Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Bilder ohne Worte: Can AI Identify Political Parties via Unspoken Visual Signatures in the 2025 German Election?

Ali Abusaleh

a.abusaleh@em-uni-frankfurt.de

Goethe University Frankfurt

Alexander Mehler

Goethe University Frankfurt

Mevlut Bagci

Goethe University Frankfurt

Research Article

Keywords: Political Communication, Multimodal Analysis, Explainable AI (XAI), Social media

Posted Date: August 31st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10681617/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Bilder ohne Worte: Can AI Identify Political Parties via Unspoken Visual Signatures in the 2025 German Election?

Ali Abusaleh[1*], Alexander Mehler[1] and Mevlut Bagci[1]

[1]Text Technology Lab (Department of Computer Science), Goethe University Frankfurt, Robert-Mayer-Straße 10, Frankfurt am Main, 60325, Hessen, Germany.

*Corresponding author(s). E-mail(s): a.abusaleh@em-uni-frankfurt.de;
Contributing authors: mehler@em.uni-frankfurt.de;
bagci@em.uni-frankfurt.de;

Abstract

Computational political communication remains largely text centric, even as party campaigning shifts to short form video and images on platforms such as X. Moreover, existing multimodal methods rely on accurate but uninterpretable embeddings, ill suited to domains such as politics where transparency and accountability are essential. We ask two questions. First, can an interpretable representation match black-box embeddings in classifying party affiliation from campaign media? Second, to what extent does party identifiability depend on subtler, coordinated aesthetic and behavioral cues when explicit branding such as logos, flags, campaign colors, and on-screen text is excluded? To address these questions, we introduce `PREF`, a Perceptual–Verbal Feature model that uses a vision language model to verbalize observable cues as a compact, human-readable feature schema mapped to party labels by an interpretable classifier. On 15,043 images and videos from the 2025 German federal election, `PREF` performs on par with both text features and 2048-dimensional embeddings while preserving interpretability through its feature model. Crucially, a controlled de-branding ablation lets us separate branding-driven from style-driven detectability, quantifying for the first time how much party identity survives the removal of explicit symbols. A parallel human-annotation study on a random subset ($\boldsymbol{n}$**=140**) yields

substantial inter-annotator agreement on the perceptual schema (Krippendorff's $\boldsymbol{\alpha}$**=0.67** overall, $\boldsymbol{\alpha}$**=0.78** for visual party identification), confirming that `PREF`'s cues are human verifiable rather than model artifacts. We frame the task as classifying institutional campaign branding, not individual ideology, and release the full pipeline for reproducible multimodal political communication research.

Keywords: Political Communication, Multimodal Analysis, Explainable AI (XAI), Social media

1 Introduction

Social media has emerged as the primary arena for political discourse, shaping public opinion beyond traditional media channels Wong et al. (2016). Earlier studies primarily focused on text content. However, the emergence of short-form video platforms such as TikTok has established multimodal content as a key vehicle for political discourse Gao et al. (2023); Sapag M. et al. (2023); Umansky and Pipal (2023). This trend also applies to social media platforms such as X (formerly Twitter), where short videos have gained increasing prominence, particularly in political contexts, with political messages originating on X often spreading to other platforms such as TikTok and Instagram Connolly (2025).

Political communication on social media is not limited to explicit verbal statements. Rather, it often occurs through implicit multimodal cues, such as facial expression, attire, setting, vocal tone, and affective display in images or videos. Consistent with this view, psychological research shows that observations of non-verbal behavior can reliably predict interpersonal outcomes Ambady and Rosenthal (1992). It also shows that brief exposure to visual cues is sufficient for humans to infer political orientation with above-chance accuracy Kosinski et al. (2024). This suggests that political affiliation is encoded not only verbally but also visually, through cues such as posture, gaze, expression, and surroundings Dan and Arendt (2021).

Recent advances in multimodal learning underscore that semantic understanding depends on cross-modal reasoning Padhi et al. (2024). This contrasts with existing embedding-level fusion methods, which exhibit limitations in three areas: The first is **semantic opacity**: like word embeddings Camacho-Collados and Pilehvar (2018), multimodal embeddings can conflate features lacking explicit grounding, e.g., encoding *cat* without distinguishing human presence. Another limitation is the **lack of interpretability**, as black-box architectures obscure their decision processes Molnar (2019). A third issue, **congruence blindness**, refers to insufficient modeling of cross-modal alignment Chang et al. (2024).

To overcome these limitations, we propose a model that verbalizes perceptual features to represent image and video content. This *perceptual-verbal feature model* (`PREF`) uses Qwen2.5-VL Bai et al. (2025) vision language models (VLMs), to extract semantic representations belonging to three categories: what is presented (vision), what is heard (in the video) (auditorys), and how do the frames of a video cohere (course of events). In a second step, we use the extracted features to represent media content and feed them into a classifier to detect the political party whose member shared the medium. Rather than attempting to infer deeply held individual ideology, our framework is designed to capture the coordinated visual and auditory branding strategies of political institutions. Our underlying classification hypothesis H1 is as follows:

H1 *A political party's distinctive communication style during an election campaign can be inferred from the visual, contextual, and auditory features of the media shared by its representatives.*

To test H1, we use 15,043[1] labeled videos and images from the 2025 German federal election, achieving 84.98% accuracy. For example, we identify conservative-party videos characterized by national flags, suits, and assertive tones. By contrast,

[1] We use 15k samples for the baseline analysis and over 17k for larger models, as the additional 2k samples exceeded smaller models' processing capacity.

progressive-party videos often show casual attire, diverse crowds, and measured tones. These perceptual-verbal features correlate with institutional campaign aesthetics. They constitute the vocabulary of a visual language through which parties express their political brand. The features that are predictive under H1 fall into two groups: explicit branding features, including logos, flags, campaign colors, and on-screen text (G1), and subtler coordinated aesthetic and behavioral cues (G2). This brings us to our second hypothesis:

H2 *H1 continues to hold when the model is restricted to G2 features. That is, after removing all G1 features (logos, flags, campaign colors, and on-screen text) from the input representation, a classifier trained on the residual G2 cues alone still predicts the posting party substantially above chance.*

The perceptual-verbal features from $G_1 \cup G_2$ help predict, but do not determine, political affiliation inferred from shared media. Rather, they interact nonlinearly to make political affiliation visually recognizable. These features constitute the vocabulary of a visual language through which parties and their representatives express political identity. `PREF` renders this visual language interpretable, while the test of H1 and H2 evaluates whether this interpretation is empirically warranted. Our contributions are threefold:

1. An **interpretable verbalized-feature framework** for party-level analysis of campaign media that matches opaque embeddings in predictive performance while remaining auditable at the feature and classifier levels (RQ1).
2. A **branding-vs-style decomposition**: a novel controlled de-branding ablation that evaluates H2 by quantifying how much party-level classifiability persists after removing logos, flags, campaign colors, and on-screen text (RQ2).
3. A **multi-model feature-elicitation** procedure that triangulates the feature schema across six language and vision language models to reduce author-specific

idiosyncrasy, together with the released PolTweetCorpus dataset and pipeline for the 2025 German federal election.

We are explicit about scope: in the H1 and H2 analyses, the predicted labels identify *the party of the representative who posted the tweet*. Thus, `PREF` models *coordinated institutional branding and production values*, not the private ideology of any individual (see section 8).

2 Related Work

2.1 Political leanings Detection

Prior work focuses on text and network analyses.

Ansari et al. (2020) perform sentiment analysis of election tweets. A title classification based on BERT, called PoLYTC, that achieves 75% accuracy on six classes is provided by AlDahoul et al. (2024). Wong et al. (2016) studies homophily-driven user graphs derived from retweets. What these approaches have in common is that they ignore the auditory, prosodic, and visual dimensions that are central to the persuasiveness of videos, especially in agitational, political communication. Notably, prior work suggests that brief observations of visual cues, such as facial expressions and bodily behavior, as well as auditory cues, can carry socially diagnostic information Ambady and Rosenthal (1992) and contribute to political perception beyond textual content alone Dan and Arendt (2021). `PREF` builds on these findings as a method for extracting such cues.

2.2 Multimodal Social Media Analysis

There is emerging research exploring cross-modal relationships: Chen et al. (2013) classify image-text congruence in tweets. Padhi et al. (2024) emphasize *semantic*

congruence for crowdfunding success. Beyond such focused applications, a new generation of benchmarks highlights the capabilities and limitations of multimodal large language models (MLLMs) in social contexts. Jin et al. (2024) introduce *MM-SOC*, a large-scale benchmark for evaluating MLLMs on multimodal social-media content spanning text, images, and videos. Guo et al. (2025) extend that line with *SNS-Bench-VL*, focusing on vision language reasoning in social-network settings. Gupta et al. (2022) present *3MASSIV*, a multilingual and multimodal dataset of short social-media videos designed to study cultural and linguistic diversity in online expression. These approaches primarily target general social-media understanding. In contrast, our framework focuses on the extraction of interpretable perceptual-verbal features from political media content to capture visual and auditory patterns that convey political identity.

2.3 Political Communication Analysis

Several studies perform multimodal analyses in political contexts.

For example, Medina Serrano et al. (2020) examine how political messages on TikTok use images, sounds, and settings to shape discourse. Marchal et al. (2021) analyze Twitter discourse during the 2019 EU parliamentary election campaign using a sample of 1,097 images and find that visuals originating from traditional political actors prevailed across the examined election-related conversations. In a longitudinal Instagram study of Brazilian presidential candidates, de Lima-Santos et al. (2025) identify consistent patterns of celebratory and personalized imagery in visual political communication during two polarized Brazilian elections. Köksal and Erisen (2023) apply multimodal discourse analysis to Turkish campaign posters, demonstrating how representational, interactional, and compositional modes produce persuasive political meaning. We extend this research by developing a generalizable framework for analyzing political videos that is independent of platforms or third-party APIs. We also

introduce a more comprehensive feature extraction method that captures a broader range of perceptual political signs.

2.4 Multimodal Representation Learning

Prior work suggests that embeddings from pretrained VLMs, such as CLIP Radford et al. (2021), may perform poorly when applied directly to highly imbalanced datasets. Wang et al. (2024) analyze CLIP's zero-shot performance on long-tailed benchmarks such as ImageNet-LT and iNaturalist-2018. They report very low accuracy (e.g., 5% on iNaturalist-18) and demonstrate that naive prompting fails on minority classes unless specialized, imbalance-aware decoders and balanced loss techniques are employed. Parashar et al. (2024) confirm this finding by systematically evaluating VLMs across long-tail visual categories. They find that, without fine-tuning, VLMs disproportionately favor head classes. They propose REtrieval-Augmented Learning (REAL), a retrieval-augmented decoder that only modestly improves tail-class accuracy. These embedding-based approaches perform poorly in domains with strong class imbalance, as demonstrated by Wang et al. (2024) and Parashar et al. (2024), who report poor accuracy on minority classes, even with zero-shot CLIP, and only slight improvements with complex balancing techniques. To overcome these limitations, we begin by extracting interpretable visual and transcript-derived cues, such as the presence of cultural symbols, the formality of clothing, and the tone of voice, by verbalizing them with Qwen2.5-VL. These perceptual-verbal features are intended to make minority-class patterns more explicit and to reduce the risk of embedding-space bias toward majority classes.

2.5 Nonverbal Political Signaling

Psychological research confirms that people can discern political ideologies from brief glimpses of behavior Ambady and Rosenthal (1992); Dan and Arendt (2021); Kosinski et al. (2024). Similarly, reviewing research on nonverbal political communication, Dumitrescu (2016) shows how nonverbal cues, including facial expressions, body language, and vocal tones, can shape perceptions of political candidates and their messages. We operationalize, computationally and at scale, the capacity to predict the party affiliation of the posting representative from nonverbal cues. This behavioral perspective is complemented by neuroimaging research showing partisanship-dependent differences in early sensory and motor responses to political content Katabi et al. (2023). People with similar political views exhibit synchronized brain activity, including in primary visual and auditory cortices, when exposed to politically aligned stimuli. These findings support our hypothesis that nonverbal political signals are associated with shared patterns of perceptual processing, suggesting a perceptual basis linked to party affiliation.

2.6 Interpretability in Machine Learning

In sensitive domains such as political analysis, interpretability is crucial for making ML model predictions transparent, contestable, and methodologically accountable. Molnar (2019) provides a comprehensive guide to interpretable ML, emphasizing the need for transparent models that can explain their decisions. Ribeiro et al. (2016) introduce techniques for explaining the predictions of classifiers. We build on these approaches to provide feature-level explanations of model predictions.

2.7 Bias and Fairness in AI

Bias and fairness are central concerns in AI applications, where biased models can distort interpretation and undermine accountability, especially in political analysis.

Table 1: Tweets from PolTweetCorpus by political party, along with mapping of Generic Group Labels to Actual Group Names

Generic Label [2]	Party	#Tweets
Group A	AfD	1,878
Group B	BSW	419
Group F	Bündnis 90/Die Grünen	3,662
Group C	CDU/CSU	1,813
Group D	Die Linke	1,376
Group E	FDP	3,567
Group G	SPD	2,500
Group H	Others	24,331
	Sum	39,546

Buolamwini and Gebru (2018) highlight disparities in intersectional accuracy in commercial gender classification systems, underscoring the importance of addressing bias in AI models. Olteanu et al. (2019) discuss methodological pitfalls and ethical boundaries in social data analysis, providing guidelines for mitigating bias in datasets. We chose a Random Forest classifier over more complex deep-learning (DL) architectures because it offers greater interpretability: feature importance and decision paths can be inspected directly. Although simpler than DL models, this choice prioritizes transparency, which is crucial for evaluating bias and fairness. Random Forests allow us to assess feature importance and identify which inputs carry the most predictive weight, facilitating the detection of potential bias sources that are often obscured in more opaque neural systems.

3 Data Collection and Preprocessing

We use the PolTweetCorpus Bagci et al. (2025) dataset. Table 1 shows the number of tweets per party in this dataset. PolTweetCorpus uses two complementary collection strategies: a *user-based* approach and a *keyword-based* approach. Together,

[2] In the main text, political groups are referred to using generic labels (Group A, Group B, etc.).

these strategies capture direct political communication from verified party members as well as a broader range of media related to political discussions. The user-based approach draws on tweets from 69 German politicians representing the seven major parties in the 2025 German federal election; see Table 1. The selected politicians held prominent positions within their respective parties and were therefore likely to contribute visibly to campaign communication. The collection period ran from November 6, 2024, when the German traffic light coalition collapsed,[3] to February 23, 2025, the date of the federal election. Table 1 shows that Group F accounts for the largest number of tweets, whereas Group B accounts for the fewest. The *Others* category is larger because it includes tweets from the keyword-based collection method, which also captured tweets that could not be associated with a specific party representative. Together, the user- and keyword-based collection methods yielded PolTweetCorpus, comprising 39,546 tweets. Of these, 17,137 contain media[4] and were published by a variety of users and organizations, mainly in English and German; we did not apply a language filter. Overall, we obtained 15,043 labeled media items, including videos and images; 12,034 were used for training and 3,009 for testing. Audio tracks were extracted with FFmpeg FFmpeg Developers (2025) to make the audio modality available for explicit downstream processing. Using VLMs (e.g., Qwen2.5-VL), we extracted interpretable visual features from each image and video frame and complemented them with transcript-derived cues from the corresponding audio track. The extracted features include quantitative indicators, such as the number of detected people and objects, as well as categorical attributes, including scene type (*indoor/outdoor*), clothing style (*formal/casual*), and speech-derived emotional tone (e.g., happiness, anger). The full list of features is provided in Appendix B. These

[3] A *traffic light coalition* (German: *Ampelkoalition*) is a German governing coalition of the SPD (red), FDP (yellow), and Greens (green), named after the colors of a traffic light.

[4] As noted above, our baseline uses 15k samples, whereas larger models were evaluated on more than 17k samples.

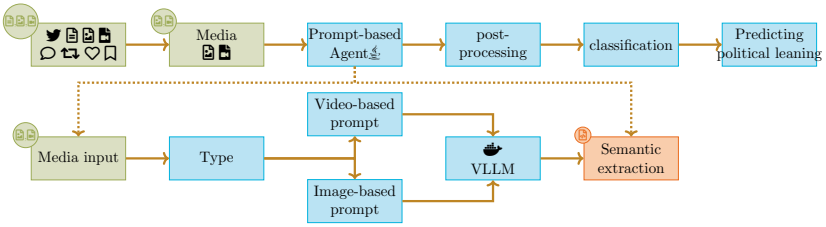


Fig. 1: PREF pipeline Overview. Media from X are encoded by a prompt-based VLM into structured representations, which an interpretable classifier uses to predict the source party's campaign aesthetics.

features are interpretable insofar as they verbalize perceptual impressions or observable states of affairs in a human-readable form. Examples include the number of visible logos associated with a given party and the tone of speech, such as "serious". We therefore refer to them as *perceptual-verbal features*.

Each video is labeled with the party affiliation of the representative who shared it. Videos shared by independent politicians were labeled with the respective politician's name. This ensures that our labels reflect (1) party affiliation for party representatives and (2) the names of individual independent politicians, while (3) the label *Others* covers all other cases.

4 Architecture

We implement PREF as a modular, multimodal processing pipeline that integrates visual, auditory, and textual cues to predict the label associated with the posting account: party affiliation, the individual name of an independent politician, or *Others*. PREF operates in three steps: (1) *Multimodal Feature Extraction*, which converts raw media into structured representations; (2) *Feature Elicitation and Selection*, which combines literature-driven and LLM-consensus methods to define the feature taxonomy; and (3) *Interpretable Classification*, which maps perceptual-verbal features to posting-account labels (Figure 1).

4.1 Multimodal Feature Extraction

The process begins with media input collected from X. Depending on the modality, a prompt-based agent generates tailored representations, which are then processed within a VLM environment to extract high-level semantic information. We use Qwen2.5-VL Bai et al. (2025) to analyze the media and generate structured, semantic descriptions. The raw model output adheres to a strict JSON schema full prompt is available Listing 3 to ensure downstream consistency.

4.2 Hybrid Feature Selection and Elicitation

To establish a theoretically grounded yet data-driven taxonomy, we employ a hybrid feature elicitation strategy. First, we identify core variables from established political communication and multimodal literature Garg and Ramsay (2011); Araujo et al. (2020). This covers **Vision** (demographics, symbols, attire), **Acoustics** (emotional tone, prosody Lausen and Hammerschmidt (2020)), and **Temporal Consistency**.

Multi-Model Consensus for Bias Mitigation

To further mitigate researcher bias and ensure feature diversity, we supplement the literature-driven approach with a Multi-Model Consensus pipeline. We leveraged an ensemble of six state-of-the-art LLMs (`GPT-5`, `Llama3.2`, `DeepSeek-R1`, `Mistral`, `Phi-4`, and `GPT-OSS`) to independently propose observable audiovisual features relevant to political discourse. These lists were aligned via canonical mapping and aggregated through majority voting (Figure 2). Only features suggested by at least two independent models were retained, ensuring the resulting schema is objective and structurally robust.

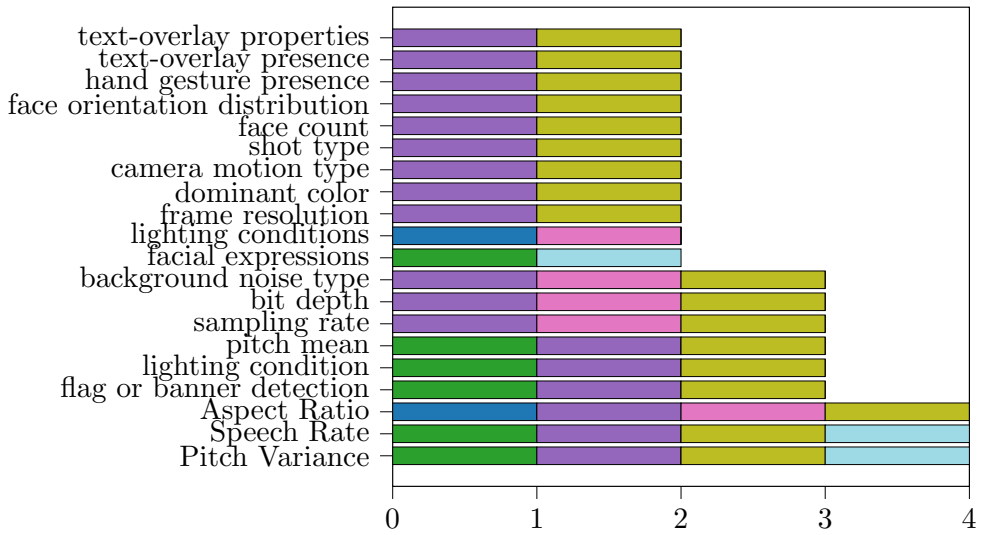


Fig. 2: Stacked bar chart showing model-wise contribution to the top aligned features. Each bar reflects one canonical feature, segmented by the LLMs that proposed it.

4.3 Semantic Abstraction and Feature Grouping

Extracted features are mapped to higher-level concepts to enhance interpretability. For example, a national flag symbolizes national identity, while formal attire conveys professionalism. Following the framework in Appendix C, we organize these into three functional groups used for our ablation studies: **Visual** (composition/objects), **Auditory** (voice/prosody), and **Contextual** (settings/events).

4.4 Interpretable Classification

We prioritize transparency by employing decision-tree-based classifiers. While deep latent models often achieve high predictive performance, their reasoning is opaque. In contrast, training on perceptual-verbal features allows us to trace predictions back to specific human-understandable cues, forming the basis for our qualitative analysis of nonverbal political communication.

5 Experiments and Results

In this section, we evaluate the effectiveness of PREF-based political classification. By unifying our experimental setup and findings, we provide a comprehensive assessment of our approach's discriminative power, robustness, and interpretability compared to traditional text- and embedding-based baselines.

5.1 Experimental Setup

Our evaluation is conducted on a held-out test set comprising 3,009 samples from our dataset (an 80/20 train-test split of the 15,043 total labeled items). We leverage the Qwen2.5-VL VLM Bai et al. (2025) to extract multimodal features from audiovisual content. To rigorously assess the discriminative power of these signals, we evaluate several semantic-based classifiers (Decision Trees, Random Forests, Logistic Regression, k-Nearest Neighbors, and Naive Bayes) and report standard metrics including accuracy, precision, recall, and weighted one-vs-rest ROC-AUC. To benchmark PREF, we compare it against three embedding-based baseline methods utilizing 2,048-dimensional Qwen2.5-VL embeddings:

- **k-NN with Cosine Similarity:** We We normalize all embeddings $\mathbf{x} \in \mathbb{R}^{2048}$ to unit length and compute cosine similarity

$$\cos(\mathbf{x}, \mathbf{x}') = \frac{\mathbf{x}^\top \mathbf{x}'}{\|\mathbf{x}\| \, \|\mathbf{x}'\|} \, .$$

  Given a test point $\mathbf{x}$, let $\mathcal{N}_k(\mathbf{x})$ be the set of its k nearest neighbors under the cosine distance $d(\mathbf{x}, \mathbf{x}') = 1 - \cos(\mathbf{x}, \mathbf{x}')$. We predict

$$\hat{y} \;=\; \arg\max_{c} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{N}_k(\mathbf{x})} \mathbb{I}\{y_i = c\} \, .$$

- **Linear Classifier with Cosine Prototype Matching:** Next, we train a one-vs-rest logistic regression on raw embeddings to learn weight vectors $\{\mathbf{w}_c\}$ for each class c. At inference, we again normalize

$$\tilde{\mathbf{x}} = \frac{\mathbf{x}}{\|\mathbf{x}\|}, \quad \tilde{\mathbf{w}}_c = \frac{\mathbf{w}_c}{\|\mathbf{w}_c\|},$$

and compute cosine similarity to each prototype Snell et al. (2017): $s_c = \cos(\tilde{\mathbf{x}}, \tilde{\mathbf{w}}_c)$. The predicted label is $\hat{y} = \arg\max_c \, s_c$ ..

- **Metric Learning via Triplet Loss:** We learn a projection $f : \mathbb{R}^{2048} \to \mathbb{R}^{128}$ (parameterized by a small neural network) using triplets $\{(a, p, n)\}$ of anchor, positive, and negative embeddings. The loss is

$$\mathcal{L} = \sum_{\text{triplets}} \big[\, \|f(a) - f(p)\|^2 \; - \; \|f(a) - f(n)\|^2 \; + \; m \big]_+$$

where $[z]_+ = \max(0, z)$ and the margin m is set to 1 Schroff et al. (2015). After training, we compute each class centroid

$$\boldsymbol{\mu}_c \; = \; \frac{1}{N_c} \sum_{i : y_i = c} f(\mathbf{x}_i)$$

in this 128-D space and classify by nearest-centroid under cosine similarity:

$$\hat{y} \; = \; \arg\max_c \; \cos\big(f(\mathbf{x}), \, \boldsymbol{\mu}_c\big) \, .$$

5.2 Bias Mitigation via LLM-Ensemble Feature Elicitation

To complement literature-informed features and reduce researcher bias, we introduce data-driven feature elicitation. We prompted a multi-model consensus ensemble of six

state-of-the-art LLMs (`GPT-5` OpenAI (2025a), `Llama3.2:latest` Meta AI (2024), `GPT-OSS:20b` OpenAI (2025b), `Mistral:latest` Jiang et al. (2023), `Phi-4:latest` Abdin et al. (2024), and `DeepSeek-R1`) with a meta-instruction to independently propose observable, non-interpretive audiovisual features suitable for political content analysis.

We aligned these feature lists using canonical mapping and aggregated them through majority voting, retaining only features that achieved cross-model consensus. When training `PREF` exclusively on this automated, LLM-derived schema, the model achieved a robust accuracy of 81.8% and an ROC-AUC of 0.765 (DT model). While slightly lower than our primary expert-informed schema (evaluated below), this proves that automated LLM ensembles can surface highly discriminative, interpretable visual and audio features. This establishes a strong, bias-mitigated baseline for political media analysis without relying on manual human descriptors.

5.3 Classification Results for `PREF`

Using our primary expert-informed perceptual-verbal features, we trained a variety of interpretable classifiers. Among all models tested, the Random Forest (RF) classifier achieved the highest overall performance, with an accuracy of 84.98%, precision of 82.12%, recall of 84.98%, and an AUC of 0.9219 (see Table 2).

Table 2: Test-set classification results for `PREF` using media-derived features from Qwen2.5 (3B).

Model	Accuracy	Precision	Recall	AUC
Decision Tree	0.8159	0.7986	0.8159	0.7630
Random Forest	**0.8498**	**0.8212**	**0.8498**	**0.9219**
Logistic Regression	0.7906	0.7906	0.7906	0.8791
K-Nearest Neighbors	0.7833	0.7516	0.7833	0.8294
Naive Bayes	0.6530	0.7882	0.6530	0.7614

These results suggest that perceptual-verbal features convey meaningful information, enabling discrimination between political groups without the need for explicit verbal content. Decision trees also perform strongly, achieving an accuracy of 81.59%, which highlights the significance of `PREF`'s feature space. Our findings confirm that party affiliation can be classified from visual and auditory cues, indicating that political parties differentiate themselves through distinctive, coordinated media presence.

5.4 Testing H2: Controlling for Label Leakage through De-Branded Ablation

This ablation is the direct empirical test of H2. If party classifiability were entirely a function of explicit branding (G_1: logos, flags, campaign colors, on-screen text), removing G_1 from the input should collapse performance to chance. If, on the other hand, parties differ in their coordinated aesthetic and behavioral style (G_2: scene setting, attire formality, speaker demographics, motion, tone), performance should remain substantially above chance even in the absence of G_1. To adjudicate between these alternatives, we systematically removed all explicit symbolic features from the dataset, specifically, those tracking `dominant_color_family`, flags, logos, and text overlays. The total number of explicit branding features removed is 2,772.

After removing these highly predictive explicit branding cues, we retrained the Random Forest classifier using only the remaining subtle perceptual features (e.g., scene setting, attire formality, speaker demographics, motion intensity, and auditory tone). **The de-branded model maintained robust performance, achieving an accuracy of 84.11% and an ROC-AUC of 0.8977.** While a performance drop from the fully featured model (84.98%, and 0.921, respectively) is expected when removing logos and text, the de-branded results remain highly discriminative and significantly above chance. This validates our core hypothesis: while explicit symbols are

utilized in campaign media, **political groups also coordinate distinct, measurable aesthetic and behavioral styles** that successfully differentiate them even in the complete absence of explicit branding. The de-branded model achieves 84.11% accuracy, only −0.87 pp below the full-feature model and well above the majority-class baseline, supporting H2: party-level classifiability is not reducible to branding shortcuts, but is carried in substantial part by G_2 style features.

5.5 Media- versus Text-based Models

To test whether perceptual-verbal features convey information that is comparable to explicit linguistic information, we conduct a classification using textual features extracted from the captions and descriptions of the respective images and videos.

Table 3: Comparing media- and text-based features (random forest)

Metric	Media-Based	Text-Based
Accuracy	84.08%	83.61%
Precision	0.8059	0.7846
Recall (Sensitivity)	0.8408	0.8361
Specificity	0.8408	0.8361
F1 Score	0.8230	0.8096
ROC-AUC	**0.9058**	0.8775

As shown in Table 3, **media-derived features perform slightly better than text-based ones across all major classification metrics**, notably achieving a higher ROC-AUC (+2.8 points). The results indicate that visual representations of different parties are clearly separable in feature space. Even without textual context, online visual media convey semantically rich and socially distinctive cues, strongly reflecting institutional branding. Public opinion is shaped not only by textual narratives but also by visual campaign strategies. Elements such as visual style, body language, and compositional framing can implicitly convey party affiliation.

5.6 Comparison with Embedding-Based Baselines

To further assess PREF, we compared it against the raw embedding-based baselines outlined in our setup.

Table 4: Comparison of embedding-based methods on the test set. Bold values indicate the best score for each metric.

Model	Accuracy	AUC	Recall	Weighted F1
RandomForest	0.8379	0.8267	0.8379	0.7799
NeuralNetwork	**0.8435**	0.8575	**0.8435**	0.8144
KNN	0.8156	0.7520	0.8156	0.7685
LogisticRegression	0.8144	0.8041	0.8144	0.7945
Triplet Network	0.7962	0.6845	0.7962	0.7712
Semantic-based (Ours)	0.8407	**0.9119**	0.8407	**0.8407**

While traditional classifiers on raw embeddings (e.g., Neural Networks) achieve high accuracy, they suffer from poor sensitivity, often failing to detect minority classes altogether in imbalanced settings. In contrast, PREF shows balanced, high performance across all metrics. It achieves the highest AUC (0.9119) and a highly balanced weighted F1 score (0.8407). This suggests that baseline models primarily rely on surface-level similarity in latent space, whereas PREF captures deeper, more interpretable distinctions relevant to political classification.

5.7 Distinctiveness and Group-Specific Media Profiles

To examine stylistic differentiation, we quantified pairwise similarities between group-specific feature profiles using three complementary metrics: *Cosine similarity*, *Jaccard similarity*, and *Jensen-Shannon divergence* (Figure 3).

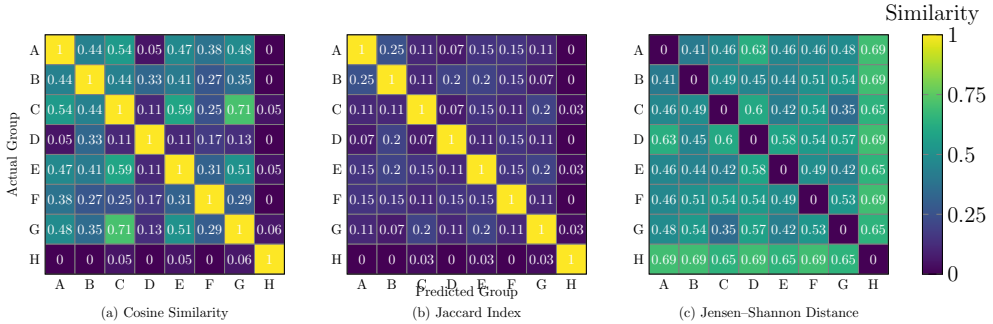


Fig. 3: Similarity between parties based on top-15 media features. From left to right: (a) Cosine Similarity, (b) Jaccard Index, and (c) Jensen–Shannon Distance.

Discussion of Similarity Metrics

Our findings suggest that groups actively construct distinct visual identities in their shared media. **Groups C, E, and G** form the most cohesive stylistic cluster (highest Cosine similarity, lowest JSD), indicating shared patterns in visual and contextual representation. **Groups A and B** show partial alignment (e.g., Jaccard overlap of 0.25), indicating partial convergence in visual attributes like attire or activity type. In contrast, **Groups D and F** occupy peripheral, distinct positions, while **Group H (Others)** emerges as a clear outlier. The combination of low overall feature overlap and selective convergence suggests that our classification performance depends on robust, class-specific cues rather than generic traits.

5.8 Feature Contributions and Media Signatures by Political Group

To interpret which nonverbal features drive model decisions, we visualized per-class feature contributions (Figure C1). Each political group follows a characteristic audio-visual pattern, reflecting systematic choices in composition, editing, and rhetorical framing:

- **Group A:** Formal attire, clear solo speeches, and low-motion podium addresses in outdoor public settings. A dominant black-blue color palette and minimal crowd presence contribute to an organized, authoritative aesthetic.
- **Group B:** Media content is characterized by indoor office scenes and static text-based images, emphasizing press-oriented communication. High speaker counts and visual aids (charts, microphones) suggest a professional media setup.
- **Group C:** Blends formal and casual dress, frequently in office or studio settings. Mixed-gender speaker configurations and sparsity of crowds point to speech-centered, institutional content.
- **Group D:** Favors indoor, static imagery and news-style text displays, relying on graphics and narration to convey informational messages.
- **Group E:** Formal and serious, often featuring male speakers and visualizations like charts in controlled office settings, aligning with a technocratic visual identity.
- **Group F:** Low-motion, text-rich visuals in office environments. Warm-toned palettes and symbolic human interactions (e.g., object handover) convey a grounded, empathetic communication approach.
- **Group G:** Formal and casual dress, low-motion speeches, and consistent indoor studio environments. A serious tone and group presence signal a blend of tradition and inclusion.

5.9 Semantic Robustness of Media-Based Features

To test whether the extracted media features capture systematic political differences or merely reflect overfitting, we conducted a robustness test with controlled label noise (randomly flipping class labels during training, evaluating on an uncorrupted test set).

As shown in Table 5, performance metrics systematically decline as label noise increases. The observed degradation suggests that `PREF` learns generalizable patterns from audiovisual data rather than exploiting spurious label-feature correlations.

Table 5: Media-based model performance under varying label noise.

Noise (%)	Accuracy	Precision	Recall	F1	ROC-AUC
0	0.8418	0.8096	0.8418	0.8254	0.9037
10	0.8229	0.7899	0.8229	0.8061	0.8799
30	0.7686	0.7618	0.7686	0.7652	0.8353
40	0.7377	0.7527	0.7377	0.7451	0.8067

This coherence–found in group-specific scene composition, framing, and appearance–implies that audiovisual content shared by political actors conveys structured political campaign signals, even without accompanying text.

5.10 Performance Across Model Architectures and Sizes

Finally, we evaluate the impact of the underlying VLM's architectural design and scale on feature extraction.

Table 6: Performance comparison across Qwen architectures and sizes (mean ± std computed over 10 runs).

Model Variant	**Accuracy**	**Precision**	**Recall**	**Sensitivity**	**Specificity**	**ROC-AUC**
3B-Instruct (Dense)	0.838 ± 0.003	0.812 ± 0.004	0.838 ± 0.003	0.838 ± 0.003	0.838 ± 0.003	0.910 ± 0.004
32B-Instruct (Dense)	0.791 ± 0.006	0.817 ± 0.004	0.791 ± 0.006	0.791 ± 0.006	0.791 ± 0.006	0.863 ± 0.004
72B-Instruct (Dense)	0.858 ± 0.002	0.830 ± 0.004	0.858 ± 0.002	0.858 ± 0.002	0.858 ± 0.002	0.905 ± 0.004
30B-A3B (MoE)	0.780 ± N/A	0.782 ± N/A	0.780 ± N/A	N/A	N/A	0.839 ± N/A

We examined the effect of LLM size on the implementation of `PREF`. Table 6 indicates a saturation effect in performance at the 3B parameter level. There is no substantial difference in accuracy between the 3B and 32B dense models ($\Delta = 0.0 \pm 0.005$),

suggesting diminishing returns with increasing model size. With the lowest computational cost, the 3B-Instruct model achieves an exceptional AUC (0.910 ± 0.004), demonstrating optimal efficiency suitable for edge applications. These results support our semantic bottleneck hypothesis: smaller models can extract task-relevant knowledge efficiently from limited data.

We additionally evaluated the *Qwen3-Omni-30B-A3B-Instruct* variant to test a Thinker-Talker Mixture-of-Experts (MoE) architecture Shazeer et al. (2017). The MoE model achieved a maximum accuracy of 78.08%, noticeably lower than the dense models. Despite their architectural sophistication and efficiency in scaling, MoE models may introduce instability during fine-grained visual-linguistic alignment. In contrast, the dense Qwen2.5-VL architectures appear better suited for the precise grounding and contextual reasoning required by `PREF`.

6 Human Validation and Inter-Annotator Agreement

We conducted a manual annotation study to answer **two questions** that neither classification accuracy nor an ablation can settle on their own. **First**, are the perceptual-verbal features on which `PREF` operates actually observable by human annotators, or are they artifacts of the VLM's internal representations? **Second**, when Qwen2.5-VL labels a given item, does its judgment align with what a human reader would extract from the same item? To address these questions, we randomly sampled n=140 media items from PolTweetCorpus and had them independently annotated across twelve semantic dimensions of the `PREF` schema, ranging from objective structural counts (people count, voice presence) through mid-level perceptual categories (scene context, dress style) to more interpretive dimensions (emotional tone, dominant color family). Each item received an average of 2.04 independent human annotations.

We report two families of metrics: (i) human–human agreement, which measures schema reliability, and (ii) LLM–human agreement, which measures how faithfully

Table 7: Human-human and LLM-human agreement on the PREF schema (n=140; avg. 2.04 human annotators per item).

	Human–Human		LLM–Human	
Feature	%	α	%	α
High reliability ($\alpha_{H-H} > 0.75$)				
Is political	96.5	0.918	95.3	0.790
Speaker gender	93.0	0.888	68.2	0.533
Voice presence	94.5	0.884	78.3	0.562
Political party (visual)	85.0	0.776	77.3	0.734
Moderate reliability ($0.55 \leq \alpha_{H-H} \leq 0.75$)				
Dress style	79.0	0.676	85.2	0.774
Scene context	86.5	0.675	54.0	0.218
People count	81.0	0.619	79.4	0.587
Motion intensity	86.0	0.579	74.4	0.484
Interpretive dimensions ($\alpha_{H-H} < 0.55$)				
Activity category	73.0	0.529	84.1	0.507
Is speaker a politician	66.0	0.529	76.5	0.572
Dominant color family	57.0	0.509	33.9	0.214
Emotional tone	63.0	0.471	34.7	0.027
Macro average	**80.0**	**0.671**	**70.1**	**0.500**

Legend: Percent agreement (%) and Krippendorff's α. Rows are sorted by human–human α.

Qwen2.5-VL reproduces human perception. For each we compute both raw percent agreement and Krippendorff's α, the latter correcting for chance agreement given the class-frequency distribution of each dimension. For dimensions that admit multiple simultaneous labels (activity category, scene context, dominant color family), we additionally report multi-label precision, recall, and F1.

6.1 Reliability of the Schema (H–H Agreement)

The macro human–human agreement of α=0.671 falls in the *substantial* range under the Landis and Koch heuristic Landis and Koch (1977), indicating that the PREF schema is a well-formed instrument rather than a set of ad-hoc labels. The dimensions

Table 8: Multi-label agreement for the three dimensions that admit overlapping labels.

	Human–Human				LLM–Human			
Feature	J	P	R	F1	J	P	R	F1
Activity category	0.80	0.81	0.87	0.83	0.88	0.92	0.88	0.89
Scene context	0.90	0.92	0.91	0.91	0.56	0.58	0.56	0.57
Dominant color family	0.65	0.69	0.69	0.68	0.48	0.62	0.50	0.53

Legend: (J: Jaccard, P: precision, R: recall, F1)

that anchor our analysis are the most reliable ones: annotators nearly always agree on whether an item is political (α=0.92), on the presence of a human voice (α=0.88), and on the speaker's apparent gender (α=0.89). Crucially for our central claim, annotators can identify the source party from the media alone with α=0.78, which independently confirms H1 at the human level: the signal that `PREF` picks up is present in the data itself, not merely in the model's internal geometry. The reliability profile also delineates the schema's soft edges. Emotional tone (α=0.47) and dominant color family (α=0.51) sit at the moderate-to-low end for humans; these are inherently interpretive dimensions, and their lower agreement is a property of the underlying construct rather than of the annotation protocol. We flag this so that downstream analyses treat these features with proportional caution.

6.2 Fidelity of the VLM (LLM–Human Agreement)

Qwen2.5-VL agrees with humans at a macro α=0.500 (*moderate*), which is a meaningful outcome for two reasons. First, the LLM-human ceiling is bounded above by human-human agreement itself: the model cannot reproduce a signal that annotators do not reproducibly recover. Second, and more informatively, the LLM's agreement pattern mirrors the humans': it is highest exactly where humans agree with each other and lowest where humans disagree. On the pivotal item of visual party identification,

the model reaches α=0.73, close to the human ceiling of α=0.78, which shows that its party-relevant readings are anchored in the same cues humans use.

Two departures from the human pattern deserve explicit notice. On dress style and activity category, LLM-human agreement (α=0.77 and 0.51) exceeds human-human agreement (α=0.68 and 0.53) because the model produces a systematic, low-variance reading close to the human majority, effectively a stable "median annotator". Conversely, on emotional tone the model degenerates (α=0.027, chance-level) and on dominant color family it agrees only weakly (α=0.21). These are the two dimensions where the LLM's output should not be taken at face value in downstream interpretation.

6.3 What the Study Adds to PREF

The validation study establishes three things that the classification experiments alone cannot. (1) PREF's features are empirically observable: the schema clears a substantial-agreement threshold at the human level. (2) The VLM's extractions are human-anchored on the features that matter for our claims, including party identification, in the same rank order humans converge on. (3) The features that PREF handles well and the features it handles poorly are the same features that humans handle well and poorly, which lets us calibrate trust in PREF's outputs on a per-feature basis rather than in aggregate. Together, these results turn interpretability from a claim about the model into a claim about the underlying construct.

7 Discussion

Our results demonstrate that the PREF framework effectively bridges the *semantic gap* between raw multimodal media and high-level political constructs. By achieving a peak accuracy of 84.98% and an AUC of 0.921, this study provides empirical evidence

that political identity is encoded in a structured, *perceptual-verbal* signature that can be decoded without relying on textual metadata.

Beyond Black-Box Embeddings

A key finding is that our interpretable feature model performs comparably to, and in some metrics exceeds, opaque embedding-based approaches. While deep latent representations offer computational scalability, they lack the transparency required for sociopolitical accountability. `PREF` proves that a manageable set of human-understandable features-attire, prosody, and environmental setting-retains enough discriminative power to classify institutional affiliation. This suggests that political *branding* is not merely a latent byproduct of media production but a systematic, perceivable strategy.

Interpretability grounded in human perception

The human-annotation study reframes what "interpretable" means for `PREF`. Interpretability here is not only that a decision tree can be read off feature-by-feature; it is that those features denote perceptual categories humans reliably identify, on the very items the model classified. The convergence of three signals: (i) an above-chance de-branded classifier (H2), (ii) humans identifying the source party from media alone at α=0.78, and (iii) LLM extractions that agree with humans on precisely the party-relevant dimensions, suggests that party-level classifiability in campaign media is a property of the communication itself, not an artifact of a component of our pipeline.

The Bias-Mitigation Advantage

The integration of a Multi-Model Consensus pipeline addresses a long-standing limitation in computational social science: researcher subjectivity. By eliciting our taxonomy from an ensemble of six LLMs, we ensured that the features used to *read* the 2025 German federal election were not biased by a single expert's perspective. The robustness

of our model under label noise further confirms that PREF captures stable campaign aesthetics rather than overfitting to transient media trends.

The Multimodal Shift in Political Discourse

Our study documents a fundamental shift in political communication: the transition from text-heavy rhetoric to *de-verbalized* visual and auditory cues. Historically, party affiliation was signaled through policy manifestos; today, it is signaled through the *Composition* of the media-the formality of an outdoor rally, the urgency of a speaker's pitch, or the specific color palettes of a video frame. PREF is designed to keep pace with this development, offering a method to process the massive volume of short-form video content on platforms like X, in a traceable manner.

8 Limitations, Implications and Future Work

While our findings demonstrate the efficacy of PREF in classifying political media, several limitations must be acknowledged. First, our dataset consists of media posted by official party members during a highly structured campaign (Germany, 2025). Consequently, the model classifies coordinated campaign aesthetics and resource-based production values (e.g., professional studios vs. grassroots content) rather than inferring individual-level political ideology. Generalizing these specific feature profiles beyond this election cycle or to informal user-generated content requires further cross-contextual validation. Second, the inclusion of explicit party markers (e.g., specific flags, logos, and dominant campaign colors) in our feature schema introduces the potential for label leakage. In official campaign contexts, these elements directly encode party affiliation, which may artificially inflate classification accuracy by providing trivial shortcuts for the model. Future work should incorporate rigorous de-branding ablation studies to isolate the predictive power of subtle nonverbal cues (like posture or tone) from explicit campaign branding. Third, our "Others" category comprises a heterogeneous mix of non-party actors and cross-national material. While it serves as a

broad catch-all baseline in this study, its conceptual incoherence limits granular analysis of non-aligned political discourse. Finally, while PREF provides greater transparency than opaque embeddings, its interpretability relies on the outputs of Qwen2.5-VL. We validate this dependence in §6 through a manual inter-annotator study. The stability of the extracted features varies systematically by dimension: high-reliability features (political nature, voice, speaker gender, visual party identification) support downstream inference with confidence, while interpretive dimensions (emotional tone, dominant color family) should be treated as suggestive rather than definitive. Analyses that hinge on the latter two categories in particular should be interpreted with proportional caution.

Implications and Future Work

Beyond the 2025 German election, the methodology presented here serves as a blueprint for *Digital Semiotic Intelligence.* The ability to derive language-like representations from visual signals allows researchers in sociology, marketing, and communication to quantify institutional identities at scale. Future work will explore cross-national validation to determine if these *unspoken signatures* are culturally specific or if globalized social media aesthetics are creating a universal visual language for political power.

Ethical and Societal Implications

The capability to automatically classify political affiliation or detect partisan media styles from audiovisual cues carries significant ethical risks. While our methodology analyzes public, official campaign communication, the underlying technology could theoretically be misapplied to informal or private user data. Such applications raise severe concerns regarding political profiling, automated surveillance, and the strategic manipulation of political messaging. Deploying multimodal inference tools in politically sensitive domains necessitates strict safeguards. It is crucial to enforce regulatory

boundaries that prevent the unauthorized scraping and ideological profiling of private citizens' visual data. Furthermore, as models become more adept at identifying subtle political cues, there is an increased risk of malicious actors reverse-engineering these insights to generate hyper-targeted, deceptive political propaganda. We urge the research community to prioritize consent, transparency, and the development of robust auditing mechanisms when applying VLMs to political discourse. To ensure conceptual clarity and avoid conflating individual psychological traits with institutional communication, we adhere to the following terminology throughout this study: **Party Affiliation:** The objective ground-truth label of the media, defined by the verified political party of the politician who shared the content. **Campaign Aesthetics / Visual Branding:** The coordinated use of visual, auditory, and contextual cues (e.g., specific settings, attire, color schemes, and symbols) deployed by a party to communicate its political message. **Stylistic Convergence:** The tendency of media shared by members of the same political party to exhibit similar multimodal feature profiles, resulting from shared campaign strategies and resources. This study does not attempt to infer the internal *political ideology* or *political leaning* of individual voters. Rather, it focuses on classifying the institutional *party affiliation* of the media source based on observable campaign aesthetics.

List of abbreviations

PREF Perceptual–Verbal Feature model

VLM Vision–Language Model

LLM Large Language Model

RF Random Forest

DT Decision Tree

MoE Mixture of Experts

ROC-AUC Area Under the Receiver Operating Characteristic Curve

IAA Inter-Annotator Agreement

ASR Automatic Speech Recognition

G1/G2 Explicit-branding / subtle-coordinated feature groups

Declarations

Availability of data and materials

The Twitter/X corpus analysed in this study is a subset of the PolTweetCorpus resource introduced by Bagci et al. (2025), which is publicly available at https://github.com/texttechnologylab/TTLabTweetCorpus For this study we use the media-bearing subset covering the 2025 German federal election period (6 November 2024 to 23 February 2025), comprising 15,043 labelled images and videos as described in §3. The exact item identifiers, train/test split, and extracted perceptual–verbal feature tables required to reproduce all results are released alongside the code (see below). In accordance with X's Developer Agreement, raw tweet content is not redistributed; we release tweet IDs and derived features, from which the original media can be rehydrated via the X API. The corpus is distributed under CC BY-NC-SA 4.0.

Code availability

All code and pipelines used for data collection, preprocessing, feature extraction with Qwen2.5-VL, classifier training, and the G1/G2 de-branding ablation will be made available and released upon acceptance using a Zenodo release.

Competing interests

The authors declare that they have no competing interests.

Funding

This research was partially funded by the German Research Foundation (DFG) within the Infrastructure Priority Programme New Data Spaces for the Social Sciences (SPP 2431), Research-driven Infrastructure for Advanced Survey-related Data (CIRCLET) measure project number 539634240.

Authors' contributions

A.A. designed and conducted the experiments, analyzed and interpreted the results, and wrote the main manuscript. A.M. proposed the research ideas, contributed to the experimental design, and participated in manuscript writing and revision. M.B. collected the dataset and contributed to the manuscript preparation. All authors read and approved the final manuscript.

Acknowledgements

We thank our colleagues at the Text Technology Lab, Goethe University Frankfurt, for their help with the manual annotation of the perceptual-verbal features used in the inter-annotator agreement study.

References

Abdin M, Aneja J, Behl H, et al (2024) Phi-4 technical report. URL https://arxiv.org/abs/2412.08905, arXiv:2412.08905

AlDahoul N, Rahwan T, Zaki Y (2024) Polytc: a novel bert-based classifier to detect political leaning of youtube videos based on their titles. Journal of Big Data 11(1):80. https://doi.org/10.1186/s40537-024-00946-1, URL https://doi.org/10.1186/s40537-024-00946-1

Ambady N, Rosenthal R (1992) Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. Psychological Bulletin 111(2):256–274. https://doi.org/10.1037/0033-2909.111.2.256, URL https://doi.org/10.1037/0033-2909.111.2.256

Ansari MZ, Aziz M, Siddiqui M, et al (2020) Analysis of political sentiment orientations on twitter. Procedia Computer Science 167:1821–1828. https://doi.org/https://doi.org/10.1016/j.procs.2020.03.201, URL https://www.sciencedirect.com/science/article/pii/S1877050920306669, international Conference on Computational Intelligence and Data Science

Araujo T, Lock I, van de Velde B (2020) Automated visual content analysis (avca) in communication research: A protocol for large scale image classification with pretrained computer vision models. Communication Methods and Measures 14(4):239–265. https://doi.org/10.1080/19312458.2020.1810648, URL https://doi.org/10.1080/19312458.2020.1810648, https://doi.org/10.1080/19312458.2020.1810648

Bagci M, Abusaleh A, Baumartz D, et al (2025) Extending a parliamentary corpus with mps' tweets: Automatic annotation and evaluation using multipartweet. URL https://arxiv.org/abs/2512.11567, arXiv:2512.11567

Bai S, Chen K, Liu X, et al (2025) Qwen2.5-vl technical report. URL https://arxiv.org/abs/2502.13923, arXiv:2502.13923

Buolamwini J, Gebru T (2018) Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Proceedings of the 1st Conference on Fairness, Accountability, and Transparency, pp 77–91

Camacho-Collados J, Pilehvar MT (2018) From word to sense embeddings: a survey on vector representations of meaning. J Artif Int Res 63(1):743–788. https://doi.

org/10.1613/jair.1.11259, URL https://doi.org/10.1613/jair.1.11259

Chang JQ, Rajan D, Vun N (2024) Multimodal few-shot classification without attribute embedding. EURASIP Journal on Image and Video Processing 2024(1):4. https://doi.org/10.1186/s13640-024-00620-9, URL https://doi.org/10.1186/s13640-024-00620-9

Chen T, Lu D, Kan MY, et al (2013) Understanding and classifying image tweets. In: Proceedings of the 21st ACM International Conference on Multimedia. Association for Computing Machinery, New York, NY, USA, MM '13, p 781–784, https://doi.org/10.1145/2502081.2502203, URL https://doi.org/10.1145/2502081.2502203

Connolly E (2025) How does social media content go viral across platforms? modelling the spread of kamala is brat across x, tiktok, and instagram. Journal of Information Technology & Politics 0(0):1–16. https://doi.org/10.1080/19331681.2025.2519052, URL https://doi.org/10.1080/19331681.2025.2519052, https://doi.org/10.1080/19331681.2025.2519052

Dan V, Arendt F (2021) Visual cues to the hidden agenda: Investigating the effects of ideology-related visual subtle backdrop cues in political communication. The International Journal of Press/Politics 26(1):22–45. https://doi.org/10.1177/1940161220936593, URL https://doi.org/10.1177/1940161220936593, https://doi.org/10.1177/1940161220936593

DeepSeek-AI, Guo D, Yang D, et al (2025) Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. URL https://arxiv.org/abs/2501.12948, arXiv:2501.12948

Dumitrescu D (2016) Nonverbal communication in politics: A review of research developments, 2005-2015. American Behavioral Scientist 60(14):1656–1675. https:

//doi.org/10.1177/0002764216678280

FFmpeg Developers (2025) Ffmpeg: Multimedia framework. URL https://ffmpeg.org/ffprobe.html, accessed: 2025-08-01

Gao Y, Liu F, Gao L (2023) Echo chamber effects on short video platforms. Scientific Reports 13(1):6282

Garg A, Ramsay A (2011) Semantic Content Analysis of Video: Issues and Trends, Springer Berlin Heidelberg, Berlin, Heidelberg, pp 443–457. https://doi.org/10.1007/978-3-642-19551-8_16, URL https://doi.org/10.1007/978-3-642-19551-8_16

Guo H, Xie Z, Cao S, et al (2025) Sns-bench-vl: Benchmarking multimodal large language models in social networking services. Withdrawn, arXiv:2505.23065

Gupta V, Mittal T, Mathur P, et al (2022) 3massiv: Multilingual, multimodal and multi-aspect dataset of social media short videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 21064–21075

Jiang AQ, Sablayrolles A, Mensch A, et al (2023) Mistral 7b. URL https://arxiv.org/abs/2310.06825, arXiv:2310.06825

Jin Y, Choi M, Verma G, et al (2024) MM-SOC: Benchmarking multimodal large language models in social media platforms. In: Ku LW, Martins A, Srikumar V (eds) Findings of the Association for Computational Linguistics: ACL 2024. Association for Computational Linguistics, Bangkok, Thailand, pp 6192–6210, https://doi.org/10.18653/v1/2024.findings-acl.370, URL https://aclanthology.org/2024.findings-acl.370/

Katabi N, Simon H, Yakim S, et al (2023) Deeper than you think: Partisanship-dependent brain responses in early sensory and motor brain

regions. Journal of Neuroscience 43(6):1027–1037. https://doi.org/10.1523/JNEUROSCI.0895-22.2022, URL https://www.jneurosci.org/content/43/6/1027, https://www.jneurosci.org/content/43/6/1027.full.pdf

Kosinski M, Khambatta P, Wang Y (2024) Facial recognition technology and human raters can predict political orientation from images of expressionless faces even when controlling for demographics and self-presentation. Am Psychol 79(7):942–955

Köksal FN, Erisen H (2023) Multimodal discourse: An analysis of two political advertising visualsmultimodal tartiŞma: İkİ sİyasİ reklam gÖrselİ Üzerİne bİr analİz. İmgelem 7:249–268. https://doi.org/10.53791/imgelem.1287019

Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33(1):159–174. URL http://www.jstor.org/stable/2529310

Lausen A, Hammerschmidt K (2020) Emotion recognition and confidence ratings predicted by vocal stimulus type and prosodic parameters. Humanities and Social Sciences Communications 7(1):2. https://doi.org/10.1057/s41599-020-0499-z, URL https://doi.org/10.1057/s41599-020-0499-z

de Lima-Santos MF, Gonçalves I, Quiles MG, et al (2025) Visual political communication on instagram: a comparative study of brazilian presidential elections. EPJ Data Science 13. https://doi.org/10.1140/epjds/s13688-024-00502-0

Marchal N, Neudert LM, Kollanyi B, et al (2021) Investigating visual content shared over twitter during the 2019 eu parliamentary election campaign. Media and Communication 9(1):158–170. https://doi.org/10.17645/mac.v9i1.3421, URL https://www.cogitatiopress.com/mediaandcommunication/article/view/3421

Medina Serrano JC, Papakyriakopoulos O, Hegelich S (2020) Dancing to the partisan beat: A first analysis of political communication on tiktok. In: Proceedings of the

12th ACM Conference on Web Science. Association for Computing Machinery, New York, NY, USA, WebSci '20, p 257–266, https://doi.org/10.1145/3394231.3397916, URL https://doi.org/10.1145/3394231.3397916

Meta AI (2024) Llama 3.2: Revolutionizing edge AI and vision with open, customizable models. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/, accessed: 2026-07-27

Molnar C (2019) Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Leanpub

Olteanu A, Castillo C, Diaz F, et al (2019) Social data: Biases, methodological pitfalls, and ethical boundaries. Frontiers in Big Data Volume 2 - 2019. https://doi.org/10.3389/fdata.2019.00013, URL https://www.frontiersin.org/journals/big-data/articles/10.3389/fdata.2019.00013

OpenAI (2025a) Chatgpt (gpt-5). https://chat.openai.com/, large language model. Accessed on September 1, 2025

OpenAI (2025b) gpt-oss-20b. https://platform.openai.com/docs/models/gpt-oss-20b, large language model. Accessed on September 2, 2025

Padhi T, Kursuncu U, Kumar Y, et al (2024) Enhancing cross-modal contextual congruence for crowdfunding success using knowledge-infused learning. https://doi.org/https://doi.org/10.1109/BigData62323.2024.10825252, URL https://arxiv.org/abs/2402.03607, arXiv:2402.03607

Parashar S, Lin Z, Liu T, et al (2024) The neglected tails in vision-language models

Radford A, Kim JW, Hallacy C, et al (2021) Learning transferable visual models from natural language supervision. URL https://arxiv.org/abs/2103.00020, arXiv:2103.00020

Ribeiro MT, Singh S, Guestrin C (2016) "why should i trust you?": Explaining the predictions of any classifier. URL https://arxiv.org/abs/1602.04938, arXiv:1602.04938

Sapag M. P, Rastrilla LP, García AR (2023) TikTok: New Media, Old Propaganda—State of the Art and Possibilities for Political Communication, Springer Nature Singapore, Singapore, pp 3–19. https://doi.org/10.1007/978-981-99-5110-9_1, URL https://doi.org/10.1007/978-981-99-5110-9_1

Schroff F, Kalenichenko D, Philbin J (2015) Facenet: A unified embedding for face recognition and clustering. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 815–823, https://doi.org/10.1109/CVPR.2015.7298682

Shazeer N, Mirhoseini A, Maziarz K, et al (2017) Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. URL https://arxiv.org/abs/1701.06538, arXiv:1701.06538

Snell J, Swersky K, Zemel R (2017) Prototypical networks for few-shot learning. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, NIPS'17, p 4080–4090

Umansky N, Pipal C (2023) Dances, duets, and debates: Analysing political communication and viewer engagement on TikTok. OSF Preprints, https://doi.org/10.31219/osf.io/g8dbu

Wang Y, Yu Z, Wang J, et al (2024) Exploring Vision-Language models for imbalanced learning. International Journal of Computer Vision 132(1):224–237

Wong FMF, Tan CW, Sen S, et al (2016) Quantifying Political Leaning from Tweets, Retweets, and Retweeters . IEEE Transactions on Knowledge & Data Engineering 28(08):2158–2172. https://doi.org/10.1109/TKDE.2016.2553667, URL https://doi.

Listing 1: Countries for downloading top 100 hashtags

```
{
  'German Speaking Countries": ["Germany", "Austria", "Switzerland"],
  "English Speaking Countries": ["United-States", "United-Kingdom", "Canada", "
      Australia", "New-Zealand"]
}
```

ieeecomputersociety.org/10.1109/TKDE.2016.2553667

The appendix is organized as follows. Appendix Appendix A presents the list of countries and hashtags used in the data collection process. Following that, Appendix B provides the main literature-informed prompt that served as the code prompt for our analysis. Appendix Figure C then outlines the conceptual framework used to associate the extracted features with higher-level semantic categories. Subsequently, Appendix Appendix D compiles the visual, audio, and multimodal features proposed by several Large Language Models, with their detailed responses shown in Appendix Appendix E. Finally, Appendix Appendix F describes the aggregation and alignment process based on majority voting across model outputs.

Appendix A List Countries and Hashtags

The Listing 1 lists the German- and English-speaking countries from which the top 100 hashtags were created. The top 100 hashtags are listed in Listing 2. To create the list of the top 100 hashtags in Listing 2, we extracted the top 100 hashtags from each country in Listing 1 via the *X API v2*. The position of the hashtags determines their weight: the lower the position, the higher the weight. We combined the top 100 hashtags from eight countries and fused the outputs, combining the positions and dividing them by the number of countries in which the hashtag is listed. The top 100 hashtags of that combined list can be found in Listing 2.

Listing 2: Top 100 Hashtags of german and english speaking countries

```
{
  'Hashtags': ['President', 'Spotify', 'Mone', 'vätern', 'bäckerei', 'Zeev Buium',
      'Cora', 'Littell', 'Martinez', 'King', 'Wachstum', 'Staatsbürger', 'Profil',
       'Eric', 'stunden tage', 'Abbot', 'The Pistons', 'Jean Silva', 'robbie
      savage', 'wooyoung', 'Dallas Stars', 'Boucher', 'Messi', '$BIG', 'Lovely', '
      trotteln', 'Wert', 'Tobias Harris', 'Bryson', 'Boom Boom', 'Justin Thomas',
      'Semana Santa', 'Jason Tindall', 'Kobbie', 'Sonia', 'mcmahon', 'Ben Keays',
      'wirkung', 'Schweiz', 'warteschlange', 'Verantwortung für Deutschland', '
      Ricarda Lang', 'abbot', 'Siakam', 'Mackie', 'The Angels', 'Ibrox', 'Vanessa'
      , 'Duran', 'Chris Bassitt', 'Cameron', 'David King', 'Covid', 'Christmas', '
      Naz Reid', 'oliver stark', 'Lopes', '#GoodBadUgly', 'Ally McCoist', 'Altay',
       'Anushka', 'Elias Pettersson', 'Origin', 'Common People', 'Santi', 'Ant Man
      ', 'Ober', '#UCFDayofGiving', 'Eddie Howe', 'Theme Park', 'mitch marner', '
      Randwick', 'Auckland', 'jimmy kimmel', 'Owen Michaels', 'Cantlay', 'Rodney
      Dunham', 'Hermes', 'Prada', '#PutThatInYourPipe', 'ASPINALL', 'Kanye', '
      Behandlung', 'Migration', 'Oettinger', 'Thibs', 'Johnny Storm', 'Bailey Ober
      ', 'Spencer Strider', 'Spurs', 'Ausar Thompson', 'Precious Achiuwa', '
      Insider Trading', 'Michaela', 'Portnoy', 'Richard Hughes', 'Kitchener', 'Maß
      nahmen', 'Kompetenz', 'Canes']
}
```

Appendix B Qwen2.5-VL Bai et al. (2025) Prompt

This appendix presents the literature-informed prompt used to extract the PREF perceptual-verbal features and to generate structured semantic descriptions. The complete prompt is provided in Listing 3. Text enclosed in double square brackets [[ ]] indicates optional elements that were included or omitted depending on the modality of the media input (for example, video or image).

Listing 3: VLM prompt used to extract features

```
Based on the following [[video and accompanying audio/image]], extract the key
    characteristics of the [[video]] content.

Please answer in **valid JSON** format with the following fields:

* "contains_people": true or false
* "number_of_people": estimated number (if visible)
* "activity": brief description of what people (if any) are doing
* "indoor_or_outdoor": "indoor" or "outdoor"
* "dominant_color_family": list of color families observed (e.g., ["blue", "
    white", "black"])
* "scene_type": brief description of the setting (e.g., "beach", "conference
    room", "farm")
```

```
* "dress_code": description of clothing styles, such as "formal", "casual", "
    traditional", or "no visible clothing"
* [["motion_intensity": "low", "medium", or "high" based on apparent movement]]
* "objects_detected": list of notable objects (e.g., "car", "microphone", "
    animals", "flag")

[[**Audio / Voice Characteristics**

* "voice_present": true or false
* "number_of_speakers": estimated count
* "speaker_genders": list of detected genders, e.g., ["male", "female"]
* "emotional_tone": general tone of the speech (e.g., "neutral", "angry", "
    excited", "calm")
* "voice_clarity": "clear", "muffled", "noisy"
* "language_spoken": detected language(s)]]

Example output:

{
  "contains_people": true,
  "number_of_people": 2,
  "activity": "interview setting with discussion",
  "indoor_or_outdoor": "indoor",
  "dominant_color_family": ["beige", "gray", "black"],
  "scene_type": "studio",
  "dress_code": "casual",
  [["motion_intensity": "medium",]]
  "objects_detected": ["sofa", "microphone", "table"],
  [["voice_present": true,]]
  "number_of_speakers": 2,
  [["speaker_genders": ["male", "female"],
  "emotional_tone": "calm",
  "voice_clarity": "clear",
  "language_spoken": "English"]]
}
```

Table C1: Feature Families and Corresponding Conceptual Categories

Feature Family	Representative Features	Associated Conceptual Category
Color	Dominant color families, brightness, hue distribution	Visual style and aesthetic tone
Voice	Pitch, speaker gender, vocal clarity	Speaker voice characteristics and communicative style
Objects	Detected objects such as flags, logos, microphones	Symbolic or thematic elements
Scene type	Indoor or outdoor scenes, environment type	Setting and contextual background
Activity	Movement intensity, gesture frequency	Behavioral and action-based cues
Dress code	Clothing type, formality level	Dress appearance and social signaling
Emotional tone	Facial expressions, prosodic cues	Emotional or affective state
Number of people	Crowd size, number of speakers or participants	Audience scale and social density
Motion	Motion intensity, camera movement	Dynamic engagement and visual energy
Other	Unclassified or residual features	Miscellaneous attributes not assigned elsewhere

Appendix C Feature-to-Concept Mapping Framework

This appendix describes the conceptual logic used to associate extracted features with higher-level semantic categories. Instead of relying on raw feature names, each feature is grouped into a family that reflects its perceptual or cognitive relevance. The mapping allows the transition from low-level visual or auditory signals to interpretable semantic dimensions used in our analysis.

The mapping framework follows an interpretive hierarchy in which perceptual features (e.g., color, motion, or sound) are first grouped by shared modality and then linked to corresponding semantic concepts (e.g., style, activity, or emotion). This

abstraction enables a systematic interpretation of multimodal data by associating measurable patterns with cognitively meaningful constructs such as symbolism, social context, and affective tone.

Appendix D Features Extraction from Large Language Models

This appendix presents a compilation of visual, audio, and multimodal features suggested by various Large Language Models (LLMs) (`GPT-5` OpenAI (2025a), `Llama3.2:latest` Meta AI (2024), `GPT-OSS:20b` OpenAI (2025b), `Mistral:latest` Jiang et al. (2023), `Phi-4:latest` Abdin et al. (2024) and `DeepSeek-R1:latest` DeepSeek-AI et al. (2025)) for the analysis of political content. The features are designed to be observable, descriptive, and potentially discriminative across different political styles or parties. Each feature is structured to support machine learning or statistical analysis, ensuring objectivity and reliability in extraction.

The prompt used to elicit these features is shown in Listing 4.

The prompt shown above (see Listing 4) was submitted to each of the selected models in a controlled setting. Each model responded with a structured list of features that it considers relevant for analyzing political video and audio content. The responses were parsed and grouped by modality (visual, audio, and multimodal).

The following sub-sections present a consolidated view of the features suggested by the models, organized according to their respective modalities.

D.1 Visual Features

- **face_presence**: Binary indicator of whether human faces are present in the frame.
- **number_of_faces**: Count of distinct faces detected per frame or segment.
- **face_orientation**: Average yaw, pitch, and roll angles of detected faces to measure gaze direction.

Listing 4: Meta-Prompt for Eliciting Extractable Visual and Audio Features in Political Media

```
You are an expert in multimodal political content analysis. Your task is to suggest the most informative and observable features that should be extracted from political videos or images (with or without accompanying audio) to support downstream classification or interpretation of political messaging, style, or affiliation.

Do not extract features from any specific content. Instead, propose a comprehensive list of descriptive visual and auditory features that can be reliably extracted from typical political content such as speeches, rallies, interviews, debates, or propaganda videos.

Each feature should be:
- Observable (visually or audibly detectable)
- Descriptive (not interpretive or speculative)
- Potentially discriminative across different political styles or parties
- Structured (suitable for machine learning or statistical analysis)

Please return your proposed features in valid JSON format using the following structure:

{
  "visual_features": [list of visual features with short descriptions],
  "audio_features": [list of audio features with short descriptions],
  "optional_multimodal_features": [list of features combining audio and visual cues]
}

Do not include political interpretation or bias in your suggestions. Focus purely on what can be objectively observed and extracted by machines.
```

- **facial_expressions**: Presence and intensity of facial Action Units (AUs) associated with basic expressions (e.g., AU12 for smiling).
- **gesture_type**: Categorical identification of hand gestures (e.g., pointing, open palms, clenched fists).
- **gesture_frequency**: Number of distinct gestures per unit time.
- **body_posture**: Estimation of upright, leaning, or slouched posture using pose detection models.
- **micro-movements**: High-frequency body or head movements captured over short time intervals.
- **scene_composition**: Categorical descriptor of the scene setup (e.g., podium, roundtable, rally, interview).

- **text_on_screen**: OCR-extracted textual elements displayed on screen (e.g., banners, subtitles, slogans).
- **logo_presence**: Binary indicator and class label for any visible organizational or political logos.
- **flag_presence**: Binary indicator and classification of flags or national symbols in the background.
- **audience_visibility**: Binary indicator of audience visibility in the frame.
- **lighting_condition**: Categorical description of lighting (e.g., bright, dim, high contrast).
- **camera_motion**: Descriptors of camera movement (e.g., static, pan, zoom, tilt).
- **color_distribution**: Histogram or dominant colors in the frame, optionally clustered.

D.2 Audio Features

- **speaker_change_rate**: Number of speaker changes per audio segment.
- **pitch_mean**: Average fundamental frequency (F0) of the speaker's voice.
- **pitch_variance**: Variance of pitch over time, indicating tonal variation.
- **speech_rate**: Number of spoken words or syllables per minute.
- **voice_intensity**: Average loudness in decibels over time.
- **pauses_per_minute**: Number of non-speech pauses per minute.
- **emotion_prosody_features**: Auditory correlates of emotion such as pitch contour, jitter, shimmer, and harmonicity.
- **background_music_presence**: Binary indicator of non-speech music audio.
- **background_music_type**: Categorical descriptor of music genre or tempo (e.g., orchestral, upbeat, solemn).
- **crowd_noise_level**: Average decibel level of crowd-related sounds (e.g., cheering, clapping).

- **applause_events**: Time-stamped events where applause is detected.
- **chanting_presence**: Binary indicator of coordinated chanting or slogan repetition.

D.3 Multimodal Features

- **lip_sync_score**: Correlation score between lip movement and speech audio to detect dubbing or manipulation.
- **gesture_speech_alignment**: Temporal alignment measure between gestural emphasis and prosodic peaks.
- **emotion_congruence**: Cross-modal consistency between facial expression and vocal tone.
- **audience_response_alignment**: Temporal alignment between speaker utterances and audience reactions (e.g., applause, laughter).
- **speaker_gaze_direction_vs_audio_focus**: Comparison of gaze direction with detected sound source or speaking individual.

Appendix E Model-Suggested Features for Multimodal Political Content Analysis

This appendix presents the structured lists of visual, audio, and multimodal features proposed by multiple language models- *GPT-5* OpenAI (2025a), *Llama3.2:latest* Meta AI (2024), *GPT-OSS:20b* OpenAI (2025b), `Mistral:latest` Jiang et al. (2023), *Phi-4:latest* Abdin et al. (2024) -in response to a prompt asking for observable and potentially discriminative features useful for analyzing political video and audio content.

Table E2: Visual features extracted by *GPT-5* OpenAI (2025a).

Feature Name	Description
face_presence	Binary indicator of whether human faces are present in the frame.
number_of_faces	Count of distinct faces detected per frame or segment.
face_orientation	Average yaw, pitch, and roll angles of detected faces to measure gaze direction.
facial_expressions	Presence and intensity of facial Action Units (AUs) associated with basic expressions.
gesture_type	Categorical identification of hand gestures.
gesture_frequency	Number of distinct gestures per unit time.
body_posture	Estimation of upright, leaning, or slouched posture.
micro-movements	High-frequency body or head movements captured over short time intervals.
scene_composition	Categorical descriptor of the scene setup.
text_on_screen	OCR-extracted textual elements displayed on screen.
logo_presence	Binary indicator and class label for visible organizational or political logos.
flag_presence	Binary indicator and classification of flags or national symbols.
audience_visibility	Binary indicator of audience visibility in the frame.
lighting_condition	Categorical description of lighting conditions.
camera_motion	Descriptors of camera movement (e.g., pan, zoom, tilt).
color_distribution	Histogram or dominant colors, optionally clustered.
clothing_type	Categorical descriptor of attire.
dominant_object_types	Classes of objects detected in the frame.

Model: *GPT-5* OpenAI (2025a)

Model: `Llama3.2:latest` Meta AI (2024)

Model: `GPT-OSS:20b` OpenAI (2025b)

Model: `Mistral:latest` Jiang et al. (2023)

Model: `Phi-4:latest` Abdin et al. (2024)

Model: `DeepSeek-R1:latest` DeepSeek-AI et al. (2025)

Appendix F Cross-Model Convergence in LLM-Suggested Multimodal Features

To quantify the degree of agreement among large language models (LLMs) in identifying salient features for multimodal political content analysis in appendix Appendix E,

Table E3: List of audio features extracted by *GPT-5* OpenAI (2025a)

Feature	Description
speech_presence	Binary indicator of whether speech is present.
speech_duration_ratio	Ratio of speech-active time to total segment length.
speaker_turn_count	Number of speaker changes per segment.
pitch_mean	Average fundamental frequency (F0).
pitch_variance	Variance of pitch over time.
speech_rate	Number of spoken words or syllables per minute.
voice_intensity	Average loudness in decibels.
pauses_per_minute	Number of non-speech pauses per minute.
emotion_prosody	Acoustic correlates of emotion such as pitch contour and jitter.
background_music_presence	Binary indicator of music audio.
background_music_type	Descriptor of music genre or tempo.
crowd_noise_level	Average decibel level of crowd sounds.
applause_events	Time-stamped events where applause is detected.
chanting_presence	Binary indicator of coordinated chanting.

Table E4: Multimodal features extracted by *GPT-5* OpenAI (2025a).

Feature Name	Description
lip_sync_score	Correlation score between lip movement and speech audio.
gesture_speech_alignment	Temporal alignment of gestures with prosodic peaks.
emotion_congruence	Cross-modal consistency between facial expression and vocal tone.
audience_response_alignment	Temporal alignment between speaker utterances and audience reactions.
speaker_gaze_direction_vs_audio_focus	Comparison of gaze direction with audio source or speaker.

we conducted a canonicalized majority voting procedure. Feature names returned by individual models were first normalized and aligned using a controlled alias mapping to account for lexical variation (e.g., `lighting_intensity` → `lighting_condition`, `mean_pitch` → `pitch_mean`).

Figure 2 visualizes the aligned output. Each bar represents a canonical feature, segmented by the models that proposed it. Multi-colored bars indicate cross-model

Table E5: Visual features extracted by `Llama3.2:latest` Meta AI (2024).

Feature Name	Description
Frame rate	N/A
Resolution (width, height)	N/A
Aspect ratio	N/A
Color palette diversity	N/A
Background noise presence	N/A
Text overlay presence and type (e.g., subtitles, logos) Camera shake or movement	N/A
Lighting conditions	N/A
Object occlusion rate	N/A
Action line velocity	N/A

Table E6: List of audio features extracted by `Llama3.2:latest` Meta AI (2024).

Feature	Description
Audio format (e.g., MP3, WAV)	NA
Sampling rate	NA
Bit depth	NA
Decibel level	NA
Speech-to-noise ratio (SNR)	NA
Music or sound effects presence and type	NA
Audio pitch range	NA

Table E7: Multimodal features extracted by `Llama3.2:latest` Meta AI (2024).

Feature Name	Description
Audio-Visual Sync	Alignment between audio and video frames measured as the mean absolute frame difference.
Text Color Contrast	Contrast ratio between subtitle or logo text and background.
Background Noise Type	Classification of audio noise, such as ambient, crowd, or instrumental.

consensus, whereas single-colored bars highlight model-specific features. The horizontal axis reflects the number of models contributing that feature.

Several features exhibit strong cross-model agreement. These include:

- **Audio:** `speech_rate`, `pitch_mean`, `pitch_variance`

Table E8: Visual features extracted by `GPT-OSS:20b` OpenAI (2025b).

Feature Name	Description
frame_resolution	Width and height in pixels of each video frame.
aspect_ratio	Ratio of width to height for each frame (e.g., 16:9, 4:3).
dominant_color	Most frequent hue/saturation value in the frame.
lighting_intensity	Average luminance value of pixels per frame.
camera_motion_type	Detected motion type: static, panning, tilting, zooming.
shot_type	Classification of frame: close-up, medium, wide, establishing.
face_count	Number of distinct faces detected in the frame.
face_orientation_distribution	Histogram of yaw angles (forward, left, right) across detected faces.
hand_gesture_presence	Binary flag indicating if any hand gesture (wave, raise) is detected.
text_overlay_presence	Boolean indicating if on-screen text (subtitles, captions) is present.
text_overlay_properties	Text font style (serif, sans-serif), size (pixels), and placement (top, bottom).
flag_or_banner_detection	Presence and dominant color of political flags or banners in the frame.
logo_detection	Presence and bounding box coordinates of recognizable logos.
crowd_density	Estimated number of people in the background beyond the main speaker.
eye_contact_frequency	Count of frames where the main speaker's gaze aligns with the camera.
speech_duration_per_frame	Time the speaker is visible speaking in each second of the video.

- **Visual:** `aspect_ratio`, `lighting_condition`, `flag_or_banner_detection`
- **Multimodal / Hybrid:** `facial_expressions`, `background_noise_type`, `sampling_rate`

This pattern suggests that, despite architectural differences and pretraining variation, LLMs converge on a core set of observable and potentially discriminative features. Meanwhile, a long tail of less frequent features (proposed by only one or two models) may capture finer-grained or model-specific priors about political media representations.

This voting-based analysis serves two purposes. First, it provides a data-driven mechanism to prioritize features that are likely to generalize across both domains and

Table E9: Audio features extracted by `GPT-OSS:20b` OpenAI (2025b).

Feature Name	Description
sampling_rate	Number of audio samples per second (Hz).
bit_depth	Number of bits used to represent each audio sample.
channel_count	Number of audio channels (mono, stereo).
average_decibel	Mean loudness level over the entire clip (dB).
speech_rate	Estimated words per minute spoken by the main speaker.
pitch_mean	Average fundamental frequency (Hz) of the speaker's voice.
pitch_variance	Standard deviation of the speaker's fundamental frequency.
voice_gender	Predicted gender of the main speaker (male, female, other).
background_noise_type	Classification of non-speech audio (crowd, applause, music, silence).
pause_duration_distribution	Histogram of silence intervals longer than 0.5 seconds.
filler_word_rate	Frequency of fillers (um, uh, like) per 100 words.
word_token_count	Total number of recognized words in the audio.
spectral_centroid	Center of mass of the spectrum, indicating perceived brightness.
temporal_entropy	Measure of unpredictability in amplitude over time.

modeling paradigms. Second, it supports the construction of a reduced yet diverse feature schema that can be used in downstream prompt design or statistical modeling for political communication studies.

F.1 Deriving an Interpretable Feature Schema from LLM-Based Voting

To guide downstream feature extraction in a scalable yet interpretable way, we constructed a data-driven prompt derived from majority voting across several large language models (LLMs). Each model was prompted to list observable, descriptive features suitable for political video and audio content analysis (see Listing 4). The resulting feature sets were aligned and aggregated into canonical form, as shown in Appendix F.

Table E10: Optional multimodal features extracted by `GPT-OSS:20b` OpenAI (2025b).

Feature Name	Description
lip_speech_sync_error	Mean absolute difference (in milliseconds) between lip-movement timestamps and spoken phoneme timestamps.
text_word_alignment	Timestamp alignment between on-screen text and spoken words (mean offset).
visual_event_audio_correlation	Correlation coefficient between detected visual events (clap, wave) and audio amplitude peaks.
crowd_noise_visual_trigger	Presence of crowd faces when background noise classification is "crowd".
light_change_audio_peak	Detection of sudden brightness changes coinciding with audio peaks greater than 20 dB.
hand_wave_audio_trigger	Occurrences where a hand wave aligns within ±200 milliseconds of a spoken keyword.

While the aggregated set included both low-level (e.g., `aspect_ratio`, `sampling_rate`) and high-level features (e.g., `flag_presence`, `facial_expressions`), not all features contribute meaningfully to interpretation or classification. We therefore excluded purely technical metadata (such as frame resolution or bit depth), which, while useful for signal processing, lack semantic discriminability across political styles or contexts.

The final schema prioritizes features that are:

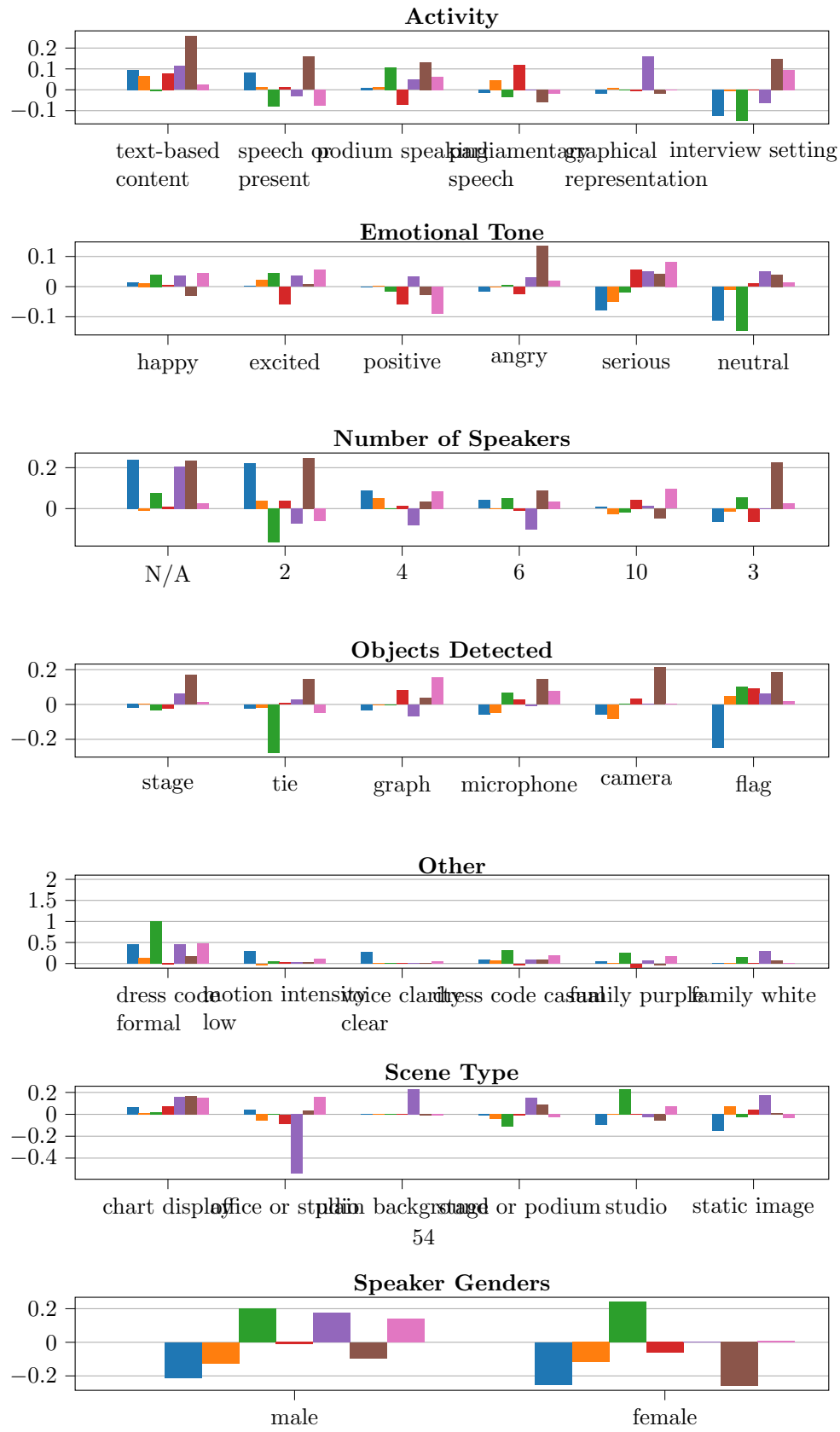
- **Observable:** Directly extractable by machine from image or audio signals
- **Interpretive:** Informative for political style, communication strategy, or audience targeting
- **Multimodal:** Derived from or aligned across visual and auditory streams

Listing 5 shows the final prompt schema, used for semantic feature extraction in our experimental setup.

Table E11: Visual features extracted by `Mistral:latest` Jiang et al. (2023).

Feature Name	Description
Frame Resolution	Width and height in pixels of each video frame.
Aspect Ratio	Ratio of width to height for each frame (e.g., 16:9, 4:3).
dominant Color	Most frequent hue/saturation value in the frame.
lighting Intensity	Average luminance value of pixels per frame.
camera motion type	Detected motion type: static, panning, tilting, zooming.
shot type	Classification of frame: close-up, medium, wide, establishing.
face count	Number of distinct faces detected in the frame.
face orientation distribution	Histogram of yaw angles (forward, left, right) across detected faces.
hand gesture presence	Binary flag indicating if any hand gesture (wave, raise) is detected.
text overlay presence	Boolean indicating if on-screen text (subtitles, captions) is present.
text overlay properties	Text font style (serif, sans-serif), size (pixels), and placement (top, bottom).
flag or banner detection	Presence and dominant color of political flags or banners in the frame.
logo detection	Presence and bounding box coordinates of recognizable logos.
crowd density	Estimated number of people in the background beyond the main speaker.
eye contact frequency	Count of frames where the main speaker's gaze aligns with the camera.
speech duration per frame	Time the speaker is visible speaking in each second of the video.

Fig. C1: Normalized feature contributions across feature categories and political groups.



Each subplot shows the top features within a single feature category (panel title) along the x-axis, with the normalized contribution values on the y-axis. Positive values indicate features that contribute positively toward the corresponding class label, while negative values indicate features that contribute negatively. Feature labels on the x-axis are shortened for readability. Colors represent different political groups (Group A: ■; Group B: ■; Group C: ■; Group D: ■; Group E: ■; Group F: ■; Group G: ■). The y-axis is normalized to the maximum absolute contribution across all features.

Table E12: Audio features extracted by `Mistral:latest` Jiang et al. (2023).

Feature Name	Description
sampling rate	Number of audio samples per second (Hz).
bit depth	Number of bits used to represent each audio sample.
channel count	Number of audio channels (mono, stereo).
average decibel	Mean loudness level over the entire clip (dB).
speech rate	Estimated words per minute spoken by the main speaker.
pitch mean	Average fundamental frequency (Hz) of the speaker's voice.
pitch variance	Standard deviation of the speaker's fundamental frequency.
voice gender	Predicted gender of the main speaker (male, female, other).
background noise type	Classification of non-speech audio (crowd, applause, music, silence).
pause duration distribution	Histogram of silence intervals longer than 0.5 seconds.
filler word-rate	Frequency of fillers (um, uh, like) per 100 words.
word token count	Total number of recognized words in the audio.
spectral centroid	Center of mass of the spectrum, indicating perceived brightness.
temporal entropy	Measure of unpredictability in amplitude over time.

Table E13: Optional multimodal features extracted by texttttMistral:latest Jiang et al. (2023).

Feature Name	Description
lip-speech sync error	Mean absolute difference (in milliseconds) between lip-movement timestamps and spoken phoneme timestamps.
text word alignment	Timestamp alignment between on-screen text and spoken words (mean offset).
visual-event-audio correlation	Correlation coefficient between detected visual events (clap, wave) and audio amplitude peaks.
crowd noise visual trigger	Presence of crowd faces when background noise classification is "crowd".
light-change audio-peak	Detection of sudden brightness changes coinciding with audio peaks greater than 20 dB.
hand-wave audio-trigger	Occurrences where a hand wave aligns within $\pm$200 milliseconds of a spoken keyword.

Table E14: Visual features extracted by `Phi-4:latest` Abdin et al. (2024).

Feature Name	Description
Speaker Appearance	Analysis of clothing style, color, and formality which may indicate political affiliation or campaign period.
Facial Expressions	Detection and classification of expressions such as smiling, frowning, raising eyebrows, etc., to assess emotional tone.
Body Language	Assessment of posture, gestures, hand movements, and proximity to others which can indicate confidence or aggression.
Background Elements	Identification of logos, flags, text banners, or other symbols that may hint at political messages or affiliations.
Audience Engagement	Evaluation of audience reactions such as applause, cheering, booing, or silence which can reflect support or dissent.
Text Overlays	Extraction and recognition of text displayed on screen for messages or slogans.
Scene Transitions	Detection of cuts, fades, wipes, etc., which can indicate editing style and pacing in the content.

Table E15: Audio features extracted by `Phi-4:latest` Abdin et al. (2024).

Feature Name	Description
Speech Rate	Measurement of words per minute to assess urgency or calmness in delivery.
Pitch Variation	Analysis of frequency changes which can indicate emotional intensity or emphasis.
Volume Dynamics	Tracking changes in loudness that may reflect passion, authority, or emphasis.
Background Noise	Detection and classification of ambient sounds such as applause, cheering, or crowd noise to assess context.
Speech Intonation Patterns	Recognition of rising or falling intonations that may signify questions, statements, or commands.

Table E16: Optional multimodal features extracted by `Phi-4:latest` Abdin et al. (2024).

Feature Name	Description
Speech Pause Temporal Patterns	Analysis of pauses in speech over time.
Speech Alignment with Visuals	Correlation between speaker emphasis and visual cues.
Body Language Features	Non-verbal communication through gestures, nodding, etc.
Text Legibility on Screen	Assessment of text clarity in on-screen content.
Voice Timbre Match	Correlation between a speaker's voice characteristics and visual cues.
Use of Digital Tools	Inclusion of digital elements like graphics, animations.

Table E17: Visual features extracted by `DeepSeek-R1:latest` DeepSeek-AI et al. (2025).

Feature Name	Description
Speaker Appearance	Analysis of clothing style, color, and formality which may indicate political affiliation or campaign period.
Facial Expressions	Detection and classification of expressions such as smiling, frowning, raising eyebrows, etc., to assess emotional tone.
Body Language	Assessment of posture, gestures, hand movements, and proximity to others which can indicate confidence or aggression.
Background Elements	Identification of logos, flags, text banners, or other symbols that may hint at political messages or affiliations.
Audience Engagement	Evaluation of audience reactions such as applause, cheering, booing, or silence which can reflect support or dissent.
Text Overlays	Extraction and recognition of text displayed on screen for messages or slogans.
Scene Transitions	Detection of cuts, fades, wipes, etc., which can indicate editing style and pacing in the content.

Table E18: Audio features extracted by `DeepSeek-R1:latest` DeepSeek-AI et al. (2025)

Feature Name	Description
Speech Rate	Measurement of words per minute to assess urgency or calmness in delivery.
Pitch Variation	Analysis of frequency changes which can indicate emotional intensity or emphasis.
Volume Dynamics	Tracking changes in loudness that may reflect passion, authority, or emphasis.
Background Noise	Detection and classification of ambient sounds such as applause, cheering, or crowd noise to assess context.
Speech Intonation Patterns	Recognition of rising or falling intonations that may signify questions, statements, or commands.

Table E19: Optional multimodal features extracted by `DeepSeek-R1:latest` DeepSeek-AI et al. (2025).

Feature Name	Description
Speech Pause Temporal Patterns	Analysis of pauses in speech over time.
Speech Alignment with Visuals	Correlation between speaker emphasis and visual cues.
Body Language Features	Non-verbal communication through gestures, nodding, etc.
Text Legibility on Screen	Assessment of text clarity in on-screen content.
Voice Timbre Match	Correlation between a speaker's voice characteristics and visual cues.
Use of Digital Tools	Inclusion of digital elements like graphics, animations.

Listing 5: LLM-Derived Prompt for Interpretable Multimodal Feature Extraction

```
Based on the following [video and accompanying audio/image], extract the key
    characteristics of the video content.

Please answer in **valid JSON** format with the following fields:

{
  "visual_features": {
    "flag_or_banner_presence": "boolean",
    "logo_presence": "boolean",
    "lighting_condition": "string (e.g., 'bright', 'dim')",
    "camera_motion_type": "enum ['static', 'pan', 'zoom']",
    "shot_type": "enum ['close-up', 'wide']",
    "face_count": "integer",
    "face_orientation_distribution": "list (e.g., ['forward'])",
    "hand_gesture_presence": "boolean",
    "facial_expressions": "list (e.g., ['smiling', 'frowning'])",
    "text_overlay_presence": "boolean",
    "text_overlay_properties": {
      "font": "string",
      "placement": "enum ['top', 'bottom']",
      "size": "integer"
    }
  },
  ["audio_features": {
    "speech_rate": "integer (words per minute)",
    "pitch_mean": "float (Hz)",
    "pitch_variance": "float",
    "background_noise_type": "string (e.g., 'crowd', 'applause')"
  }],
  "multimodal_features": {
    "lip_speech_sync_error": "float (ms)",
    "text_word_alignment": "float (ms)",
    "facial_expressions": "repeat from visual/audio if applicable",
    "background_noise_type": "repeat if cross-modal"
  }
}
```