MM-AU:Towards Multimodal Understanding of Advertisement Videos

Digbalay Bose
University of Southern California
Los Angeles, CA, USA
dbose@usc.edu

Rajat Hebbar
University of Southern California
Los Angeles, CA, USA
rajatheb@usc.edu

Tiantian Feng
University of Southern California
Los Angeles, CA, USA
tiantianf@usc.edu

Krishna Somandepalli*
Google Research
New York, NY, USA
ksoman@google.com

Anfeng Xu
University of Southern California
Los Angeles, CA, USA
anfengxu@usc.edu

Shrikanth Narayanan
University of Southern California
Los Angeles, CA, USA
shri@ee.usc.edu

Figure 1: Schematic diagram showing illustrative examples of various tasks in the MM-AU (Multi-modal ads understanding) dataset. Multimodal understanding of ads along the lines of (a) Topic categorization (18 classes) (b) Tone transition (c) Social message detection, i.e. Absence/Presence of social message. Image sources: [47], [8], [27]

ABSTRACT

Advertisement videos (ads) play an integral part in the domain of Internet e-commerce, as they amplify the reach of particular products to a broad audience or can serve as a medium to raise awareness about specific issues through concise narrative structures. The narrative structures of advertisements involve several elements like reasoning about the broad content (topic and the underlying message) and examining fine-grained details involving the transition of perceived tone due to the sequence of events and interaction among characters. In this work, to facilitate the understanding of advertisements along the three dimensions of topic categorization, perceived tone transition, and social message detection, we introduce a multimodal multilingual benchmark called ***MM-AU*** comprised of 8.4 K videos (147hrs) curated from multiple web-based sources. We explore multiple zero-shot reasoning baselines through the application of large language models on the ads transcripts. Further, we demonstrate that leveraging signals from multiple modalities, including audio, video, and text, in multimodal transformer-based supervised models leads to improved performance compared to unimodal approaches.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; **Computer vision**; • **Applied computing** → **Media arts**; • **Information systems** → *Multimedia databases.*

KEYWORDS

multimodal learning, media understanding, advertisements

ACM Reference Format:
Digbalay Bose, Rajat Hebbar, Tiantian Feng, Krishna Somandepalli, Anfeng Xu, and Shrikanth Narayanan. 2023. MM-AU:Towards Multimodal Understanding of Advertisement Videos. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3581783.3612371

1 INTRODUCTION

Media content aims to portray rich human-centered narrative structures through various forms and outlets, including advertisements, movies, news, and increasingly as digital shorts on social media.

*The work was done while the author was at USC.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10.
https://doi.org/10.1145/3581783.3612371

As a primary media source for disseminating information, advertisements (ads for short) have been utilized to promote products or convey messages such as about extant social or political issues. The utility of advertisements as a media source has been amplified by the wide variety of platforms like radio, television, newspaper print, video streaming, and social networking sites, presenting significant influences, whether direct or indirect, to viewers from diverse backgrounds [52]. The rising importance of advertisements in the current socio-economic scenario is evident from the expected increase in media ad spending [46] from 225.79 in 2020 to 322.11 billion dollars in 2024.

Objectively understanding the rich content in ads, and their impact on the viewer experience and behavior is hence of great interest. However, in enabling computational media understanding [62], advertisements present unique challenges in the form of condensed narrative structures [32]. Due to their relatively short duration when compared to feature length movies, an advertisement video showcases a particular narrative structure in a tightly-integrated sequence with different formats, including slice-of-life [42], drama [35], and transformational [54]. Further, reasoning about the narrative structure requires multi-scale understanding of the underlying topic and fine-grained elements, including the sequence of events (of characters and interactions) and related messages. As shown in Fig 1, the key elements associated with the narrative structure of ads and related challenges are listed as follows:

Topic: understanding enables personalized categorization and retrieval for customers along with key insights into the representation of genders [19] and different demographic groups with respect to target classes like healthcare, retail, travel, etc. Topic categorization involves the handling of both inter and intra-topic diversity between the videos in terms of human-object interactions and a wide variety of items/products, as shown in Fig 1 (a).

Tone transition: Affective tone associated with an advertisement video refers to the perceived feeling by a viewer [69]. Associating the appropriate tone with an ad enhances its persuasiveness, thus enabling the associated brand to expand its reach to a wide range of customers. While the positive tone centers around optimistic elements associated with hope and success [7], portrayals of negative tone are tied to sad narratives involving fear and suffering. However, due to the narrative structure, the perceived affective tone exhibits transitions during the duration of an advertisement video, accompanied by changes in visuals and background music. In Fig 1 (b), the video starts on a positive note with a happy person taking a picture, followed by a perceived negative tone in the middle due to the suffering of the person. The advertisement ends on a positive note, with the person being saved by an incoming vehicle.

Social message: Advertisements act as a major source of information about pressing social issues for consumers. Brands conveying messages about various social issues, including but not limited to gender inequalities, racial discrimination, and environmental conservation, are viewed favorably by consumers across different age groups [7]. In terms of advertisement videos, social message portrayal is characterized by a huge diversity in depiction, as shown in Fig 1 (c) due to underlying categories like smoking, accidents, gun violence etc.

In this work, we introduce a multilingual multimodal benchmark called ***MM-AU*** for the computational understanding of advertisement videos across the tasks of topic categorization, social message, and tone transition detection. Due to the inherent structure of the videos involving transitions in scenes, audio soundscapes, and diverse interactions among characters (text transcripts), we adopt a multi-modal approach to advertisement understanding. We outline our contributions below:

- **Topic classification:** We merge existing taxonomies for topic annotations from prior ads datasets and publicly available websites like Ads of the world [47] to obtain a condensed set of topic labels for the advertisement videos.
- **Tone Transition detection:** We introduce a novel benchmark task of tone transition detection in advertisement videos by obtaining crowdsourced feedback from human annotators.
- **Social message detection:** We provide weak human expert-based labels for detecting the presence/absence of social messages in advertisement videos.
- **Language-based reasoning:** We explore zero-shot baselines for the three benchmark tasks through applications of large-language models on ad transcripts.
- **Multimodal/Unimodal modeling:** We provide multiple unimodal and multimodal baselines to benchmark the performance for the three tasks (topic classification, transition detection, and social message) and highlight future possibilities of explorations.

2 RELATED WORK

Narrative understanding: Narratives [16] play an important role in enabling effective human communication and organizing the daily sequence of events. Advertisements centered on narratives [15] influence consumers by providing a concrete story arc centered around specific themes, protagonists and their actions. Kim et al. [32] introduced an integrated theory of understanding narratives in advertisements based on key variables like emotic response, ad hedonic value, ad credibility, and perceived goal facilitation. Lien et al. [38] explored narrative ads from the lens of persuasion and the relationship with different advertisement mediums - verbal or visual. In the realm of computational narrative understanding, prior works have focused on language-based approaches for marking high-level structures in short stories [36], most reportable events (MRE) in Reddit comment threads [50] and primary processes in movie scripts, newspaper articles, etc [6].

Affect modeling in videos: Advertisement brands tend to invoke emotional reactions [25] in viewers by influencing their actions i.e., purchasing a particular product. In the domain of television commercials, a combination of physiological, symbolic, and self-report measures was explored in [43] to determine the emotional responses of viewers. The role of facial expressions in decoding the preferences of viewers, including purchase intent, and smile responses, has been explored through large-scale studies in [41], [66]. Apart from facial expressions, the role of CNN-based audio-visual and EEG descriptors from the viewers has been explored in a multi-task setup [57–59] for arousal and valence prediction in advertisement videos. Existing video-based affect datasets like DEAP [34], VideoEmotion [31] also focused on single arousal, valence, and dominance ratings as well as discrete emotion labels for music

and user-generated videos, respectively.
In the domain of continuous affect modeling, datasets with frame-level annotations have been introduced across a wide variety of domains, including naturalistic and induced clips (HUMAINE [14]), movies (COGNIMUSE [76], LIRIS-ACCEDE [4]) and online videos (EEV [64]). Further extensions of continuous affect modeling based on independent and self-reports have been explored for daily emotional narratives in SENDv1 dataset [48]. For advertisements, climax annotations (presence + rough timestamps) were provided by human annotators on a subset of the Video Ads dataset [27] in [71] along with climax-driven modeling strategies to predict sentiment labels at the video level. In our proposed benchmark ***MM-AU***, based on the standard definition in [7], we ask human annotators to denote the perceived tone in the advertisement video across segments approximately marking the start, middle, and end. The perceived tone transition enables tracking of the narrative dynamics in advertisement videos by considering the interactions between different modalities, including audio, visual, and narrations/spoken interactions (through transcripts).

Advertisement datasets: While there has been progress in terms of movie understanding due to the introduction of large-scale multimodal datasets like Condensed Movies [3], MovieNet [26], MAD [61], Movie-cuts [51], MovieCLIP [5] and SAM-S [22], only few datasets have focused on large-scale understanding of advertisements across broad and fine-grained content. Hussain et al. [27] introduced the benchmark Video-Ads dataset to faciliate understanding of images and videos along the lines of broad topics, induced sentiments and action/intent reasoning. The images in the Video-Ads dataset were utilized in [60] for computational modeling of persuasion across 21 categories in marketing domain. Regarding large scale ads understanding, Tencent-AVS dataset [30] was proposed to enable multi-modal scene level categorization into semantic classes like presentation, places and styles. While the previously mentioned datasets focused on classification tasks, E-MMAD [74] introduced the task of informative caption generation from advertisements across 120k e-commerce videos.

MM-AU, our curated multilingual dataset utilizes videos from Ads-of-the-world [47] along with a subset from Video-Ads dataset and an in-house video catalog from Cannes Lion archive [8]. We provide 18 broad topic categories by combining existing taxonomies (Cannes Lion, Ads-of-the-world and Video-Ads dataset). Further, we rely on human expert annotators to label transitions in perceived tone along with the absence/presence of social messages in 8.4K advertisement videos. A comparative overview of ***MM-AU*** and other advertisement datasets is shown in Table 1.

Semantic video understanding: Existing large-scale video datasets including Kinetics[9], Moments-in-time [44], ActivityNet [23], AVA [20] have focused mainly on classifying entity driven actions from in-the-wild short videos. Higher-level semantic labels beyond actions like topics, concepts, events and video types were exploited for large-scale video-level categorization in datasets like Youtube-8M [1], Holistic-visual understanding (HVU) [13] and 3MASSIV [21]. Our proposed benchmark ***MM-AU*** explores the domain of semantic video understanding in ads by considering broad categories like topic, presence/absence of social message, and fine-grained affective labels of perceived tone transition.

Multimodal representation learning: Multimodal representation learning [37] centers around the fusion of information from different modalities at multiple scales, including early, late, and mid-fusion. Prior works related to ads have utilized multimodal-LSTMs [68], segment-level autoencoders [63] or joint cross-modal embedding [72] approaches for learning multimodal representations for a variety of tasks. With the advent of transformer [67] based multimodal models like PerceiverIO [29], attention bottlenecks[45], and VATT [2], information fusion at the input token space, followed by joint encoders, have become more prevalent. A multi-task attention-based approach was explored in [73] for jointly predicting topic and sentiment labels associated with advertisement images. A NextVLAD [39] based approach [70] combined with global-local attention was utilized for predicting scene-specific labels in the Tencent [30] ads benchmark dataset.

3 MM-AU DATASET

3.1 Data sources

We consider multiple ads-specific sources for curating our proposed MM-AU dataset. As a primary source, we consider Ads-of-the-world (AOW) [47] video hosting website since it contains a richly-curated catalog of ads in various formats like film, print, digital, and video spanning across multiple countries. As auxiliary sources, we consider additional videos from the Cannes Lion Film Festival [8] archive and Video-Ads dataset [27]. We filter the videos based on unique video ids associated with their public links to ensure no duplicates across three sources. The share of different sources in curating the combined list of 8399 advertisement videos is shown in Fig 2 We employ a semi-automatic process for tagging

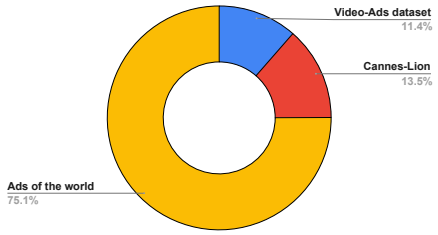


Figure 2: Share of different ad sources in MM-AU dataset. Ads of the world (6304 videos), Cannes Lion (1135), Video-Ads dataset (960)

the advertisement videos with broad topic categories. For the detection tasks of tone transition and social message, we use Amazon Mechanical Turk [1] to obtain responses from a pool of 36 human annotators. For selecting a pool of workers with the requisite expertise, we hosted an initial pilot study where the workers are instructed to mark the tone transition labels and presence/absence of social message in the given set of videos. Further, in the final annotation phase, three annotators independently annotate each sample for the tone-transition and social message detection tasks. An outline of the annotation process associated with tone transition and social message detection is provided in the Supplementary (Section 2.2).

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

Dataset	Annotation type	Duration	#Samples	#Shot	#Class	Modalities	Languages	Tasks
Video Ads Dataset (I)	H	NA	64832	NA	38 (T), 30 (S), AR (OE), H(2), Ex(2)	Images	English	Image level classification
Video Ads Dataset (V)	H	144.87	3477	NA	38 (T), 30 (S), AR (OE), H(2), Ex(2)	Video	English	Video level classification
Tencent-AVS	H	142.1h	12k	121.1k	25 (Pr), 34 (St), 23 (Pl)	Video/Audio/ASR/OCR	Chinese and English	Scene level classification
Ads-persuasion dataset	H + AL	NA	3000	NA	21 (PS)	Images	English	Image level classification
E-MMAD	SG descriptions	1021.6 h	120984	NA	4863 (PC)	Video	Chinese and English	Video level captioning
MM-AU	**H + SA**	**147.8h**	**8399**	**216.4k**	**18 (T), 3 (Tone), 2 (SM)**	**Video/Audio/ASR**	**Multilingual (65 languages)**	**Video level classification**

Table 1: Comparison of *MM-AU* with other available advertisement datasets across different modalities. Annotation type: H: Human annotation, AL: Active Learning, SA: Semi-automatic, SG: Store generated. Duration: NA: Not applicable for images; Mentioned in hours(h). #Samples: Number of video clips or images. #Shot: NA: Not applicable for images; Number of shots detected from all the video samples. #Class: T: Topic, S: Sentiment, OE: Open-Ended, H: Humor, Ex: Exciting, Pr: Presentation, St: Style, Pl: Place, PS: Persuasion strategy, PC: Product categories, SM: Social messages

The annotation process details associated with the respective tasks are listed below:

Tone transition: The annotators are instructed to mark the perceived tone labels associated with the start, middle, and ending segments of the advertisement videos. To reduce the burden associated with the task, no instructions are provided to mark the timestamps associated with the respective segments. Based on the tone definition considered in [7], we provide the following descriptions to aid the annotation process:

- **Positive tone:** An advertisement video segment has a positive tone if it contains: *optimistic elements portraying hope and success* or *positive imagery based on uplifting music and visuals*. Examples include girls overcoming negative stereotypes and succeeding in sports or a blind person being able to navigate easily through city roads by using an app.
- **Negative tone:** An advertisement video segment has a negative tone if it contains: *sad narrative showing suffering, fear, destruction* or *depressing music and distressing visuals*. Salient themes associated with negative tone include domestic violence, environmental damage, human trafficking, crisis from wars etc.

If a segment does not contain the above-mentioned characteristics, the annotators are instructed to mark the perceived tone as neutral. To determine the reasoning involved in marking the tone labels associated with the segments, the annotators are also asked to provide explanations regarding their choices.

Social message detection: For social message detection, the annotators are instructed to check for the absence/presence of social messages in the given video. Based on the social message framing in ads [7], we provide the following definition to guide the annotation process:

- An advertisement video has a social message if it provides awareness about any social issue. Examples include *gender equality, drug abuse, police brutality, workplace harassment, domestic violence, child labor, homelessness, hate crimes* etc.

To simplify the annotation process, we ask the annotators to mark Yes/No for indicating the presence/absence of social messages in the videos instead of marking the exact categories in the curated list of social issues [11].

Topic categorization: We annotate topic categories using the existing taxonomies from Ads-of-the-world (AOW), Cannes Lions Film Festival [8], and Video-Ads [27] datasets. We denote the taxonomies associated with Cannes Lions Film Festival and Video-Ads datasets as Cannes-coding [**CC**] and Video-Ads [**VA**] coding schemes. We extract the available tags associated with 6304 videos in Ads-of-the-world [**AOW**] and retain the top 40 tags based on frequency. Then we manually merge the filtered topic tags from AOW with similar labels in Cannes-coding [**CC**] and Video-Ads [**VA**] coding schemes. Some examples of merged labels from different sources are listed as follows with the final parent topic category:

- **Publications media:** Media & Publications [**CC**]; Media and arts [**VA**]; TV Promos, Music, Media, Movies [**AOW**]
- **Games:** Games and toys [**VA**]; Gaming [**AOW**]
- **Sports:** Sports equipment and activities [**VA**]; Sports [**AOW**]
- **Clothing:** Clothing, Footwear & Accessories [**CC**]; Clothing and accessories [**VA**]; Personal Accessories [**AOW**]

A detailed list of mapping between the AOW, CC and VA coding schemes is included as part of the Supplementary (Section 2.1). Our final merged topic taxonomy consists of 18 categories as follows:

- *Games, Household, Services, Misc, Sports, Banking, Clothing, Industrial and agriculture, Leisure, Publications media, Health, Car, Electronics, Cosmetics, Food and Drink, Awareness, Travel and transport, Retail*

Dataset Filtering: During the annotation process, we employ certain checks to maintain the quality of the annotated data. We reject those tone transition annotations with very short explanations (single words) or long generic descriptions of ads copied from the internet. Further, we also flag tone-transition annotations with the copied content across the start, middle, and end segments. For topic categorization, we merge categories with low frequencies, i.e., Alcohol and Restaurant, into the broad category of Food and Drink.

3.2 Dataset analysis

MM-AU consists of 8399 annotated videos with a total of 147 hours of curated data. A detailed overview of MM-AU with total duration, number of tags, and year coverage is shown in Table 2. The distribution of topics is shown in Fig 4, with Food and Drink, Awareness, and Electronics being the top-3 dominant categories. In the case of perceived tone labels, we obtain a high majority agreement among annotators in marking the start (**91.2%**), middle (**91.6%**), and the ending (**94.5%**) segments of the videos with perceived tone labels.

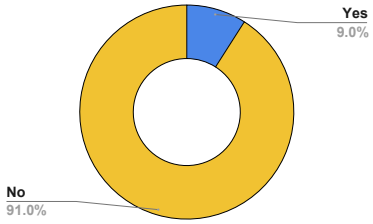


Figure 3: Distribution of the social message absence (No) and presence (Yes) labels in MM-AU

Attribute	Value
#videos	8399
#explanations	74970
#topics	18
#social msg labels	25197
#tone labels	75,591
#duration	147.79 hrs
#avg duration	63.35s
year	2006-2019
#annotators	36
#countries	99

Table 2: Data statistics of MM-AU dataset. #social msg labels: total number of labels

Since annotating the presence/absence of social messages is a comparatively less subjective task than perceived tone labeling, we obtain a majority agreement (**99%**) among the annotators. In terms of tone labels for start, middle, and end segments, we can see from Fig 5, that the dominant perceived tone for the advertisements is positive, with its share rising from **60.2%** (start) to **81.3%** (end). This can be explained due to the fact that advertisements are primarily designed to persuade viewers to buy certain products or act toward certain social issues. However, from Fig 5, we can see that the share of negative tone labels increases from **15.5%** to **19.6%** due to the narrative structure of the ads, where the middle segment portrays negative elements like human suffering, environmental damage, etc to set up the final conclusion. From Fig 3, we can see that **9.0%** of the videos, i.e. 759 contain social messages, as marked by **Yes** label. Out of 759 videos, **62.5%** exhibit transition in perceived tone with the share of negative tone rising from **32.3%** to **43.3%** in the middle segments. Further intersectional views of social message presence with different topics and different segments (start, middle, and end) are included as part of the Supplementary (Figures 5 and 6).

4 MULTIMODAL REPRESENTATIVE TASKS

For defining the multimodal representative tasks, we denote the audio, video, and text representations associated with the ith video sample as e_a, e_v, e_t.

Social message detection: Based on the social message annotations (majority), the presence/absence of social message (SM) for ith video is defined as:

$$SM_i = \begin{cases} 0 & \text{No (Absence of SM)} \\ 1 & \text{Yes (Presence of SM)} \end{cases} \tag{1}$$

Our aim is to learn a multimodal network $f_{SM}(e_a, e_v, e_t)$ to predict social message presence/absence $\hat{SM}_i$ for the ith video. The above definition results in 759 and 7640 videos marked with the presence (1) and absence (0) of social messages, respectively.

Tone transition: Based on the start, middle, and end perceived tone labels (majority), we define the transition for ith video as follows:

$$Tr_i = \begin{cases} 0 & Start_i = Mid_i = End_i \\ 1 & \text{else} \end{cases} \tag{2}$$

Our aim is to learn a multimodal network $f_{Tr}(e_a, e_v, e_t)$ to predict binary tone transition $\hat{Tr}_i$ for the ith video. MM-AU dataset has 3854 and 4545 videos marked with Transition (1) and No Transition (0), respectively.

Topic categorization: For topic categorization, we aim to learn a multimodal network $f_{Topic}(e_a, e_v, e_t)$ to predict $\hat{Topic}_i$ for the ith video out of 18 target categories.

5 PROPOSED METHOD

We propose a two-stage method to combine multiple modalities i.e., audio, video, and text, in a transformer-based framework. For the transformer encoder, we use the PerceiverIO [29] architecture as part of our design choice due to its generalization capabilities to a wide variety of inputs. For ith sample consisting of video (**v**), audio (**a**), text (**t**), the two-stage operation can be summarized as follows:

Stage 1: Stage 1 involves complete finetuning of the T-V and A-V transformer encoder models indicated by $\mathbf{Tx_{TV}}$ and $\mathbf{Tx_{AV}}$. The text (**t**), audio (**a**), video (**v**) are passed through the respective encoders f_{text}, f_{audio} and f_{visual} respectively. The sequence of operations in Stage 1 can be summarized as follows:

$$\begin{aligned} e_{text} &= Text_{proj}(f_{text}(t)) \\ e_{vis} &= Vis_{proj}(f_{visual}(v)) \\ TV_{logits} &= \mathbf{AvgPool}(\mathbf{Tx_{TV}}([e_{text}; e_{vis}])) \\ e_{aud} &= Aud_{proj}(f_{audio}(a)) \\ AV_{logits} &= \mathbf{AvgPool}(\mathbf{Tx_{AV}}([e_{vis}; e_{aud}])) \end{aligned} \tag{3}$$

Here $Text_{proj}$, Vis_{proj}, Aud_{proj} are linear projection layers used for mapping the outputs of respective modality-specific encoders to the same dimensions.[.;.] indicates concatenation along the temporal dimension.

Stage 2: In stage 2, we freeze the two transformer encoder models $\mathbf{Tx_{TV}}$ and $\mathbf{Tx_{AV}}$ and combine the respective logit outputs through the following strategies:

- **A-max:** $pred_{class} = \arg\max_i (TV_{logits}(i) + AV_{logits}(i))/2$
- **D-max:** $pred_{class} = \arg\max_i max\{(TV_{logits}(i), AV_{logits}(i))\}$

Here $i \in \{1, ...Class\}$, where $Class$ refers to the number of classes associated with the task. A detailed outline of our proposed approach is shown in Fig 6.

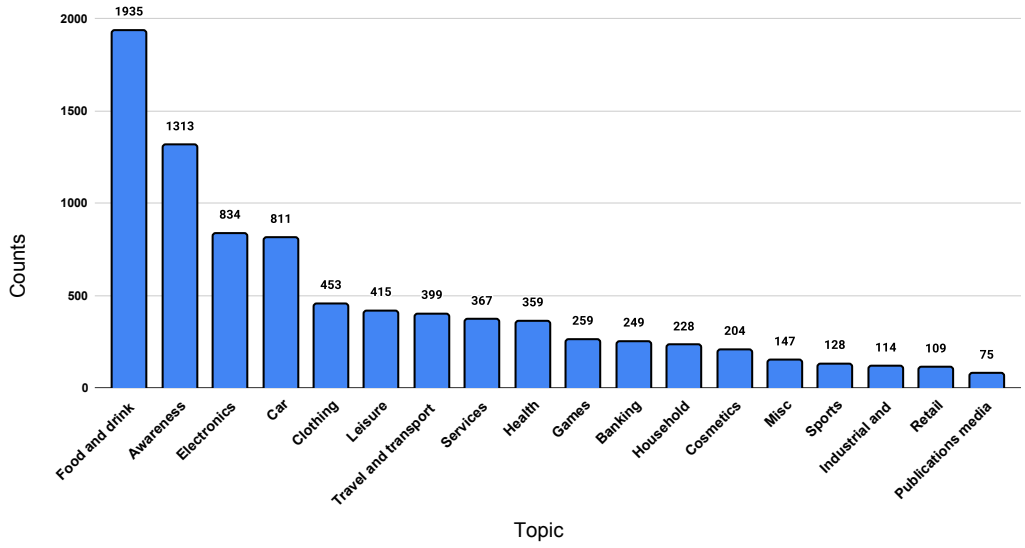
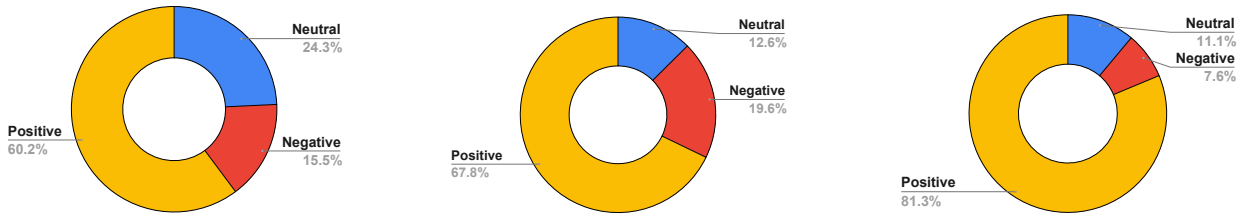


Figure 4: Distribution of topics in MM-AU dataset



(a) Start segment **(b) Middle segment** **(c) End segment**

Figure 5: Distribution of majority perceived tone labels (among 3 annotators) across (a) start, (b) middle, and (c) ending segments in MM-AU dataset (8399 videos)

6 EXPERIMENTS

6.1 Experimental Setup

For training, validation, and testing purposes, we consider a split of 5877 (70%), 830 (10%), and 1692 (20%) videos. For the visual modality, we segment the videos into shots using `PySceneDetect`[2] followed by extraction of frame-wise features at 4 fps using CLIP's [55] pretrained visual encoder (f_{visual}), `ViT-B/32`. Further, we average pool the frame-wise visual representations to obtain shot representations.

For the text modality, we use Whisper [56] multilingual model (`large`) to extract the transcripts. A detailed split of the languages available with the transcripts is shown in the Supplementary (Figure 9). We translate the multilingual transcripts to English using `GPT-4` [49] (*temp=0.02, max tokens=2048*) by providing the following translation-specific prompt: ***Please provide an English translation of this transcript***. We use the pretrained BERT[12] model as the text encoder (f_{text}). For the audio modality, we use the Audio-Spectrogram transformer (AST) [18] model (f_{audio}) pretrained on AudioSet [17] for extracting features at 10-sec intervals with a step size of 512. We conduct our experiments in a distributed manner using the Pytorch [53] framework on 4 2080ti GPUs. For evaluation metrics, we use accuracy and macro-F1.

[2]https://github.com/Breakthrough/PySceneDetect

6.2 Language-based reasoning

We investigate the zero-shot capabilities of foundational large language models [75] by applying `GPT-4` [49], `Opt-IML` [28], `Flan-T5` (XXL,XL,L) [10] and `Alpaca` [65] on the translated transcripts. For zero-shot evaluation, we report the results on 1670 non-empty transcripts out of the test split of 1692 samples. For `GPT-4` we use the following task-specific prompts:

- **SM:** *An advertisement video has a social message if it provides awareness about any social issue. Example of social issues: gender equality, drug abuse, police brutality, workplace harassment,*

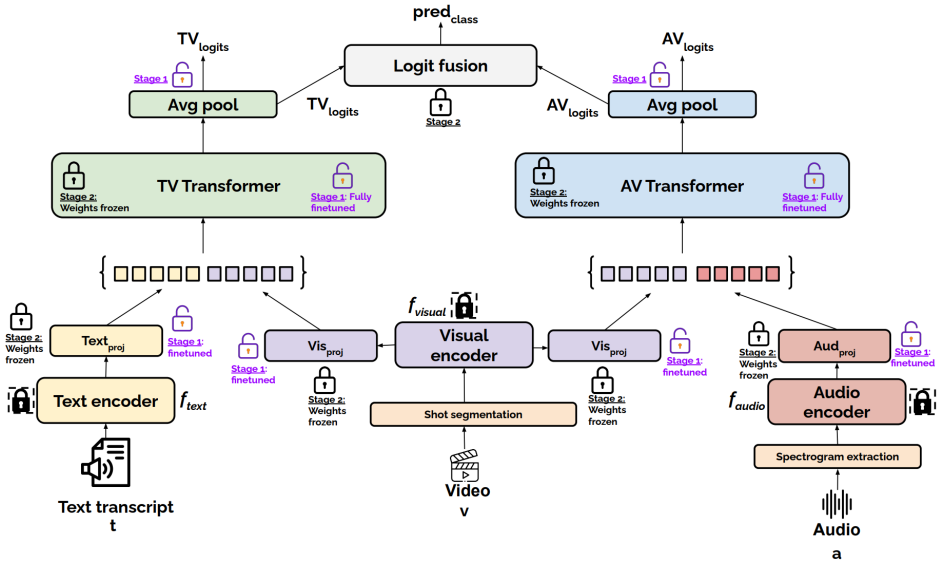


Figure 6: The proposed two-stage approach for fusing audio-visual (AV) and text-visual (TV) transformers. Stage 1: Full finetuning. Stage 2: Weights are completely frozen. The text, visual and audio encoders are kept completely frozen.

domestic violence, child labor, environmental damage, homelessness, hate crimes, racial inequality etc. Based on the given text transcript, determine if the advertisement has any social message. Please provide answers in ***Yes*** *and* ***No.***

- **TT:** *Based on the given text transcript from the advertisement, determine if the advertisement has any transitions in tones. Possible tone labels are: positive, negative, and neutral. Please respond by saying* ***Transition*** *or* ***No transition***
- **Topic:** *Associate a single topic label with the transcript from the given set: <Topic list>*

Here SM, TT, and Topic refer to the benchmark tasks of Social message detection, tone transition, and topic categorization. <Topic list> refers to the condensed list of 18 topic categories curated for MM-AU dataset. Further details about the prompting strategies are included as part of the Supplementary (Section 3.1).

6.3 Unimodal and Multimodal baselines

For the supervised unimodal baselines we consider the following model choices:

- **LSTM [24]:** 2 layers and hidden dimension = 256
- **MHA [67]:** 4 layers, 4 heads and hidden dimension = 256

For multimodal models, we use Perceiver IO structure [29] as the attention-based transformer encoder to explore combinations of different paired modalities i.e., audio-visual ($\mathbf{Tx}_{\mathbf{AV}}$), text-visual ($\mathbf{Tx}_{\mathbf{TV}}$), audio-text ($\mathbf{Tx}_{\mathbf{AT}}$). For the perceiver IO blocks, we adopt a light-weight structure composed of 4 encoder layers, 16 latent vectors, 8 heads, and a hidden latent dimensionality of 256. We use binary cross-entropy for social message and tone transition detection tasks and multi-class cross-entropy for topic categorization. For training the unimodal and multimodal models, we use a batch size of 16 with Adam [33] or AdamW [40] as optimizers and $lr \in \{1e-4, 1e-5\}$.

While training the supervised models, we fix the maximum sequence lengths for visual(shots), audio, and text modalities at 35, 14, and $\{256, 512\}$, respectively. During multimodal fusion, when text is missing in the transcripts due to the absence of speech, we replace the text with a string of `[MASK]` tokens. A detailed breakdown of model-wise hyperparameters is included in the Supplementary (Section 3.2).

Configurations		SM		TT		Topic	
Model	**Params**	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**
GPT-4 [49]	NA^*	**87.6**	**65.66**	**58.56**	**58.33**	**33.29**	**29.21**
Flan-T5-XXL [10]	11B	85.69	62.7	54.79	44.15	30.54	24.23
Flan-T5-XL [10]	3B	65.51	49.31	54.67	42.39	27.18	24.1
Alpaca [65]	7B	10.77	10.56	46.88	39.1	11.19	11.68
Opt-IML [28]	1.3B	37.07	32.49	54.37	35.22	22.22	19.08
Flan-T5-L [10]	780M	8.32	7.76	54.43	35.25	26.82	19.42
Random Baseline	-	49.57	39.61	49.95	49.88	5.71	4.71
Majority Baseline	-	90.96	47.63	54.11	35.11	23.04	2.08

Table 3: Zero shot performance comparison between various LLMs on MM-AU dataset. Tasks: *SM*: Social message detection, *TT*: Tone transition, *Topic*: Topic categorization. NA^* : Information not available. F1: Macro-F1. Best performing results are marked in bold for respective tasks.

7 RESULTS

7.1 Language-based reasoning

Based on results in Table 3, we can see that `GPT-4` exhibits superior zero-shot performance (F1 and Accuracy) across all the tasks when compared to other large language models. Further, there is a trend towards improved model performance with model scaling, except for `Alpaca`. The poor performance of `Alpaca` can be attributed to a lack of self-instruct data associated with complex reasoning

Configurations			Acc	F1	Acc	F1	Acc	F1
Model	**Features**	**Modality**	**Social message**		**Tone transition**		**Topic categorization**	
Random	NA	NA	49.57±0.28	39.61±0.28	49.95±0.30	49.88±0.30	5.71±0.16	4.71±0.24
Majority	NA	NA	90.96	47.63	54.11	35.11	23.04	2.08
Unimodal								
LSTM	CLIP-S	V	90.57±0.47	68.65±1.70	61.93±0.37	61.65±0.37	52.86±0.43	36.48±1.58
MHA	AST	A	89.07±2.02	55.33±3.96	59.78±1.56	58.60±0.96	25.11±1.39	15.48±1.57
MHA	CLIP-S	V	90.41±2.24	72.28±1.66	61.74±1.07	61.48±1.11	61.30±0.89	47.74±1.27
Multimodal								
$\mathbf{Tx_{AT}}$	AST + BERT	A + T	90.24±0.81	64.02±0.81	62.98±0.56	62.29±0.83	42.99±0.65	30.77±0.96
(1) $\mathbf{Tx_{AV}}$	CLIP-S +AST	A + V	91.62±0.58	70.05±0.67	64.01±0.54	63.72±0.66	61.62±0.46	48.67±0.64
(2) $\mathbf{Tx_{TV}}$	CLIP-S +BERT	T + V	**92.23±0.57**	**74.03±1.00**	63.96±0.99	63.48±0.84	63.27±0.59	50.58±1.32
A-Max(1,2)	CLIP-S +BERT +AST	A + V + T	92.51±0.46	73.17±1.00	**65.05±0.36**	**64.67±0.33**	**65.92±0.54**	**54.22±1.14**
D-Max(1,2)	CLIP-S +BERT +AST	A + V + T	92.52±0.46	73.21±0.98	65.01±0.39	64.63±0.32	65.51±0.58	53.67±1.24

Table 4: Comparative results between different unimodal and multimodal models across different tasks: Social message and Tone transition, Topic categorization. *CLIP-S*: Shot level features extracted using CLIP. *Modality:* A: Audio, V: Visual, T: Text. Results are reported as an average of 5 runs with randomly selected seeds. Best performing results are marked in bold for respective tasks.

tasks from transcripts. Instruction finetuning coupled with model scaling improves the zero-shot performance (F1) of T-5 models from **49.31%** to **62.7%** and **42.39%** to **44.15%** for social message and tone transition tasks respectively.

7.2 Unimodal and Multimodal baselines

From Table 4, we can see that supervised unimodal models (MHA, LSTM) show improved performance as compared to simple random and majority baselines. In terms of unimodal models, MHA model trained on shot-level visual features (denoted by CLIP-S) perform far better than audio features (AST) in social message detection (**F1: 72.28** vs **F1: 55.33**) and topic categorization tasks. (**F1: 47.74** vs **F1: 15.48**). For the tone transition task, MHA model trained on audio features (AST) shows close performance (**F1:58.60** vs **F1: 61.48**) as compared to visual features, due to the dependence of the tone transition task on the ambient music.

For multimodal models, we observe that the fusion of text and visual modalities through a Perceiver-IO-based encoder ($\mathbf{Tx_{TV}}$) performs better (**F1:74.03**) for social message detection compared to audio-visual or audio-text fusion. This can be attributed to the presence of socially-relevant descriptors in the transcripts and video shots. However, the fusion of audio with visual signals ($\mathbf{Tx_{AV}}$) improves the performance in the tone-transition detection task (**F1: 63.72**). For topic categorization, we find that the fusion of text and visual modalities ($\mathbf{Tx_{TV}}$) performs better than other paired modalities (**F1:50.58**) due to topic-specific identifiers in transcripts and shots.

Our proposed approach based on logit fusion strategies: Average-Max (**A-Max**) and Dual-Max (**D-Max**) exhibits similar performance across all tasks. We obtain gain for tone-transition (**F1:64.67**) based on the **A-Max** fusion strategy of $\mathbf{Tx_{AV}}$ and $\mathbf{Tx_{TV}}$ models. In terms of class-wise metrics, **A-Max** fusion improves the transition class average F1-score to **61.28%** as compared to $\mathbf{Tx_{AV}}$ (**60.72%**) and $\mathbf{Tx_{TV}}$ (**60.18%**). In the case of topic categorization, logit fusion through **A-Max** of $\mathbf{Tx_{AV}}$ and $\mathbf{Tx_{TV}}$ models results in the best performance (**F1: 54.22**). Further, **A-Max** fusion results in improvements over $\mathbf{Tx_{AV}}$ and $\mathbf{Tx_{TV}}$ across 15 topic categories (out of 18), with noticeable gains in the minority categories i.e., Retail (∼**6.2%**), Industrial & Agriculture (∼**5%**), Household (∼**7%**). Similar trends can be observed for **D-Max** fusion, with improvements obtained across 14 topic categories (out of 18). A detailed class-wise breakdown for topic categorization is included as part of the Supplementary (Figure 13).

8 CONCLUSIONS

We introduced a multimodal multilingual ads dataset called MM-AU, and presented the core benchmark tasks of topic categorization, tone transition, and social message detection in advertisements. We use human experts to annotate the presence of underlying social messages and perceived tone for different segments of advertisement videos. For a broader content understanding, we also propose a condensed taxonomy of topics by combining information from multiple ads-specific expert sources. We also investigate language-based reasoning through the use of large-language models on the ads transcripts. Furthermore, we show the utility of exploiting multiple modalities (audio, video and text) in a transformer-based fusion approach to obtain state-of-the-art results in the MM-AU benchmark tasks. Future directions would include: (a) expansion to additional benchmark tasks e.g., prediction of user intent, (b) understanding of causal mechanisms associated with tone transition and the (c) exploration of multimodal foundational models to tag and summarize ad videos with possible explanations.

ACKNOWLEDGMENTS

We would like to thank the Guggenheim Foundation for supporting the study.

REFERENCES

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Apostol Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. YouTube-8M: A Large-Scale Video Classification Benchmark. *ArXiv* abs/1609.08675 (2016).

[2] Hassan Akbari, Linagzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. 2021. VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text. In *Neural Information Processing Systems.*

[3] Max Bain, Arsha Nagrani, Andrew Brown, and Andrew Zisserman. 2020. Condensed movies: Story based retrieval with contextual embeddings. In *Proceedings of the Asian Conference on Computer Vision*.

[4] Yoann Baveye, Emmanuel Dellandréa, Christel Chamaret, and Liming Luke Chen. 2015. LIRIS-ACCEDE: A Video Database for Affective Content Analysis. *IEEE Transactions on Affective Computing* 6 (2015), 43–55.

[5] Digbalay Bose, Rajat Hebbar, Krishna Somandepalli, Haoyang Zhang, Yin Cui, Kree Cole-McLaughlin, Huisheng Wang, and Shrikanth Narayanan. 2023. MovieCLIP: Visual Scene Recognition in Movies. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2083–2092.

[6] Ryan L. Boyd, Kate G. Blackburn, and James W. Pennebaker. 2020. The narrative arc: Revealing core narrative structures through text analysis. *Science Advances* 6 (2020).

[7] Mary Elizabeth Brooks, Clay M. Craig, and Shannon L. Bichard. 2020. Exploring Ads of the World: How Social Issues Are Framed in Global Advertisements. *Howard Journal of Communications* 31 (2020), 150 – 170.

[8] Cannes. 2017. Cannes Lions. https://www.canneslions.com/

[9] João Carreira and Andrew Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), 4724–4733.

[10] Hyung Won Chung, Le Hou, S. Longpre, et al. 2022. Scaling Instruction-Finetuned Language Models. *ArXiv* abs/2210.11416 (2022).

[11] James Ciment. 2006. *Social Issues in America: An Encyclopedia.* M.E. Sharpe, Armonk, NY.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv* abs/1810.04805 (2019).

[13] Ali Diba, Mohsen Fayyaz, Vivek Sharma, Manohar Paluri, Jürgen Gall, Rainer Stiefelhagen, and Luc Van Gool. 2020. Large scale holistic video understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*. Springer, 593–610.

[14] Ellen Douglas-Cowie, Roddy Cowie, Ian Sneddon, Cate Cox, Orla Lowry, Margaret McRorie, Jean-Claude Martin, Laurence Devillers, Sarkis Abrilian, Anton Batliner, Noam Amir, and Kostas Karpouzis. 2007. The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data. In *Affective Computing and Intelligent Interaction*.

[15] Jennifer Edson Escalas. 1998. ADVERTISING NARRATIVES: What are they and how do they work?

[16] Walter R. Fisher. 1987. Human Communication As Narration: Toward a Philosophy of Reason, Value and Action.

[17] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio Set: An ontology and human-labeled dataset for audio events. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2017), 776–780.

[18] Yuan Gong, Yu-An Chung, and James Glass. 2021. AST: Audio Spectrogram Transformer. In *Proc. Interspeech 2021*. 571–575. https://doi.org/10.21437/Interspeech.2021-698

[19] Google. [n. d.]. Diversity and inclusion in advertisement videos. https://www.thinkwithgoogle.com/feature/diversity-inclusion/?vertical=All

[20] Chunhui Gu, Chen Sun, Sudheendra Vijayanarasimhan, Caroline Pantofaru, David A. Ross, George Toderici, Yeqing Li, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. 2017. AVA: A Video Dataset of Spatio-Temporally Localized Atomic Visual Actions. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2017), 6047–6056.

[21] Vikram Gupta, Trisha Mittal, Puneet Mathur, Vaibhav Mishra, Mayank Maheshwari, Aniket Bera, Debdoot Mukherjee, and Dinesh Manocha. 2022. 3MASSIV: Multilingual, Multimodal and Multi-Aspect dataset of Social Media Short Videos. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), 21032–21043.

[22] Rajat Hebbar, Digbalay Bose, Krishna Somandepalli, Veena Vijai, and Shrikanth S. Narayanan. 2023. A dataset for Audio-Visual Sound Event Detection in Movies. *ArXiv* abs/2302.07315 (2023).

[23] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. ActivityNet: A large-scale video benchmark for human activity understanding. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), 961–970.

[24] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9 (1997), 1735–1780.

[25] Morris B. Holbrook and John J. O'Shaughnessy. 1984. The role of emotion in advertising. *Psychology & Marketing* 1 (1984), 45–64.

[26] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. 2020. Movienet: A holistic dataset for movie understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*. Springer, 709–727.

[27] Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. 2017. Automatic Understanding of Image and Video Advertisements. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), 1100–1110.

[28] Srinivas Iyer, Xiaojuan Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, Xian Li, Brian O'Horo, Gabriel Pereyra, Jeff Wang, Christopher Dewan, Asli Celikyilmaz, Luke Zettlemoyer, and Veselin Stoyanov. 2022. OPT-IML: Scaling Language Model Instruction Meta Learning through the Lens of Generalization. *ArXiv* abs/2212.12017 (2022).

[29] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Andrew Brock, Evan Shelhamer, Olivier J. H'enaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and João Carreira. 2021. Perceiver IO: A General Architecture for Structured Inputs & Outputs. *ArXiv* abs/2107.14795 (2021).

[30] Jie Jiang, Zhimin Li, Jiangfeng Xiong, Rongwei Quan, Qinglin Lu, and Wei Liu. 2022. Tencent AVS: A Holistic Ads Video Dataset for Multi-Modal Scene Segmentation. *IEEE Access* 10 (2022), 128959–128969.

[31] Yu-Gang Jiang, Baohan Xu, and X. Xue. 2014. Predicting Emotions in User-Generated Videos. In *AAAI Conference on Artificial Intelligence*.

[32] Eunjin Anna Kim, Srinivasan Ratneshwar, and Esther Thorson. 2017. Why Narrative Ads Work: An Integrated Process Explanation. *Journal of Advertising* 46 (2017), 283 – 296.

[33] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[34] Sander Koelstra, Christian Mühl, M. Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and I. Patras. 2012. DEAP: A Database for Emotion Analysis ;Using Physiological Signals. *IEEE Transactions on Affective Computing* 3 (2012), 18–31.

[35] Siew Meng Leong, Swee Hoon Ang, and Lynn Heng. 1994. Using Drama to Persuade: the Effects of Involvement and Ad Form on Persuasion. *ACR Asia-Pacific Advances* (1994).

[36] Boyang Albert Li, Beth Cardier, Tong Wang, and Florian Metze. 2017. Annotating High-Level Structures of Short Stories and Personal Anecdotes. *ArXiv* abs/1710.06917 (2017).

[37] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2022. Foundations and Recent Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions. *ArXiv* abs/2209.03430 (2022).

[38] Nai-Hwa Lien and Yi-Ling Chen. 2013. Narrative ads: The effect of argument strength and story format. *Journal of Business Research* 66 (2013), 516–522.

[39] Rongcheng Lin, Jing Xiao, and Jianping Fan. 2018. NeXtVLAD: An Efficient Neural Network to Aggregate Frame-level Features for Large-scale Video Classification. In *ECCV Workshops*.

[40] Ilya Loshchilov and Frank Hutter. 2017. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

[41] Daniel J. McDuff, Rana El Kaliouby, Jeffrey F. Cohn, and Rosalind W. Picard. 2014. Predicting Ad Liking and Purchase Intent: Large-Scale Analysis of Facial Responses to Ads. *IEEE Transactions on Affective Computing* 6 (2014), 223–235.

[42] David Mick. 1987. Toward a Semiotic of Advertising Story Grammars.

[43] Anca C. Micu and Joseph T. Plummer. 2010. Measurable Emotions: How Television Ads Really Work. *Journal of Advertising Research* 50 (2010), 137 – 153.

[44] Mathew Monfort, Bolei Zhou, Sarah Adel Bargal, Alex Andonian, Tom Yan, Kandan Ramakrishnan, Lisa M. Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, and Aude Oliva. 2018. Moments in Time Dataset: One Million Videos for Event Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42 (2018), 502–508.

[45] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. 2021. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems* 34 (2021), 14200–14213.

[46] José Gabriel Navarro. 2023. Media advertising spending in the United States from 2020 to 2024. https://www.statista.com/statistics/272314/advertising-spending-in-the-us/#:~:text=According%20to%20market%20estimates%2C%20total,grow%20to%20322%20billion%20dollars...

[47] Ads of the World. [n. d.]. Ads of the World. https://www.adsoftheworld.com/

[48] Desmond C. Ong, Zhengxuan Wu, Zhi-Xuan Tan, Marianne C. Reddan, Isabella Kahhalé, Alison Mattek, and Jamil Zaki. 2019. Modeling Emotion in Complex Stories: The Stanford Emotional Narratives Dataset. *IEEE Transactions on Affective Computing* 12 (2019), 579–594.

[49] OpenAI. 2023. GPT-4 Technical Report. *ArXiv* abs/2303.08774 (2023).

[50] Jessica Ouyang and Kathleen McKeown. 2015. Modeling Reportable Events as Turning Points in Narrative. In *Conference on Empirical Methods in Natural Language Processing*.

[51] A. Pardo, Fabian Caba Heilbron, Juan Le'on Alc'azar, Ali K. Thabet, and Bernard Ghanem. 2021. MovieCuts: A New Dataset and Benchmark for Cut Type Recognition. In *European Conference on Computer Vision*.

[52] J Carol Pardun. 2013. *Advertising and Society : an Introduction.* John Wiley & Sons, Inc., New York, NY, USA.

[53] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning

Library. In *Neural Information Processing Systems*.

[54] Christopher P. Puto and William D. Wells. 1984. Informational and Transformational Advertising: the Differential Effects of Time. *ACR North American Advances* (1984).

[55] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*.

[56] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. *ArXiv* abs/2212.04356 (2022).

[57] Abhinav Shukla, Shruti Shriya Gullapuram, Harish Katti, M. Kankanhalli, Stefan Winkler, and Subramanian Ramanathan. 2019. Recognition of Advertisement Emotions With Application to Computational Advertising. *IEEE Transactions on Affective Computing* 13 (2019), 781–792.

[58] Abhinav Shukla, Shruti Shriya Gullapuram, Harish Katti, Karthik Yadati, M. Kankanhalli, and Subramanian Ramanathan. 2017. Affect Recognition in Ads with Application to Computational Advertising. *Proceedings of the 25th ACM international conference on Multimedia* (2017).

[59] Abhinav Shukla, Shruti Shriya Gullapuram, Harish Katti, Karthik Yadati, M. Kankanhalli, and Subramanian Ramanathan. 2017. Evaluating content-centric vs. user-centric ad affect recognition. *Proceedings of the 19th ACM International Conference on Multimodal Interaction* (2017).

[60] Yaman Kumar Singla, Rajat Aayush Jha, Arunim Gupta, Milan Aggarwal, Aditya Garg, Ayushi Bhardwaj, Tushar, Balaji Krishnamurthy, Rajiv Ratn Shah, and Changyou Chen. 2022. Persuasion Strategies in Advertisements: Dataset, Modeling, and Baselines. *ArXiv* abs/2208.09626 (2022).

[61] Mattia Soldan, Alejandro Pardo, Juan León Alcázar, Fabian Caba, Chen Zhao, Silvio Giancola, and Bernard Ghanem. 2022. MAD: A Scalable Dataset for Language Grounding in Videos From Movie Audio Descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5026–5035.

[62] Krishna Somandepalli, Tanaya Guha, Victor R. Martinez, Naveen Kumar, Hartwig Adam, and Shrikanth Narayanan. 2021. Computational Media Intelligence: Human-Centered Machine Analysis of Media. *Proc. IEEE* 109, 5 (2021), 891–910. https://doi.org/10.1109/JPROC.2020.3047978

[63] Krishna Somandepalli, Victor Martinez, Naveen Kumar, and Shrikanth Narayanan. 2018. Multimodal Representation of Advertisements Using Segment-Level Autoencoders. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction* (Boulder, CO, USA) *(ICMI '18)*. Association for Computing Machinery, New York, NY, USA, 418–422. https://doi.org/10.1145/3242969.3243026

[64] Jennifer J. Sun, Ting Liu, Alan S. Cowen, Florian Schroff, Hartwig Adam, and Gautam Prasad. 2020. EEV Dataset: Predicting Expressions Evoked by Diverse Videos. *ArXiv* abs/2001.05488 (2020).

[65] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, et al. 2023. Stanford Alpaca: An Instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca.

[66] Thales S. Teixeira, Rosalind W. Picard, and Rana El Kaliouby. 2014. Why, When, and How Much to Entertain Consumers in Advertisements? A Web-Based Facial Tracking Field Study. *Mark. Sci.* 33 (2014), 809–827.

[67] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 6000–6010.

[68] Nikhita Vedula, Wei Sun, Hyunhwan Lee, Harsh Gupta, Mitsunori Ogihara, Joseph Johnson, Gang Ren, and Srinivasan Parthasarathy. 2017. Multimodal Content Analysis for Effective Advertisements on YouTube. *2017 IEEE International Conference on Data Mining (ICDM)* (2017), 1123–1128.

[69] Ekant Veer and Simon Pervan. 2008. How the tone and wording of advertisements interact. *International Journal of Advertising* 27 (2008), 191 – 207.

[70] Zejia Weng, Lingjiang Meng, Rui Wang, Zuxuan Wu, and Yu-Gang Jiang. 2021. A Multimodal Framework for Video Ads Understanding. *Proceedings of the 29th ACM International Conference on Multimedia* (2021).

[71] Keren Ye, Kyle Buettner, and Adriana Kovashka. 2018. Story Understanding in Video Advertisements. In *British Machine Vision Conference*.

[72] Keren Ye, Narges Honarvar Nazari, James Hahn, Zaeem Hussain, Mingda Zhang, and Adriana Kovashka. 2019. Interpreting the Rhetoric of Visual Advertisements. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43 (2019), 1308–1323.

[73] Huaizheng Zhang, Yong Luo, Qiming Ai, Yonggang Wen, and Han Hu. 2020. Look, Read and Feel: Benchmarking Ads Understanding with Multimodal Multitask Learning. In *Proceedings of the 28th ACM International Conference on Multimedia* (Seattle, WA, USA) *(MM '20)*. Association for Computing Machinery, New York, NY, USA, 430–438. https://doi.org/10.1145/3394171.3413582

[74] Zhipeng Zhang, Xinglin Hou, Kai Niu, Zhongzhen Huang, Tiezheng Ge, Yuning Jiang, Qi Wu, and Peifeng Wang. 2022. Attract me to Buy: Advertisement Copywriting Generation with Multimodal Multi-structured Information. *ArXiv* abs/2205.03534 (2022).

[75] Wayne Xin Zhao, Kun Zhou, Junyi Li, et al. 2023. A Survey of Large Language Models. *ArXiv* abs/2303.18223 (2023).

[76] Athanasia Zlatintsi, Petros Koutras, Georgios Evangelopoulos, Nikos Malandrakis, Niki Efthymiou, Katerina Pastra, Alexandros Potamianos, and Petros Maragos. 2017. COGNIMUSE: a multimodal video database annotated with saliency, events, semantics and emotion with application to summarization. *EURASIP Journal on Image and Video Processing* 2017 (2017), 1–24.