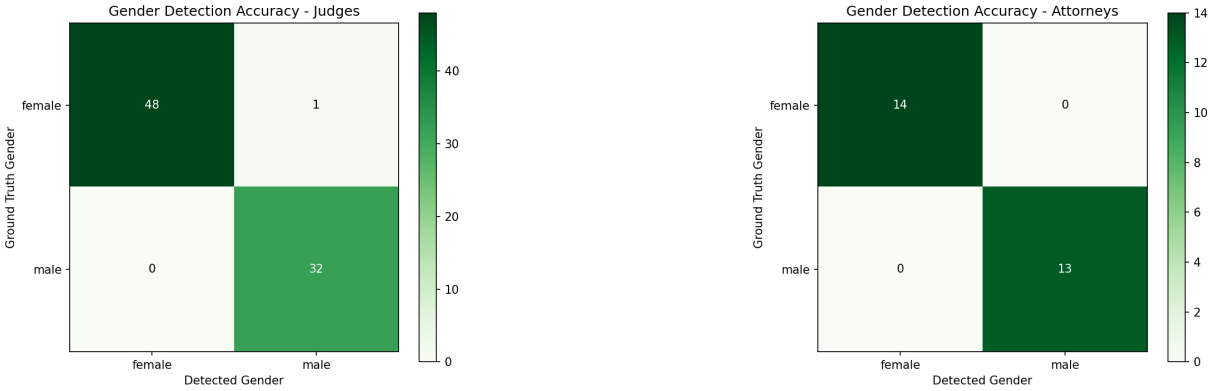
where α_c is a case fixed effect, $\text{FemaleJudge}_i = 1$ if the judge is coded as female, $\text{FemaleAttorney}_{jc} = 1$ if the arguing attorney in segment j of case c is coded as female, and SegmentTwo_{jc} indi-

Figure 6: Demographic Detection Accuracy Using Synthetic Validation

(a) Panel A: Gender

(b) Judges

(c) Attorneys



(d) Panel B: Race

(e) Judges

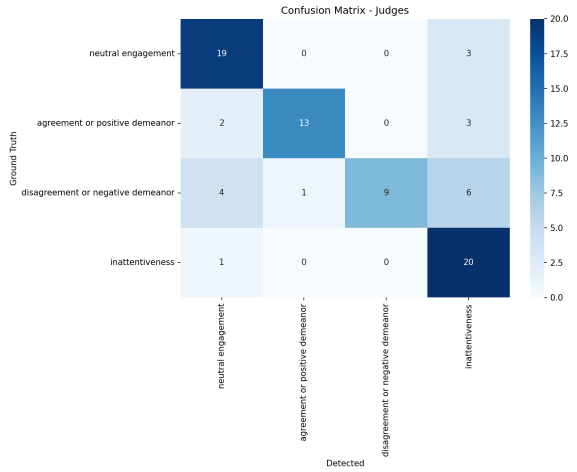
(f) Attorneys



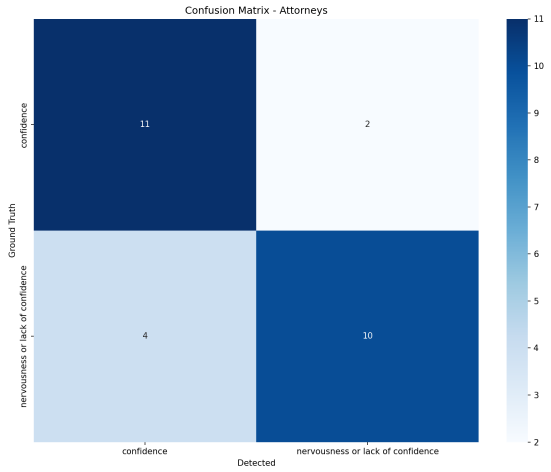
Notes: Confusion matrices for demographic classification of judges and attorneys in synthetic courtroom video clips generated by Veo. Each clip holds the underlying interaction constant while varying depicted gender and race. Ground truth is the demographic specification used in video generation.

Figure 7: Communicative Intent Detection Accuracy Using Synthetic Validation

(a) Judge Communicative Intent



(b) Attorney Communicative Intent



Notes: Confusion matrices for communicative intent classification in synthetic courtroom video clips. Ground truth is the intended affective state specified during video generation. Judges are classified into four categories (neutral, agreement, disagreement, inattentiveness); attorneys into two (confidence, nervousness).

cates the appellee's argument. We estimate Equation (1) separately for each of the behavioral outcomes by OLS.

The coefficients of interest are β_1 (whether female judges differ from male judges in baseline engagement), β_2 (whether judges engage differently with female versus male attorneys), and β_3 (whether the female-attorney differential depends on judge gender). Case fixed effects absorb all time-invariant case characteristics—type, subject matter, complexity, and panel composition.

The design identifies within-case conditional associations between attorney gender and judge behavior. Case fixed effects absorb case-level confounders, and the within-case comparison—the same panel hearing the same case, but different attorneys arguing for different sides—provides useful variation. But the results should not be interpreted as causal. Judge and attorney gender may correlate with other characteristics of the judge, attorney, and case.

6 Results

6.1 Gender and Judicial Engagement

Before turning to regression analysis, we document the basic features of the data. The behavioral probability distributions are heavily right-skewed: most observations receive high neutral scores, with agreement and inattentiveness activated only episodically. Judges spend most of oral argument listening attentively without strong affective signals.

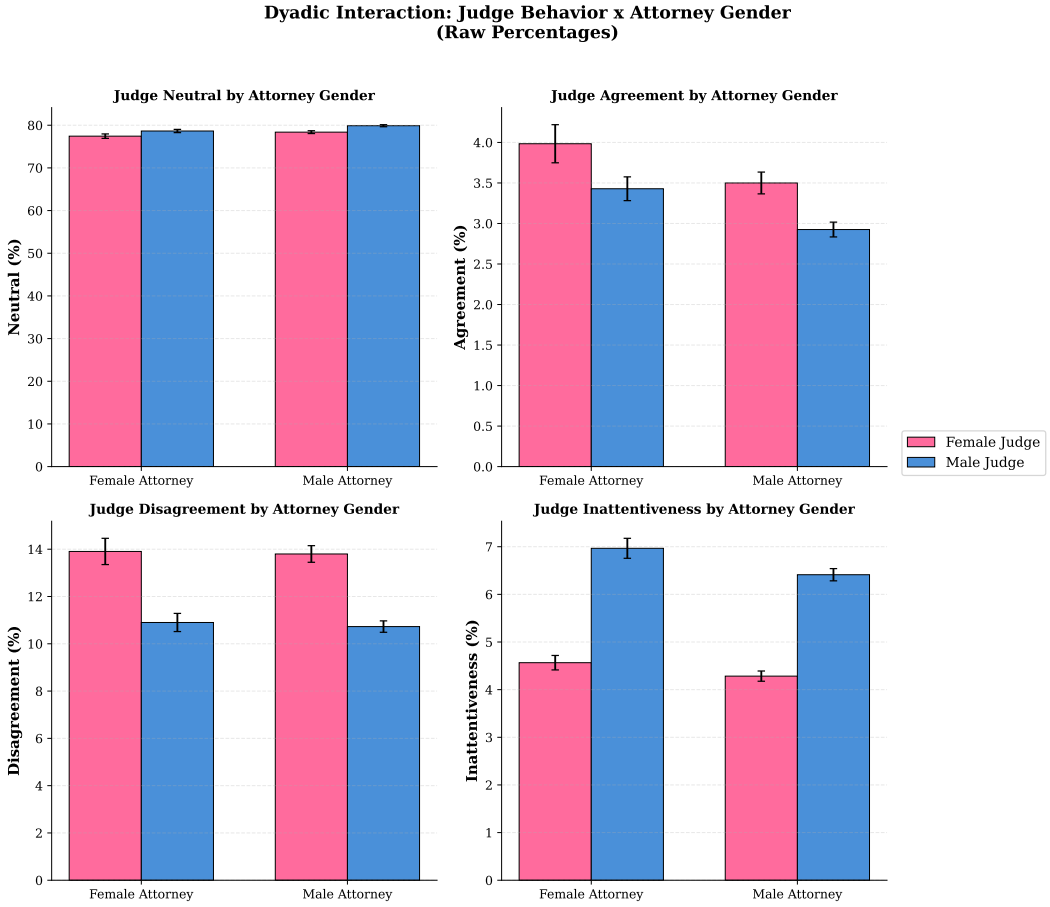
Figure 8 provides a descriptive preview of the main finding. Mean agreement and inattentiveness scores are plotted separately by judge gender and attorney gender. Female judges display higher agreement than male judges across both attorney genders, and the gap is largest when the attorney is also female. The reverse pattern holds for inattentiveness: female judges are less inattentive overall, with the smallest inattentiveness scores observed in female-judge–female-attorney pairs.

Table 4 presents the main regression results. Each panel reports estimates for one of the four behavioral outcomes. Model (1) is the baseline with video fixed effects. Model (2) adds controls for case type, modality, and year. The most striking finding is a substantial and highly significant reduction in inattentiveness among female judges (Panel C). Given that these behavioral categories sum to 100, this reduction in disengagement represents a reallocation toward more active participation; female judges exhibit significantly higher rates of both agreement (Panel A) and disagreement (Panel B) compared to their male colleagues. These results suggest that female judges maintain a more active presence on the bench, a pattern that remains robust after controlling for case characteristics, modality, and time trends.

The data further show that while judges do not systematically vary their level of disagreement or inattentiveness based on the gender of the attorney, there is a distinct same-gender effect in affirmative engagement. The positive and significant interaction term in Panel A indicates that the tendency for female judges to signal agreement is further amplified when interacting with female attorneys. Collectively, these results point to a systematic difference in judicial style, characterized by higher overall engagement from female judges that is particularly pronounced in same-gender dyads.

We find that the most important dimension of heterogeneity is the attorney's institutional affiliation. As reported in Table 5, the baseline gender-attention pattern is driven entirely by interactions with private practitioners and disappears when the arguing attorney represents a government entity. A triple-interaction specification confirms this difference, yielding a

Figure 8: Effect of Judge and Attorney Gender on Judge Nonverbal Behavior



Notes: Mean agreement and inattentiveness probabilities (0–100 scale) by judge gender and attorney gender, with 95% confidence intervals. Aggregated to judge-segment-video cells ($N = 2{,}646$).

Table 4: Gender and Judicial Attention: Main Regression Results

	(1) Baseline	(2) Controls
Panel A: Agreement		
Female judge	0.871***	0.871***
	(0.258)	(0.259)
Female attorney	0.057	0.057
	(0.281)	(0.281)
Female judge × female attorney	0.987**	0.987**
	(0.351)	(0.351)
Panel B: Disagreement		
Female judge	3.835*	3.835*
	(1.675)	(1.678)
Female attorney	−0.463	−0.463
	(0.912)	(0.913)
Female judge × female attorney	−0.630	−0.630
	(1.423)	(1.425)
Panel C: Inattentiveness		
Female judge	−3.007***	−3.007***
	(0.639)	(0.640)
Female attorney	0.004	0.004
	(0.286)	(0.287)
Female judge × female attorney	−0.638	−0.638
	(0.513)	(0.514)
Video (case) fixed effects	Yes	Yes
Case type, modality, year	No	Yes
SE clustering	Judge	Judge
N	2,646	2,646
Videos (FEs)	447	447

Notes: OLS estimates. The unit of observation is a judge × argument-segment × video cell mean. Outcome variables are probability scores (0–100 scale) from the behavioral classification model; the four outcomes sum to 100 by construction. All specifications absorb video (case) fixed effects via within-group demeaning and include a second-segment indicator. Column (2) additionally controls for case-type dummies, modality dummies (Zoom, hybrid; baseline: in-person), and a mean-centered year trend. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

negative and significant coefficient for the female-pair interaction when the attorney is a government representative.

These results suggest that institutional identity—such as representing the Department of Justice or a State Attorney General—serves as an "imprimatur" that overrides individual gender signals. When attorneys appear as institutional agents, the bench's engagement style becomes uniform. However, when private practitioners appear, judicial engagement revert to the gendered patterns documented in our main results. This finding also serves as a critical validation of our empirical strategy: if our AI-based measurement pipeline were merely reflecting algorithmic bias against female voices, we would expect to see consistent patterns across both subsamples. The fact that the effect vanishes in a theoretically predictable context—where professional role supersedes personal identity—strongly suggests our pipeline is capturing genuine behavioral variation.

6.2 The Hostile Interruption Penalty

We next examine how judicial intervention styles influence attorney behavior. Table 6 reports the results of an interruption-level analysis, where the outcome is the shift in attorney confidence following a judicial interruption. The results indicate that the affective tone of a judge's interruption is a primary determinant of subsequent attorney confidence. Figure 9 displays the distribution of confidence effects across tone categories, showing that the concentration of confidence declines is strongest in the hostile category. Figure 10 shows that hostile interruptions reduce confidence for both male and female attorneys, but the reduction is visibly larger for female attorneys. Hostile and skeptical tones are both associated with statistically significant declines in confidence. The effect of a hostile tone is particularly substantial, representing a sharp downward shift in an attorney's composure relative to neutral interruptions.

The impact of these interruptions varies significantly by attorney gender. While hostile judicial tones reduce confidence across the board, this effect is significantly amplified for female attorneys. As shown in Table 6, the interaction between female attorney and hostile tone is negative and statistically significant, suggesting that women experience a disproportionately larger confidence decline when facing hostile interventions.

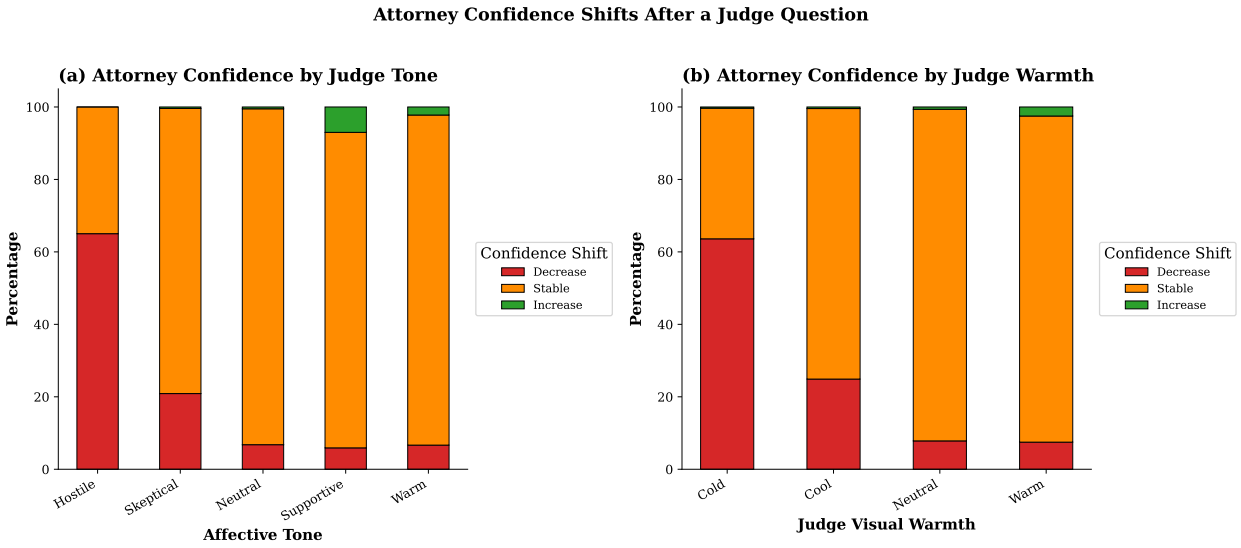
These patterns are robust to the inclusion of judge gender controls, which do not materially alter the baseline relationship between hostile interruptions and attorney response. While the magnitude of the interaction effect is notable, we characterize this finding as suggestive given the limited number of hostile interactions involving female attorneys in the

Table 5: Heterogeneity by Attorney Counsel Type: Gender and Judicial Agreement

	(1) Full	(2) Govt Atty	(3) Private Atty	(4) No Govt	(5) Triple Interaction
Outcome: Prob. Agreement (0–100 scale)					
Female Judge	0.871***	0.649	0.928**	0.901***	0.999**
	(0.258)	(0.731)	(0.422)	(0.287)	(0.436)
Female Attorney	0.057	6.152***	−0.854	−0.243	−0.378
	(0.281)	(1.508)	(0.934)	(0.357)	(0.734)
Female Judge × Female Atty	0.987***	−0.940	1.771**	1.433***	1.516**
	(0.351)	(0.861)	(0.805)	(0.413)	(0.771)
Triple interaction terms (Col. 5 only)					
Atty = Government	—	—	—	—	−0.597
					(0.572)
Female Judge × Govt	—	—	—	—	−0.257
					(0.777)
Female Attorney × Govt	—	—	—	—	1.387
					(0.994)
F.Judge × F.Atty × Govt	—	—	—	—	−2.769***
					(1.028)
N	2,646	311	708	2,039	1,019
Videos	447	101	158	345	205
Judge clusters	266	131	171	232	195
Video FEs	Yes	Yes	Yes	Yes	Yes
SE: clustered by judge	Yes	Yes	Yes	Yes	Yes

Notes: OLS with video fixed effects. Unit of observation is a judge × argument-segment × video cell mean. Col. (1) is the full estimation sample. Col. (2) restricts to observations where the attorney presenting is classified as a government attorney (appellant = government if argument_segment = attorney_1, etc.). Col. (3) restricts to private-firm attorneys. Col. (4) restricts to videos with no government attorney on either side. Col. (5) estimates a triple interaction model on the government/private-firm subsample. Standard errors clustered by judge. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$. Results marked with *fragile* have fewer than 30 judge clusters.

Figure 9: Attorney Confidence Shifts After a Judge Question



Notes: Distribution of attorney confidence shifts (increased, stable, decreased) following judicial interruptions, by affective tone category. Based on 23,668 interruption events.

sample. Nevertheless, the direction of the result is consistent with existing research regarding gendered responses to evaluative pressure in high-stakes environments.

7 Conclusion

This paper develops a scalable, multimodal approach to measuring judicial behavior during appellate oral argument. Using publicly available Ninth Circuit courtroom videos from 2010 through 2025, we combine visual, audio, and textual information to recover who speaks, how they speak, and how interactions unfold in real time. These measures capture dimensions of judging that are largely absent from written opinions and transcripts, yet central to persuasion, authority, and professional evaluation in appellate courts.

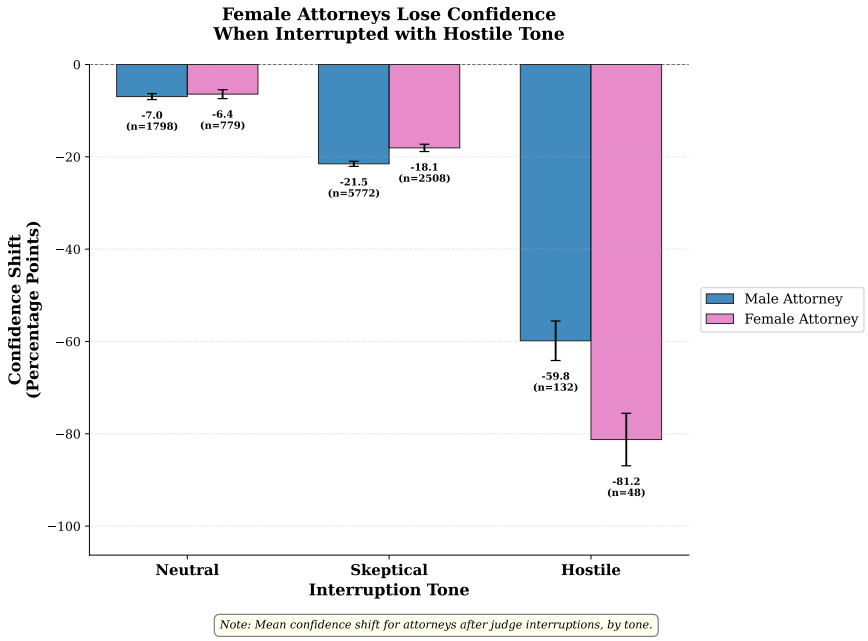
Our findings show that nonverbal judicial behavior is systematic and socially differentiated. Female judges display higher rates of visible agreement, while male judges exhibit higher rates of inattention, particularly toward female attorneys. Judicial style also has immediate behavioral consequences. Hostile interruptions are associated with sharp declines in attorney confidence, with substantially larger effects for female attorneys. Using a synthetic validation design that holds interactions fixed while varying visible identity, we find that the measurement pipeline responds to observable behavior rather than demographic stereo-

Table 6: Attorney Confidence Following Judicial Interruptions

	(1) Baseline	(2) Judge Gender	(3) Pooled Tone	(4) PooledJudge
Female Attorney	0.0243	0.0107	0.0227	0.0079
	(0.0187)	(0.0217)	(0.0189)	(0.0220)
Hostile Tone	-0.3225^{***}	-0.3230^{***}	—	—
	(0.0508)	(0.0508)		
Skeptical Tone	-0.0561^{***}	-0.0563^{***}	—	—
	(0.0142)	(0.0142)		
Warm Tone	0.0361	0.0356	—	—
	(0.0235)	(0.0235)		
Supportive Tone	0.0390	0.0380	—	—
	(0.0301)	(0.0301)		
Female × Hostile	-0.1871^{**}	-0.1841^{**}	—	—
	(0.0853)	(0.0854)		
Female × Skeptical	0.0200	0.0200	—	—
	(0.0184)	(0.0184)		
Negative Tone (Hostile + Skep.)	—	—	-0.0649^{***}	-0.0651^{***}
			(0.0140)	(0.0140)
Female × Negative Tone	—	—	0.0183	0.0184
			(0.0190)	(0.0190)
Female Judge (Interrupter)	—	−0.0051	—	−0.0027
		(0.0147)		(0.0146)
Female Judge × Female Att.	—	0.0298	—	0.0321
		(0.0227)		(0.0231)
Turn mechanics FE	Yes	Yes	Yes	Yes
Engagement type FE	Yes	Yes	Yes	Yes
Video FE	Yes	Yes	Yes	Yes
SE clustering	Video	Video	Video	Video
N	11620	11620	11620	11620
R^2	0.171	0.171	0.161	0.161

Notes: OLS estimates. Outcome is attorney confidence shift (−1=decrease, 0=stable, 1=increase) following a judicial interruption. Sample includes all judge→attorney transitions. Baseline affective tone is neutral. Female Attorney and Female Judge are binary indicators. Models (2) and (4) restricted to observations with known interrupter gender. Standard errors clustered by case (video) in parentheses. Associations, not causal estimates. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Figure 10: Effect of Judge Tone on Attorney Confidence, by Attorney Gender



Notes: Mean attorney confidence shift (-1 = decrease, 0 = stable, $+1$ = increase) following judicial interruptions, by attorney gender and affective tone category. Hostile interruptions reduce confidence for both genders, with a larger decline for female attorneys.

types, while still requiring careful interpretation given the distinctive norms of courtroom interaction.

Substantively, these results point to interactional inequality as a meaningful dimension of appellate adjudication. Differences in attention, tone, and interruption style shape how attorneys experience oral argument and how they respond in the moment, even when formal legal rules and outcomes are held constant. This suggests that inequality in judicial institutions can operate through everyday interactional processes rather than through explicit doctrinal distinctions or case outcomes. By making these processes observable, the paper contributes to a more complete understanding of how authority and evaluation are produced in appellate courts.

The findings also speak to active policy and practice debates. The expanding use of cameras in courtrooms has raised questions about transparency, judicial behavior, and institutional legitimacy. By linking video access to measurable patterns of engagement and interaction, this work provides evidence that can inform discussions about courtroom recording and appellate procedure. For legal practitioners, the results offer insight into how judicial style shapes attorney responses, with implications for advocacy strategies and professional

training. More broadly, in a period of fragile trust in courts, systematic evidence on how judging looks and sounds in practice may help bridge the gap between legal institutions and the public.

Methodologically, the paper introduces a general toolkit for annotating and analyzing audiovisual court data that can be extended to other courts and adapted to settings such as legislatures, administrative hearings, or public deliberative bodies. By showing that multimodal records can be transformed into validated social-science measures, the study opens new avenues for interdisciplinary research on law, institutions, and interaction, and provides a foundation for future work on interactional dynamics that have long been central to theory but difficult to measure empirically.

References

Ash, E., S. Asher, A. Bhowmick, S. Bhupatiraju, D. Chen, T. Devi, C. Goessmann, P. Novosad, and B. Siddiqi (2022). In-group bias in the indian judiciary.

Ash, E., D. L. Chen, and A. Ornaghi (2023). Gender attitudes in the judiciary: Evidence from us circuit courts. *American Economic Journal: Applied Economics*.

Ash, E., D. L. Chen, and A. Ornaghi (2024, January). Gender attitudes in the judiciary: Evidence from us circuit courts. *American Economic Journal: Applied Economics 16*(1), 314–50.

Ash, E., J. Krümmel, and J. B. Slapin (2025). Gender and reactions to speeches in german parliamentary debates. *American Journal of Political Science 69*(3), 866–880.

Blanck, P. D. (1991). What empirical research tells us: Studying judges' and juries' behavior. *American University Law Review 40*(4), 775–799. Symposium Issue on the Selection and Function of the Modern Jury.

Dietrich, B. J. (2021). Using Motion Detection to Measure Social Polarization in the U.S. House of Representatives. *Political Analysis 29*(2), 250–259.

Dietrich, B. J., R. D. Enos, and M. Sen (2018). Emotional arousal predicts voting on the us supreme court. *Political Analysis 26*(3), 305–320.

Dietrich, B. J., R. D. Enos, and M. Sen (2019). Emotional Arousal Predicts Voting on the U.S. Supreme Court. *Political Analysis 27*(2), 237–243.

Dietrich, B. J., M. Hayes, and D. Z. O'Brien (2019). Pitch Perfect: Vocal Pitch and the Emotional Intensity of Congressional Speech. *American Political Science Review 113*(4), 941–962.

Dupas, P., A. Handlan, A. S. Modestino, M. Niederle, M. Seré, H. Sheng, J. Wolfers, S. D. Collective, et al. (2025). Gender differences in economics seminars. *American Economic Review*.

Epstein, L. and J. Knight (1997). *The choices justices make*. Sage.

Epstein, L. and J. F. Kobylka (1992). *The Supreme Court and Legal Change: Abortion and the Death Penalty*. Chapel Hill, NC: University of North Carolina Press.

Epstein, L., W. M. Landes, and R. A. Posner (2013). *The Behavior of Federal Judges: A Theoretical and Empirical Study of Rational Choice*. Cambridge, MA: Harvard University Press.

Fagan, J. and E. Ash (2017). New policing, new segregation: from ferguson to new york. *Geo. LJ Online 106*, 33.

Glynn, A. N. and M. Sen (2015). Identifying judicial empathy: Does having daughters cause judges to rule for women's issues? *American Journal of Political Science 59*(1), 37–54.

Harris, A. P. and M. Sen (2019). Bias and judging. *Annual Review of Political Science 22*, 241–259.

Jacobi, T. and M. Sag (2019). The new oral argument: Justices as advocates. *Notre Dame Law Review 94*, 1161.

Jacobi, T. and M. Sag (2023). Supreme court interruptions and interventions: The changing role of the chief justice. *Boston University Law Review 103*, 1741.

Johnson, T. R. (2001). Information, oral arguments, and supreme court decision making. *American Politics Research 29*(4), 331–351.

Johnson, T. R., P. J. Wahlbeck, and J. F. Spriggs (2006). The influence of oral arguments on the u.s. supreme court. *American Political Science Review 100*(1), 99–113.

Kastellec, J. P. (2011). Panel composition and voting on the us courts of appeals over time. *Political Research Quarterly 64*(2), 377–391.

Koenecke, A., A. S. G. Choi, K. X. Mei, H. Schellmann, and M. Sloane (2024). Careless whisper: Speech-to-text hallucination harms. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, New York, NY, USA, pp. 1672–1681. Association for Computing Machinery.

Marder, N. S. (2016). The Supreme Court's Transparency: Myth or Reality? *Georgia State University Law Review 32*.

Martin, A. D. and K. M. Quinn (2002). Dynamic ideal point estimation via markov chain monte carlo for the us supreme court, 1953–1999. *Political analysis 10*(2), 134–153.

Miles, L. K., L. K. Nind, and C. N. Macrae (2009). The rhythm of rapport: Interpersonal synchrony and social affiliation. *Journal of Experimental Social Psychology 45*(3), 585–589.

Miller, M. G. and J. L. Sutherland (2023). The effect of gender on interruptions at congressional hearings. *American Political Science Review 117*(1), 103–121.

Nadler, J. (2012). Rapport in negotiation and conflict resolution. *Marquette Law Review 96*(1), 191–220.

Ostberg, C. and M. E. Wetstein (2023). Panel effects on courts around the world.

Posner, R. A. (2010). *How judges think*. Harvard University Press.

Rawls, J. (1973). *A Theory of Justice*. Oxford: Oxford University Press.

Salinas, A., A. Haim, and J. Nyarko (2025). What's in a name? auditing large language models for race and gender bias.

Samaha, A. M., M. Heise, and G. C. Sisk (2020). Inputs and outputs on appeal: An empirical study of briefs, big law, and case complexity. *Journal of empirical legal studies 17*(3), 519–555.

Searcy, M., S. Duck, and P. Blanck (2005). Nonverbal communication in the courtroom and the "appearance" of justice. In R. E. Riggio and R. S. Feldman (Eds.), *Applications of Nonverbal Communication*, pp. 41–61. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Segal, J. A. (2002). *The Supreme Court and the Attitudinal Model Revisited*. Cambridge University Press.

Segall, E. J. (2016). Invisible justices: How our highest court hides from the american people. *Georgia State University Law Review 32*(4).

Spriggs, J. F. and P. J. Wahlbeck (1997). Amicus curiae and the role of information at the supreme court. *Political Research Quarterly 50*(2), 365–386. Accessed 3 Jan. 2025.

Stephenson, M. C. (2009). Legal realism for economists. *Journal of Economic Perspectives 23*(2), 191–211.

Sunstein, C. R., D. Schkade, and L. Ellman (2007). *Are judges political?: an empirical analysis of the federal judiciary*. Bloomsbury Publishing USA.

A Appendix A: Multimodal Measurement Pipeline

This appendix provides a complete technical description of the pipeline used to extract behavioral measures from oral argument videos.

A.1 Overview and Architecture

The pipeline processes one oral argument video at a time. Each video is a publicly available YouTube recording from the Ninth Circuit Court of Appeals. The full pipeline consists of two primary extraction stages, a metadata enrichment stage, and a merge step that produces the analysis-ready dataset.

Stage 1 (Behavior Analysis). Each video is divided into approximately 10-second observation windows, one per visible actor. For each window, a prompt is sent to the Gemini API with the video clip, a structured set of behavior definitions, and instructions for producing probability-weighted classification outputs. The output for each window is a vector of six behavioral probabilities: neutral, agreement, disagreement, inattentiveness, confidence, and nervousness. These probabilities sum to 100 and are stored as integers in the range $[0, 100]$. The primary behavior category is the argmax of this distribution. The final Stage 1 dataset contains one row per actor-window observation.

Stage 1c (Interruption Detection). A separate batch pass processes each video and extracts discrete interruption events. Each event records the timestamp, the identities and roles of the interrupter and the interrupted party, and a rich set of behavioral covariates capturing the tone, mechanics, and paralinguistic context of the exchange. This stage uses a prompt that incorporates both the video stream and the automatically generated transcript to improve detection accuracy. The output contains one row per detected interruption event.

Metadata enrichment. A lookup script queries a case information database using extracted docket numbers to attach information on case type, outcome (affirmed, reversed, or remanded), and litigant identity.

Analysis-ready merge. A build script merges Stage 1 behavioral observations with case-level metadata, producing the primary input to all downstream analysis.

A.2 Gemini API and Model Version

Production data extraction uses the `gemini-2.5-pro-preview` model via the Google Generative AI Python client. Data are sent to the API using the Gemini Files API, which accepts video files directly without requiring manual frame extraction. Responses are returned as structured JSON and parsed immediately.

The synthetic-video validation exercise (Section B) used an earlier model version (`gemini-2.5-pro-prev`). Subsequent model updates may have altered classification tendencies. The validation accuracy figures should be interpreted as a lower bound on production pipeline performance.

A.3 Stage 1 Prompt Structure

The Stage 1 prompt presents the model with a video clip and a structured behavioral coding scheme consisting of three components.

Role and task description. The model is instructed to act as an expert observer of courtroom proceedings and to classify the visible behavior of the specified actor during the window.

Behavior category definitions. Six exhaustive behavioral categories are defined:

1. **Neutral**: attentive listening without visible agreement or challenge.
2. **Agreement or positive engagement**: leaning in, nodding, positive facial cues.
3. **Disagreement or negative engagement**: shaking head, furrowing brow, visibly skeptical.
4. **Inattentiveness**: looking away, writing while another speaker talks, distracted posture.
5. **Confidence** (attorney-relevant): standing posture, direct eye contact, controlled gestures.
6. **Nervousness** (attorney-relevant): fidgeting, hesitation, filler words, avoidant gaze.

Output format instruction. The model returns a JSON object containing: (a) probability weights across categories summing to 100; (b) a free-text description of observed behaviors; (c) primary and secondary categories with probabilities; (d) a certainty flag; and (e) a reasoning field.

A.4 Stage 1c Prompt Structure

The interruption detection prompt provides the model with a video segment and the corresponding automated transcript. The model identifies all moments at which one speaker begins speaking while another has not finished, and classifies each event along several dimensions:

- *Interrupter and interrupted parties*: name, role, gender, and race.
- *Turn type*: judge-to-attorney, attorney-to-judge, or judge-to-judge.
- *Turn mechanics*: clean yield, overlap, or cutoff.
- *Engagement type*: challenge, press, hypothetical, clarification, or information request.
- *Affective tone*: hostile, skeptical, neutral, supportive, or warm.
- *Paralinguistic covariates*: vocal intensity, speech rate, intonation (judge-side); vocal confidence, speech fluency, filler words, confidence shift, hedging language, and response latency (attorney-side).
- *Visual covariates*: posture, gesture, gaze, facial expression, and overall demeanor.

A.5 Batch Processing

Videos are processed in parallel batches using the Gemini Batch API. Each batch submission contains up to 100 video requests. The pipeline maintains a file cache mapping local video filenames to Gemini upload URIs. Total API cost was approximately \$200–\$400 for the behavior extraction stage and a similar amount for the interruption detection stage.

A.6 Quality Control

JSON parsing failures. Responses that cannot be parsed as valid JSON are logged and skipped. Parsing failures affect fewer than 2% of requests.

Missing or malformed probability vectors. The pipeline checks that probability values sum to approximately 100 (within ± 5 points). Rows failing this check are flagged and excluded from analyses relying on the full probability distribution.

Actor identity. The pipeline attempts to match each observed actor to a ground-truth roster using face recognition and name matching heuristics. The `gt_matched` flag records successful matches (3.6% of rows). Unmatched actors are assigned position-based identifiers.

B Appendix B: Validation Details

B.1 Synthetic Video Generation

To validate the pipeline's behavioral classification accuracy, we generated synthetic courtroom videos using the Veo video generation API (Google DeepMind). Each video depicts a three-judge panel and one attorney in a stylized courtroom setting. The generation prompt specifies the race, gender, and approximate age of each figure, the behavioral category to be displayed, and the specific physical action to be depicted. A total of 32 synthetic videos were generated, each containing four figures, yielding 128 actor-video observations.

B.2 Validation Procedure

Each synthetic video was passed through the same Stage 1 pipeline used for production data. The pipeline's output classification was compared against the ground-truth behavioral label assigned at generation time. A classification is considered correct if the model's primary behavior category matches the ground-truth label.

B.3 Validation Results

Table 7: Validation Accuracy on Synthetic Videos

Sample	**N**	**Accuracy**
All actor-observations	128	64.1%
Successfully detected rows	108	75.9%
Failed detection (no output)	20	0.0%

Notes: Each row is one actor figure in one synthetic video. Ground-truth behavioral labels were specified at video generation time. The validation run used `gemini-2.5-pro-preview`; production data extraction used `gemini-2.5-pro-preview`. Accuracy figures should be interpreted as a lower bound on production pipeline performance.

The model performs best on neutral engagement and disagreement categories, and weakest on confidence versus agreement discriminations. Classification accuracy does not vary systematically by the race or gender of the depicted figure, suggesting that the model does not apply different classification standards to different demographic groups.

B.4 Validation Limitations

The synthetic videos are stylized rather than photorealistic. The ground-truth labels involve an imperfect mapping from specified physical actions to behavioral categories. The validation used an older model version than the production pipeline. The sample is small (32 videos, 128 observations), limiting statistical precision. These limitations motivate our emphasis on reporting associations with appropriate uncertainty rather than causal interpretations.

C Appendix C: Additional Tables

C.1 Heterogeneity by Case Type

Table 8: Heterogeneity by Case Type: Female Judge × Female Attorney Interaction

	(1) Civil	(2) Criminal	(3) Other	(4) Immigration	(5) Admin.
Panel A: Agreement					
Interaction	1.359*	0.731	0.968	0.317	2.569*
	(0.714)	(0.871)	(0.631)	(1.033)	(1.469)
Panel B: Disagreement					
Interaction	−0.683	0.527	−0.917	2.151	−4.333
	(2.601)	(2.937)	(2.719)	(3.601)	(9.794)
Panel C: Neutral					
Interaction	0.379	−2.525	0.778	−1.321	3.590
	(2.721)	(2.836)	(2.507)	(3.341)	(9.709)
Panel D: Inattentiveness					
Interaction	−1.114	1.263	−0.870	−1.135	−1.826
	(0.891)	(1.564)	(0.719)	(1.000)	(1.725)
N	1,020	411	927	174	114
Judge clusters	164	110	160	62	42

Notes: Each column reports the coefficient on Female Judge × Female Attorney from separate subsample regressions by case type. All specifications include video fixed effects with standard errors clustered by judge. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

C.2 Heterogeneity by Panel Composition

Table 9: Heterogeneity by Panel Composition: Female Judge × Female Attorney Interaction

	(1) All-male panel	(2) Mixed/female panel
Panel A: Agreement		
Interaction	0.000	0.888**
	(0.000)	(0.392)
Panel B: Disagreement		
Interaction	0.000	−0.195
	(0.000)	(1.551)
Panel C: Neutral		
Interaction	0.000	−0.193
	(0.000)	(1.559)
Panel D: Inattentiveness		
Interaction	0.000	−0.527
	(0.000)	(0.562)
N	604	2,042
Judge clusters	108	224

Notes: Each column reports the coefficient on Female Judge × Female Attorney from separate subsample regressions. All-male panels have zero female judges by construction, making the interaction mechanically zero. All specifications include video fixed effects with standard errors clustered by judge. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table 10: Heterogeneity by Time Period: Female Judge × Female Attorney Interaction

	(1) Pre-2020	(2) 2020+
Panel A: Agreement		
Interaction	0.512	1.065***
	(1.125)	(0.406)
Panel B: Disagreement		
Interaction	−10.727**	0.734
	(5.099)	(1.327)
Panel C: Neutral		
Interaction	7.384	−0.735
	(4.836)	(1.380)
Panel D: Inattentiveness		
Interaction	2.789***	−1.091*
	(0.965)	(0.564)
N	363	2,283
Judge clusters	99	224

Notes: Each column reports the coefficient on Female Judge × Female Attorney from separate subsample regressions by time period. Pre-2020 has limited power ($N = 363$, 99 clusters). All specifications include video fixed effects with standard errors clustered by judge. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table 11: Heterogeneity Analysis: Female Judge × Female Attorney Interaction

	Agree		Disagree		Neutral		Inatt.	
	Coef.	SE	Coef.	SE	Coef.	SE	Coef.	SE
civil $N = 1020$ ($J = 164$)	1.359*	(0.714)	−0.683	(2.601)	0.379	(2.721)	−1.114	(0.891)
criminal $N = 411$ ($J = 110$)	0.731	(0.871)	0.527	(2.937)	−2.525	(2.836)	1.263	(1.564)
other_indeterminate $N = 927$ ($J = 160$)	0.968	(0.631)	−0.917	(2.719)	0.778	(2.507)	−0.870	(0.719)
immigration $N = 174$ ($J = 62$)	0.317	(1.033)	2.151	(3.601)	−1.321	(3.341)	−1.135	(1.000)
administrative $N = 114$ ($J = 42$)	2.569*	(1.469)	−4.333	(9.794)	3.590	(9.709)	−1.826	(1.725)
in-person $N = 1944$ ($J = 217$)	1.271***	(0.446)	−2.093	(1.745)	1.130	(1.760)	−0.349	(0.592)
zoom $N = 498$ ($J = 135$)	−1.099	(0.822)	6.068*	(3.307)	−3.286	(3.279)	−1.683	(1.513)
hybrid $N = 204$ ($J = 55$)	2.516	(1.565)	0.180	(4.229)	−1.179	(4.476)	−1.449	(0.909)
Democrat-appointed $N = 1460$ ($J = 143$)	1.885***	(0.522)	0.376	(1.658)	−0.690	(1.750)	−1.602**	(0.709)
Republican-appointed $N = 1093$ ($J = 110$)	0.088	(0.595)	2.669	(3.150)	−3.023	(3.107)	0.226	(0.643)
All-male panel $N = 604$ ($J = 108$)	0.000	(0.000)	0.000	(0.000)	0.000	(0.000)	0.000	(0.000)
Mixed/has-female panel $N = 2042$ ($J = 224$)	0.888**	(0.392)	−0.195	(1.551)	−0.193	(1.559)	−0.527	(0.562)
Pre-2020 $N = 363$ ($J = 99$)	0.512	(1.125)	−10.727**	(5.099)	7.384	(4.836)	2.789***	(0.965)
2020+ $N = 2283$ ($J = 224$)	1.065***	(0.406)	0.734	(1.327)	−0.735	(1.380)	−1.091*	(0.564)

Notes: Coefficients on the Female Judge × Female Attorney interaction from Model 1 (video FEs, judge-clustered SEs) estimated on each subsample separately. J = number of judge clusters. † Fewer than 30 judge clusters; treat as fragile. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table 12: Attorney Behavior by Gender

Variable	N (F)	Mean (F)	SD (F)	N (M)	Mean (M)	SD (M)	p-value
P(neutral behavior)	4,539	0.000	0.000	10,114	0.000	0.000	—
P(agreement)	4,539	0.000	0.000	10,114	0.000	0.000	—
P(disagreement)	4,539	0.000	0.000	10,114	0.000	0.000	—
P(confidence)	4,539	0.826	0.109	10,114	0.838	0.097	0.000
P(nervousness)	4,539	0.174	0.109	10,114	0.162	0.097	0.000
Behavior entropy	4,539	0.615	0.180	10,114	0.595	0.184	0.000
Primary behavior prob	4,539	0.832	0.088	10,114	0.841	0.086	0.000
Age (years)	148	44.291	9.476	516	50.959	9.961	0.000

Notes: Unit of observation is a 10-second observation window for one attorney. Probability variables are rescaled to [0, 1]. p-values are from two-sample Welch t-tests (unequal variance). F = female attorneys; M = male attorneys.

C.3 Heterogeneity by Time Period

C.4 All Heterogeneity Results

C.5 Gender Comparison of Attorneys

D Appendix D: Variable Definitions

`prob_neutral` Probability (0–100) that the judge displays neutral engagement: attentive but noncommittal listening posture. Judge-row mean: 79.7.

`prob_agreement` Probability that the judge displays agreement or positive demeanor: nodding, leaning forward, verbal affirmation, warmth. Mean: 3.3.

`prob_disagreement` Probability that the judge displays disagreement or negative demeanor: head-shaking, furrowing, skeptical expression. Mean: 11.3.

`prob_inattentiveness` Probability that the judge displays inattentiveness: gaze averted, writing, looking at materials. Mean: 5.7.

`prob_confidence` Probability that the attorney displays confidence: upright posture, steady voice, maintained gaze. Attorney-row mean: 83.4.

`prob_nervousness` Probability that the attorney displays nervousness: voice faltering, fidgeting, averted gaze. Mean: 16.6.

`female_judge` Binary indicator (1 = female, 0 = male) for judge gender as classified by Gemini (96.4%) or administrative records (3.6%). Mean among judge rows: 0.379.

`female_attorney` Binary indicator for attorney gender, Gemini-estimated for all observations. Mean among attorney rows: 0.310.

`democrat_appointed` Binary indicator for whether the appointing president was a Democrat. Mean among judge rows with known party: 0.587.

`appellant_wins` Binary case outcome (1 = appellant wins, 0 = appellee wins). Defined for 251 of 501 cases. Mean: 0.183.

`polarity_score` For judges: prob_agreement−prob_disagreement. For attorneys: prob_confidence−prob_nervousness. Range: $[-100, 100]$.

`prob_entropy` Shannon entropy of the probability distribution across behavioral categories. Higher values indicate greater model uncertainty.

E Appendix E: Data Construction Notes

Judge name standardization. The behavioral extraction pipeline produces 387 distinct string values for judge-role actors. Stripping honorific prefixes ("Judge," "Justice," "Hon.") and normalizing capitalization reduces this to 277 unique judge names. Some residual duplication remains (e.g., "Andrew D. Hurwitz" and "Andrew Hurwitz" are treated as separate identities). This imprecision attenuates judge fixed-effect estimates and may introduce noise in judge-clustered standard errors.

Demographic coding. Race values were standardized to five categories: white, asian, hispanic, black, and other/unknown. Party labels were harmonized from "democrat"/"democratic" to "democrat." All demographic attributes for attorneys and 96.4% of judge demographics are Gemini-estimated, not sourced from administrative records.

Case outcome linkage. Case outcome data are linked to behavioral data using the video file name as the join key, producing a 100% match rate for the 352 audited videos. Binary outcomes are defined conservatively: mixed, vacated, and indeterminate dispositions are excluded.

Interruption sample coverage. The interruption detection stage covers 357 of 501 videos (71.3%). The 144 missing videos were predominantly dropped due to batch API failures and file-format incompatibilities. Diagnostic analyses show that missing videos do not differ significantly from covered videos on modality or gender composition.