Supplementary File
PALADIN: Understanding Video Intentions in Political Advertisement Videos

Hong Liu[1], Yuta Nakashima[1], Noboru Babaguchi[1,2]
[1]Osaka University, Japan
[2]Fukui University of Technology, Japan
{hliu, n-yuta, babaguchi}@ids.osaka-u.ac.jp

1. Annotation system

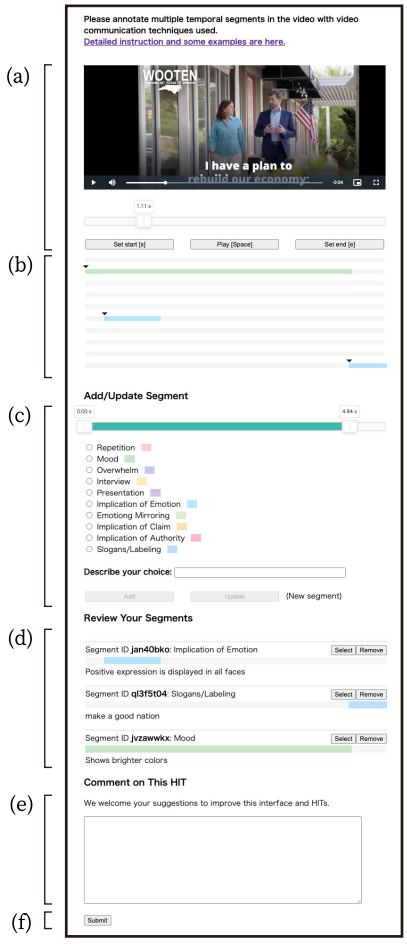

Figure 1. The interface of the annotation system. The annotators can watch the video and label the communication techniques in the video. Then, they can submit the labels to the server.

We develop an annotation system to label the communication techniques in the political advertisement videos. The interface is shown in Fig. 3, consisting of six blocks. The instructions are provided as an external website and another PDF file is also attached in our Supplementary. On the instruction website, we offer 11 example videos with annotated segments that cover all communication techniques.

Block (a) is the video viewer with a seek bar as well as a play/pause toggle button. It also provides buttons to set the start and end times of a segment.

Block (b) provides the summary of the segments that the annotator already made on the timeline for each communication technique. The annotator can select the segment on the timeline to modify the segment.

Block (c) is the main panel for annotation. An annotator first specifies the start and end times of the segment using the seek bar, where the start and end times can also be specified by pressing the corresponding buttons in Block (a). Then, the annotator identifies the communication technique for the segment. They are also asked to describe their choice in a short sentence. This description is merely to reduce the possibility of random annotations. The annotator watches the video and adds as many segments as they want using Blocks (a) and (c) back and forth.

After the annotator is satisfied with their annotations, they use Block (d) to review their annotations. The annotator can modify the annotation by pressing the "select" button or remove it by pressing the "remove" button.

Block (e) is used when the annotator finds any problem in the video (*e.g.*, a corrupted video). Finally, the annotator presses "submit" button in Block (f) to send the annotations to the AMT server.

2. The Intention classification task

Let $D = \{(v, S)\}$ be the dataset where v is a video and S is the corresponding intention labels. The intention classification task is to identify the communication techniques S given a video v. Since we have three annotators for labeling each video, we use the majority vote to determine the final ground truth label for each video. Specially, we determine

the longest segment of the video across the three annotators, which can be formulated as:

$$\hat{s}_i = \underset{s_{i,j} \in S}{\arg\max}(b_{i,j} - a_{i,j}), \tag{1}$$

where $a_{i,j}$ and $b_{i,j}$ are the start and end time of the j-th segment in i-th video with its corresponding label $s_{i,j}$, respectively. After this process, we can get the dataset $D' = \{(v, \hat{S})\}$. As a result, we can train a classifier f to predict the intention label $\hat{S}$ given a video v.

In the experiments of the main paper (See Tabel 2 in Section 4.1), we conduct the experiments on the dataset D', and report the Top-1 and Top-5 accuracy of the classifier f. The Top-1 accuracy is the percentage of videos where the predicted label is the same as the ground truth label. The Top-5 accuracy is the percentage of videos where the ground truth label is in the top 5 predicted labels. For the comparison, we select three state-of-the-art methods, including C2D [3], I3D [1], and SlowFast [2], as the backbone networks for the intention classification task. Moreover, we also consider the random classification test to establish a baseline for the intention classification task. We use the same training and testing strategy for all the methods, and report the results in Table 2 in the main paper.

Since the communication techniques are not uniformly distributed in the dataset (see Figure 4 in main paper), we have three annotators for labeling the segments of each video, thus each video has different ground truth labels. Suppose that i-th annotator labels the video with the communication techniques $\overline{S} = \{\overline{s_1}, ..., \overline{s_N}\}$, where $\overline{s}_i$ is the ground truth label vector for the i-th video that is a one-hot vector with the length of 10. This ground truth vector can be calculated as follows:

$$\overline{s_i} = \frac{s_i^1 + s_i^2 + s_i^3}{3}, \tag{2}$$

where s_i^j is the one-hot vector of the j-th annotator's label for the i-th video.

Therefore, we can build a new dataset that can be formulated as $D'' = \{(v, \overline{S})\}$. We can train a classifier f to predict the intention label $\overline{s_i}$ given a video v_i, and then we evaluate the performance of the classifier f on the dataset D''. Finally, we report the KL divergence between the predicted label $f(v_i)$ and the ground truth label $\overline{s_i}$ for the intention classification task. Moreover, we also report the mean average precision (mAP) over all test data. Similar to previous experiments, we also use the same baseline methods and report these results in Section 4.1 of the main paper. Where we can see that the KL divergence of the random classification is 0.4515, and the KL divergence of the I3D, C2D, and SlowFast are 0.4268, 0.4346, and 0.4322, respectively. And the results of mAP have the same tendency as KL divergence.

At last, in the main paper, we report the result of most frequent communication technique of segments, such as

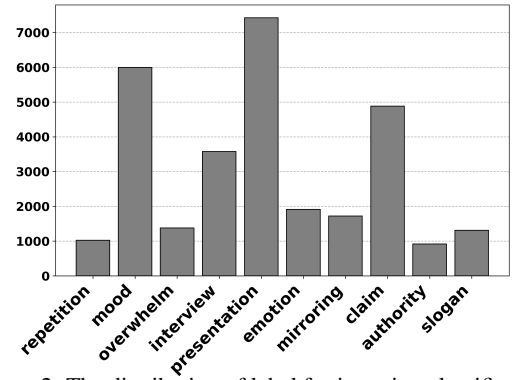

Figure 2. The distribution of label for intention classification.

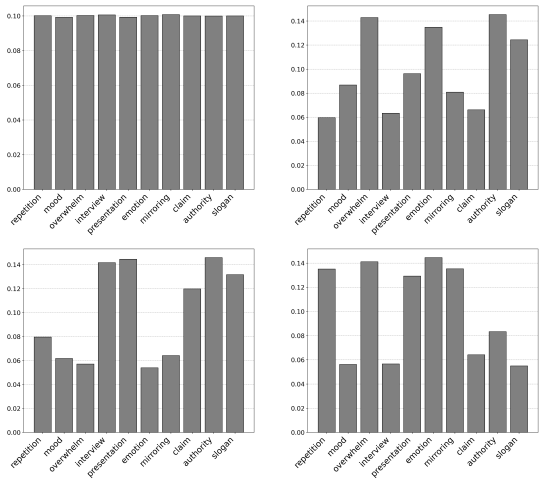

Figure 3. The predicted label distribution of different models on intention classification

`Presentation`. Refer to the Figure 2 that shows the distribution of the label for the intention classification task. Moreover, we also show the distribution of the predicted labels (Top-5 prediction) within four compared baselines, such as random guess, C2D, I3D, and slowfast. See Figure 3. We observe that the top-5 predicted classes actually cover most videos, while having a similar trend to the ground-truth distribution shown in Figure 2. For example, most trained models output `Presentation` with high probability. On the other hand, the distribution of the random guess is equal to a uniform distribution, where the probability of the Top-5 accuracy is around 87.5%, which is on par with our result (*i.e.,* 84.56%).

When we take dataset D' into account, we statistically analyze the category with the highest probability output by the model, which is also called the most frequent communication technique that the model consistently predicts. This is the `Implication of emotion` with the probability of 51.59%. After that, we select the top-5 most frequent communication techniques that the models usually predict in

Table 1. Annotation agreements by mAP at various tIoU thresholds.

Label		tIoU 0.3	0.4	0.5	0.6	0.7
Slogan	$\mathcal{A}_1$ and $\mathcal{A}_2$	58.39	10.97	3.09	1.59	1.78
	$\mathcal{A}_1$ and $\mathcal{A}_3$	62.24	8.66	4.93	2.84	2.24
	$\mathcal{A}_2$ and $\mathcal{A}_3$	41.46	16.03	7.32	8.01	3.48
Presentation	$\mathcal{A}_1$ and $\mathcal{A}_2$	39.02	24.39	7.32	3.66	2.44
	$\mathcal{A}_1$ and $\mathcal{A}_3$	40.63	12.50	6.25	6.25	3.13
	$\mathcal{A}_2$ and $\mathcal{A}_3$	37.14	24.29	4.29	8.57	2.86
Claim	$\mathcal{A}_1$ and $\mathcal{A}_2$	38.45	13.68	10.14	8.05	5.07
	$\mathcal{A}_1$ and $\mathcal{A}_3$	35.84	14.25	10.71	6.55	4.96
	$\mathcal{A}_2$ and $\mathcal{A}_3$	25.50	16.21	14.84	10.59	8.29
Interview	$\mathcal{A}_1$ and $\mathcal{A}_2$	38.78	18.18	8.69	5.45	5.66
	$\mathcal{A}_1$ and $\mathcal{A}_3$	41.81	15.07	7.34	3.95	3.20
	$\mathcal{A}_2$ and $\mathcal{A}_3$	25.87	20.43	15.65	4.56	3.91
Emotion	$\mathcal{A}_1$ and $\mathcal{A}_2$	37.47	15.14	9.68	8.19	4.22
	$\mathcal{A}_1$ and $\mathcal{A}_3$	35.07	12.69	12.19	5.97	3.98
	$\mathcal{A}_2$ and $\mathcal{A}_3$	25.00	16.85	12.50	8.42	4.89
Authority	$\mathcal{A}_1$ and $\mathcal{A}_2$	16.04	11.00	12.96	10.57	11.06
	$\mathcal{A}_1$ and $\mathcal{A}_3$	12.19	12.19	12.25	12.88	10.69
	$\mathcal{A}_2$ and $\mathcal{A}_3$	14.68	15.18	15.32	13.12	10.50
Mirroring	$\mathcal{A}_1$ and $\mathcal{A}_2$	17.06	13.92	12.55	7.65	9.80
	$\mathcal{A}_1$ and $\mathcal{A}_3$	10.29	8.76	10.10	11.62	9.90
	$\mathcal{A}_2$ and $\mathcal{A}_3$	17.05	15.25	14.21	10.34	10.59
Overwhelm	$\mathcal{A}_1$ and $\mathcal{A}_2$	29.03	13.98	8.60	2.15	5.38
	$\mathcal{A}_1$ and $\mathcal{A}_3$	13.25	14.45	4.82	2.41	10.84
	$\mathcal{A}_2$ and $\mathcal{A}_3$	12.66	24.05	12.55	7.65	9.80
Mood	$\mathcal{A}_1$ and $\mathcal{A}_2$	17.36	10.54	7.05	6.75	6.52
	$\mathcal{A}_1$ and $\mathcal{A}_3$	15.54	8.58	8.30	8.30	8.66
	$\mathcal{A}_2$ and $\mathcal{A}_3$	16.42	10.56	11.18	8.78	6.92
Repetition	$\mathcal{A}_1$ and $\mathcal{A}_2$	43.37	12.05	3.61	3.61	3.61
	$\mathcal{A}_1$ and $\mathcal{A}_3$	46.60	9.71	5.83	3.88	2.91
	$\mathcal{A}_2$ and $\mathcal{A}_3$	31.13	16.98	8.49	7.55	3.77

our dataset, which are `Interview`, `Implication of Claim`, `Slogan`, `Mood`, and `Presentation`. Our experiments also show that the model can usually predict the most frequent communication technique with a high probability, which is consistent with the results of the main paper (see Figure 4 (e)).

3. More discussion on annotation agreement

Here we compute the per-class agreement (as the Table 1 shown), `Slogan`, for instance, is consistent. This is because `Slogan` comes with concrete visual and acoustic cues, and there is less subjectivity in spotting them. On the other hand, the `Slogan` segments always appear at the end of political advertisement videos and are usually very short (often lasting only 1 second). This brevity makes them difficult for annotators to label accurately. Consequently, the agreement at tIoU=0.3 is higher for these segments. But, from an overall analysis, labels with concrete visual and acoustic cues typically achieve higher annotation consistency, such as `Slogan` and `Presentation`.

In addition to the consistency statistics based on segment overlap (see Table 1), we also calculated the voting consistency among different annotators for the same video. That is, at least two annotators assigned the same label to the same video. We report the results as follows:

Slogan	49.36%	Presentation	58.43%
Authority	25.57%	Repetition	25.79%
Mood	55.43%	Overwhelm	18.57%
Interview	42.40%	Emotion	38.10%
Mirroring	40.80%	Claim	67.67%

These results show the same trend as the previously mentioned consistency analysis results. On the other hand, `Implication of Claim` is the most common label in the annotations, so the voting consistency will also have a relatively high score.

From this fact, we would argue that our annotators worked well, while some classes offer inherent subjectivity. There can be many such tasks, like facial expression recognition, where there can be differences in emotional expressions on faces and subjectivity in received emotion (some people see a smiling face, while others may see a crying face in the same image, though the current datasets remove such examples for annotators' consistency). As performance is saturated for basic tasks, it may be time to dive into such subjective tasks. And yes, we still need to explore the evaluation metric and algorithms for such tasks in our future work, as the performance is capped because of the inconsistency, as pointed out. We still believe that the paper serves as a baby step toward AI that can handle subjectivity.

4. Results on majority voting for intention classification

Our initial labeling method used the longest segment of each annotator, assuming it represented the communication technique that was the most used. We then adopt a majority vote system among the three annotators per video. Note that for the initial labeling method, each video contains three labels, while each video only contains one single label for the majority voting method. The comparative results are presented in the table below.

Method	Random	C2D	I3D	SlowFast
Top-1	8.36	8.48	9.60	9.92
Top-5	49.56	63.32	66.64	67.76

5. Comparison with different annotators

We evaluate our method on the annotations from each annotator (*i.e.*,$\mathcal{A}_1$, $\mathcal{A}_2$, or $\mathcal{A}_3$). The results on annotations of $\mathcal{A}_1$ achieves highest performance with the average score of 15.17%, which is slightly higher than the model evaluated on all annotations (*i.e.*, 14.68%). But it is better than the model evaluated on the annotations of both $\mathcal{A}_2$ and $\mathcal{A}_3$.

Table 2. Results of the our multi-modal model with different Feature extractors.

Features	tIoU=0.3	tIoU=0.4	tIoU=0.5	tIoU=0.6	tIoU=0.7	Avg.
VGGish	15.61	13.46	11.23	9.44	7.94	11.54
AST	17.35	15.29	13.02	10.70	8.36	12.94
I3D	18.93	16.65	14.11	11.79	9.68	14.23
VideoMAE	19.22	17.05	14.73	12.38	9.92	14.66
VGGish + Text	15.77	13.57	11.29	9.57	8.11	11.66
AST + Text	17.79	15.67	13.69	11.53	9.23	13.58
I3D + Text	19.19	16.70	14.23	11.80	9.69	14.33
I3D + VGGish	19.37	16.74	14.38	11.98	9.84	14.46
I3D + AST	19.31	16.52	14.19	11.80	9.72	14.31
VideoMAE + AST	19.31	16.57	14.24	11.83	9.74	14.34
I3D + VGGish + Text	19.68	17.08	14.50	12.08	10.06	14.68
I3D + AST + Text	19.68	17.12	14.46	12.10	10.19	14.71
VideoMAE + AST + Text	19.64	17.22	14.76	12.45	10.32	14.88

Table 3. Comparison of different annotators on the Three modalities. The results are reported in terms of the mAP@tIoU metric. A1, A2, and A3 represent the three annotators, respectively. And "All" means that the model is trained on the annotations of all annotators.

Method	tIoU=0.3	tIoU=0.4	tIoU=0.5	tIoU=0.6	tIoU=0.7	Avg.
All$\rightarrow\mathcal{A}_1$	20.24	17.63	15.08	12.78	10.13	15.17
All$\rightarrow\mathcal{A}_2$	16.97	13.60	10.69	8.51	6.39	11.23
All$\rightarrow\mathcal{A}_3$	19.37	16.74	14.26	11.79	9.75	14.38

6. More discussion with our simple multi-modal model

ViT features outperform I3D features, while AST features surpass VGGish features, though both audio features underperform visual features. Text features consistently improve performance when combined with visual and audio features, highlighting their importance in understanding political ad intentions. Notably, I3D+VGGish achieves comparable results, likely due to high feature complementarity. However, incorporating textual information with stronger visual and audio features yields even better results. In sum, the combination of visual, audio, and textual features is crucial for understanding political ad intentions, with visual features playing a significant role in enhancing model performance.

7. Limitations and further discussion

Our dataset is collected from YouTube, which is a global and popular platform. To analyze the communication technique used in existing political advertisement videos, we have collected videos from mainly from the USA. This is often studied in related research of political analysis and media analysis. Moreover, our annotators are all English speakers, so our dataset is mainly focused on English-speaking countries. We agree with your comments that the geographical variations in political video understanding are important and interesting. We will consider this in our future work.

In fact, slogans are a key element in political ad videos, and are often considered as a concise message that can quickly convey the intentions of the video. We think that the slogan can be a useful signal for the model to understand the intentions of the video. But our paper focuses mainly on building a dataset and designing a simple model to study the potential of understanding the intentions of political advertisement videos. We believe that using other techniques like OCR can further improve the performance of the model. And we will leave this for future work and promote communities to explore this direction.

Currently, we can see low scores, suggesting disagreement of annotations. This is the nature of the task, and we believe that the errors and discrepancies can provide some ideas about video intentions. Although the current annotations show high consistency for relatively certain techniques, there is significant uncertainty in the annotations for more subjective techniques. Therefore, future work will explore using additional annotators or advanced large multi-modal models to enhance annotation quality.

References

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 2

[2] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 2

[3] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 2