3MASSIV: Multilingual, Multimodal and Multi-Aspect dataset of Social Media Short Videos

Vikram Gupta[1,*], Trisha Mittal[2,*], Puneet Mathur[2], Vaibhav Mishra[1],
Mayank Maheshwari[1], Aniket Bera[2], Debdoot Mukherjee[1], Dinesh Manocha[2]
[1]ShareChat, India
[2]University of Maryland, College Park, USA
{vikramgupta, vaibhavmishra, mayankmaheshwari, debdoot}@sharechat.co
{trisha, puneetm, bera, dmanocha}@umd.edu
Project URL: https://sharechat.com/research/3massiv

Abstract

We present 3MASSIV*, a multilingual, multimodal and multi-aspect, expertly-annotated dataset of diverse short videos extracted from short-video social media platform - Moj.* 3MASSIV *comprises of* 50k *short videos (*20 *seconds average duration) and* 100K *unlabeled videos in* 11 *different languages and captures popular short video trends like pranks, fails, romance, comedy expressed via unique audio-visual formats like self-shot videos, reaction videos, lip-synching, self-sung songs, etc.* 3MASSIV *presents an opportunity for multimodal and multilingual semantic understanding on these unique videos by annotating them for concepts, affective states, media types, and audio language. We present a thorough analysis of* 3MASSIV *and highlight the variety and unique aspects of our dataset compared to other contemporary popular datasets with strong baselines. We also show how the social media content in* 3MASSIV *is dynamic and temporal in nature, which can be used for semantic understanding tasks and cross-lingual analysis.*

1. Introduction

Semantic understanding of videos has been a well-researched problem but still continues to garner a lot of attention from the computer vision and multimedia research communities because videos encode rich information which can be understood across different dimensions using various tasks. Notable progress has been made in terms of analyzing these video for tasks like action classification [34, 40, 64], action localization [16, 85], video description [11, 75], video question answering [41, 66, 81], object and scene understanding [85], etc. The majority of these tasks are focused on recognizing visual aspects present/happening in the video, e.g., action, scene, object detection, and classification.

Figure 1. **3MASSIV:** We highlight three videos uploaded by a particular user. The concept labels for these are *festival*, *couple romance*, *comedy* respectively. We also see the diversity in the video types, *self-shot*, *split-screen* and *special effects*. We also observe how the content is temporally aligned to real-word events; for instance, the festivals. 3MASSIV has 50K such annotated videos across 11 languages with masked user identifiers and timestamps for deeper semantic analysis of social media content. *Faces have been blurred for preserving privacy.*

Detecting these visual aspects helps in answering *what occurs in a video?* But, it does not capture *how viewers interpret the video?* and *which concept(s) the creator of the video wishes to convey?* In this work, we investigate the semantic understanding of videos uploaded on short-video social media platform - *Moj* [1] from the perspective of *creators* and *viewers* of these videos, which has not been explored before, primarily due to the lack of large-scale annotated video datasets. Considering the rapid adoption of social media, a holistic understanding of the creation, con-

*The first two authors contributed equally to this work.

[1]https://mojapp.in

sumption, and popularity dynamics of these videos forms an important and timely research direction.

To facilitate this under-explored research direction, we present a novel dataset, 3MASSIV, built from short videos posted on the short-video platform - *Moj*. Even though existing datasets for semantic understanding source videos from social media (*e.g.* YouTube [1], Vine [48], Facebook [52]), they are not suitable for our task. We highlight the key challenges and elaborate on how 3MASSIV addresses them:

- **Taxonomy:** Prior datasets [1, 19, 34] adopt a top-down approach of constructing a vocabulary of visual concepts from domain-independent taxonomies (*e.g.* freebase) and mining videos from social media using this vocabulary. However, this vocabulary is not exhaustive and fine-grained for capturing popular concepts in social media discourse. Moreover, this method generates "easy videos" as search engines prioritize them first [52]. [52] adopt uniform sampling to address this problem while we construct a comprehensive bottom-up taxonomy using popularity-based sampling of videos for bridging this gap.
- **Novel video types:** Existing datasets do not capture novel and challenging video formats like split-screen videos, special effects (masks/graphics overlaid on faces), portrait videos, lip-syncing to pre-recorded audio, etc. (Figure 1) which are dominant on social media platforms. 3MASSIV curates the videos from a short video platform - *Moj* and annotates them for these media types for filling this gap.
- **Video Narrative:** Broadly speaking, there are three distinct kinds of videos on social media: a) ***Micro Narrative***: Videos which are short in duration [48] (5-6 secs) or are clipped out from longer videos [15,46,85], b) ***Long Narrative***: Longer videos [1, 10, 78], usually more than 1-2 minutes, which tell a more detailed narrative or story c) ***Short Narrative***: These are longer than micro-videos (typically 10-20 secs) and provide authors and content creators more flexibility in terms of time limits. Despite the explosive growth of short video platforms like Tiktok, Reels, Youtube Shorts, and Moj, short videos have not been explored in detail in the Computer Vision and AI communities, primarily because of the lack of a large-scale labeled dataset. 3MASSIV contains complete videos created with a short and concise narrative presenting an opportunity to understand this new avenue of video understanding.
- **Sparse/Noisy Hashtags:** Since expert annotation is expensive, large datasets often use hashtags added by the creators [48]. However, hashtags are usually sparse - 56% of videos did not have hashtags in MV-58 [48]. Also, they can be noisy, as shown in (A.6). Our dataset, 3MASSIV, addresses this by manually annotating the videos using expert annotators.
- **Linguistic Diversity:** Existing datasets for semantic understanding of videos are not motivated towards exploring linguistic diversity while 3MASSIV comprises of videos from 11 languages, annotated with the language of the audio for facilitating multilingual semantic understanding of videos.

3MASSIV contains concept, affective states, audio type, video type and language annotations for understanding the *creator's* and *viewer's* perspectives. We label the videos with the following annotations for modeling the **viewer's perspective:**

- **Concept:** Each video is annotated for a concept (across 34 labels) by expert annotators. Our dataset contains widely popular and unique social media concepts like *pranks, fails, romance, philanthropy, comedy, etc.* Figure 2 shows some examples which demonstrate that understanding these videos, which are very human-centric, self-shot with a short story goes beyond detecting and classifying the audio-visual aspects and makes 3MASSIV challenging.
- **Affective States:** We provide annotations for 11 emotion categories present in these videos.

Similarly, to understand the **creator's perspective**, we provide annotations for media types that content creators use to convey their point. Figure 1 shows some of the examples.

- **Audio Types:** The audio types are unique and diverse with recorded/self-sung songs, dialogues, monologues, instrumentals, etc.
- **Video Types:** Video formatting comprises of slideshows, animations, split-screens, self-shot, movie/TV-serial clips, etc. which are very popular on short video platforms.

Additionally, our dataset 3MASSIV can be used for various tasks and applications, such as:

- **Multilingual Modeling:** We provide annotations for the 11 different languages, opening opportunities for multilingual semantic understanding.
- **Creator Modeling:** We also provide masked creator identifiers and recent videos uploaded by these creators (100k videos), opening up exciting user modeling ideas inspired by semantic video understanding.
- **Temporal Analysis:** Social media content has a very short life span and is very dynamic. To enhance understanding here, we provide timestamps of these videos, which can help model temporal dynamics of the nature of popular content on such platforms. Moreover, we provide masked user profiles to identify videos from the same creators to analyze the shift in their perspectives over time.

To the best of our knowledge, 3MASSIV is the first human-annotated large-scale dataset of short videos that can

be used for modeling concepts, affective states, and media types across 11 languages, presenting a unique opportunity for understanding social media content. Overall, 3MASSIV contains 900 hours of video data uploaded by 23121 creators with 50K expertly annotated videos and 100K unlabeled videos with an average duration of around 20 seconds. We also present baseline results to empirically establish that 3MASSIV is challenging and unique in Section 4. In Section 5, we discuss the application of 3MASSIV over various research problems.

2. Related Work

We review related datasets for semantic understanding of videos from social media and summarise them in Table 1.

2.1. Semantic Understanding Datasets

Various datasets and tasks have been proposed for video understanding.

Action classification is a popular research problem for which benchmark datasets like [8, 10, 16, 20, 32, 34, 36, 40, 43, 46, 64, 85] have been proposed.

Concept Understanding: Going beyond action classification, detection, and segmentation of visual elements, *theme/concept* classification datasets focus on modeling interplay between the visual and audio elements for understanding the overall theme/concept represented by the videos. For instance, YouTube-8M [1] focuses on classifying videos into categories like *fashion, games, shopping, animals, etc.*. The taxonomy has been curated manually to capture purely *visual* categories, and the dataset has been machine annotated using the YouTube Video Annotation system for collecting videos. Similarly, Holistic Video Understanding (HVU) [15] annotate videos from [1,34,85] for concepts along with scenes, objects, actions, attributes, and events using Google Vision API and Sensifai Video Tagging API[2]. **MicroVideos** [48] contributes videos collected from a micro-video application - Vine and interpret user-generated hashtags as annotations. More recently, datasets for understanding **Intent** and **Motivation** from social media posts are being investigated [29, 39, 60, 69, 70, 76, 82].

Other Video Understanding Tasks: [12, 45, 51, 53, 56, 72] have been proposed for object detection, segmentation and tracking from videos. At the intersection of vision and language, datasets for video description [71, 75], question-answering [41, 66, 81], video-object grounding [9, 84] and text-to-video retrieval [4,42] have been proposed. SVD [30] contribute a dataset for near-duplicate video retrieval.

2.2. Affective Analysis of Social Media Content

Understanding perceived emotions of individuals using verbal and non-verbal cues is an important problem in both AI and psychology for various applications. One such application is for understanding the projected [80] and evoked emotions [33, 44] from multimedia content like advertisements and movies. There is vast literature in inference of perceived emotions from a single modality or a combination of multiple modalities like facial expressions [2, 58], speech/audio signals [59], body pose [47], walking styles [7] and physiological features [35]. There has been a shift in the paradigm, where researchers have tried to fuse multiple modalities to perform emotion recognition, also known as Multimodal Emotion Recognition. Fusion methods like early fusion [62], late fusion [21], and hybrid fusion [63] have been explored for emotion recognition from multiple modalities.

2.3. Research Problems with Social Media Content

Multilingual Analysis of Videos: Multilingual analysis of images and videos has been studied previously. Harwath et al. [25] proposed a bilingual dataset comprising English and Hindi captions. Ohishi et al. [49] extended this dataset to include Japanese captions and proposed a trilingual dataset. Approaches for bilingual video understanding include [6, 31, 50]. On the other hand, several datasets for multilingual video understanding [57, 71] along with techniques for analyzing them [55] have been proposed, although they lack diversity in audio language.

User Modeling of Social Media Content: People are increasingly relying on social media platforms for sharing their daily lives, which reflect their personality traits and behavior. User modelling based on their online persona and activity has been successfully leveraged for digital marketing [3, 77] and content recommendation [73, 79]. Not only on the consumer side, but user profiling is also helpful for helping content creators on such social media platforms [5, 27]. To further research in these directions, we provide masked user identifications.

Temporal Analysis of Social Media Content: A unique characteristic of social media content is the short life span of posts [17]. Such dynamically and temporally evolving content is evident and can be mapped to major festivals, celebrations, political events, news, and trends [24]. Such dynamic and temporally evolving content can be helpful to understand social media platforms better.

3. Our Dataset: 3MASSIV

In this section, we introduce 3MASSIV and elaborate on the dataset collection and annotation process.

3.1. Taxonomy

We annotate our datatset for the following taxonomies. A detailed description of all the annotation labels of the taxonomy is presented in Appendix A.1.

Concept: Creation of a taxonomy for concepts is a non-trivial exercise, requiring both comprehensiveness as well as frequency coverage. We adopted a bottom-up approach

[2]https://cloud.google.com/vision, https://sensifai.com/

	Datasets	Size	Duration	Source	Labels	Audio Types	Video Types	Affective	Focus	Lang	Year
Image	Intentonomy [29]	14k	-	Flickr	HA	-	-	NA	Understanding Intent of Social Media Posts	-	'20
Video	Sports-1M [32]	1M	4 min	YouTube	MG	NA	NA	NA	Sports Activity Classification (487 Classes)	NA	'14
	ActivityNet [78]	27801	5-10 mins	Web	HA	NA	NA	NA	HAR (203 classes)	NA	'15
	MV-58K [48]	260k	6 secs	Vine	MG	NA	NA	NA	Activity, Objects, Platform Specific Classes	NA	'16
	Charades [61]	10k	30 secs	CrowdSourced	HA	NA	NA	NA	HAR + Object Classification (157 classes)	NA	'16
	YouTube-8M [1]	8M	2-10 mins	YouTube	MG	NA	NA	NA	Video Topic Classification	NA	'16
	Kinetics [34]	300k	10 secs	YouTube	HA	NA	NA	NA	HAR (400/600/700 classes)	NA	'17
	Something-Something [19]	100k	2-6 secs	CrowdSourced	HA	NA	NA	NA	HAR (174 classes)	NA	'17
	Epic-Kitchens [10]	39594	1-55 mins	CrowdSourced	HA	NA	NA	NA	Actions in Kitchen	Yes	'18
	SOA [52]	562k	10 secs	Facebook	HA	NA	NA	NA	Scenes, Objects, Actions	NA	'18
	MomentsInTime [46]	1M	3 secs	10 sources	HA	NA	NA	NA	339 action classes	NA	'19
	HACS [85]	1.5M	2 secs	YouTube	HA	NA	NA	NA	HAR (200 classes)	NA	'19
	HVU [15]	500k	$\leq$ 10 secs	YT8M, Kinetics, HACS	MG	NA	NA	NA	Actions, Objects, Concepts, Events, Attributes, Scenes	NA	'20
Affects (videos)	CMU-MOSEI [80]	23K	7.3 secs	YouTube	HA	NA	NA	Yes	Perceived EC (6 classes)	NA	'18
	EEV [65]	23k	1-25 mins	Online Videos	MG	NA	NA	Yes	Evoked EC (15 classes)	NA	'20
	Ours	50k(+100k)	-	Social Media Platform	HA	Annotated for 7 classes	Annotated for 8 classes	Yes	Concept, Affective States, Media Type, Language	Annotated for 11 languages	'21

† NA, MG, HA indicate "not annotated", "machine generated", and "human annotated", respectively.

Table 1. Comparison of 3MASSIV with related image and video datasets. Our dataset has exhaustive and expertly annotated annotations for concepts, audio/video types, affective states and audio language for social media short videos. Majority of the other datasets focus on specific tasks like action classification and do not annotate for other dimensions. YT8M, SOA and HVU adopt more holistic annotations. We report the range or average duration of videos for the datasets.

to model social media behavior rather than mining videos for an existing taxonomy. To achieve this, we employed a team of digital social media experts for scanning 1.5 million popular posts and assigned a label that concisely describes a post. The taxonomy grew to more than 1000 concepts and was pruned to 34 popular labels covering more than 75% of the videos for this study. Some of these concepts like *fails, pranks, comedy, romance, philanthropy* are unique to our dataset and are illustrated in Figure 2. We illustrate the distribution across these concepts in Figure 3a.

Affective States: We provide annotations for the projected affective labels for the videos. Inspired by [13], we adopt a 11 label taxonomy for affective states. We present the distribution across these affective states in Figure 3b.

Audio Type: Social media creators use a variety of audio styles like lip-syncing to pre-recorded songs, monologues, dialogues, self-sung songs, or instrumental music. We present a taxonomy of 7 labels to cover the broad spectrum of audio content type (Figure 3c).

Video Type: We provide annotations for classifying video types based on how the video was created/edited (Figure 3d). The videos can be conventionally sourced from Movie or TV-Show clips or be self-shot on personal handheld devices. The videos also contain slideshows, still images, and split screens. Additionally, many creators also publish videos with text to add a linguistic message to enhance the audio-visual effect.

Language: We annotate audio language for our videos and highlight the linguistic diversity of our dataset in Figure 3e.

3.2. Data Collection

We collect our dataset from a leading short video application supporting over 15 languages. The platform contains short videos uploaded by professional and amateur content creators on which users can view, like, share and comment. We extracted more than 1.5M videos uploaded over 9 months (Feb, 2021 to Oct, 2021) across 11 languages and share $50k$ labeled and $100k$ unlabeled from this set. These videos were shortlisted based on platform engagement metrics after removing near-duplicates. The duration of videos ranges between $4.5 - 116$ seconds (averaging 20 seconds). Videos reported to be of sensitive nature and those containing nudity, violence, and abuse were removed. Additional steps about data collection are mentioned in A.2.

3.3. Data Annotation

We employed domain experts in the field of social media who provided labels for the 50K videos. Annotators were selected to ensure that we can label every video, across 11 different languages, by experts who are fluent writers and speakers of the dominant language of the video. The annotators were provided with guidelines, which comprised of instructions about each task, definitions of class labels (Appendix A.4, Table 7) and a few worked-out examples to familiarize them with the annotation task.

Annotator Onboarding: We followed a strict annotator onboarding mechanism. We provided new candidates with a set of 100 posts that have been pre-annotated by expert reviewers and benchmarked against other candidates. Candidates not adhering to the benchmarks were not allocated further posts, and their responses were discarded.

Inter-Annotator Agreement: We evaluated inter-annotator agreements across all labels in different concepts using Krippendorff's alpha (K-alpha) [38] to account for labeling reliability amongst multiple annotators. All annotations were performed by 3 annotators each, and their majority vote was accepted as the ground truth label. In case of a three-way disagreement, an expert annotator resolved the conflict and assigned the final label. The K-alpha values for the 4 taxonomies, concept, audio type, video type, and affective states are 0.77, 0.59, 0.62, and 0.40, respectively. We present detailed per-label annotator agreement in Table 6. We observe strong agreements for most of the tasks. 3MASSIV is finally split into train,

(a) **Prank Scene**: A man is trying to prank the lady by putting an adhesive on her footwear with the intent of creating a funny situation for the viewers. Deep semantic understanding is required to understand the spatio-temporal-audio context of the scene to classify as "prank" because detection of visual or audio aspects is not sufficient.

(b) **Fail Scene**: Kid is trying to perform a summersault using a small trampoline but fails to complete the flip. For correct classification, model needs to focus on the unplanned fall at the end of the video to classify it as a "fail" video.

(c) **Philanthropy Scene**: A man meets and greets needy strangers and surprises them with a gift. In order to recognize this as a gesture of kindness, our model needs to understand the economical situation and emotional state of the subjects in the videos and focus on the exchange of tokens.

(d) **Comedy Scene**: A funny and sarcastic verbal exchange between two friends. Both display a range of emotions during the act but the overall outcome of the video is a comedic situation. Focussing on facial emotions or human pose might not be sufficient for understanding the scene.

Figure 2. **Unique Concepts present in 3MASSIV:** Our theme taxonomy comprises of several unique topics popular in social media domain but unexplored in literature: (a) Prank videos showing planned mischievous acts aimed to elicit reactions from co-creators [28]; (b) Fail videos that record unsuccesful attempts resulting in harm-joy [54]; (c) Philanthropy videos portraying acts of helpful service, moral assistance or charitable deeds; (d) Scripted and natural comedy videos which can be further categorized based on the inter-agent relationships between the actors - couple, family, kids, friends, etc. *Faces have been blurred for preserving privacy.*

validation, and test sets in a ratio of $60:20:20$.

3.4. Dataset Analysis

3MASSIV contains 55262 annotated videos and 100K unlabeled videos with a total of 910 hours of video data. Figure 3a – 3e show the exhaustive taxonomy and distribution of 3MASSIV.

Concept: As evident from Figure 3a, *comedy* and *romance* have a higher frequency than other labels, while *pets* has the least frequency. This is expected given the trends in short video social media platforms that incentivize creators to create content with wide appeal.

Affective States: Figure 3b shows the 11 affective states found in the corpus. We observe class imbalance that mirrors the distribution of natural human emotions.

Audio Type: Figure 3c highlights an interesting phenomenon wherein more than 50% of the videos borrow the background music from a pre-recorded source while self-spoken dialogues and monologues are comparatively less. This alludes to the fact that a large majority of creators are more comfortable in visual mode of expression. Similarly, lip-syncing to existing audio is the second-most popular way of video creation.

Video Type: As evident in Figure 3d, more than two-third of videos sampled in the dataset are self-shot. Advances in photography have aided creators in adding visual as well

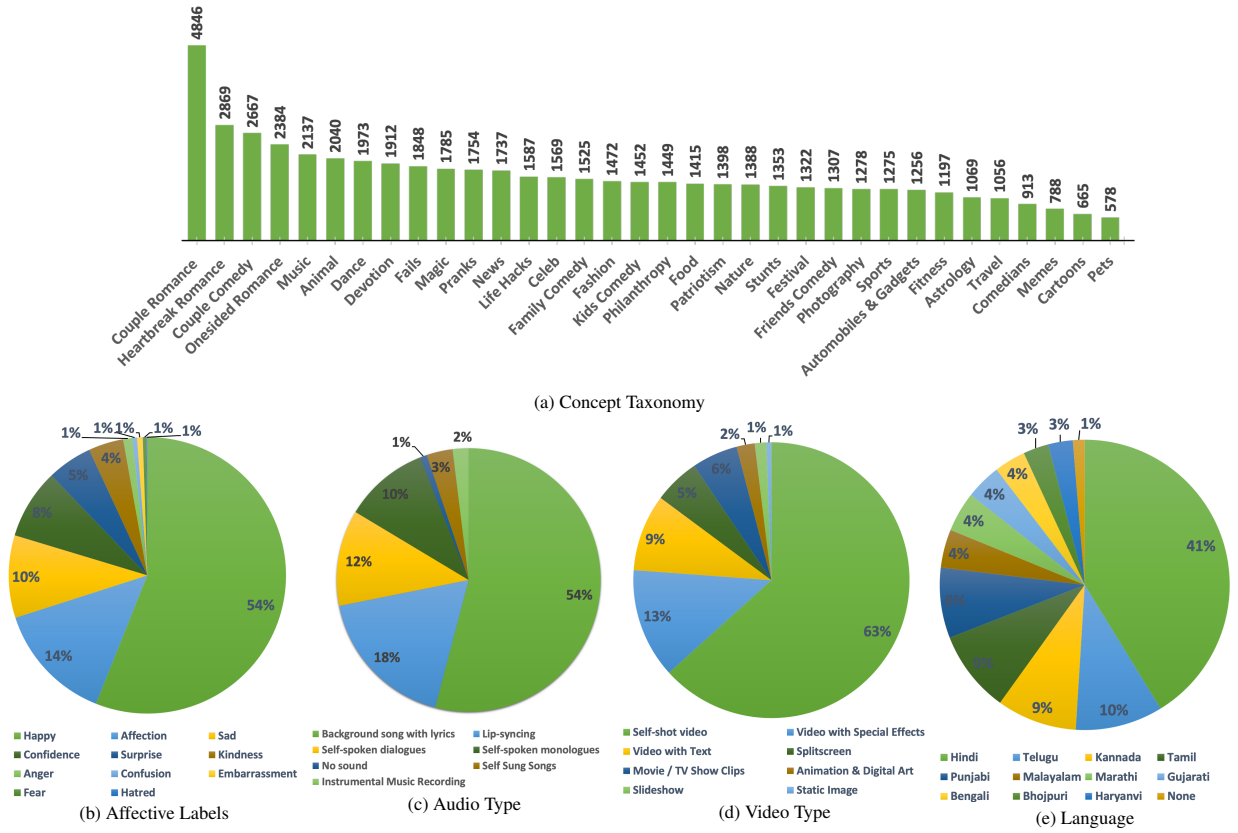


Figure 3. **3MASSIV Taxonomy:** Sub-figures 3a – 3e show the taxonomy and label distributions in the proposed 3MASSIV dataset for concept, affective states, audio type, video type and language in anti-clockwise direction.

Data Description	Value
# Concept	34
# Languages	11
# Affective States	11
# Audio Types	7
# Video Types	8
# Creators	23121
# Annotators	95
# Labelled Videos	55262
# Unlabelled Videos	100K
Total Duration Labelled	310 hours
Total Duration Unlabelled	600 hours
Average Duration	20.2 ($\pm$9.5) seconds
Min/Max Duration	4.5/116 seconds

Table 2. **3MASSIV Statistics**

as textual effects to the videos, making them the next most popular video formats.

Languages: The dataset comprises videos in 11 languages with Hindi as the majority language.

Duration: 3MASSIV comprises of videos ranging from 4.5s-116s with an average duration of 20 seconds.

Creators: 3MASSIV comprises of videos from 23121 unique creators. A large majority of these creators (15998) contribute only one video in our dataset, while 7133 contributed more than one video. This demonstrates the immense diversity of our dataset in terms of creators.

Taxonomy Correlation: In Appendix A.5, Figure 5, we present the correlation between concepts and affective states/media types. We observe that *heartbreak romance* videos predominantly have *sad* affective state; *philanthropy* is strongly linked with *kindness*. Similarly, we observe that videos with *magic* label are linked with *surprise* affective state; *couple romance* shows the strongest predisposition towards *affection*. These correlations provide insights that 3MASSIV comprises of videos that depict strong correlation with other underlying aspects and this correlation can be leveraged for better semantic understanding.

4. Baseline Experiments

We perform baseline experiments to highlight the unique and challenging aspects of 3MASSIV.

4.1. Concept Classification

We report the results for concept classification using different modalities individually and in combination using late

fusion in Table 3a. We report top-1, top-3, and top-5 accuracy for all the experiments.

Audio-Visual Representation: We experiment with 3D ResNet [23] backbones trained over Kinetics700 [8] for spatio-temporal modelling. We also evaluate deeper (R3D-101) and depth-wise separable architecture (R(2+1)D-50) [67] but did not observe gains. Hence we use R3D-50 for all our experiments. For audio modelling, we leverage pretrained VGG [26] model and CLSRIL23 [22]. VGG is trained for sound classification ([18]) and CLSRIL23 is trained over speech data of 23 Indic languages. We freeze the audio-visual backbones and train the classifier and multimodal fusion layers.

Results and Discussion: From Table 3a, we observe that the performance of visual modality is higher than audio, which highlights the importance of visual modality for our dataset 3MASSIV. On combining the modalities using late-fusion, we observe a gain of 4% (Row 6 and 7). This demonstrates the multimodal nature of the dataset. By combining both VGG and CLSRIL23 features with visual modality, we notice further gains showing complementary information in both these audio representations (Row 8). This is not surprising because our dataset contains a wide variety of audio types like *songs, monologues, and dialogues.* While VGG has been trained for modeling sounds (music, vehicle, creek, instrument, etc.), CLSRIL23 is more specialized for understanding human speech. We expand on the training details and hyperparameters in Appendix B.1.1.

Error Analysis: We analyze error cases for different media types in Figure 4b and Figure 4a. We notice comparatively less performance on *images, reaction videos, and slide-shows*, which showcases the novelty of these types in video datasets. *Reaction videos* contain split-screens and are complex as the model needs to focus on the salient parts. Similarly, slide shows contain a lot of abrupt scene changes making it extremely challenging. On audio-types, we notice the model shows less accuracy for classes like *lip-sync, instrumental, and silence/noise*. This is not unexpected as these do not provide relevant signals about the concept. Similarly, *lip-sync* encodes the majority of the semantic information in the audio channel. These observations strongly highlight the unique challenges of our dataset 3MASSIV, which have not been explored before. In Figure 7a (in Appendix B.1.2), we plot the confusion matrix of the audio-visual model. We notice confusion among the concept labels like *memes, kids, family, friends*, and *couple comedy*, demonstrating the challenges in semantic understanding of such content. We also study the impact on accuracy of concept categories using the audio-visual modalities in Figure 7b (Appendix B.1.2).

4.2. Affective State Classification

We select two state-of-the-art affective state classification models and benchmark them on 3MASSIV. The results are summarized in Table 3b. We report top-1, top-3 accuracy scores. Also, because there is an imbalance in the number of data points per affective label, we also report F1 score. The first method, Kosti et al. [37] is an emotion recognition model which uses the facial expressions of the dominant subject in the video and the background context. Tsai et al. [68] is a multimodal transformer-based model that uses both visual and audio modalities and has shown high performance on other emotion recognition datasets. We observe that the performance of these models on 3MASSIV is not very high. On further analysis of these models, we notice that videos associated with human-centric concept labels *pranks, fails* often get misclassified. Similarly, videos with *static images* and *animations* often get misclassified.

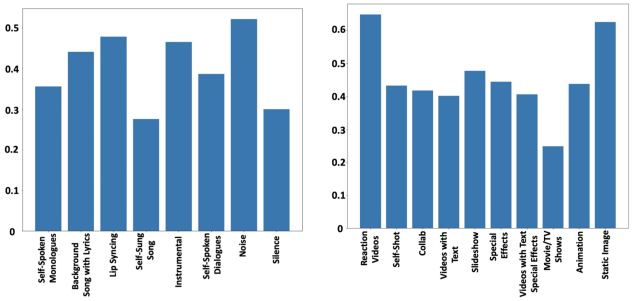
Modality	Backbone	Top-1	Top-3	Top-5
Visual	R(2+1)D-50	50.6	72.3	81.4
Visual	R3D-50	52.7	74.5	83.6
Visual	R3D-101	52.6	74.1	83.3
Audio	VGG	31.6	50.5	60.9
Audio	CLSRIL23	31.2	50.1	60.6
Visual, Audio	R3D-50 + VGG	54.9	74.9	82.4
Visual, Audio	R3D-50 + CLSRIL23	54.9	75.4	82.9
Visual, Audio	R3D-50 + VGG + CLSRIL23	**56.5**	**76.5**	**83.8**

(a) Concept Classification

Method	Modality	Top-1	Top-3	F1
Kosti et al. [37]	visual	35.08	81.92	0.19
Tsai et al. [68]	audio	27.10	66.67	0.21
	audio-visual	38.05	83.90	0.29

(b) Affective State Classification

Table 3. **Baseline Experiments:** Baseline experiments for concept and affective state classification on 3MASSIV using different modalities and combinations.



(a) Error Rate in Concept Classification per Audio Type label

(b) Error Rate in Concept Classification per Video Type label

Figure 4. **Challenging Audio and Video Types:** We present an in-depth analysis into the misclassifications for concept classification. We try to understand the relation between the audio type and the video type of the incorrectly classified videos.

5. Social Media Content Analysis

Creator User Profile Modeling: We leverage affinity of creators towards concepts for improving semantic under-

Method	#Posts	Top-1	Top-3	Top-5
Audio-Visual	-	56.5	76.5	83.8
ProbDist	1	56.9	77.2	84.1
ProbDist	5	58.5	77.7	84.7
ProbDist	20	**59.3**	**78.9**	**85.7**
ProbMax	20	58.8	78.1	85.2

Table 4. **Creator Profiling:** Concept classification with semantically inferred creator profile with audio-visual representations.

Target	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5
Hindi	40.1	61.5	70.5	61.2	79.1	85.5
Telugu	48.1	72.1	81.6	54.9	78.2	86.1
Tamil	45.8	66.6	78.1	51.0	73.8	82.3
Kannada	48.5	72.7	79.6	56.8	78.4	84.4
Punjabi	39.9	62.2	72.9	45.7	69.5	79.4

Table 5. **Cross-lingual Experiments:** We train the audio-visual concept classification model on all the languages apart from the target language and evaluate on target language (green column) columns; all languages are used for training (blue column).

standing in Table 4. For every creator, we mine recent videos uploaded by them and use our audio-visual semantic model for predicting the concept probabilities for these posts. We average the predicted probability distributions and use them for representing the creator (ProbDist). Creator representation is then combined with audio-visual features via late fusion for training the model. We observe gains of 5% over the audio-visual baseline by incorporating creator profile as prior for semantic understanding. We vary the number of recent posts and observe gains by increasing the number of posts (Row $2, 3, 4$), showing that longer creation history is helpful in modeling the creators. We also experiment with maximum prediction (ProbMax) for each post instead of probability distribution (Row 5). This simple yet effective baseline motivates further investigation for modeling creator user profiles using only semantics.

Cross-Lingual Analysis: We also explore 3MASSIV for cross-lingual analysis over 5 popular languages in Table 5. For each target language, we remove it from the training set and train an audio-visual model using other languages. We evaluate this model on the target language to obtain zero-shot results. We present the top-1, top-3, and top-5 accuracy for concept classification with this experiment in *green* columns. In *blue* columns, we use all 5 languages for training and testing. We can see that the performance gap between *green* and *blue* columns is significant, indicating that 3MASSIV can be useful for advancing the state-of-the-art in cross-lingual video understanding tasks.

Temporal Analysis: We explore another interesting aspect of 3MASSIV- temporally evolving content. We notice a strong link to real-world events (Figure 8). We extract top-performing 50K posts based on views from 10 weeks (29th August - 7th November 2021) and analyze the predictions for these posts using our models. We observe an increasing content related to *sports* concept because of an upcoming major sports league. Similarly, we see some peaks in *celebrations* concept because of the recent festive season.

6. Ethics, Data and User Privacy

Respecting User Privacy: The videos collected for the dataset are all publicly available on *Moj*. Informed consent of the users has been taken by the platform for public usage of these videos. The user identifiers and exact publication date have been masked to protect privacy.

Respecting Intellectual Property: Creators have the complete freedom to take down their content. Our dataset provides direct URL links to access the videos, while the platform holds the rights to these videos. This would allow the users to delete the videos on the platform, thus deactivating the links. Our data collection and dissemination efforts abide by platform guidelines.

Opt-out form: Users may choose to have their video removed from the dataset upon request through an opt-out form is available on the dataset homepage.

Handling Misuse: Adequate caution was taken to not store any user information, videos (raw or processed), or metadata on permanent storage outside the computing infrastructure of the social media platform. We aim to disseminate the data upon request and log all access to the dataset, which will only be available for research purposes.

License: We release 3MASSIV for research purposes only (i.e. no commercial usage).

Annotator Compensation: We ensured that all annotators were fairly compensated on an hourly basis and they were apprised of potential social media fatigue [83] resulting from long exposure to social media content.

7. Conclusion

We presented 3MASSIV, a multilingual, multimodal and multi-aspect, human-annotated dataset of social media short videos extracted from a social media platform. 3MASSIV comprises of 50K labeled short videos and 100K unlabeled short videos from a popular social media platform in 11 different languages. 3MASSIV is useful to further semantic understanding of social media content which embodies unique characteristics and nuances. We presented an in-depth analysis and showed the challenges and uniqueness of the dataset using baseline comparisons. We also present some applications of 3MASSIV for various user-modeling tasks and cross-lingual tasks.

8. Acknowledgements

Mittal, Mathur, Bera and Manocha were supported, in part by ARO Grants W911NF1910069 and W911NF2110026.

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

[2] Kingsley Oryina Akputu, Kah Phooi Seng, and Yun Li Lee. Facial emotion recognition for intelligent tutoring environment. In *IMLCS*, pages 9–13, 2013.

[3] Ali Abdallah Alalwan, Nripendra P Rana, Yogesh K Dwivedi, and Raed Algharabat. Social media in marketing: A review and analysis of the existing literature. *Telematics and Informatics*, 34(7):1177–1190, 2017.

[4] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017.

[5] Arturo Arriagada and Francisco Ibáñez. "you need at least one picture daily, if not, you're dead": Content creators and platform evolution in the social media ecology. *Social Media+ Society*, 6(3):2056305120944624, 2020.

[6] Emmanuel Azuh, David Harwath, and James R Glass. Towards bilingual lexicon discovery from visually grounded speech audio. In *INTERSPEECH*, pages 276–280, 2019.

[7] Uttaran Bhattacharya, Trisha Mittal, Rohan Chandra, Tanmay Randhavane, Aniket Bera, and Dinesh Manocha. Step: Spatial temporal graph convolutional networks for emotion perception from gaits. *arXiv preprint arXiv:1910.12906*, 2019.

[8] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019.

[9] Zhenfang Chen, Lin Ma, Wenhan Luo, and Kwan-Yee K Wong. Weakly-supervised spatio-temporally grounding natural sentence in video. *arXiv preprint arXiv:1906.02549*, 2019.

[10] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(11):4125–4141, 2021.

[11] Pradipto Das, Chenliang Xu, Richard F Doell, and Jason J Corso. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2634–2641, 2013.

[12] Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. Tao: A large-scale benchmark for tracking any object. In *European conference on computer vision*, pages 436–454. Springer, 2020.

[13] Munmun De Choudhury, Michael Gamon, and Scott Counts. Happy, nervous or surprised? classification of human affective states in social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 6, 2012.

[14] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*, 2020.

[15] Ali Diba, Mohsen Fayyaz, Vivek Sharma, Manohar Paluri, Jurgen Gall, Rainer Stiefelhagen, and Luc Van Gool. Holistic large scale video understanding. *arXiv preprint arXiv:1904.11451*, 38:39, 2019.

[16] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia 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.

[17] Martin S Fiebert, Azadeh Aliee, and Hoda Yassami. The life span of a facebook post: age & gender effects. *International Review of Social Sciences and Humanities*, 7(2):140–143, 2014.

[18] 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.

[19] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017.

[20] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[21] Hatice Gunes and Massimo Piccardi. Bi-modal emotion recognition from expressive face and body gestures. *Journal of Network and Computer Applications*, 30(4):1334–1345, 2007.

[22] Anirudh Gupta, Harveen Singh Chadha, Priyanshi Shah, Neeraj Chimmwal, Ankur Dhuriya, Rishabh Gaur, and Vivek Raghavan. Clsril-23: Cross lingual speech representations for indic languages. *arXiv preprint arXiv:2107.07402*, 2021.

[23] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018.

[24] Giannis Haralabopoulos, Ioannis Anagnostopoulos, and Sherali Zeadally. Lifespan and propagation of information in on-line social networks: A case study based on reddit. *Journal of network and computer applications*, 56:88–100, 2015.

[25] David Harwath, Galen Chuang, and James Glass. Vision as an interlingua: Learning multilingual semantic embeddings of untranscribed speech. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4969–4973. IEEE, 2018.

[26] 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 *International conference on acoustics, speech and signal processing*. IEEE, 2017.

[27] Lauri Huotari, Pauliina Ulkuniemi, Saila Saraniemi, and Minna Mäläskä. Analysis of content creation in social media by b2b companies. *Journal of Business & Industrial Marketing*, 2015.

[28] Yosra Jarrar, Ayodeji Olalekan Awobamise, Sheila Ogochukwu Nnabuife, and Gabriel E. Nweke. Perception of pranks on social media: Clout-lighting. *Online Journal of Communication and Media Technologies*, 2019.

[29] Menglin Jia, Zuxuan Wu, Austin Reiter, Claire Cardie, Serge Belongie, and Ser-Nam Lim. Intentonomy: a dataset and study towards human intent understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12986–12996, 2021.

[30] Qing-Yuan Jiang, Yi He, Gen Li, Jian Lin, Lei Li, and Wu-Jun Li. Svd: A large-scale short video dataset for near-duplicate video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5281–5289, 2019.

[31] Herman Kamper and Michael Roth. Visually grounded cross-lingual keyword spotting in speech. *arXiv preprint arXiv:1806.05030*, 2018.

[32] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[33] Karim Sadik Kassam. *Assessment of emotional experience through facial expression*. Harvard University, 2010.

[34] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[35] R Benjamin Knapp, Jonghwa Kim, and Elisabeth André. Physiological signals and their use in augmenting emotion recognition for human–machine interaction. In *Emotion-oriented systems*, pages 133–159. Springer, 2011.

[36] Quan Kong, Ziming Wu, Ziwei Deng, Martin Klinkigt, Bin Tong, and Tomokazu Murakami. Mmact: A large-scale dataset for cross modal human action understanding. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[37] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Emotic: Emotions in context dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 61–69, 2017.

[38] Klaus Krippendorff. Computing krippendorff's alpha-reliability. 2011.

[39] Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. Integrating text and image: Determining multimodal document intent in instagram posts. *arXiv preprint arXiv:1904.09073*, 2019.

[40] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011.

[41] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018.

[42] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 447–463. Springer, 2020.

[43] Ang Li, Meghana Thotakuri, David A Ross, João Carreira, Alexander Vostrikov, and Andrew Zisserman. The ava-kinetics localized human actions video dataset. *arXiv preprint arXiv:2005.00214*, 2020.

[44] Anca Cristina Micu and Joseph T Plummer. Measurable emotions: How television ads really work: Patterns of reactions to commercials can demonstrate advertising effectiveness. *Journal of Advertising Research*, 50(2):137–153, 2010.

[45] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016.

[46] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(2):502–508, 2019.

[47] Costanza Navarretta. Individuality in communicative bodily behaviours. In *Cognitive Behavioural Systems*, pages 417–423. Springer, 2012.

[48] Phuc Xuan Nguyen, Gregory Rogez, Charless Fowlkes, and Deva Ramanan. The open world of micro-videos. *arXiv preprint arXiv:1603.09439*, 2016.

[49] Yasunori Ohishi, Akisato Kimura, Takahito Kawanishi, Kunio Kashino, David Harwath, and James Glass. Trilingual semantic embeddings of visually grounded speech with self-attention mechanisms. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4352–4356. IEEE, 2020.

[50] Yasunori Ohishi, Akisato Kimura, Takahito Kawanishi, Kunio Kashino, David Harwath, and James R Glass. Pair expansion for learning multilingual semantic embeddings using disjoint visually-grounded speech audio datasets. In *INTERSPEECH*, pages 1486–1490, 2020.

[51] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 724–732, 2016.

[52] Jamie Ray, Heng Wang, Du Tran, Yufei Wang, Matt Feiszli, Lorenzo Torresani, and Manohar Paluri. Scenes-objects-actions: A multi-task, multi-label video dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 635–651, 2018.

[53] Esteban Real, Jonathon Shlens, Stefano Mazzocchi, Xin Pan, and Vincent Vanhoucke. Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5296–5305, 2017.

[54] Ira J. Roseman and Amanda K. Steele. Concluding commentary: Schadenfreude, gluckschmerz, jealousy, and hate—what (and when, and why) are the emotions? *Emotion Review*, 10:327 – 340, 2018.

[55] Andrew Rouditchenko, Angie Boggust, David Harwath, Samuel Thomas, Hilde Kuehne, Brian Chen, Rameswar Panda, Rogerio Feris, Brian Kingsbury, Michael Picheny, et al. Cascaded multilingual audio-visual learning from videos. *arXiv preprint arXiv:2111.04823*, 2021.

[56] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[57] Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. How2: a large-scale dataset for multimodal language understanding. *arXiv preprint arXiv:1811.00347*, 2018.

[58] Jason M Saragih, Simon Lucey, and Jeffrey F Cohn. Face alignment through subspace constrained mean-shifts. In *ICCV*, pages 1034–1041. IEEE, 2009.

[59] Klaus R Scherer, Tom Johnstone, and Gundrun Klasmeyer. Vocal expression of emotion. *Handbook of affective sciences*, pages 433–456, 2003.

[60] Behjat Siddiquie, Dave Chisholm, and Ajay Divakaran. Exploiting multimodal affect and semantics to identify politically persuasive web videos. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 203–210, 2015.

[61] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526. Springer, 2016.

[62] Karan Sikka, Karmen Dykstra, Suchitra Sathyanarayana, Gwen Littlewort, and Marian Bartlett. Multiple kernel learning for emotion recognition in the wild. In *ICMI*, pages 517–524. ACM, 2013.

[63] Karan Sikka, Karmen Dykstra, Suchitra Sathyanarayana, Gwen Littlewort, and Marian Bartlett. Multiple kernel learning for emotion recognition in the wild. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 517–524. ACM, 2013.

[64] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[65] Jennifer J Sun, Ting Liu, Alan S Cowen, Florian Schroff, Hartwig Adam, and Gautam Prasad. Eev dataset: Predicting expressions evoked by diverse videos. *arXiv e-prints*, pages arXiv–2001, 2020.

[66] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640, 2016.

[67] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[68] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access, 2019.

[69] Carl Vondrick, Deniz Oktay, Hamed Pirsiavash, and Antonio Torralba. Predicting motivations of actions by leveraging text. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[70] Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. Persuasion for good: Towards a personalized persuasive dialogue system for social good. *arXiv preprint arXiv:1906.06725*, 2019.

[71] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4581–4591, 2019.

[72] Yi Wang, Pierre-Marc Jodoin, Fatih Porikli, Janusz Konrad, Yannick Benezeth, and Prakash Ishwar. Cdnet 2014: An expanded change detection benchmark dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 387–394, 2014.

[73] Le Wu, Junwei Li, Peijie Sun, Richang Hong, Yong Ge, and Meng Wang. Diffnet++: A neural influence and interest diffusion network for social recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[74] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. *arXiv preprint arXiv:1611.05431*, 2016.

[75] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016.

[76] Keren Ye, Narges Honarvar Nazari, James Hahn, Zaeem Hussain, Mingda Zhang, and Adriana Kovashka. Interpreting the rhetoric of visual advertisements. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[77] S Yogesh, N Sharaha, and S Roopan. Digital marketing and its analysis. *International Journal of Innovative Research in Computer and Communication Engineering*, 5(7):201957007, 2019.

[78] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134, 2019.

[79] Fajie Yuan, Xiangnan He, Alexandros Karatzoglou, and Liguang Zhang. Parameter-efficient transfer from sequential behaviors for user modeling and recommendation. In

Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 1469–1478, 2020.

[80] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, 2018.

[81] Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun. Leveraging video descriptions to learn video question answering. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[82] Mingda Zhang, Rebecca Hwa, and Adriana Kovashka. Equal but not the same: Understanding the implicit relationship between persuasive images and text. *arXiv preprint arXiv:1807.08205*, 2018.

[83] Shiyi Zhang, Yanni Shen, Tao Xin, Haoqi Sun, Yilu Wang, Xiaotong Zhang, and Siheng Ren. The development and validation of a social media fatigue scale: From a cognitive-behavioral-emotional perspective. *PLoS ONE*, 16, 2021.

[84] Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10668–10677, 2020.

[85] Hang Zhao, Zhicheng Yan, Lorenzo Torresani, and Antonio Torralba. Hacs: Human action clips and segments dataset for recognition and temporal localization. *arXiv preprint arXiv:1712.09374*, 2019.