VidSum-Reason: A Dataset for Text-Queryable Video Summarization with Reasoning

Overview

VidSum-Reason is a user-guided video summarization dataset designed to evaluate and benchmark models that perform zero-shot, training-free, and text-queryable video summarization. Each video is paired with user queries that demand reasoning, multi-step understanding, or long-tailed concept recognition.

Key Features

• Text-guided Summarization: Each video is annotated with one or more textual queries reflecting user intent.

• Reasoning-based Queries: Queries require deeper understanding such as cause-effect, temporal reasoning, or multi-event tracking.

• Multi-annotator Vetted Labels: Labels and query relevance scores are vetted by multiple annotators to ensure consistency and reliability.

• Class Diversity: Query prompts are grouped into diverse semantic classes (e.g., Actions, Reactions, Outcomes), allowing category-based evaluation.

• Evaluation-Friendly: Frame-level ground-truth annotations are provided for evaluation of both general and query-focused summarization methods.

Dataset Structure

• Videos: Natural, diverse videos sourced from public domains, varying in length (2–6 minutes).

• Queries: Natural-language questions or prompts targeting specific moments or patterns in the video.

• Labels: fine-grained rubric scale (1–5) per frame based on query relevance.

• Query classes: (i) Reasoning, (ii) Reasoning with General Knowledge, (iii) Standard, and (vi) Standard with Special Attribute

Statistics

• # Videos: 9

• # Queries:20

• # Query Classes : 4

• # Annotators: 1 annotator per video-query pair

Annotation Protocol

Each video-query pair is annotated by one human annotator (authur) who:

1. View the video and query.

2. Segment the video into non-overlapping fragment, each 2 seconds in duration. Each segment was rated on a scale from 1 (not relevant) to 5 (very relevant) based on how well it matched the content of a specific user query.

If a segment was found not relevant to the query, then, it was evaluated based on its overall importance to the main theme, message, or core event of the video, using the same 1–5 scale.

Applications

• Query-Focused Video Summarization (QFVS)

• Zero-Shot Video Understanding

• Video-Language Model Benchmarking

• Scene Reasoning Evaluation

Example Categories

• Reasoning : "Exclude scenes with negative emotions"

• Reasoning with General Knowledge : "Focus on scenes where the score changes"

• Standard : "Show scenes with explosions"

• Standard with Special Attribute : "Focus on cars with bright colors"

VidSum-Reason : Strcture

The VidSum-Reason dataset is organized to support robust evaluation of query-focused, frame-level video summarization. It includes the following components:

1. Videos

• A curated collection of videos spanning various categories and difficulty levels, with an emphasis on diverse events, multi-step reasoning, and long-tailed concepts.

• Each video is queryable via multiple textual prompts to allow for varied summarization perspectives.

2. Ground Truth Annotations

• Each video-query pair is annotated with fine-grained frame-level scores.

3. Evaluation Splits

• The dataset is partitioned into five randomly generated splits, enabling robust and reproducible evaluation.

• Additionally, we provide a designated training/testing split to standardize comparison across methods.

4. Mapping File

• A central mapping file is included, which links each video-query pair to its corresponding:

• Video file

• Textual query

• Query class (e.g., entity-focused, temporal reasoning)

• Ground-truth annotation file

VidSum-Reason : Evaluation Protocol

To evaluate summarization performance on VidSum-Reason, we adopt a fragment-based comparison scheme using both ground-truth and predicted frame-level relevance scores. The procedure is as follows:

1. Fragmentation:

Each video is uniformly segmented into non-overlapping fragments, where each fragment covers d% of the total video duration (e.g., d = 2%).

2. Scoring Fragments:

•For each fragment, its relevance score is computed as the mean of the frame-level scores (either from ground truth or model predictions) it contains.

•This converts dense frame-level importance into a more interpretable, segment-level representation.

3. Summary Generation (Knapsack Selection):

•To form the summary under a maximum budget constraint of M% (e.g., M = 15%), we apply the Knapsack algorithm to select a subset of

fragments.

•The objective is to maximize total fragment importance while respecting the total duration constraint.

4. Evaluation Metrics:

•The selected predicted summary is compared against the ground-truth summary (generated from annotated scores using the same budget).

•Standard evaluation metrics such as F1-score, Precision, and Recall are computed based on overlap with the ground-truth summary.

After generating both the ground-truth and machine-predicted summaries, we compute the final accuracy for each video-query pair using the F1-score, which balances Precision and Recall to assess the quality of the predicted summary against the ground truth.

This protocol allows consistent and fair benchmarking of both supervised and zero-shot summarization methods across variable-length videos and diverse query types.