8. Supplementary

8.1. MTurk Study

Participants for the study were from the pool of US resident users of Amazon MTurk as described in Section 5. Participants were shown a series of pages starting with an instructions page that outlined the process. They were told that the task would involve watching a series of videos and would take approximately 10 minutes. They were then given the following specific instructions for how the task would proceed:

1. Read a short prompt about the video,
2. Watch the video,
3. Answer questions about the video and your impressions

They were finally instructed that they must enable their audio and answer all questions in order to complete the task and receive payment. Participants were not informed about the number of videos they would view, their subject matter, or that some of the videos were edited using an automated system.

Each participant was shown a total of 6 videos, one per topic in order of "Republican National Convention", "Democratic National Convention, "Covid", "USPS", "Hurricane", and "Protests". Each topic had two corresponding storylines that were chosen at random for each topic. The storylines correspond to the voice-over audio of the human edited videos (Table 1).

Within each topic, the viewer was randomly shown either the human edited video condition or the CMVE video condition. Every possible condition (Topic, version number, and editing condition) was tested at least once in the study.

For each video, the participant was shown three separate pages. The first page contained an instruction "Please read the following description and click "View Video" to watch the video and the text of the storyline for that video.

On the second page, the video and audio played automatically. No video controls were displayed for the participant. Below the video was an instruction to click "Next" once the video had finished. This button only appeared after the video concluded, and the participant was informed they would not be able to return to this screen.

At the top of the final page for each video was the text storyline, shown again for the viewer. Below were the questions described in Section 8.2. Questions 1 and 2 were presented as a 1-10 slider, with 5 set as the default. The remaining questions were presented as a set of 5 options, of which they could only choose one. Those options were "Strongly Disagree", "Disagree", "Undecided", "Agree", and "Strongly Agree".

Participants were able to proceed at their own pace, and when they clicked next, they were taken to the first page of the next video. A brief demographic survey was collected at the end of the study, which included the following questions:

1. Age
2. Gender
3. Approximately [how] many hours per week of print news content (e.g. newspaper, online news articles) do you consume?
4. Approximately how many hours per week of video news content (e.g. TV, online video) do you consume?

8.2. Questions

The complete list of questions posed to participants on MTurk, in order:

1. Please rate the overall quality of the video.
2. Please rate the overall quality of the audio.
3. I would expect to see the video on a news channel.
4. The video matches the text content.
5. The video matches the audio content.
6. The video was informative.
7. The video was interesting.
8. All clips in the video were relevant to the text.

8.3. Algorithms

Below we provide algorithms `RankClips`, `SplitVideos` and `EntityVector`. These outline the overall procedure for ranking a set of video clips $\{C^{(m)}\}$, for splitting a video $V^{(i)}$ into clips, and for computing $s_{\mathbf{e}}$, the entity score for a clip.

Topic	Version	Storyline
Republican National Convention	1	Vice President Mike Pence forcefully defended law enforcement during a speech to the Republican National Convention. He did not mention Black Americans killed by police, or the recent shooting of Jacob Blake in Kenosha, Wisconsin.
	2	The Republican Party has formally nominated President Donald Trump for second term, kicking off the opening day of the GOP convention. Trump questioned the integrity of the election and took issue with mail-in voting, which experts say has proven remarkably safe, during his acceptance speech.
Democratic National Convention	1	California Senator Kamala Harris formally became the Democratic Party's vice presidential nominee on Wednesday night. She addressed the virtual Democratic Convention from Wilmington, Delaware, speaking about the nation's long struggle for racial justice and condemning President Trump's leadership.
	2	Former presidential nominee Hillary Clinton addressed the virtual Democratic Convention, saying she wished Donald Trump were a better president. Former President Barack Obama said Trump won't grow into the job of president, quote, 'because he can't.'
COVID-19	1	Texas has now confirmed more than ten thousand coronavirus deaths. About 80 percent of them have occurred since June 1st, when the state was weeks into one of the country's fastest reopenings.
	2	Democratic vice presidential nominee Kamala Harris plans a Thursday speech condemning President Donald Trump for his handling of the coronavirus pandemic, and she'll give it just hours before Trump is set to accept renomination for a second term. Joe Biden's campaign tells the associated press that Harris will detail a 'profound failure of leadership' from Trump and highlight Biden's proposals to control the virus and confront the economic fallout.
USPS	1	Postmaster General Louis DeJoy told lawmakers that he has warned allies of President Trump that the president's repeated attacks on mail-in ballots are 'not helpful,' but denied that changes at the Postal Service are tied to the November elections. Democrats fear they're being done to harm Joe Biden's chances in November.
	2	Facing public backlash, Postmaster General Louis DeJoy is testifying that it's his 'sacred duty' to ensure election mail delivery. A Senate committee is digging into service changes he made ahead of the November election, just as millions of Americans are expected to vote by mail.
Hurricane	1	Tropical Storm Laura is now expected to become a major hurricane. Laura is forecast to make landfall Wednesday night or Thursday near the Texas-Louisiana border.
	2	Hurricane Laura made landfall in southwest Louisiana early Thursday morning as an 'extremely dangerous' Category Four storm. The storm made landfall with top sustained winds of 150 miles an hour, and a massive storm surge is also a major concern.
Protests	1	Protesters marched, drove and honked horns in Kenosha, where things were reported to be peaceful. There were no groups patrolling with long guns, unlike Tuesday night, when two people were shot to death.
	2	Kenosha, Wisconsin, has become the nation's latest flashpoint for racial unrest after a 29 year old black man was shot by police numerous times – apparently in the back – as he leaned into his SUV. Three officers at the scene have been placed on administrative leave – pending an investigation.

Table 1. **MTurk Study Storylines** - Storylines per topic, per version for each video in the study. These correspond to the audio voice-over that was identical for both the CMVE- and Human-edited conditions of each video.

Characteristic	N = 80[1]
Age	42.8 (22-71)
Unknown	4
Gender	
Male	50 (68%)
Female	23 (32%)
Unknown	7
Print News Consumed (hours/week)	4.9 (1-35)
Unknown	2
Video News Consumed (hours/week)	5.5 (1-35)
Unknown	2

Table 2. Participant demographics. Statistics presented: Mean (Minimum-Maximum); n (%)

Question	CMVE Mean (SE)	Human Mean (SE)
1	7.55 (0.10)	7.60 (0.09)
2	7.51 (0.10)	7.74 (0.09)
3	3.87 (0.05)	4.00 (0.04)
4	3.89 (0.06)	4.14 (0.05)
5	3.92 (0.06)	4.11 (0.04)
6	3.90 (0.05)	3.99 (0.05)
7	3.88 (0.06)	3.95 (0.05)
8	4.00 (0.06)	4.11 (0.05)

Table 3. Response mean (SE) from MTurk study. Note questions 1 and 2 were on the 1-10 scale and questions 3-8 were on a 1-5, Likert scale.

Algorithm 1: `RankClips`

Input: Input text $w^{(\mathcal{I})}$, set of videos $\mathcal{D}$ set of constraints $\mathcal{I}_c$, reference video score vector $\mathbf{s_w}$
Output: Ranked list of clips N

$T_1 = \mathcal{M}_K(\mathcal{D})$ `// Keyword search, Eq. 2`
$T_2 = \mathcal{M}_E(T_1)$ `// Rank by Universal Sentence Encoder embedding, Eq. 3`
$\{C^{(m)}\}$ = **SplitVideos**$(T_2, \mathcal{I}_c)$ `// Algorithm 2`
initialize N, empty list of clips
for *sentence* $w_j^{(\mathcal{I})} \in w^{\mathcal{I}}$ **do**
    compute minimum clip length L for $w_j^{(\mathcal{I})}$
    initialize C_j; empty list of clips for sentence j
    for $C^{(i)} \in \{C^{(m)}\}$ **do**
        if *duration*$(C^{(i)}) > L$ **then**
            add $C^{(i)}$ to C_j
    `// Eliminate clips that do not meet duration threshold.`
    $E_j^{(\mathcal{I})}$ is a list of entities recognized in $w_j^{(\mathcal{I})}$
    initialize S_E, empty list of entity scores
    $\mathbf{e_w}$, reference video entity dependency weight vector
    for $C^{(i)} \in C_j$ **do**
        with $E^{(i)}$, entity metadata for clip $C^{(i)}$
        $\mathbf{e_c}^{(i)}$ = **EntityVector**$(C^{(i)}, E^{(i)}, E_j^{(\mathcal{I})})$ `// Algorithm 3`
        $s_{\mathbf{e}}^{(i)} = \mathbf{dot}(\mathbf{e_w}, \mathbf{e_c}^{(i)})$
        Add $s_{\mathbf{e}}^{(i)}$ to S_E
    `// Calculate entity score for each clip`
    $C^{(R)} = \arg\max_C \left(s_{\mathbf{e}}^{(i)} \in S_E\right)$
    `// Select as reference clip the clip with the highest entity score`
    remove $C^{(R)}$ from C_j
    for $C^{(i)} \in C_j$ **do**
        calculate $s_{\mathbf{p}}^{(i)}$, $s_{\mathbf{b}}^{(i)}$, $s_{\mathbf{u}}^{(i)}$ using Eq(5, 6), Eq(7), Eq(8) respectively
    $s^{(i)} = \mathbf{s_w} \cdot < s_{\mathbf{p}}^{(i)}, s_{\mathbf{b}}^{(i)}, s_{\mathbf{u}}^{(i)}, s_{\mathbf{e}}^{(i)} >$
    T_3 = **rank** C_j by $s^{(i)}$, descending
    $N_j = \{C^{(R)}\} \cup T_3$
    add N_j to N
return N

Algorithm 2: `SplitVideos`

Input: Set of videos $V^{(i)}$, a set of input constraints $\mathcal{I}_c$
Output: $\{\mathrm{C}^{(m)}\}$, set of clips

$n = |V^{(i)}|$
for $i = 0$ ***to*** n **do**
    $\{C_1^{(1)}, \ldots, C_j^{(1)}\}$ = **split** $V^{(i)}$; **where** video i has j clips
    drop clips by **minimum confidence threshold** and **minimum shot length** from $\mathcal{I}_c$
return $\{\{C_1^{(1)}, \ldots, C_a^{(1)}\} \cup \{C_1^{(2)}, \ldots, C_b^{(2)}\} \cup \cdots \cup \{C_1^{(i)}, \ldots, C_z^{(i)}\}\}$
`// Number of clips per video may vary`

Algorithm 3: `EntityVector`

Input: clip C, entity information for the clip E_c, entity information for a sentence E_s
Output: $s_\mathbf{e}$, score vector for clip C

initialize vector $s_\mathbf{e}$
for *entity* $e_i \in E_s$ **do**
    $count$ = **count**$(e_i \in E_c)$
    with dependency tag d_i for e_i
    add $count$ to vector $s_\mathbf{e}$ at index corresponding to d_i
return s_e

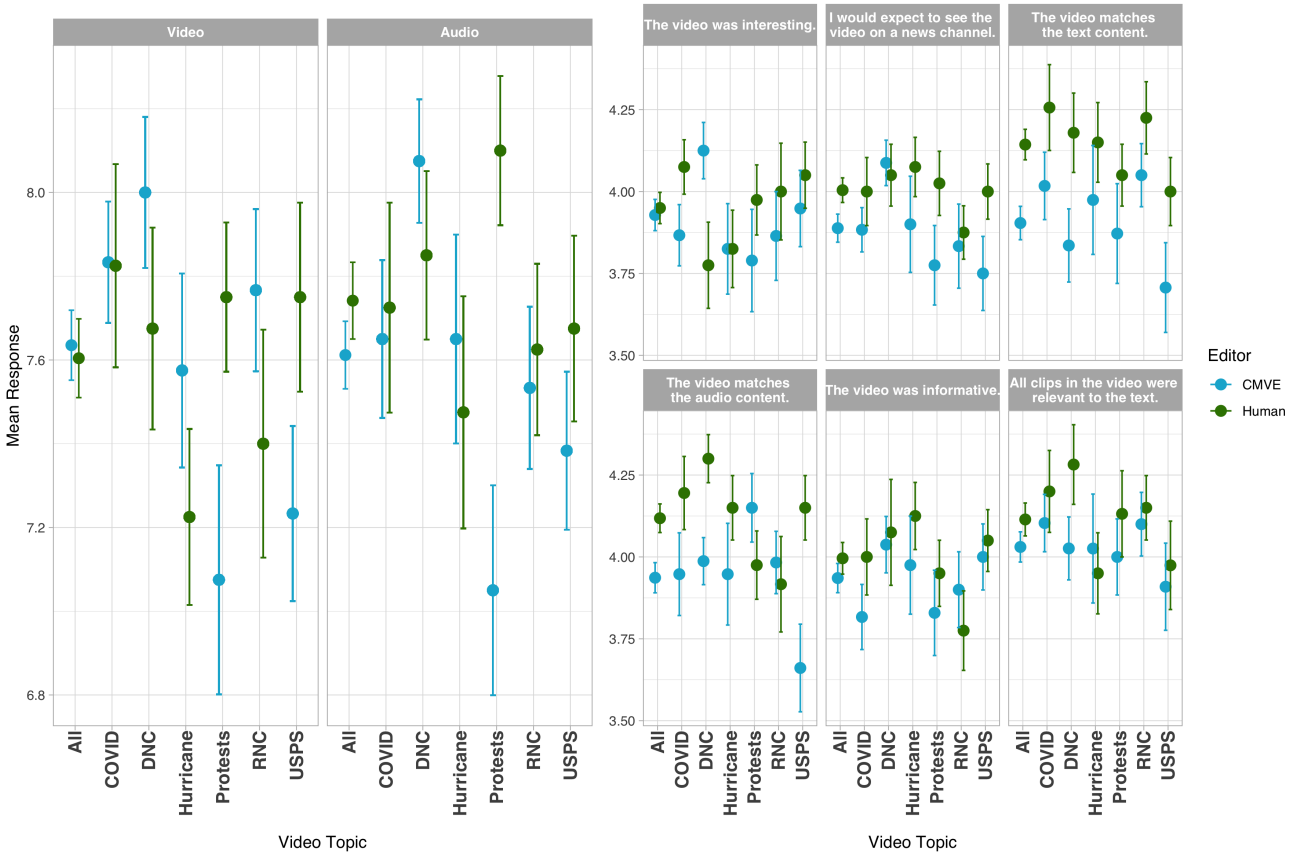

Figure 7. **Question-Specific Results** - from the MTurk study. In total, participants were asked 8 questions. Questions 1 and 2 were rating quality, and were scaled 1-10. The rest were posed on the Likert scale.

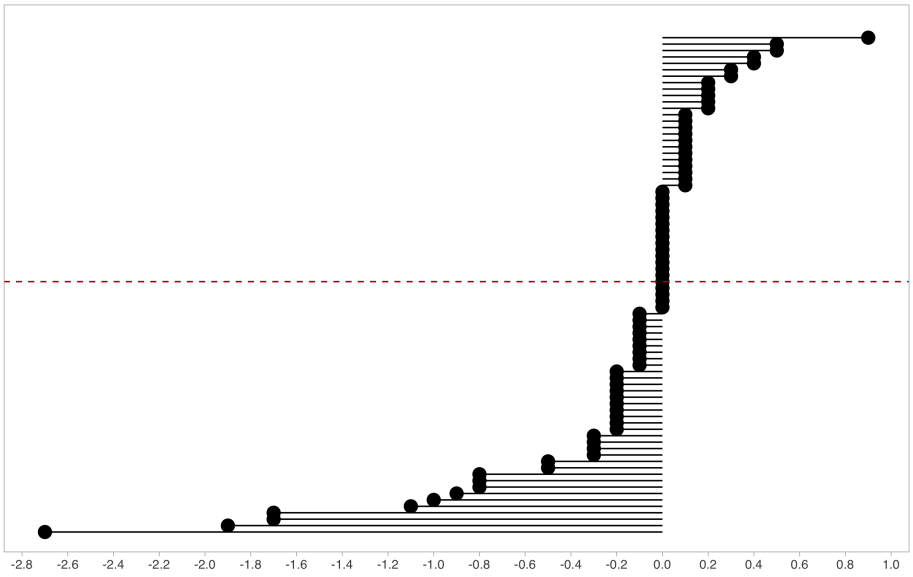

Figure 8. **Subject Preferences** - Preference for CMVE edits was computed as average score, across all questions, as CMVE minus human edited. Subjects with a greater average preference for human edits are negative on the X-axis. Each participant is represented by one line and dot.

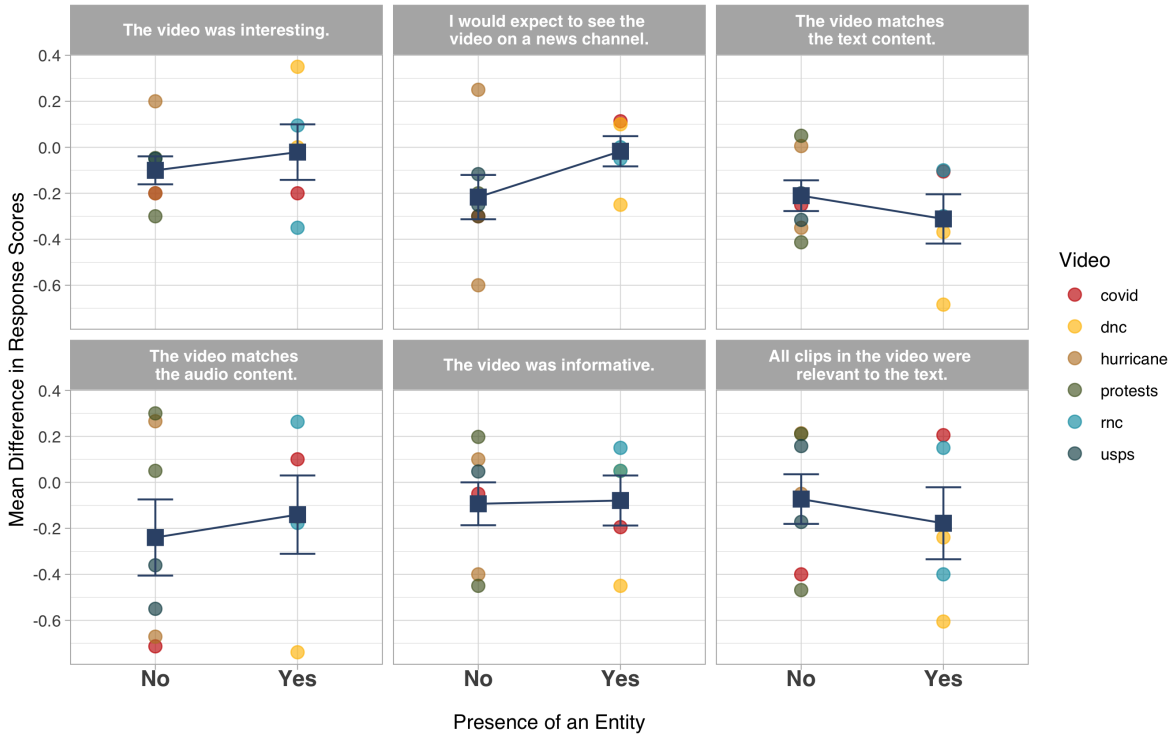

Figure 9. **Entity vs. Edit Preference** - Results from relevant questions, grouped by whether there was an entity present in the input query, and it exists at least one selected clip ("Yes"). If there was no known entity detected in by the object detection model, the video was classified as "No." The mean difference in response scores were computed as CMVE minus human edited ratings (preference for human professional edits is positive on the X-axis).

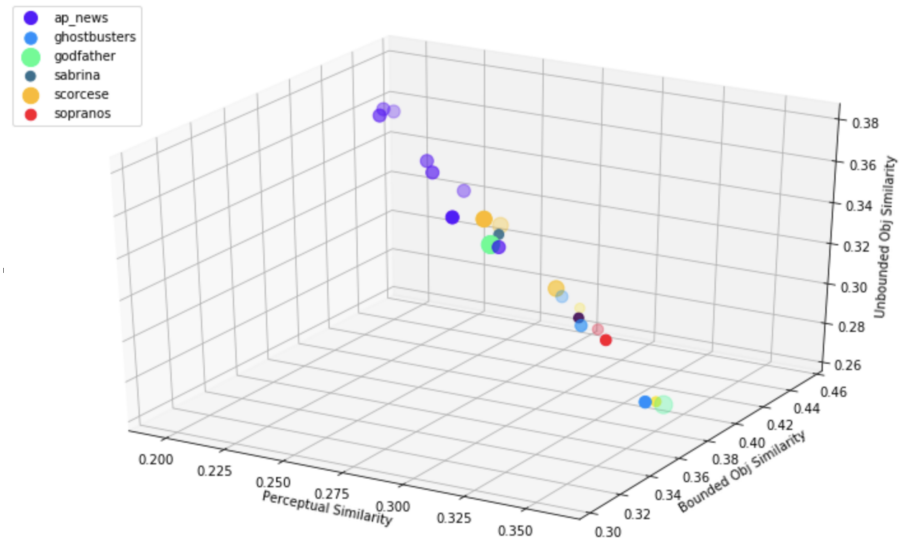

Figure 10. **Similarity Measures** - clustering of NC dataset (purple) when compared to test samples of videos (approximately 1 minute in duration) from other genres. Each dot represents a video comprised of a reference clip and at least 1 other clip. The similarity measures were averaged across non-reference clips.

Global:	
$w^{(\mathcal{I})}$	input text query, possibly comprised of multiple sentences
$\mathcal{I}_f$	input film rules (e.g. transitions)
$\mathcal{I}_c$	input constraints (e.g. preferred duration of clips)
$\mathcal{D}$	collection of videos
$w^{(i)}$	human-annotated text description attributed to video i in $\mathcal{D}$
$\|x\|$	size of variable x
Indices:	
$w^{(\mathcal{I})}$	number of sentences ($j \in \{1, ..., j\}$)
$\mathcal{D}$	number of videos ($i \in \{1, ..., i\}$)
N	number of bounded objects ($n \in \{1, ..., N\}$)
Y	number of unbounded objects ($i \in \{1, ..., Y\}$)
P	number of dependency tags ($\|P\| \in \{1, ..., p\}$)
S	number of score classes (s $\in \{1, ..., S\}$)
C	number of clips ($m \in \{1, ..., m\}$)
$C^{(R)}$	reference clip
Models:	
$\mathcal{M}_E$	sentence encoding model returning list of videos ranked by distance between two text descriptions,
$\mathcal{M}_C$	shot (i.e. clip) segmentation model that splits a video into clips,
$\mathcal{M}_O$	object detector and recognizer used for extracting bounded and unbounded object labels from clips,
$\mathcal{M}_S$	video editor model that stitches together clips for a sentence and applies transitions defined in $\mathcal{I}_f$
Scores:	
$s_{\mathbf{p}}^{(m)}$	training clip perceptual similarity,
$s_{\mathbf{b}}^{(m)}$	training clip bounded object similarity,
$s_{\mathbf{u}}^{(m)}$	training clip unbounded object similarity,
$e_{\mathbf{c}}$	average entity counts, by dependency tag, across sentences,
$s_{\mathbf{e}}^{(m)}$	clip entity overlap,
$\mathbf{s_w}$	score weight vector derived from training sample, softmax of $< s_{\mathbf{p}}^{(m)}, s_{\mathbf{b}}^{(m)}, s_{\mathbf{u}}^{(m)}, s_{\mathbf{e}}^{(m)} >$
$s^{(m)}$	clip score, computed as the dot product of $\mathbf{s_w}$ and score vector $< s_{\mathbf{p}}^{(m)}, s_{\mathbf{b}}^{(m)}, s_{\mathbf{u}}^{(m)}, s_{\mathbf{e}}^{(m)} >$

Table 4. Selected Notation