An Image is Worth 40.38 Words: Partisanship and Attention in Videos*

Andrea Ciccarone

This version: November 2, 2025

Click here for the latest version

Abstract

Political news today is consumed in a fragmented, attention-scarce, and predominantly video-based media environment. Yet existing measures of media partisanship focus almost exclusively on text and assume attentive information processing. This paper analyzes how partisanship emerges and operates in video news when information is conveyed jointly through images and text and attention is limited. I develop a multimodal framework that embeds text and images of a video in a shared semantic space to quantify partisan signals, training it on political advertisements and applying it to video news clips. The analysis shows that attention governs the relative informational strength of each modality: images dominate when exposure is brief, conveying partisanship rapidly through emotional cues, while text conveys it more slowly through issue-based content when attention is sustained. A survey experiment with real news footage corroborates these patterns: partisan images elicit rapid emotional and behavioral responses, while partisan text shifts policy attitudes only with sustained exposure. Together, the results characterize partisanship in modern video media as multimodal and attention-dependent, with different modalities influencing different types of viewers.

*I am grateful to Andrea Prat for his invaluable support and advice. I thank Luigi Caloi, Alessandra Casella, Mark Dean, Laura Doval, Amir Feder, Suresh Naidu, Jacopo Perego, Jonah Rockoff, Jesse Schreger, and Andrey Simonov for helpful feedback and guidance, as well as discussants and participants at the Columbia Applied Microeconomics Colloquium, the Columbia Business School Economics Seminar, and the Columbia Political Economy Seminar. All errors are my own.

1 Introduction

Partisan media play a central role in shaping political attitudes, policy preferences, and voting behaviors (DellaVigna and Kaplan, 2007; Gentzkow and Shapiro, 2010; Martin and Yurukoglu, 2017; Simonov et al., 2022; Djourelova, 2023; Ash and Galletta, 2023). Today, partisan information is consumed in a highly fragmented and intensely competitive media landscape (Newman et al., 2025). The proliferation of digital platforms, streaming services, and social networks has vastly expanded the supply of political content while diffusing attention across countless channels. In this "information-rich world" (Simon, 1971), platforms compete for scarce attention in formats optimized for speed and engagement, rather than deliberation.[1]

A second feature characterizes the contemporary media environment: it is overwhelmingly video-based. In 2025, video networks were one of the two primary sources of news consumption—more than one-third of U.S. adults reported regularly getting news from YouTube, and the share obtaining news on TikTok rose from 3 percent to 20 percent over the past five years (Pew Research Center, 2025). The other main source of news remains television, another video-based medium. Most political media content is thus distributed multimodally, through a combination of images and words.[2]

Together, these patterns define a media environment in which political information is both low-attention and visual: it is processed quickly and conveyed through images as well as text. Understanding partisanship in this environment thus requires tools that can quantify ideological signals embedded in both modalities, and assess how these signals operate when attention is scarce. Existing measures of media partisanship and its effects, however, rely almost exclusively on text—articles, transcripts, or posts—under the implicit assumption of full information and attentive processing (Gentzkow and Shapiro, 2010; Martin and Yurukoglu, 2017; Gentzkow et al., 2019).

This paper fills this gap by analyzing how partisanship emerges and functions in a fragmented, video-based media environment, explicitly accounting for the role of attention and modality. To this aim, I address the following questions. How does attention scarcity shape the expression of partisanship in this environment? How do different modalities—text and images—contribute to these partisan signals and their persuasive effects? Through which channels do visual and textual signals convey partisanship?

The central finding of this paper is that partisanship in modern video media is multimodal and attention dependent. Text and images convey distinct partisan signals, and their relative importance varies with effective exposure length. Even under full attention—that is, when the entire

[1] As Simon (1971, p. 40-41) early argued, "a wealth of information creates a poverty of attention and a need to allocate that attention efficiently among the overabundance of information sources that might consume it."

[2] According to Newman et al. (2025), 2025 was the first year in which the share of Americans reporting social and video networks as a source of news in the previous week (54%) exceeded that of television (50%).

informational content of a video is processed—measures based solely on text systematically underestimates partisan slant, reflecting the omission of visual information. This understatement of partisanship becomes increasingly more severe as attention is limited. Under attention constraints, images convey partisan meaning far more effectively than text, which requires time to accumulate and unfold.

A second result concerns the mechanism through which images convey partisanship. Visuals are not just a faster form of text; they operate through a different channel: emotion. The partisan component of images is carried disproportionately by affective cues such as fear, anger, empathy, and pride, whereas textual signals rely primarily on issue content. The role of emotion is particularly relevant in this setting, as affective cues play a central role in modern media consumption, where emotionally charged content captures engagement more effectively than neutral or deliberative messages (Lee and Theokary, 2021; Sánchez-Fernández and Jiménez-Castillo, 2021).[3]

To establish these results, I proceed in three steps. First, I develop a scalable multimodal framework to measure the partisanship of political video content by jointly analyzing its textual and visual components. I train a neural network on more than 15,000 televised political advertisements from U.S. House, Senate, and Presidential elections between 2016 and 2020, labeled by party affiliation of the ad's sponsor.[4] For each ad, I extract video frames and transcript segments, embed both in a shared semantic space using CLIP (Radford et al., 2021)—a state-of-the-art multimodal neural network trained on 400 million image–text pairs—and train separate classifiers for each modality. The resulting image- and text-based predictions achieve high accuracy on held-out ads but are far from perfectly correlated. When combined in a joint multimodal model, the two modalities yield higher predictive accuracy than either alone, implying that each captures partisan information not contained in the other.

Second, I transfer the models to political news clips drawn from the official YouTube channels of major U.S. networks—including Fox News, MSNBC, and CNN—as well as a set of local networks and digital outlets. Each video is represented through CLIP embeddings for both text (transcripts) and images (frames), the same semantic space used for political ads. I then apply the classifier trained on political advertisements to these representations, in order to estimate the partisanship of news videos. This approach yields a quantitative measure of partisan orientation for news content derived separately from text, from images, or from a combined specification that integrates both modalities. This exercise not only provides an out-of-sample validation of the partisanship model, but also enables the analysis of how partisanship operates in news videos under varying levels of attention and across modalities.

[3]Recent work further highlights the role of emotion in shaping how individuals process information, form partisan attachments, and translate exposure into political behavior (eg. Algan et al. (2025)).

[4]The idea of using party sponsorship to infer the partisan orientation of political advertisements follows Fowler et al. (2021), who use a similar party label classification strategy based on specifically coded ad features.

I begin by examining how each modality signals partisanship under full attention—that is, when the full informational content of a segment is available and processed. Under full attention, the text-only measure understates the partisan gap between Fox News and MSNBC by about 5.5 percentage points relative to the joint multimodal estimate. This gap arises because visuals contain partisan information that text omits, even when all content is observed. Still, when attention is unconstrained, the textual channel remains the stronger single predictor of partisanship. Ignoring visuals therefore yields an incomplete—but not fundamentally distorted—picture of partisan slant.

The relative signaling strength of the text and image channels reverses once attention becomes limited. To study this, I simulate limited exposure by randomly sampling short segments—or "chunks"—from each video. For each chunk, I embed both modalities in the shared image–text space and aggregate the resulting representations to re-estimate partisanship at the video level. This approach approximates the information available to viewers who consume political content in brief, fragmented clips rather than through sustained viewing.[5] As exposure shortens, the two modalities lose information at different rates: images retain most of their partisan signal, while text rapidly loses predictive power. Under these conditions, visuals dominate the transmission of partisan meaning, while textual cues require sustained exposure to convey it effectively. When exposure falls to just 10 seconds (two chunks), the understatement of the Fox News-MSNBC partisan gap in the text-only measure becomes three times larger than under full attention.

Finally, I examine whether images—beyond their speed of transmission—encode different kinds of information than text. I map the image and text embeddings of videos onto interpretable "concepts" adapting the sparse concept embedding method from Bhalla et al. (2024). This allows me express each modality within a video as a linear combination of topics (e.g., immigration, jobs, healthcare) and emotions (e.g., fear, anger, joy).[6] This decomposition serves two aims. First, it allows direct comparison of how strongly each modality loads on emotional versus issue content. Second, it measures how strongly each modality depends on individual concepts—for example, how much the visual component of a video shifts when "fear" is removed; larger shifts indicating greater dependence on that concept. I find that images load more heavily on emotion concepts, whereas text is primarily topic-based. Furthermore, removing an emotion concept alters image representations far more than textual ones, showing that visual content depends more heavily on emotional cues.

Removing individual concepts from image and text representations effectively creates counterfactual versions of each modality—showing how an image or transcript would appear in the absence

[5]Because text and images are represented in the same semantic space, this framework allows me to express their informational content on a common scale. The title of the paper reflects this equivalence: the first image of a video convey as much partisan information as 15 seconds of text, approximately 40 words.

[6]The emotion vocabulary follows Cowen and Keltner (2017), while the topic vocabulary is derived from the political ad metadata.

of a given concept. I use this approach to open the model's black box and examine how specific concepts shape partisan predictions. For each concept, I estimate how the model's prediction of party classification changes when that concept is removed—the Concept Treatment Effect (CTE). Emotional concepts generate larger effects in images than in text, and these effects are largely uncorrelated across modalities, indicating that emotions influence partisan classification through distinct channels. By contrast, topic-level CTEs are highly correlated across modalities, with larger magnitudes in texts. Together, these results show the visual and textual channels also differ in substance: images convey partisanship primarily through emotion, while text does so through issues.

The third and final step tests whether the modality–attention relationship identified in the empirical analysis of videos translates into meaningful effects of partisan video news exposure on viewers. To identify these effects, I conduct a survey experiment in which respondents are randomly assigned to view either a short ($\approx$30 seconds) or long ($\approx$2 minutes) news clip about the same immigration event. Clip length exogenously varies effective exposure, while text and image partisanship are independently manipulated within each length in a $2 \times (2 \times 2)$ design.

Guided by the partisanship model, the clips are drawn from real Fox News and MSNBC broadcasts, providing the variation in modality partisanship. I refer to these as Republican and Democratic text and image treatments. Transcripts are cleaned to remove identifying cues and re-dubbed using voice generation to standardize delivery across treatments. Visuals are preprocessed to remove channel identifiers while retaining the authentic immigration-related footage, ensuring that partisan cues derive from content rather than recognizable branding. The footage is presented in its original order.

The experiment shows that images and text influence viewers in distinct ways, consistent with their empirical properties. Visuals act quickly and dominate under short exposure. Republican images increase the perceived Republican leaning of the broadcast by 0.2 standard deviations, and the effect remains stable across both short and long clips. They also heighten negative emotions such as fear, anger, and disgust, while reducing sympathy. In short clips, I find evidence that, beyond their emotional effects, images also influence behavior: Republican images reduce the probability of donating to a pro-immigrant charity. Yet visuals do not shift immigration policy attitudes at either length. Their strength lies in shaping perceptions, emotions, and immediate actions rather than issue-based beliefs.

By contrast, text affects policy attitudes only under sustained exposure. In long clips, Republican text increases anti-immigration attitudes. In short clips, however, textual cues have no measurable effect on attitudes. I find weak evidence that text conveys a partisan signal under long exposure, but this effect disappears when exposure is brief. Overall, text influences attitudes—and, to a limited extent, perceptions of partisanship—only when viewers have time to process it, with little impact on emotion and behaviors.

I next examine heterogeneity in these effects by party affiliation and find marked differences in how different types of viewers respond to each modality. Democrats are significantly responsive to textual persuasion under sustained exposure: Republican text in long clips significantly increases their anti-immigration attitudes. Republicans, by contrast, show little attitudinal response to text but react emotionally and behaviorally to visuals—Republican images in short clips reduce their likelihood of donating to a pro-immigrant charity. In general, Republicans are largely unresponsive to long exposure on any dimension. Overall, Democrats respond to long text exposure, whereas Republicans respond more strongly to short image exposure.

Taken together, the empirical and experimental results reveal a division of labor across modalities. Visuals dominate when attention is scarce, operating through an affective and instantaneous channel. Text regains importance only with sustained exposure: its propositional content accumulates more slowly and, over time, shifts attitudes. Visuals have a comparative advantage in speed—immediately recognized as partisan and capable of triggering emotion even under brief exposure—whereas text operates more slowly but through substance, altering beliefs only when attention persists. The distinct properties of the two modalities therefore produce measurable differences in how video news shapes both rapid, emotion-driven reactions and slower, content-driven preferences.

This paper builds on a large literature in media economics that examines the ideological slant of news content and its effects on political beliefs and behavior (Gentzkow and Shapiro (2010); Martin and Yurukoglu (2017); Simonov et al. (2022); Djourelova (2023); Ash and Galletta (2023), Ash and Poyker, 2024). That work has shown how variation in media exposure can shift policy preferences, voting patterns, and even political engagement. Methodologically, the paper connects to research on causal concepts and interpretability in deep learning (Kim et al. (2018); Feder et al. (2022); Bhalla et al. (2024)), applying these tools to the domain of political communication to identify and interpret the features that drive partisan classification in both text and images.

While social psychology and communication studies have long emphasized the distinctive emotional and cognitive effects of visual stimuli (Nelson et al. (1976); Baddeley (1992); Iyer and Oldmeadow (2006); Dahmen (2015)), only recently has work in political economy begun to quantify partisan content in images (Ash et al. (2021); Boxell, 2021 Caprini (2024); Torres (2024)). The closest related work is probably Caprini (2024), which develop a dictionary-based method to measure visual bias in online news images and show that partisan visuals can shift opinions independently of accompanying text.

My paper departs from this literature both substantively and methodologically. Conceptually, it is the first to study partisanship in images and text jointly and in comparison, framing their relationship through the lens of attention. Methodologically, it analyzes videos rather than static images, using a unified approach that embeds text and visuals in the same semantic space. This shift is

essential to study the attention–modality relationship that lies at the core of the paper. The design captures ideological signals across modalities without relying on hand-coded visual features. In doing so, the paper also differs from the approach in Fowler et al. (2021), which measures partisanship in political advertisements using supervised classification on specifically coded features.

The findings also relate to research on the role of emotion in shaping political attitudes (Andries et al. (2024); Algan et al. (2025)) and to experimental work on political persuasion. The experimental design is inspired by Afrouzi et al. (2024), which emphasizes the persuasive role of leader identity and message source. In contrast, the focus of our experiment is on message modality. By manipulating partisan cues in images and text independently within real television news coverage, the design isolates their separate and joint causal effects on viewers' preferences, behaviors and emotional responses.

The remainder of the paper proceeds as follows. Section 2 introduces the partisanship model and describes its construction and training on political advertisements. Section 3 applies the model to video news from YouTube, and examines how attention shapes the expression of partisanship across modalities. Section 4 analyzes the conceptual structure of partisanship through the concept decompositions. Section 5 presents the survey experiment testing the modality–attention relationship on viewers. Section 6 concludes with implications for media bias measurement, political persuasion, and policy design.

2 Partisanship Model of Political Ads

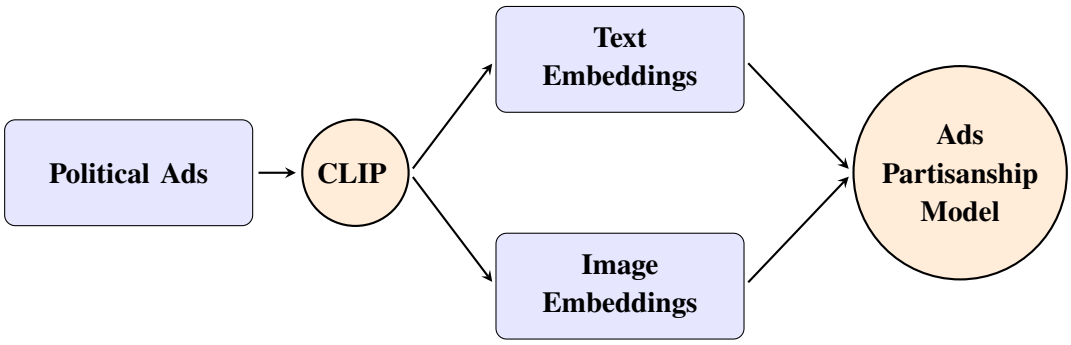


Figure 1: Political Ads Partisanship Model

Notes: Political ads are passed through the CLIP model to generate aligned embeddings for both text and image content. These embeddings are then used as inputs to classification models, which predict the party affiliation of the ad sponsor.

The first building block of this paper is a methodology to quantify the partisanship of political video content, both visual and textual. It treats text and images as complementary signals, ex-

tracting and comparing the ideological cues each contains. The starting point follows a tradition in political economy that estimates ideological positioning from language, such as Congressional speeches (Gentzkow and Shapiro (2010); Jensen et al. (2012)), but extends it to the multimodal domain of videos. Specifically, I define partisanship as the degree to which a video's content predicts the political affiliation of the sponsoring candidate. This definition preserves the standard interpretation of partisanship as content distinctiveness across party lines, while enabling a unified empirical strategy to compare the informational contribution of images and text. Figure 1 illustrates schematically how the partisanship estimation model is built.

The key innovation I introduce is to treat the raw visual and textual content of each ad as high-dimensional semantic objects, rather than as bundles of selected features. Prior work in Fowler et al. (2021) has constructed partisanship scores of political ads by first extracting specific features from transcripts and frames. This feature set, including both manually coded attributes and outputs from pretrained models, is then then used to estimate predictive models.

Differently from previous approaches, our methodology operates directly on the full semantic representations of text and image content. These representations are obtained from a pretrained multimodal embedding model and mapped into a shared vector space, which enables direct comparison and integration of visual and textual information. By placing text and images in the same representation space, the methodology allows for direct comparison of their predictive power, integration in a pooled specification (considering text and images jointly), and scalable application to relevant domains like TV news—where predefined features may be unavailable or difficult to harmonize across formats.

2.1 Political Ads Data

The political advertising data come from the Wesleyan Media Project (Fowler et al., 2021), which systematically records televised political ads aired on local, national, and cable television across all 210 U.S. media markets. The underlying source is Kantar/CMAG, a commercial firm specializing in monitoring and classifying political advertisements. For each airing, the dataset reports detailed metadata including the television station, market, airing time, and estimated cost. Furthermore, it provides video recordings of each unique advertisement ("creatives"). The Wesleyan Media Project further codes the ads along several qualitative dimensions, such as tone and issue focus.

Our sample comprises the universe of unique political advertisements aired during the 2016, 2018, and 2020 U.S. federal election cycles, spanning Presidential, Senate, and House races. Using sponsor metadata, each ad comes with a partisan label (Republican or Democratic) based on the party affiliation of the sponsoring candidate or political action committee. Ads sponsored by independent or nonpartisan entities are excluded to ensure unambiguous partisan identification.

The resulting dataset consists of 15,427 unique ads, of which 7,731 are labeled as Democratic and 7,696 as Republican. For each ad, I extract the full transcript and sample one video frame every five seconds. To maintain alignment across modalities, I segment the transcript into the same number of textual chunks as there are extracted frames, creating matched image–text pairs. These aligned pairs form the raw inputs to the modality-specific embedding models. The Appendix details the extraction and preprocessing pipeline. The next subsection describes how these inputs are mapped into a shared semantic space using the CLIP multimodal model.

2.2 CLIP Embeddings: A Common Representation for Images and Text

To compare the partisan content of text and images within the same analytical framework, I require a representation in which both modalities can be expressed in the same semantic space. I use the Contrastive Language-Image Pretraining (CLIP) model introduced by Radford et al. (2021).[7] CLIP is a state-of-the-art multimodal embedding model trained on over 400 million image–text pairs, learning to embed images and texts into a shared vector space by maximizing similarity between matched pairs while minimizing similarity between mismatched ones. This training objective produces embeddings that capture rich semantic relationships and generalize well to domains outside the training corpus, making CLIP particularly suited to this application.

CLIP maps each input into a common 512-dimensional embedding space, allowing text and images to be treated symmetrically in downstream analyses. To aggregate information at the advertisement level, I impose a linearity assumption on embeddings, following Bhalla et al. (2024) and the linear representation hypothesis[8]: the embedding of the entire ad can be approximated by the average of its constituent frame or text-snippet embeddings. This parallels the intuition of bag-of-words models in text analysis, where the order of tokens is ignored and the aggregate signal is treated as additive. In Appendix B, I explicitly define the necessary assumptions for my representation.

A challenge in multimodal analysis is the modality gap: image and text embeddings tend to occupy different regions of the shared space (Liang et al. (2022)). Without adjustment, this limits their direct comparability and can bias pooled estimates. I address this in three steps: (i) normalize each ad's embeddings to lie on the unit sphere; (ii) compute the average embedding across all ads for each modality; (iii) subtract the modality-specific global mean from each embedding and re-

[7] More precisely, I use the publicly available CLIP ViT-B/32 model released by OpenAI. ViT-B/32 uses a Vision Transformer with a patch size of 32 pixels as the image encoder, paired with a transformer-based text encoder, both mapping into a 512-dimensional embedding space.

[8] The linear representation hypothesis, in this context, refers to the idea that semantic information in embedding spaces can be meaningfully composed through simple linear operation (e.g. Mikolov et al. (2013)). See also the original CLIP paper by Radford et al. (2021), which uses itself linear aggregation of CLIP embeddings (i.e., averaging across multiple captions) in downstream tasks.

normalize.

For ad A in modality mod $\in \{\text{img}, \text{txt}\}$, with normalized embeddings $\bar{\mathbf{x}}_A^{\text{mod}}$ and global mean $\boldsymbol{\mu}_{\text{mod}}$, the adjusted embedding is

$$\tilde{\mathbf{x}}_A^{\text{mod}} = \frac{\bar{\mathbf{x}}_A^{\text{mod}} - \boldsymbol{\mu}_{\text{mod}}}{\|\bar{\mathbf{x}}_A^{\text{mod}} - \boldsymbol{\mu}_{\text{mod}}\|_2}, \quad \text{mod} \in \{\text{img}, \text{txt}\}.$$

This centering ensures that both modalities share a common reference point, enabling meaningful comparison and integration in subsequent analyses. Intuitively, this adjustment controls systematic differences between how CLIP encodes text and images, so that any remaining variation in the embeddings reflects content rather than the modality itself.

Having mapped each ad's text and images into a shared semantic space, I can use the adjusted CLIP embeddings to predict whether an ad was sponsored by a Republican or Democrat. Separate models for text and image embeddings measure each channel's standalone predictive power, while a pooled specification captures their complementary signals. The following subsection details the classification framework.

2.3 Partisanship Classification Model

Defining the models. Our CLIP embedding strategy represents each political ad at the video level through two aligned feature vectors: one containing information from its images ($\tilde{\mathbf{x}}^{\text{img}}$) and one from its text ($\tilde{\mathbf{x}}^{\text{txt}}$). These representations provide the inputs to estimate the partisan signal contained in each modality. The task is a binary classification problem: predict whether an ad was sponsored by a Republican ($y = 1$) or a Democrat ($y = 0$) based solely on its visual or textual content. I estimate separate models for images and text to assess each modality's specific predictive power, and a pooled specification to evaluate their combined informational content.

I split the dataset into training and test sets (80/20), stratified by party to preserve class balance. For each modality, I train a deep learning classifier on the adjusted CLIP embeddings, where the predicted probability from each model represents the estimated likelihood that an ad is Republican given its image or text content. To then combine information from both modalities, I train a stacked model that takes as input the predicted probabilities from the image and text model, following Ludwig and Mullainathan (2024). This pooled model captures complementary signals while retaining interpretability.[9]

Formally, I consider the stacked data $\mathbf{x} = [\tilde{\mathbf{x}}^{\text{txt}}, \tilde{\mathbf{x}}^{\text{txt}}]$ and I let let $m_{\text{img}}(\tilde{\mathbf{x}}^{\text{img}})$ and $m_{\text{txt}}(\tilde{\mathbf{x}}^{\text{txt}})$ denote

[9]Because the training uses standard machine learning methods, I relegate the rest of the technical discussion to Appendix B.

the predicted probabilities from the image-only and text-only models. The pooled prediction is:

$$m_{\text{p}}(\mathbf{x}) = \sigma\left(\alpha_0 + \alpha_1 m_{\text{img}}(\tilde{\mathbf{x}}^{\text{img}}) + \alpha_2 m_{\text{txt}}(\tilde{\mathbf{x}}^{\text{txt}})\right),$$

where $\sigma(\cdot)$ is the logistic function. The coefficients $(\alpha_0, \alpha_1, \alpha_2)$ are estimated to maximize classification accuracy on the training set.

Models Evaluations. I evaluate predictive performance using the area under the receiver operating characteristic curve (AUC), a standard measure of ability to discriminate between labels. The AUC summarizes how well a model rank-orders Republican and Democratic ads across all possible thresholds: a value of 0.5 corresponds to random guessing, and a value of 1.0 indicates perfect separation. Intuitively, the AUC is the probability that a randomly selected Republican ad receives a higher predicted score than a randomly selected Democratic ad.

To quantify the information contained in each modality, I evaluate out-of-sample performance using ten-fold cross-validation (Figure 2). The image-based model attains an average AUC of 0.911 and the text-based model 0.909. For reference, the pooled benchmark that combines the two modality-specific scores attains an AUC of 0.945. Relative to this benchmark, the AUC-based "signal shares" are respectively $\frac{0.911-0.5}{0.945-0.5} \approx 92.3\%$ for images, and $\frac{0.909-0.5}{0.945-0.5} \approx 91.8\%$ for text.

Following Ludwig and Mullainathan (2024), these results can also be appreciated in R^2 terms. Because each model outputs a single predicted probability for each observation, the observed partisan label can be regressed directly on these scalar predictions. Whereas the AUC measures only the model's ability to rank observations correctly, R^2 s also sensitive to the distance between the predicted probabilities and the actual outcomes. In the pooled specification including both modality-specific predictions, the R^2 is 0.631. Using only the image-based prediction yields $R^2 = 0.516$, and using only the text-based prediction yields $R^2 = 0.507$. Relative to the pooled model, this corresponds to 81.81% of the explained variance being captured by the image prediction and 80.29% by the text prediction.

In sum, regardless of whether accuracy is measured by AUC or R^2, each modality on its own accounts contributes significantly to the overall signal - between 80% and 90%.

Images and text tell different (yet complementary) stories. What does this first look at the partisanship model reveal? First, images clearly matter: either modality on its own captures the majority of the predictive signal in the pooled model, indicating that both visual and textual content contain strong partisan cues. Second, the improvement from combining them points to complementarities between the two channels. When image- and text-based predictions are used jointly, the model achieves higher accuracy than with either input alone, suggesting that each contains dimensions of

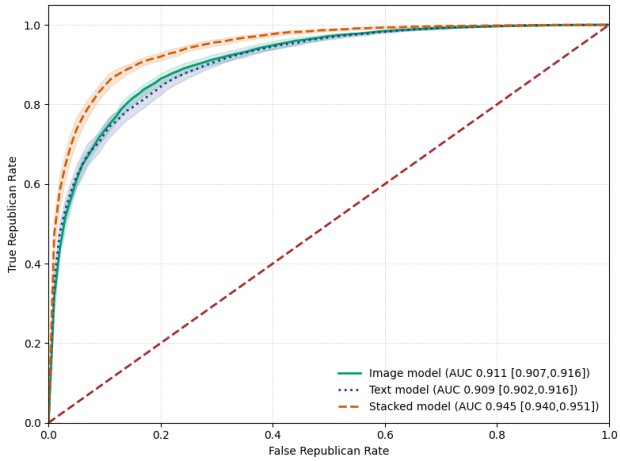


Figure 2: Out-of-Sample Classification Performance (AUC)

Notes: Average AUC from 10-fold cross-validation for each model. The text-only and image-only models achieve AUCs of 0.909 and 0.911, respectively. The stacked model achieves an AUC of 0.945. Shaded areas represent 95% confidence intervals across folds.

partisan presentation not fully reflected in the other.

The high percentages of pooled-model signal explained by each modality indicate substantial overlap in the information they provide. However, overlap in predictive performance does not necessarily imply reliance on the same underlying features. Even if two sets of features were entirely distinct, their predictions could be highly correlated if the components they capture are themselves perfectly correlated in partisan terms. The fact that the pooled model outperforms either modality alone shows that the signals are not perfectly overlapping, but without further analysis it remains unclear which aspects of partisanship are unique to each. This motivates a closer examination of the specific features each modality leverages to produce its predictions.

To examine this further, I compare the partisanship scores generated by the image-only and text-only models. Figure 1 plots the predicted probability that an ad is Republican according to each modality. If both models captured exactly the same underlying signals, the points would lie along the 45-degree line. Instead, the correlation is moderate ($R^2 = 0.443$), indicating meaningful divergence between the modalities.

he analysis so far shows that the model captures partisan distinctions reliably and that text and images encode partly distinct dimensions of partisan presentation. Yet the purpose of building the model is not to explain political advertisements per se, but to develop a scalable measure of partisanship that can be used on the video-based news environment that motivates this study. Political ads provide a labeled and controlled setting in which the partisan structure of visual and textual content can be learned; news videos are where those signals are produced and consumed at scale. The next section applies the ad-trained classifiers to news videos, testing whether the learned partisan fea-

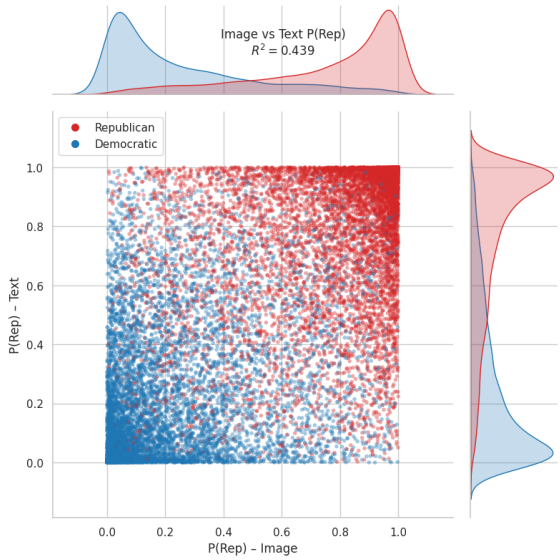


Figure 3: Predicted Republican Probability: Image vs. Text Models

Notes: Each point represents one political ad. Horizontal and vertical axes show the predicted Republican probability from the image and text models, respectively. Points are colored by ground-truth party label.

tures generalize across formats and examining how attention constraints shape their expression in this low-attention medium. This exercise highlights a central advantage of the embedding approach: by representing both modalities in a shared, continuous semantic space, embeddings make it possible to measure partisanship consistently across arbitrarily short segments—a necessary feature for studying how attention shapes partisan communication.

3 From Political Advertising to Video News

The media partisanship literature has largely relied on text-based measures of slant designed to travel across corpora. Following this tradition, I apply the ad-trained multimodal model, without re-estimation, to a new out-of-sample domain: news videos. This exercise provides both a validation of the model's generalizability and a first application to the low-attention, video-based media environment that motivates the broader analysis.

Even in this new environment, I represent each video as two parallel sequences—one of text snippets and one of frames—each mapped into CLIP embeddings in the same semantic space. Because embeddings can be averaged over any subset of elements, any portion of a video—whether a few seconds or the full segment—can be collapsed into a single vector and assigned a partisanship score. This design makes the model applicable to segments of any length and allows separate estimation for images, text, or their combination.

To study how partisan information builds with inattentive exposure, I simulate limited atten-

tion by randomly sampling K short chunks from each video, averaging their embeddings, and re-estimating partisanship at the video level. Small K represents brief or distracted viewing; large K corresponds to sustained attention.

Before analyzing how partisan information accumulates with exposure length, I first describe the news video data and assess how well the ad-trained model transfers to this new domain. This establishes the baseline performance of the model on news content before turning to the dynamics of information flow.

3.1 TV News Data

I assemble a corpus of news videos by scraping the official YouTube channels of major U.S. outlets—like Fox News, MSNBC, and CNN—together with a set of local and digital networks (full list in Appendix Table 3). Using a standardized API query, I request videos labeled as "medium" duration (5–20 minutes) and containing the keyword "immigration" in their metadata fields (title, description, or tags). This query biases the sample toward immigration-related coverage—a domain where partisan divisions are salient—but it does not restrict it exclusively to that topic. Because the YouTube search API relies on channel-supplied metadata rather than transcript content, the resulting corpus includes broader political news segments in which immigration appears alongside other issues. The final dataset comprises 5,156 videos, totaling more than 50,000 minutes of content.

The analysis centers on immigration because it provides a strong test case for multimodal partisanship. Immigration is highly salient to voters, visually rich, and sharply polarized across party lines, making it an ideal setting to evaluate how partisan signals manifest in text and images. The goal, however, remains to measure general partisanship rather than issue-specific slant.[10] The corpus is not limited to immigration alone—videos retrieved through the keyword query often include broader political coverage within the same segment. In Appendix Figure 11, I show that the estimated partisanship of image and text clusters is not confined to the immigration-related subset, confirming that the model captures general partisan orientation rather than topic effects.

An important feature of this application is that television news constitutes a strictly out-of-sample domain: the multimodal classifier is trained exclusively on political advertisements and applied to news videos as is, without any fine-tuning or validation on the new corpus. For each video, I take the complete set of captions and extract one frame every five seconds, pairing each frame with the corresponding text snippet. This ensures that the representation and prediction procedures remain identical to those used in the political ads sample.

[10] As I show in Section 4, the the model assigns slightly higher Republican scores to immigration-related content, being immigration associated to Republican ads in the training data. This bias is unimportant for my purposes, as the analysis focuses on relative differences across channels and on how partisan information is carried by text versus images.

Image preprocessing. Unlike political advertisements—where frames can be processed directly through CLIP—news videos present a more complex visual layout. A single frame often combines multiple layers: the main scene, overlaid banners, scrolling tickers, split screens, and studio backdrops (Appendix Figure 9). It is not obvious ex ante which portion should count as the image for measuring visual content.

To address this, I isolate the main on-screen scene and discard the surrounding broadcast scaffolding. For each representative frame, I detect strong horizontal and vertical edges (via edge detection followed by a Hough transform) and identify rectangular regions. I then retain the largest contiguous scene area—and, if present, a secondary region of comparable size—cropping away lower-thirds, side panels, and other overlays. This procedure preserves the substantive visual content perceived by viewers while removing design elements that carry little informational value. Details are provided in Appendix D.

This preprocessing step ensures that visual representations in the news corpus are defined consistently with those in political ads, allowing the multimodal classifier to operate on a comparable notion of "image content".

Text preprocessing. Extracting transcripts from TV news is more challenging than for political ads, which are shorter, more structured, and generally cleaner. For TV news, I rely on the captions provided directly by YouTube, which offer a time-aligned textual record of the broadcast. I clean these raw captions using the procedures detailed in Appendix D, which include removing artifacts and non-speech tokens, restoring punctuation, and recovering sentence boundaries.

I then split each cleaned transcript into the same number of segments as retained frames, allocating words uniformly across segments. Each segment is embedded with the CLIP text encoder, producing one embedding per segment and a video-level average. This minimal preprocessing is aligned with the one used for political ads, so that the text is clean and ready for the encoder.

3.2 Does Ad Partisanship Transfers into Video News?

The next step tests whether the partisan patterns learned from political ads extend to news videos. While ads and news differ in purpose and style, both communicate ideology through images and language. Applying the ad-trained classifier to news content without re-estimation provides a direct test of transfer: if it still separates partisan sources, the model captures general visual and verbal features of partisanship rather than artifacts of campaign advertising.

For each video v I compute one embedding for images $\tilde{\mathbf{x}}_v^{\text{img}}$, and one embedding for text $\tilde{\mathbf{x}}_v^{\text{txt}}$. Feeding these into the corresponding classifiers yields predicted Republican probabilities at the

P(Rep) = 0.01 P(Rep) = 0.99

Figure 4: Example of image-based partisanship predictions in TV news coverage.

Notes: The figure displays two frames from immigration coverage on different channels, respectively MSNBC (left) and Fox News (right). Predicted probabilities are obtained from the image-only classifier trained on political ads.

video level. That is, I compute video–level predictions:

$$p_{\text{img}}(v) = m_{\text{img}}\big(\tilde{\mathbf{x}}_v^{\text{img}}\big), \qquad p_{\text{txt}}(v) = m_{\text{txt}}\big(\tilde{\mathbf{x}}_v^{\text{txt}}\big), \qquad p_{\text{p}}(v) = m_{\text{p}}\big(\tilde{\mathbf{x}}_v\big),$$

where m_{img}, m_{txt} and m_{p} are the ads classifiers. Comparing these predictions across outlets provides an immediate assessment of whether partisan structure in the ad space transfers to the news domain.

The partisanship model transfers into video news I begin by testing whether the partisan patterns learned from political ads extend to news videos. The classifier transfers cleanly: when applied to this new corpus, both the text- and image-based models produce systematic differences in predicted partisanship across outlets. These differences align with established assessments of channel slant.[11]

Before presenting the average predictions by channel, it is instructive to illustrate the image model's operation at the single frame level. Figure 4 shows two frames from immigration coverage on different networks. The image classifier assigns the first frame, from MSNBC, a predicted Republican probability of 0.01, and the second frame, from Fox News, a probability of 0.99. The first depicts a small group of migrants walking through shallow water, while the second shows a larger group standing beside a border barrier. Although both address the same issue, their visual composition—the setting, the prominence of subjects, and the depicted structures—differs in ways the model has learned to associate with opposing partisan signals. In the cluster analysis (Appendix Figure 10), this immigration-related cluster corresponds to Cluster 1, and exhibits a clear partisan gap on average.[12]

[11]For instance, the classification is consistent with the media bias ratings of independent rankings such as AllSides. See https://www.allsides.com/media-bias.

[12]Although the partisan gap in the immigration scenes cluster is not larger than other clusters, confirming that the estimated partisanship differences are not driven solely by immigration scenes.

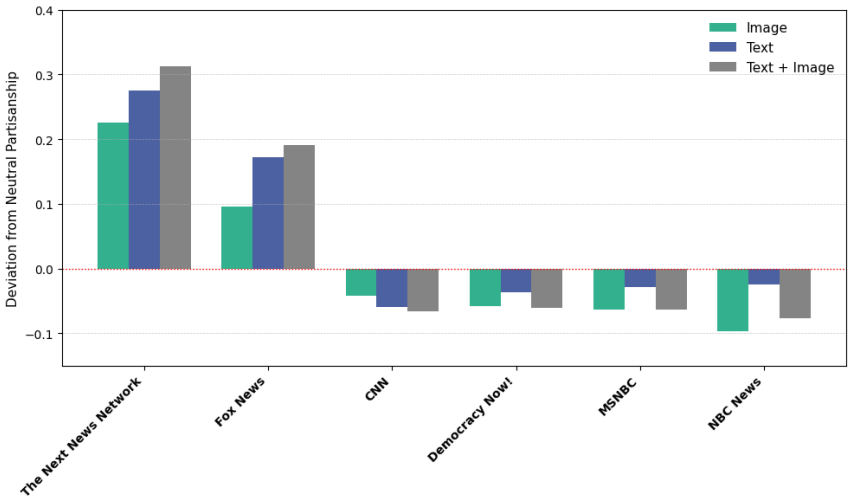


Figure 6: Average predicted partisanship across channels and modalities, under full attention.

Notes: The figure reports the mean predicted Republican probability for each news outlet, centered at the neutral baseline of 0.5. Bars represent predictions from the image-only, text-only, and joint models. Predictions are at the video-level and consider the full length of the video (full attention).

The model's scalability allows predictions at the frame or text-snippet level, but the relevant unit of analysis is the video. For each video in the dataset, I aggregate the corresponding image and text embeddings and generate predicted partisanship scores. Appendix Figure 10 plots the distribution of predicted Republican probabilities for Fox News and MSNBC, separately for the image-only and text-only models. In both modalities, the distributions are clearly shifted in opposite directions: Fox News videos receive higher predicted Republican probabilities, while MSNBC videos receive lower ones. The separation is somewhat larger in the text model, indicating that textual content provides stronger channel discrimination on average. At the same time, the substantial overlap across channels in both modalities shows that neither alone fully captures partisan differences, leaving scope for additional information when the two are combined. To gauge what this additional information contributes, I aggregate the predictions at the channel level and compare the resulting averages.

Figure 6 summarizes the average predicted Republican probabilities across channels, centered at the neutral baseline of 0.5. These averages, shown separately for the image-only, text-only, and joint models, should be read as illustrative rather than absolute: the model is trained on political ads, so only relative distances across outlets are meaningful. The ranking is intuitive—Fox News and right-leaning digital outlets lie well above the neutral baseline, while MSNBC, CNN, and public broadcasters fall below.

Does text underestimates partisanship under full attention? Figure 6 also allows us to quantify, under full attention, how much the text-only model misses relative to the joint text-and-image benchmark. Take the conservative outlet The Next News Network as an example. Its predicted

Republican probability rises from 0.775 when using text alone to 0.811 when both modalities are combined—a difference of 0.037. This difference measures the prediction error that arises when visual information is omitted. Turning to the partisan gap, consider Fox News and MSNBC. The corresponding predictions are 0.672 versus 0.691 for Fox News, and 0.472 versus 0.437 for MSNBC. In other words, the text-only model predicts a partisan gap of 0.200, while the joint model corrects it to 0.254. Text alone therefore understates the Fox–MSNBC gap by 5.5 percentage points—roughly 20% of the total separation observed in the multimodal benchmark.

A complementary way to assess how much partisan information is lost when focusing on a single modality is to compare model performance using the area under the ROC curve (AUC). Appendix Figure 13 plots the ROC curves for the image-only, text-only, and joint models, treating channel identity as a proxy for partisanship—labeling Fox News as Republican-leaning and MSNBC as Democratic-leaning. The image model achieves an AUC of 0.690, the text model 0.768, and the pooled model 0.795. All three perform well above chance, confirming that both modalities contain meaningful partisan signal. Text remains the stronger predictor when full information is available, but the higher AUC of the pooled model shows that images and text encode complementary cues. Omitting one therefore leads to a systematic understatement of partisan slant, even under full attention.

In summary, under full attention, the text-only model outperforms the image model and loses little relative to the multimodal benchmark. But this comparison gives text its strongest possible advantage: complete information processing. In practice, viewers rarely consume ten or twenty uninterrupted minutes of news. As discussed earlier, news is typically encountered in short clips, often while scrolling or multitasking, and attention to language can fade long before a segment ends. When attention is constrained, the picture changes sharply. For example, restricting the analysis to about ten seconds of exposure causes the partisan gap between Fox News and MSNBC in the text-only model to collapse from roughly 0.20 to just 0.04. With so little information, text becomes almost uninformative—short clips contain too few linguistic cues to convey partisanship reliably. This observation motivates the central result of the paper, which I discuss in the next section: in the world of inattention, the relationship between text and image reverses, and visuals become the dominant and more effective carrier of partisan meaning.

3.3 Video Partisanship under Limited Attention, or "An Image is Worth 40.38 Words"

So far I have given text a fundamental advantage: several uninterrupted minutes to accumulate cues, assuming the viewer listens carefully throughout. Real viewing is rarely so focused. On television, topics shift, field reports interrupt, and scenes change; on social media, exposure often comes in

short clips. In these settings, the competitive margin must be how much partisan information each modality can convey within brief windows of content.

Strategy to capture the time dimension. I can analyze how quickly each modality accumulates partisan information thanks to the embedding representation strategy. The task would be infeasible with raw text or video data. Each video is represented as a sequence of five-second "chunks", where each chunk consists of one frame and the corresponding transcript segment (about fifteen words).[13]

The five-second window is a conservative choice for images: visual scenes in news videos typically change roughly every 2–5 seconds, (Lang et al., 2000; Grabe et al., 2003), so equating one image to five seconds of exposure gives text proportionally more information per chunk. In this sense, the comparison slightly penalizes the image modality, making any observed advantage in visual speed a lower-bound estimate of its true informational efficiency.

For a fixed K, I repeatedly draw K chunks for each video at random, average the embeddings within modality, and feed these aggregated representations to the partisanship classifiers. This sampling is done many times per video to smooth out idiosyncratic variation in individual draws, producing a distribution of predicted partisanship scores for each modality at that exposure length. Performance is evaluated by comparing predictions to the "true" channel affiliation using the same AUC metric as in Section 3.2. To trace how quickly accuracy rises, I vary K from 1 to 30 chunks—up to 150 seconds of material.

Larger K approximates attentive, sustained viewing, while $K = 1$ the information extractable from a single five-second exposure. This design quantifies the speed at which image- and text-based signals converge to their long-run performance, isolating the visual channel's advantage in conveying partisan meaning under limited attention. In other words, the random chunk aggregation can be interpreted as a stylized model of inattentive viewing: small K mimics fragmented exposure, large K approximates full engagement. Comparing the rate of convergence across modalities thus provides a direct measure of their efficiency in transmitting partisan cues when attention is scarce.

Images dominate for short segments, text dominates for longer ones. Figure 7 plots the AUC of the image, text, and joint classifiers as a function of the number of seconds of content pooled per video. With only five seconds of material, the stacked model's AUC is almost identical to that of the image-only classifier, implying that at very short exposures almost all partisan information comes from images. As duration increases, the text model begins to gradually close the gap. At the 20-second mark, the image model has already captured about 95 percent of the partisan information it will ever extract, whereas the text model has reached only about 75 percent of its eventual accuracy.

[13]Because there is one frame per chunk, the average number of words in five seconds can be approximated by the ratio of total words to total frames (see Table 3). In the sample, this average is approximately 15.

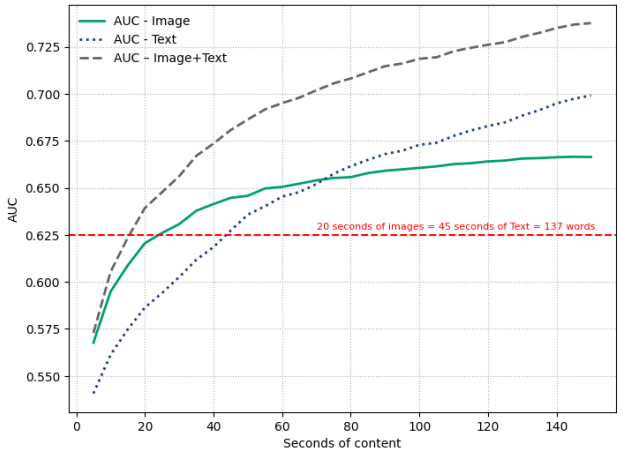


Figure 7: AUC as a function of the number of seconds of content pooled per video

Notes: In short segments, images dominate: at 5 seconds, the image model outperforms text by over four AUC points. By 20 seconds, images capture 95% of their eventual peak accuracy, compared to only 80% for text, and remain ahead until about 90 seconds.

In other words, images deliver almost their full contribution within the first few seconds, while text accumulates signal more gradually and continues to improve well beyond the point where images have plateaued.

The shapes of the accumulation curves explain this pattern. Image accuracy rises steeply at first and then flattens after roughly 100 seconds. Part of this plateau reflects an intrinsic limit on how much partisan information visuals can convey relative to text. A mechanical factor amplifies this: television news relies heavily on repeated visuals, so as viewing time increases, many of the "new" frames are near-duplicates of earlier ones, reducing the marginal contribution of additional visual data.

Text, by contrast, evolves continually: anchors and reporters introduce new sentences and topics, sustaining information growth over time. As a result, textual accuracy continues to climb, eventually overtaking images. When that happens, the stacked model's advantage over text reflects the persistent, complementary signal in images even after they have plateaued. In other words, images keep contributing significantly on top of the text signal, even as length increases.

The red dashed line provides a concrete measure of this speed advantage. The AUC from 20 seconds of images matches that from 45 seconds of text (about 137 words). In short or inattentive viewing windows, images alone deliver the same classification power as more than twice as much textual content, while combining both yields the highest accuracy at all durations.

An image is worth 40.38 words. I can build on the conversion-rate idea to ask a broader question: at any given content length, how valuable are images relative to text? The idea is to map the image AUC curve onto the text curve, finding for each number of image chunks the amount of text that

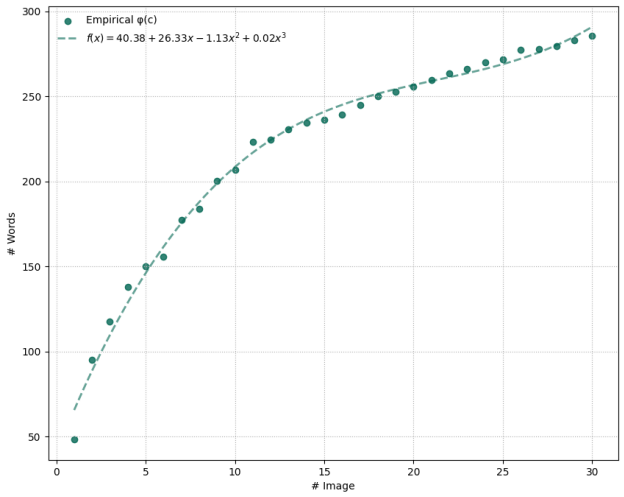


Figure 8: Estimated words equivalence of images.

Notes: The figure plots the empirical mapping between the number of images and their equivalent transcript length in words, based on matching the AUC of the image model to that of the text model. The dashed line shows a polynomial fit to the empirical points; the intercept, corresponding to one image, is 40.38 words.

delivers the same classification accuracy. This mapping is expressed in words per image, using the average words per 5-second text chunk in my data. The result is a function $\phi(K)$ that captures, for any K, the textual equivalent of K images in terms of predictive power.

To present this relationship clearly, I fit a smooth polynomial to the empirical $\phi(K)$ curve (Figure 8). The intercept of this polynomial—my estimate of $\phi(1)$— is 40.38 words. This is the constant that gives the paper its title: in my model, a single 5-second visual glimpse contains as much partisan signal as roughly forty words of transcript. The curve is concave, with diminishing marginal value for additional images, consistent with the repetition dynamics discussed above.

Are images just "quick text"? The evidence so far shows that when attention is scarce, images dominate. They convey partisan meaning almost immediately, while text requires sustained exposure to reach comparable accuracy. This inversion is central: in a media environment built on short clips and divided attention, partisanship is transmitted primarily through visuals. But this result also raises a deeper question: what kind of information are images conveying so effectively?

The difference in partisan signal between text and images may not be merely a matter of speed. Images may encode partisanship through different forms of information: not just faster language, but distinct conceptual or semantic domains. In the next section, I decompose the representation of images and text in video news into combinations of topics and emotions. I focus on emotions as a central component of both persuasion and audience engagement in the modern low attention media environment. I study whether the partisan signal in images arises from emotional features relatively more than text. As a result, I make explicit what the models rely on to predict partisanship.

4 Concept Decomposition of Image and Text Partisanship

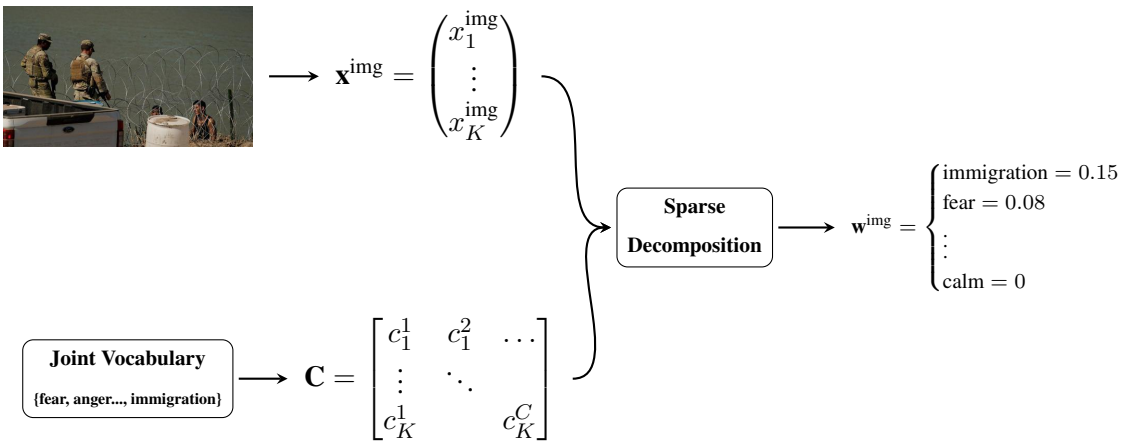


Figure 9: Semantic Splice Decomposition of a Video Frame

The attention-modality relationship outlined in Section 3 raises a natural question: are images just faster text? Or do they rely on different concepts in their partisan signaling? This section opens the black box of the image and text classifiers and shows that the two modalities exploit distinct, interpretable cues. Given my discussion of emotions as a driver of media engagement and inspired by the social psychology literature on the emotional load of images (see e.g. Iyer and Oldmeadow (2006)), I test in particular whether the image model is significantly more responsive to the emotions embedded in the videos.

The analysis proceeds in three steps. First, I construct a joint concept vocabulary combining both emotion and topic dimensions. Emotion concepts follow Cowen and Keltner (2017), and topic concepts are derived from the political ad metadata. I then map both image and text embeddings onto this shared vocabulary using the sparse concept embedding method of Bhalla et al. (2024). Figure 9 illustrates the decomposition for a single frame, but the decomposition in this analysis is performed at the video level. This representation expresses each modality as a weighted combination of emotion and topic concepts, where I use topic as a benchmark for comparison. Second, I use this joint mapping to examine how strongly each modality loads on emotional versus topic dimensions. Third, I conduct a counterfactual exercise that removes one concept at a time from the representation to quantify how much the model's predictions depend on that concept.

I use this idea of removing concepts to build counterfactual representations of images and text in videos to quantify how individual concepts shape partisan predictions. I compute a *Concept Treatment Effect* (CTE) for each element of the joint vocabulary. For every video, I compare the model's predicted Republican probability using the full embedding with the prediction obtained after removing a given concept from the representation. The difference between the two captures

the marginal contribution of that concept to the model's output. Averaging these differences across all videos yields the CTE for that concept and modality. A positive CTE indicates that the presence of the concept increases the predicted probability of Republican partisanship, while a negative one indicates the opposite.

4.1 Sparse Concepts Decomposition

Concept Representation. As previously discussed, the CLIP embedding approach represents the visual and textual components of each ad in a shared 512-dimensional vector space. To reduce dimensionality and facilitate interpretation, I apply the SpLICE procedure (Bhalla et al., 2024), which projects these vectors, subject to a sparsity constraint, onto a "concept vocabulary" of natural language embeddings. Intuitively, once each concept is expressed as a vector in the CLIP space, the goal is to represent the image and text embeddings of each ad as a sparse linear combination of these concept vectors. This representation captures the extent to which each concept is present in the ad, separately for the visual and textual modalities.

Formally, let the input embedding for a given modality (text or image) be $\mathbf{x} \in \mathbb{R}^{512}$. I can define a concept vocabulary $\mathbf{C} \in \mathbb{R}^{512 \times C}$, where C is the number of concepts in the vocabulary. In the main application, this vocabulary consists of 160 topics or emotions in the vocabulary. For example, the word "fear" – represented by its mean-centered CLIP representation – is an element of $\mathbf{C}$. Let $\sigma(x)$ denote the mean-centering operation described in Section 2.2. Given a list of concept terms $x^{con} =$ ["fear", "anger", "immigration"...], the concept dictionary $\mathbf{C}$ is:

$$\mathbf{C} = [\sigma(x_1^{con}), \ldots, \sigma(x_C^{con})] \,.$$

Centering the vocabulary embeddings ensures that all concepts are expressed relative to a common origin in the representation space, making their coefficients directly comparable.

Sparse decomposition. The preceding setup allows the image or text embedding of each ad to be expressed as a weighted combination of a subset of concept vectors in $\mathbf{C}$. The weights quantify the presence of each concept in the ad, separately for each modality, and form the basis for the subsequent analysis.

In order to find these weights, I compute the non-negative vector $\mathbf{w}$ that minimizes the squared distance between the original embedding $\mathbf{x}$ and its linear combination of concept vectors, subject to a λ_1 sparsity penalty:

$$\mathbf{w}^* = \arg\min_{\mathbf{w} \geq 0} \left\| \mathbf{C}\mathbf{w} - \mathbf{x} \right\|_2^2 + 2\lambda_1 \|\mathbf{w}\|_1. \tag{1}$$

This procedure is applied to both the image and text embeddings for every ad in the dataset. I

can interpret an element $\mathbf{w}_c^*$ as a measure of the prominence of concept c in embeddings $\mathbf{x}$. The embedding can then be reconstructed as as:

$$\hat{\mathbf{x}} = \mathbf{C}\mathbf{w}^*.$$

This representation enables the construction of counterfactual embeddings by setting the weight on a specific concept c to zero ("removing" the concept from the representation) and feeding the modified embedding into the partisanship model. Comparing predictions before and after this removal yields the concept-specific average treatment effect (CATE from now on).

Emotion Vocabulary The "emotion space" I use for representation is spanned by the 27 affective buckets introduced by Cowen and Keltner (2017). For each category (e.g. *anger*, *calm*) I select a small set of canonical tokens and embed them with the usual embedding strategy. Stacking these embeddings column-wise yields a 512×79 matrix C.

Topic Vocabulary The topic vocabulary is constructed starting from the issue metadata of every political ad, available directly from the Wesleyan Media Project ads dataset. I select all those issues that appear in at least 1% of the ads, and use those issues as the vocabulary. This procedure yields 81 topics. The details of the creation of the topic vocabulary and its elements are detailed in the Appendix.

4.2 Intermezzo: Emotion Load in Images and Text

Before estimating concept-level effects, I ensure that image and text decompositions are comparable. The sparse decomposition uses an λ_1 penalty to control the number of concepts with nonzero weight that I use for tuning. Because the two modalities differ in scale and structure, a common sparsity penalty λ_1 may yield incomparable representations. I therefore tune λ_1 separately by modality to equalize reconstruction quality: for each modality, I tweak λ_1 until the reconstruction similarity (cosine similarity between the original embedding and its sparse reconstruction) plateaus at approximately 0.34.[14] This choice equalizes approximation accuracy across modalities and prevents downstream differences in Concept Treatment Effects (CTEs) from being artifacts of looser or tighter reconstructions in one channel.

At this matched reconstruction similarity, images respond more strongly to concepts. The distribution of active weights (nonzero concept coefficients) differs markedly across modalities: images activate about 35 concepts on average, while text activates about 18 (see the CDF in Appendix

[14]The sparsity coefficients associated to these results are $\lambda_1 = 0.05$ for images and $\lambda_1 = 0.15$ for text.

Figure 3). This wider support in images reflects conceptual dispersion in visuals rather than a tuning artifact. To verify this, I repeat the entire analysis tuning sparsity fixing the average number of active concepts across modalities, instead of matching reconstruction similarity; the qualitative implications are unchanged. The cross-modal differences reported below are not driven by sparsity calibration. They reflect substantive differences in how images and text load on the joint emotion-topic vocabulary and how those loads influence partisan predictions.

4.3 Representation of Emotions and Topics across Modalities

To illustrate how the decomposition operates in practice, Appendix Figure 7 plots the top twenty concept groups with the largest partisan differences in average weight, separately for image and text representations of YouTube news videos. Each bar shows the normalized contribution of a concept group-either emotion or topic-to the overall representation of Fox News and MSNBC videos. It is worth noticing how much weight the text decomposition loads on the "immigration" topic.

I now turn to two exercises illustrating how emotions shape image representations. The first tests whether images load more heavily on emotional concepts than text, relative to topics. The second removes individual concepts from the representation to show how much each contributes to the overall meaning.

Do images represent more emotions? Having validated the decomposition, I first examine how strongly each modality loads on emotional relative to topical concepts. For each video, I can write its reconstructed embedding as the weighted sum of emotion and topic components,

$$\hat{x} = C_T w_T + C_E w_E,$$

where C_T and C_E denote the topic and emotion vocabularies, and w_T and w_E are the corresponding nonzero concept weights.

I use this representation to compute the *emotion share*, defined as the ratio of the absolute weight on emotion concepts to the total absolute weight across both domains. Intuitively, this measure captures how much of a representation's conceptual mass is devoted to emotions rather than issues. In the illustrative example from Figure 9, the emotion share equals $0.16/(0.27 + 0.16) = 0.37$. This exercise provides a simple but informative summary of whether images systematically embed a larger emotional component than text.

Figure 10 compares the distribution of emotion shares in image and text representations. The image distribution (green) is shifted markedly to the right of the text distribution (blue), indicating that visual embeddings allocate a larger fraction of their conceptual mass to emotional content. The median emotion share for images is 0.45, compared to 0.31 for text, and the entire right tail of the

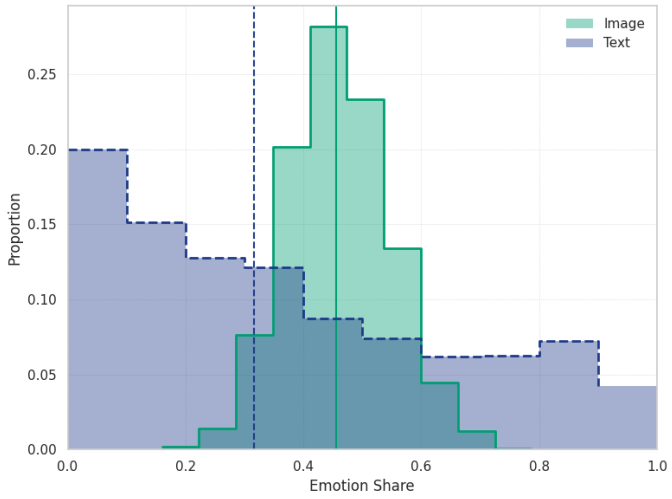


Figure 10: Distribution of emotion shares in image and text representations.

Notes: The figure plots the distribution of emotion shares in news videos, defined as the ratio of absolute weight on emotion concepts to total absolute weight across all concepts in the sparse decomposition. The average emotion share is 0.45 for images and 0.31 for text. Higher values indicate a greater relative presence of emotional concepts in the representation.

image distribution remains thicker throughout. In other words, visuals are not only slightly more emotional on average-they are consistently more likely to encode emotion across the board. This difference highlights the richer affective load of images, even when both modalities are projected onto the same joint emotion–topic space.

In other words, not only is the emotion share higher for images, but emotional concepts also appear to play a more structural role in how visual representations are organized. I next remove individual concepts from the embeddings and measure how much the model's predictions change. This concept removal exercise strengthen the argument that emotions are more relatively more represented in the visual component of news videos.

What happens when we remove a concept? I next examine how much individual concepts contribute to each modality's representation. The idea is straightforward: if we remove one concept-say *fear*-and reconstruct the embedding without it, how much does the representation change? Formally, for each video I compute a counterfactual embedding

$$\hat{x}_{-f} = C_T w_T + C_E w_{E,-f},$$

where the weight on concept f is set to 0. Comparing $\hat{x}$ and $\hat{x}_{-f}$ reveals how strongly that concept shapes the representation. Larger deviations imply greater conceptual dependence. This exercise provides a direct measure of how integral each emotion or topic is to the structure of image and text

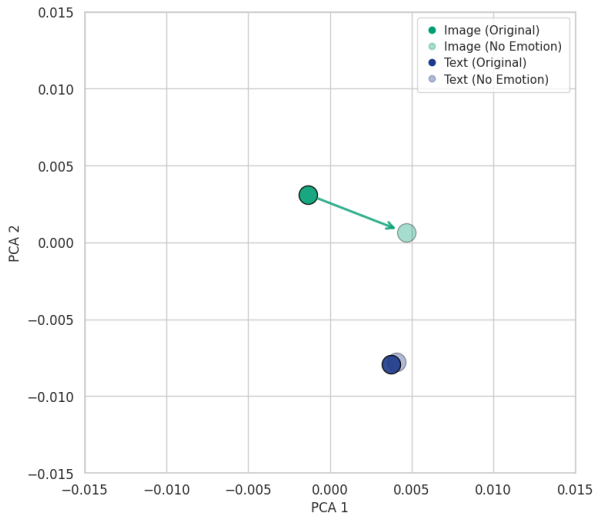


Figure 11: Shift of representative video image and text embeddings after removing the concept *fear*.

Notes: Each dot represents the average embedding across all videos for the image or text modality, projected onto the first two principal components. Arrows indicate the change in the average embedding when the weight on *fear* is set to zero. The average image embedding shifts visibly, while the text embedding remains almost unchanged, showing that visual representations are more sensitive to the removal of the *fear* concept.

embeddings.

I illustrate this idea with a simple example. Suppose we are interested in what happens to the visual or textual content of a video when we remove the concept of *fear*. For each video, I set the weight on *fear* to zero, reconstruct the embedding, and then average these counterfactual embeddings across videos within each modality. The same averaging is done for the original embeddings. I then project both averages into the two-dimensional PCA space computed from the full set of video embeddings. Appendix Figure 8 displays the overall PCA structure for all videos.[15]

Figure 11 shows the shift between the original and no-*fear* averages. When the weight on *fear* is removed, the average image embedding moves visibly, while the text embedding remains almost unchanged. The shift is evident in the PCA projection, showing that even the removal of a single emotional concept significantly alters visual representations. In contrast, the text embedding barely shifts, indicating that emotional concepts contribute little to the structure of textual representations.

To generalize this result, I repeat the exercise for every concept-emotions and topics-without PCA reduction. For each video, I compute the cosine similarity between the original and counterfactual embeddings and define the *representation shift* as $1 - \cos(\hat{x}, \hat{x}_{-f})$, where cos is cosine similarity. Averaging these shifts by concept yields a consistent pattern. Removing a single emotion concept shifts image representations by 0.014 on average (a rotation of about $9.6°$ in embedding space), compared to 0.006 for text (about $6.3°$). Images are also slightly more responsive to topic

[15]No visible clustering across modalities is expected or desired given the mean-centering of embeddings.

removals-13° versus 12°-but the gap is much larger for emotions. Overall, images depend more heavily on emotional cues, while text remains primarily anchored in topical structure.

The counterfactual exercise shows that removing a single concept-like *fear*-can change how images are represented. But does it also change how partisan the model thinks the video is? To test this, I take the original embedding $\hat{x}$ and its counterfactual version without a concept $\hat{x}_{-f}$, feed both into the partisanship model, and compare the predicted probabilities. The difference, $m(\hat{x}) - m(\hat{x}_{-f})$, measures how much that concept contributes to the model's partisan classification. This is the intuition behind the Concept Treatment Effect (CTE).

4.4 Concept-Level Causal Analysis (CTE)

Defining CTE. Each video representation can be expressed as a sparse combination of interpretable concepts. This allows counterfactual interventions that remove one concept at a time and quantify its contribution to the model's partisan prediction. Formally, let the representation of video i in modality mod $\in \{\text{img}, \text{txt}\}$ be:

$$\hat{\mathbf{x}}_i^{\text{mod}} = \mathbf{C}\mathbf{w}_i^*,$$

where $\mathbf{C}$ is the concept vocabulary matrix and $\mathbf{w}_i^*$ the estimated nonnegative concept weights. The counterfactual representation with concept c removed is obtained by setting its weight to zero:

$$\mathbf{w}_{i,-c}^*(k) = \begin{cases} \mathbf{w}_i^*(k), & k \neq c, \\ 0, & k = c, \end{cases} \qquad \hat{\mathbf{x}}_{i,-c}^{\text{mod}} = \mathbf{C}\mathbf{w}_{i,-c}^*.$$

This preserves the influence of all other concepts while excluding c.[16]

Let $m^{\text{mod}} : \mathbb{R}^d \to [0,1]$ be the trained partisanship model returning the probability that a video is Republican. The Concept Treatment Effect (CTE) of concept c in modality mod is defined as:

$$\Delta_c^{\text{mod}} = \frac{1}{|I_c|} \sum_{i \in I_c} \Big[m^{\text{mod}}(\hat{\mathbf{x}}_i) - m^{\text{mod}}(\hat{\mathbf{x}}_{i,-c}) \Big], \qquad I_c = \{i : w_{i,c}^* > 0\}.$$

A positive Δ_c^{mod} indicates that the presence of c raises the predicted Republican probability, holding all other concepts fixed; a negative value indicates the opposite. The magnitude of Δ_c^{mod} captures how much partisan meaning the model attributes to that concept.

Concepts in $\mathbf{C}$ are not statistically orthogonal, so removing one may indirectly affect correlated dimensions. The exercise therefore does not identify a structural causal parameter but measures the model's internal attribution: how much the predicted partisanship changes when a specific

[16] Because weights are not normalized to sum to one, removing a concept alters the total magnitude of $\mathbf{w}_i$. Appendix results show that estimated effects are not mechanically proportional to $w_{i,c}$ or to changes in $|\mathbf{w}_i|$.

semantic element is removed. This provides an interpretable decomposition of partisan meaning at the concept level within each modality.

Because image and text representations differ in both the number of activated concepts and their relative weights, using modality-specific representations for each model would conflate model properties with differences in conceptual coverage. To ensure comparability, I compute Δ_c^{mod} using the image representation and its counterfactual version for both the image and text models. Results are unchanged if the text representation is used instead. This equivalence arises because the CTE responds to perturbations within the shared embedding space: it reflects characteristics of the models, not of the underlying representation.

Results: Model Sensitivity to Emotions and Topics. To compare concepts at a meaningful level, I group them into broader semantic categories-for example, combining "anger'' and "disgust'' into a single emotion bucket, or "tax'' and "budget'' into a policy bucket. For each bucket, I then average the estimated change in predicted Republican probability separately for the image and text models. The resulting pairs $\left(\bar{\Delta}_b^{\text{img}}, \bar{\Delta}_b^{\text{txt}}\right)$ summarize, for each emotion or topic b, how strongly it shifts partisan predictions in each modality.

Figure 12 plots these aggregated estimates for emotions (left) and topics (right), with the regression line reporting the cross-modal correlation r. Each point corresponds to a concept bucket, colored by semantic category. This procedure yields a direct comparison between how the same underlying concepts influence partisan predictions across modalities, holding constant the shared embedding space.

Emotional concepts generate larger effects in images than in text: the average absolute CTE is 0.021 for images and 0.016 for text, with a cross-modal correlation of $r = -0.01$. These effects are thus largely uncorrelated across modalities, indicating that emotions influence partisan classification through distinct channels in the two models. By contrast, topic-level CTEs are highly correlated ($r = 0.76$) and somewhat larger in text (0.033 vs. 0.025). Together, these patterns show that the visual and textual channels differ not only in speed but in substance: images convey partisanship primarily through emotional cues, while text does so through issue content.

From Intrinsic Properties to Effects on Viewers Two distinctive properties emerge from the analysis so far. First, images convey partisan information more rapidly and remain predictive under limited attention, whereas text requires longer exposure to accumulate accuracy. This suggests that visual signals are more resilient to inattention and dominate when cognitive resources are scarce. Second, the concept-level analysis shows that image partisanship is primarily driven by emotional cues, while text relies on issue content. Visual and verbal channels therefore differ not only in speed but also in the kind of information they encode.

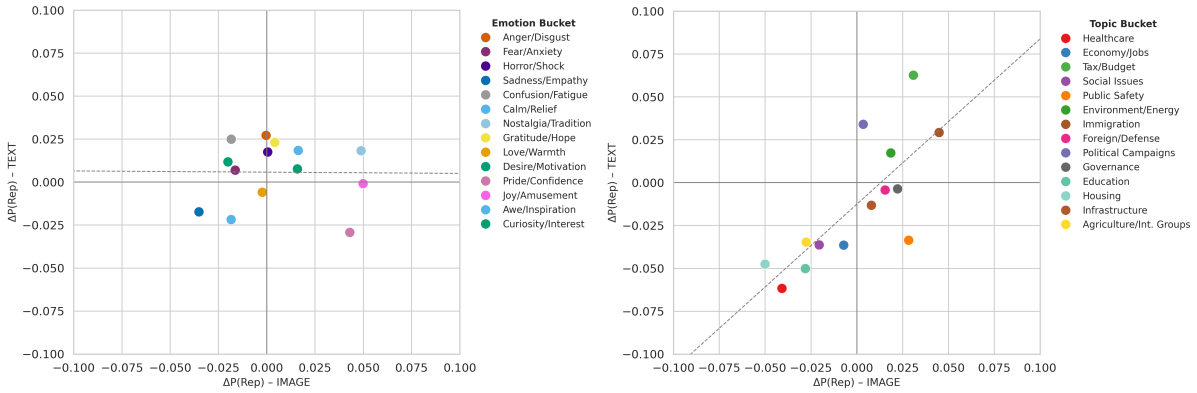


Figure 12: Cross-modal Concept Treatment Effects in Video News

Notes: Each point represents an emotion (left) or topic (right) bucket, summarizing the average change in predicted Republican probability when that concept is removed from the representation. The x–axis reports image-based effects; the y–axis reports text-based effects. Dashed lines show cross-modal regressions; r denotes the correlation across buckets. Points are colored by semantic category.

Images are therefore faster and more emotional; text is slower and more topic-based. These differences raise a natural question: do they matter for how viewers respond to partisan content? If visuals transmit faster and more affective signals, we should expect them to shape emotions even under brief exposure, whereas text may influence beliefs and attitudes only with sustained attention. The next section tests these implications through a survey experiment that independently varies the partisan direction and duration of visual and textual exposure.

5 Survey Experiment

The preceding analysis shows that images and text differ along two dimensions. Images convey partisan information more rapidly and remain accurate under limited attention. Furthermore, emotions as concept are more represented in images, which drive their partisan signal. On the other hand, text partisanship is relatively more topic-based. These differences imply distinct mechanisms of influence: visuals should affect perception and emotion quickly, even under short exposure, whereas text should influence attitudes only when viewers have time to process it.

I run a preregistered survey experiment testing these predictions by exposing respondents to video news segments that vary independently in the partisanship of visuals and transcripts, and in the duration of exposure. The design thus isolates how each modality affects perception, emotion, and persuasion across short and long viewing conditions. Three results summarize the findings. First, images dominate percieved partisanship. Republican visuals increase the perceived Republican leaning of the video, and this effect remains stable across exposure lengths. Textual cues have little effect on perceived partisanship: they are recognized when evaluated in isolation but are

largely crowded out by visuals in integrated assessments.

Second, images shape emotions. Republican visuals elicit anger, fear, and disgust, and reduce empathy toward immigrants, relative to Democratic visuals. This is consistent with the model's finding that visual partisanship is transmitted through affective cues. These emotional responses arise immediately and persist regardless of clip length.

Third, modalities differ in persuasive reach. Text influences policy attitudes only under sustained exposure. In long clips, Republican transcripts shift immigration attitudes toward more restrictive positions—particularly among Democrats—but have no detectable effect in short clips. Images, by contrast, do not alter policy attitudes but affect behavior: in short clips, Republican visuals reduce the probability of donating to a pro-immigrant charity, especially among Republican respondents.

Together, these results reveal a sharp asymmetry between modalities. Text, tied to topics and propositional content, alters beliefs under sustained attention. Images, tied to emotions and immediacy, shape perceptions, emotions, and short-run behavior. The two modalities therefore operate on distinct channels—cognitive versus affective—and under different time horizons (or attention conditions).

The experimental design builds directly on the partisanship model. Two broadcasts covering the same immigration event are selected, one from Fox News and one from MSNBC. From these, I construct four videos by independently assigning the partisanship of the visuals and of the text to be either Republican or Democratic. Each version is produced in both short (30-second) and long (120-second) form, yielding eight treatment arms in a $2 \times (2 \times 2)$ design. The next sections detail the construction of these treatments, randomization, and outcome measures.

5.1 Experimental Protocol

Event selection: April 3, 2024 Texas Senate Bill 4 The selection of the event covered in the experimental broadcasts follows the procedure detailed in the Appendix. Briefly, I restricted attention to MSNBC and Fox News segments and identified pairs of transcript snippets that: (i) contained immigration-related terms; (ii) lay at opposite ends of the text-based partisanship distribution; and (iii) were similar in their embedding representation. From this candidate set, I selected the event and specific broadcast pair through manual inspection. The remainder of this section describes the event and the broader context of the chosen broadcasts.

The event selected is the April 3, 2024 appellate hearing on Texas Senate Bill 4 (SB4). SB4 was a contested immigration law that authorizes state authorities to arrest and deport individuals suspected of crossing the U.S.–Mexico border illegally, a power traditionally reserved for federal agencies. The April 3 hearing before the Fifth Circuit Court of Appeals followed a lower-court injunction blocking the law's implementation, and drew national media attention due to its impli-

cations for federal state authority over immigration enforcement. For the experiment, I selected two broadcasts covering the hearing that are similar in length and semantic content but lie at opposite ends of the partisanship distribution for both text and images (see Appendix Figure 15). The videos were drawn from Fox News' America's Newsroom 11 a.m. program and MSNBC's José Díaz-Balart Reports 8 a.m. program, both aired on April 3, 2024.

The two videos share a very similar structure and have two distinct components. In the first, the host announces the issue and introduces the topic. In the second, a reporter provides a more detailed description and explanation of the issue. The first part lasts about 30 seconds, and the second about 90 seconds. This structure allows me to create one short treatment arm, including only the first 30 seconds, and one long treatment arm, including the whole clip.

Treatments. From each broadcast, I extract and clean the full transcript, removing elements that could directly identify the speaker (e.g., the reporter's name). I then process it through an AI voice generator (ElevenLabs) to produce realistic voiceovers that closely match the style and tone of the original journalists. This approach serves two purposes. First, it standardizes delivery across treatments, ensuring that differences in perceived partisanship are not driven by variation in vocal characteristics. Second, it allows me to precisely control timing so that all audio tracks have the same length. The same voices are used consistently across all lines of narration.

For the visual treatment, I include exclusively images depicting immigration scenes. This removes visual cues that could identify the channel and mirrors the preprocessing approach used in the TV News analysis. To align the visual and audio tracks, I retain the original immigration-related footage from the selected segment and then supplement it with additional imagery drawn from another broadcast on the same topic from the same network—specifically, America's Newsroom 8 a.m. for Fox News and Ana Cabrera Reports 7 a.m. for MSNBC. This augmentation ensures that the visual content remains authentic to each network's coverage style while providing enough material to fully match the duration of the redubbed audio.[17]

Using these materials, I construct the experimental treatments by independently varying the partisanship of the audio (text) and the visuals (image). For each length condition—short (approximately 30 seconds) and long (approximately two minutes)—I create four versions: Republican visuals with Republican text, Republican visuals with Democratic text, Democratic visuals with Republican text, and Democratic visuals with Democratic text. Crossing the four video types with the two lengths yields a total of eight treatment arms. All other features of the videos—including a short, self-made three-second intro clip—are held constant, ensuring that any differences in outcomes can be attributed solely to the partisan orientation of the text and images.

[17] I provide the full transcripts used, as well as the frame sequence of the videos, in Appendix C.

Experimental procedure and measurement. I recruit $\sim$ 4000 participants (3430 after sample restrictions) on Prolific, randomly assigned to each of 8 treatment groups. The sample is split 50-50 across Republican and Democratic responders, by design. Prior to treatment, respondents complete a pre-treatment questionnaire that records demographics, baseline political attitudes, and media consumption habits. After viewing the assigned video, participants answer a set of post-treatment questions designed to capture four dimensions of response: (i) immigration policy preferences, (ii) emotional reactions, (iii) perceived partisanship of the content, and (iv) behavioral choice over donating to an immigration charity.

To assess randomization balance, I regress each baseline covariate on the full set of treatment dummies and report the mean, standard deviation, and joint F-test for equality of means across treatment arms. The p-values in Appendix Table 4 represent to these joint tests. Out of 26 covariates, 2 show differences significant at the 5% level. Overall, the treatment groups are well balanced across demographic and media-consumption characteristics.

The first outcome, immigration policy preferences, measures persuasion at the attitudinal level. I construct an anti-immigration index as the average of items covering key policy domains—deportation, border patrol expansion and restrictions on regularization.[18] The items are coded on a 5-point scale where higher values indicate more restrictive views. The index and all other outcomes I present are standardized to enable comparison of effect sizes across measures.

The second outcome captures emotion responses. Participants self-report the intensity of emotions such as anger, fear, disgust, sympathy, sadness, indifference, and sadness, as well as sympathy on a 7-point scale. These measures provide a direct test of the model's prediction that images load heavily on emotions.

To test how actual viewers interpret the partisan signal coming from images and text, I measure perceived partisanship in two complementary ways. Respondents are first asked to classify the whole video on a global left–right scale (a 5-point Likert scale from Strongly Democratic to Strongly Republican) and to rate separately the perceived partisan leaning of the visuals and of the transcript (here the scale is from -100 to 100 for each). This outcome allows me to assess whether viewers recognize partisan signals, and whether recognition differs across modalities and lengths.

The final outcome moves beyond self-reports to observed behavior. After treatment, participants are asked to choose whether to allocate a monetary lottery prize to a pro-immigrant charity, and if so, how much. Respondents make the choice before knowing whether they won the lottery or not. This choice provides a behavioral measure of the downstream consequences of exposure, allowing me to test whether partisan signals affect not only perceptions and attitudes but also concrete actions.[19]

[18] The construction of this index follows Afrouzi et al. (2024).

[19] The charity question is asked after the block of emotion questions. This ordering is deliberate: it avoids priming respondents with an explicit behavioral task before they report their feelings about the video.

Together, these four outcomes span the main channels through which partisan signals in video news may matter: shifts in policy opinions, emotion reactions, perceptions of partisan bias, and behavioral decisions. This allows me to show the slower, content-based influence of textual cues and the faster, emotion-driven influence of visuals, consistent with the theoretical framework developed above.

5.2 Empirical strategy

Respondents are first assigned to watch either a short ($\approx$ 30s) or long ($\approx$ 2m) clip. Within each length, I independently vary the partisanship of the visuals and of the narrated transcript: the images shown are either Republican-leaning or Democratic-leaning, and the voiceover text is likewise either Republican- or Democratic-leaning. This design yields four combinations of image–text partisanship for each clip length. Let $\text{ImageR}_i \in \{0, 1\}$ indicate Republican (vs. Democratic) images for respondent i, and let $\text{TextR}_i \in \{0, 1\}$ indicate Republican (vs. Democratic) text.

Outcomes Y_i span the four group of post-treatment question: (i) immigration policy preferences; (ii) emotion responses; (iii) perceived partisanship of the content; (iv) behavioral choice. All outcomes are standardized, allowing coefficients to be compared. The pre-treatment covariate vector X_i includes demographics and standardized indices for prior views on gun control, abortion, healthcare, and taxes.

Main specification. My main specification is

$$Y_i \;=\; \beta_0 \;+\; \beta_1 \,\text{ImageR}_i \;+\; \beta_2 \,\text{TextR}_i \;+\; X_i^{\top}\gamma \;+\; \varepsilon_i, \tag{2}$$

estimated by OLS with robust standard errors.

By design, β_1 captures the effect of switching images from Democratic to Republican while holding text Democratic (visual channel), and β_2 the analogous effect for text while holding images Democratic (textual channel). Although randomization theoretically guarantees unbiasedness even without controls, the covariate vector X_i is still helpful to improves precision. Equation (2) is estimated separately for long and short exposure conditions.

The preferred specification interprets β_1 and β_2 as the average effects of switching the partisanship of one modality from Democratic to Republican, averaging across the partisanship of the other. This approach is appropriate given the absence of a pure control, as it captures the average marginal contribution of each modality's partisanship rather than effects conditional on a specific text–image pairing. A potential concern is that mixed treatments (e.g., Republican text with Democratic images) may introduce misalignment effects. While this is limited in the present setting—since all segments concern immigration coverage—I verify robustness by estimating a regression directly

comparing misaligned treatments (RD vs. DR), reported in the Appendix.

I present results in three steps. I begin with perceived partisanship, testing whether viewers recognize the partisan orientation of the treatments and whether this recognition varies with exposure length. I then compare affective and attitudinal responses to isolate how images and text differ in their effects: images are expected to move emotions, while text shifts policy attitudes. Third, I examine behavior using the donation decision as a measure of whether these cues translate into concrete actions. Exploiting the balanced sample of Democratic and Republican respondents, I analyze throughout heterogeneity in responses for the outcomes, in order to assess whether different types of viewers react differently to each modality.

5.3 Experimental Results

Perceived Partisanship Republican visuals strongly increase the perceived Republican leaning of the broadcast. As shown in Table 1, Republican images raise perceived Republican partisanship by about 0.2 standard deviations, and this effect remains stable across exposure lengths. Textual cues are weaker: Republican transcripts produce a small, imprecise increase in perceived partisanship under long exposure and no detectable effect in short clips.

Republican Partisanship		
	Long	Short
Image (Rep)	0.202***	0.210***
	(0.047)	(0.049)
Text (Rep)	0.072	-0.029
	(0.047)	(0.049)
Obs.	1748	1682

Table 1: Effect on Perceived Republican Partisanship

These results indicate that viewers primarily infer political orientation from visuals. Image signals are immediate and retain their influence when attention is limited, whereas textual signals require sustained processing and quickly lose salience. Table Appendix 5 shows that when respondents are asked specifically about the transcript rather than the full video, they recognize some partisan differences, but this recognition also disappears under short exposure.

I next examine whether these patterns differ by respondents' party affiliation. Table 2 reports the effects separately for Republicans and Democrats. The contrast is striking. Among Republican respondents, visuals matter under short exposure: Republican images increase perceived Republican partisanship by roughly 0.2 standard, but no effect in long clips or from text.

Democratic respondents are actually responsive to textual cues. Republican transcripts increase

Republican Partisanship				
	Republicans		Democrats	
	Long	Short	Long	Short
Image (Rep)	0.050	0.207***	0.317***	0.220***
	(0.069)	(0.065)	(0.063)	(0.072)
Text (Rep)	-0.026	-0.205***	0.168***	0.138*
	(0.068)	(0.066)	(0.064)	(0.071)
Obs.	788	869	960	813

Table 2: Effect on Perceived Republican Partisanship, by party affiliation of viewer

perceived partisanship by about 0.17 standard deviations in long clips and remain marginally significant in short ones, while losing power and magnitude. These differences suggest that viewers interpret partisan signals selectively—Republicans rely on visual cues, while Democrats attend more to text—and that attention constraints amplify this asymmetry.

These patterns mirror the earlier empirical results: visuals dominate under limited attention, while text requires sustained processing to shape partisan interpretation. The next section examines whether this asymmetry extends beyond perception by comparing how images and text affect emotions and policy attitudes.

Text (Slowly) Affects Topic Attitudes; Images (Quickly) Affect Emotions. The analysis of video content showed that images are "short and emotional", while text is "long and topic-based". If these properties govern how viewers process partisan information, images should influence emotions quickly and under limited attention, whereas text should shape policy attitudes only with sustained exposure.

(a) Anti-Immigration		
	Long	Short
Image (Rep)	-0.037	0.005
	(0.029)	(0.028)
Text (Rep)	0.071**	-0.013
	(0.029)	(0.028)
Obs.	1748	1682

(b) Negative Emotions		
	Long	Short
Image (Rep)	0.120**	0.188***
	(0.048)	(0.046)
Text (Rep)	0.001	0.031
	(0.048)	(0.046)
Obs.	1748	1682

Table 3: Effect on Immigration Topic Attitudes and Emotions

To test this implication, I focus on two outcomes reflecting these distinct channels of influence. The first is an anti-immigration index, constructed as the average of items on deportation policies,

border-patrol expansion, and restrictions on regularization. The second is a negative-emotion index, defined as the average of self-reported fear, anger, and disgust after treatment (results are robust to including sadness).

Table 3 presents the core result. Panel (a) shows that text persuades, but only with enough exposure (attention): Republican text increase the anti-immigration index by about 0.07 standard deviations in long clips (s.e. 0.029), with no effect in short clips. The effect is larger on items closest to the issue (e.g., ~ 0.17 s.d. for border-patrol hiring), further evidence that textual persuasion operates through propositional content. Images, by contrast, do not shift policy attitudes significantly at either length.

Panel (b) shows the complementary pattern for emotions: images move emotions, and especially under limited attention. Republican visuals raise negative emotions by 0.12–0.19 standard deviations, with the largest and most robust effect in short clips. Text has no reliable impact on emotions.[20].

Taken together, the table delivers the experiment's main message: text changes attitudes only with long exposure, while images immediately shift emotions—and do so precisely when attention is scarce. This mirrors the modality–attention evidence: the textual channel is slower and cognitive; the visual channel is fast and affective.

(a) Anti-Immigration				
	Republicans		Democrats	
	Long	Short	Long	Short
Image (Rep)	-0.001	0.048	-0.059	-0.044
	(0.039)	(0.035)	(0.041)	(0.043)
Text (Rep)	0.008	0.024	0.118***	-0.054
	(0.039)	(0.035)	(0.040)	(0.042)
Obs.	788	869	960	813

(b) Negative Emotions				
	Republicans		Democrats	
	Long	Short	Long	Short
Image (Rep)	-0.002	0.173***	0.197***	0.201***
	(0.067)	(0.058)	(0.068)	(0.072)
Text (Rep)	0.004	0.047	-0.033	0.005
	(0.066)	(0.058)	(0.068)	(0.072)
Obs.	788	869	960	813

Table 4: Effect on Topic Attitudes and Emotions, by Party Affiliation of Viewer

We now turn to heterogeneity by party affiliation. Table 4 shows that the patterns mirror those for perceived partisanship. The persuasive effect of text is concentrated among Democratic respondents: Republican transcripts increase anti-immigration attitudes by about 0.12 standard deviations in long clips, but have no effect in short ones. Among Republicans, textual cues have no detectable influence on policy attitudes at either length.

For emotions, the pattern changes. Republican visuals raise negative emotions for both groups, and Republicans are affected under short exposure—about 0.17 standard deviations. Democrats respond emotionally to Republican images as well, though their reactions are similar across exposure

[20]Text does have some moderate effect on sympathy, as shown in the Appendix

lengths. Consistent with the earlier evidence, Republican viewers are affected only by visuals and only when attention is limited.

Not Just Emotions? Image Effect on Charity Choice. Are these effects limited to emotions, or do they extend to behavior? Emotions are central to political communication because they shape engagement, attention, and action. To test whether the emotional channel translates into concrete behavior, I use a simple behavioral measure: respondents are asked whether they would donate part of a $25 lottery prize to a pro-immigration charity. The outcome is binary but standardized for comparability.

Charity Choice		
	Long	Short
Image (Rep)	0.062	-0.094**
	(0.044)	(0.044)
Text (Rep)	0.023	0.023
	(0.044)	(0.044)
Obs.	1748	1682

Table 5: Effect on Pro-immigration Charity Choice

Table 6 shows that only images affect behavior. Republican visuals reduce the probability of donating by about 0.09 standard deviations in short clips, while text has no effect. The image effect flips in long exposure, and become positive while still not statistically significant.

While the results on charity choice are less definitive than the ones presented above, they still reinforce the interpretation of the visual channel as fast and affective: under limited attention, images alter both emotions and —at least to some extent— behavior, while text does not.

The heterogeneity results shed light on these patterns. Table 6 shows that the behavioral effect of images is concentrated among Republican respondents in the short treatment: in short clips, Republican visuals reduce their likelihood of donating to the pro-immigration charity by about 0.10 standard deviations. Democrats, in contrast, show no clear behavioral response in short clips but display a small, opposite-signed effect in long clips—weakly increasing donations when exposed to Republican images. Neither group responds to textual cues.

These results align with the broader evidence on modality and attention. Republican viewers react emotionally and behaviorally to partisan visuals, but only under limited attention. Democratic viewers are less sensitive to the emotional cue itself and may engage in compensatory reasoning under longer exposure, slightly reversing the direction of the effect. Together, these patterns still suggest that the behavioral response to partisan content arises primarily from short-run affective reactions to images.

Charity Choice				
	Republicans		Democrats	
	Long	Short	Long	Short
Image (Rep)	0.034	-0.101**	0.099	-0.092
	(0.054)	(0.051)	(0.066)	(0.071)
Text (Rep)	0.061	0.041	-0.010	-0.013
	(0.055)	(0.051)	(0.066)	(0.072)
Obs.	788	869	960	813

Table 6: Effect on Pro-immigration Charity Choice, by party affiliation of viewer

5.4 Interpretation and Takeaways

The experiment confirms a systematic asymmetry in how images and text operate. Visuals act fast: they are immediately recognized as partisan cues, raise negative emotions, and under short exposure even alter behavior in the charity task. These effects appear quickly and persist across clip lengths, consistent with images transmitting partisanship through an affective channel that dominates when attention is scarce.

Text persuades slowly. It shifts topic attitudes only in long clips, when viewers have time to process the content. Republican transcripts increase anti-immigration attitudes by about 0.07 standard deviations, with effects concentrated among Democratic respondents. This gradual accumulation reflects the propositional nature of textual information—persuasion occurs only with sustained attention.

Together, the results reveal a clear division of labor between modalities. Images drive emotions and short-run behavior; text drives policy attitudes. The heterogeneity patterns reinforce this interpretation: short emotional effects are concentrated among Republicans, while longer attitudinal shifts arise mainly among Democrats. Visual and verbal channels thus shape persuasion on different attention horizons—one fast and affective, the other slow and cognitive.

6 Conclusions

This paper develops a multimodal scalable framework to measure partisanship in political video content. By separately modeling text and images and combining them in a unified specification, I show that the two modalities capture distinct and complementary partisan signals. Images are more tightly linked to emotional content and dominate predictive performance under limited attention, while text reflects topic-based information and becomes relatively more important as exposure length increases.

Experimental evidence using real television footage shows how these informational differences translate into differences in influence. Visual partisanship acts fast: it shapes perceptions, triggers emotional responses, and, under short exposure, alters concrete behavior in the charity task. Textual partisanship acts slow: it requires sustained attention to affect attitudes toward policy. Together, these results suggest that the mechanisms of persuasion in audiovisual media are both modality- and attention-dependent.

These findings carry three broader implications. First, for media bias measurement, relying solely on transcripts systematically understates the ideological content of video news—particularly in the short, attention-limited segments that dominate modern media consumption. Incorporating visual analysis is crucial in these contexts to provide an accurate measure of partisan slant.

Second, for political persuasion, the results reveal a clear division of labor between modalities. Images dominate in low-attention environments, transmitting affective information that shapes perception, emotion, and behavior. Text regains importance only under sustained attention, shifting attitudes through its propositional content.

Finally, for media policy design—including debates on disclosure, labeling, and content moderation—these results underscore that approaches focused only on text overlook a central channel of political communication. In an increasingly video-based and attention-scarce information environment, understanding partisanship requires accounting for the speed, emotional force, and distinct informational role of images alongside text.

References

Afrouzi, Hassan, Carolina Arteaga, and Emily Weisburst (2024). "Is it the message or the messenger? examining movement in immigration beliefs". *Journal of Political Economy Microeconomics* 2.2.

Algan, Yann, Eva Davoine, Thomas Renault, and Stefanie Stantcheva (2025). *Emotions and policy views*. Working Paper.

Andries, Marianne, Leonardo Bursztyn, Thomas Chaney, Milena Djourelova, and Alex Imas (2024). *In their shoes: Empathy through information*. Tech. rep. National Bureau of Economic Research.

Ash, Elliott, Ruben Durante, Maria Grebenschikova, and Carlo Schwarz (2021). *Visual representation and stereotypes in news media*. CEPR Discussion Paper No. DP16624.

Ash, Elliott and Sergio Galletta (2023). "How cable news reshaped local government". *American Economic Journal: Applied Economics* 15.4.

Ash, Elliott and Michael Poyker (2024). "Conservative news media and criminal justice: evidence from exposure to the fox news channel". *The Economic Journal* 134.660.

Baddeley, Alan (1992). "Working memory". *Science* 255.5044.

Bhalla, Usha, Alex Oesterling, Suraj Srinivas, Flavio Calmon, and Himabindu Lakkaraju (2024). "Interpreting clip with sparse linear concept embeddings (splice)". *Advances in Neural Information Processing Systems* 37.

Boxell, Levi (2021). "Slanted images: measuring nonverbal media bias during the 2016 election". *Political Science Research and Methods*.

Caprini, Giulia (2024). *Visual bias*. Working Paper.

Cowen, Alan S and Dacher Keltner (2017). "Self-report captures 27 distinct categories of emotion bridged by continuous gradients". *Proceedings of the national academy of sciences* 114.38.

Dahmen, Nicole Smith (2015). "Watchdog, voyeur, or censure? an eye-tracking research study of graphic photographs in the news media". *Journalism Practice* 9.3.

DellaVigna, Stefano and Ethan Kaplan (2007). "The fox news effect: media bias and voting". *The Quarterly Journal of Economics* 122.3.

Djourelova, Milena (2023). "Persuasion through slanted language: evidence from the media coverage of immigration". *American Economic Review* 113.3.

Feder, Amir, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E Roberts, et al. (2022). "Causal inference in natural language processing: estimation, prediction, interpretation and beyond". *Transactions of the Association for Computational Linguistics* 10.

Fowler, Erika Franklin, Michael M Franz, Gregory J Martin, Zachary Peskowitz, and Travis N Ridout (2021). "Political advertising online and offline". *American Political Science Review* 115.1.

Gentzkow, Matthew and Jesse M Shapiro (2010). "What drives media slant? evidence from us daily newspapers". *Econometrica* 78.1.

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

Iyer, Aarti and Julian Oldmeadow (2006). "Picture this: emotional and political responses to photographs of the kenneth bigley kidnapping". *European Journal of Social Psychology* 36 (5).

Jensen, Jacob, Ethan Kaplan, Suresh Naidu, Laurence Wilse-Samson, David Gergen, Michael Zuckerman, and Arthur Spirling (2012). "Political polarization and the dynamics of political language: evidence from 130 years of partisan speech". *Brookings Papers on Economic Activity* 2012.2.

Kim, Been, Martin Wattenberg, Justin Gilmer, Carrie J. Cai, James Wexler, Fernanda Viegas, and Rory Sayres (2018). "Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)". *Proceedings of the 35th International Conference on Machine Learning*. Vol. 80. Proceedings of Machine Learning Research.

Lee, Michael T and Carol Theokary (2021). "The superstar social media influencer: exploiting linguistic style and emotional contagion over content?" *Journal of Business Research* 132.

Liang, Victor Weixin, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou (2022). "Mind the gap: understanding the modality gap in multi-modal contrastive representation learning". *Advances in Neural Information Processing Systems* 35.

Ludwig, Jens and Sendhil Mullainathan (2024). "Machine learning as a tool for hypothesis generation". *The Quarterly Journal of Economics* 139.2.

Martin, Gregory J. and Ali Yurukoglu (2017). "Bias in cable news: persuasion and polarization". *American Economic Review* 107 (9).

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013). "Efficient estimation of word representations in vector space". *arXiv preprint arXiv:1301.3781*.

Nelson, Douglas L, Valerie S Reed, and John R Walling (1976). "Pictorial superiority effect." *Journal of experimental psychology: Human learning and memory* 2.5.

Newman, N, A Ross Arguedas, CT Robertson, RK Nielsen, and R Fletcher (2025). *Digital news report 2025*. Tech. rep.

Pew Research Center (2025). *Social Media and News Fact Sheet*. https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news/.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever (2021). "Learning Transferable Visual Models from Natural Language Supervision". *Proceedings of the 38th International Conference on Machine Learning*. Vol. 139. Proceedings of Machine Learning Research.

Sánchez-Fernández, Raquel and David Jiménez-Castillo (2021). "How social media influencers affect behavioural intentions towards recommended brands: the role of emotional attachment and information value". *Journal of Marketing Management* 37.11-12.

Simon, Herbert A. (1971). "Designing organizations for an information rich world". *Computers, communications, and the public interest*.

Simonov, Andrey, Szymon Sacher, Jean-Pierre Dubé, and Shirsho Biswas (2022). "Frontiers: the persuasive effect of fox news: noncompliance with social distancing during the covid-19 pandemic". *Marketing Science* 41.2.

Torres, Michelle (2024). "A framework for the unsupervised and semi-supervised analysis of visual frames". *Political Analysis* 32.2.

A Additional tables and figures

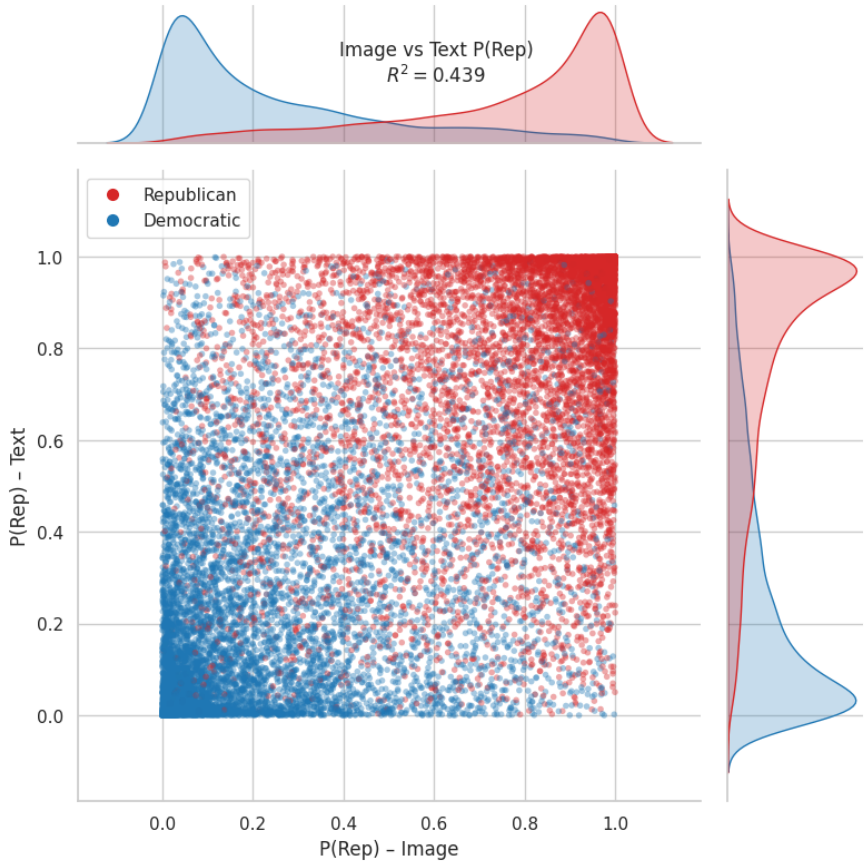


Figure 1: Predicted Republican Probability: Image vs. Text Models

Notes: Each point represents one political ad. Horizontal and vertical axes show the predicted Republican probability from the image and text models, respectively. Points are colored by ground-truth party label.

Bucket	Tokens	Bucket	Tokens
Admiration	*admiration, respect*	Contempt	*contempt, scorn, disdain*
Adoration	*adoration, devotion*	Contentment	*contentment, satisfied, fulfilled*
Aesthetic Appreciation	*aesthetic, beauty, gorgeous*	Craving	*craving, desire, yearning*
Amusement	*amusement, humorous, funny*	Disgust	*disgust, repulsed, nauseated*
Anger	*anger, enraged, furious*	Empathic Pain	*empathy, sympathy, sorrow*
Anxiety	*anxiety, uneasy, worried*	Entrancement	*entranced, captivated, mesmerized*
Awe	*awe, awestruck, wonder*	Excitement	*excitement, exhilarated, thrilled*
Boredom	*boredom, bored, uninterested*	Fear	*fear, terrified, frightened*
Calm	*calm, serene, tranquil*	Gratitude	*gratitude, thankful, appreciative*
Confusion	*confusion, confused, perplexed*	Guilt	*guilt, guilty, remorse*
Horror	*horror, horrified, appalled*	Interest	*interest, intrigued, curious*
Joy	*joy, joyful, ecstatic*	Nostalgia	*nostalgia, nostalgic, wistful*
Pride	*pride, proud, triumphant*	Relief	*relief, relieved, reassured*
Romantic Love	*love, loving, romantic*		

Table 1: 27 Dimensions of Emotions from Cowen and Keltner, 2017.

Bucket	Tokens	Bucket	Tokens
Healthcare	*healthcare, medicare, pro-aca, anti-insurance, anti-obama plan, anti-ahca, anti-aca, health insurance reform, prescription drugs, prescription drugs: cost, prescription drugs: anti-industry, coronavirus*	Economy/Jobs	*economy, jobs/unemployment, outsourcing, minimum wage, manufacturing/construction, trade, trade: china, financial services, financial reform, retirement, union*
Tax/Budget	*taxes, tax reform, budget/government spending, social security*	Social Issues	*social issues, abortion, women's rights, drugs, civil rights, opioids/opiates, faith/religion, guns, birth control, human rights*
Public Safety	*public safety, gun control, anti-gun control, pro-gun control, terrorism*	Environment	*energy/environment, oil, oil-anti, green energy, global warming, coal-pro, coal, oil-pro*
Immigration	*immigration, immigration: anti, immigration: pro*	Foreign/Defense	*international affairs, china, national defense, defense/aerospace, iraq/afghan war, veterans affairs*
Political Campaigns	*anti-trump, pro-trump, anti-biden, pro-biden, anti-clinton, pro-clinton, anti-obama message, pro-obama message, anti-sanders, pro-sanders, campaign finance reform, call to action, impeachment*	Governance	*corruption, supreme court, term limits*
Education	*education*	Housing	*housing/home ownership*
Infrastructure	*transportation, telecommunications*	Interest Groups	*food/agriculture, aarp: 50+ voters, aarp mention*

Table 2: 14 Dimensions of Topics Used in the Joint Vocabulary.

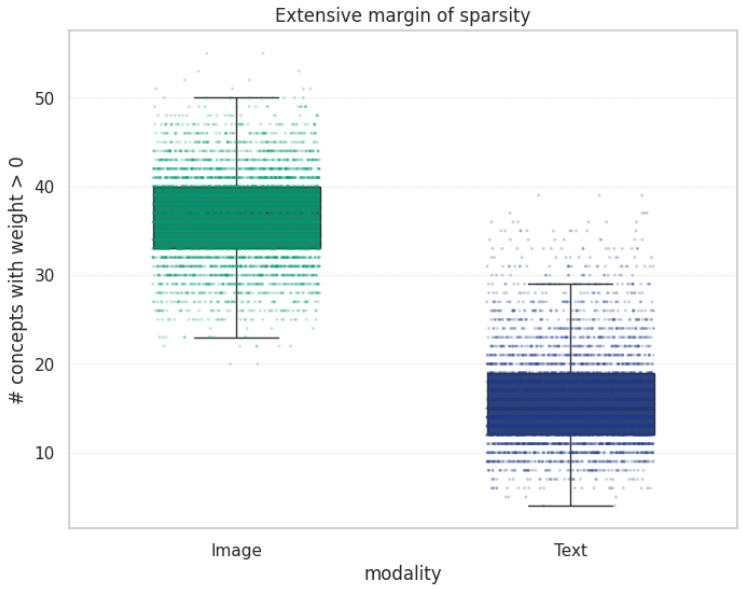


Figure 2: Distribution of non-zero concepts across image and text modalities, fixing similarity across modalities

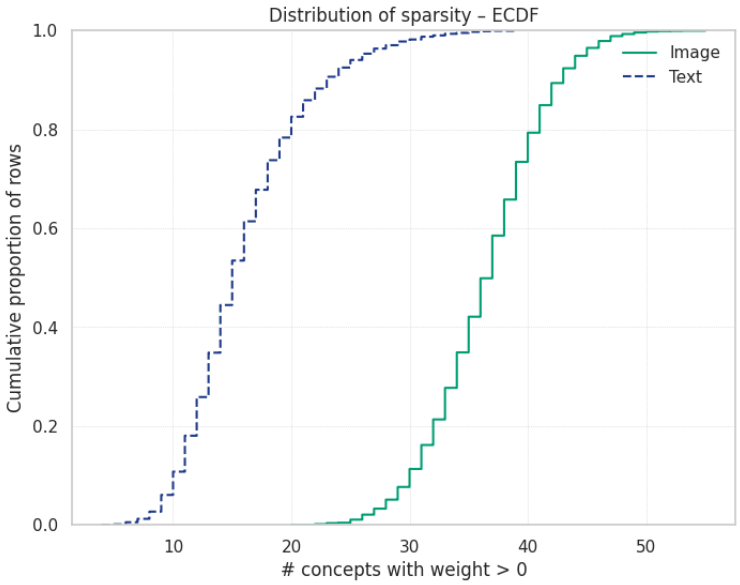


Figure 3: Cumulative distribution of the number of active concepts with nonzero weight per video.

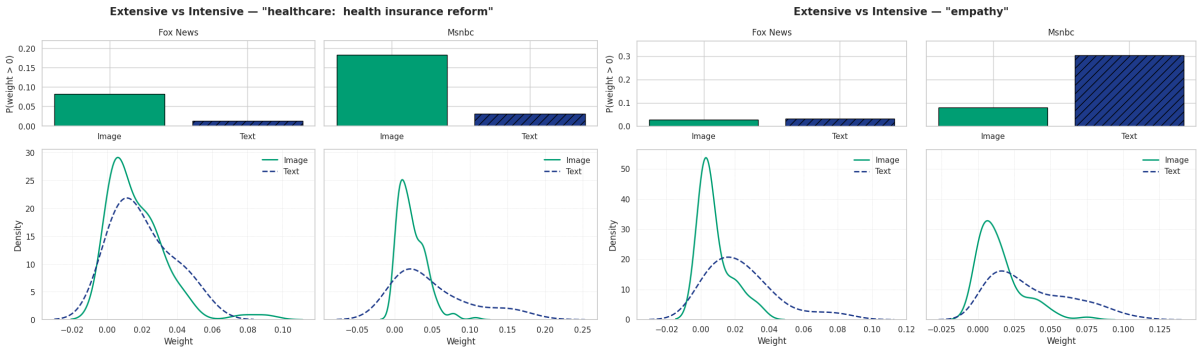


Figure 4: Extensive and intensive margins for two example concepts: *Health Insurance Reform* and *Empathy*.

Notes: The top panels report the share of ads in which the concept has a strictly positive weight in the sparse decomposition (extensive margin). The bottom panels plot the distribution of weights when the concept is present (intensive margin), separately for image and text modalities and by party.

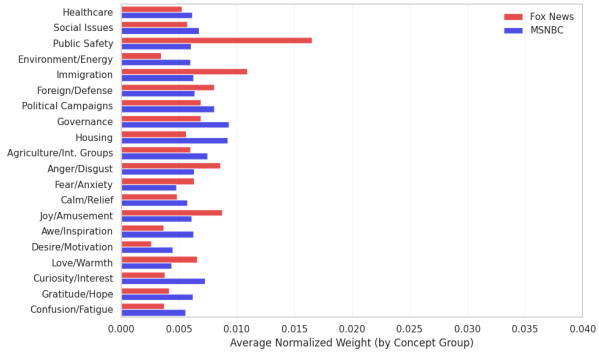


Figure 5: Average Weights in Images

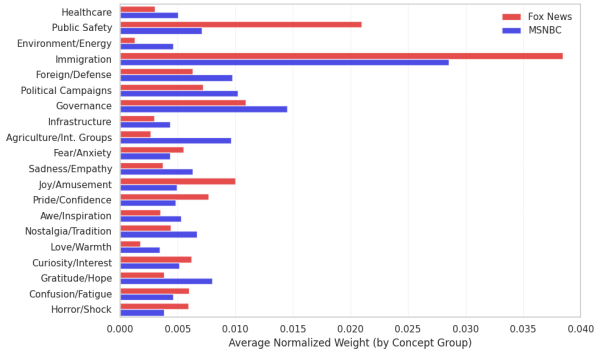


Figure 6: Average Weights in Text

Figure 7: Average Concept Weights in Image and Text Representations of YouTube News Videos

Notes: Each bar reports the average normalized weight of concepts in the joint SpLICE representation, aggregated at the group level (emotion or topic) and averaged across videos from Fox News (red) and MSNBC (blue). Buckets are ordered according to their position in the joint vocabulary. Weights represent the share of each modality's representation explained by a given conceptual group.

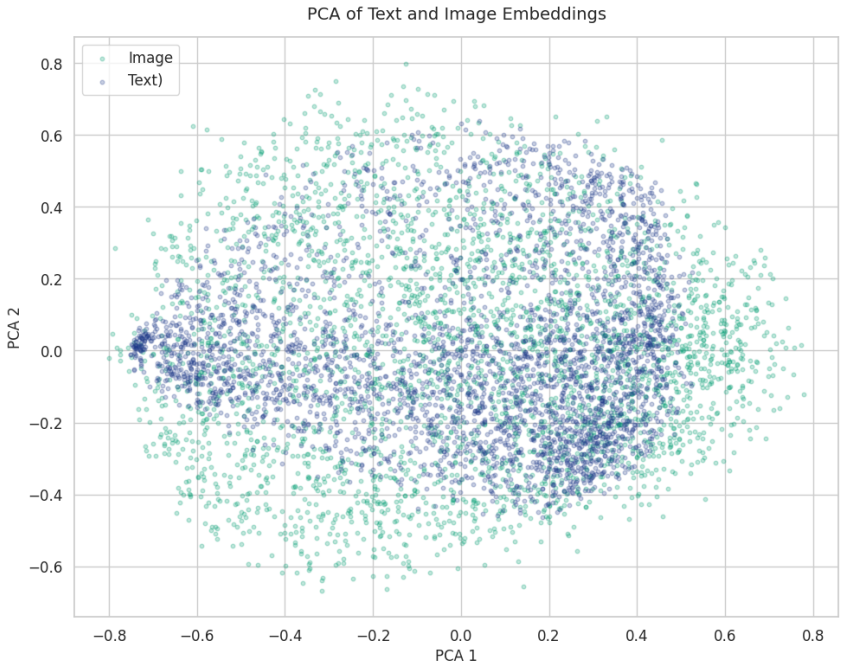


Figure 8: Principal component structure of image and text embeddings across all videos.

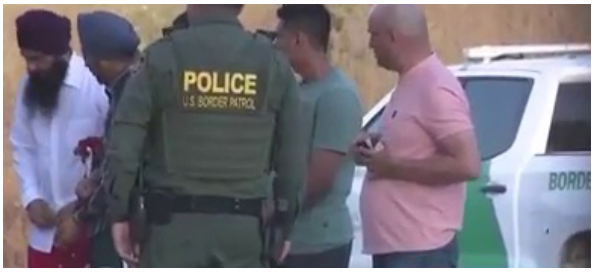
Notes: The figure shows the first two principal components of the original image and text embeddings for all videos in the dataset. The lack of visible clustering across modalities is expected given the mean-centering of embeddings. This projection provides a reference for the movement observed in Figure 11.

Table 3: TV News Sample Summary (All Channels)

Channel	# Videos	Avg Words/Video	Avg Images/Video	Avg Views/Video
ABC News	372	1,675.1	118.5	184,200
CBS News	493	936.1	80.0	34,792
CNN	595	1,281.5	105.3	467,424
Democracy Now!	368	1,787.7	147.8	178,746
FOX 4 Dallas-Fort Worth	67	1,113.6	104.7	82,781
Fox News	955	1,092.9	91.1	350,153
KTLA 5	34	816.7	75.3	34,213
MSNBC	947	2,060.5	104.2	182,954
NBC News	455	1,627.9	110.2	116,952
PBS NewsHour	495	879.9	96.8	73,203
The Next News Network	96	2,040.2	151.8	52,658
WGN News	279	1,033.8	80.2	3,586
Total	**5,156**	**1,408.6**	**102.99**	**146,805**

Note: Averages are per video. Channel upload date ranges are reported in the Appendix along with scraping and preprocessing details.

Raw TV news frame with graphics (banners, ticker, split screen).



On-scene crop retained via line detection; overlays removed.

Figure 9: TV news selection: defining the "image." The left panel shows a typical broadcast frame that mixes story content and on-screen graphics; the right panel shows the content-focused crop used for embedding. The procedure is applied uniformly across channels and videos.

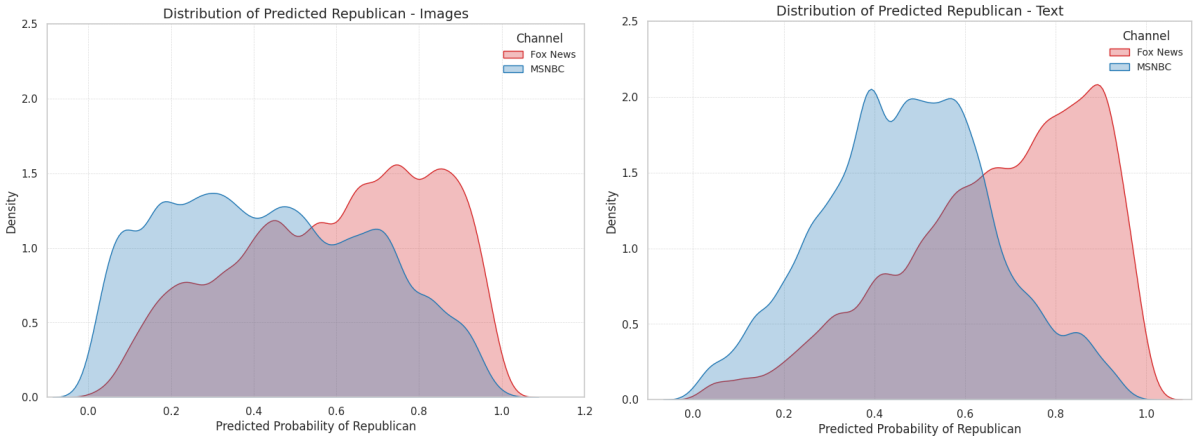


Figure 10: Distribution of predicted Republican probabilities for Fox News and MSNBC videos.

Notes: Each panel shows the kernel density of predicted probabilities from the image-only and text-only classifiers, respectively.

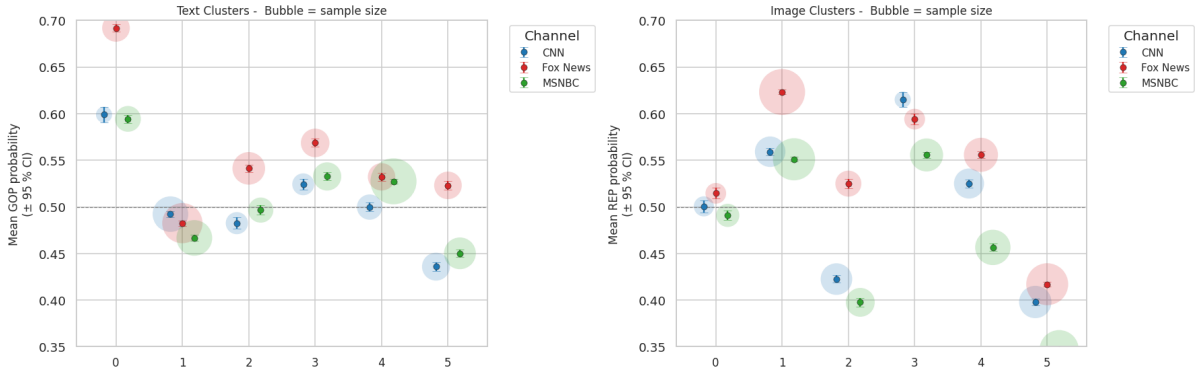
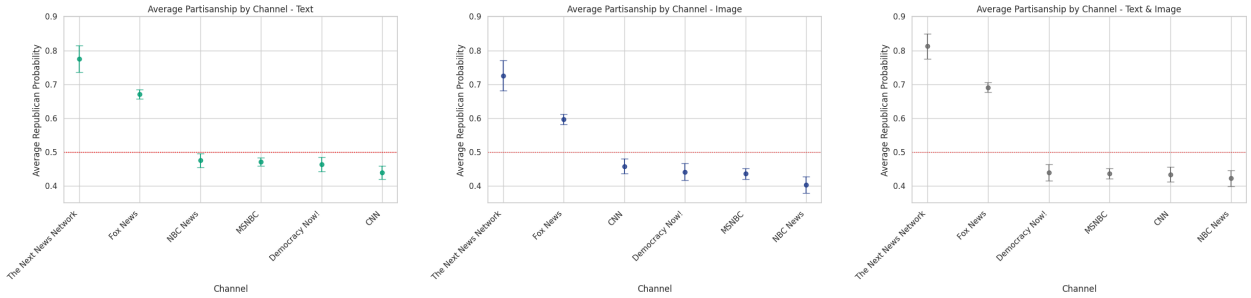


Figure 11: Average predicted Republican probability by channel and content cluster.

Notes: Each point shows the mean predicted Republican probability for a cluster of semantically similar text (left) or image (right) embeddings, with 95% confidence intervals. Bubble size reflects the number of observations within each cluster. Predictions are obtained from the ad-trained classifier applied to the news video corpus.



Text only. Image only. Text & Image.

Figure 12: Average predicted Republican probability by channel at the video level, for the 3 models.

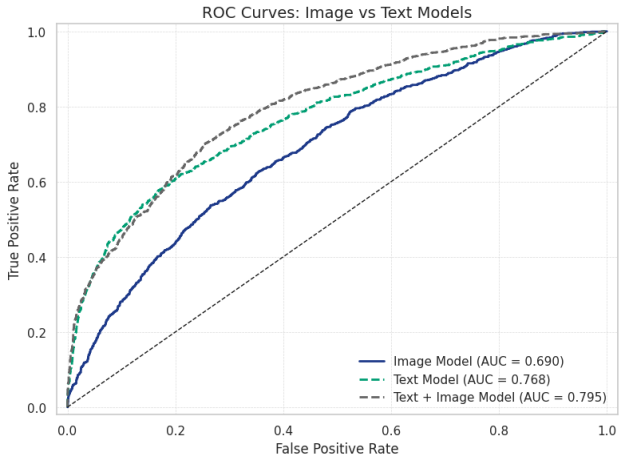


Figure 13: ROC curves for image-only, text-only, and pooled models at the video level.

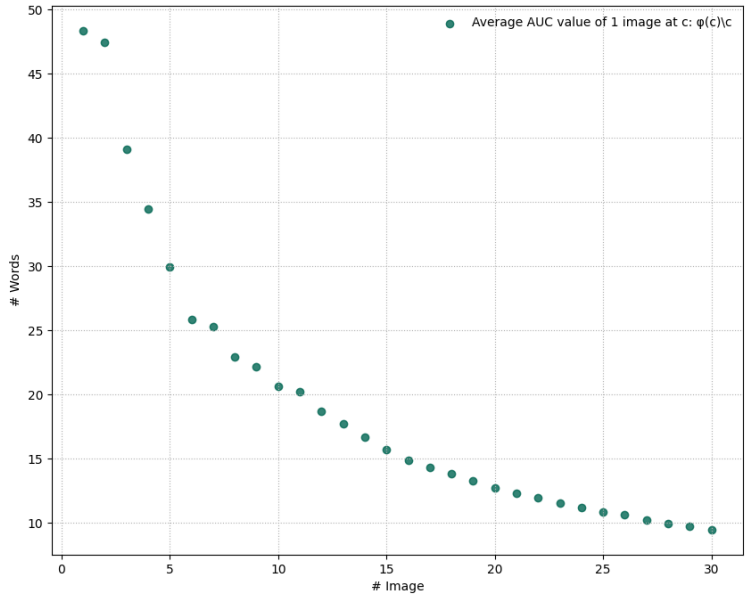


Figure 14: Image to words equivalence. Each point maps c images to words as length increases. Early images are especially informative.

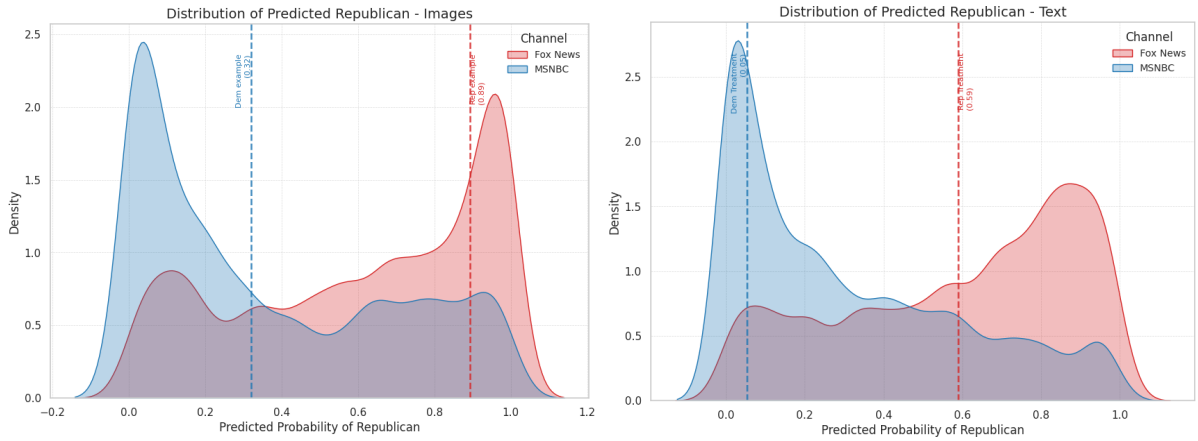


Figure 15: Distribution of predicted Republican probabilities for Fox News and MSNBC videos.

Notes: Each panel shows the kernel density of predicted probabilities from the image-only and text-only classifiers, respectively. The distribution are re-estimated within the immigration-only clusters using a partisanship prediction model trained *only* on immigration-focused political ads.

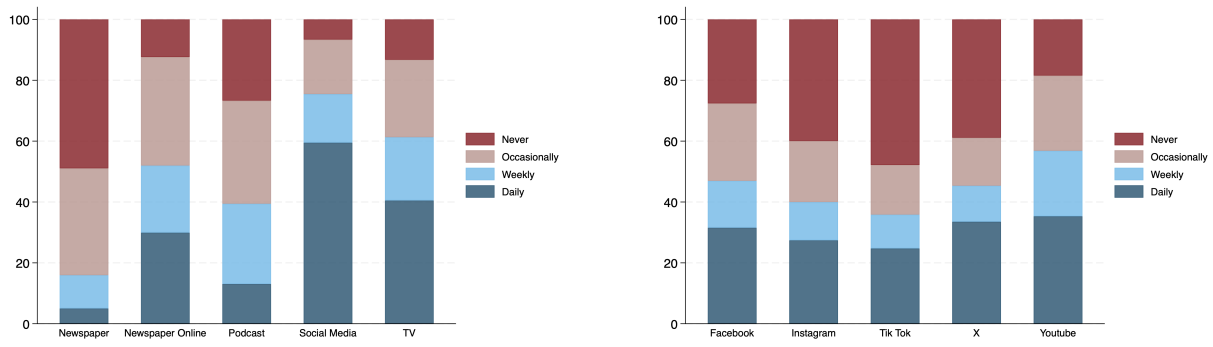


Figure 16: News Consumption Frequency by Platform

Notes: The left panel shows the frequency of news consumption across broad media platforms (TV, social media, newspapers, podcasts). The right panel shows the frequency of news consumption across individual social media platforms. Social media and television are the most widely used sources overall. YouTube is the dominant platform for news consumption within social media, followed by X and Facebook. Data from original survey ($N = 4000$).

Table 4: Balance Across Treatment Groups

Notes: The table reports mean characteristics and balance tests across the eight experimental treatment groups. Each row displays the sample mean, standard deviation, and the F-test and p-value for equality of means across treatments. No systematic imbalances are observed across demographic or media consumption variables. $N = 3{,}147$.

	Full Sample			
	Mean	S.D.	F-Test	P-Value
Female	0.496	(0.500)	1.361	0.217
White	0.755	(0.430)	1.564	0.141
Mixed	0.056	(0.230)	1.423	0.191
Black	0.112	(0.316)	1.069	0.380
18–24 years	0.067	(0.250)	1.225	0.285
25–34 years	0.239	(0.427)	0.895	0.509
45–64 years	0.431	(0.495)	1.355	0.220
College or More	0.609	(0.488)	1.763	0.090
Full-time Employed	0.552	(0.497)	0.898	0.507
News (Weekly+): Newspaper	0.026	(0.159)	1.557	0.144
News (Weekly+): Newspaper Online	0.257	(0.437)	0.754	0.626
News (Weekly+): TV	0.309	(0.462)	0.502	0.833
News (Weekly+): Social Media	0.537	(0.499)	0.511	0.827
News (Weekly+): Podcast	0.111	(0.314)	2.242	0.028
News (Weekly+): NY Times	0.106	(0.308)	0.445	0.874
News (Weekly+): CNN	0.127	(0.333)	1.541	0.148
News (Weekly+): MSNBC	0.076	(0.264)	0.494	0.840
News (Weekly+): Newsmax	0.029	(0.168)	1.923	0.062
News (Weekly+): Facebook	0.239	(0.426)	2.662	0.010
News (Weekly+): Twitter	0.206	(0.405)	0.531	0.812
News (Weekly+): Instagram	0.181	(0.385)	0.408	0.898
News (Weekly+): TikTok	0.186	(0.389)	0.526	0.816
News (Weekly+): YouTube	0.293	(0.455)	1.029	0.408
Top Issue: Healthcare	0.148	(0.356)	0.362	0.925
Voted (2024)	0.917	(0.276)	0.947	0.469
Voted for Trump (2024)	0.467	(0.499)	2.390	0.019
N	3147			

	(Log of) Video Partisanship					
	Full Sample		Republicans		Democrats	
	Long	Short	Long	Short	Long	Short
Image (Rep)	0.036***	0.030***	0.021	0.025**	0.045***	0.035***
	(0.009)	(0.009)	(0.013)	(0.012)	(0.011)	(0.013)
Text (Rep)	0.004	-0.011	-0.010	-0.037***	0.018	0.013
	(0.009)	(0.009)	(0.013)	(0.012)	(0.011)	(0.013)
Observations	1475	1414	635	735	840	679

(a) General video partisanship: outcome is $\log(1 +$ video republican partisanship$)$

	(Log of) Text Partisanship					
	Full Sample		Republicans		Democrats	
	Long	Short	Long	Short	Long	Short
Image (Rep)	0.029***	0.009	0.019*	0.004	0.034***	0.014*
	(0.006)	(0.006)	(0.010)	(0.009)	(0.008)	(0.008)
Text (Rep)	0.023***	-0.006	0.012	-0.023**	0.032***	0.011
	(0.006)	(0.006)	(0.010)	(0.009)	(0.008)	(0.008)
Observations	1475	1414	635	735	840	679

(b) Text partisanship: outcome is $\log(1 +$ text republican partisanship$)$

	(Log of) Image Partisanship					
	Full Sample		Republicans		Democrats	
	Long	Short	Long	Short	Long	Short
Image (Rep)	0.072***	0.055***	0.040***	0.049***	0.092***	0.064***
	(0.007)	(0.007)	(0.012)	(0.011)	(0.009)	(0.010)
Text (Rep)	0.025***	-0.013*	0.018	-0.032***	0.031***	0.006
	(0.007)	(0.008)	(0.012)	(0.011)	(0.009)	(0.010)
Observations	1475	1414	635	735	840	679

(c) Image partisanship: outcome is $\log(1 +$ image republican partisanship$)$

Table 5: Regression on perceived Republican partisanship of the video and of each modality separately. For the whole video, outcomes are originally on a 1 to 5 scale, for the individual modalities outcomes are originally on a -100 to 100 slider scale. Higher scores indicate higher Republican leaning. I normalize the scores and take the logs.

Increase Border Patrol						
	Full Sample		Republicans		Democrats	
	Long	Short	Long	Short	Long	Short
Image (Rep)	-0.045	-0.027	-0.046	-0.025	-0.030	-0.027
	(0.034)	(0.034)	(0.040)	(0.037)	(0.052)	(0.056)
Text (Rep)	0.170***	0.009	0.077*	0.033	0.250***	-0.019
	(0.035)	(0.034)	(0.041)	(0.037)	(0.052)	(0.055)
Obs.	1748	1682	788	869	960	813

Table 6: Effect on Increase Border Patrol

Anger						
	Full Sample		Republicans		Democrats	
	Long	Short	Long	Short	Long	Short
Image (Rep)	0.093*	0.136***	-0.049	0.172***	0.183***	0.094
	(0.048)	(0.046)	(0.066)	(0.060)	(0.068)	(0.071)
Text (Rep)	0.003	0.021	0.032	0.018	-0.052	0.011
	(0.048)	(0.046)	(0.065)	(0.060)	(0.068)	(0.072)
Obs.	1748	1682	788	869	960	813

Table 7: Effect on Anger

Disgust						
	Full Sample		Republicans		Democrats	
	Long	Short	Long	Short	Long	Short
Image (Rep)	0.162***	0.158***	0.044	0.132**	0.242***	0.185**
	(0.048)	(0.046)	(0.066)	(0.059)	(0.068)	(0.073)
Text (Rep)	-0.004	0.033	-0.018	0.066	-0.023	-0.008
	(0.048)	(0.046)	(0.064)	(0.058)	(0.068)	(0.073)
Obs.	1748	1682	788	869	960	813

Table 8: Effect on Disgust

(a) Anti-Immigration		
	Long	Short
RD vs DR	-0.100**	-0.021
	(0.042)	(0.041)
Obs.	874	831

(b) Charity Choice		
	Long	Short
RD vs DR	0.028	-0.126**
	(0.063)	(0.063)
Obs.	874	831

(c) Negative Emotions						
	Full Sample		Republicans		Democrats	
	Long	Short	Long	Short	Long	Short
RD vs DR	0.106	0.157**	-0.026	0.108	0.204**	0.194*
	(0.070)	(0.068)	(0.095)	(0.101)	(0.104)	(0.107)
Obs.	874	831	402	420	472	411

(d) Heterogeneity by Party				
	Republicans		Democrats	
	Long	Short	Long	Short
Anti-Immigration (RD vs DR)	0.008	0.020	-0.180***	0.013
	(0.055)	(0.047)	(0.060)	(0.063)
Charity Choice (RD vs DR)	-0.034	-0.154**	0.114	-0.067
	(0.073)	(0.074)	(0.090)	(0.102)

Table 9: Modality Misalignment: Effects of Cross-Partisan Pairings (RD vs DR)

Notes: Each panel compares videos where visual and textual partisanship are misaligned (Republican text with Democratic images, RD, versus Democratic text with Republican images, DR). Panel (a) reports effects on anti-immigration attitudes; panel (b) on charity choice; panel (c) on negative emotions; and panel (d) on heterogeneity by party affiliation.

B Partisanship Model Details

Details of CLIP embedding representation Let $\mathbf{x}_j^{\text{img}} \in \mathbb{R}^{d_i}$ and $\mathbf{x}_j^{\text{txt}} \in \mathbb{R}^{d_t}$ represent the j-th image frame and text snippet extracted from advertisement A. Define the CLIP image and text encoders as mappings $f : \mathbb{R}^{d_i} \to \mathbb{R}^d$ and $g : \mathbb{R}^{d_t} \to \mathbb{R}^d$, respectively.

Assumption 1. *The embedding representing an entire advertisement A can be approximated by the average of the embeddings of its individual constituent frames or snippets:*

$$\mathbf{x}_A^{mod} \approx \frac{1}{N} \sum_{j=1}^{N} f^{mod}(\mathbf{x}_j^{mod}), \quad mod \in \{img, txt\},$$

where N is the number of frames or snippets, and f^{mod} denotes the modality-specific CLIP encoder.

The main limitation of this assumption is that it ignores sequence effects and higher-order interactions among elements of the ad; we provide empirical evidence in Appendix [...] that averaging closely approximates a full-context embedding.

A challenge in multimodal analysis is the modality gap: image and text embeddings tend to occupy different regions of the shared space (Liang et al. (2022)). Without adjustment, this limits their direct comparability and can bias pooled estimates. We address this in three steps: (i) normalize each ad's embeddings to lie on the unit sphere; (ii) compute the average embedding across all ads for each modality; (iii) subtract the modality-specific global mean from each embedding and re-normalize.

For ad A in modality mod $\in \{\text{img}, \text{txt}\}$, with normalized embeddings $\bar{\mathbf{x}}_A^{\text{mod}}$ and global mean $\boldsymbol{\mu}_{\text{mod}}$, the adjusted embedding is

$$\tilde{\mathbf{x}}_A^{\text{mod}} = \frac{\bar{\mathbf{x}}_A^{\text{mod}} - \boldsymbol{\mu}_{\text{mod}}}{\|\bar{\mathbf{x}}_A^{\text{mod}} - \boldsymbol{\mu}_{\text{mod}}\|_2}, \quad \text{mod} \in \{\text{img}, \text{txt}\}.$$

This centering ensures that both modalities share a common reference point, enabling meaningful comparison and integration in subsequent analyses. Intuitively, this adjustment controls systematic differences between how CLIP encodes text and images, so that any remaining variation in the embeddings reflects content rather than the modality itself.

C Experiment Text & Image Treatments

Text Treatment.

Democratic Text:

"» Right now a federal appeals court in New Orleans is hearing arguments again about a controversial new Texas immigration law. This law, which the court has currently put on hold, would let the state arrest and deport migrants for illegally crossing the border. The State and Federal government have been in and out of court for months over this, arguing about whether the state should have that power. let's get right to our homeland security correspondent. We've been following this law through so many twists and turns, what should we know about this hearing today?

» So, I've been listening to the oral arguments and there's something you should know, because Texas is actually changing the way they said this law would work. Just as you pointed out, they've always said they would arrest and deport them. Their plan was to walk migrants back into Mexico and change the territory on which these migrants stood. Now, they're arguing, before this judge panel in the 5th circuit, that they would simply turn these migrants back over to the custody of the United States by taking them to a port of entry on the U.S. side, and handing them over to customs and border protection. That's similar to something Texas already has been doing for a number of years, after they charged someone for trespassing on Texas state property. And so it really left the justice department, arguing on behalf of the federal government, in a strange situation as judges then started questioning them: well, could this stand? Would this be okay, if Texas did it this way? They simply pointed to the language that was already in the law and then the arguments to say: that's not how they said they would do this. So, a lot of confusion going on inside this courtroom right now. It seems that texas might need to rework the way the law is worded and the justice department would need more time to figure out how they would respond to a law that looked very different than what they set out in the first place."

"Republican" Text:

"» The showdown between texas and the biden administration, now going one step higher on the legal ladder. appeals court set to hear oral arguments about a law that allows the lone star state to arrest and deport illegals, only weeks after a lower court put that law on hold. Where to now? Our correspondent live in new orleans to cover this today. Good morning.

» Oral arguments expected to go about an hour, though they could go a little bit longer given this issue, very controversial, this will be a three-judge appeals panel here in new orleans. The reason we're in new orleans talking about a texas immigration law is because new orleans fifth

court of appeals covers the state of texas. Now, the fifth circuit Court of Appeals here in downtown new orleans is a traditionally conservative appeals court here in the united states. One of the most in the united states, if not the most. We won't know exactly how the court is leaning until these arguments are underway. You may remember last month, the supreme court allowed the Texas law that allows local Texas law enforcement to arrest anyone who is suspected of crossing the border illegally to be enforced, and to arrest those people. Just hours later a lower court put a pause, or a stay in place. The justice department will be here today, fighting to make this law unconstitutional arguing that immigration control is an entirely federal power and U.S. border patrol agents alone control the border and make arrests. Governor Greg Abbott in Texas says that the law is needed because of the state's sovereignty from an invasion of illegal immigrants bringing crime and drugs. He says the border patrol agents are understaffed and overwhelmed and the State should be able to make arrests. Abbott and his team says that the border patrol not doing their job, so he wants to do his job by having the state enforce this law. We're not going to have a final decision today, they're not gonna rule from the bench. But no matter what happens today in New Orleans we expect to fight in Washington back at the steps of the Supreme Court on this very issue."

Image Treatment.

Video: Democratic Images

Video: Republican Images

D Preprocessing Details

Text. We standardize and represent video transcripts in four steps.

(i) Normalization. Transcripts are lowercased, control characters and simple markup are removed, and excess whitespace is collapsed. To reduce caption noise, we drop common non-lexical fillers (e.g., *um*, *uh*, *erm*, *like*, *you*, *know*). Immediate word repetitions generated by automatic speech recognition (ASR) are de-duplicated.

(ii) Punctuation restoration. Because ASR captions often lack sentence boundaries, we restore punctuation with a pre-trained restoration model. Processing is performed in short, fixed-length slices to respect the model's token limits; slices are then rejoined and sentence starts capitalized.

(iii) Alignment to frames. Let N denote the number of retained frames for the video after image filtering. We partition the cleaned transcript into N contiguous segments of (approximately) equal word count. This yields a one-to-one alignment between text segments and frames without relying on timestamps.

(iv) Representation. Each segment is tokenized and embedded with the CLIP text encoder, producing a segment-level vector used in chunk-based analyses. We also compute a video-level average text embedding for aggregation.

(v) Quality control. We exclude videos with missing or empty transcripts or with zero retained frames. For included videos we enforce $N_{\text{text}} = N_{\text{frames}}$ and report counts of exclusions and alignment failures.

Design rationale. We intentionally avoid lemmatization, stemming, and broad stopword removal

beyond the filler list. Because CLIP uses subword tokenization, preserving surface forms maintains fidelity to the encoder's pretraining distribution and avoids altering token boundaries in ways that may distort embeddings.

Images. We process extracted frames to isolate the primary content region and suppress broadcast overlays (tickers, lower-thirds, pillarboxes).

(i) Structural detection. Each frame is converted to grayscale and lightly smoothed. We compute edges and detect straight lines using standard computer-vision primitives (Canny edges and a probabilistic Hough transform), restricting attention to near-horizontal and near-vertical lines. Detected lines are bucketed by screen half (top/bottom for horizontals; left/right for verticals) to exploit typical broadcast layouts.

(ii) Crop boundary selection. From each bucket we select the interior-most candidate to propose top/bottom and left/right crop boundaries. To avoid degenerate crops, candidate boundaries must be separated by at least 10% of the corresponding image dimension; if candidates are absent or too close, we fall back to the full extent on that axis. Bounds are clipped to the image size.

(iii) Subregion selection. The resulting boundaries partition the frame into up to four rectangles. We retain the largest rectangle as the primary content crop. When a second rectangle's area is at least 90% of the largest (e.g., split-screen interviews), we also retain it. If no valid partition is produced, the original frame is kept.

(iv) Representation and QA. Cropped images are embedded with the CLIP image encoder using its standard preprocessing (resize and center-crop as implemented by the encoder). We compute segment-level embeddings and a video-level average image embedding. Corrupted frames are skipped; processing is idempotent, and logs record exclusions and fallbacks.

Design rationale. We favor geometric cropping based on detected layout rather than inpainting or heavy denoising. This removes consistent studio graphics while keeping images close to the encoder's training distribution, limiting the risk of artifacts that could contaminate embeddings.

E A Simple Framework: Modality Choice under Attention Constraints

Viewers have an effective attention window $T > 0$ seconds. A media outlet allocates a share $s \in [0, 1]$ of its partisan signal to visuals (with $1 - s$ to text). The persuasive impact from modality $m \in \{v, t\}$ over window T is

$$P_m \;=\; \beta_m \, g_m\big(s_m T\big), \qquad s_v = s, \;\; s_t = 1 - s,$$

where g_m is increasing and concave with $g_m(0) = 0$, and $\beta_m > 0$ measures per-second effectiveness ("modality strength"). Visuals are assumed *faster* at short horizons:

$$g_v'(0) > g_t'(0), \qquad \text{with } g_t \text{ catching up as } T \uparrow .$$

The outlet chooses s to maximize

$$\max_{s\in[0,1]} \ \Pi(s;T) \equiv \beta_v g_v(sT) + \beta_t g_t((1-s)T) - c(s),$$

where c is convex with $c'(0) = 0$. The interior FOC is

$$\beta_v T\, g_v'(sT) \ - \ \beta_t T\, g_t'((1-s)T) \ = \ c'(s).$$

Implications. If $c''(\cdot) \geq 0$, $g_m'' < 0$, and the "visual-speed" difference $g_v'(x) - g_t'(T-x)$ is decreasing in T for $x \in (0,T)$, then the optimal visual share $s^\star(T)$ is strictly decreasing in T:

$$\frac{ds^\star}{dT} < 0.$$

For short attention windows, visuals deliver higher marginal returns ($g_v'(0) > g_t'(0)$), so $s^\star$ is high. As T grows, text's marginal returns decline more slowly, shifting the optimal allocation toward text.

(i) *Speed.* In short segments, visual signals should account for more partisan variation than text; the gap shrinks with T. (ii) *Mechanism.* If visual effectiveness loads on affect (higher β_v for emotion-rich content), then shocks to emotional tone generate larger changes in partisan predictions for images than for text. (iii) *Measurement.* Text-only bias measures understate slant when T is short; the understatement diminishes as T increases. (iv) *Supply.* In markets where typical T is short (e.g., social video), equilibrium $s^\star$ is higher, predicting a relative over-supply of emotionally charged visuals compared to long-form formats.