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

Andrea Ciccarone

Columbia University

4th AI + Economics Workshop Zurich
December 6, 2025

Motivation: Rethinking Partisan Media

- Partisan media shapes beliefs, policy preferences, and behaviors
  Martin and Yurukoglu, 2017; Djourelova, 2023; Ash et al., 2024
- So far, focus on text-based media slant – both in text-only and multimodal environments

Motivation: Rethinking Partisan Media

- Partisan media shapes beliefs, policy preferences, and behaviors
  Martin and Yurukoglu, 2017; Djourelova, 2023; Ash et al., 2024
- So far, focus on text-based media slant – both in text-only and multimodal environments
- Today most political content is consumed in **low-attention video environments**:
  - YouTube most used platform for news consumption in 2025 Pew Research (2025)
  - Share of U.S. adults regularly getting news on Tiktok from 3% to 20% in 5 years

Motivation: Rethinking Partisan Media

- Partisan media shapes beliefs, policy preferences, and behaviors
  Martin and Yurukoglu, 2017; Djourelova, 2023; Ash et al., 2024
- So far, focus on text-based media slant – both in text-only and multimodal environments
- Today most political content is consumed in **low-attention video environments**:
  - YouTube most used platform for news consumption in 2025 Pew Research (2025)
  - Share of U.S. adults regularly getting news on Tiktok from 3% to 20% in 5 years

Motivation: Rethinking Partisan Media

- Partisan media shapes beliefs, policy preferences, and behaviors
  Martin and Yurukoglu, 2017; Djourelova, 2023; Ash et al., 2024
- So far, focus on text-based media slant – both in text-only and multimodal environments
- Today most political content is consumed in **low-attention video environments**:
  - YouTube most used platform for news consumption in 2025 Pew Research (2025)
  - Share of U.S. adults regularly getting news on Tiktok from 3% to 20% in 5 years

What does partisanship look like in **low-attention + multimodal** environments?

What does this imply for partisan persuasion?

This Paper

Combination of ML framework + survey experiment

This Paper

Combination of ML framework + survey experiment

Intrinsic characteristics of image signal makes it more efficient in low-attention environments:

- Text signal is **slow and deliberative**: issue-based; takes time to accumulate
- Image signal is **fast and emotional**: based on emotional cues; conveys information quickly

This Paper

Combination of ML framework + survey experiment

Intrinsic characteristics of image signal makes it more efficient in low-attention environments:

- Text signal is **slow and deliberative**: issue-based; takes time to accumulate
- Image signal is **fast and emotional**: based on emotional cues; conveys information quickly

Effects on viewers: In low-attention video environments, the text channel of partisan persuasion breaks $\rightarrow$ main vehicle of partisan content is images affecting emotional responses

Plan for Today

1. **Measurement:** Scalable model to quantify partisanship from text & images in videos – trained on video political ads

2. **Empirical Properties:** Measure and compare partisanship in text and images in **news videos** + Characteristics of the signals

3. **Effects on Viewers:** Survey experiment using real immigration news clips to study effect on viewers varying partisanship at different exposure lengths

1. **Measurement:** Partisanship Model of Political Ads

Political ads and Partisanship

Idea:

- Use text and images in video political ads to create a multimodal measure of partisanship
  Gentzkow and Shapiro (2010)

Data:

- **Wesleyean Media Project:** universe of political ads aired on U.S. TV channels in (House, Senate and Presidential) elections from from 2016 to 2020
- 15,427 unique ads ($\sim$ 7.7k Dem, $\sim$ 7.7k Rep)

Political ads and Partisanship

Idea:

- Use text and images in video political ads to create a multimodal measure of partisanship
  Gentzkow and Shapiro (2010)

Data:

- **Wesleyean Media Project:** universe of political ads aired on U.S. TV channels in (House, Senate and Presidential) elections from from 2016 to 2020
- 15,427 unique ads ($\sim$ 7.7k Dem, $\sim$ 7.7k Rep)

Image and Text embeddings:

- To compare text and images: represent in the same vector (embedding) space via **CLIP**
- Use embeddings as inputs for Neural Network binary classification model (Rep vs Dem)
  Why CLIP? Representation Details

CLIP Representation of Images and Text - Political Ads

$\downarrow$

$x^1_{img} \in \mathbb{R}^{512}$

...

$\downarrow$

$x^t_{img} \in \mathbb{R}^{512}$

...

$\downarrow$

$x^N_{img} \in \mathbb{R}^{512}$

"Dad helped build this business from scratch..."

$\downarrow$

$x^1_{txt} \in \mathbb{R}^{512}$

...

"... protect Medicare and Social Security, especially with COVID..."

$\downarrow$

$x^t_{txt} \in \mathbb{R}^{512}$

...

"My friends, my family, my community, that's why I'm running for Congress..."

$\downarrow$

$x^N_{txt} \in \mathbb{R}^{512}$

CLIP Representation of Images and Text - Political Ads

$\downarrow$

$x^1_{img} \in \mathbb{R}^{512}$

$\downarrow$

$x^t_{img} \in \mathbb{R}^{512}$

$\downarrow$

$x^N_{img} \in \mathbb{R}^{512}$

...

...

...

"Dad helped build this business from scratch..."

$\downarrow$

$x^1_{txt} \in \mathbb{R}^{512}$

"... protect Medicare and Social Security, especially with COVID..."

$\downarrow$

$x^t_{txt} \in \mathbb{R}^{512}$

"My friends, my family, my community, that's why I'm running for Congress..."

$\downarrow$

$x^N_{txt} \in \mathbb{R}^{512}$

- Text: $\bar{x}_{txt} \rightarrow Pr(\text{Republican}|\bar{x}_{txt}) = ?$
- Image: $\bar{x}_{img} \rightarrow Pr(\text{Republican}|\bar{x}_{img}) = ?$

CLIP Representation of Images and Text - Political Ads

$\downarrow$

$x^1_{img} \in \mathbb{R}^{512}$

...

$\downarrow$

$x^t_{img} \in \mathbb{R}^{512}$

...

$\downarrow$

$x^N_{img} \in \mathbb{R}^{512}$

"Dad helped build this business from scratch..."

$\downarrow$

$x^1_{txt} \in \mathbb{R}^{512}$

...

"... protect Medicare and Social Security, especially with COVID..."

$\downarrow$

$x^t_{txt} \in \mathbb{R}^{512}$

...

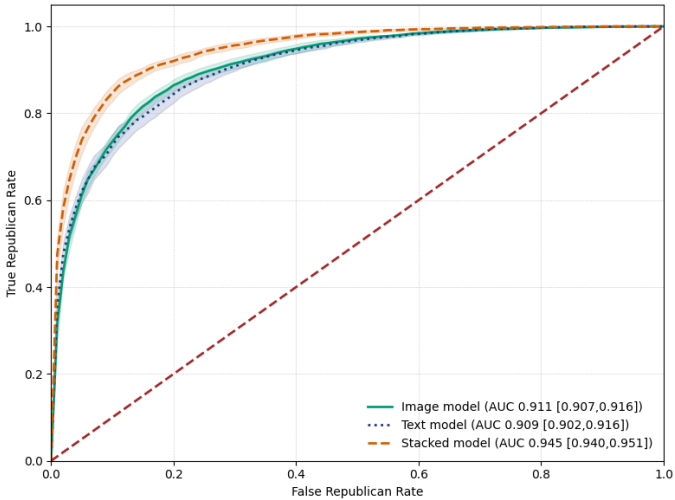
"My friends, my family, my community, that's why I'm running for Congress..."

$\downarrow$

$x^N_{txt} \in \mathbb{R}^{512}$

- Text: $\bar{x}_{txt} \rightarrow Pr(\text{Republican}|\bar{x}_{txt}) = ?$
- Image: $\bar{x}_{img} \rightarrow Pr(\text{Republican}|\bar{x}_{img}) = ?$
- Joint: $\rightarrow Pr(\text{Republican}|\bar{x}_{img}, \bar{x}_{txt}) = ?$

Images captures as much partisanship as text from ads...



...Image and Text Signals Do Not Perfectly Overlap

2. **Properties:** Transferring from Political Ads to Video News

From Political Ads to YouTube News Videos

Does this measure of slant apply across corpora?

From Political Ads to YouTube News Videos

Does this measure of slant apply across corpora?

Data: YT Videos from Fox, MSNBC, CNN + other outlets (YouTube uploads, 5–20 mins, focus on immigration) Sample Summary

Approach:

- Process each segment into image (frames) + text (transcripts) embeddings Preprocessing
- Aggregate embeddings for each video and feed into ad-trained models *as is*
- Obtain predicted probability that the video is Republican (Democratic)

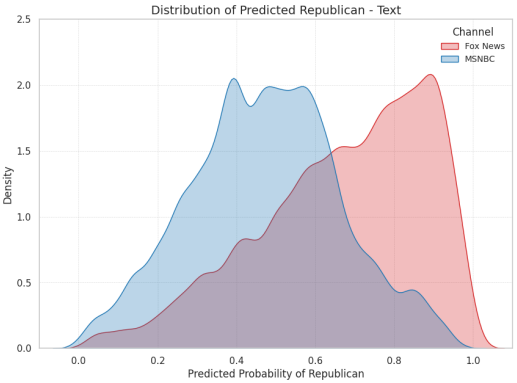
Example of image-based partisanship predictions

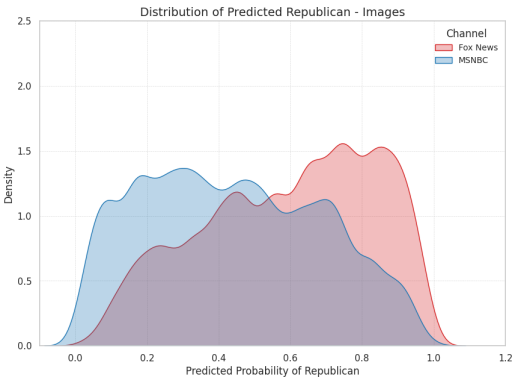
P(Rep) = 0.01

P(Rep) = 0.99

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.

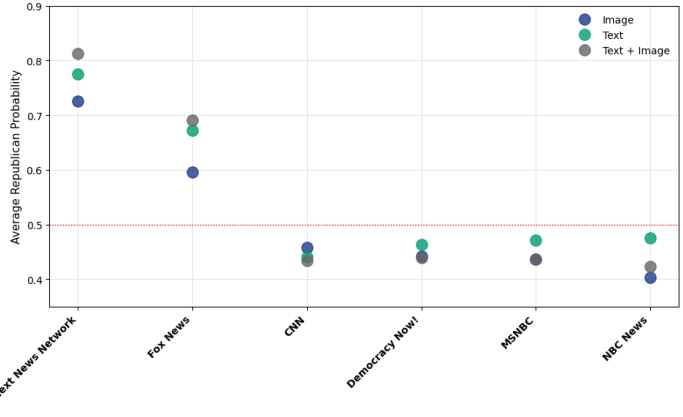
Distribution of Predictions - FNC and MSNBC





All channels

Partisanship Model Transfer into YouTube News Channels



Text to Multi-Modal $\rightarrow$ FNC-MSNBC partisan gap increase by 5.5 p.ps $\sim$ 20% of benchmark

Image to Multi-modal $\rightarrow$ FNC-MSNBC partisan gap increase by $\sim$ 40% of benchmark

Attention: “An Image is Worth 40.38 Words”

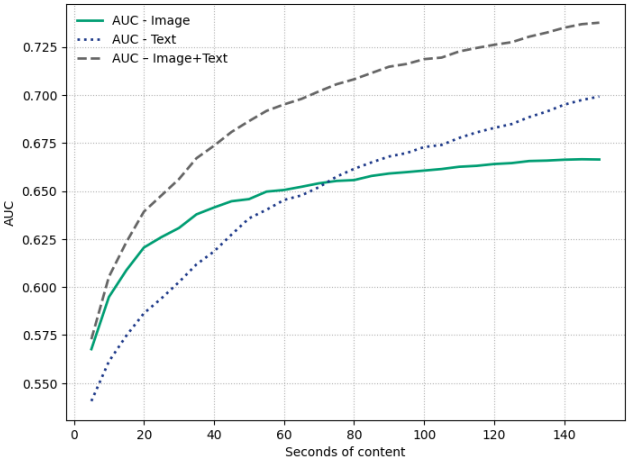
Attention, or Speed of Information Transfer

- **Question:** How images vs. text convey partisan cues under inattention?

Attention, or Speed of Information Transfer

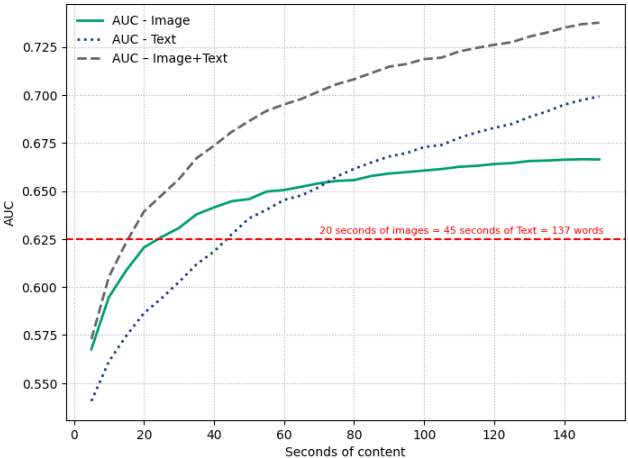
- **Question:** How images vs. text convey partisan cues under inattention?
- **Strategy:** Details
  - Break segments into 5s "chunks" (1 frame + ∼15 words)
  - Randomly sample K chunks, average embeddings, predict partisanship
  - Vary K:
    - $K = 1$: inattentive viewer (5s glimpse)
    - Large K: sustained viewing
- **Key idea:** Compare how fast AUC rises across modalities → quantify advantage of images under low attention

The Speed of Information Transfer, or "An Image is Worth 40.38 Words"



Notes: In short lenghts, 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.

The Speed of Information Transfer, or "An Image is Worth 40.38 Words"



Notes: In short lenghts, 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.

Emotions: Opening the Black Box of the Partisanship Model

Opening the Black Box: Emotions vs. Topics

- Image signal is quicker, but are images just “quick text”?

Opening the Black Box: Emotions vs. Topics

- Image signal is quicker, but are images just "quick text"?
- Embedding representation allows us to study **concepts** captured in text/image
  1. Images and text represent different groups of *concepts*
  2. Partisan signal from text and images rely on different *concepts*

Opening the Black Box: Emotions vs. Topics

- Image signal is quicker, but are images just "quick text"?
- Embedding representation allows us to study **concepts** captured in text/image
  1. Images and text represent different groups of *concepts*
  2. Partisan signal from text and images rely on different *concepts*
- **Approach (SpLICE):** Bhalla et al. (2024) Details
  1. Define 2 concept vocabularies: "emotions" and "topics" Emotions Topics
  2. Map video embeddings onto these vocabulary $\rightarrow$ text & images as combinations of concepts

Intuitively, we are representing videos (image and text) as linear combinations of emotions and/or topics

SpLICE Decomposition on Joint Vocabulary

$$\longrightarrow \mathbf{x}^{\text{img}} = \begin{pmatrix} x_1^{\text{img}} \\ \vdots \\ x_K^{\text{img}} \end{pmatrix}$$

Joint Vocabulary

{fear, anger..., immigration}

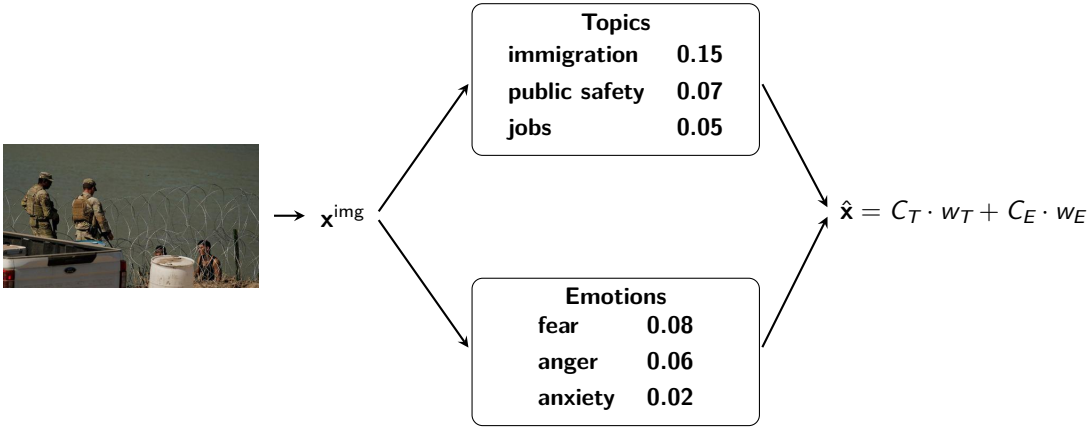
$$\longrightarrow \mathbf{C} = \begin{bmatrix} c_1^1 & c_1^2 & \cdots \\ \vdots & \ddots & \\ c_K^1 & & c_K^C \end{bmatrix}$$

Sparse Decomposition

$$\longrightarrow \mathbf{w}^{\text{img}} = \begin{cases} \text{immigration} = 0.15 \\ \text{fear} = 0.08 \\ \vdots \\ \text{calm} = 0 \end{cases}$$

Details | Emotions: Fox News vs MSNBC | Topics: Fox News vs MSNBC

1. What happens to the representation if we remove a concept?



1. What happens to the representation if we remove a concept?

1. What happens to the representation if we remove topics?

$\rightarrow \mathbf{x}^{\text{img}}$

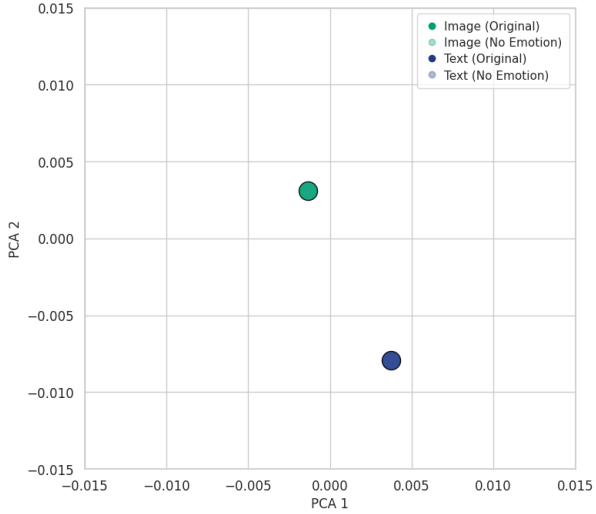
Topics	
immigration	0.15
public safety	0.07
jobs	0.05

Emotions	
fear	0.00
anger	0.06
anxiety	0.02

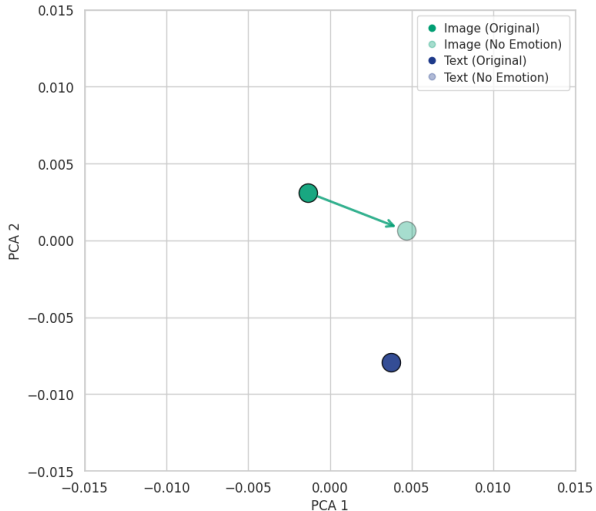
$$\hat{\mathbf{x}}_{-f} = C_T \cdot w_T + C_E \cdot w_{E\_f}$$

1. What happens to the representation if we remove a concept?
$\rightarrow$ We can compare $\hat{\mathbf{x}}_{-f}$ and $\hat{\mathbf{x}}$

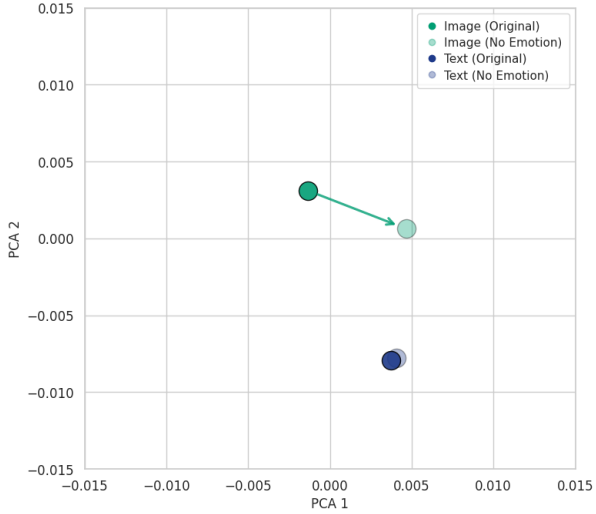
1. What happens if we remove a concept? Removing “Fear”



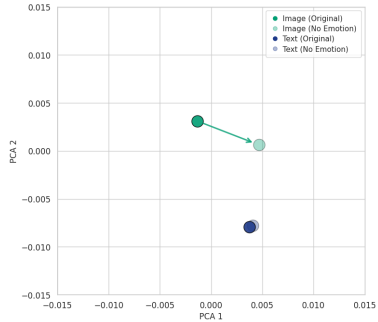
1. What happens if we remove a concept? Removing “Fear”



1. What happens if we remove a concept? Removing “Fear”



1b. What happens if we remove a concept? Removing "Fear"



→ On average, removing a single emotion concept from images shift the representation by 0.014 in cosine distance $\simeq 9.6^\circ$ rotation in the embedding space

→ For text, representation shifts by 0.006 $\simeq 6.3^\circ$ rotation

→ Images are slightly more responsive to topic shifts, but magnitude is comparable (13° vs 12°)

Explaining Partisanship: Concept Treatment Effect

$\rightarrow$ Decomposition allows to study which concepts are used in the partisanship model

Explaining Partisanship: Concept Treatment Effect

$\rightarrow$ Decomposition allows to study which concepts are used in the partisanship model

- Take $\hat{\mathbf{x}}$ and no-fear counterfactual $\hat{\mathbf{x}}_{-f}$
- Take partisanship model m^{img}
- $m^{\text{img}}(\hat{\mathbf{x}}) - m^{\text{img}}(\hat{\mathbf{x}}_{-f}) \rightarrow$ effect of adding/removing "fear" on predicted partisanship

Explaining Partisanship: Concept Treatment Effect

$\rightarrow$ Decomposition allows to study which concepts are used in the partisanship model

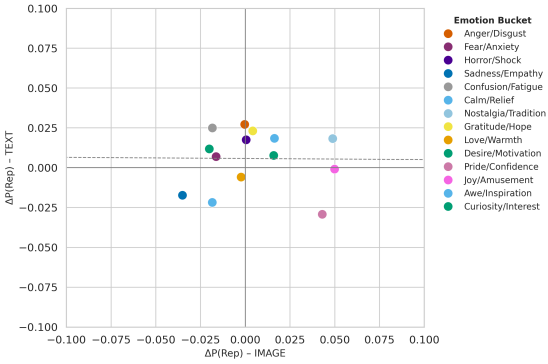
- Take $\hat{\mathbf{x}}$ and no-fear counterfactual $\hat{\mathbf{x}}_{-f}$
- Take partisanship model m^{img}
- $m^{\text{img}}(\hat{\mathbf{x}}) - m^{\text{img}}(\hat{\mathbf{x}}_{-f}) \rightarrow$ effect of adding/removing "fear" on predicted partisanship

Concept Treatment Effect:

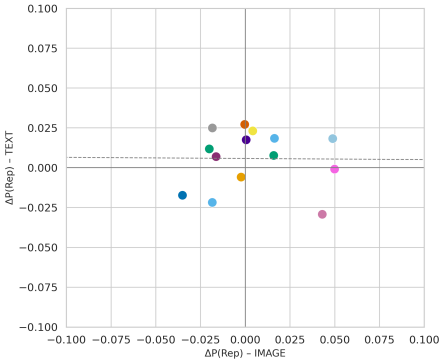
$$\Delta_c^{\text{mod}} = \frac{1}{|I_c|} \sum_{i \in I_c} \left[m^{\text{mod}}(\hat{\mathbf{x}}_i) - m^{\text{mod}}(\hat{\mathbf{x}}_{i,-c}) \right]$$

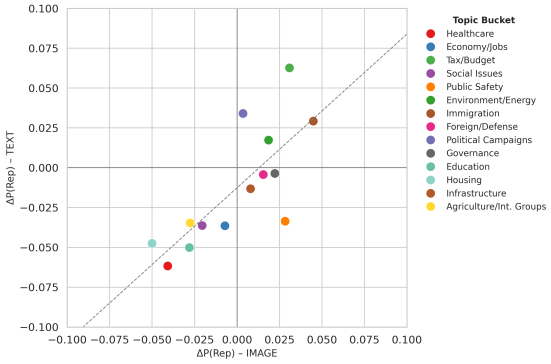
A positive Δ_c indicates that the presence of c raises the predicted Republican probability

2. Does image partisan signal rely more on emotions?

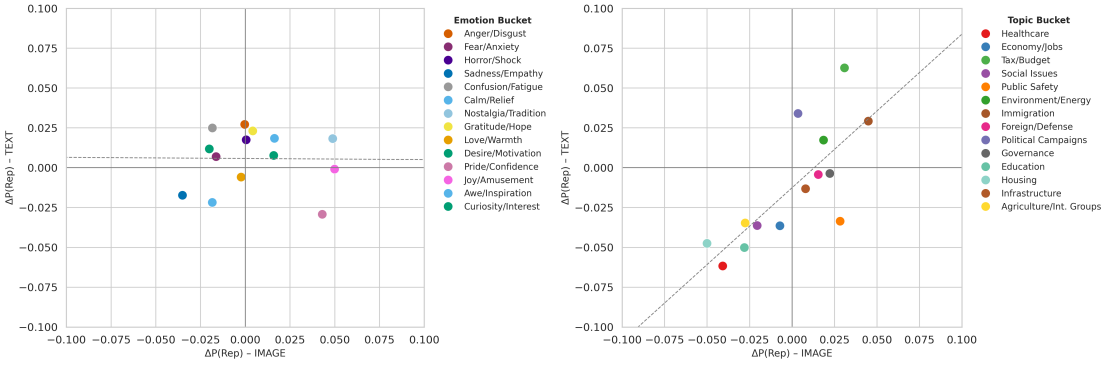


2. Does image partisan signal rely more on emotions?





2. Does image partisan signal rely more on emotions?



→ Avg. absolute **emotions** CTE is 0.021 for image and 0.016 for text; $r = -0.01$

→ Avg. absolute **topics** CTE is 0.025 for image and 0.033 for text; $r = 0.76$

What do we Learn?

Two distinctive properties of images relative to text in videos:

1. **Attention**: Images convey partisan information quickly; more resilient to inattention
2. **Emotions**: Emotions more represented in images and have more partisan predictive power

What do we Learn?

Two distinctive properties of images relative to text in videos:

1. **Attention**: Images convey partisan information quickly; more resilient to inattention
2. **Emotions**: Emotions more represented in images and have more partisan predictive power

Do these properties have real implications for how different modalities affect viewers?

Do they translate into different effects on persuasion?

3. **Effects:** Evidence from a Survey Experiment

Experimental Design: Treatment Setup[1]

- **Context:** $\sim$ 2 mins TV-news segments on the *April 3, 2024* hearing on Texas SB 4 Event Selection
- **Design:** $(2 \times 2 \times 2)$ Treatment Arms
  - Visual partisanship: *Republican* vs. *Democratic* images
  - Text partisanship: *Republican* vs. *Democratic* text
  - Short and Long versions
- **Treatment:**
  - *Channels:* MSNBC and Fox News morning coverage of the hearing.
  - *Text:* Transcript $\rightarrow$ remove cues that reveal channel or identity $\rightarrow$ re-dub
  - *Images:* Retain only on-scene immigration footage; drop studio/host shots and presenter close-ups.

Treatment T | Treatment I | Partisanship T | Partisanship I

[1] The design is largely based on Afrouzi et al., 2023

Experimental Design: Treatment Setup[1]

- **Context:** $\sim$ 2 mins TV-news segments on the *April 3, 2024* hearing on Texas SB 4 Event Selection
- **Design:** $(2 \times 2 \times 2)$ Treatment Arms
  - Visual partisanship: *Republican* vs. *Democratic* images
  - Text partisanship: *Republican* vs. *Democratic* text
  - Short and Long versions
- **Treatment:**
  - *Channels:* MSNBC and Fox News morning coverage of the hearing.
  - *Text:* Transcript $\rightarrow$ remove cues that reveal channel or identity $\rightarrow$ re-dub
  - *Images:* Retain only on-scene immigration footage; drop studio/host shots and presenter close-ups.

Treatment T | Treatment I | Partisanship T | Partisanship I

[1]The design is largely based on Afrouzi et al., 2023

	Text	
Image	Republican	Democratic
Republican	RR	RD
Democratic	DR	DD

2 Lengths:

- Short version $\sim$ 30 seconds: introduction of the report
- Long version $\sim$ 120 seconds: introduction + actual report

Main Specification

I recruit $\sim$ 4000 participants randomly assigned to each of 8 treatment groups (3430 after sample restrictions).

I standardize each outcome and estimate (separately for Long and Short):

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

Robustness: Modality Misalignment

Main Specification

I recruit $\sim$ 4000 participants randomly assigned to each of 8 treatment groups (3430 after sample restrictions).

I standardize each outcome and estimate (separately for Long and Short):

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

Robustness: Modality Misalignment

Consider first this set of outcomes:

1. **Immigration topic attitudes** (eg. "Do you favor allowing U.S. military to assist immigration raids?")
2. **Emotions**: ask to rate emotions (eg. fear, anger, disgust) Davoine et al. (2025)

1. Text (Slowly) Affects Attitudes; Images (Quickly) Affect Emotions

(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

1. Text (Slowly) Affects Attitudes; Images (Quickly) Affect Emotions

(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

$\Rightarrow$ Republican text shifts anti-immigration attitudes in long clips by $\sim$ 0.07 s.d.

Stronger text effects closer to the issue (e.g. 0.17 s.d. on border patrol) Border Patrol

1. Text (Slowly) Affects Attitudes; Images (Quickly) Affect Emotions

(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

$\Rightarrow$ Republican text shifts anti-immigration attitudes in long clips by ~ 0.07 s.d.
Stronger text effects closer to the issue (e.g. 0.17 s.d. on border patrol) Border Patrol

$\Rightarrow$ Republican images shift negative emotions, especially in short clips (~ 0.19 s.d.)

2. Not Just Emotions: Image Effect on Charity Choice

Back

Charity choice: "If you win the lottery, you can choose a prize of $25 dollars or you can donate a portion of this amount..." to pro-immigration charity

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

$\Rightarrow$ Republican images in short decrease probability of donating by $\sim$ 0.1 standard deviations

$\Rightarrow$ Negative effect on charity comes mostly from Republicans Charity Heterogeneity

3. Modalities and Perceived Partisanship

"On the U.S. political spectrum, how would you classify the video?"

3. Modalities and Perceived Partisanship

"On the U.S. political spectrum, how would you classify the video?"

Republican Partisanship		
	Long	Short
Image (Rep)	0.036***	0.030***
	(0.009)	(0.009)
Text (Rep)	0.004	-0.011
	(0.009)	(0.009)
Obs.	1475	1414

$\Rightarrow$ Republican images increase perceived Republican slant by $\sim$ 0.3 - 0.4 standard deviations

$\Rightarrow$ Images don't lose predictive power in short exposure

Lessons from the Survey Experiment

The key properties of text and images have important implications for political communication:

1. **Visuals act fast:** Images immediately shift perceptions and emotions, and in short clips affect likelihood of charity donation
2. **Text persuades slow:** Text moves topic attitudes, but only in long clips
3. **Division of labor:** Images drive *emotions & behavior*; text drives *policy attitudes*
4. **Partisan Heterogeneity:** Short emotional effects are concentrated among Republicans; long attitudinal shifts come mainly from Democrats

Conclusions

Conclusions

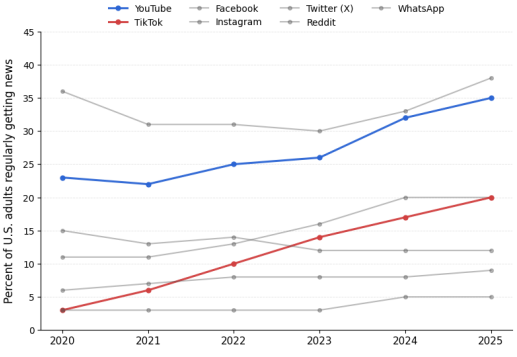
- Implication $\rightarrow$ shift to short-form video changes the nature of political persuasion: filters out text effect on beliefs, leaving only the image effect on emotional responses
- Text-only measures are not equipped to study modern media environment
- Policies focusing only on text miss a central channel of political communication

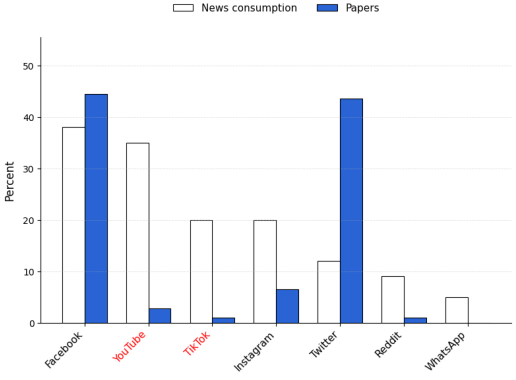
In a low-attention video-first world, it is mostly about images

Appendix

A New (and Understudied) Media Environment

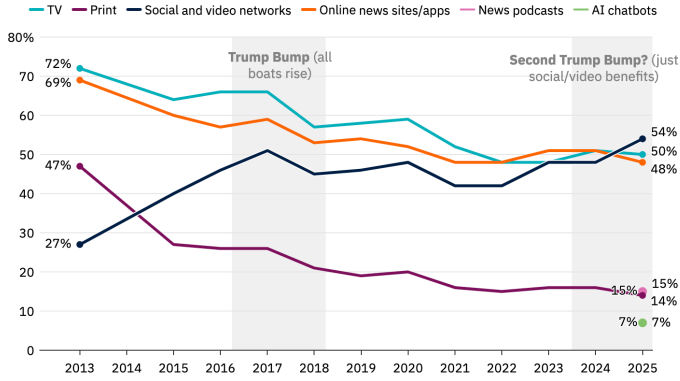
Back





Reuters Institute, Source of News in the Last Week

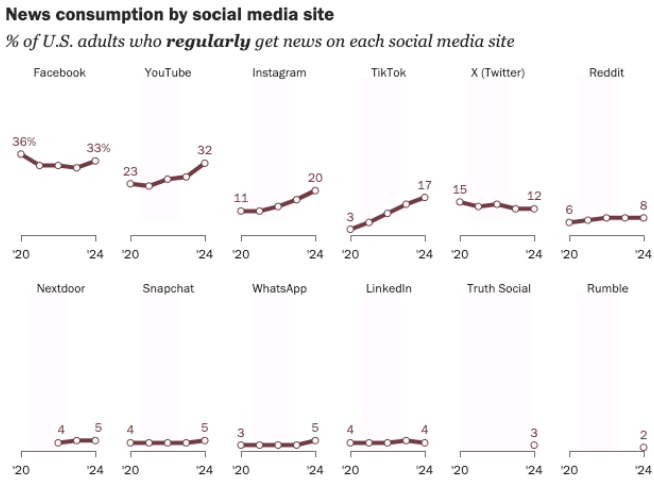
Back



Q3. Which, if any, of the following have you used in the last week as a source of news? *Base: Total sample in each year ≈ 2000. Note: No data for 2014. There was a sampling and weighting change from 2021 onwards.*

Pew Research, Social Media and News

Back



Why CLIP? Back

- **CLIP** (Radford et. al, 2021) was released by OpenAI and is a state of the art ML model for multimodality
- Pre-trained to predict if an image and a text snippet are paired together in its dataset (400 millions image-text pairs)
- Embeddings are created to maximize similiarity of image and text embeddings of the real pairs while minimizing the similarity of the incorrect pairings
- For each ad:
  1. Average embeddings over frames and transcript segments
  2. Normalize + mean-center to control for modality gap

CLIP Representation Details

Back

- Maps images + text into a shared 512-dim space (ViT-B/32). Vision Transformer, patch 32px + transformer text encoder.
- Ad-level embedding: average over frames/snippets (linear representation hypothesis).

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

1. Normalize embeddings to unit sphere.
2. Subtract modality-specific mean $\boldsymbol{\mu}_{\text{mod}}$.
3. Re-normalize:

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

Partisanship Classification Model

Back

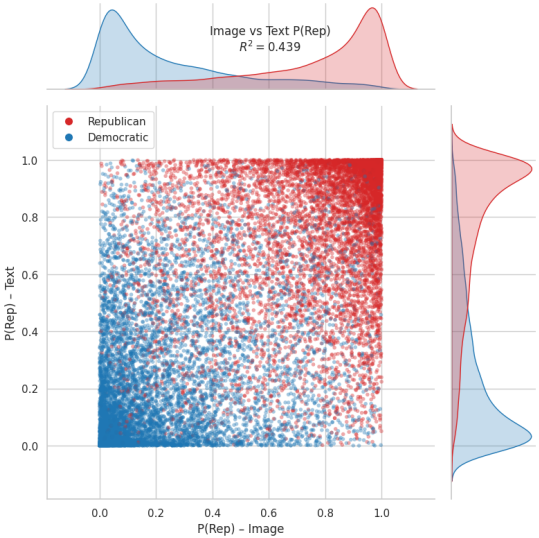
- Task: binary classification (R vs. D) at ad level.
- Inputs: adjusted embeddings $\tilde{\mathbf{x}}^{\text{img}}$, $\tilde{\mathbf{x}}^{\text{txt}}$.
- Models:
  1. Image-only
  2. Text-only
  3. Pooled (stacked predictions)
- Dataset: 80/20 train–test split, stratified by party.

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

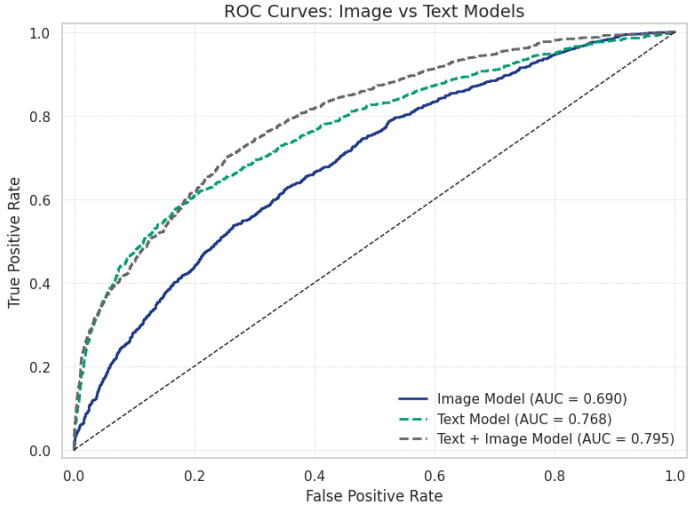
- $\sigma(\cdot)$: logistic function; $(\alpha_0, \alpha_1, \alpha_2)$ estimated on training data.
- Captures complementary partisan signals across modalities.

...Image and Text Signals Do Not Perfectly Overlap

Back



Are we Underestimating Partisanship? Back

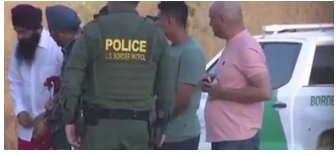


Notes: ROC curves for image-only, text-only, and pooled models at the video level. This implies etc etc...

YouTube News: What should we use? Back

- YouTube News "images" are intrinsically, complex:

- Fix a topic (immigration) and extract only relevant immigration scenes:



Crop & clustering strategy Back

Step 1: Extracting Sub-Images

- Use edge detection and Hough transform to identify dominant structural lines
- Segment frames into sub-images based on detected lines
- Retain the largest sub-images

Step 2: Image Embedding

- Use CLIP to generate feature embeddings
- Normalize and preprocess embeddings for clustering

Step 3: Clustering

- Apply K-means clustering with $k = 30$ to group similar images
- Identify cluster centroids and rank images based on proximity

Step 4: Selecting Relevant Clusters

- Identify clusters corresponding to immigration-related content
- Filter out irrelevant clusters
- Save filtered images and embeddings for downstream analysis

YouTube News Sample Summary

Back

Channel	# Videos	Avg Words	Avg Images	Avg Views
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 News	955	1,092.9	91.1	350,153
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**	**1,409**	**103.0**	**146,805**

Random Sampling of Embeddings

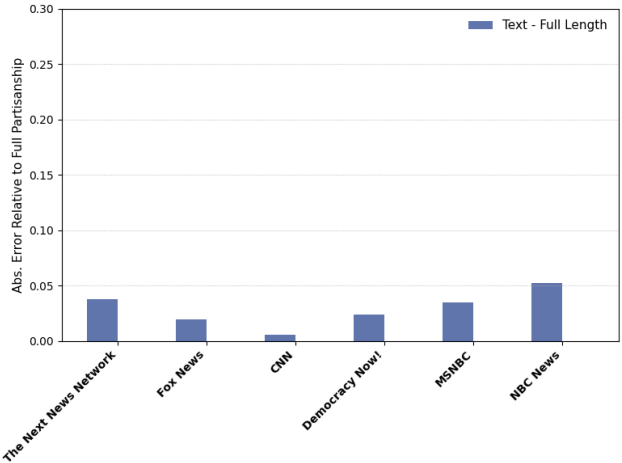
Back

Strategy: Fix a number of chunks C. 1 YouTube News chunk is 5 seconds, and is represented by 1 image and ~ 11.33 words

1. **Embeddings per Video:**
$$\mathcal{X}(v) = \{\mathbf{x}_1, \ldots, \mathbf{x}_{n_v}\}, \quad \mathcal{T}(v) = \{\mathbf{t}_1, \ldots, \mathbf{t}_{n_v}\}.$$
(Image embeddings in $\mathcal{X}(v)$, text embeddings in $\mathcal{T}(v)$.)
2. **Random Subset:** For each video v, sample
$$\Omega_v^X \subset \{1, \ldots, n_v\}, \quad |\Omega_v^X| = C,$$
similarly for text with Ω_v^T.
3. **Average Embeddings:**
$$\widehat{\mathbf{x}}(v) = \frac{1}{|\Omega_v^X|} \sum_{i \in \Omega_v^X} \mathbf{x}_i, \quad \widehat{\mathbf{t}}(v) = \frac{1}{|\Omega_v^T|} \sum_{j \in \Omega_v^T} \mathbf{t}_j.$$
4. **Partisanship Estimation:** For each video, we feed $\widehat{\mathbf{x}}(v)$ and $\widehat{\mathbf{t}}(v)$ in the classification model and obtain partisanship estimates.

Outcome: We repeat this sampling M times,so each video ends with M distinct predictions for the "averaged" image embeddings and text embeddings.

Text Only : Are we Underestimating Partisanship? Back

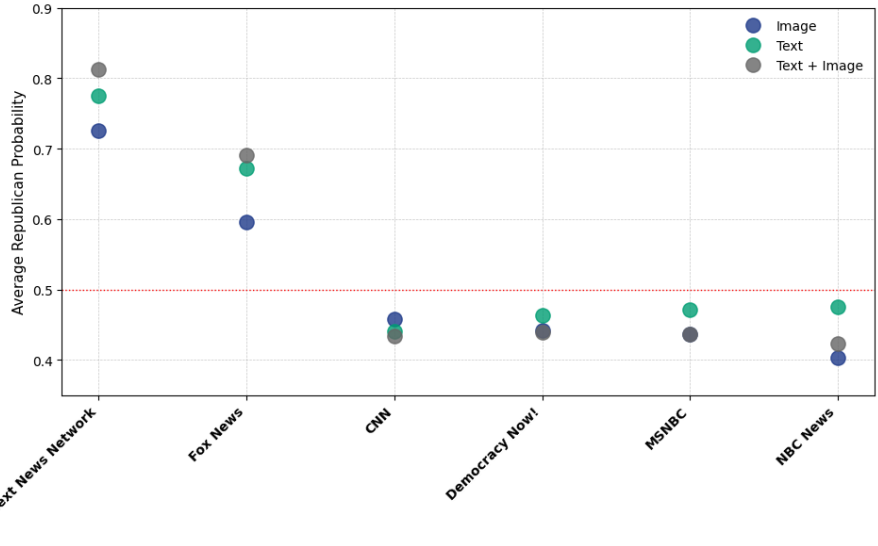


FNC: 0.67 $\rightarrow$ 0.69; MSNBC: 0.47 $\rightarrow$ 0.44 $\implies$ partisan gap increases by $\sim$ 5.5 percentage points

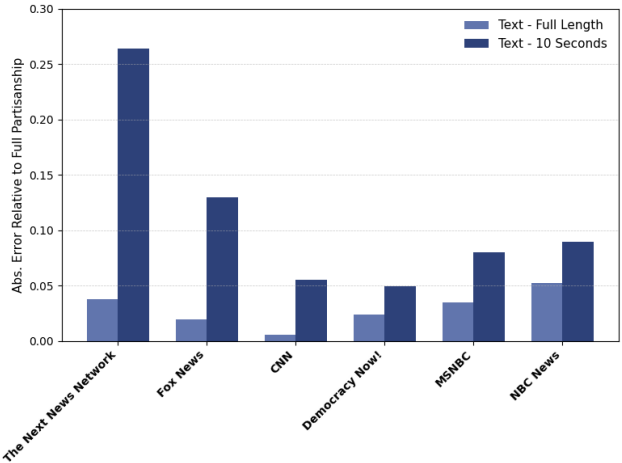
$\rightarrow$ This is under no information loss on more than 10 minutes of content... 10 Second Predictions

Partisanship Model Transfer into YouTube News Channels

Back



Text Only : Are we Underestimating Partisanship? Back

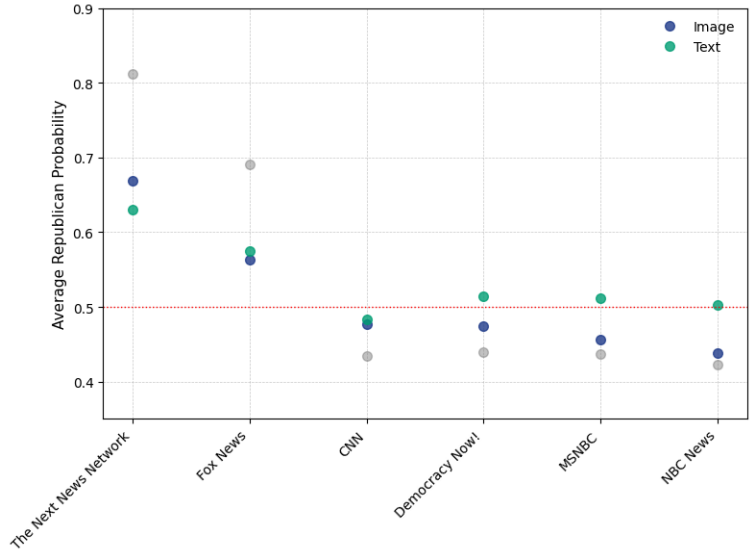


With 10 secs of text, partisan gap underestimated by $\sim$ 15 points (75% of total partisanship!)

When attention is low or content it short $\rightarrow$ **images!** 10 Second - Images

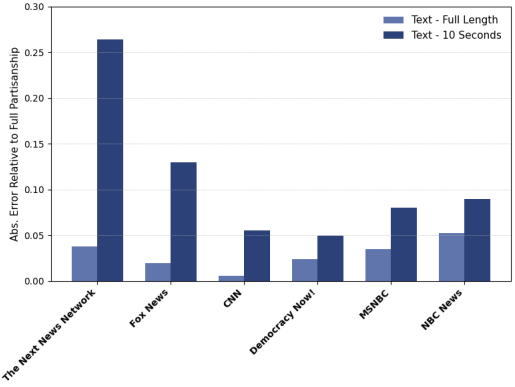
10 Seconds Predictions of Partisanship

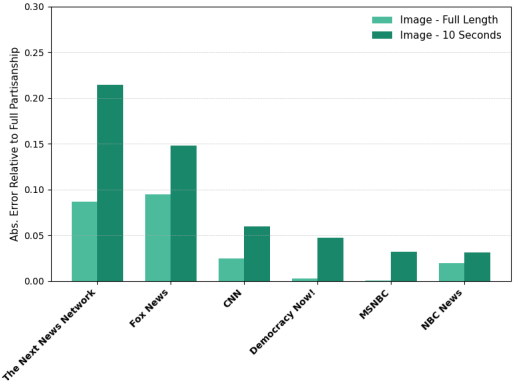
Back



Measuring Partisanship under Inattention

Back





Varying Chunk Size and Computing AUC

Back

Startegy: We now let $C \in \{1, \ldots, 50\}$ and observe how performance (AUC) changes with the number of sampled embeddings per video.

1. **Chunk Size Variation:** For each C we get probabilities $\{p_{\text{img}}(v), p_{\text{txt}}(v)\}$
2. **Align with Labels:** We assume partisanship labels $y(v)$ - MSNBC is Dem (0), FNC is Rep (1)
3. **Compute AUC:**

$$\text{AUC}_{\text{img}}(C) = \text{AUC}\Big(\{p_{\text{img}}(v)\}, \{y(v)\}\Big), \quad \text{AUC}_{\text{txt}}(C) = \text{AUC}\Big(\{p_{\text{txt}}(v)\}, \{y(v)\}\Big).$$

4. **Plot and Compare:** Finally, plot

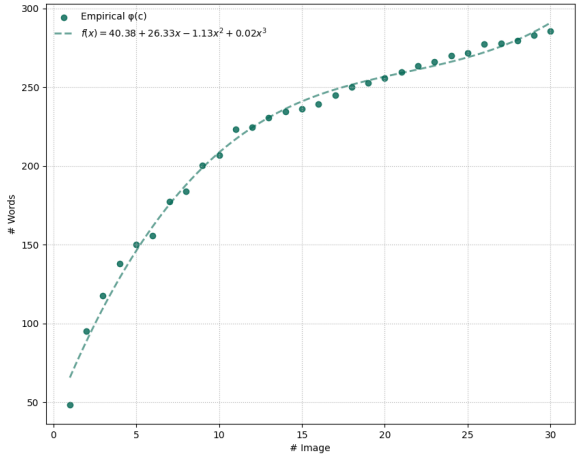
$$\text{AUC}_{\text{img}}(C) \quad \text{and} \quad \text{AUC}_{\text{txt}}(C)$$

vs. C for $C = 1, \ldots, 50$. This reveals the impact of chunk size on performance.

Conclusion: This allows us to compare how much partisanship information we get from image vs text at each chunk size C

The Speed of Information Transfer, or "An Image is Worth 40.38 Words"

Back



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.

Sparse Encoding and Representation

Back **Sparse Encoding:** Bhalla et al., 2024

$$\text{Image Representation: } \mathbf{w} = \arg\min_{\mathbf{w} \geq 0} \|\mathbf{C}\mathbf{w} - \mathbf{x}\|_2^2 + 2\lambda\|\mathbf{w}\|_1$$

- $\mathbf{x}$: Embedding of the image $\mathbf{x}^{\text{img}}$
- $\mathbf{C}$: Concept vocabulary matrix $\rightarrow$ 1st decompositon on Emotions; 2nd on Topics
- $\mathbf{w}$: Sparse weights representing the image in emotion space
- λ: Sparsity parameter

Once we have the optimal $\mathbf{w}^*$ we can reconstruct the image as $\hat{\mathbf{x}}^{\text{img}} = \mathbf{C}\mathbf{w}^*$
This allow us to define image **CTE** as:

$$\Delta_{\text{fear}}^{\text{img}} = f\left(\hat{\mathbf{x}}^{\text{img}}\right) - f\left(\hat{\mathbf{x}}_{-\text{fear}}^{\text{img}}\right)$$

importantly this is on the **extensive margin!**

SpLiCE Representation and Interventions

This approach allows us to interpret our model in terms of concepts:

- Suppose img has positive weight on the "illegal immigration" concept, i.e. $\mathbf{w}^*(\text{illegal immigration}) > 0$
- We can define $\mathbf{w}^*_{-\text{illegal immigration}}$ as:

$$\mathbf{w}^*_{-\text{illegal immigration}}(c) = \begin{cases} \mathbf{w}^*(c), & \text{for } c \neq \text{illegal immigration} \\ 0, & \text{for } c = \text{illegal immigration} \end{cases}$$

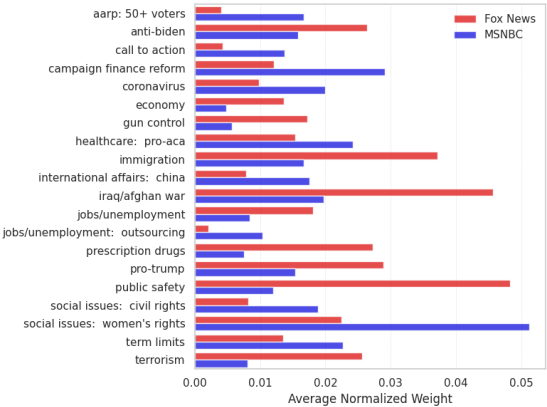
- And we can reconstruct $\hat{\mathbf{x}}^{\text{img}}_{-\text{illegal immigration}} = \mathbf{C}\mathbf{w}^*_{-\text{illegal immigration}}$
- Given our partisanship prediction model, the **Concept Treatment Effect** on Predicted Partisanship of "including" the illegal immigration concept is:

$$\Delta^{\text{img}}_{\text{illegal immigration}} = f\left(\hat{\mathbf{x}}^{\text{img}}\right) - f\left(\hat{\mathbf{x}}^{\text{img}}_{-\text{illegal immigration}}\right)$$

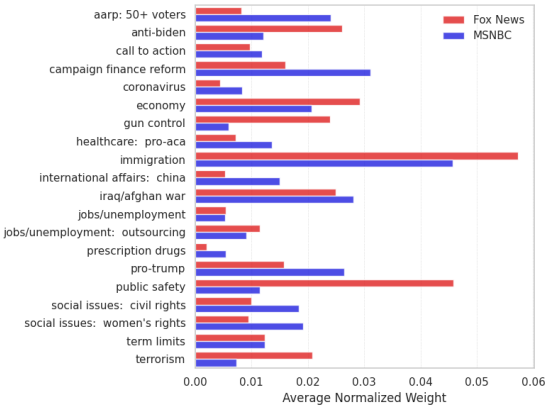
Back

SpLICE Topics: Fox News vs MSNBC

Back



Average Weights in Images



Average Weights in Text

SpLICE Emotions: Fox News vs MSNBC

Back

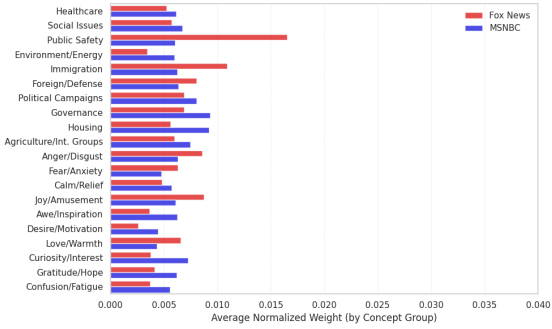


Figure: Average Weights in Images

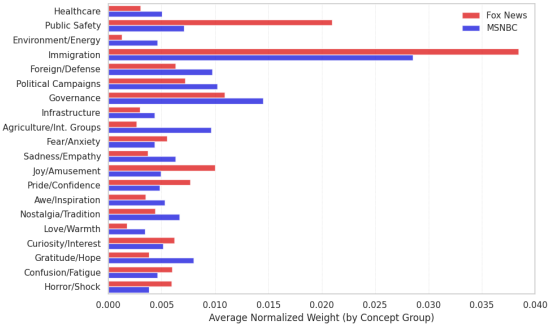


Figure: Average Weights in Text

Emotion Vocabulary (27 buckets, 81 tokens)

Back

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*		

Note. Each bucket groups semantically related tokens; weights are computed per token and aggregated at the bucket level for interpretability. The 27 buckets are adapted from Cowen & Keltner (2017).

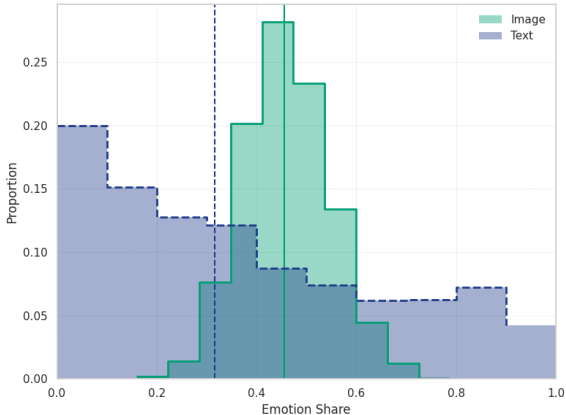
Topic Vocabulary (14 buckets, 82 tokens) Back

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	*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*	Finance Housing	*housing/home ownership*
Infrastructure	*transportation, telecommunications*	Agriculture/Interest Groups	*food/agriculture, aarp: 50+ voters, aarp mention*

Note. Each bucket groups semantically related tokens; weights are computed per token and aggregated at the bucket level for interpretability.

1a. Do images represent more emotions? Back

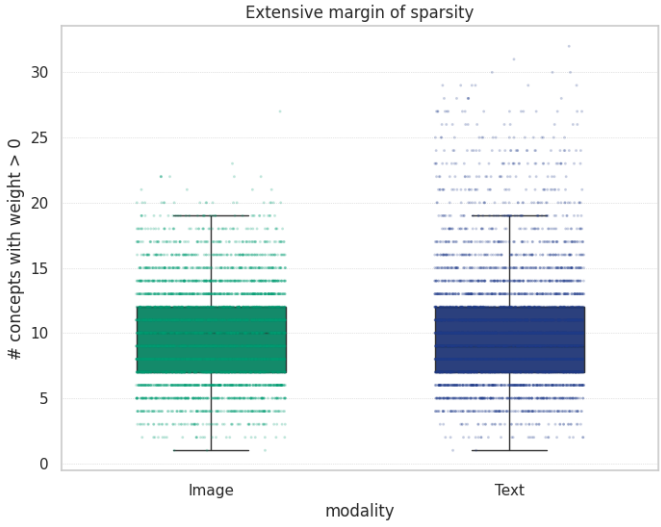
- Comparing the **relative load** of text vs images on emotions versus topics



→ Images are 50/50 mixtures of topics and emotions; text is mostly topics-only

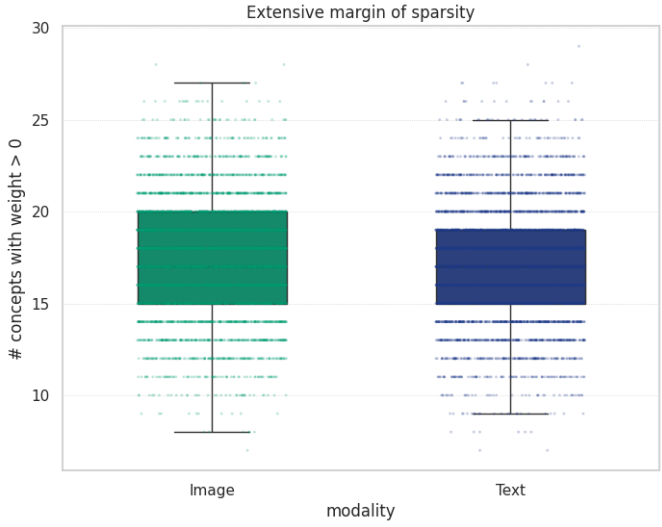
Joint Vocabulary: Sparsity Distribution

Back



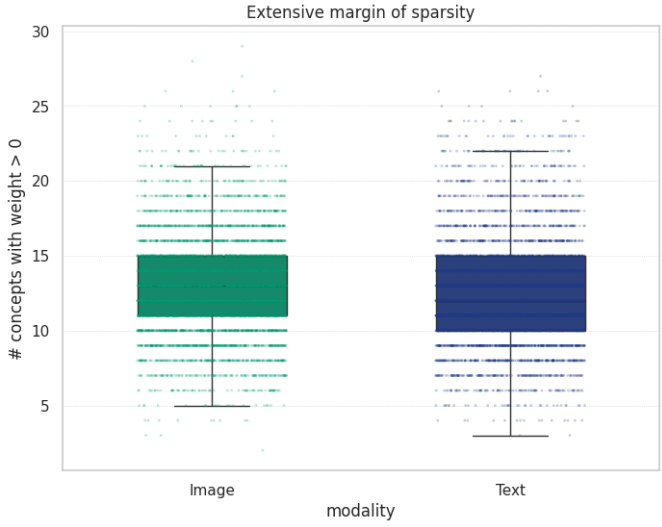
Emotions: Sparsity Distribution

Back



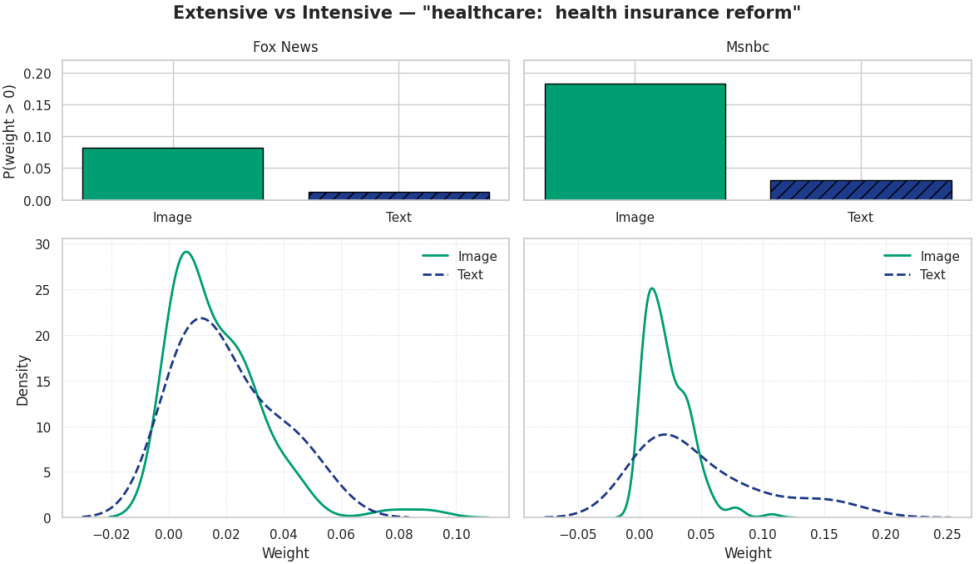
Topics: Sparsity Distribution

Back

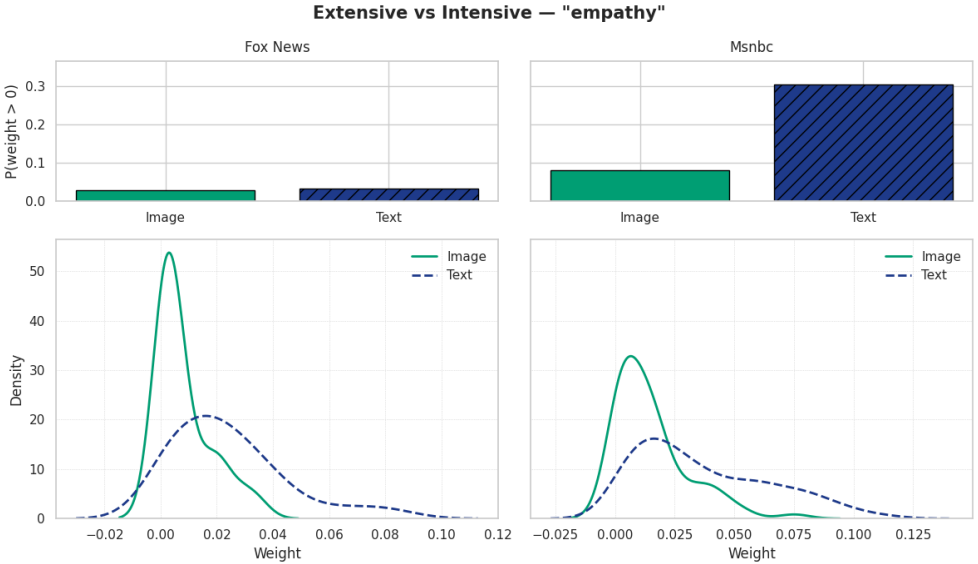


Example Topic: “Health Insurance Reform”

Back



Example Emotion: “Empathy’ Back

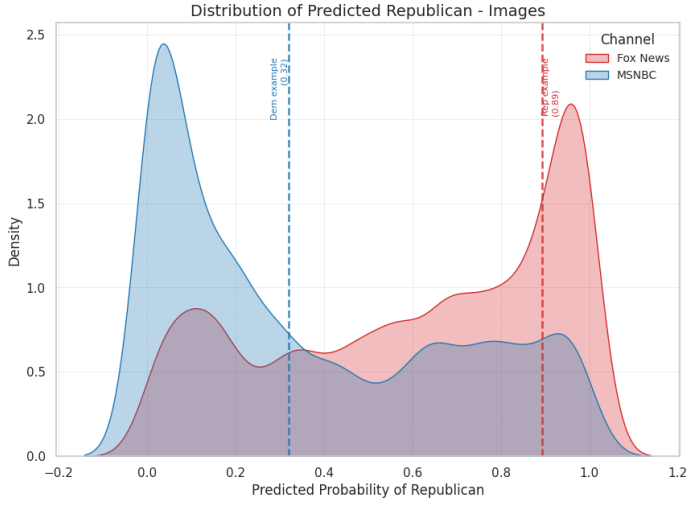


Event Selection: Texas SB4 Hearing (April 3, 2024) Back

- Focused on MSNBC and Fox News coverage of immigration.
- Selected pairs of transcript snippets that:
  1. contained immigration-related terms,
  2. lay at opposite ends of the text-based partisanship distribution,
  3. were similar in embedding representation.
- Manual inspection $\rightarrow$ chose April 3, 2024 appellate hearing on Texas Senate Bill 4 (SB4).
- SB4: law authorizing state arrests/deportations of border crossers (blocked by lower court; argued before Fifth Circuit).
- Two segments used:
  - Fox News *America's Newsroom*, 11 a.m.
  - MSNBC *José Díaz-Balart Reports*, 8 a.m.
- Structure of segments:
  - $\sim$30s host introduction $\rightarrow$ short treatment
  - $\sim$90s reporter explanation $\rightarrow$ added to form long treatment

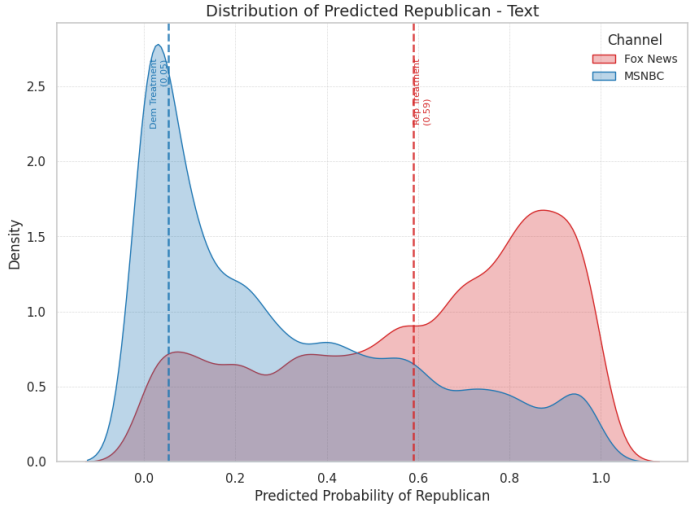
Image Treatment Partisanship

Back



Text Treatment Partisanship

Back



Experimental Design: Image Treatment

Back

Video: Democratic Images (MSNBC)

Video: Republican Images (Fox News)

Experimental Design: Text Treatment

Back

Video: Democratic Text (MSNBC)

"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 [...] 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."

Video: Republican Text (Fox News)

"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 [...] 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."

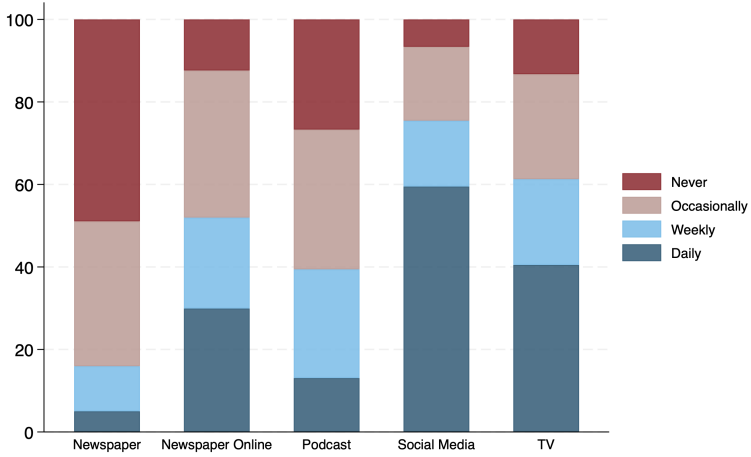
Balance Across Treatment Groups

Back

	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			

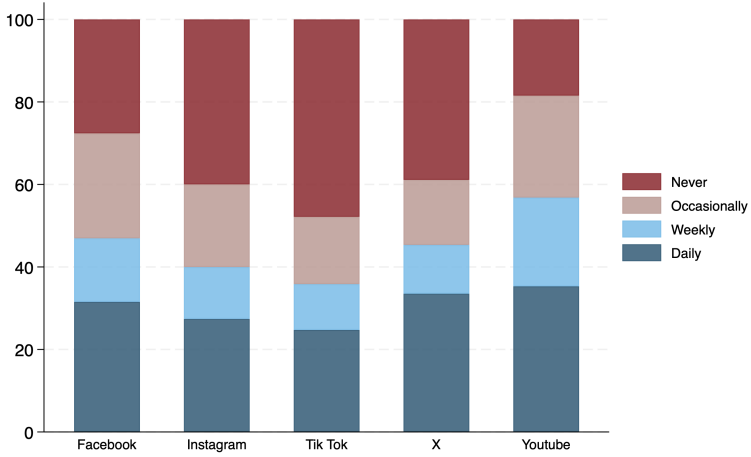
Media consumption in the survey

Back



Media consumption in the survey - Social Media

Back



Partisanship

Back

Republican Partisanship						
	Full Sample		Republicans		Democrats	
	Long	Short	Long	Short	Long	Short
Image (Rep)	0.201***	0.210***	0.050	0.207***	0.317***	0.220***
	(0.047)	(0.048)	(0.069)	(0.065)	(0.063)	(0.072)
Text (Rep)	0.072	-0.029	-0.026	-0.205***	0.168***	0.138*
	(0.047)	(0.049)	(0.068)	(0.066)	(0.064)	(0.071)
Obs.	1748	1682	788	869	960	813

Increase Border Patrol

Back

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

Anger

Back

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

Disgust

Back

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

Modality Misalignment: RD vs DR

Back

(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

Heterogeneity by Party: Anti-Immigration (RD vs DR)

	Republicans		**Democrats**	
	Long	Short	Long	Short
RD vs DR	0.008	0.020	-0.180***	0.013
	(0.055)	(0.047)	(0.060)	(0.063)
Observations	402	420	472	411

Heterogeneity by Party: Charity Choice (RD vs DR)

	Republicans		**Democrats**	
	Long	Short	Long	Short
RD vs DR	-0.034	-0.154**	0.114	-0.067
	(0.073)	(0.074)	(0.090)	(0.102)
Observations	402	420	472	411

Negative Emotions (RD vs DR)

	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)	(0.017)
Observations	874	831	402	420	472	411

3. Not Just Emotions: Image Effect on Charity Choice

Back

Charity choice: "If you win the lottery, you can choose a prize of $25 dollars or you can donate a portion of this amount..." to pro-immigration charity

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

$\Rightarrow$ Republican images in short decrease probability of donating by $\sim$ 0.1 standard deviations

$\Rightarrow$ Negative effect on charity comes mostly from Republicans Charity Heterogeneity

3. Heterogeneity by Party: Charity Choice

Back

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

$\Rightarrow$ Negative effect on charity comes mostly from Republicans