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
https://lyn-l.github.io/politicalad/

Abstract

*In this paper, we introduce a novel task for video understanding that focuses on detecting editing intentions in political advertisement videos. Political advertisement videos are edited with some intentions (*e.g.*, "associating some candidates with negative emotions") of making people unthinkingly believe the messages in the videos, potentially ending up with some irrational bias. Detecting such intentions is thus the primary step toward fairer decision-making based on the messages themselves. To this end, we classify such editing intentions into 10 categories (referred to as communication techniques) in consultation with a professional editor as well as based on communication techniques presented in the natural language processing community, and build a dataset of 12,526 political advertisement videos, each of which are annotated with several communication technique segments. We also explore the capability of existing video understanding models in detecting editing intentions over the dataset, which identifies new dimensions of challenges to be addressed.*

1. Introduction

In pursuit of effective and impressive content to convey some messages, video creators usually employ editing techniques [6], such as compiling various video segments, incorporating sound effects, color adjustment, and overlaying captions and logos. As illustrated in a possible conceptual mental model of a video creator's editing process in Figure 1, a video creator may initially decide on the *message* (*e.g.*, "Party B's policy is not good") to communicate in their videos. Then, the creator thinks about how to impress the message to make the video effective and impactful so that people (unthinkingly) believe the message. For example, the creator may wish to associate the party with a negative emotion to highlight an undesirable consequence of Party B's policy. This step involves making up the creator's *intention* for editing the video. To actualize this intention, creators build up a more concrete plan for editing the video, or *agendea*, finding video clips from a video database or shooting some, and deciding their composition and various editing techniques (*e.g.*, color changes, transitions, and adding logos/subtitles) to use. Finally, the creator compiles the video clips to get the final cut.[1]

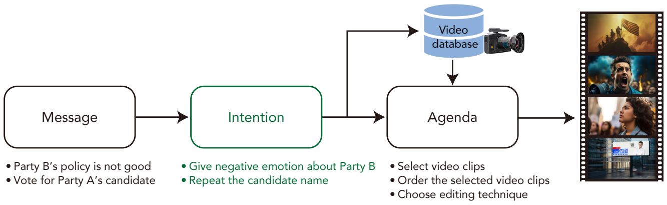

Video database
Message
Intention
Agenda
• Party B's policy is not good
• Vote for Party A's candidate
• Give negative emotion about Party B
• Repeat the candidate name
• Select video clips
• Order the selected video clips
• Choose editing technique

Figure 1. A conceptual model for video creators' editing process.

The message and intention are thus essential for understanding videos, especially for edited ones, at the higher level.[2] Meanwhile, the majority of recent research in video comprehension has concentrated on lower-level (perception) tasks, such as action recognition [26], object detection [24], and video captioning [48], in which a model's ability to recognize visual elements matters.[3] Identifying (and localizing) video intentions and messages can involve deeper understanding of all modalities in videos (*i.e.*, frames, audio, speech, *etc.*), which remains underexplored.

Edited videos are often associated with strong intentions to convince their viewers to have certain thoughts, which may even be seen as *propaganda* [12,19,27,35]. Most edited videos can exhibit such an aspect, and one prominent example is political advertisement videos, in which politicians depict themselves positively while their opponents negatively, trying to maximize the impact according to their intentions. Recognizing these underlying intentions of a video allows viewers to critically evaluate the video's message.

[1] This model is simplified; a real editing process goes back and forth between the iteration and agenda steps.

[2] We believe inferring the agenda (or detecting, *e.g.*, editing techniques) is a low-level task as the agenda is directly reflected in the edited video.

[3] Video captioning requires generating text, though this is mostly about a language model.

In this paper, we present a novel task in the realm of video understanding, with a specific focus on identifying intentions, particularly within the context of political advertisement videos. To this end, we collect a new dataset, called PALADIN (**P**olitic**AL** **AD**vertisement video **IN**tention), which contains more than 12,526 short political advertisement videos[4]. We define a taxonomy of intentions with ten categories based on the taxonomy for natural language text [4,10], such as `interview`, `presentation`, `slogans`, *etc*., which are extensively used in political advertisement videos. For each video, our annotators identify multiple video segments that reflect intentions. We present some baselines based on the state-of-the art action detection methods to detect intentions in political advertisement videos to show the challenges in our task. Moreover, we also introduce a simple multi-modal model that can leverage both visual and audio information to detect intentions in videos. The results show that these models fall short of effectively finding the intentions in videos, which show potentially new dimensions of challenges to be addressed in video understanding.

Contributions. From the application side, we present a new dataset and associated task, called PALADIN, to localize the creator's intentions in political advertisement videos. Intention localization can provide a meta-perspective to be critical of the messages in the video. Technically, the task poses two new dimensions of challenges to be addressed, *i.e*., the subjectivity/ambiguity of intentions and deeper comprehension of a video (detailed in Section 3.1).

2. Related Work

What we call by *editing intentions* has been studied in the context of propaganda detection, especially in natural language processing. In the early stages, propaganda detection was done on the basis of entire text units, which is still being studied extensively [5, 12]. Davidson *et al*. [7] constructed a dataset of about 24,000 tweets labeled in the categories of hate speech, offensive language, and neither to distinguish them. Rashkin *et al*. [34] created a corpus of news articles from eight different sources and classified them into four categories: propaganda, hoax, trusted, and satire.

In recent years, the study has been extended to the goal of detecting fine-grained propaganda techniques at the fragment level [5]. This study defined 18 different propaganda techniques, and they manually annotated a corpus of news articles at the fragment level. Other studies include building models to detect and classify propaganda techniques using the PTC corpus [28] and propaganda detection in multimodal data [9]. In recent years, Sprenkamp *et al*. [37] investigated the effectiveness of large language models, such as GPT3 and GPT4, for propaganda detection.

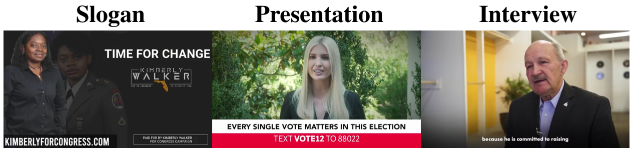


Figure 2. Examples of communication techniques in political advertisement videos, such as slogan, presentation, interview.

The development of social media networks with the capability of sharing videos has increased the risk of propaganda being carried out using videos as well as text [38, 47]. There are fewer propaganda detection methods for videos than for text data. Even these methods do not use a video itself as input for detection models but its captions [23] or metadata [21]. To our knowledge, there are no existing studies for videos that address propaganda technique detection at the fragment level. In this study, we focus on identifying propaganda used in fragmentary segments within videos.

3. Dataset

To investigate video intentions employed in political advertisement videos, we construct the PARADIN dataset, which is the first dataset dedicated to understanding intentions within the context of political advertisement videos. To this end, we first define a taxonomy of intentions borrowed from fine-grained propaganda definition [4] (Section 3.1). The video creators typically convey multiple messages, and each message may be emphasized with multiple intentions, even in a short video (as shown in Figure 1). Therefore, we designed our task to localize intention segments in videos, thereby annotating videos with multiple video segments associated with intention labels.

3.1. Taxonomy of Video Intentions

The definition of intentions in Figure 1 is strongly tied to propaganda, which is defined in [30] as:

> *The systematic dissemination of information, esp. in a biased or misleading way, in order to promote a particular cause or point of view...*

With this definition, propaganda is an act of disseminating the messages that the video creator wishes to convey. Various video editing techniques can be used to achieve this, which can be seen as agendas in Figure 1. There is still a gap between the messages and agendas, which is how to impress the messages so that the creator can choose suitable video editing techniques. This is called *propaganda techniques*. As in the definition above, the term *propaganda* is used in negative contexts. Meanwhile, we argue that the messages are not necessarily biased or for deception (even in political advertisement videos). We, therefore, use the term *intentions* to mean how to impress the message and refer to the individual strategies as *communication techniques*, instead of propaganda techniques, to avoid negative impressions.

[4] Here, we define a short video as a video that is less than 20 seconds.

1. `Repetition`: Repeat the same message until the audience accepts it.
2. `Mood`: Show scenes with some colors or modification of the dominating colors in the scenes so that the scenes give the impressions/emotions associated with the colors. For example, dark colors may imply negative impressions/emotions; red may imply anger or heat.
3. `Overwhelm`: Show many visual elements (*e.g.*, changing scenes rapidly or packing many visual elements like images, subtitles, etc.) in a scene), sometimes in a chaotic way, which may prevent the viewer from thinking.
4. `Interview`: Show an interview with some people (typically a single person) talking about some ideas as their own thoughts. By this, the viewer may think that the ideas are accepted by (a certain group of or a general) people.
5. `Presentation`: Discuss or present some ideas to convince the viewer.
6. `Implication of emotion`: Show some visual elements that imply certain emotions or impressions so that the topic of (the corresponding part of) the video can be associated with these emotions/impressions. For example, talking about a certain candidate in a political advertisement video with showing a person with raising his/her clenched fist may imply confidence.
7. `Emotion mirroring`: This is similar to `Implication of emotion`, but with showing facial expressions (instead of arbitrary visual elements).
8. `Implication of claim`: Show some visual elements that support a certain idea in (the corresponding part of) the video. Visual elements are not necessarily associated with the context. For example, showing an image of a candidate together with an image of bills may give the impression that the candidate is mean about money.
9. `Implication of authority`: Show some visual elements that imply authority to give an impression that the idea is authorized, which may not necessarily be the case.
10. `Slogans`: Show a slogan (as a visual element like a subtitle) to impress it.

Figure 3. Our taxonomy of communication techniques.

In the literature, there are many different definitions and taxonomies of such techniques. The earliest study by Miller [29] identified seven communication techniques. Recently, the natural language processing community has identified 22 types of communication techniques, such as slogans, repetition, appeal to authority, *etc*. [4, 10]. These techniques are mainly for text, and they are not always suitable to describe communication techniques in videos. For example, the technique `straw man` is well-known, but it involves only the verbal modality and does not affect the choice of editing techniques.[5]

We refer to the literature for tips on how to classify the communication techniques with some informal discussion with a professional video editor and borrow 10 different techniques in Figure 3 that are extensively used in political advertisement videos. Figure 2 shows examples of some communication techniques in our dataset. We denote the set of the communication technique labels as $\mathcal{L} = \{l\}$, where $|\mathcal{L}| = 10$.

3.2. Data Collection and Annotation

Recent years have witnessed the extensive use of communication techniques for manipulation, especially during political campaigns, as discussed in the Wikipedia page[6]. We thus chose political advertisement videos for benchmarking models' ability of video intention comprehension. Google's Ads Transparency Center[7] provides various metadata of advertisement videos. We use political advertisement videos aired in the United States through YouTube. We selected shorter videos ranging from 3 to 16 seconds (the vast majority is 15 seconds) to reduce annotators' burden. This data collection process ended up with $N = 12,526$ short videos in total. We denote a set of these videos by $\mathcal{V} = \{v\}$, where $|\mathcal{V}| = N$.

We developed an interface system for the annotation of temporal video segments. After reading the instructions for the annotation process, along with some examples of each communication technique, an annotator watches a video and annotates all segments in which the intentions are encoded. The annotator then asked to assign communication technique label $t_i \in \mathcal{L}$ for all segment i with a short description[8]. We denote the set of annotations for a video as $\mathcal{S} = \{(b_i, e_i, t_i)\}_i$, where b_i and e_i are the start and end times of the segment i.

We used Amazon Mechanical Turk (AMT)[9] to deploy our annotation jobs. One job (referred to as a HIT in AMT) corresponds to one short video. The annotators were compensated with 0.5 USD per job so that AMT's standard could be met. As this annotation process requires understanding English, we recruited only annotators who could speak fluent English. To ensure that annotators were familiar with our jobs, we asked them to take a qualification test, in which they were asked to annotate a certain video (excluded from our dataset). We manually checked all qualification test results by annotator candidates, and those who made reasonable annotations were adopted.

Each video was annotated by three annotators, where we deployed a single job for each $v \in \mathcal{V}$ as a batch, and the three batches were chronologically separated. We denote the set of annotations by $\mathcal{A} = \bigcup_{k=1}^{3} \mathcal{A}_k$, where $\mathcal{A}_k = \{S\}$ is the set of the annotations for all videos in $\mathcal{V}$ for batch k.

It should be noted that this process allows annotators to give multiple segments simultaneously *i.e.*, the segments can overlap with each other. This is because a video can encode

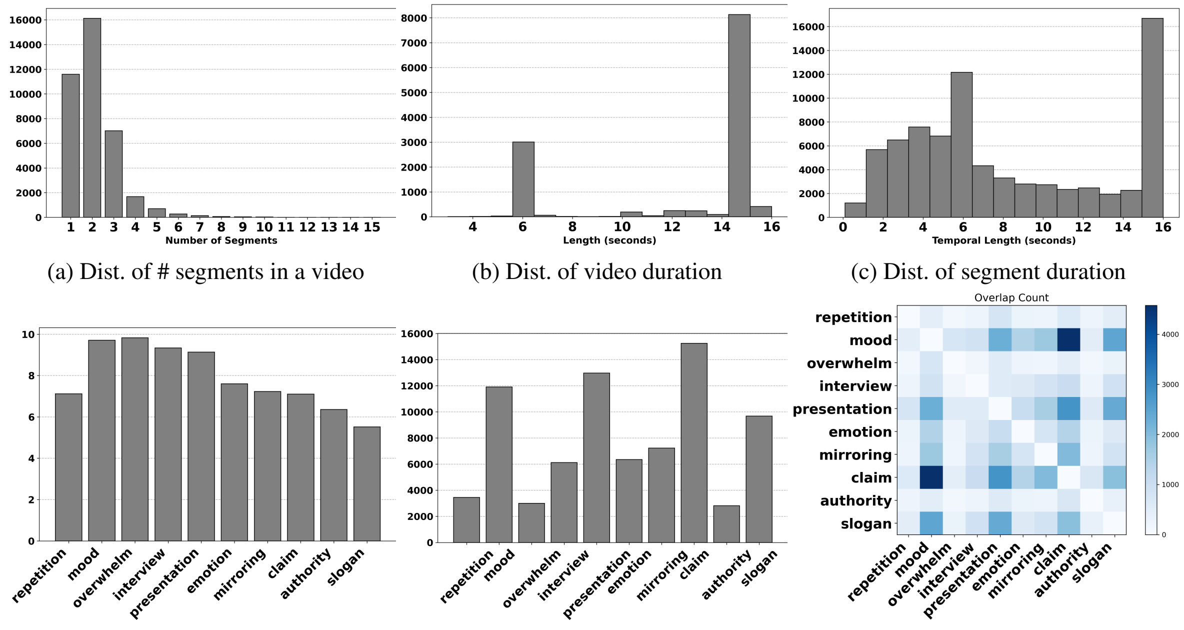
[5] A straw man fallacy can be used in a video, but we argue that it is not a technique for video since it is closed in the verbal modality.

[6] https://en.wikipedia.org/wiki/Video_manipulation

[7] https://adstransparency.google.com/

[8] This is merely to suppress random annotations and is not used in the paper though included in our dataset.

[9] https://www.mturk.com



(a) Dist. of # segments in a video

(b) Dist. of video duration

(c) Dist. of segment duration

(d) Avg. segment duration per technique

(e) Dist. of labels

(f) Co-occurrence of labels

Figure 4. Various statistics on the PALADIN dataset.

more than one intention simultaneously since a video can have a hierarchical structure [39]. For example, a video in the Presentation style can have a segment that shows a happy face, which can be seen as Implication of emotion, Emotion mirroring, or Mood.

3.3. Statistical Analysis

In total, PALADIN contains a total of 78,790 video segments, distributed over 40 hours of videos, covering annotations from three annotators. Figure 4 shows the statistics on our dataset. Note that some statistics as well as distributions are computed over $\mathcal{A}$ to show the characteristics of our dataset; therefore, the number of videos (denoted by "count") for each entry sums up to $3N$.

Number of Segments in a Video. As shown in Figure 4 (a), 69.21% of the videos contain more than one segment, indicating that the videos are edited with multiple intentions. The average number of segments in a video is 2.10, and the maximum is 15, which seems an outlier. Among the segments, there are $7,573$ videos that contain overlapping segments, which means that the videos are edited with multiple intentions at the same time.

Durations. Figure 4 (b) shows the distribution of the duration of videos. We can see two noticeable peaks at 6 and 15 seconds, where 15 is the vast majority, occupying roughly 75% of videos. Recent studies [1] have shown that the short duration of political advertisement videos (such as 15 seconds long) is effective in delivering messages, which is the most common choice for video creators to show their intentions. More recently, a 6-second long video, also called Bumper Ads[10], has also been used for the same purpose. So our dataset primarily comprises 6 and 15-second videos, aligning with current trends. Intriguingly, the distribution of the segment duration forms two peaks as in Figure 4 (c): one is around 6 seconds, and the other is at 15 seconds. We can guess that some intentions, such as Interview and Presentation, can only be actualized by using entire videos (as the videos themselves are already short). This guess is supported by the average segment duration per technique in Figure 4 (d), which shows the average durations of segments with Interview, Presentation, Mood, and Overwhelm labels are near 10 seconds.

Distribution of Communication Technique Labels. Figure 4 (e) shows the distribution over ten communication technique labels $\mathcal{L}$. We can see that Implication of claim and Interview are the two most frequently used techniques, while Implication of authority and Overwhlem are the least used. This may echo the fact that the Implication of claim is often used to implicitly convey some negative aspects of the opponent candidate in the political campaigns, and Slogan can easily impress a candidate both positively and negatively. Meanwhile, Overwhelm may not be often used because it may confuse the viewers [4, 20].

Co-occurrence between Communication Techniques. Figure 4 (f) shows the co-occurrence counts between all pairs of the communication techniques. Mood is often used together with Implication of claim, Slogan, and Presentaiton, while the Presentation is often used

[10] https://megadigital.ai/en/blog/youtube-bumper-ads/

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

	tIoU				
	0.3	0.4	0.5	0.6	0.70
$\mathcal{A}_1$ and $\mathcal{A}_2$	32.27	12.77	9.06	7.05	6.18
$\mathcal{A}_1$ and $\mathcal{A}_3$	27.81	11.82	9.74	8.00	6.84
$\mathcal{A}_2$ and $\mathcal{A}_3$	21.31	15.36	13.43	9.94	7.74

together with `Slogan`, `Implication of claim`, and `Mood`. This is also in line with our intuition about political advertisement videos, where a presentation of some ideas may be accompanied by, *e.g.*, a slogan to emphasize them.

Agreement of Annotations. Due to the ambiguities in the definitions of communication techniques and the subjectivity of a viewer's impression when watching a video, we presume that the agreement among different rounds can be low. We evaluated the agreement by computing mAP@tIoU scores for three pairs (*i.e.*, $(\mathcal{S}_1, \mathcal{S}_2)$, $(\mathcal{S}_2, \mathcal{S}_3)$, and $(\mathcal{S}_1, \mathcal{S}_3)$, where $\mathcal{S}_k \in \mathcal{A}_k$ for a single video), taking their average, and then taking the average over all videos in $\mathcal{V}$.

Table 1 summarizes the scores. We can see low scores, suggesting disagreement of annotations. We consider that this is the inherent nature of our task. For example, `Implication of Emotion` and `Emotion mirroring` are similar to each other, whereas `Emotion mirroring` involves facial expressions. Yet, `Implication of Emotion` can also use facial images. Their boundary is not always obvious. Moreover, although the disagreement is shown in the table, we further calculate the agreement rates of the same label among the annotators. Where "Presentation" (58.43%) and "Slogan" (49.36%) show higher consistency, while "Implication of Authority" (25.75%) and "Repetition" (25.79%) show lower consistency with lower agreement rates. More details refer to supplementary. This variance stems from the relative subjectivity and ambiguity of the communication techniques. These findings indicate reliable annotations for certain techniques but highlight difficulties with more subjective ones. This is because the annotations can be affected by the annotators' personal experiences and backgrounds, which can lead to different interpretations of the same video. This fact poses an additional challenge in communication technique detection, *i.e.*, how to deal with inherent label disagreement.

From this annotator agreement analysis, we decided to provide all annotations in $\mathcal{A}_1$, $\mathcal{A}_2$, and $\mathcal{A}_3$ with anonymized annotator IDs, instead of finding ones that are consistent in a video in some criteria, because of the subjectivity and ambiguity of the annotation task itself. The dataset can come with annotation errors (*e.g.*, due to misunderstanding of the instructions and examples), though discrepancies in the annotations also stem from the subjectivity and ambiguity. We believe that the errors and discrepancies can provide some ideas about video intentions. Making full use of multiple annotations (perhaps from different perspectives of different annotators) offers a challenge in [13, 17, 42].

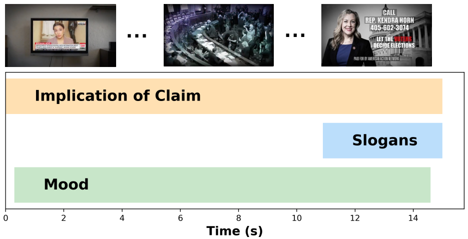


Figure 5. Example of the temporal detection task, encompassing boundary detection and intention classification for each segment.

3.4. The tasks

Let $\mathcal{D} = \{(v, \mathcal{S})\}$ denote the PALADIN (training) dataset, consisting of pairs of video v and the corresponding annotation $\mathcal{S}$. We define two tasks over PALADIN, *i.e.*, *intention classification* and *temporal detection* tasks. The intention classification task aims to identify the communication technique label of the longest segment in a given video v, assuming that it is the most prominent in v. PALADIN offers a set $\mathcal{D}^{\text{int}} = \{(v, l_{i^\star}) | (v, \mathcal{S}) \in \mathcal{D}_{\text{train}}\}$ for training, where $i^\star = \arg\max_i (e_i - b_i)$ in $\mathcal{S}$. A model f predicts $l_{i^\star}$ given v as $\hat{l} = f(v)$. The temporal detection task aims to identify all segments encoding intentions within a video and predict their corresponding labels. For this task, the PALADIN dataset $\mathcal{D}$ is directly used for training a model g, which gives a predicted set of segments as $\hat{\mathcal{S}} = g(v)$.

4. Experiments

We evaluated existing methods for video classification and action recognition in the PALADIN dataset as our baselines. The baselines were implemented using PyTorch [31] and trained on NVIDIA A100 GPUs. We follow the standard training settings for each baseline as provided in their original papers. All reported results are averaged over three runs with different random seeds. We use the Top-1, Top-3, and Top-5 accuracy as evaluation metrics for the intention classification task and the mAP@tIoU for the temporal detection task. The dataset is split into training and evaluation sets ($\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{eval}}$ with a ratio of 4:1, where $\mathcal{D} = \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{eval}}$. For this, we followed the multi-label split strategy [40] that targets a well-balanced data distribution in these two subsets. Thus, the training set contains $10,021$ videos, while the testing set contains $25,06$ videos.

4.1. Intention Classification Task

First, we conducted a random classification test to establish a baseline for the intention classification task. The motivation behind this test is to ensure that the training based method can achieve better performance than random guessing, thus confirming its ability to learn meaningful intention patterns from the data. We implemented the random clas-

Table 2. Performance of baselines on the PALADIN dataset. This table shows the Top-1 and Top-5 accuracy for the intention classification task. “Random” is the results of random classification. We report the results of I3D, C2D, SlowFast, and VideoMAE (ViT-B).

Model	Top-1	Top-3	Top-5	mAP
Random	22.88	61.52	84.56	56.32
I3D	26.98	63.96	83.05	57.99
C2D	26.98	63.64	83.04	57.89
SlowFast	29.25	64.64	84.64	59.51
VideoMAE	31.33	64.92	85.12	60.46

sifier by assigning labels to each video segment based on a uniform probability distribution across all available categories. After running the random classification on our test set that contains three annotations, we obtained an average accuracy of 22.88% for the Top-1 and 84.56% for the Top-5, see the results of “Random (ALL)” in Table 2. These results approximately align with the expected outcome for a ten-class classification problem with different ground-truths from three annotators, which serve as a lower bound for the following evaluation.

Since the `Presentation` is the most frequent communication technique used in the classification task (see Figure in the supplementary), we report the results of the model that always outputs this most frequent communication technique or not. This can be seen as a binary classification task, and the result is reported in Table 2. This model achieves an average accuracy of 76.60%, which is higher than other compared baselines. We think that this communication technique shows concrete visual and acoustic cues, which is easy for annotators to identify.

Second, we select C2D [45], I3D [3], and SlowFast [11], which are three classical supervised video recognition models, as our baselines. In our experiments, all these methods utilize ResNet-50 as the backbone, which is pre-trained on ImageNet [8]. Then, we fine-tune the model on $\mathcal{D}^{\text{int}}_{\text{train}}$ with supervised learning and evaluate it on the test set. We also use VideoMAE [43] as another baseline, which is a self-supervised video recognition model. We use ViT-Base as the backbone, which is pre-trained on Kinetics-400 [25]. We extract the features from the last layer of the backbone and add a linear probe layer trained on $\mathcal{D}^{\text{int}}_{\text{train}}$ for prediction.

Table 2 shows the performance scores of our baselines for the intention classification task. We observe that C2D and SlowFast achieved the top-1 accuracy of 26.98% and 29.25%, and the top-5 accuracy of 81.36% and 81.64%. Also, the results of VideoMAE are higher than the random classification results, which indicates that the model can learn meaningful intention patterns. This verifies that the self-supervised learning can be effective for the intention classification task. On the other hand, the performance of these trained models is comparable to the random guessing results, which indicates that our task is challenging. We hope the future work can improve the performance of the intention classification task by designing more sophisticated models or using more data.

Since each video in our dataset usually contains multiple communication techniques, we further evaluated the model’s output for predicting the multiple communication techniques in a video. To this end, we transform the ground truth labels to a ten-dimensional vector where each element represents the presence of a communication technique in the video. Thus, this label vector can be seen as the label distribution of communication techniques. We evaluated the performance of the three baseline models and random classification by calculating the Kullback-Leibler (KL) divergence between the predicted probability distribution and the ground truth distribution. We also report the mean Average Precision (mAP) score for the intention classification task. The results are shown in the following table:

Method	Random	C2D	I3D	SlowFast
KL divergence	0.4515	0.4268	0.4346	0.4322
mAP	52.22	53.94	53.74	53.87

We observe that all trained models achieve a lower KL divergence score and a higher mAP score than random classification, indicating that the model can learn meaningful patterns of intention. Despite SlowFast’s superior Top-1 accuracy (Table 2), it underperforms C2D in both KL divergence and mAP metrics. This suggests that the C2D model is relatively more suitable for multi-label communication technique recognition. To this end, based on the C2D model, we observed communication techniques with four highest output probabilities: `Implication of authority`, `Overwhelm`, `Implication of emotion`, `Slogan`, and `Presentation`. This is because the visual style of these techniques is very significant, which usually contains a similar template and font style.

4.2. Temporal Detection Task

This task is similar to the traditional temporal action localization task [44], but it is more challenging because the communication techniques are subjective and ambiguity. We follow the benchmark in the temporal action detection task [46], and use ActionFormer [49], TemporalMaxer [41], and TriDet [36] as our baselines. For all these detection models, we extract the visual feature vectors using the two-stream I3D model [3] and VideoMAE model [43], which are pre-trained in Kinetics-400 [25]. Since these baseline models are designed mainly for action localization tasks with only video data, we also follow the multi-modal model [15] that can handle both video and audio data. We extract audio feature vectors using the VGGish model [18] and AST model [16] pre-trained on AudioSet [14]. Moreover, we also use whisper [33] to recognize speech information and then use the CLIP text encoder [32] to extract text features.

Table 3. Results of different video understanding models on the PALADIN dataset. "Random Guess 1" is the result of randomly guessing the communication technique labels for each segment of ground truth. "Random Guess 2" shows the results when we randomly guess the boundaries of the segments and the labels corresponding to the communication technique. "Random Guess 3" is the result of the random guessing of the labels of the communication technique for each segment that the TemporalMaxer model achieves. "Most Frequent #" employs the same prediction scheme as "Random Guess #", with the main difference being that we assign the majority label (*i.e.*, `Implication of Claim`) to the corresponding segments.

Method	tIoU=0.3	tIoU=0.4	tIoU=0.5	tIoU=0.6	tIoU=0.7	Avg.
Random Guess 1	—	—	—	—	—	9.15
Most Frequent 1	-	—	—	—	—	25.93
Random Guess 2	3.11	2.24	1.47	0.87	0.41	1.62
Most Frequent 2	10.16	14.65	10.23	6.54	3.64	10.85
Random Guess 3	11.75	9.71	7.68	6.02	4.63	7.96
Most Frequent 3	28.52	20.54	13.94	9.00	5.99	15.60
ActionFormer	17.35	14.98	12.71	10.56	8.96	12.91
TemporalMaxer	19.06	16.64	14.14	11.93	9.73	14.30
TriDet	18.93	16.65	14.11	11.79	9.68	14.23
Ours	19.68	17.08	14.50	12.08	10.06	14.68

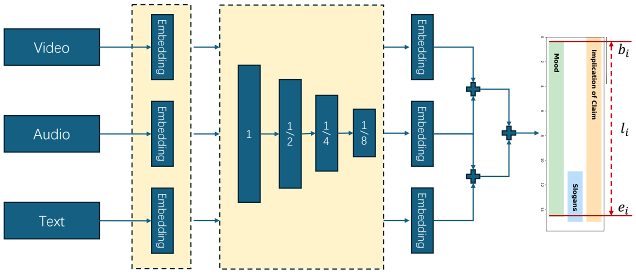


Figure 6. A simple framework of the multi-modal fusion method for temporal detection task. The audio, video and text features are projected into a shared embedding space via a 1D convolution. Then, max-pooling is applied to downsample features, which further input to a simple concatenated embedding.

4.2.1 A Simple Multi-modal Model

We consider that the audio/speech and its corresponding visual information are both crucial to identifying the communication techniques in the political advertisement videos. To this end, we design a simple multi-modal model that combines the video, audio, and text features for this task. Figure 6 shows the framework of the multi-modal fusion method.

Given a video features $V \in \mathbb{R}^{K \times D_v}$, audio features $A \in \mathbb{R}^{K \times D_a}$, and text features $T \in \mathbb{R}^{K \times D_t}$, where K is the number of frames, D_v, D_a, and D_t are the dimensions of the video, audio, and text features, respectively. First, for features $F \in \{V, A, T\}$, the downsampled feature $F' \in \mathbb{R}^{T' \times D'}$ is obtained by applying a 1D convolution operation. Second, the max-pooling is applied to downsample the features via max-pooling with stride of 2 to get the feature $\hat{F} = \text{MaxPool}(F')$, which is similar to the operation in the TemporalMaxer model [41]. Third, the three downsampled features are fused via a simple concatenated embedding, which is shown in Figure 6. Finally, we use the same training objective as the TriDet model to train the multi-modal model on the PALADIN dataset, which contains a classification loss to predict the communication techniques and a regression loss to predict the temporal boundaries of the video segments that contain communication techniques. Since the model is easy to overfit to the training set, we use the early stopping strategy to prevent overfitting. Based on our experiences, we usually early stop the training process at 10-th epoch. Other settings are the same as the implemented details of the original methods.

During inference, the model gives the probability p_{lt} of label l for temporal index t, as well as the onset (start time) d_t^{s} and offset (end time) d_t^{e}, such that the predicted label c_t for t is given by

$$c_t = \arg\max_l p_{lt}, \tag{1}$$

and the start (τ_t^{s}) and end (τ_t^{e}) temporal boundaries are

$$\tau_t^{\mathrm{s}} = t - d_t^{\mathrm{s}} \quad \tau_t^{\mathrm{e}} = t + d_t^{\mathrm{b}}. \tag{2}$$

We use the same evaluation metric as the temporal detection model to evaluate the performance on the PALADIN dataset. Specifically, we report mAPs at different tIoU thresholds.

4.2.2 Temporal Detection Results

Since video edits often appear in segments, our first goal is to predict the type of video editing present in a given short video, as well as determine where it starts and ends. This is similar to the temporal action localization task. Following the benchmark work in [46], we study the ActionFormer model [49], TemporalMaxer model [41], and TriDet model [36] as our baselines, which are the classical video understanding models for temporal action localization tasks. All these models are trained on our dataset under supervised learning and evaluated on the test set. Table 3 show the results of the temporal detection task on the PALADIN dataset.

Random Detection Results. We evaluate the performance of the random guess model on the PALADIN dataset, which serve as a lower bound for the following evaluation.

Table 4. 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
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
VideoMAE + AST + Text	19.64	17.22	14.76	12.45	10.32	14.88

First, we randomly guess the communication technique labels for each segment of the ground truth, and the mAP is 9.15%. This results aligns with the expected outcome for a ten-class classification problem.

When we randomly guess the boundaries of the segments and the corresponding communication technique labels, the average mAP@tIoU is 1.62%, which is much lower than the results of the model with training-based methods. Moreover, we use the TemporalMaxer model as the basic model for temporal boundary localization to predict the temporal boundaries of the video segments that contain communication techniques. And we randomly assign the communication technique category for each video segment. The results are shown in Table 3, where the random guess model achieves the average mAP@tIoU of 7.96%. Moreover, we report the results of assigning the majority label to the random generated, ground-truth, and predicted segments, respectively. The results show better results comparing to random labels. We also find that training-based models consistently outperform the random guess model, which indicates that the models can learn visual cues to predict communication techniques.

We observe that all the models evaluated achieve comparable performance in the PALADIN dataset, and all these methods have a similar tendency in terms of the mAP@tIoU metric. This is most likely caused by the fact that the models share a similar architecture, inspired by the architecture of the ActionFormer. However, most of the results are not satisfactory, compared to the results of the THUMOS-14 dataset [22] and the ActivityNet-1.3 dataset [2]. This shows that our task is more challenging than the traditional temporal action localization task. Furthermore, we compare the performance of our simple multi-modality model with these baselines, and the results are shown in Table 3. It has an average mAP@tIoU of 14.68%, which has a slight improvement over the TemporalMaxer model that is the second best model in the baselines. This indicates that the multi-modality information can provide additional cues to predict the communication techniques in the political advertisement videos. This trend also suggests that future work should focus on exploring multi-modal information fusion to better detect communication techniques.

Then, Table 4 includes the results of our method with different features. First, we observe that the VGGish features achieve the lowest performance, which indicates that the audio information is not powerful enough to predict the communication techniques in the political advertisement videos. This is also in line with our impression of the communication techniques in political advertisement videos, where the audio information is usually similar from the beginning to the end of the video, but the visual information is often changed significantly. When using video features, the performance is improved, indicating that visual information is more important than audio information in detecting the intention in political advertisement videos.

By combining the text features with the video or audio features, the performance has improved slightly, indicating that the text information can provide additional information to predict the communication techniques. But combining video and audio features can achieve more improvement, which indicates that video and audio information can complement each other in predicting communication techniques. Finally, by combining the VideoMAE, AST, and text features, the performance is the best, indicating that the introduced multimodal model can learn the cues from the political advertisement videos to predict the communication techniques.

5. Conclusion

This paper presents a new task in video understanding focused on detecting editing intentions in political advertisement videos. We defined these editing intentions as "communication techniques" of propaganda and classified them into ten distinct categories with the help of psychology experts. To support this task, we introduced a dataset comprising 12,526 political advertisement videos, each featuring multiple segments that illustrate various communication techniques. Additionally, we studied current video understanding models for identifying these editing intentions. Our findings demonstrate that the introduced dataset is instrumental in detecting communication techniques within political advertisement videos.

Acknowledgement This work was supported in part by JST CREST under Grant JPMJCR20D3.

References

[1] Porismita Borah, Erika Fowler, and Travis Nelson Ridout. Television vs. youtube: political advertising in the 2012 presidential election. *Journal of Information Technology & Politics*, 15(3):230–244, 2018. 4

[2] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 8

[3] 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. 6

[4] Giovanni Da San Martino, Yu Seunghak, Alberto Barrón-Cedeno, Rostislav Petrov, Preslav Nakov, et al. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 5636–5646. Association for Computational Linguistics, 2019. 2, 3, 4

[5] Giovanni Da San Martino, Yu Seunghak, Alberto Barrón-Cedeno, Rostislav Petrov, Preslav Nakov, et al. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 5636–5646. Association for Computational Linguistics, 2019. 2

[6] Ken Dancyger. *The technique of film and video editing: history, theory, and practice*. Routledge, 2018. 1

[7] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515, 2017. 2

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[9] Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. Semeval-2021 task 6: detection of persuasion techniques in texts and images. *arXiv preprint arXiv:2105.09284*, 2021. 2

[10] Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. Detecting propaganda techniques in memes. In *ACL*, pages 6603–6617, 2021. 2, 3

[11] 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. 6

[12] Christian Fuchs. Propaganda 2.0: Herman and chomsky's propaganda model in the age of the internet, big data and social media. In *Propaganda Model Today: Filtering Perception and Awareness: Filtering Perception and Awareness*, pages 71–91. University of Westminster Press London, 2018. 1, 2

[13] Zhengqi Gao, Fan-Keng Sun, Mingran Yang, Sucheng Ren, Zikai Xiong, Marc Engeler, Antonio Burazer, Linda Wildling, Luca Daniel, and Duane S Boning. Learning from multiple annotator noisy labels via sample-wise label fusion. In *European Conference on Computer Vision*, pages 407–422. Springer, 2022. 5

[14] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017. 6

[15] Tiantian Geng, Teng Wang, Jinming Duan, Runmin Cong, and Feng Zheng. Dense-localizing audio-visual events in untrimmed videos: A large-scale benchmark and baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22942–22951, 2023. 6

[16] Yuan Gong, Yu-An Chung, and James Glass. AST: Audio Spectrogram Transformer. In *Interspeech*, 2021. 6

[17] Marek Herde, Denis Huseljic, and Bernhard Sick. Multi-annotator deep learning: A probabilistic framework for classification. *Transactions on Machine Learning Research*, 2023. 5

[18] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017. 6

[19] Kristina Hristakieva, Stefano Cresci, Giovanni Da San Martino, Mauro Conti, and Preslav Nakov. The spread of propaganda by coordinated communities on social media. In *Proceedings of the 14th ACM Web Science Conference 2022*, pages 191–201, 2022. 1

[20] Kung-Hsiang Huang, Kathleen McKeown, Preslav Nakov, Yejin Choi, and Heng Ji. Faking fake news for real fake news detection: Propaganda-loaded training data generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 14571–14589, July 2023. 4

[21] Muhammad Nihal Hussain, Serpil Tokdemir, Nitin Agarwal, and Samer Al-Khateeb. Analyzing disinformation and crowd manipulation tactics on youtube. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 1092–1095. IEEE, 2018. 2

[22] Haroon Idrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The thumos challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding*, 155:1–23, 2017. 8

[23] Raj Jagtap, Abhinav Kumar, Rahul Goel, Shakshi Sharma, Rajesh Sharma, and Clint P George. Misinformation de-

tection on youtube using video captions. *arXiv preprint arXiv:2107.00941*, 2021. 2

[24] Licheng Jiao, Ruohan Zhang, Fang Liu, Shuyuan Yang, Biao Hou, Lingling Li, and Xu Tang. New generation deep learning for video object detection: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 33(8):3195–3215, 2021. 1

[25] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset, 2017. 6

[26] Yu Kong and Yun Fu. Human action recognition and prediction: A survey. *International Journal of Computer Vision*, 130(5):1366–1401, 2022. 1

[27] Ariel Victoria Lieberman. Terrorism, the internet, and propaganda: A deadly combination. *J. Nat'l Sec. L. & Pol'y*, 9:95, 2017. 1

[28] G Martino, Alberto Barrón-Cedeno, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. Semeval-2020 task 11: Detection of propaganda techniques in news articles. *arXiv preprint arXiv:2009.02696*, 2020. 2

[29] Clyde R Miller. The techniques of propaganda. from "how to detect and analyze propaganda," an address given at town hall. *The Center for learning*, 1939. 3

[30] Oxford English Dictionary. *propaganda, n.* Oxford University Press, July 2023, https://doi.org/10.1093/OED/1010698303. 2

[31] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 5

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 6

[33] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023. 6

[34] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2931–2937, 2017. 2

[35] Julian Richards. Extremist propaganda and the" politics of the internet". *The Journal of Intelligence, Conflict, and Warfare*, 3(3):22–33, 2021. 1

[36] Dingfeng Shi, Yujie Zhong, Qiong Cao, Lin Ma, Jia Li, and Dacheng Tao. Tridet: Temporal action detection with relative boundary modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18857–18866, 2023. 6, 7

[37] Kilian Sprenkamp, Daniel Gordon Jones, and Liudmila Zavolokina. Large language models for propaganda detection. *arXiv preprint arXiv:2310.06422*, 2023. 2

[38] S Shyam Sundar, Maria D Molina, and Eugene Cho. Seeing is believing: Is video modality more powerful in spreading fake news via online messaging apps? *Journal of Computer-Mediated Communication*, 26(6):301–319, 2021. 2

[39] Dídac Surís, Ruoshi Liu, and Carl Vondrick. Learning the predictability of the future. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12607–12617, 2021. 4

[40] Piotr Szymański and Tomasz Kajdanowicz. A scikit-based python environment for performing multi-label classification. *arXiv preprint arXiv:1702.01460*, 2017. 5

[41] Tuan N Tang, Kwonyoung Kim, and Kwanghoon Sohn. Temporalmaxer: Maximize temporal context with only max pooling for temporal action localization. *arXiv preprint arXiv:2303.09055*, 2023. 6, 7

[42] Ryutaro Tanno, Ardavan Saeedi, Swami Sankaranarayanan, Daniel C Alexander, and Nathan Silberman. Learning from noisy labels by regularized estimation of annotator confusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11244–11253, 2019. 5

[43] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022. 6

[44] Binglu Wang, Yongqiang Zhao, Le Yang, Teng Long, and Xuelong Li. Temporal action localization in the deep learning era: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 6

[45] 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. 6

[46] Jan Warchocki, Teodor Oprescu, Yunhan Wang, Alexandru Dămăcuş, Paul Misterka, Robert-Jan Bruintjes, Attila Lengyel, Ombretta Strafforello, and Jan van Gemert. Benchmarking data efficiency and computational efficiency of temporal action localization models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3008–3016, 2023. 6, 7

[47] Chloe Wittenberg, Ben M Tappin, Adam J Berinsky, and David G Rand. The (minimal) persuasive advantage of political video over text. *Proceedings of the National Academy of Sciences*, 118(47):e2114388118, 2021. 2

[48] Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10714–10726, 2023. 1

[49] Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, pages 492–510. Springer, 2022. 6, 7