Video Repurposing from User Generated Content: A Large-scale Dataset and Benchmark

Yongliang Wu[1,2*], Wenbo Zhu[2*†], Jiawang Cao[2*], Yi Lu[2,3], Bozheng Li[2,4], Weiheng Chi[2,5], Zihan Qiu[2], Lirian Su[2], Haolin Zheng[2], Jay Wu[2], Xu Yang[1‡]

[1]Southeast University
[2]Opus AI Research
[3]University of Toronto
[4]Brown University
[5]National University of Singapore
{yongliangwu, xuyang_palm}@seu.edu.cn

Abstract

The demand for producing short-form videos for sharing on social media platforms has experienced significant growth in recent times. Despite notable advancements in the fields of video summarization and highlight detection, which can create partially usable short films from raw videos, these approaches are often domain-specific and require an in-depth understanding of real-world video content. To tackle this predicament, we propose Repurpose-10K, an extensive dataset comprising over 10,000 videos with more than 120,000 annotated clips aimed at resolving the video long-to-short task. Recognizing the inherent constraints posed by untrained human annotators, which can result in inaccurate annotations for repurposed videos, we propose a two-stage solution to obtain annotations from real-world user-generated content. Furthermore, we offer a baseline model to address this challenging task by integrating audio, visual, and caption aspects through a cross-modal fusion and alignment framework. We aspire for our work to ignite groundbreaking research in the lesser-explored realms of video repurposing.

Code — https://github.com/yongliang-wu/Repurpose

Introduction

Driven by the rapid growth of social media platforms such as Instagram, TikTok, and YouTube Shorts, short-form videos become the primary medium for sharing daily life and conveying information. Even creators who traditionally focus on long-form content, such as live streams, interviews, and vlogs, now shift towards producing captivating short-form videos for these platforms (Wang et al. 2023b). This shift underscores the importance of identifying the most engaging clips within longer videos while maintaining logical coherence.

This brings us to the critical task of video repurposing (Singh 2004). We introduce the task of long-to-short video repurposing, a sophisticated process that transforms user-generated videos, typically longer than 30 minutes, into a series of engaging clips around 60 seconds each, referred to as repurposed clips. This process involves extracting and highlighting key segments to ensure the overall narrative remains complete and engaging. It includes setting a length limit for the final video and re-editing or reorganizing the retrieved content to form a coherent and engaging clip. The goal is to produce a compelling narrative suitable for direct publication on social media platforms.

The challenge of automatically repurposing videos remains an unresolved issue. Existing methods, such as shot boundary detection (SBD) (Zhu et al. 2023), temporal event localization (TEL) (Geng et al. 2023; Wu et al. 2024), video chapter (Yang et al. 2024), and temporal action detection (TAD) (Yang et al. 2023), serve as auxiliary tools for video repurposing but do not fully address the core problem. Two other related tasks are video highlight detection (Badamdorj et al. 2022) and video summarization (Zhang et al. 2016; Hu et al. 2023). The former involves autonomously identifying the most captivating moments within a raw video, while the latter aims to create a concise synopsis that succinctly summarizes the video content by selecting its most informative and pivotal segments. However, it is important to note that previously proposed datasets and methodologies are often domain-specific and do not necessarily meet the unique requirements of video repurposing, which demands a comprehensive understanding of the content (Sun, Farhadi, and Seitz 2014; Potapov et al. 2014; Panda et al. 2017; Song et al. 2015; Pei et al. 2022; Wang et al. 2022; Wu et al. 2023). The difference between these tasks is shown in Figure 1.

Based on this, we compile a large-scale dataset named Repurpose-10K, which includes more than 10,000 video samples and more than 120,000 annotated clips, making it a formidable benchmark for video repurposing. Previous research (Demartini, Roitero, and Mizzaro 2021) highlights the inherent limitations of traditional reliance on human annotators, such as the introduction of bias and potential inaccuracies from untrained individuals. Given the challenges associated with precise clip annotation in videos, we develop an approach that involves acquiring annotations

[*]These authors contributed equally.
[†]Project leader
[‡]Corresponding author

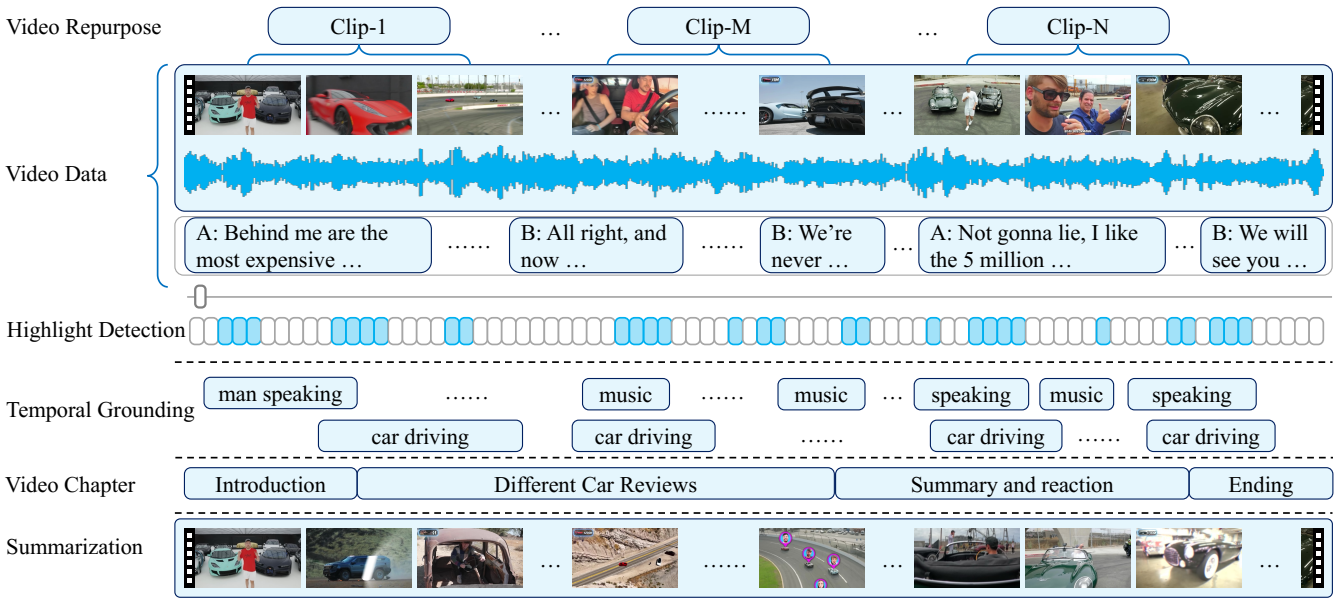

Figure 1: The distinction between Video Repurposing and other similar tasks. From top to bottom: Video Repurposing, Highlights Detection, Temporary Event Localization (Temporal Grounding), Video Chapter, and Video Summarization.

from authentic User Generated Content (UGC) in the real world. The annotation methodology consists of three distinct stages. Initially, repurposing tools based on large language models are utilized to transform long-form videos into shorter clips, establishing preliminary edits with coarse-cut boundaries. Users then express their preferences for each clip by marking them as "like" or "dislike". Clips marked as "like" are selected from the video. Finally, content creators are invited to meticulously refine the start and end timestamps for each clip through manual annotation, resulting in the creation of the final annotated dataset.

To establish robust baseline models to address this challenge, we introduce an end-to-end transformer encoder-decoder architecture that conceptualizes video repurposing as a joint classification and regression problem (Geng et al. 2023; Li et al. 2024d, 2023b; Shi, Dao, and Cai 2024; Shi, Hayat, and Cai 2024). To fully leverage the information from all the modalities within the video, we introduce the Caption Enhancement Module. This module utilizes captions generated by an Automatic Speech Recognition (ASR) model to aid in the alignment of visual and audio modalities. Given that the model may rely on information from a single modality to make judgments, for instance, the audio branch might select segments with engaging music, while the visual branch might choose visually striking segments, relying solely on fused features can lead to insufficient integration of modal features. To address this, we propose the Multi-Modal Align Guider. This module is designed to guide each modality towards selecting the appropriate segments, ensuring consistency and alignment across the different modalities. Through comprehensive experimental analysis and comparison with other state-of-the-art video highlight detection models, we demonstrate the superior performance and practical applicability of our proposed model for this task.

Our contributions are as follows: (1) We introduce Repurpose-10K, a large-scale dataset specifically curated for the video repurposing task. Unlike traditional crowdsourcing methods for annotation, Repurpose-10K compiles annotations from user-generated content (UGC). To the best of our knowledge, this is the first large-scale dataset developed specifically for the challenging video repurposing task. (2) We propose a baseline model tailored for the video repurposing task. This model effectively integrates multimodal information from videos and includes the design of a Multi-Modal Align Guider module, which aims to maximize the agreement and ensure consistency across the different modalities. (3) Through a comprehensive experimental analysis, we validate the effectiveness and practical feasibility of our model. The results demonstrate superior performance and practical applicability in the task of video highlight detection.

Related Work

UGC Video Long to Short

In recent years, video podcasts surge in popularity (Wang et al. 2023b). Platforms like Instagram Reels, TikTok, and YouTube Shorts offer new opportunities to enhance podcast visibility. Successful podcasters often use teasers, brief highlights from their episodes, to attract new listeners. The strategy of repurposing, which involves editing and reshaping existing videos for different objectives or audiences, becomes a prominent method for converting long-form content into short, digestible formats. This technique helps viewers to assimilate the information presented in the videos more efficiently (Truong et al. 2021). The key to repurposing content lies in identifying valuable segments within a video that

meet the criteria for short-form video platforms. While previous video annotations, such as highlights (Sun, Farhadi, and Seitz 2014; Lei, Berg, and Bansal 2021), focus on the visual appeal of images, they often do not require high integrity of user-generated content and are thus unsuitable for direct repurposing. Addressing video editing tasks that demand deeper semantic understanding reveals a significant gap in large-scale, labeled datasets and end-to-end methods.

Related Tasks

Temporal Event Localization. Temporal Event Localization (TEL) (Tian et al. 2018; Tian, Li, and Xu 2020; Lee et al. 2021; Geng et al. 2023; Zhang et al. 2024; Li et al. 2024a; Zhang, Zhang, and Zhou 2024) is a video understanding task that aims to identify and pinpoint the start and end times of events or actions of interest within untrimmed videos. This task proves crucial for areas such as multimedia content analysis, video surveillance, and sports event analysis. The supervised learning paradigms of TEL are categorized into two primary methodologies: two-stage approaches (Zhao et al. 2017; Li et al. 2023a, 2024c) and single-stage frameworks (Buch et al. 2019; Zhang, Wu, and Li 2022; Yan et al. 2023; Li et al. 2024b; Liu, Li, and Yu 2024). Recently, UnLoc (Yan et al. 2023) proposes a unified framework for video localization tasks, consisting of a two-tower CLIP model, with output features fed into a video-text fusion module and feature pyramid. Unlike TEL, which typically focuses on identifying explicit timestamps, video repurposing revolves around identifying complete and engaging segments within lengthy videos, without explicitly marking a specific starting point, as a temporal localization task (Liu et al. 2024).

Video Highlight Detection. Video highlight detection (Sun, Farhadi, and Seitz 2014) aims to automatically identify the most compelling and quintessential segments of a video. While earlier methods consider only visual features (Sun, Farhadi, and Seitz 2014), recent works start to incorporate both audio and visual components to retrieve highlights from the video (Badamdorj et al. 2021). Query-Dependent DETR (Moon et al. 2023) employs a cross-attention transformer encoder to generate more potent multi-modal video representations, deliberately infusing text queries into the feature representation. However, highlight detection particularly emphasizes moments that are exceptionally striking or exhilarating, rather than necessarily providing a coherent summary of the video content. For example, a moment is considered a highlight if at least 70% of its frames match the edited video in the training of YouTube Highlights (Sun, Farhadi, and Seitz 2014). This approach to combining highlights poses challenges in crafting a comprehensive narrative thread, making it difficult for direct publication on short-form social media platforms due to the lack of a cohesive storyline. Additionally, repurposing video highlights often requires post-processing to form coherent segments, introducing further effort.

Video Summarization. Video summarization (Song et al. 2015) aims to extract key information from lengthy video content and generate a concise summary version. This task preserves the most informative and representative parts of the video, allowing users to quickly grasp the main content without watching the entirety of the original video. The key parts within a video could either be a series of key frames or a collection of one or more key segments. To better capture temporal correlations in video content, a variety of sophisticated deep learning models (Zhang et al. 2016; Li et al. 2022) develop to map dependencies across time, using either localized or holistic approaches. Video summarization prioritizes comprehensiveness, often overlooking the engaging aspect of video content. In contrast, video repurposing does not mandate conveying the entire essence of the video.

Dataset	Domain	Sizes	Avg. Length	Modality	TB
AVE	TEL	4,143	10s	AV	✓
LLP	TEL	11,849	10s	AV	✓
ACAV100M	TEL	100M	10s	AV	✗
UnAV-100	TEL	10,790	42.1s	AV	✓
OVP	VS	50	1.5 mins	AV	✗
SumMe	VS	25	2.4 mins	AV	✗
TVSum	VS	50	4.2 mins	AV	✗
YT-Highlights	VH	433	143s	AV	✓
QVHighlights	VH	10,148	150s	AV	✓
MultiSeg	TSLLV	1,000	78 mins	AVC	✓
Repurpose-10K	VR	11,210	32 mins	AVC	✓

Table 1: Comparison with related multi-modal video datasets. A: audio; V: visual; C: caption; TB: temporal boundaries; TEL: Temporal Event Localization; VH: Video Highlights; VS: Video Summarization; TSLLV: Temporal Segmentation of the Long Livestream Videos; VR: Video Repurposing.

Repurpose-10K Dataset

Overview

We propose the Repurpose-10K dataset. By leveraging a repurposing SaaS platform, we collect annotations from real users. For each long video, the annotations include timestamps marking the start and end of multiple clips, with each clip encapsulating a self-contained sub-topic, such as a compilation of actions or a dialogue exchange.

The comparison of our dataset with related datasets is presented in Table 1. We argue that the quantity of videos plays a critical role in User Generated Content (UGC) video understanding tasks. When comparing video summarization datasets, their quantities are relatively small, which may not sufficiently represent the diversity and complexity needed for repurposing challenges. Furthermore, addressing the challenges posed by real-world repurposing tasks, which involve the processing of untrimmed videos, we provide an extensive collection of long-format videos, averaging 32 minutes each. Additionally, given the prevalence of captions, we include them as part of the dataset to facilitate a comprehensive understanding of the multi-modal content.

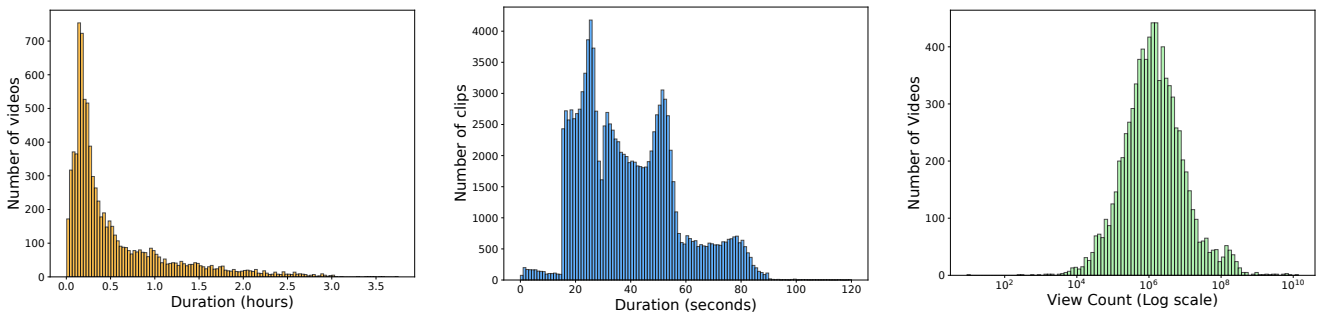

Figure 2: Histogram of Repurpose-10K videos duration. From left to right: collection videos duration, repurpose clips duration (y-axis: number of videos), log-scale distribution of collection videos view counts.

Data Collection

To gain a deeper insight into user preferences regarding UGC, we conduct an extensive collection of real-world user interaction data. Initially, leveraging the topic segmentation capabilities of the OPUS platform[1], a versatile video repurposing tool, users divide YouTube videos into multiple short sections related to various topics, creating a preliminary edit with coarse-cut boundaries. Then, these users personally select their preferred clips from the entire collection of these sections by clicking the "like" or "post" button. We crawl 8,398 YouTube videos posted by 114,883 content creators. In these videos, there are at least one or more repurposed clips that users mark as "like" or "post". To ensure the accuracy of the timestamps, we selectively choose clips that have undergone user re-editing for the start and end timestamp refinement, resulting in 120,925 finely annotated video clips. Strictly limited to clips that are manually annotated by users and subsequently endorsed as their preferred choices, our dataset comprises a curated selection of user-preferred content.

Data Analysis

The Repurpose-10K dataset includes an extensive collection of 8,398 videos. The total duration of these videos amounts to 4,539.94 hours, with an average length of 0.54 hours per video. To manage the annotation workload, videos exceeding 30 minutes are divided into multiple clips, resulting in a total of 11,210 videos. We also compile view counts using yt-dlp[2], with an average of 18.9 million views per video, and a 90% confidence interval ranging from 62,900 to 38.6 million views. The distribution, illustrated in Figure 2, suggests that these long videos have inherent dissemination value, making them suitable for repurposing tasks.

Human Evaluation

To assess the annotation accuracy in video repurposing within UGC content, we conduct a comparative user study on various repurposing outcomes. We randomly select a sample of 110 videos, yielding a total of 462 labeled results with an average duration of 58.66 seconds. Additionally, we extract 462 clips with random timestamps from the original videos, each ranging from 30 to 90 seconds. Two professional short video creators are also employed to identify the clips within the videos that hold the highest potential for repurposing, producing 3 short clips for each video. These clips are then randomized, and 10 participants are invited to partake in a survey to evaluate the clips. Participants cast votes and rate the clips from different perspectives (engagement level and completeness level, each ranging from 1 to 5). The conclusive scores are presented in Table 2. The results reveal a close alignment between user-generated labels and those provided by professional editors. Notably, the user-generated labels outperform randomly assigned clips by a significant margin.

Group	Duration(s)			Avg.Score		
	Min	Max	Avg	E	C	Total
Random	33.39	88.90	59.55	1.623	1.508	3.131
Professional Editor	31.93	89.25	53.27	3.938	3.738	7.676
Repurpose-10K	20.42	154.41	58.66	3.868	3.418	7.286

Table 2: Human evaluation results from different annotation. (E: Engaging level, C: Completeness level)

Method

Preliminary

We approach video repurposing by jointly learning classification and regression tasks, taking video elements (segments, audio, captions) as input and output the time range for selected clips. The classification task determines if a segment should be selected, while the regression task pinpoints the exact temporal offsets for the clip's start and end. For an input video sequence with visual and audio tracks, we segment it into visual and audio pairs, represented as $\{V_t, A_t\}_{t=1}^{T}$, where T is the number of segments. In cases involving human speech, we use whisperX (Bain et al. 2023) to convert audio segments into time-stamped captions, C_t, enhancing our segment pairs to $\{V_t, A_t, C_t\}_{t=1}^{T}$. For each video, we select a set of clips, compiling temporal intervals denoted by $R = \{(t_{s,n}, t_{e,n})\}_{n=1}^{N}$, where N is the number of curated excerpts.

[1]https://www.opus.pro/

[2]https://github.com/yt-dlp/yt-dlp

In the classification task, segments within selected clips are labeled 1, and those outside are labeled 0, creating a binary classification framework. In the regression task, we simplify by calculating the temporal offset from the segment's timestamps to the start and end of the clip. This is expressed as $Y = \{d_{s,t}, d_{e,t}, c_t\}_{t=1}^{T}$, where $d_{s,t}$ and $d_{e,t}$ are the temporal distances to the start and end of the clip, and c_t is the binary classification label (0 or 1). T represents the total number of labeled segments in the video.

Framework

Feature Extractor. As shown in Figure 3, we begin by extracting three modalities of data using dedicated pre-trained models. This results in $F_V = \{f_t^v\}_{t=1}^{T}$ for the visual modality, $F_A = \{f_t^a\}_{t=1}^{T}$ for the audio modality, and $F_C = \{f_t^c\}_{t=1}^{T}$ for the caption modality. We then use three distinct MLP layers to map these features to a unified dimension d.

Caption Enhancement Encoder. Initially, we employ three separate self-attention modules, each with N_s layers, to capture the temporal information within each modality. We treat caption features as an auxiliary component, noting the positive correlation between audio and caption features. Captions often explicitly describe spoken language and visually referenced objects, accompanied by corresponding audio cues. For example, if a car appears in the scene, the caption might describe it while the audio captures the engine's sound. We leverage caption features to align and discern visual and audio attributes. To achieve this, we introduce two cross-attention modules with N_c layers to facilitate information exchange and synergy among the modalities. Finally, we use N_f-layer cross-attention modules to promote dynamic interaction between visual and audio modalities. The fused features are then concatenated, mapped to lower dimensions, and fed into the decoder module.

Multi-Modal Alignment Guider. The use of multi-modal data holds great potential for enhancing learning by integrating features from various domains, thereby improving model performance. However, the fusion process can sometimes be ineffective in bridging the gaps between different modalities, leading to suboptimal results. Direct fusion for predictions can introduce instability during the model's learning process. To address this issue, we introduce the Multi-Modal Alignment Guider. This module is designed to guide each modality in selecting appropriate segments, ensuring consistency and alignment across the different modalities. For each branch, we implement a classification head composed of a three-layer Multilayer Perceptron (MLP) with a sigmoid activation function to predict outcomes for the current video sequence. We assume that the uni-modal branches select segments based on the specific information available within their respective modalities. For instance, the audio branch may choose segments with compelling music, while the visual branch selects visually striking segments. In contrast, the multi-modal branch makes predictions based on the fused features of both modalities. During this process, we use focal loss (Lin et al. 2017), denoted as $\mathcal{L}_{focal}$, to predict each segment individually. The classification loss can be formulated as follows:

$$\mathcal{L}_{\text{focal}}^{uni} = -(\sum_{i=1}^{T} c_i \log y_i^v + \sum_{i=1}^{T} c_i \log y_i^a), \tag{1}$$

$$\mathcal{L}_{\text{focal}}^{mul} = -\sum_{i=1}^{T} c_i \log y_i^m, \tag{2}$$

where y_i^a, y_i^v, and y_i^m represent the predicted outcomes for audio, visual, and multi-modal fusion, respectively.

To further promote inter-modality fusion and ensure stability in the joint optimization process, we introduce an alignment loss to align the predictions from multiple branches towards consistency. This loss adopts the Kullback-Leibler divergence loss (KLDivLoss) (Wang et al. 2023a) to align the probability distributions of uni-modal predictions with the multi-modal fusion predictions.

$$\mathcal{L}_{\text{KL}}^{v \to m} = \sum_{i=1}^{T} y_i^v \log \left(\frac{y_i^v}{y_i^m} \right), \tag{3}$$

$$\mathcal{L}_{\text{KL}}^{a \to m} = \sum_{i=1}^{T} y_i^a \log \left(\frac{y_i^a}{y_i^m} \right). \tag{4}$$

Training Objective. To more accurately identify the clips that should be selected, we employ a customized 1-D IoU-based loss (Rezatofighi et al. 2019), denoted as $\mathcal{L}_{iou}$, which refines our temporal predictions. This approach is similar to a variant of the Temporal Event Localization framework (Zhang, Wu, and Li 2022) and is related to the task of audio-visual event detection (Geng et al. 2023). To achieve this goal, we design a regression head with a structure similar to that of the classification head, but incorporating a ReLU activation layer at its final stage.

Formally, the composite objective function for a given video is represented as follows:

$$\lambda_1 \mathcal{L}_{\text{focal}}^{uni} + \lambda_2 \mathcal{L}_{\text{focal}}^{mul} + \lambda_3 (\mathcal{L}_{\text{KL}}^{v \to m} + \mathcal{L}_{\text{KL}}^{a \to m}) + \lambda_4 \mathcal{L}_{iou}, \tag{5}$$

where the hyper-parameters λ_{1-4} are equilibrium constants.

Experiments

Setting

Feature Extraction. Each video is sampled at one frame per second. Each frame is then processed using the CLIP ViT-B/32 model (Radford et al. 2021) to derive 512-dimensional visual features. In parallel, each one-second audio segment is input into a pre-trained PANN model (Kong et al. 2020) trained on AudioSet (Gemmeke et al. 2017), yielding 2048-dimensional audio features. For caption features, we use the WhisperX (Bain et al. 2023) model for Automatic Speech Recognition (ASR) with timestamps, followed by the all-MiniLM-L6-v2 model from Sentence-Transformers (Reimers and Gurevych 2019) to extract 384-dimensional sentence features. We associate transcript sentences with corresponding video segments to address the temporal misalignment between transcripts and video frames. If a sentence overlaps multiple segments, we duplicate it to create segment-transcript pairs. For empty captions

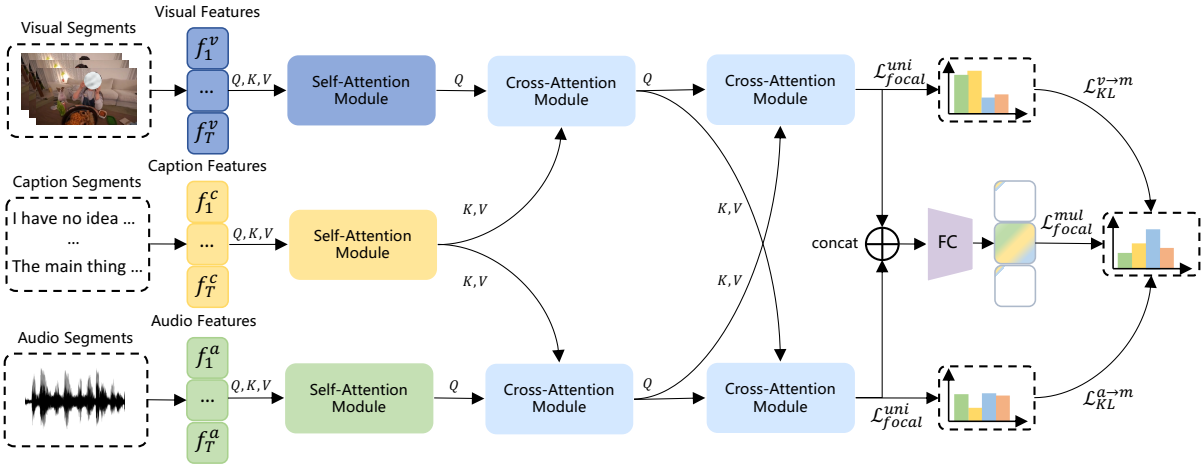

Figure 3: The overall architecture of our proposed baseline model consists of two main components: the Caption Enhancement Encoder and the Multi-Modal Align Guider. Q: Query. K: Key. V: Value. FC: Fully Connected Layer.

Method	Modality	0.5	0.6	0.7	0.8	0.9	Avg.
SL-Module	V	21.89	14.11	8.48	3.91	0.97	9.87
Moment-DETR	V	23.29	14.70	8.25	3.45	0.98	10.13
QD-DETR	V	23.02	14.47	8.01	3.54	0.99	10.01
UMT	V&A	21.46	13.06	7.52	3.49	1.04	9.31
QD-DETR	V&A	23.30	14.64	7.98	3.87	1.01	10.16
Ours	V&A&C	**24.94**	**16.35**	**10.14**	**4.86**	**1.57**	**11.57**

Table 3: Comparison with State-of-the-Art Temporal Grounding Models: SL-Module (Xu et al. 2021), Moment-DETR (Lei, Berg, and Bansal 2021), UMT (Liu et al. 2022), and QD-DETR (Moon et al. 2023). The numbers in the first row of the table represent the thresholds. To denote the best results, we use bold text, and for the second-best results, we underline them.

Modality	UF	AL	0.5	0.6	0.7	0.8	0.9	Avg.
A	-	-	22.20	13.99	8.13	3.34	0.63	9.66
V	-	-	21.91	13.87	8.08	3.14	0.60	9.52
C	-	-	22.30	14.02	7.98	3.06	0.57	9.59
A&V	-	-	23.28	14.95	8.69	3.71	0.90	10.31
A&V	✓	-	23.90	15.54	9.10	3.97	1.02	10.71
A&V	✓	✓	24.13	15.71	9.47	4.30	1.21	10.96
A&V&C	-	-	23.74	15.47	9.13	4.26	1.19	10.75
A&V&C	✓	-	24.51	15.97	9.86	4.63	1.45	11.28
A&V&C	✓	✓	**24.94**	**16.35**	**10.14**	**4.86**	**1.57**	**11.57**

Table 4: Ablation study of the impact of various modalities and the primary results on the Repurposing-10K test set. A: Audio modality. V: Visual modality. A&V: Audio modality combined with visual modality. A&V&C: Audio modality and visual modality integrated with caption modality. UF: Uni-modal focal loss. AL: Alignment loss. The numbers in the first row of the table represent the thresholds.

within a time window, we use a special [empty] token as a placeholder. Through these steps, visual, audio, and caption features align at the segment level.

Implementation Details. We partition the dataset into train/val/test splits at a ratio of 8/1/1. The embedding dimension of the model is set to $d = 512$, and the number of layers N_s, N_c, and N_f is set to 3. We utilize the Adam optimizer with a learning rate of 1e-4 for 100 epochs, which is adjusted using cosine learning rate decay, while the first 5 epochs employ linear warm-up to facilitate stable learning. The hyper-parameters λ_{1-4} are set to 0.1, 0.3, 0.1, and 0.7, respectively. All experiments are conducted on two A100 GPUs within the PyTorch framework.

Evaluation Metrics. We initially apply a threshold of 0.5 to filter out segments with low confidence. Subsequently, we employ soft-nms (Bodla et al. 2017) to obtain the top-k clips. The value of k is dynamically determined based on statistical results from the training set, where we set it as 3 clips every ten minutes on average. We present the results using tIoU with thresholds of [0.5, 0.9, 0.1].

Results and Analysis

Comparison With Baselines. For comparison, we select several representative temporal grounding models as baselines (Xu et al. 2021; Lei, Berg, and Bansal 2021; Liu et al. 2022; Moon et al. 2023). The results, presented in Table 3, show that our model significantly outperforms these baselines across various IoU thresholds, consistently achieving top scores and demonstrating superior performance.

Modality Fusion Ablations. As shown in Table 4, in uni-modal scenarios, the model using caption features achieves the best average performance. Notably, it outperforms the visual-based model at lower thresholds, highlighting the effectiveness of caption features when a single modality is considered. Additionally, performance consistently improves when both audio and visual features are utilized to-

Figure 4: Two visual examples of video repurposing results. Blue: Ground Truth. Yellow: Predictions.

gether. This improvement is further enhanced with the integration of the Caption Enhancement Module, underscoring the importance of multimodal learning for the video repurposing task.

Modality Alignment Ablations. As shown in Table 4, incorporating auxiliary multimodal constraints enhances the model's robustness. Allowing uni-modal branches to make independent predictions can improve performance to some extent. However, this improvement is significantly amplified when their distributions are explicitly aligned with multimodal predictions. This suggests that the alignment loss not only indirectly reinforces information from individual modalities but also directly guides the multimodal branches to learn more effectively.

Modality Fusion Layers Ablations. We experiment with varying the number of layers in distinct modules, denoted as N_s, N_c, and N_f. As shown in Table 5, a balanced configuration of 3/3/3 layers produces the best results. Interestingly, highly imbalanced configurations, such as 7/1/1, perform worse than models using a single modality. This highlights the importance of not only considering interactions within individual modalities but also fostering effective cross-modal interactions. Such an approach is essential for fully leveraging the potential of features across different modalities. Additionally, our findings indicate that shallow layers for attention mechanisms are insufficient for the model to establish meaningful connections between various modalities.

Qualitative Results. As illustrated in Figure 4, there is a notable alignment between our model's predictions and the actual ground truth. In the first scenario, the ground truth annotation provides a straightforward and concise description, leading to the decision to explore these "strange places." In contrast, our model's prediction starts with an engaging B-roll news clip and ends with the main character's declaration, "Let the adventure begin." In the second scenario, while the annotation offers a detailed introduction to the product of the video creator, our predictions more effectively highlight the video's concluding section. Both selected clips begin with intriguing content and end with fitting remarks, demonstrating their ability to engage and provide a complete narrative.

$N_s/N_c/N_f$	0.5	0.6	0.7	0.8	0.9	Avg.
1 / 4 / 4	23.37	15.01	8.62	4.10	1.29	10.48
3 / 3 / 3	**24.94**	**16.35**	**10.14**	**4.86**	**1.57**	**11.57**
5 / 2 / 2	22.51	13.60	7.23	3.25	0.80	9.48
7 / 1 / 1	19.88	12.41	6.77	3.08	0.73	8.57

Table 5: Ablation study to explore the selection of various numbers of transformer modality fusion layers. N_s: Number of self-attention layers. N_c: Cross-attention layer for caption enhancement. N_f: Fusion layer for integrating audio and visual features. The numbers in the first row of the table represent the thresholds.

Conclusion

In this paper, we curate the Repurpose-10K dataset for the challenging task of video repurposing. The dataset comprises over 10,000 videos and more than 120,000 clips, annotated with User Generated Content (UGC) in the real world. Furthermore, we present an end-to-end model designed to address the video repurposing challenge by jointly learning classification and regression tasks. Through extensive experiments and ablation studies, we demonstrate the effectiveness of the proposed model.

Acknowledgments

This work is supported by the National Science Foundation of China (62206048), the Natural Science Foundation of Jiangsu Province (BK20220819), and the Fundamental Research Funds for the Central Universities (2242024k30035). The Big Data Computing Center of Southeast University also supports this research work.

References

Badamdorj, T.; Rochan, M.; Wang, Y.; and Cheng, L. 2021. Joint visual and audio learning for video highlight detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8127–8137.

Badamdorj, T.; Rochan, M.; Wang, Y.; and Cheng, L. 2022. Contrastive learning for unsupervised video highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14042–14052.

Bain, M.; Huh, J.; Han, T.; and Zisserman, A. 2023. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. *INTERSPEECH 2023*.

Bodla, N.; Singh, B.; Chellappa, R.; and Davis, L. S. 2017. Soft-NMS–improving object detection with one line of code. In *Proceedings of the IEEE international conference on computer vision*, 5561–5569.

Buch, S.; Escorcia, V.; Ghanem, B.; Fei-Fei, L.; and Niebles, J. C. 2019. End-to-end, single-stream temporal action detection in untrimmed videos. In *Procedings of the British Machine Vision Conference 2017*. British Machine Vision Association.

Demartini, G.; Roitero, K.; and Mizzaro, S. 2021. Managing Bias in Human-Annotated Data: Moving Beyond Bias Removal. *arXiv preprint arXiv:2110.13504*.

Gemmeke, J. F.; Ellis, D. P.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R. C.; Plakal, M.; and Ritter, M. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 776–780. IEEE.

Geng, T.; Wang, T.; Duan, J.; Cong, R.; and Zheng, F. 2023. Dense-Localizing Audio-Visual Events in Untrimmed Videos: A Large-Scale Benchmark and Baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22942–22951.

Hu, Z.; Wang, Z.; Song, Z.; and Hong, R. 2023. Dual Video Summarization: From Frames to Captions. In *IJCAI*, 846–854.

Kong, Q.; Cao, Y.; Iqbal, T.; Wang, Y.; Wang, W.; and Plumbley, M. D. 2020. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28: 2880–2894.

Lee, S.; Chung, J.; Yu, Y.; Kim, G.; Breuel, T.; Chechik, G.; and Song, Y. 2021. ACAV100M: Automatic Curation of Large-Scale Datasets for Audio-Visual Video Representation Learning. In *ICCV*.

Lei, J.; Berg, T. L.; and Bansal, M. 2021. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34: 11846–11858.

Li, B.; Liu, M.; Wang, G.; and Yu, Y. 2024a. Frame Order Matters: A Temporal Sequence-Aware Model for Few-Shot Action Recognition. *arXiv preprint arXiv:2408.12475*.

Li, H.; Ke, Q.; Gong, M.; and Zhang, R. 2022. Video joint modelling based on hierarchical transformer for co-summarization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 3904–3917.

Li, P.; Zhang, Y.; Yuan, L.; Zhao, J.; Xu, X.; and Zhang, X. 2023a. Adversarial attacks on video object segmentation with hard region discovery. *IEEE Transactions on Circuits and Systems for Video Technology*.

Li, R.; Zhang, C.; Wang, Z.; Shen, C.; and Lin, G. 2024b. Self-Supervised 3D Scene Flow Estimation and Motion Prediction using Local Rigidity Prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Li, X.; Wang, T.; Zhao, J.; Mao, S.; Wang, J.; Zheng, F.; Peng, X.; and Li, X. 2024c. Two in One Go: Single-stage Emotion Recognition with Decoupled Subject-context Transformer. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 9340–9349.

Li, Y.; Wang, X.; Xiao, J.; Ji, W.; and Chua, T.-S. 2023b. Transformer-empowered invariant grounding for video question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Li, Y.; Zhao, N.; Xiao, J.; Feng, C.; Wang, X.; and Chua, T.-s. 2024d. LASO: Language-guided Affordance Segmentation on 3D Object. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14251–14260.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Liu, M.; Li, B.; and Yu, Y. 2024. OmniCLIP: Adapting CLIP for Video Recognition with Spatial-Temporal Omni-Scale Feature Learning. *arXiv preprint arXiv:2408.06158*.

Liu, M.; Wu, F.; Li, B.; Lu, Z.; Yu, Y.; and Li, X. 2024. Envisioning Class Entity Reasoning by Large Language Models for Few-shot Learning. *arXiv preprint arXiv:2408.12469*.

Liu, Y.; Li, S.; Wu, Y.; Chen, C.-W.; Shan, Y.; and Qie, X. 2022. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3042–3051.

Moon, W.; Hyun, S.; Park, S.; Park, D.; and Heo, J.-P. 2023. Query-dependent video representation for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23023–23033.

Panda, R.; Das, A.; Wu, Z.; Ernst, J.; and Roy-Chowdhury, A. K. 2017. Weakly supervised summarization of web videos. In *Proceedings of the IEEE international conference on computer vision*, 3657–3666.

Pei, S.; Xu, S.; Yuan, Y.; Feng, J.; Shen, X.; and Jin, X. 2022. Global Prototype Encoding for Incremental Video Highlights Detection. *arXiv preprint arXiv:2209.05166*.

Potapov, D.; Douze, M.; Harchaoui, Z.; and Schmid, C. 2014. Category-specific video summarization. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, 540–555. Springer.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Reimers, N.; and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; and Savarese, S. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 658–666.

Shi, H.; Dao, S.; and Cai, J. 2024. LLMFormer: Large Language Model for Open-Vocabulary Semantic Segmentation. *International Journal of Computer Vision*.

Shi, H.; Hayat, M.; and Cai, J. 2024. Unified Open-Vocabulary Dense Visual Prediction. *IEEE Transactions on Multimedia*.

Singh, G. 2004. Guest editor's introduction: Content repurposing. *IEEE MultiMedia*, 11(01): 20–21.

Song, Y.; Vallmitjana, J.; Stent, A.; and Jaimes, A. 2015. Tvsum: Summarizing web videos using titles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5179–5187.

Sun, M.; Farhadi, A.; and Seitz, S. 2014. Ranking domain-specific highlights by analyzing edited videos. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, 787–802. Springer.

Tian, Y.; Li, D.; and Xu, C. 2020. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, 436–454. Springer.

Tian, Y.; Shi, J.; Li, B.; Duan, Z.; and Xu, C. 2018. Audio-visual event localization in unconstrained videos. In *Proceedings of the European conference on computer vision (ECCV)*, 247–263.

Truong, A.; Chi, P.; Salesin, D.; Essa, I.; and Agrawala, M. 2021. Automatic generation of two-level hierarchical tutorials from instructional makeup videos. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16.

Wang, D.; Hu, Z.; Zhou, Y.; Hong, R.; and Wang, M. 2022. A text-guided generation and refinement model for image captioning. *IEEE Transactions on Multimedia*, 25: 2966–2977.

Wang, J.; Wang, C.; Wang, X.; Huang, J.; and Jin, L. 2023a. CocaCLIP: Exploring Distillation of Fully-Connected Knowledge Interaction Graph for Lightweight Text-Image Retrieval. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, 71–80.

Wang, S.; Ning, Z.; Truong, A.; Dontcheva, M.; Li, D.; and Chilton, L. B. 2023b. PodReels: Human-AI Co-Creation of Video Podcast Teasers. *arXiv preprint arXiv:2311.05867*.

Wu, M.; Cao, M.; Bai, Y.; Zeng, Z.; Chen, C.; Nie, L.; and Zhang, M. 2023. An Empirical Study of Frame Selection for Text-to-Video Retrieval. *arXiv preprint arXiv:2311.00298*.

Wu, Y.; Hu, X.; Sun, Y.; Zhou, Y.; Zhu, W.; Rao, F.; Schiele, B.; and Yang, X. 2024. Number it: Temporal Grounding Videos like Flipping Manga. *arXiv preprint arXiv:2411.10332*.

Xu, M.; Wang, H.; Ni, B.; Zhu, R.; Sun, Z.; and Wang, C. 2021. Cross-category video highlight detection via set-based learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7970–7979.

Yan, S.; Xiong, X.; Nagrani, A.; Arnab, A.; Wang, Z.; Ge, W.; Ross, D.; and Schmid, C. 2023. Unloc: A unified framework for video localization tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13623–13633.

Yang, A.; Nagrani, A.; Laptev, I.; Sivic, J.; and Schmid, C. 2024. Vidchapters-7m: Video chapters at scale. *Advances in Neural Information Processing Systems*, 36.

Yang, M.; Chen, G.; Zheng, Y.-D.; Lu, T.; and Wang, L. 2023. Basictad: an astounding rgb-only baseline for temporal action detection. *Computer Vision and Image Understanding*, 232: 103692.

Zhang, C.; Zhang, L.; Wu, J.; Zhou, D.; and He, Y. 2024. Causal prompting: Debiasing large language model prompting based on front-door adjustment. *arXiv preprint arXiv:2403.02738*.

Zhang, C.; Zhang, L.; and Zhou, D. 2024. Causal Walk: Debiasing Multi-Hop Fact Verification with Front-Door Adjustment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 19533–19541.

Zhang, C.-L.; Wu, J.; and Li, Y. 2022. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, 492–510. Springer.

Zhang, K.; Chao, W.-L.; Sha, F.; and Grauman, K. 2016. Video summarization with long short-term memory. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, 766–782. Springer.

Zhao, Y.; Xiong, Y.; Wang, L.; Wu, Z.; Tang, X.; and Lin, D. 2017. Temporal action detection with structured segment networks. In *Proceedings of the IEEE international conference on computer vision*, 2914–2923.

Zhu, W.; Huang, Y.; Xie, X.; Liu, W.; Deng, J.; Zhang, D.; Wang, Z.; and Liu, J. 2023. AutoShot: A Short Video Dataset and State-of-the-Art Shot Boundary Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2237–2246.