

# DeVAn: Dense Video Annotation for Video-Language Models

Tingkai Liu<sup>1\*</sup>, Yunzhe Tao<sup>1</sup>, Haogeng Liu<sup>2,3</sup>, Qihang Fan<sup>2,3</sup>, Ding Zhou<sup>1</sup>,  
Huaibo Huang<sup>2</sup>, Ran He<sup>2</sup>, Hongxia Yang<sup>1</sup>

<sup>1</sup>ByteDance, Inc.

<sup>2</sup>MAIS & CRIPAC, Institute of Automation, Chinese Academy of Sciences, China

<sup>3</sup>School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China

## Abstract

We present a novel human annotated dataset for evaluating the ability for visual-language models to generate both short and long descriptions for real-world video clips, termed **DeVAn** (Dense Video Annotation). The dataset contains 8.5K YouTube video clips of 20-60 seconds in duration and covers a wide range of topics and interests. Each video clip is independently annotated by 5 human annotators, producing both captions (1 sentence) and summaries (3-10 sentences). Given any video selected from the dataset and its corresponding ASR information, we evaluate visual-language models on either caption or summary generation that is grounded in both the visual and auditory content of the video. Additionally, models are also evaluated on caption- and summary-based retrieval tasks, where the summary-based retrieval task requires the identification of a target video given *excerpts* of a given summary. Given the novel nature of the paragraph-length video summarization task, we compared different existing evaluation metrics and their alignment with human preferences and found that model-based evaluation metrics provide more semantically-oriented and human-aligned evaluation. Finally, we benchmarked a wide range of current video-language models on DeVAn, and we aim for DeVAn to serve as a useful evaluation set in the age of large language models and complex multi-modal tasks. Code is available at <https://github.com/TK-21st/DeVAn>.

## 1 Introduction

With billions of active users on video content platforms such as YouTube and TikTok, there has been an unprecedented need for automated complex video understanding. Classically, video understanding has focused on captioning and/or retrieval tasks on short videos with brief (sentence-long) captions. The concise nature of both the videos

selected and captions labeled has partly been the result of model limitations, where detailed and nuanced multi-sentence video descriptions have not been possible with lightweight text decoders. With the recent leaps in large language models (LLMs), however, vision-language models (VLMs) now have the opportunity to tap into the immense natural language capabilities of models such as LLaMA (Touvron et al., 2023a,b) and ChatGPT (Ouyang et al., 2022; OpenAI, 2023). With tens to hundreds of billions of parameters, these LLMs are able to write entire essays with details and poise that mimic human to an unprecedented extent. With video conversational models such as ImageBind-LLM (Han et al., 2023), Video-LLaMA (Zhang et al., 2023), Video-ChatGPT (Maaz et al., 2023) and VideoChat (Li et al., 2023b) claiming to be able to generate detailed and fine-grained descriptions of video inputs, we believe the time is ripe for an evaluation benchmark that matches the capabilities of modern VLMs powered by LLMs.

In the current work, we focus on videos with multi-shot compositions containing diverse information streams such as dialogues, background music, and complex visual sequences. We developed **DeVAn**, a novel task and dataset for dense long-form video descriptions. This new multi-modal dataset contains 8.5K video clips carefully selected from previously published YouTube-based video datasets (YouTube-8M (Abu-El-Haija et al., 2016) and YT-Temporal-1B (Zellers et al., 2022)) that integrate visual and auditory information. Over the span of 10 months, a team of 24 human annotators (college and graduate level students) created 5 short captions (1 sentence each) and 5 long summaries (3-10 sentences) for each video clip, resulting in a rich and comprehensive human-annotated dataset that serves as a robust ground truth for subsequent model training and evaluation (See Figure 1 for example).

As opposed to short video captions where N-

Corresponding: [tingkai.liu@columbia.edu](mailto:tingkai.liu@columbia.edu)

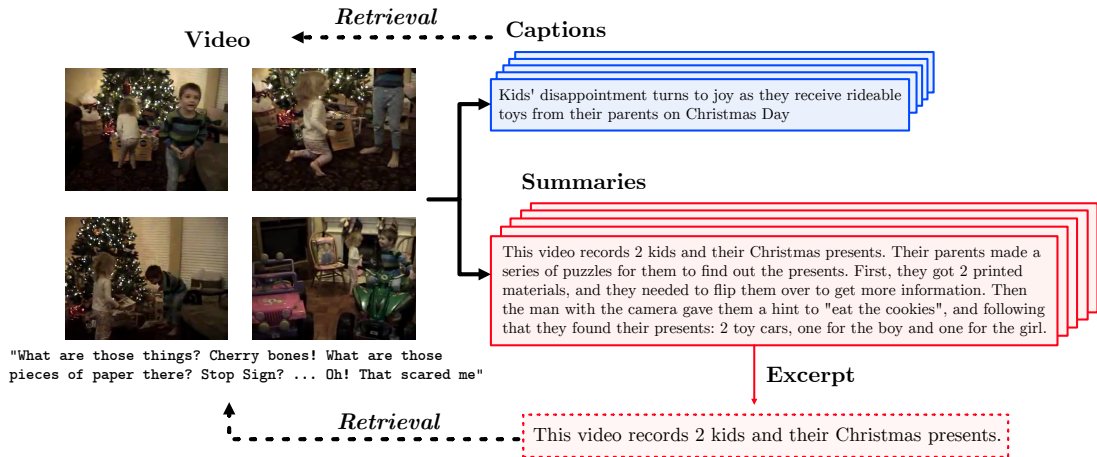


Figure 1: **Example of DeVAN dataset.** For each video, 5 captions and 5 summaries are independently annotated based on both visual and auditory information of the selected videos. Text-to-video retrievals from video summaries are performed by randomly sampling single-sentence excerpts.

gram based metrics such as CIDEr (Vedantam et al., 2015) offer good alignment with human preferences, it is not immediately apparent as how to evaluate long-form video summarization. Additionally, poor instruction-following capabilities of VLMs with frozen LLMs may result in significant variability in the lengths of video descriptions generated, leading to deceptively low N-gram based metric values even for short video captioning tasks. By comparing N-gram-based metrics and model-based metrics (e.g. BLEURT (Sellam et al., 2020), BERTScore (Zhang et al., 2020)) to human preferences, we find that model-based metrics are better able to capture semantic similarities between model generated responses and human annotations.

Finally, we evaluate different types of VLM architectures on our test set, aiming to provide a comprehensive landscape of what is currently feasible and effective on our task. Specifically, we compare a wide range of recent models featuring frozen LLMs (e.g., VideoChatGPT, ImageBind-LLM, Video-LLaMA) to an end-to-end foundation model developed based on VideoCoCa (Yan et al., 2023). By training an instance of VideoCoCa using our own training set designed for long-form video summarization, we aim to provide an effective end-to-end baseline model that covers all three tasks at hand.

Our contributions are summarized as follows:

- We introduce a new dataset of human annotated video caption (1 sentence) and summaries (3-10 sentences) to gauge the ability of VLMs to perform long form summary of video content. To the best of our knowledge, DeVAN is the first com-

prehensive human-annotated evaluation dataset for long-form open-domain video summaries.

- We compare different evaluation metrics for long-form video summarization task and find that model-based metrics offer better alignment to human preference.
- We evaluate a wide range of video-language models on our DeVAN benchmark including both models with frozen LLMs and end-to-end foundation models, and compare their performances with/without audio information.

## 2 Related Work

The endeavor to understand and provide textual descriptions of video content has been the subject of numerous research initiatives. We briefly review recent models and datasets relevant to DeVAN.

**Video-Language Models** In the video-language model landscape, two model architectures are prevalent. The first category encompasses the *end-to-end trainable models* such as BLIP (Li et al., 2022) and VideoCoCa (Yan et al., 2023), which are designed to learn representations from both videos and text simultaneously, without any frozen modules. In contrast, pioneered by models including BLIP-2 (Li et al., 2023a), *models with frozen modules* have dominated the video-language scene since the introduction of powerful LLMs like ChatGPT (Ouyang et al., 2022). Models like BLIP-2 (Li et al., 2022), Video-LLaMA (Zhang et al., 2023), Video-ChatGPT (Maaz et al., 2023), and VideoChat (Li et al., 2023b) augment pre-trained frozen lin-

guistic components with additional trainable components, often a lightweight visual backbone. Effectively, such models take advantage of the natural language capabilities of LLMs by providing soft prompts encoded by a lightweight trainable multimodal adaptor. As LLMs are capable of both consuming and generating texts with hundreds if not thousands of words, models in this category are often capable to generate long and detailed video descriptions. We hope that our DeVAn benchmark will contribute the continued advancements in these powerful video-language models.

**Video-Language Datasets** Datasets in this domain can be broadly categorized based on their domain specificity and downstream tasks. Refer to Table 1 for comparison. Under *domain specificity*, datasets like MSVD (Chen and Dolan), MSR-VTT (Xu et al., 2016), YouTube-8M (Abu-El-Haija et al., 2016), YT-Temporal-1B (Zellers et al., 2022), HD-Vila-100M (Xue et al., 2022) provide a panoramic view of diverse video content, fostering a comprehensive model understanding. In contrast, datasets such as How2 (Sanabria et al., 2018) YouCook2 (Zhou et al., 2017) and HowTo100M (Miech et al., 2019) predominantly focus on instructional content.

In terms of *task orientation*, open-domain datasets mentioned above are often focused on video-to-text generation and retrieval tasks. In contrast, datasets such as Kinetics-700 (Carreira et al., 2022), ActivityNet (Caba Heilbron et al., 2015) and ActivityNet Captions (Krishna et al., 2017) focus on more specialized downstream applications such as activity detection.

As our dataset is designed to primarily gauge the ability for models to accurately capture a balanced understanding of overall content and details in a given video, we focused on generation and retrieval tasks for open domain videos.

### 3 DeVAn Dataset

In this section, we describe the procedure with which DeVAn was constructed and how generation and retrieval task performances are evaluated on DeVAn.

#### 3.1 Evaluation Dataset

The dataset utilized in this study is an amalgamation of YouTube videos, which were source from two previously available large-scale video datasets: YouTube-8M and YT-Temporal-1B.

The selection of videos for human annotation was focused on English videos with high quality and diversity, and saw one significant evolution during the course of the annotation process. Refer to Figure 6 in Appendix A for examples of relevant metadata information used during the video selection process.

**First Phase: 2.3K Videos** In the **first phase** of the data curation process, videos are selected from YouTube-8M and YT-Temporal-1B datasets based solely video metadata with the following criteria:

- Video title, description and subtitles (if applicable) must be primarily in English;
- Video must contain Chapter information, which is video keyframe information provided by video uploaders;
- Video clips, when segmented based on chapter information, should be between 20 to 60 seconds.

We find that of all videos in the YouTube-8M and YT-Temporal-1B datasets, roughly 1% satisfied our constraint. Based on the “category” metadata information of the videos, we uniformly sampled around 2.3K video segments were curated following this procedure, which form the **first phase** of our data annotation process. Note that the 100K training dataset mentioned later in this paper was curated in tandem with the **first phase** evaluation dataset, as such the distribution of our training dataset aligns best with this portion of the test set (see Section 3.4 for more information).

**Second Phase: 6.2K Videos** In the **second phase** of the data annotation process, we adjust the criteria to favor videos for which visual-grounding is *necessary* for accurate annotation. In particular, previous selection criteria (most significantly, the requirement for “Chapter” information) led to a bias towards News and Instruction type videos, for which speech information contents were dominant.

As a result, human annotators often heavily favored, for example, the *content* of the News articles being broadcast over the actual *visual setting* of the broadcasting room. This has the undesired consequence that visual-language models with strong language capabilities but weaker visual grounding can potentially have better performance than the more visually-grounded counterparts. Additionally, to avoid videos that are montages of static images,

Dataset	Annotation	Source	Duration (hrs)	Domain	Videos	Clips	Clip Length	Description Length (words)
HowTo100M	Automatic	YouTube	134,472	Instruction	1.221M	136M	6.5 min	-
ActivityNet Captions		YouTube	849	Open	100k	100k	180 sec	-
MSVD	Manual	YouTube	4.13	Open	-	2k	10 sec	7 (test)
MSR-VTT		Commercial Engine	41.2	Open	7,180	10k	10-30 sec	9.3 (test)
Vatex		Kinetics-600 + YouTube	1,300	Open	-	41.3k	20 sec	14.5 (test)
YouCook2		YouTube	175.6	Instruction	-	2k	315 sec	8.8 (train/val)
<b>DeVAn</b>	Manual	YouTube	95.9	Open	6,709	8.5k	40.7 sec	11 (caption) / 54 (summary)

Table 1: **Video-Language Datasets Comparison.** Refer to Figure 7 in Appendix B for detailed information on DeVAn.

we apply an additional frame-embedding based filter to select videos with high inter-frame variability. The video selection process is as follows.

1. The audio content of each video is first processed by Whisper-Base (Radford et al., 2022) to generate automatic speech recognition (ASR) content, followed by an entropy computation, where only videos with entropy lower than 4.2 are kept.
2. The visual content of each video is evaluated by uniformly sampling 8 frames and computing embedding of each frame using CLIP (Radford et al., 2021);  $L_2$  distances between embeddings of neighboring frames are computed and averaged, where only videos with average inter-frame  $L_2$  distance above 5.5 are kept.
3. Video title, description and subtitles (if applicable) must be primarily in English.
4. Instead of “Chapter” information, videos are segmented using key-frames detected via TransNet (Souček et al., 2019). Only segments that satisfy the 20-60 second duration requirement are kept.

As in the First Phase, the automatically filtered videos were again sampled uniformly based on the “category” metadata information. Selected video clips are filtered manually during the annotation process where annotators were provided the option to discard a given video if it is deemed of poor quality: non-English or does not contain sufficient information content for summarization in 3-10 sentences. We find that roughly 20% of automatically selected videos were filtered by human annotators.

The 10-months-long annotation process is divided into multiple rounds, with each round covering 500-1500 videos. After each round of annotation, 20% of videos are randomly selected for quality control independent of the original annotators. If systematic problems are detected in a batch

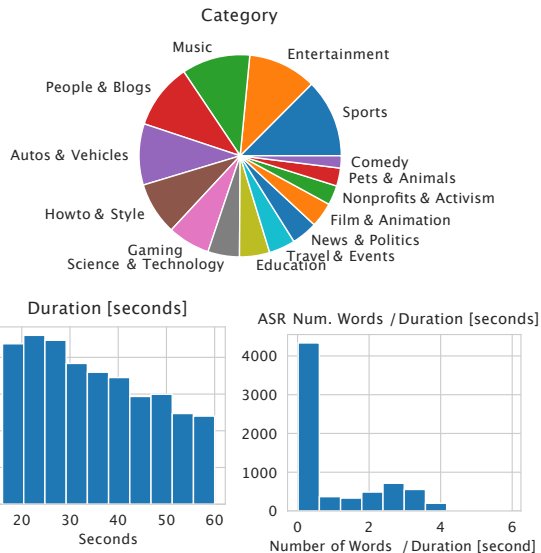


Figure 2: **Diversity of DeVAn dataset.** Our dataset contains English videos covering a diverse range of topics uploaded across the past 17 years.

of annotations, the entire batch is returned to annotators for revision before going through another round of quality control. In later rounds, as quality of annotation stabilized, the percentage of videos selected for independent quality control is adjusted downwards to a minimum of 7.5%. This process is repeated until the batch at question is deemed of satisfactory quality, and every batch went through at least one round of revision. Refer to Appendix F for more details on the annotation process.

As shown in Figure 2, the final 8.5K evaluation dataset contains videos covering a wide range of topics and interests. The statistics of the videos and annotations are shown in Table 1 and Figure 7 in Appendix B. For more qualitative examples of the dataset, refer to Appendix H.

### 3.2 Tasks and Evaluation Metrics

**Video-to-Text Generation Task** For video-to-text generation tasks, we compared both N-gram-based and model-based evaluation metrics. For



N-gram-based metrics, following prior works such as MSR-VTT (Xu et al., 2016), we report commonly used metrics including BLEU (Papineni et al., 2001), ROUGE-L (Lin, 2004) and CIDEr (Vedantam et al., 2015) to gauge the quality of the model-generated captions. Evaluation of these metrics follow the implementation used in CLIP-Score (Hessel et al., 2021), where the Stanford CoreNLP’s PTBTokenizer (Manning et al., 2014) is used for text pre-processing. These metrics have proven effective for evaluating the lexical overlap and syntactic structure in brief captions, which are relatively straightforward and independent of language models. In addition to N-gram-based metrics, we also report model-based metric, BLEURT (Sellam et al., 2020), which we found to have a better agreement to human preferences (see Section 3.3 for more details) especially for long-form video summarization tasks. As such, while we report a wide range of evaluation metrics, the model-based BLEURT metric serves as our primary method of evaluation for video-to-text generation tasks.

**Text-to-Video Retrieval Task** For evaluating the efficacy of models in the video retrieval tasks, we follow the classic retrieval accuracy metrics at different levels of granularity: Recall @1, @5, and @10. While this is sufficient for standard text-to-video retrieval tasks using one-sentence video descriptions, it is not directly applicable for retrieval via multi-sentence video summaries. For such task, we introduce a new evaluation methodology where the recall is evaluated using individual sentences from a given video summary. The overall summary-to-video retrieval performance is the averaged recall from each sampled sentences. Note that only sentences with more than 5 words were used to ensure that excerpts of video summaries contain sufficient information for retrieval. This task mimics the common scenario where viewers may desire to search for videos based on memories of partial information. We report Recall @1, @5 and @10 for both caption-to-video and summary-to-video retrieval tasks.

### 3.3 Alignment of Evaluation Metrics to Human Preference

Given the novel nature of long-form video summarization task, we sought to compare the alignment of different evaluation metrics to human preferences.

To start, we computed the Spearman Rank Correlation between different evaluation metrics on both captioning and summarization tasks as shown in Figure 8. Note that Spearman Rank Correlation was chosen over Pearson Correlation to emphasize pairwise consistency between evaluation metrics. Using one annotated response as prediction and all other responses as ground truths, we computed N-gram based (BLEU-4, ROUGE-L, CIDEr) and model-based (BLEURT) metrics for both captioning and summarization tasks. We observe that across all annotators and all videos, the correlation between N-gram-based and model-based metrics is 0.65 for captioning task, while the minimum correlation within N-gram metrics is 0.73. This result suggests that N-gram based metrics may offer more consistent evaluation for captioning task. In contrast, N-gram metric such as CIDEr appear poorly correlated with other metrics for long-form video summarization tasks.

To determine the most suitable metric for evaluating long-form summarization task, we intuited that the alignment of a metric to human preferences can be measured by the metric’s ability to tightly cluster annotations of the same video created by different labelers. Formulating this intuition as a text-to-text retrieval problem, we compared recall performances of identifying the same annotations from different annotators using different evaluation metrics, and found that while CIDEr and BLEURT have similar recall performances for video captions, BLEURT significantly outperforms all N-gram based metrics for video summaries by over 14% (see Table 5 in Appendix C). This provides indirect support that the model-based BLEURT metric may be better aligned to human preferences especially for long-form video summaries. To further validate this result, we randomly selected 20 videos and manually ranked summaries from annotators 1 and 5 by their perceived quality and relatedness to the video content. We then compared this human labeled ranking result to the ranking created by aforementioned metrics, computed using label 1/5 as predictions and label 2/3/4 as ground truths. We observe a 56% alignment of human ranking to ranking by CIDEr score and a 67% alignment to that by BLEURT, reinforcing the previous finding that BLEURT is better aligned to human preferences.

It is worth noting that while the difference in human alignment between BLEURT and CIDEr

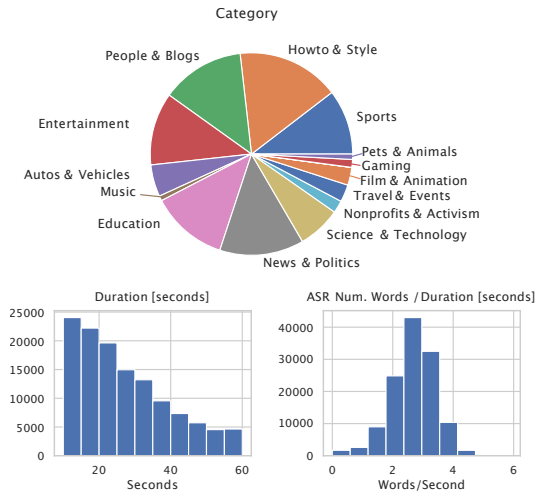


Figure 3: **Diversity of training dataset.** Our training dataset contains captions and summaries for 100K ASR-rich video segments. Note that as opposed to test dataset in Figure 2, the ASR Number of Words normalized by video duration does not have a significant concentration around 0, indicating that all videos in the training dataset contains a significant amount of ASR information.

appears significant, it is biased by preferences by human labelers and limited by the number of videos sampled during comparison. Nevertheless, the difference between CIDEr and BLEURT for video summaries point to the differences between semantic and lexical similarities when evaluating paragraph-length textual descriptions. We chose BLEURT as our metric for video summarization task but it remains unclear how better alignment to human preferences should be evaluated and implemented for complex multi-modal tasks with long form text generation.

### 3.4 Training Dataset

In addition to the 8.5K evaluation dataset, we also prepared a training dataset with 100K video clips whose captions and summaries were generated using metadata information (e.g., title, description, category, ASR, etc.). The training dataset is comprised of 100,000 videos segments, selected via the same procedure as in the **first phase** of evaluation dataset curation (see Section 3.1). Note due to time constraint, we were not able to collect another training set based on the selection criteria in the **second phase** of evaluation dataset development. Consequently, the training dataset is skewed towards ASR-rich videos (see Figure 3).

Similar to videos in the evaluation dataset, videos in the training dataset were segmented based on the video chapter information. A prompt tem-

plate was then used to create queries for an LLM to generate 5 captions and summaries for each video (see Appendix D). To exercise control over the text generation process and maintain consistency, each prompt was prefixed with “This video”. At the time of dataset creation, the `gpt-3.5-turbo` (Ouyang et al., 2022) model was observed to have significantly more issues related to hallucinations as compared to `text-davinci-003` (Brown et al., 2020), thereby motivating the choice of the seemingly less powerful but more “reliable” `text-davinci-003` for our specific requirements.

It is important to note that the main focus of the current work is on the *zero-shot* performances of current models on our DeVAn *evaluation* dataset, and, as later discussed in Section 4.3, only the end-to-end VideoCoCa model was fine-tuned on the DeVAn training dataset. This is due to the fact that our VideoCoCa model was initialized from the CoCa-ViT-L-14 ckpt (see Section 4.3 and Appendix E for more details), which was originally trained on COCO-styled short captions. Direct inference of such model for video summarization tasks resulted in catastrophically poor performance as it lacks the capability to be prompted via natural language to generate video descriptions of different lengths (see Appendix G). However, as the DeVAn training dataset was created using only the metadata information of the videos, the training dataset differ significantly from the evaluation dataset in terms of category, audio content, description length, and vocabulary, thereby constituting a zero-shot evaluation of the VideoCoCa model.

## 4 Experiments

In this section, we describe both human performance and model performance on the DeVAn evaluation set. To comprehensively evaluate the capabilities and limitations of various architectures for video captioning and retrieval, we consider two distinct types of models, each with its own set of advantages and drawbacks.

### 4.1 Human Performance

To establish a human performance benchmark, we use a strategy similar to that described in Section 3.3 where one human annotation is used as the “prediction”, while the remaining four annotations are used as “references”. The aggregated results across annotators for both captioning and

Model	Audio	Caption						Summary							
		Generation Metrics			Retrieval Metrics			Generation Metrics			Retrieval Metrics				
		B4	R	C	B-RT	R@1	R@5	R@10	B4	R	C	B-RT	R@1	R@5	R@10
Human (Avg)	Raw	6.3	32.1	53.9	50.5	-	-	-	15.7	34.5	36.9	55.6	-	-	-
Human (Min)	Raw	4.5	29.5	47.1	48.6	-	-	-	12.4	32.1	30.9	53.6	-	-	-
ImageBind-LLM	N/A	0.3	20.0	2.1	34.0	-	-	-	1.5	22.7	1.1	45.8	-	-	-
Video-LLaMA2 13B	N/A	0.1	7.9	0.0	<b>47.2</b>	-	-	-	0.5	18.2	0.0	39.9	-	-	-
	Raw	0.1	7.9	0.0	47.1	-	-	-	0.5	18.2	0.0	40.0	-	-	-
Video-LLaMA2 7B	N/A	0.1	10.8	0.0	43.6	-	-	-	0.5	19.1	0.0	43.9	-	-	-
	Raw	0.1	10.8	0.0	43.6	-	-	-	0.5	19.1	0.1	43.9	-	-	-
VideoChatGPT	N/A	0.4	19.9	2.0	40.5	-	-	-	<b>2.9</b>	<b>24.4</b>	<b>5.8</b>	<b>46.7</b>	-	-	-
VideoCoCa	N/A	0.2	13.2	2.3	17.6	32%	50%	58%	0.9	16.4	3.3	23.9	25%	41%	48%
	ASR	<b>0.8</b>	<b>20.3</b>	<b>9.2</b>	21.9	<b>36%</b>	<b>53%</b>	<b>59%</b>	2.0	21.6	5.5	22.9	<b>27%</b>	<b>42%</b>	<b>48%</b>

Table 2: **Results for generation and retrieval tasks of DeVAn evaluation dataset.** For evaluation of human performance, annotation from each annotator is used as *prediction* and computed against *ground truth* results from all other 4 annotators. The overall metrics are then aggregated via Average and Minimum. Note that only results for VideoCoCa models are shown for retrieval tasks, since other VLMs tested in the current work do not supported retrieval. **B4**: BLEU-4, **R**: Rouge-L, **C**: CIDEr, **B-RT**: BLEURT. **R@k**: Recall at k.

Task	Comparison	Metric	
		C	B-RT
Caption	Across Annotators	31	-
	Across Time	41	-
Summary	Across Annotators	50	60
	Across Time	38	56

Table 3: **Same annotator re-labeling summaries of the same video twice is equivalent to the same video summary labeled by different annotators.** 150 randomly selected captions and summaries were relabeled by the same annotator at least 1 month apart from the original annotation. We observe that annotations by the same annotator 1 month apart have similar level of discrepancy as compared to annotations by different annotators.

summarization tasks are shown in Table 2. For detailed metrics of each annotator, see Table 4 in Appendix C.

While the above evaluation establishes the consistency of human annotations across annotators, we also wanted to evaluate the consistency across time. To that end, we compared similarities of annotations for the same video created by the same annotator more than one month apart. As shown in Table 3, compared to other annotators, same annotator is able to achieve a higher CIDEr score for captioning task but a lower CIDEr and BLEURT score for summarization task. This indicates that annotators are able to maintain a much higher consistency for captioning task over summarization task, which is reasonable given the extensive nature of long-form video summarization. In fact, comparison of BLEURT scores of summarization in Table 3

indicates that annotation consistency *across time* is similar that *across annotators*, suggesting the high subjective difficulty of the long-form video summarization task. It is therefore crucial to have multiple (5 in our case) summary annotations for each video to ensure the diversity of ground truths.

## 4.2 Trainable Visual Encoder with Adaptor on Frozen LLMs

Recently, many methods combining visual encoders with frozen LLMs have emerged. Pioneered by models including BLIP-2, recent additions to this category of models include ImageBind-LLM, Video-LLaMA2 and VideoChatGPT, which all involve projecting video encoding (visual and optionally auditory information) into soft tokens that serve as prefix for frozen LLMs. We performed zero-shot instruction-based evaluations of these models with default system prompts from demos in their corresponding repositories, and using instructions “Describe this video in ONE sentence.” for captioning task and “Summarize this video in three to ten sentences.” for summarization task.

Surprisingly, as shown in Table 2, we found that models like Video-LLaMA2 have very poor video-to-text generation performances for both captioning and summarization when evaluated using N-gram-based metrics. Upon closer examination of the generated results, we realize that due to poor instruction-following (or a lack of prompt engineering), models with frozen LLMs generated responses with highly variable length as shown in Figure 4. Pathological examples of such cases with close to 0 CIDEr but high BLEURT scores are

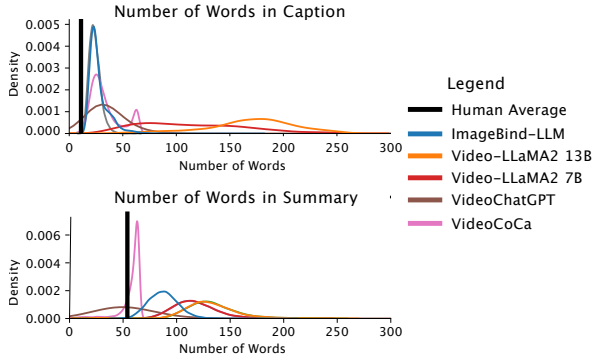


Figure 4: **Distribution of Number of Words in Captions and Summaries.** Note that for legibility, only distributions for models without audio signals were shown. However, we found that the distributions of caption and summary lengths do not vary significantly with the introduction of audio signals.

shown in Appendix I, which all corresponding to very long, sometimes repeating, generated captions. As such, while the results generated are semantically related to the video and human annotations (as found both via human evaluation and BLEURT), they have very low performances in N-gram-based evaluations.

### 4.3 End-to-End Foundation Model

Despite recent success of VLMs with frozen LLMs, they often cannot support retrieval tasks due to the lack of a dedicated language encoder. To provide a baseline model that is able to achieve competitive results in both video-to-text generation and text-to-video retrieval tasks, we developed an end-to-end trainable foundation model based on the VideoCoCa architecture (Yan et al., 2023) that includes both a visual encoder and ASR encoder (see Figure 5 for model architecture). As no open source VideoCoCa implementation was available, we created an implementation from scratch following the **Attention Pooler** type model described in the VideoCoca manuscript.

Following CoCa (Yu et al., 2022) and VideoCoCa, the text encoder takes in Caption or Summary as input where a special [CLS] token is suffixed to all input sequences. The text encoder is evenly divided into unimodal (bottom) and multi-modal parts, where the output encoding of the [CLS] token by the unimodal encoder is used to compute contrastive training objective against other modalities.

On the visual encoder side, 8 frames are uniformly sampled from each video and encoded in-

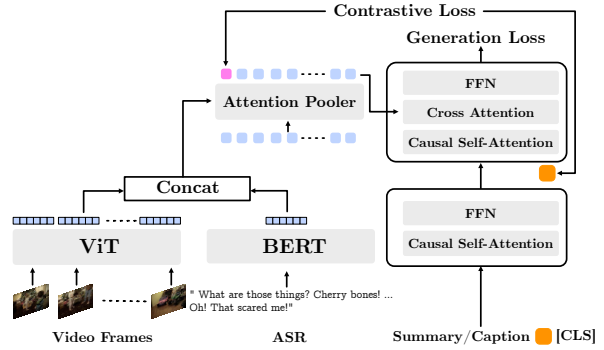


Figure 5: **End-to-End Model Architecture.** Our End-to-End model combines VideoCoCa model architecture with an additional ASR encoder. Frame-level embeddings of VideoCoCa and ASR embeddings are concatenated before passing through the Attention Pooler. Note that the contrastive loss is computed using the first output embedding of the Attention Pooler.

dependently by ViT. The output visual encoding of each frame is concatenated to form the overall representation of the visual information in an input video. The visual encoder output is further concatenated with BERT encoding of the ASR information. The encodings are then integrated and compressed by the “Attention Pooler” module with 256 query tokens. The first output token of Attention Pooler is used to compute contrastive loss against encoding of [CLS] token mentioned above.

Following CoCa, the training objective is a combination of generation loss and contrastive loss. All parameters are initialized from the OpenCLIP (Ilharco et al., 2021) implementation of CoCa, except for the ASR encoder which was initialized from BERT-base (Devlin et al., 2019). Results of the video-to-text generation and text-to-video retrieval on DeVAn can be found in Table 2. We note that the integration of audio information via ASR significantly improves both retrieval and generation performances of our model. This is consistent with the observation that roughly 25% of the evaluation dataset contains videos rich in speech content.

## 5 Conclusion

In this paper, we introduce a new multi-modal dataset curated from a diverse range of YouTube videos, designed to gauge the capabilities of visual-language models in long-form video summarization tasks. Through a carefully orchestrated annotation process involving multiple human annotators, multiple rounds of video selection and quality control, we ensure that the dataset is comprised of



high-quality and diverse captions and summaries.

We show that while qualities of one-sentence captions can be accurately evaluated by N-gram, multi-sentence summaries require a more semantically aligned metric such as BLEURT. Using such metrics, we provide an extensive benchmarking of the current state-of-the-art video-to-text generation models. However, current visual-language models with frozen LLMs often do not support retrieval tasks due to lack of language encoders. To provide a baseline for text-to-video retrieval tasks, we finetuned VideoCoCa using our automatically generated training dataset, which achieved reasonable retrieval, but comparatively weaker generation performances.

As the field of video summarization and captioning continues to evolve, it is imperative that datasets and evaluation metrics keep pace. Our work aims to serve as a stepping stone in this direction, providing a balanced approach to video understanding.

## 6 Limitations

While DeVAn strives to become a comprehensive and objective evaluation benchmark for video-to-text generation and text-to-video retrieval tasks, certain limitations warrant consideration.

- A limitation of the current work both in the creation of training dataset and in the benchmarking of existing models is that of prompt engineering. It has long been observed that LLMs can be highly sensitive to prompt design (Liu et al., 2021), which may contribute to the high variability in the length of the generated responses. While we tried to mitigate this problem by staying close to the default instructions in used in the corresponding works, we still observed problems of poor instruction following.
- Additionally, despite an effort to compare different available metrics, neither CIDEr nor BLEURT achieved higher than 80% alignment to human preferences. While it is feasible to pose the problem in a multiple choice format to mitigate this problem, this would usually require finetuning models to following MCQ instructions. Alternatively, a Reward Model for video annotations may be trained based on human preferences (similar to finetuning BLEURT). However, this would require significantly more human annotations which was unfortunately infeasible.

Due to time and resource constraints, these limitations were not addressed in the current work. However, we do expect performance gains of the evaluated models by addressing these limitations.

## 7 Ethics Statement

This research aims to provide an objective and comprehensive benchmark for video-to-text generation and text-to-video retrieval tasks from open domain videos. Although our research does not involve human subjects directly, it is important to acknowledge and discuss the broader ethical implications.

**Data Bias and Fairness:** Both training and testing videos in DeVAn are selected from YouTube-8M and Youtube-Temporal-1B datasets, and follow their data curation and anonymization practices. While these datasets are widely used, we acknowledge that we cannot fully ascertain the extent to which they may contain discriminatory, biased, or sensitive material. Given that our finetuned VideoCoCa model inherits the biases present in our training datasets, there exists the risk of perpetuating or even amplifying existing societal biases. Despite the broad acceptance of these datasets, caution should be exercised.

**Responsible Usage:** Like all open domain video datasets, it is imperative to implement safety mechanisms to ensure that models trained or evaluated on our dataset do not inadvertently produce outputs that could disclose sensitive or personal information.

**AI Assistant Usage:** ChatGPT was used for grammatical correction in the current manuscript.

## References

- Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. [YouTube-8M: A Large-Scale Video Classification Benchmark](#). ArXiv:1609.08675 [cs].
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#).

- Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. 2022. [A Short Note on the Kinetics-700 Human Action Dataset](#). ArXiv:1907.06987 [cs].
- David Chen and William Dolan. Collecting Highly Parallel Data for Paraphrase Evaluation.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [Bert: Pre-training of deep bidirectional transformers for language understanding](#).
- Jiaming Han, Renrui Zhang, Wenqi Shao, Peng Gao, Peng Xu, Han Xiao, Kaipeng Zhang, Chris Liu, Song Wen, Ziyu Guo, Xudong Lu, Shuai Ren, Yafei Wen, Xiaoxin Chen, Xiangyu Yue, Hongsheng Li, and Yu Qiao. 2023. [Imagebind-llm: Multi-modality instruction tuning](#).
- Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: a reference-free evaluation metric for image captioning. In *EMNLP*.
- Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. [Openclip](#). If you use this software, please cite it as below.
- Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. [Dense-Captioning Events in Videos](#). ArXiv:1705.00754 [cs].
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. [BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models](#). ArXiv:2301.12597 [cs].
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. [BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation](#). ArXiv:2201.12086 [cs].
- KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023b. [Videochat: Chat-centric video understanding](#).
- Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries.
- Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. [Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing](#).
- Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. [Video-chatgpt: Towards detailed video understanding via large vision and language models](#).
- Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. [The Stanford CoreNLP natural language processing toolkit](#). In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.
- Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. [Howto100m: Learning a text-video embedding by watching hundred million narrated video clips](#).
- OpenAI. 2023. [Gpt-4 technical report](#).
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#).
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. [BLEU: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, page 311, Philadelphia, Pennsylvania. Association for Computational Linguistics.
- 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](#).
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. [Robust Speech Recognition via Large-Scale Weak Supervision](#).
- Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. [How2: A large-scale dataset for multimodal language understanding](#).
- Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. [BLEURT: Learning Robust Metrics for Text Generation](#). ArXiv:2004.04696 [cs].
- Tomáš Souček, Jaroslav Moravec, and Jakub Lokoč. 2019. [TransNet: A deep network for fast detection of common shot transitions](#). ArXiv:1906.03363 [cs].

- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. [Llama: Open and efficient foundation language models](#).
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. [Llama 2: Open foundation and fine-tuned chat models](#).
- Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. [CIDEr: Consensus-based image description evaluation](#). In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, Boston, MA, USA. IEEE.
- Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. [MSR-VTT: A Large Video Description Dataset for Bridging Video and Language](#). In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5288–5296, Las Vegas, NV, USA. IEEE.
- Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. 2022. [Advancing High-Resolution Video-Language Representation with Large-Scale Video Transcriptions](#). In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5026–5035, New Orleans, LA, USA. IEEE.
- Shen Yan, Tao Zhu, Zirui Wang, Yuan Cao, Mi Zhang, Soham Ghosh, Yonghui Wu, and Jiahui Yu. 2023. [Video-Text Modeling with Zero-Shot Transfer from Contrastive Captioners](#). ArXiv:2212.04979 [cs].
- Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. [CoCa: Contrastive Captioners are Image-Text Foundation Models](#). ArXiv:2205.01917 [cs].
- Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. 2022. [MERLOT Reserve: Neural Script Knowledge through Vision and Language and Sound](#). ArXiv:2201.02639 [cs, eess].
- Hang Zhang, Xin Li, and Lidong Bing. 2023. [Video-llama: An instruction-tuned audio-visual language model for video understanding](#).
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [BERTScore: Evaluating Text Generation with BERT](#). ArXiv:1904.09675 [cs].
- Luowei Zhou, Chenliang Xu, and Jason J. Corso. 2017. [Towards automatic learning of procedures from web instructional videos](#).

## A YouTube Metadata information

YouTube metadata information refers to all information related to an uploaded video except for the video and audio content. This includes information such as title, description, category, tags, asr, chapter, playlist, subtitles, etc. In figure 6, we highlight a few important metadata information that is relevant for video selection in our work.

We highlight ‘‘Chapter’’ information, which corresponds to keyframe information with human annotated segment subtitles. These chapter information naturally divide video into semantically meaningful chunks that we used to split longer videos into 20-60 second clips.

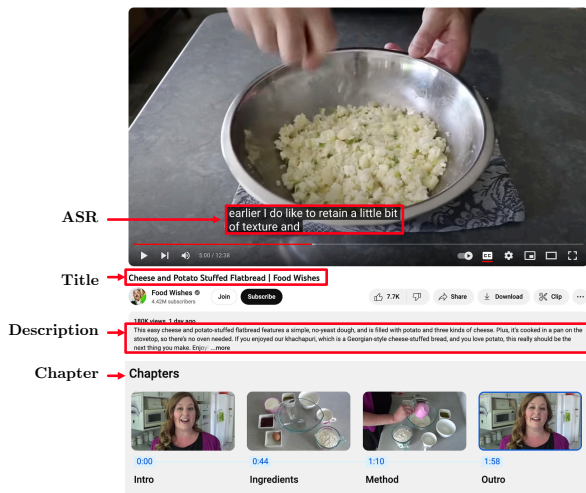


Figure 6: **Example of YouTube metadata information.** Note that only metadata that are relevant for video selection is shown. ‘‘Chapter’’ information contains keyframe information with headers provided by the video uploader.

## B Detailed Statistics of DeVAn Evaluation Dataset

Here we provide detailed statistics on the evaluation dataset. In particular, in Figure 7 second column from left, we show that the ASR content, when normalized by duration of video, demonstrates a clear bimodal distribution, corresponding to videos with high and low ASR content selected during the two phases of annotation process. Additionally, we observe that a significant portion of video summaries are longer than 100 words, which is considerably longer than video annotations in previously available datasets.

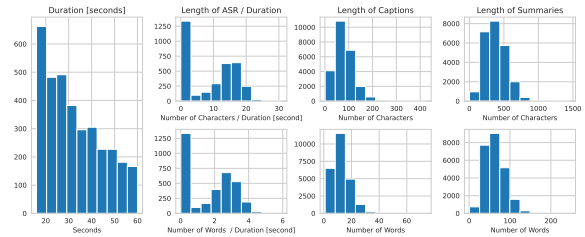


Figure 7: **Distribution of video duration, ASR content, length of captions and length of summaries of DeVAn dataset.** The *Length of ASR* normalized by duration of video clip indicates that our dataset covers videos ranging from no speech content to high speech content. Note that number of words are calculated by counting the number of white-space-separated character groups, and may contain punctuation.

## C Detailed Generation Metrics for each Annotator

Here we provide metrics comparing each annotator’s caption & summaries to all other annotators. Each row in Table 4 correspond to metric evaluated using the given annotator’s caption/summary as prediction and all others’ as ground truth. We computed both Average and Minimum values for each metric for reference, where the minimum value indicates the lower-bound of human performance.

Annotator	Caption				Summary			
	B4	R	C	B-RT	B4	R	C	B-RT
1	6.5	33.2	55.3	51.4	17.3	35.8	40.0	56.4
2	6.7	33.5	57.1	51.4	16.7	35.5	37.1	56.5
3	7.3	33.0	57.0	50.9	16.8	35.3	41.0	56.4
4	6.7	31.3	52.9	50.2	15.1	33.8	35.7	55.1
5	4.5	29.5	47.1	48.6	12.4	32.1	30.9	53.6
Avg	6.3	32.1	53.9	50.5	15.7	34.5	36.9	55.6
Min	4.5	29.5	47.1	48.6	12.4	32.1	30.9	53.6

Table 4: **Human Performance of Video-to-Text Generation Task.** Annotation from each annotator is used as *Prediction* and computed against Ground Truth results from all other 4 annotators. The overall metrics are then aggregated via Average and Minimum.

To contrast the similarities between annotations of the same video across annotators to that between different annotations, we performed a text-to-text retrieval task. The motivation behind this experiment is that a good evaluation metric should produce high similarity between annotations created by different human labelers for the same video, thereby resulting in a high text-to-text retrieval performance. Guided by this intuition, we randomly selected 100 videos, and computed pairwise metrics between annotator 1’s caption/summary to an-



notations from all other annotators. We then perform a text-to-text retrieval task using each metric and report recall@1,5,10. As shown in Table 5, for the captioning task, CIDEr and BLEURT give similar recall performance. However, for the summarization task, BLEURT metric provides much higher recall than all other N-gram based metrics, providing an indirect support for the claim that the BLEURT metric is better aligned to human preferences.

	Caption			Summary		
	R@1	R@5	R@10	R@1	R@5	R@10
BLEU-4	45%	66%	71%	64%	80%	89%
ROUGE-L	37%	59%	66%	69%	84%	88%
CIDEr	65%	80%	87%	60%	83%	92%
BLEURT	61%	87%	90%	83%	96%	99%

Table 5: Comparison of text-to-text retrieval performances across metrics.

We also computed Spearman Rank Correlation of all annotations from all annotators across evaluation metrics. As shown in Figure 8. Note that Spearman Rank Correlation was chosen to emphasize rank consistency between metrics and loosens the linearity assumption implicit in Pearson Correlation.

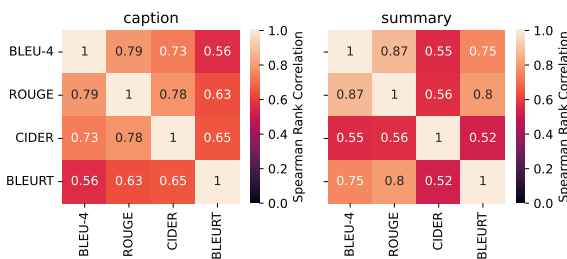


Figure 8: Spearman Rank Correlation between evaluation metrics of human annotators. Spearman Rank Correlation is evaluated between all human performance values for each video clip across annotators. Spearman Rank Correlation is chosen over Pearson Correlation to emphasize the ranking consistency using different evaluation metrics.

## D Prompt of Training Data Generation

The following prompt template is used for generating video summaries in training set. Here `cc` refers to the video subtitles.

```
summary_template = """Please write a
    ↪ summary in 3 to 10 sentences that
accurately summarizes the video's
    ↪ content and captures
its essence based on the following
    ↪ information.
```

```
Title: {{ title }}|{{ chapter_title }}
Category: {{ category }}
Description: {{ description }}
Closed Captions: {{ cc }}
SUMMARY: This video"""
```

The following prompt template is used for generation video captions in training set. Here `cc` refers to the video subtitles.

```
caption_template = """Please write a 1-
    ↪ sentence caption that
accurately summarizes the video's
    ↪ content and captures
its essence based on the following
    ↪ information.
```

```
Title: {{ title }}|{{ chapter_title }}
Category: {{ category }}
Description: {{ description }}
Closed Captions: {{ cc }}
ONE-SENTENCE CAPTION: This video"""
```

## E VideoCoCa Training Details

VideoCoCa models are trained using 64 V100-32G GPUs with global batchsize of 256 (3 per card). Training 100K dataset for 5 epochs consumes roughly 5 hours.

Learning rate of VideoCoCa follows Linear Warmup with Cosine decay pattern, with warmup LR of  $1e-7$  (500 warmup steps), peak LR of  $1e-5$  and a minimum LR of  $1e-6$ .

Both VideoCoCa and VideoCoCa w/o ASR models are trained with 8 uniformly sampled input video frames for 5 epochs.

## F Details of Human Annotation

Annotation of DeVAn occurred over a 10-months period, divided into multiple rounds, with each round covering 500-1500 videos. Prior to annotation, annotators were recruited and evaluated based on their performances on 200 held out videos. All accounted for, 24 human annotators (college and graduate level students) were recruited and performed their tasks on an online platform as shown in Figure 9. Annotators are required to fill in one caption and one summary for each video, or alternatively mark the video as invalid if it resembles a slideshow recording, is not in English, or does not

have sufficient visual information to support a 3-10 sentence video description.

After each round of annotation, 20% of videos are randomly selected for quality control independent of the original annotators. If systematic problems are detected in a batch of annotations, the entire batch is returned to annotators for revision before going through another round of quality control. In later rounds, as quality of annotation stabilized, the percentage of videos selected for independent quality control is adjusted downwards to a minimum of 7.5%. This process is repeated until the batch at question is deemed of satisfactory quality, and every batch went through at least one round of revision.

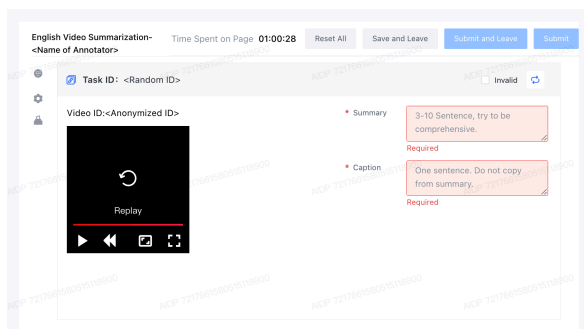


Figure 9: Screenshot of annotation platform.

## G Performance of VideoCoCa model without finetuning

In Table 6, we show the performance of VideoCoCa model (CoCa-ViT-L-14) evaluated on the DeVAn summarization task. Due to the lack of instruction following capabilities of VideoCoca, it had a catastrophically low performance on the out-of-domain video summarization task without finetuning on video summarization dataset.

Architecture	Training	BLEU-4	ROUGE	CIDEr	BLEURT
VideoCoCa	DeVAn Training Set	2.9	16.4	3.3	23.9
VideoCoCa	No Training	7.5	7.5	0.2	11.0

Table 6: Video summarization performances of VideoCoCa with and without training on DeVAn training set.

## H Examples from DeVAn Dataset

In this section, we provide more qualitative examples from DeVAn dataset. Note that though each video has correspondingly 5 captions and summaries, only one caption and one summary is shown for brevity.



ASR: Hey there, this is Derek Mithog with Voltaire Cycles Franchises and today I...  
Caption: Introducing Voltaire Cycles' hassle-free bike and trike delivery service.  
Summary: In this video, a man from Voltaire Cycles franchises is introducing their door-to-door, fully assembled delivery service for bikes and trikes. Many customers are hesitant to purchase a bike or trike online due to the assembly process involved. Most sites claim that the bike comes mostly assembled, but that still means a lot of work for the customer. Voltaire Cycles' delivery service takes away all of the guesswork and hassle by delivering the bike or trike fully assembled to the customer's door.

Figure 10: Example DeVAn data annotated from YouTube video zheszz4PLUw.



ASR: None.  
Caption: Videos from MTV that were popular in 2008 and 2009.  
Summary: In this video, a mobile phone is playing scenes from music videos that were popular on MTV in the years 2008 and 2009. This gives a nostalgic vibe, as those years were around the peak of MTV's music video era. The silver color of the phone is a characteristic detail, and it's interesting how technology has evolved since then.

Figure 11: Example DeVAn data annotated from YouTube video -da7ZrCiupo

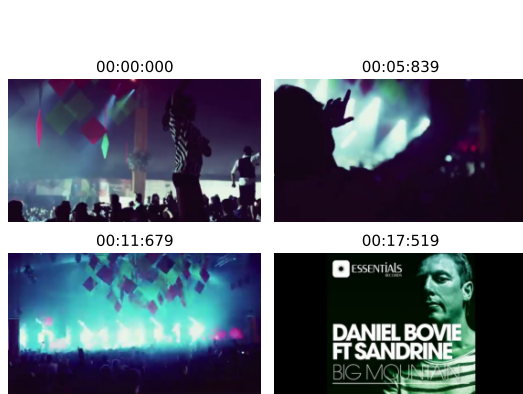
## I Pathological Examples Captions Generated from Video-Language Models

We noted that models like Video-LLaMA2-Instruct had very low CIDEr scores but high BLEURT scores. In this section, we show some pathological examples where model generated responses show close to 0 CIDEr score but very high BLEURT scores.



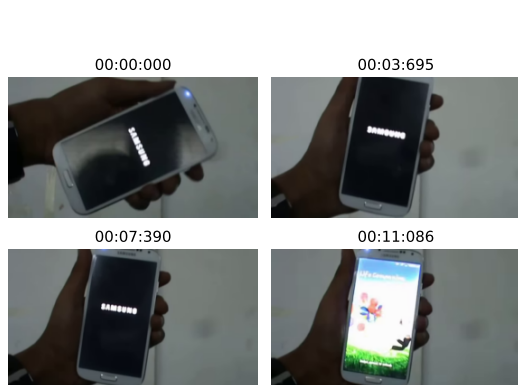
ASR: 47, empty cars. 40, dogs on a leash. 35, unseer Girlfriend is probably tryi...  
 Caption: A grey vintage car is showcased.  
 Summary: A man takes the wheel and showcases the car in its entirety. A scene unfolds in an underground garage, with a silver-colored sedan approaching slowly. The sedan features black front covers, headlights, and an exquisite exterior design. Close-up shots accentuate the front, sides, and rear of the vehicle. As a final touch, smoke billows from the exhaust.

Figure 12: Example DeVAn data annotated from YouTube video 0Cmr0ISWU44.



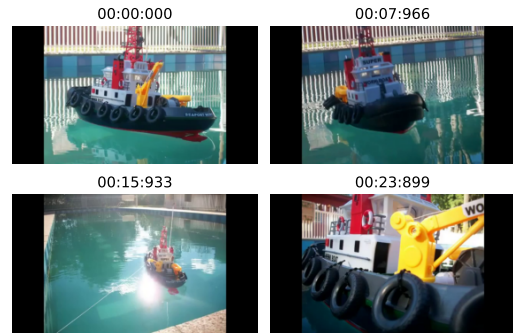
ASR: None.  
 Caption: People are dancing at a nightclub.  
 Summary: People are dancing and reveling at a nightclub. Women are dancing on stage, while people below either wave their hands to the music or dance along. The stage is illuminated with flashing lights, with a DJ spinning records. Everyone joyfully moves their bodies in rhythm with the music. The stage lights change colors. The poster of the last person appears.

Figure 13: Example DeVAn data annotated from YouTube video 6CLbKXgWivU.



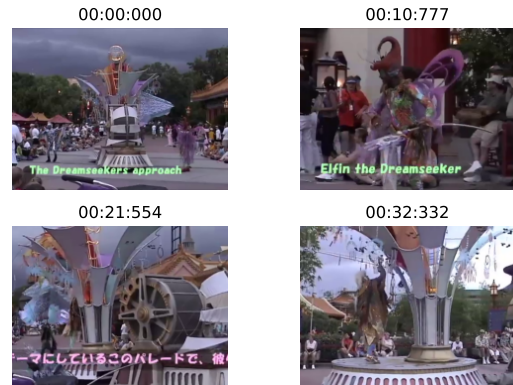
ASR: None.  
 Caption: Restarting a Samsung phone.  
 Summary: In this video, a hand is shown holding a Samsung cellphone. It takes a while to show the home screen. After several minutes, the phone finally displays the home screen wallpaper.

Figure 14: Example DeVAn data annotated from YouTube video 38jPK8IRIb0.



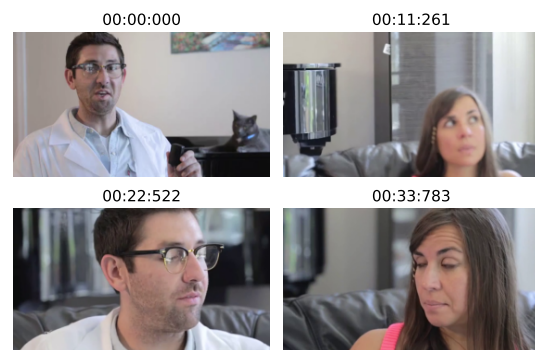
ASR: What amplified the protests? the F.C.C. and the Senate The remains of the ci...  
 Caption: Discover the Seaport Workboat: Remote-controlled sailing fun in the pool!  
 Summary: The video showcases images of a Seaport Workboat, a remote control boat, as it sails in a pool. The boat's design and features are on display, demonstrating its ability to navigate through water with ease. The visuals provide a clear insight into the boat's appearance and performance, offering viewers a glimpse of its functionality.

Figure 15: Example DeVAn data annotated from YouTube video cb\_HvxX80sE.



ASR: None.  
 Caption: A float parade is held at an amusement park.  
 Summary: This is a video of a float parade in an amusement park. It starts with a festooned vehicle with some actors called "Dream Seekers" next to it. We can see an actress wearing a red tiara with purple feathers behind her. The float looks like a fountain, and then all kinds of floats appear. The second float looks like a giant clock.

Figure 16: Example DeVAn data annotated from YouTube video HVmx56OSjDE.



ASR: You get used to the rhythm. See? It's kinda nice. May I present to you... th...  
 Caption: Innovative cat bell offers a solution for feline needs, but training manual not included.  
 Summary: In this video, the speaker introduces the "cat bell" as a solution to address a cat's needs such as hunger, thirst, or attention. The idea is to train the cat to hit the bell, but the product does not include training instructions. The speaker suggests that interested buyers can search for training methods online, but the product being sold is essentially just a bell.

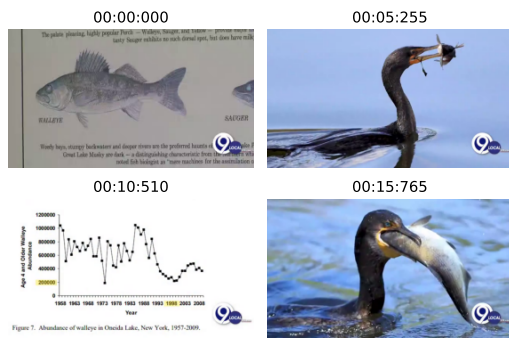
Figure 17: Example DeVAn data annotated from YouTube video Hvo6APy2f5A.





ASR: Under the radar sports media the hottest sports videos in the nation So  
Caption: Athletes are competing on the field.  
Summary: In the video, athletes are fighting on the field. The video begins with the logo of a sports report and scores. Many wonderful moments are shown in the video.

Figure 18: Example DeVAN data annotated from YouTube video LCmfVX2S8zs.



ASR: Walleye population hasn't always been this abundant. Cormorants, a fish-eati...  
Caption: Walleye's population decline in Oneida Lake which is caused by cormorants, leading to a decrease to around 200,000 adult Walleye about 20 years ago.  
Summary: The population of Walleye fish in Oneida Lake faced a threat from cormorants, a bird species known for preying on fish. Approximately 20 years ago, there were only around 200,000 adult Walleye in the lake. During the late 1980s and early 1990s, the fishery suffered a significant decline due to the impact of cormorants. However, there was uncertainty about whether the Walleye population would recover to reach one million.

Figure 19: Example DeVAN data annotated from YouTube video MhZRLhwcsJg.



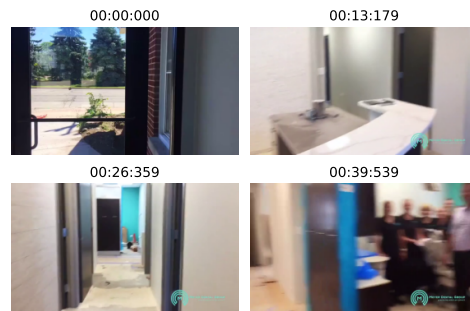
ASR: This is your Shipmate Action Figure Therapy's Drunk Sailor, taking a moment ...  
Caption: Navy is so much better than the Army.  
Summary: In this video the host talks about how the Navy is so much better than the Army. He first wished the football team good luck in the coming games. After that he listed a few reasons why Navy is better and Navy team will win the game against Army. He made fun of the Army over their losing 9 games in a row, and he made some sexual jokes on them.

Figure 20: Example DeVAN data annotated from YouTube video pKC23tWTi-k.



BLEU-4: 0.00, ROUGE: 3.23, CIDER: 0.00, BLEURT: 73.79  
Annotated Caption: Random videos of cats doing various activities were the story of the video.  
Generated Caption: First, we see a little brown puppy sitting on a plate on the floor with food in its bowl. Next, we see a black puppy with a white dot on its face sitting on a table with food in its bowl. Then, we see a black and white puppy sitting on a plate on the floor with food in its bowl. The camera then cuts to a brown puppy sitting on a plate on the floor with food in its bowl. In the next frame, we see a black puppy with a white dot on its face sitting on a table with food in its bowl. After that, we see a brown puppy sitting on a table with food in its bowl. The following frame shows a black and white puppy sitting on a plate on the floor with food in its bowl. The camera then cuts to a brown puppy with a white dot on its face sitting on a table with food in its bowl. Finally, we see a little brown puppy sitting on a plate on the floor with food in its bowl. The video ends with a black puppy with a white dot on its face sitting on a table with food in its bowl.

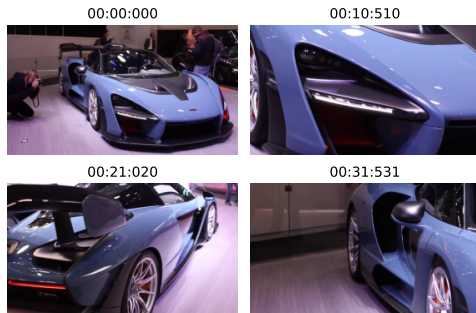
Figure 21: Pathological example caption data generated by Video-LLaMA2-Instruct for video clip take from YouTube video 1iJwajSfE6g.



BLEU-4: 0.00, ROUGE: 9.07, CIDER: 0.00, BLEURT: 72.64  
Annotated Caption: A sneak peek of a new office location with finishes and equipment coming in, ready to go.  
Generated Caption: First, we see a man and a woman standing in a modern dental office with blue walls and a black-framed window. They are looking at a TV and a man is standing next to them. Then, we see a man and woman standing in a modern dental office with blue walls and a black-framed window. They are looking at a blue TV screen and a man is standing next to them. Next, we see a man and woman standing in a modern dental office with blue walls and a black-framed window. They are looking at a blue TV screen, and the man is standing next to them. Then, we see a man and woman standing in a modern dental office with blue walls and a black-framed window. They are looking at a blue TV screen, and the man is standing next to them. Finally, we see a man and woman standing in a modern dental office with blue walls and a black-framed window. They are looking at a blue TV screen, and the man is standing next to them. Overall, the video shows people enjoying themselves in a dental office and looking at TV screens.

Figure 22: Pathological example caption data generated by Video-LLaMA2-Instruct for video clip take from YouTube video 1sRqzVs3etI.





BLEU-4: 0.00, ROUGE: 4.71, CIDER: 0.00, BLEURT: 72.36

Annotated Caption: Unleashing the Power: A Closer Look at McLaren's Latest Road-Going Marvel.

Generated Caption: First, we see a man standing in front of a blue car on display at a trade show. He is wearing a suit and is an older man. There is a person in the background with an umbrella. The man is wearing a black tie and the car is parked in front of a white wall. The car has a silver body and wheels, and the background is mostly white. Next, we see a man in a suit standing next to a silver blue car on display at a trade show. He is older and has a black tie. In the background, there are people with bags, and there is a white and yellow umbrella. The car has a red body and wheels, and there is a silver rim around the tires. The man is wearing a black suit, and the car is parked in front of a white wall. The background is mostly white. Then, we see a man in a suit standing next to a silver blue car at a trade show. He is an older man and has a black tie. In the background, there are people with bags, and there is a white and yellow umbrella. The car has a red body and wheels, and there is a silver rim around the tires. The man is wearing a black suit, and the car is parked in front of a white wall. The background is mostly white. Finally,

**Figure 23: Pathological example caption data generated by Video-LLaMA2-Instruct for video clip take from YouTube video [\\_6st0uhrcBZU](#).**



BLEU-4: 0.00, ROUGE: 4.46, CIDER: 0.00, BLEURT: 72.90

Annotated Caption: This video showcases a modded version of GTA San Andreas.

Generated Caption: First, a black and white picture of a car on a street is shown. Next, a black and white image of a city at night is displayed, with a car parked on the side of the road and a black and white sign that reads, "stop" in the background. Then, a black and white picture of a car on a street is shown again, followed by a black and white photo of a city at night with a car parked on the side of the road and a black and white sign that reads, "stop" in the background. A man wearing a black and white shirt and black and white shoes walks down a street, followed by a black and white picture of a city at night with a car parked on the side of the road and a black and white sign that reads, "stop" in the background. The video then shows a black and white picture of a city at night with a car parked on the side of the road and a black and white sign that reads, "stop" in the background, followed by a black and white image of a city at night with a car parked on the side of the road

**Figure 24: Pathological example caption data generated by Video-LLaMA2-Instruct for video clip take from YouTube video [eoIJuYUA0UI](#).**