NEPTUNE: THE LONG ORBIT TO BENCHMARKING LONG VIDEO UNDERSTANDING

Arsha Nagrani*, Mingda Zhang*, Ramin Mehran*, Rachel Hornung,
Nitesh Bharadwaj Gundavarapu, Nilpa Jha, Austin Myers, Xingyi Zhou,
Boqing Gong, Cordelia Schmid, Mikhail Sirotenko, Yukun Zhu, Tobias Weyand
Google Research †

ABSTRACT

This paper describes a semi-automatic pipeline to generate challenging question-answer-decoy sets for understanding long videos. Many existing video datasets and models are focused on short clips (10s-30s). While some long video datasets do exist, they can often be solved by powerful image models applied per frame (and often to very few frames) in a video, and are usually manually annotated at high cost. In order to mitigate both these problems, we propose a scalable dataset creation pipeline which leverages large models (VLMs and LLMs), to automatically generate dense, time-aligned video captions, as well as tough question answer decoy sets for video segments (up to 15 minutes in length). Our dataset Neptune covers a broad range of long video reasoning abilities and consists of a subset that emphasizes multimodal reasoning. Since existing metrics for open-ended question answering are either rule-based or may rely on proprietary models, we provide a new open source model-based metric (GEM) to score open-ended responses on Neptune. Benchmark evaluations reveal that most current open-source long video models perform poorly on Neptune, particularly on questions testing temporal ordering, counting and state changes. Through Neptune, we aim to spur the development of more advanced models capable of understanding long videos. The dataset is available at https://github.com/google-deepmind/neptune.

1 INTRODUCTION

Videos are experiencing an *explosion* moment online, with new research constantly pushing the frontier for video and language tasks such as video question answering (VideoQA) (Xu et al., 2017; Zhong et al., 2022; Xiao et al., 2021; Yang et al., 2021; Mangalam et al., 2023). Early video and language models, while adept at VideoQA, have largely focused on short, trimmed clips (less than 1 minute long (Yu et al., 2019a; Xiao et al., 2021)). The recent release of powerful, longer context multimodal models (eg. Gemini 1.5 (Reid et al., 2024) and GPT4 (Achiam et al., 2023)), however, has ushered in the promise of models being able to reason over millions of tokens, covering longer stretches of videos (many minutes long).

While promising, these claims are often evidenced by qualitative examples, or results on small-size datasets – for example the 1H-VideoQA (Reid et al., 2024) benchmark, which while valuable, only consists of 125 questions. Popular video benchmarks for question answering still tend to focus on short, trimmed clips (*e.g.*, Next-QA (Xiao et al., 2021)). Other datasets that *do* contain longer videos are often 'short-term' benchmarks disguised as long-term ones, evidenced by models that are able to solve them with a single (or a few) frames (*e.g.* some tasks on the LVU dataset (Wu & Krahenbuhl, 2021) such as scene prediction of movies). Other long video datasets may contain strong linguistic biases in multiple choice evaluation, as shown by MoreVQA (Min et al., 2024), which gets strong performance on EgoSchema (Mangalam et al., 2023) without access to the video at all, or can be solved with external internet knowledge, such as those made from popular movies (Li et al., 2023d).

A key challenge in creating a truly long form video understanding dataset is the significant manual cost required to select, watch, understand and annotate long videos with free-form natural language.

[*] Equal Contribution
[†] Authors now at Google DeepMind

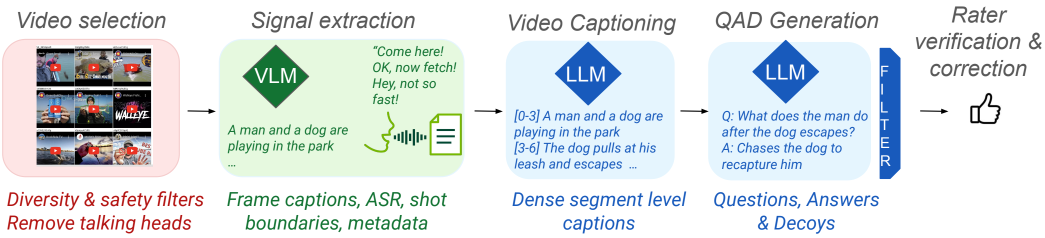
Video selection
Signal extraction
Video Captioning
QAD Generation
Rater verification & correction
VLM
"Come here! OK, now fetch! Hey, not so fast!
A man and a dog are playing in the park …
LLM
[0-3] A man and a dog are playing in the park
[3-6] The dog pulls at his leash and escapes …
LLM
Q: What does the man do after the dog escapes?
A: Chases the dog to recapture him
FILTER
Diversity & safety filters
Remove talking heads
Frame captions, ASR, shot boundaries, metadata
Dense segment level captions
Questions, Answers & Decoys

Figure 1: **Pipeline Overview:** Our pipeline consists of 5 key stages - (i) Video selection, where suitable videos are identified from YouTube, (ii) Signal extraction, (iii) Video level captioning, (iv) Question, answer and decoy (QAD) generation and (v) Manual rater verification. The first four stages are entirely automatic. Before rater verification, we automatically filter out QADs that can be solved by an LLM without access to the video content.

Answering challenging questions about longer videos is often a *multimodal* (as it may involve listening to the audio track in addition to watching the video), and *non-sequential* endeavour (as sometimes it is necessary to rewind and rewatch key parts to answer a question). Proposing suitable high-level questions that are not trivially solved by a few frames is also tricky for humans to do consistently and with adequate diversity. The key aim of this paper is to solve this challenge by leveraging automatic tools to reduce rater effort while at the same retaining quality. Inspired by EgoSchema, we do this by proposing a scalable dataset creation pipeline (Fig. 1) that leverages strong foundational Video Language Models (VLMs) and Large Language Models (LLMs) with carefully designed prompts. We first generate dense, time-aligned video captions automatically, from which tough question-answer-decoy (QAD) sets can be automatically derived. This is done by extracting image captions, automatic speech recognition (ASR), shot boundaries and video metadata, and combining these signals with multi-stage, chain of thought prompting of an LLM. Our pipeline can be applied to any video on YouTube (Fig. 1).

While most of the pipeline is automatic, a comprehensive rater verification stage at the end ensures quality. While other dataset pipelines that are entirely manual (Zhou et al., 2024; Fang et al., 2024; Wang et al., 2024), our verification stage is lightweight, which we show by ablating the automatic part of the pipeline, and measuring the time taken by raters to propose QAs for videos from scratch. Results show that our semi-automatic pipeline almost halves rater effort. Our dataset is called Neptune[1], and covers a diverse range of videos, is multimodal (requires audio and visual information), and poses challenging questions for videos that test a variety of reasoning abilities over long time horizons. Neptune allows for two modes of evaluation: multiple-choice and open-ended question answering. Since existing metrics for open-ended question answering are either rule-based and derived from captioning (WUPS (Wu & Palmer, 1994), CIDEr (Vedantam et al., 2015), etc) or are LLM-based evals that rely on proprietary APIs (such as ChatGPT[2]), we finetune an open source model on a generic answer equivalence dataset (Bulian et al., 2022) to score question answering results and evaluate it as a metric on a manually annotated answer equivalence dev set. We call this new metric Gemma Equivalence Metric (GEM).

To summarise, we make the following contributions: (i) We propose a scalable pipeline to generate complex QAD annotations for any video that halves rater time compared to manual annotation. (ii) We use this pipeline to generate the Neptune evaluation-only dataset, which consists of 3,268 QAD annotations for 2,405 videos. We also release a *challenging* subset, NEPTUNE-MMH for which *vision* plays an important role. (iii) We provide both multiple choice and open-ended evaluation metrics. For the latter, we propose a new open-ended metric called Gemma Equivalence Metric (GEM) which outperforms rule-based metrics on a manually annotated answer equivalence dataset; and finally (iv) We provide benchmarking and ablations of state-of-the-art VideoQA models on the Neptune sets. Benchmarking shows a significant gap between open-source video models and properietary models such as Gemini-1.5 and GPT-4. All data will be released publicly to the research community.

[1] Named after the planet with the longest orbit

[2] https://openai.com/index/chatgpt/

2 Related Works

Video Question Answering: Video Question-Answering (VideoQA) is an important task for assessing multimodal video understanding systems' ability to reason about videos (Xu et al., 2017; Zhong et al., 2022; Xiao et al., 2021; Yang et al., 2021; Mangalam et al., 2023). Vision and language models for this task can be broadly classified into three categories: (i) early end-to-end VLMs for this task which typically consists of strong vision and language encoders/decoders, such as Flamingo (Alayrac et al., 2022), BLIP2 (Li et al., 2023b), Video-Llama (Zhang et al., 2023a), GIT2 (Wang et al., 2022) and PALI (Chen et al., 2022; 2023a;b). These typically are moderate sized models, and memory limits often lead to significant downsampling: *e.g.* temporally sampling a few frames with large strides (Wang et al., 2022; Chen et al., 2023a) or spatially subsampling each frame to a single token (Yang et al., 2023; Zhou et al., 2018; Wang et al., 2021); (ii) Socratic style models (Zeng et al., 2022), which consists of combining various specialised *frozen* models with carefully prompted state-of-the-art VLMs and LLMs (eg. MoreVQA (Min et al., 2024)) and (iii) end-to-end large multimodal models such as Gemini (Gemini Team Google, 2023) and GPT-4 (Achiam et al., 2023), which have long context lengths and can ingest multimodal data, including video, sound and text.
Video QA Benchmarks: Key datasets have pushed towards assessing reasoning for temporal questions (Grunde-McLaughlin et al., 2021; Xiao et al., 2021; Wu et al., 2021), longer videos (Yu et al., 2019a; Mangalam et al., 2023), as well as focusing on diverse domains like instructional (Yang et al., 2021) and egocentric videos (Gao et al., 2021; Mangalam et al., 2023). We summarise existing VideoQA benchmarks in a table provided in the appendix. Most datasets either focus on shorter videos (less than 100s), or are short video datasets 'in disguise', and can actually be solved with a few frames (*e.g.* ActivityNet-QA (Yu et al., 2019b) or MovieQA (Tapaswi et al., 2016)). 1H-VideoQA (Reid et al., 2024) consists of videos longer than 1 hour, but is limited to 125 questions and is closed-source. Like Neptune, ActivityNet-RTL (Huang et al., 2024), CinePile (Rawal et al., 2024) and EgoSchema (Mangalam et al., 2023) are generated by prompting LLMs, but cover only limited domains and rely on existing annotations while Neptune covers a much broader spectrum of video types and its pipeline is applicable to arbitrary videos. Most importantly, EgoSchema also has strong linguistic biases, while Neptune mitigates these through filtering (Sec. 5). Unlike other benchmarks which come with their own training sets (eg. MSR-VTT (Xu et al., 2016), ActivityNet (Yu et al., 2019a)), we propose a generalisation-focused *zero-shot* evaluation regime. The goal for Neptune is to benchmark any model, pre-trained with any external dataset or task, in order to assess real-world domain transfer. Hence we release *test* sets only. More discussion on related datasets and dataset pipelines is provided in the appendix.

Metrics for open-ended VideoQA: Earlier QA datasets consisted of short answers (Xiao et al., 2021) (sometimes a single word), typically from a closed set, and therefore metrics such as accuracy or accuracy with exact match (EM) can be applied. As datasets have evolved with more real-world annotation (longer, open-set answers), designing a metric becomes challenging. Existing rule-based metrics for captioning, such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004) and CIDEr (Vedantam et al., 2015) can be applied, however they all primarily measure n-gram overlap, and do not capture the inherent subjectivity of the task, where different phrasing is often equally valid. Other metrics for captioning include SPICE (Anderson et al., 2016) (adds action and object relationships), while model-based metrics using earlier language models or image-language models include BERT-Score (Zhang et al., 2020), BERT-Score++ (Yi et al., 2020) (fine-tunes BERT for image captioning), LEIC (Cui et al., 2018), NUBIA (Kane et al., 2020), TIGEr (Jiang et al., 2019), CLIPScore (Hessel et al., 2021), and EMScore (Shi et al., 2022). For answer equivalence specifically, token F1 and exact match (EM) have been used, but suffer many of the same shortcomings that rule-based metrics do, and EM is often too strict for open-ended eval. BEM (Bulian et al., 2022) finetunes BERT on an answer-equivalence dataset, and shows that this provides a better score for QA. Recently, LLMs trained with reinforcement learning from human feedback (RLHF) that already exhibit strong human alignment (Bubeck et al., 2023) are used in works such as VideoChatGPT (Maaz et al., 2023) and MovieChat (Song et al., 2023) (LLM-as-a-judge). A challenge here is that the models used (ChatGPT) are called via proprietary APIs, where the underlying model may be non-static, thereby leading to non-reproducibility in the metric. Instead, we take a state-of-the-art open-sourced lightweight language model (Team et al., 2024a) and finetune it on a public answer equivalence dataset (Bulian et al., 2022), to create an open-source, static, model-based evaluation metric.