MASTER'S THESIS

Out of Order: Evaluating MLLMs on Reordering Shuffled Video Segments, Temporal Logic, and Multimodal Event Understanding

Author:
Wilfred, Okajevo

Supervisors:
Dr. Mohamad Ballout
Prof. Dr. Elia Bruni

A thesis submitted in fulfillment of the requirements for the degree of Master of Science in Cognitive Science

August 20, 2025

Eigenständigkeitserklärung / Declaration of Authorship

_____

Name, Vorname (Druckbuchstaben)/ Full Name (block letters)

_____

Matrikelnummer / student number

Die Inhalte der hier vorgelegte Arbeit geben meinen eigenen Wissensstand, mein eigenes Verständnis und meine eigene Auffassung zum bearbeiteten Thema wieder. Falls KI-Tools eingesetzt wurden, habe ich deren Einsatzweise und -zweck transparent angegeben. Darüber hinaus habe ich alle meine Quellen akademischen Standards entsprechend ausgewiesen. Ich bin bereit und fähig, die hier erläuterten Inhalte zu erklären und die entwickelten Standpunkte zu vertreten. Die vorliegende Leistung wurde weder zum Teil noch vollständig an dieser oder einer anderen Universität eingereicht.

The content of this thesis represents my own knowledge, my own understanding and my own perspective on the topic. In case artificial intelligence tools were used, their way and purpose of usage has been made transparent. Moreover, I have cited all my sources in accordance with academic standards. I am ready and able to explain and defend the positions developed in this thesis. This thesis has not been submitted, either in part or whole, at this or any other university.

_____

Datum und Unterschrift / Date and Signature

Statement of Contribution

I, **Wilfred, Okajevo**, declare that this thesis titled, "Out of Order: Evaluating MLLMs on Reordering Shuffled Video Segments, Temporal Logic, and Multimodal Event Understanding" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University. The research presented herein was conducted specifically for the fulfillment of my Master's thesis requirements.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.

The work is grounded in a collaborative research project for which I was a co-first author. My individual contributions, undertaken during my thesis period, included both foundational research and practical implementation:

- The initial research and brainstorming that shaped the methodological approach.
- The practical creation and curation of the SPLICE benchmark.
- Serving as one of the expert annotators in the human validation protocol.
- The design and implementation of the complete experimental framework.
- The execution of all model inference experiments and the collection of results.
- Actively analysing and participating in the analysis of the experimental results.

A significant portion of this research forms the basis of a manuscript submitted for peer review, with details as follows:

Manuscript Title: Can you SPLICE it together? A Human Curated Benchmark for Probing Visual Reasoning in VLMs
Authors: Mohamad Ballout*; Wilfred, Okajevo*; Seyedalireza Yaghoubi; Muhammad Abdelmoneim; Julius Mayer; Elia Bruni
Status: Accepted and Pending Camera-ready version for publication in the Findings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP).
**Denotes equal contribution as co-first authors.*

Acknowledgments

I express my deepest gratitude to my supervisors **Dr. Mohamad Ballout** and **Prof. Dr. Elia Bruni** for their invaluable guidance, support, and encouragement throughout the course of this research. Their expertise and mentorship have been instrumental in shaping both the direction and quality of this work. I also thank my colleagues and fellow researchers for their collaboration, insightful discussions, and constructive feedback, which enriched the development of this thesis.

I express my heartfelt thanks to my family and friends for the continuous support and patience during this journey. Most Specially, I am deeply grateful to my brother, **Dr. Onome Okajevo**, who has always been there for me with steadfast encouragement and advice throughout the course of my studies and far beyond - Migwo Oloto.

Statement on the Use of AI Tools

During the preparation of this thesis, I used the following AI-assisted writing and research tools in accordance with university guidelines:

- **Grammarly:** This tool was used exclusively for grammar checking, spelling correction, and proofreading to improve the clarity and readability of the text.
- **NotebookLM:** This tool was used to summarize and digest academic papers and was never used to write this thesis.

I affirm that these tools were not used to generate original content, concepts, or analysis for this thesis. All intellectual contributions, research design, analyses, and interpretations presented herein are entirely my own.

Abstract

Universität Osnabrück

Master of Science in Cognitive Science

Out of Order: Evaluating MLLMs on Reordering Shuffled Video Segments, Temporal Logic, and Multimodal Event Understanding

by Wilfred, Okajevo

The rapid advancement of Multimodal Large Language Models (MLLMs) has pushed capabilities beyond static images into the complex domain of video understanding. However, a significant "evaluation deficit" exists, as current benchmarks often fail to robustly assess a model's grasp of temporal logic and causal event structure. Many benchmarks are susceptible to linguistic shortcuts or focus on simple classification, inadequately probing deep, structural reasoning. This thesis confronts this challenge by proposing a novel evaluation methodology: a video segment reordering task. To instantiate this, I introduce SPLICE (Sequential Processing for Learning and Inference in Chronological Events), a human-curated benchmark derived from 3,381 instructional videos from the COIN dataset, segmented into 11,423 coherent event clips.

An extensive evaluation of leading MLLMs (including the Gemini and Qwen families) on SPLICE reveals a substantial performance gap, with the best-performing model achieving a perfect sequence match accuracy of only 51%, compared to a human baseline of approximately 85%. Crucially, results show that while textual annotations significantly improve model performance, they have no effect on human accuracy, indicating a strong reliance on language priors over genuine visual understanding in current MLLMs. Further analysis reveals that models perform relatively better on tasks dominated by causal and temporal logic than on those requiring contextual or spatial reasoning, and they struggle significantly when presented with visually similar but logically distinct event steps.

Ultimately, this work not only quantifies the current limitations of MLLMs in multimodal event understanding but also validates the reordering task as a rigorous and diagnostic tool for driving future progress in building more capable and human-like AI systems.

Contents

List of Abbreviations

AI Artificial Intelligence

COIN Comprehensive Instructional Video Analysis

ConvGRU Convolutional Gated Recurrent Units

CUVA Causation Understanding of Video Anomalies

EMNLP Conference on Empirical Methods in Natural Language Processing

GPU Graphics Processing Unit

LCS Longest Common Subsequence

LLM Large Language Model

MBA Multiple Binary Accuracy

MECD Multi-Event Causal Discovery

MLLM Multimodal Large Language Model

MLVU Multi-task Long Video Understanding

MoE Mixture-of-Experts

NExT-QA Next-Event Question Answering

RNN Recurrent Neural Network

SPLICE Sequential Processing for Learning and Inference in Chronological Events

STAR Situated Reasoning in Real-world Videos

SVLTA Synthetic Vision-Language Temporal Alignment

ToT Test of Time

TRN Temporal Relation Network

VLM Vision-Language Model

VQA Visual Question Answering

List of Figures

List of Tables

Chapter 1

Introduction

1.1 From Language to Multimodal Understanding

The field of artificial intelligence is currently undergoing a period of extraordinary transformation. Much of this progress can be traced back to the development of the Transformer architecture—a type of neural network model introduced by Vaswani et al [1]. that excels at handling sequential data, such as sentences, by capturing relationships across long distances in the input. This architectural breakthrough paved the way for the emergence of Large Language Models (LLMs), which are models like GPT-3 and PaLM [2, 3] that have been trained on vast collections of internet text, enabling them to understand, generate, and reason with human language. The success of these models has established a new and powerful paradigm: the "foundation model," a single, large-scale system that can serve as a versatile reasoning engine for a multitude of downstream applications.

The next step in this paradigm is to extend these reasoning capabilities into vision. This has led to the field of Multimodal Large Language Models (MLLMs), which process both visual and textual information in an integrated manner. Recent models, such as Flamingo and BLIP-2, have demonstrated that this approach is effective for static images [4, 5], enabling tasks like visual question answering (VQA) and image-grounded dialogue.

Our world is a dynamic flow of events[6, 7]. To build artificial intelligence that can understand and interact with it, we must move from recognizing static images to mastering video understanding. This requires not only identifying objects, but also grasping processes, change, and the structure of events. The main challenge is reasoning about visual events, which is the focus of this investigation.

1.2 The Implicit Grammar of Events: Temporality, Causality, and Structure

An event, as perceived by a human, is not merely a sequence of disconnected sensory inputs. It is a coherent narrative, imbued with an underlying structure governed by the twin principles of temporal progression and causal relationships. Consider the simple act of preparing a meal. This event is a structured process, composed of sub-events—such as chopping vegetables, heating a pan, and sautéing the ingredients—that are linked by a precise and often non-negotiable logic. The temporal order is critical; one must chop the vegetables before they can be cooked. The causal links are inextricable; heating the pan is a causal prerequisite for the act of sautéing.

Humans infer the temporal and causal "grammar" of events, which is central to intelligence[8]. This allows making predictions, explaining outcomes, and planning. For artificial agents to reach a similar understanding, they must go beyond recognizing video components and reason about the structure binding them together. Modern LLMs model sequences in text effectively, but their ability to apply this to video remains unclear. The core question motivating this thesis is: To what extent can modern MLLMs comprehend the temporal and causal structure of visual events?

1.3 The Evaluation Deficit: A Gap Between Capability and Measurement

MLLMs have evolved rapidly, but measurement methods have not kept pace. Current benchmarks for video understanding track specific progress, yet often fail to clearly assess structural reasoning. This thesis argues that the evaluation gap results from three main limitations in existing paradigms.

First is a focus on classification over process. Many prominent benchmarks, such as UCF101 and Kinetics (a large dataset for human action recognition in videos), are structured as action classification tasks, meaning the model must identify specific actions from video clips. A model can often achieve high accuracy on these datasets by identifying discriminative keyframes or short-term motion patterns, such as recognizing "playing tennis" by spotting a racket and a ball, without needing to comprehend the long-range temporal narrative or the sequence of actions that constitutes the event[9].

Second is the **pitfall of linguistic shortcuts**. In benchmarks that utilize a question-answering format, MLLMs can often leverage their powerful linguistic priors to deduce the correct answer from the text alone, thereby bypassing the need for deep visual grounding. As the work of Ko et al. [10] demonstrates, an LLM can answer complex temporal and causal questions with surprising accuracy without ever "seeing" the video, leading to an overestimation of its true multimodal reasoning abilities. This phenomenon, often referred to as "hallucination" or "ungrounded guessing," [10] is a significant obstacle to accurate evaluation.

Third is a lack of structural probing. There are a few tasks that directly test a model's ability to reconstruct the logical and causal structure of a multi-step event from disordered evidence. Without these tasks, we must infer reasoning indirectly, which is confounded by existing biases and shortcuts. This evaluation gap makes it challenging to pinpoint MLLM failures, identify genuine sources of success, and accurately measure progress in event understanding. We require a new evaluation paradigm that directly tests structural reasoning.

1.4 Research Aim and Guiding Questions

The central objective of this thesis is to systematically and in-depth evaluate the capability of modern MLLMs for understanding complex events, with a focus on temporal logic and sequential reasoning. I move beyond surface-level metrics to dissect the implicit understanding of temporal logic and causality. To this end, I developed a video segment reordering task as a diagnostic tool. My investigation is guided by the following research questions:

1. How proficiently can current MLLMs reconstruct the correct chronological sequence of shuffled video segments depicting a coherent event, and how does this performance compare to a human baseline?

2. What is the relative importance of visual information versus accompanying textual descriptions for MLLMs in solving event reordering tasks, and how does this compare to human reliance on these modalities?

3. To what extent can MLLMs discern and utilize different facets of reasoning—such as temporal, causal, contextual, and spatial logic—when interpreting and ordering event segments?

4. What are the characteristic failure modes of MLLMs in event sequence understanding, and what do these reveal about their current limitations in processing temporal logic and complex multimodal event narratives?

5. Can a carefully designed video segment reordering task serve as an effective methodology for benchmarking and diagnosing MLLM capabilities in multimodal event understanding?

1.5 Thesis Contributions

This thesis makes several main contributions to multimodal AI, stemming from a collaborative project in which I was co-first author. I led the groundwork, designed experiments, curated data, and evaluated models, guided by my supervisor. My primary roles were framing, analysis, and discussion. The main contributions are:

First, I propose and validate a novel evaluation method: video segment reordering for assessing an MLLM's comprehension of event structure, temporal logic, and multimodal coherence. Designing this experiment was a central aspect of my work.

Second, this work introduces the SPLICE Benchmark[11]—a concrete implementation detailing the design and curation of SPLICE (Sequential Processing for Learning and Inference in Chronological Events), a human-validated dataset of 3,381 videos segmented into 11,423 event clips. I was directly responsible and active from searching and eliminating datasets, and reaffirming the COIN dataset to data preparation, segmentation, and human annotation.

Third, we conducted a comprehensive MLLM and human evaluation. I evaluated top MLLMs on SPLICE, wrote code implementations, managed inference, and gathered results to be analyzed. I was also responsible for the human validation protocol experiment. We show human performance baseline to show the current limits of AI.

Fourth, I present an in-depth analysis of reasoning abilities within this thesis. By examining results, I identify strengths (such as handling clear causal cues) and weaknesses (like overreliance on visual similarity or trouble with context) in MLLMs.

Finally, to ensure the reproducibility of these findings and to facilitate future research, the complete experimental framework and evaluation code have been made publicly available.[1]

[1]The codebase for this research is available at: https://github.com/prokajevo/OoOMLLM

Chapter 2

Literature Review

2.1 Introduction: Beyond Static Frames

Video understanding is a major cornerstone of modern artificial intelligence, moving beyond static image understanding. A video is not just a collection of frames, but a narrative with an observable structure guided by Time and Causality. [7]To understand a video is to grasp how, when, and why things happen, not just what is present. This chapter will trace the development from foundational models of coherence to advanced multimodal large language models, aiming to understand the rules of video-visual reasoning.

Within this chapter, I outline the scholarly arguments that underpin this thesis, beginning with early ideas of leveraging temporal order as a self-supervised learning signal to teach models without labels. I then discuss architectures designed to explicitly understand temporal relationships and map dynamics in videos. The chapter expands into multimodality and causality, which are necessary for deeper video understanding and for developing architectures capable of advanced event reasoning.

The chapter will culminate with an examination of today's era of powerful VLMs and MLLMs, critically analyzing their application in video understanding. This discussion will naturally lead to the challenge of accurately evaluating these models' reasoning abilities, encapsulated by the 'Evaluation Deficit' problem. Addressing this deficit, the introduction of the video reordering task becomes not just relevant but necessary. Thus, the chapter frames the current research not as an isolated step but as an integral part of the continuing evolution toward revealing—and testing—the reasoning capabilities of modern models, bridging past advances with future needs.

2.2 Learning from Time Itself: The "Order as Supervision" Paradigm

How can a model learn without millions of human-provided labels? This foundational challenge in early video understanding research led to investigating the inherent structure of time. Researchers leveraged the fact that time in video flows in one direction, allowing temporal order to serve as a persuasive, free teaching signal. The concept labeled in this thesis as the 'Order as Supervision' paradigm posits that requiring a model to reconstruct chronological order in visual data fosters high-level semantic learning—encompassing physics, actions, and object permanence. This core notion lays the groundwork for basing the thesis evaluation methodology on reordering, providing a thread that ties this

research back to the foundational challenges described earlier.

2.2.1 Learning from Temporal Coherence and Frame-Level Ordering

Earlier work to tackle the Order as Supervision studied the issue from a close frame coherence perspective. In its simplicity, frames that are next to each other in a video are almost identical. Mobahi et. al. in "Deep Learning from Temporal Coherence in Video" [12] stated this finding. They pioneered a learning method for deep architectures that made use of the natural coherence found in unlabelled video recordings, by training it to see consecutive frames as "similar" since they are likely to contain the same object(s) undergoing minor transformation and non-consecutive frames as "different" and by enforcing these similarity and dissimilarity, the model could learn representation invariant to changes like pose, translation and rotation. Mobahi et al. used a deep convolutional network architecture and a temporal coherence regularizer to demonstrate that the self-supervised signal enhanced the performance of recognition tasks on datasets such as COIL100. The findings, without a doubt, established a baseline principle: the inherent continuity of time in video is a powerful teacher.

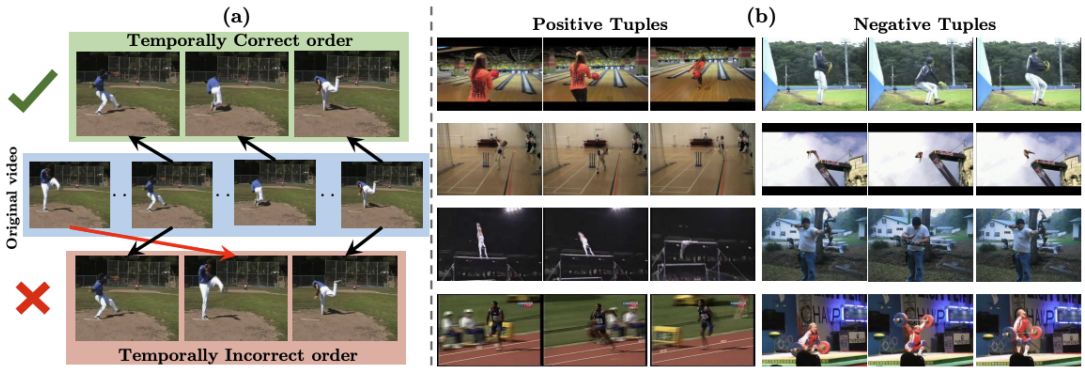

Figure 2.1: Illustration of the temporal order verification task for unsupervised learning. The model learns to distinguish between correct and shuffled frame sequences, thereby learning meaningful spatiotemporal representations without semantic supervision. Adapted from [13].

The subsequent advancement was to move past adjacent frame comparison to analyzing frames spaced further apart. This is where the seminal "Shuffle and Learn" by Misra et. al. becomes pivotal [13]. Their methodology, a precursor to this thesis, involved unsupervised tasks that verified whether a sequence of three frames followed the correct temporal order.

In their work, they trained a triplet siamese network on a tuple of three frames and a binary classification task of whether they are "temporally valid or not". There were tricky issues that needed to be addressed; the major one handled was situations involving cyclical motion. They addressed this issue by focusing on sampling from temporal windows with high motion, using optical flow as a proxy. It is worth noting that the Siamese architecture, having shared weights across three network stacks, was designed to focus on the temporal signal itself since only the middle sampled frame of a tuple would differ between a positive (e.g., frames b, c, d) and a negative (e.g., frames b, a, d) example.

The important finding from their work was that by solving this simple verification task, the underlying CNN learned a powerful visual representation that captured temporally varying information, such as human pose and their qualitative analysis, using nearest neighbor retrieval, showed that while an ImageNet-pretrained model focused on scene semantics, their unsupervised model focused on the human pose itself. This showed a crucial concept that unsupervised temporal learning captures information that is complementary to that learned from supervised static image datasets, and when used as a pre-training method for action recognition on UCF101[14] and HMDB51[15], their approach showed significant gains over learning from random initialization, establishing the viability of order verification as a powerful self-supervision technique.

2.2.2 From Frame Verification to Clip-Based Prediction

Given two frames of a gymnast on a balance beam, is it possible to tell which came first? Most of the time, it is difficult to accurately determine this without seeing the motion between them, which highlights a limitation of using only frames. Addressing this ambiguity, Xu et al. proposed in "Self-supervised Spatiotemporal Learning via Video Clip Order Prediction" extending the order prediction task from frames to video clips [16]. This step builds directly on previous frame-level approaches, aiming to capture richer dynamics inherent in video.

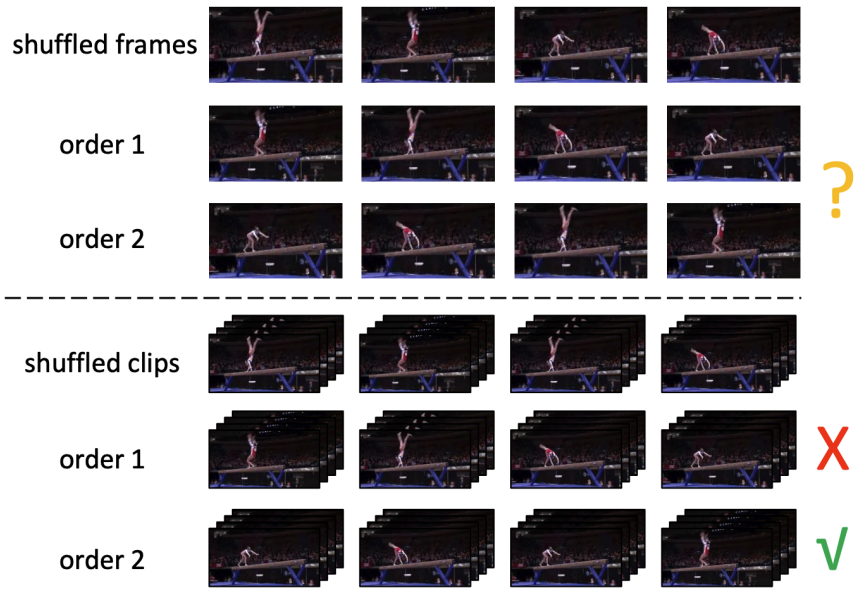

Figure 2.2: Comparison of frame-based versus clip-based temporal order prediction. The figure demonstrates how clip-based ordering resolves ambiguities present in frame-based approaches, particularly for cyclical motions, making it more effective for learning spatiotemporal representations [16].

The rationale for using clips was that, unlike frames, clips possess richer motion information, revealing scene evolution and making chronological placement less ambiguous. This offers a stronger learning signal for spatiotemporal representation. Their framework sampled fixed-length, non-overlapping clips, shuffled them, and trained a 3D CNN [17, 18] to predict the correct permutation as a multiclass classification, demonstrating that

clip-order prediction pre-trains 3D CNN [17, 18]s, which are otherwise challenging to optimize without extensive labels.

The payoff of their work and the proof that the method works were evident in the evaluation of the learned representations, which was two-fold. Using a nearest-neighbor retrieval experiment on UCF101, they first showed that their self-supervised models were able to retrieve semantically similar videos, indicating a high level of video understanding. For example, they could find videos with similar meanings, such as retrieving "uneven bars" for a "balance beam" query. Above all, when these self-supervised models were used as a pretrained initialization and finetuned for action recognition tasks, they outperformed training from scratch and other self-supervised methods. Their R(2+1)D model, for example, achieved a 13.8% improvement on UCF101 compared to the state-of-the-art model from Büchler et al. The work provides a powerful validation for the core design choice this thesis uses, that is, using video segments as the base unit for the reordering task and it very much establishes that clips are the right level of abstraction in order to learn higher dynamics and probe deeper reasoning, that would not be only spatio-temporal, but include other reasoning.

2.3 Building a Sense of Time: Architectures for Explicit Temporal Reasoning

Models learning about time and the dynamics surrounding it form the basis of the 'Order as Supervision' paradigm, achieved through surrogate tasks. Recent work has shifted toward explicitly designing architectures that model temporal relationships. This approach aims to move beyond feature averaging across frames, building models that capture complex long-range connections—the explicit narrative structure of video. This section connects previous self-supervised advances to the architectural innovations required for more robust representation and deeper understanding.

2.3.1 Recurrent Models for Sequential Data

Recurrent Neural Networks (RNNs) are naturally the most intuitive architectural choice for sequential data like videos, and this is why the work of Dwibedi et. al. used this approach for their work on temporal reasoning in videos using a Convolutional Gated Recurrent Units (ConvGRUs) [19] which is a variant of LSTMs to experiment on action recognition and their central finding was that the usefulness of memory-based models depended entirely on the difficulty of the task which in turn is defined by the nature of the dataset. With datasets like UCF101 and Kinetics, RNNs did not help much because the temporal order was not essential to solve the task, as a model could guess "playing tennis" just because a tennis was seen in the first frame without needing to learn or remember the player's motion over time. However, for the more recent 20BN-Something-Something dataset, this was much harder to do, as the dataset contained fine-grained, object-agnostic human-object interactions, such as "Opening Something" vs. "Closing Something", which not only required recognition but also temporal reasoning was critical.

On the 20BN-Something-Something dataset[20], which is a more demanding dataset, their recurrent models provided a performance boost, as their qualitative analysis showed that the RNNs' hidden states were encoding more fine-grained, visually meaningful state transitions, for example, the visual state of an object transitioning from "closed" to

"open". Fundamentally, this means that to truly test temporal reasoning, the benchmark itself must be one that resists shortcuts.

The findings from the Dwibedi et al.[19] suggest that future benchmarks should select datasets that are extensively taxing, allowing models to be tested beyond just object recognition and to probe other reasoning dynamics. Instructional videos would be ideal for such benchmarks because they are inherently rich in temporal dependencies, process-oriented, and less prone to object recognition shortcuts.

2.3.2 The Temporal Relation Network (TRN)

Sequential models, such as RNNs, though powerful, struggle to capture relationships between non-consecutive frames and distant moments in time, as they process frames individually. To address this, Temporal Relation Networks (TRNs) were proposed as a solution to the long-term decay issue. Zhou et al. introduced TRN as a simple and efficient module designed to learn and reason about temporal dependencies between frames across different time scales [21].

TRNs excelled at datasets that required deep temporal reasoning, and it is for this reason that their models performed well on datasets like Something-Something, Jester, and Charades, as these were temporally heavy. The qualitative summary of their work showed that TRNs learned to identify the most representative frames that capture the essence of an action, including its beginning, middle, and end. TRN-equipped networks outperformed baseline networks that used simple frame averaging.

Thus, benchmarking models to reason across timescales and relationships is important[21]. Segment reordering would be a logical extension, as it would challenge models to perform this kind of relational reasoning implicitly to accurately reconstruct coherent video sequences.

2.4 Beyond Sequence: The Roles of Multimodality and Causality

Human perception operates in multiple modalities: when watching a video, we also process audio or text through captions, subtitles, or descriptions. This integration, known as multimodality, enriches our understanding beyond simple visual observation. Furthermore, humans naturally progress from perceiving temporal sequences to inferring causality, understanding not only that A precedes B but that A may cause B. Advanced AI systems must similarly address multimodality and causal inference to move beyond temporal ordering toward a more comprehensive understanding of events, building conceptually on the progression outlined thus far.

2.4.1 Multimodal Fusion for Temporal Ordering

Enter Sharma et. al., one of the key researchers who make a strong case for not only using video frames, but also including audio and text. They demonstrated that combining these multiple modalities yields a better understanding, proposing a model to learn multimodal clip representations by utilizing the temporal ordering of unordered video clips as a proxy[22].

They encoded all modalities into a joint representation and trained a model in a self-supervised paradigm to infer the order of unordered video clips. Their findings show that

different modalities complement one another, and combining them significantly improves performance[23]. This suggests that future benchmarks should test models with both video and video+text modalities to measure how linguistic cues influence task accuracy and whether models rely on them over visual reasoning.

2.4.2 Towards Causal Reasoning in Video

Climbing the cognitive complexity ladder from temporal to causal reasoning marks a significant leap. Unlike simply knowing the sequence of events, understanding the causal links between them is a more intricate challenge. As the next key direction in the field, a growing body of research is now tackling this essential hurdle.

Causal discovery have been ongoing and one of the early work of Li et. al.[24] is a foundational example of some of these works even though it was in a simple domain, and in their paper Causal Discovery in Physical Systems from Videos, their goal was to discover latent causal structures for example whether two balls were connected by a rod or string just by observing a video of their physical interactions. The components they used were quite direct, and they included a perception model used to extract keypoints, an inference module to determine the causal graph, and a dynamics module to predict the future. The main takeaway was that to infer causal mechanisms from raw visual data was a challenge. In recent times, Chen et al. introduced the MECD benchmark[25],

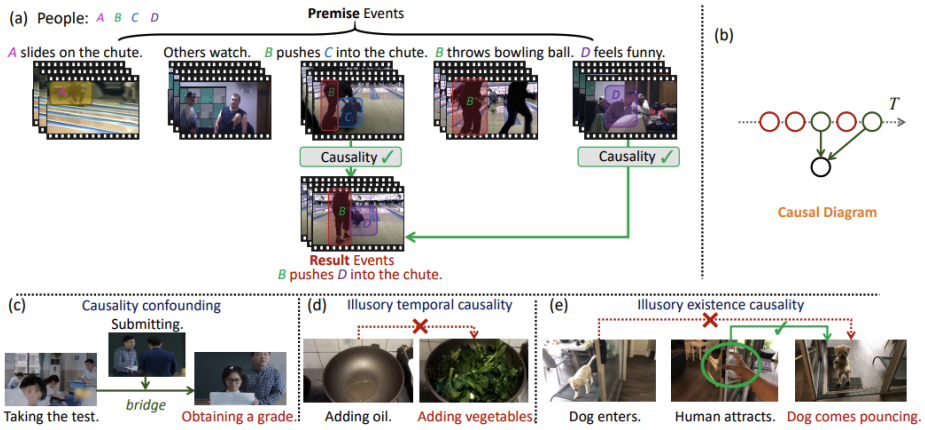

Figure 2.3: Conceptual overview of Multi-Event Causal Discovery (MECD) in video reasoning. The diagram illustrates the task formulation and examples of causality confounding and illusory causality that challenge current video understanding models [25].

which focuses on unlocking Multi-Event Causal Discovery in Video Reasoning. As their task and dataset address causal reasoning in long, real-world videos, and moving beyond simple question-answering frameworks, this in fact formalizes the idea of using multiple event segments as this is a multi-event scenario, which provides a great bed for reasoning about complex causal chains.

Guo et. al. in 2025 also introduced a causal event modelling framework, TRACE ("Temporal Grounding Video LLM via Causal Event Modeling")[26], which represented a video as a sequence of structured events that include timestamps, captions, and salient

scores, and the model predicts the next events based on the previous events and instructions.

Causal reasoning and these works set a grand picture of what visual reasoning should look like, and this "high bar" is needed for capable causal mapping and video understanding models. Future benchmarks should include tasks that are highly causal in nature and require deep causal inference (e.g., recognizing that mixing flour and water causes the formation of dough, or blowing on a dandelion would cause the seeds to disperse and fly off in the wind), to probe whether MLLMs can move beyond simple temporal dynamics to this more complex logic of cause and effect.

2.5 The VLM Revolution: A New Era of Capability and the Persistent Challenge of Evaluation

The current landscape of MLLMs, particularly VLMs, is the result of the convergence of several research threads discussed thus far, including temporal modeling, causal reasoning, and multimodality. These important concepts and paradigms are central to modern MLLMs. Works like Video-LLaMA and VideoLLM exemplify architectures that fuse a vision encoder—such as a 3D CNN [17, 18] or ViT [27] —with powerful pretrained LLMs, achieving remarkable abilities like video dialog, temporal grounding, and even zero-shot understanding, which was a flagship feature of VideoInsta[28]. The findings of Ko et al. support this, showing that LLMs possess an intrinsic prior enabling deep temporal and causal reasoning, which can be leveraged for Video Question and Answer tasks[10]. The state-of-the-art models that are the subject of this thesis include the Gemini family of models, Qwen2-VL, LLaVa-OneVision, and InternVL 2.5[29, 30, 31, 32], which, in a broad sense, represent the culmination of past advancements in the field of video understanding. In fact, these models are not just large; some have variants with as few as 3B parameters and up to 70B+ parameters.

The Gemini family has pushed the boundary of multimodal understanding through a well-bred, sparse Mixture-of-Experts (MOE) transformer-based architecture. It is interesting to see the performance in the reason the Gemini models hold, as they are capable of processing and reasoning across videos that are hours long, due to their huge context windows in the million-range, and the 2.0 flash experimental is no slouch in this regard as well.

On the other hand, the open-source community also has some great contenders. The Qwen2-VL series[30] stands out as one of the most capable open-source variants, introducing a novel technique for handling visual data in model training. Its unified "Native Dynamic Resolution" approach processes videos and visual data across various resolutions, while "Multimodal Rotary Position Embedding (M-RoPE)" fuses positional information from all modalities—text, images, and video frames. Unlike Gemini, which also processes audio, the Qwen-2 VL series does not, but its unified paradigm enables a coherent framework for images and video.

LLaVA-OneVision [31] is the final spot in the LLaVA lineage, as the research line focuses more on creating cost-efficient recipes and facilitating easy transfer of visual tasks. This concept involves a model trained on images achieving a high level of zero-shot performance on video tasks. Their major contribution reaffirmed that there is power in high-quality instruction data and training strategies that are scalable. Finally, the InternVL 2.5 [32] also adopted a scalable paradigm, achieving results comparable to those

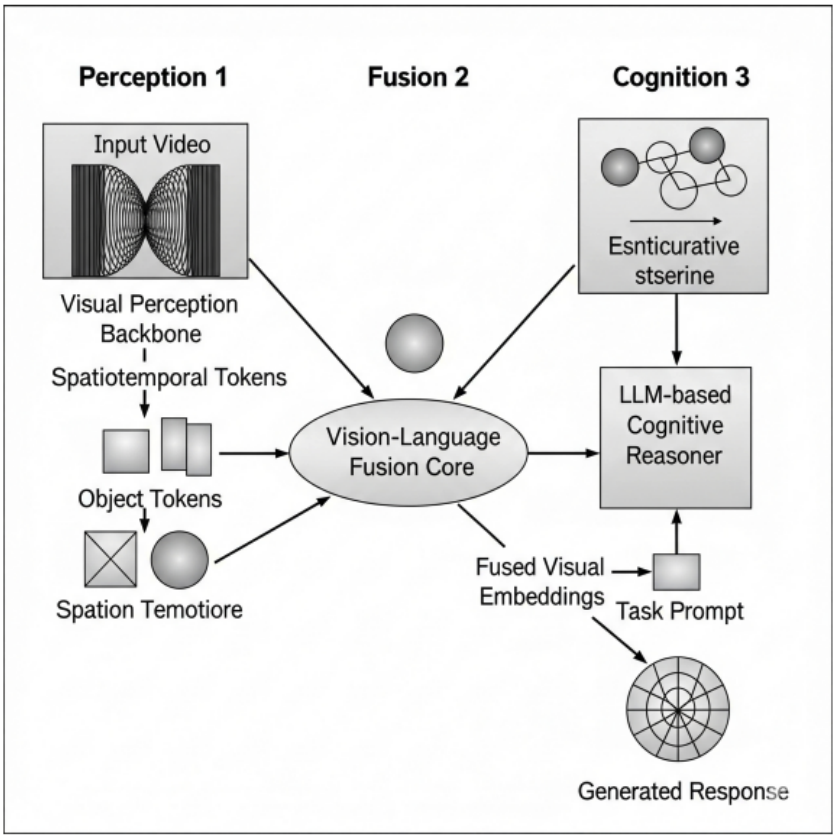

Figure 2.4: Architecture of an LLM-based framework for video event reasoning and prediction. The system processes raw video through visual perception, vision-language fusion, and LLM-based cognitive reasoning components for explicit temporal reasoning [10].

of the leading open-source and closed-source models. They systematically improved their vision encoder, the LLM, and the quality and quantity of training data.

However, the sophistication, architectural design, training data, and techniques that make these MLLMs so capable also make them difficult to evaluate or probe. It remains challenging to determine whether they are truly solid reasoners and to what extent they visually understand video streams. This challenge is heightened by the added complexity of the Video Encoders + LLM stack. As I mentioned earlier in this thesis, these models have created an "evaluation deficit," as their advanced linguistic capabilities can create an illusion of understanding without deep visual comprehension. Recall that the work of Ko et al. demonstrates that LLMs possess an intrinsic prior for temporal and causal reasoning, which they may use as a linguistic shortcut on many benchmarks[10].

The Evaluation Deficit represents a critical challenge that needs to be addressed in the field. This logical gap persists because many current benchmarks focus on simple recognition tasks or those prone to linguistic shortcuts. What is needed is an intuitive yet challenging evaluation paradigm that can directly probe structural reasoning capabilities.

An ideal evaluation methodology should be designed to reduce the limitations of previous benchmarks for evaluating MLLMs, this benchmark should likely be:

Human Curated - A human curated benchmark enables meaningful, undistracting, and unambiguous event-based scene annotations. It could be timestamped and rigorously filtered, resulting in a Dataset that is a a more detailed, unambiguous, event-segment-

based collection of clips under a single thought or process.

Event-Based Structure - To prevent models from relying on simple shortcuts, the can be highly event-based, allowing for discrete event reasoning through this segmentation. For example, in a karate fight, you have an opening salutation, a fight, and a closing salutation. A shortcut, such as matching similar frames' opening and closing salutations, would most of the time result in failure.

Multi-faceted reasoning - The task should be highly diverse and require a basket of reasoning; temporal, causal, spatial, contextual, and even common sense. Controlled Modality: It should be evaluated in two modes: "video-only" and "Video+text," where annotations of segments can be provided to measure the impact of language prior.

Such an approach would need to demonstrate that even SOTA models struggle compared to human performance, highlighting the complexity of video understanding and true visual reasoning. This would validate the reordering task as a rigorous and necessary evaluation paradigm for the future development of MLLMs that excel in visual reasoning.

2.6 A Critical Review of Video Understanding Benchmarks

Progress in artificial intelligence depends strongly on the quality of datasets and, perhaps more importantly, on benchmarks that evaluate models. A well-designed benchmark does more than produce a score; it highlights specific weaknesses in models and points to areas for improvement. These benchmarks are crucial for identifying the limitations of current MLLMs and, by extension, other vision-language models. High-performing models require not just good data and techniques but also benchmarks that thoroughly assess their internal workings. Yet, since deep neural networks often function as black boxes, it is difficult to benchmark them effectively. To advance video understanding, benchmarks must go beyond surface tests and thoroughly examine visual reasoning, including temporal, causal, spatial, and commonsense contextual reasoning.

To fully appreciate the necessity of a deeper benchmark in testing SOTA MLLMs, it is important to first conduct a critical review of present benchmarks, provide a brief overview of their performance, and highlight their strengths. Therefore, benchmarks will be categorized based on the specific cognitive dynamics they probe, from temporal to causal to multimodal, as well as additional reasoning tasks. Their effectiveness will be discussed, recognizing that this is not an exhaustive list. This approach will illustrate the Evaluation Deficit currently facing the field, thereby motivating a need for a deeper diagnostic tool that uses these deeper level of probing.

2.6.1 The "What" Question: Foundational Benchmarks for Action Recognition

The first challenge in the scope of video understanding was scenarios and benchmarks that tested the "What" questions. What action is happening in frames; however, they often ended up testing static object recognition instead of event understanding. These earliest benchmarks, which ultimately became classification tasks, set the stage for what the field currently has. The early foundational datasets were UCF101 & HMDB51 [14, 15]; however, as stated by Misra et. al [13], these benchmark datasets contained short

clips where it was obvious to predict an action from a single frame without having to understand the other dynamics from other frames, as shown in section 2.2.1 of this thesis.

The Kinetics dataset introduced scale to the field by offering hundreds of thousands of video clips across numerous action classes, enabling the development of larger models [33]. However, Dwibedi et. al [19] found that, despite its large scale, achieving high performance on this benchmark did not require a nuanced understanding of temporal order. Instead, a simple baseline using temporal averaging of frames yielded comparable results. This indicated that many tasks could be solved by recognizing the presence of objects rather than understanding the sequence of events, thereby emphasizing classification over more complex temporal reasoning.

The next benchmark dataset advanced from basic action recognition to capturing more nuanced actions. The 20BN-Something-Something and Jester datasets were introduced to address the limitations of earlier datasets [20]. They were specifically designed to require models to consider temporal order as a non-negotiable aspect, featuring fine-grained, object-agnostic interactions. For example, distinguishing 'pushing something from left to right' from 'pushing something from right to left' demands attention to sequence, and networks equipped with TRNs have demonstrated improved performance on these tasks.

Although these benchmarks contributed to foundational model development, they fall short of probing the depth of reasoning skills[34] emphasized by this thesis. They focus mainly on identifying "what" occurs in a video event—tasks that can be completed without engaging with the "how" or "when" an event happens. Thus, they are insufficient for evaluating deeper understanding.

2.6.2 The "When" Question: Advancing to Temporal Reasoning

To probe deeper into temporal order understanding, the field moved on to more demanding datasets and tasks that required an explicit understanding of time and the ability to localize events within untrimmed videos. ActivityNet and Charade-STA were an important duo for this stage [35, 36]. ActivityNet, in particular, introduced the task of temporal action localization, where the goal was to identify the start and end of an event within a longer video. Charade-STA took it a step further by probing temporal grounding via natural language queries. The limitation was that the task was framed as a retrieval or proposal-based problem, rather than a generative one, as the models simply select the best candidate segment rather than reconstructing the segment boundaries, and an event occurs instead of being derived from first principles.

The next set of sophisticated temporal benchmarks includes REXTIME, TDDiscourse, YouCook2, and Test of Time (ToT)[37, 38, 39, 40]. These go beyond simple localization to test more complex temporal skills. For example, REXTIME measures cross-event cause-effect, TDDiscourse tests global event ordering, YOUCOOK2 evaluates procedural steps, and TOT serves as a diagnostic suite of specific temporal abilities.

While these "When" datasets represent a major step forward in temporal reasoning, they still do not address the "Why" question at the heart of this thesis. Often solved with discriminative approaches, these benchmarks also rely on natural language queries, which introduce linguistic shortcuts that will be discussed in the next section.

2.6.3 The "Why" Question: The Challenge of Causal Reasoning

Therefore, the pursuit of true event understanding necessitated shifting from "what" happened and "when" to "why" it happened. This requires the model to understand the

concept of cause-and-effect; recent benchmarks have been designed to probe this deeper level of comprehension. Benchmarking high-level and specialized causality awareness includes Next-Event Question Answering (NExT-QA) and Situated Reasoning in Real-world Videos (STAR) [41, 42], which were designed to evaluate a model's ability to answer complex questions about causal relationships and temporality. An example of a Question in NExT-QA often takes the form of "What happens next?" or "Why did the person do X?" which requires not only predictive but also explanatory reasoning. STAR focuses on situational reasoning, encompassing the feasibility of actions and the interactions between objects.

Another recent benchmark, CausalVQA[43], addresses causal reasoning in real-world, physically grounded scenes. It employs counterfactual, hypothetical, and descriptive questions, reducing reliance on linguistic cues by including paired questions with perturbed distractors. Causation Understanding of Video Anomalies (CUVA)[44] provides a nuanced benchmark for understanding the causes of anomalies in video, introducing a novel metric called MMEval that aligns with human preferences for multimodal inputs [44]. CUVA supplies annotations for the "What," "Why," and "How" of an anomaly, as well as a free-text explanation of the cause and effect.

Multi-Event Causal Discovery (MECD) [25] takes it a step further by evaluating causality in settings involving multiple events. It requires a model to decide if one event causes another that follows. MECD uses Top-1 accuracy for evaluating causal relation chains and Structural Hamming Distance for comparing causal graphs[25]. Other text-based benchmarks include COPES and GLUCOSE[45, 46], both of which focus on extracting event causality from text. COPES uses F1 scores, while GLUCOSE uses BLEU, BERTScore, and sentence similarity. Research suggests that models perform better on downstream tasks, such as evaluating story quality and aligning video and text, when event causality information is provided.

Although these benchmarks advance evaluation of causal reasoning, most use a question-answer format vulnerable to the "Linguistic Shortcut Problem." Foundational research by Ko et al.[10] shows strong MLLMs can answer many causal questions using prior knowledge and language patterns, often without video evidence. For instance, for questions like "Why did the person add flour to the bowl?" with choices "(A) to make a cake" and "(B) to wash the dishes," models may pick the right answer based solely on cooking context, not visual content.

2.6.4 The Modern Era: Comprehensive Suites and the Limits of Simulation

Current evaluation approaches have split into two trends: (1) comprehensive benchmarks that test a broad range of model abilities, and (2) specialized synthetic benchmarks that focus on specific reasoning skills within controlled environments. Both trends reflect the growing need for better assessments of advanced multimodal large language models (MLLMs) and vision-language models (VLMs).

Benchmarks like MM-Bench evaluate a model across a wide array of single and multi-image tasks [47], while VideoChatGPT[48] Bench uses the SOTA GPT-4 as a judge to evaluate conversational responses about a video across multiple dimensions, including correctness, details, and temporal understanding. The Multi-task Long Video Understanding (MLVU) Benchmark also utilizes GPT-4 as a judge, incorporating longer videos, a diverse set of evaluation tasks, and a wider range of video genres[49]. Another notable benchmark is TemporalBench[50], which evaluates fine-grained temporal understanding

in videos. It is worth noting that this benchmark handles linguistic shortcuts by implementing Multiple Binary Accuracy (MBA), as the benchmark is based on a plethora of question-answer pairs. It has shown a significant performance gap between SOTA models and humans. Other benchmarks to name a few include VideoLLM Benchmarks[51, 52], VRPTEST [53], and PSALM [54] Benchmarks.

While these benchmarks are great for assessing general quality, they lack targeted diagnostic probing of structural event reasoning, as they rely on open-ended or multiple-choice formats that do not force the model to reconstruct the temporal and causal backbone of an event, nor the plethora of other reasonings.

The last of the two directions is synthetic benchmarks, such as CLEVRER and CATER [55, 56], designed to test causal reasoning in a controlled, physical environment. Synthetic Vision-Language Temporal Alignment (SVLTA)[57] provides a fair diagnostic framework by controlling the temporal distribution of events and addressing biases found in real-world datasets. For video prediction and temporal Action Segmentation, researchers used KTH, BAIR, Human3.6M, and UCF101 datasets, as well as 50Salads, Breakfast, BEOID, and GTEA [58, 59, 60, 14, 61, 62, 63, 64]

2.7 A New Direction: The Rationale for Event Reordering as a Diagnostic Paradigm

The evaluation deficit for probing deep visual understanding and the broad spectrum of underlying reasoning is increasingly significant in the context of current MLLMs, highlighting the need for a new methodological direction. A video segment reordering task offers a powerful diagnostic tool. The rationale for this task is threefold: its strong foundation in representation learning, its inherent robustness against the pitfalls of previously discussed benchmarks, and its strength as a direct test of structural, visual reasoning, and understanding.

The video segment reordering task is underpinned by well-established concepts in machine learning and is influenced by how humans reason about visual data. The principle of order as supervision, discussed in Chapter 2.2, posits that the natural, chronological flow of time in a video serves as a powerful, freely available teaching signal. Training a model to predict the correct temporal sequence compels learning of high-level semantic features without explicit labeling. Approaches such as 'Shuffle and Learn' [13] and 'Self-supervised Spatiotemporal Learning via Video Clip Order Prediction'[16] support this method. This thesis builds on such approaches, positioning the task as not only a pre-training strategy but also as a direct evaluation methodology to probe the very representation the model is intended to develop.

Secondly, the task directly addresses the flaws identified throughout this chapter, as it is robust against benchmark pitfalls. It addresses the "Classification over Process" problem, as observed in benchmarks such as UCF101[14] and Kinetics[65]. The reordering task makes the temporal narrative the essence of the problem, as it forces the model to comprehend the entire process, the change, and the sequence of actions, rather than simply succeeding by recognizing an object in a frame. It comprehends not just what, but the how and why events unfold as they do. The other benchmark pitfall it handles is the "Linguistic Shortcut" problem, which is prominent across the most advanced question-answering benchmarks. A Benchmark dataset with a primary input of disordered visual data and text that is descriptive, rather than interrogative, bypasses this drawback. Ko

et al.[10] exposed this shortcoming, and reordering tasks are a more authentic test of multimodal reasoning, as the model must succeed by grounding its logic in the visual evidence.

Thirdly, the task of video reordering is more than a clever puzzle, it is a direct and generative probe of a model's internal understanding of logical structure, so, instead of asking a model to discriminatively choose the best answer from a predefined list, it demands generative reconstruction of a coherent narrative from disordered evidence, this takes a high level of reasoning and understanding, for example, while preparing a meal, ingredients must be chopped and prepared before they are cooked also, the pan must be heated before sautéing. These logics are non-negotiable to accurately complete such reordering, and therefore serve as a great way to probe reasoning. The diagnostic power of this lies in the errors or successes the models achieve, for example, a model that fails to correctly place a sub-event can reveal where its understanding breaks down, and we can then probe "why" it failed, does it struggle with long-range dependencies, or confuse similar but logically distinct steps? Or does it fail to recognize causal links between events?.

Thus, the diagnostic power of the reordering task also provides a clearer and more detailed look into a model's deep reasoning capabilities than a simple accuracy score on a classification task or question-answer task could ever offer, making it the ideal paradigm to push the field forward.

2.8 Conclusion: Establishing the Ground for a New Inquiry

From the foundational concept of learning from temporal coherence through order as supervision, to today's sophisticated multimodal systems, this review of the literature has extensively explored video understanding. It has revealed the clear progression from implicit learning of temporal order to models and architectures designed to learn relational reasoning explicitly.

Despite this progress, the literature reveals a crucial gap. As models become more powerful, evaluating and probing their deep reasoning abilities becomes more difficult, especially with today's multimodal large language models. Current benchmarks and datasets have limitations that create an Evaluation Deficit. Many can be exploited by models that use object recognition instead of process understanding, relying on static cues. There is also a vulnerability to linguistic shortcuts, where models use language skills to answer questions without really understanding the video. Furthermore, there is a lack of generative tasks or benchmarks that force a model to reconstruct the logical structure of an event.

This precisely defined gap is the reason for my thesis. The Video Segment Reordering task is not just another benchmark, but a direct and targeted response to the evaluation deficit problem the field faces. The following chapters demonstrate how this simple, intuitive, and yet deeply challenging paradigm can be implemented in a benchmark capable of probing the very capabilities that are essential for the next generation of artificial intelligence.

Chapter 3

Methodology

3.1 Introduction

The preceding chapter traced the evolution of video understanding and highlighted a persistent evaluation gap: existing MLLM benchmarks struggle to probe deeper reasoning. This motivates an evaluation paradigm based on video segment reordering — a generative probe robust against known shortcut strategies.

This chapter details the instantiation of that paradigm through the SPLICE (Sequential Processing for Learning and Inference in Chronological Events) benchmark [11]. Developed collaboratively and submitted to EMNLP 2025, SPLICE was designed as a human-validated, event-based dataset. My specific contributions included: (i) designing the reordering task, (ii) curating the dataset, (iii) implementing the evaluation framework, (iv) managing human validation, and (v) executing experiments and active participation in analysis. The following sections describe this methodology in detail.

3.2 The SPLICE Benchmark: Design and Curation Pipeline

The main objective of the SPLICE benchmark [11] is to provide a human-validated dataset that is not only diverse, but challenging enough to test MLLMs, by subjecting it to a reordering task where the model has to generatively reconstruct the order of segments of a video based on a logical and temporal multistep events. The multistage development process of SPLICE as illustrated in Figure 3.1 was done in a carefully curated setting to ensure quality, coherence and the diagnostic power of the final benchmark in probing visual understanding.

3.2.1 Stage 1 & 2: Sourcing and Selection of Foundational Data

The foundation of the SPLICE benchmark is the COIN (Comprehensive Instructional Video Analysis) dataset [66]. The COIN dataset was chosen after an extensive research, review, and even multiple sandboxed experimental reasoning setups for multiple datasets. COIN was strategically chosen for its high suitability to the reordering task, and what made it superior to other datasets like YouCook2, Kinetics, Something-Something[39, 65, 20] was the inherent procedural logic and its granular even-based annotation which made it possible to logically create human-vetted event-based segmentation.

My responsibility in this initial stage aside the initial dataset elimination and sandboxing experiments was the systematic filtering of the 3,600-video subset from the COIN dataset to be highly diverse and representative of the entire dataset. This involved a manual review process to exclude instances with high repetition or visual ambiguity, thereby ensuring the foundational quality of the benchmark.

It is worth noting that the COIN dataset wasn't just procedural and capable of being used for deeper reasoning probing, it was also large-scale and diverse and its three-level hierarchical structure made it easy to get an holistic view of the Domain, Task and Steps, which is a powerful feature that required no extra annotating to accomplish.

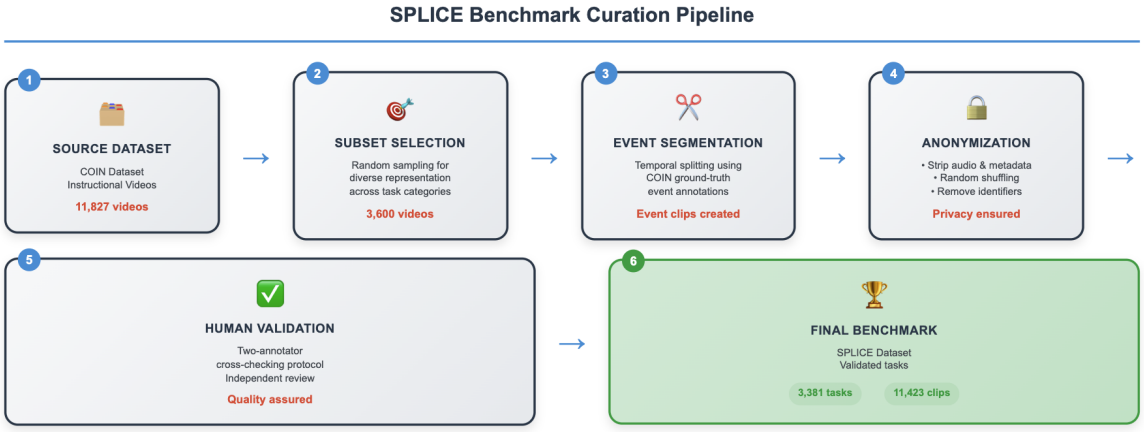

Figure 3.1: The SPLICE Benchmark [11] Curation Pipeline. This diagram illustrates the six-stage process for creating the benchmark dataset. Starting from the source COIN dataset (11,827 instructional videos), videos undergo: (1) random sampling for diverse representation yielding 3,600 videos, (2) temporal segmentation using COIN's ground-truth event annotations to create distinct event clips, (3) anonymization and standardization to isolate visual reasoning capabilities, (4) rigorous human validation where tasks are excluded only when both independent annotators mark them as unsolvable. The final benchmark contains 3,381 validated tasks distributed across 11,423 event clips.

3.2.2 Stage 3 & 4: Event-Based Segmentation and Anonymization

Event-centric segmentation is a strategic, non-arbitrary decision for creating SPLICE. Unlike fixed-length clipping, the event-based segmentation methodology forces the model and humans to reason about high-level even logic. Shifting the evaluation away from a test of low-level visual continuity to one of procedural logical reasoning is non-negotiable to probe MLLMs performance. The event-based segmentation was done using the human annotation and timestamps the COIN dataset comes with. Fig 3.2 shows a sample, in this case a "Make French Fries" Video has its Start and End timestamp for each individual logical procedural segment of the task.

Closer inspection reveals that by matching these human annotated timestamp of event, we can have situations where the videos are also not sliced to cover the entire length of the video, just the specific logical segments in. the bigger picture. Segment S1 starts at 00:02 and ends at 00:18 but S2 does not start immediately, instead it begins at 00:22

and ends 00:26. This simple but intuitive segmentation makes it possible to probe visual reasoning and contributes in structurally avoiding scenarios where frames are completely adjacent, include a shuffling mechanism at the clip level, then this becomes a powerful probe.

A strict anonymization and standardization was necessary to be able to create a dataset that was fair across the board. To accomplish this and to prevent any iota of metadata and order leak, I first and foremost stripped all the event clips of their audio track to isolate visual reasoning and to ensure fairness across all MLLMs and even humans. For example, some of the models that were tested possess audio capability, and this may create a shortcut as a model could possibly clip segments together based on what was said or the audio itself, this would be an unfair comparison to models that do not possess an audio representation, and also, it would make true visual reasoning probing messy.

Aside audio stripping, all original metadata were removed including filenames and source timestamps, were completely removed. Subsequently, the clips within each task were randomly shuffled and renamed to a generic, anonymized format and standardized into the .mp4 format.

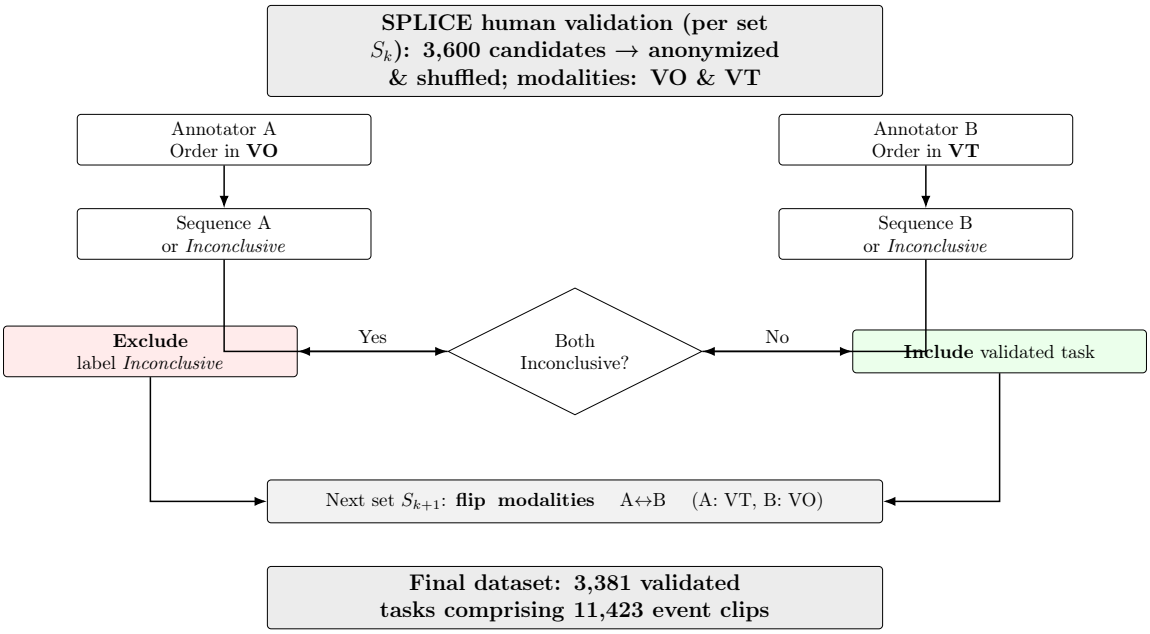

Figure 3.2: Cross-checking with modality flip: exclude a task only if *both* annotators mark it *Inconclusive*.

3.2.3 Stage 5: Human Validation for Quality Control and Baseline Performance

The final and most critical stage in the curation of the SPLICE benchmark was a rigorous human validation process, the goal is to establish the scientific rigor of the SPLICE benchmark. The correct chronological order of every clip was already known as this was derived from the COIN dataset time-stamped annotation as shown in Fig 3.2, but this was not enough. Verifying human **solvability and unambiguity** of each reordering task was crucial thereby filtering out flawed, trivial or unfair tasks.

The 3600 videos initially filtered from section 3.2.1, was split into sets for validation. The validation was done by a team of four annotators, and each set consisted of one PhD

and Master's student in cognitive science. This was done to ensure high consistency and quality of the task itself, and for this reason a crowd-source was not used, as the COIN dataset was easily available and may become game-able and this would have affected the quality of the benchmark.

Each team was assigned 1800 tasks, already shuffled and anonymized. They received a detailed instruction brief on how to validate and complete the task, which was to attempt to reconstruct the correct sequence and also identify tasks that they deemed unsolvable. No external sources policy was enforced as they solved each task based on the visual evidence they had, just like the models would have to. The workflow for an Annotator is shown in Fig 3.9

For a given set of videos, Annotator A would order them in the video-only modality, while Annotator B would independently order the exact same set of videos in the video+text modality, after which the modalities would be flipped for the next set. A video was labeled as "inconclusive" and excluded from the final benchmark only if both annotators independently agreed that a coherent sequence could not be determined as shown in Fig 3.2. For the sake of explicit accuracy, If one annotator provided an order and the other marked it "inconclusive," the task was kept as this acknowledges that a solvable path might exist even in a complex task. It never matters if the annotators gets a task wrong, it is kept at all time as long as both annotators did not independently mark inconclusive. At the end of the process, the final curated dataset for SPLICE was 3,381 human validated video tasks. The human baseline performance was calculated by comparing against the ground truth from COIN timestamped annotation and this is the gold standard for performance, representing expert human ability on a rigorously filtered and validated set of tasks.

3.3 A Taxonomy of Reasoning in SPLICE

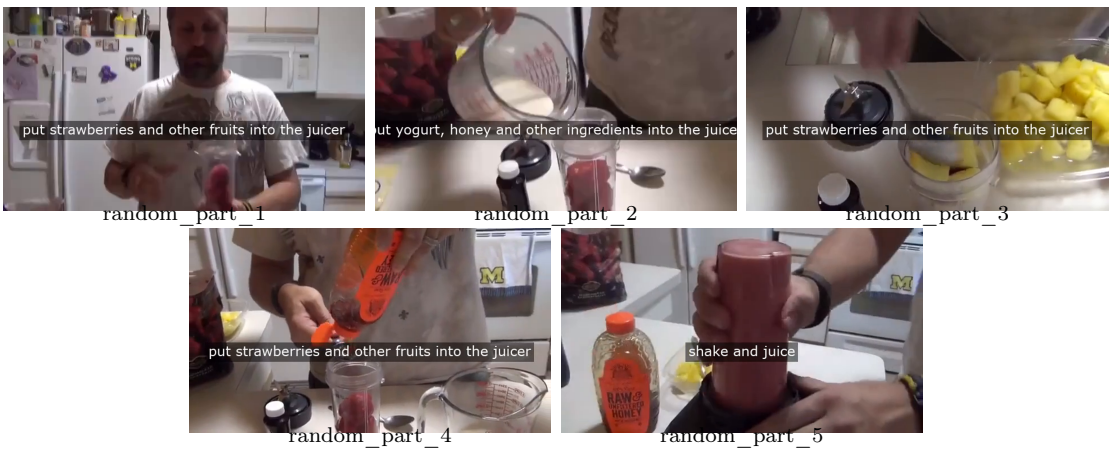

Figure 3.3: Temporal–Contextual example (Video 54). Shown here in raw order (`random_part_1 ... random_part_5`). The correct chronological order is: `1→4→2→3→5`.

The procedural and instructional nature of the source video ensures that the SPLICE benchmark is not a simple reordering task based on simple temporal perception, to succeed on splice, a human or model must be capable of an interconnected plethora of reasoning skills. This section provides the five prominent reasoning dimension that SPLICE probes, with an example from the SPLICE dataset.

SPLICE probes five dimensions of reasoning:

1. **Temporal reasoning** – tracking object states across time. Example: Video 54 (smoothie-making).

2. **Causal reasoning** – inferring cause–effect relations. Example: Video 111 (cooking noodles).

3. **Contextual reasoning** – using environment/process dependencies. Example: Video 140 (PC assembly).

4. **Spatial reasoning** – interpreting trajectories/orientations. Example: Video 458 (parallel parking).

5. **Commonsense reasoning** – applying prior world knowledge. Example: Video 538 (medical injection).

Temporal reasoning is the ability to grasp the correct linear sequence of events and this is possible by understanding the changing states of objects over time. The sample Video 54 task shows this temporal reasoning dynamics, as it shows the process of "Making a Smoothie". For a model or human to precisely reorder a shuffled clip like this, it must be able to track the state of the blender's contents to know that adding honey must happen when only strawberries are present and adding more fruit must happen after milk is present.

Causal Reasoning recognizes the cause and effect relationship that governs why events happen in a particular order. It is broad and very much intertwined with and inspiring other reasoning types. This is illustrated in Video 111 (Cooking Noodles) and Fig 3.4, where the model or human must infer that the visible softer texture of the noodle is an effect of it being cooked longer. **Contextual reasoning** is also a process-dependent reasoning, and it encompass understanding the logical dependency of a process based on the state of the environment. Consider Video 140, the model must the visual context to infer that installing a power supply into an empty case must precede installing a fan into a partially filled case.

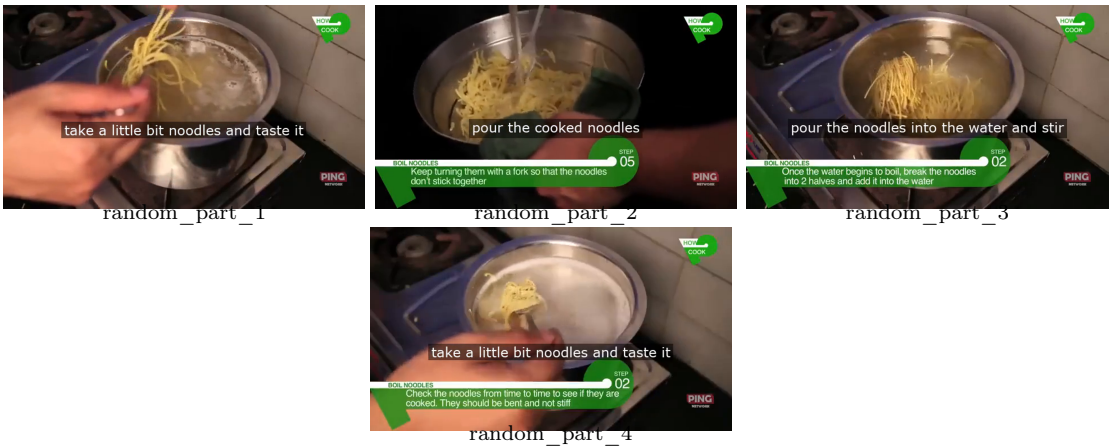

Figure 3.4: Causal reasoning (Video 111). Shown here in raw order (`random_part_1 ... random_part_4`). The correct chronological order is: `3→1→4→2`.

Spatial Reasoning interprets an object's trajectory and orientation in 3D space over time, and this is clearly illustrated in Video 458 which is about parallel "Parallel Parking" as the sequence will be dictated entirely by the car's spatial movements relative to the parking space, i.e., driving parallel, reversing, adjusting, and leaving as shown in Fig 3.5

Figure 3.5: Spatial reasoning (Video 458). Shown here in raw order (`random_part_1 ... random_part_4`). The correct chronological order is: `2→4→1→3`.

Lastly, **General and Commonsense Reasoning** is applying external real-world knowledge to infer a plausible sequence when visual cues are insufficient, and Video 538 from the SPLICE benchmark which describes "Medical Injection" scenario. The model must access its internal script and commonsense understanding of medical procedure to infer that sterilization must precede the injection itself and not the other way around.

Figure 3.6: CommonSense reasoning (Video 538). Shown here in raw order (`random_part_1 ... random_part_3`). The correct chronological order is: `2→3→1`.

A look at the taxonomy of reasoning within SPLICE, it becomes obvious that the task does not only probe one form of reasoning, but a plethora as for a model to succeed sometimes, it may have to have a round table visual reasoning.

3.4 Evaluation Framework

This section is to formally define the experimental framework used to measure and compare the performance of MLLMs on the SPLICE benchmark. The design of this framework, writing the code for the evaluation, managing the model inference pipelines, and collecting and organizing all of the subsequent results used for analysis was my responsibility.

3.4.1 Evaluation Metrics

Given a ground-truth sequence of clip indices y and a model's predicted sequence $\hat{y}$, performance is assessed using four primary metrics. Binary Accuracy is most stringent

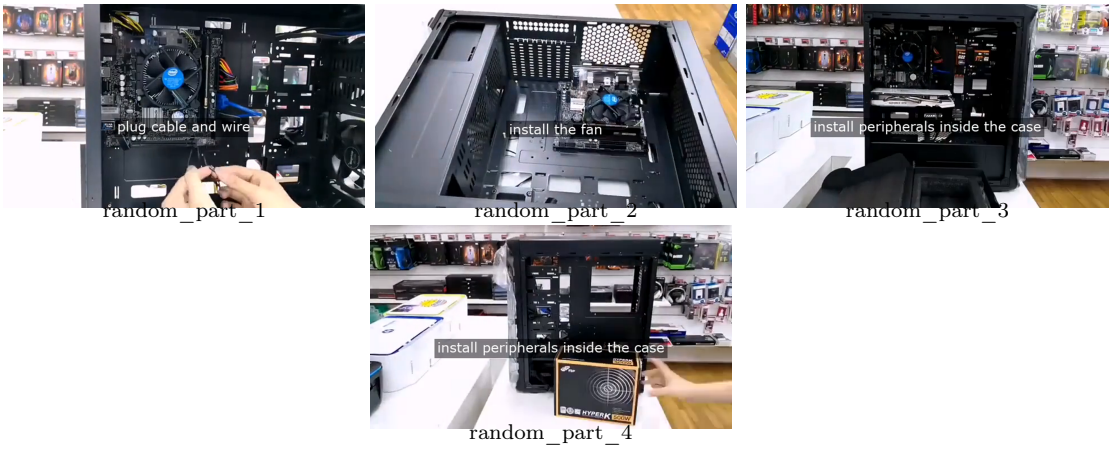

Figure 3.7: Contextual reasoning (Video 140). Shown here in raw order (`random_part_1` ... `random_part_4`). The correct chronological order is: `4→2→1→3`.

and a prediction is considered correct only if the entire sequence exactly matches the ground truth.

For a more granular measure, Position-Wise (Hamming) Accuracy calculates the proportion of correctly placed clips. To award partial score for capturing the core logical flow, the Longest Common Subsequence (LCS) metric measures the longest sequence of clips that appear in the correct relative order. Finally, Edit Distance (Levenshtein Distance) [67] quantifies the degree of error by calculating the minimum number of single-clip edits required to transform the predicted sequence into the ground truth.

$$\text{Binary Accuracy} = \begin{cases} 1 & \text{if } \hat{\mathbf{y}} = \mathbf{y}, \\ 0 & \text{otherwise.} \end{cases}$$

Position-Wise (Hamming) Accuracy The proportion of correctly placed clips:

$$\text{Hamming Accuracy} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(\hat{y}_i = y_i)$$

where $\mathbb{I}(\cdot)$ is the indicator function.

3.4.2 Baselines for Comparison

To properly interpret the performance of advanced MLLMs, their results are contextualized against two crucial baselines. The first, a Random Baseline, establishes the performance floor by calculating the expected scores for a model that is guessing randomly. The second and most important baseline is the Human Baseline. As described in Section 3.3, the aggregated performance of the expert human annotators during the validation protocol serves as this baseline, providing a direct point of comparison to quantify the gap between current machine capabilities and human-level event understanding.

Original Instructional Video

The temporal segmentation process begins with an input video from the COIN dataset, accompanied by ground-truth temporal annotations. Figure 3.8 illustrates the temporal structure of the example video.

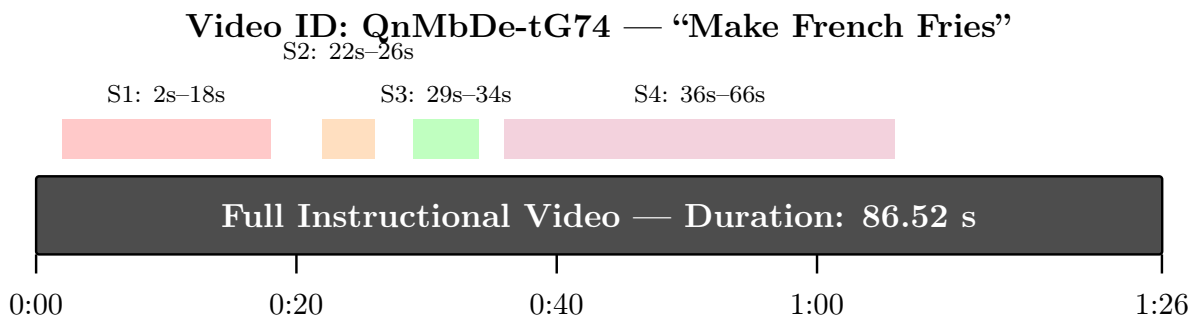

Figure 3.8: Temporal timeline of input video QnMbDe-tG74 with annotated segments.

COIN Ground Truth Temporal Annotations

The COIN dataset provides precise temporal annotations for each instructional step within the video. These annotations serve as ground truth for the segmentation process, ensuring semantic consistency and temporal accuracy.

```
{
  "video_id": "QnMbDe-tG74",
  "domain": "Dish",
  "class": "MakeFrenchFries",
  "duration": 86.517542,
  "annotations": [
    {
      "segment_id": "203",
      "start": 2.0,
      "end": 18.0,
      "label": "cut potato into strips"
    },
    {
      "segment_id": "204",
      "start": 22.0,
      "end": 26.0,
      "label": "soak them in water"
    },
    {
      "segment_id": "205",
      "start": 29.0,
      "end": 34.0,
      "label": "dry strips"
    },
    {
      "segment_id": "206",
      "start": 36.0,
      "end": 66.0,
      "label": "put in the oil to fry"
    }
  ]
}
```

Listing 3.1: COIN annotation structure for video QnMbDe-tG74

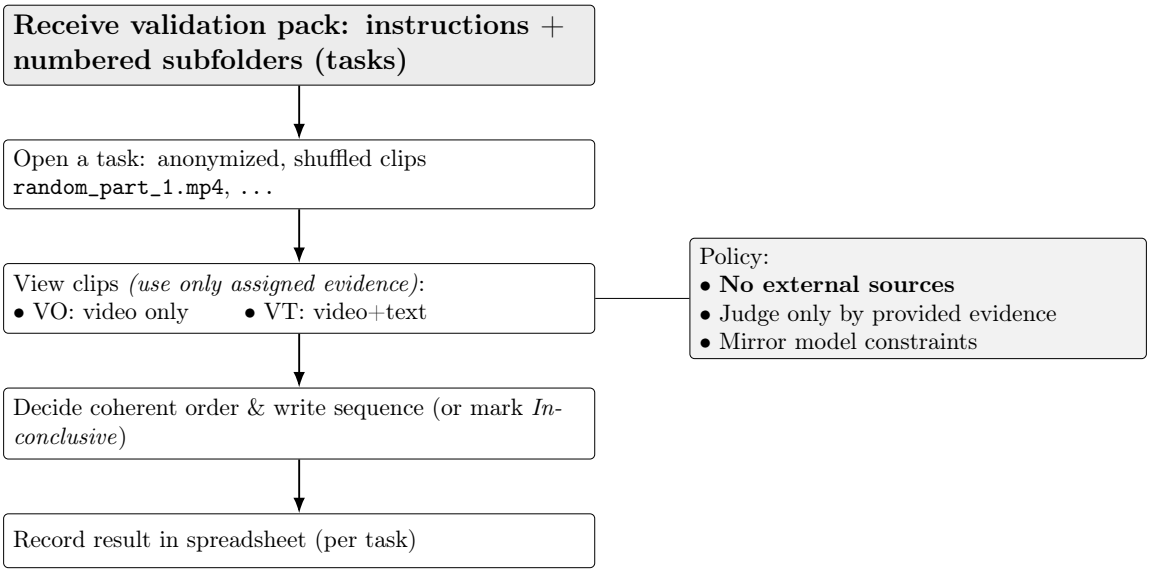

Figure 3.9: Annotator task flow per video task (legible version).

Chapter 4

Experimental FrameWork

Having established the methodological foundation of the SPLICE benchmark in the previous chapter, this chapter details the specific experimental framework designed to implement it. The objective is to create a controlled, rigorous, and reproducible environment for evaluating the capability of Multimodal Large Language Models (MLLMs) on the reordering task of SPLICE. The chapter outlines the four core components of the experimental design, which include the selection and description of the MLLMs under evaluation, the precise zero-shot protocol and prompting strategy, the establishment of essential performance baselines, and, finally, the final evaluation protocol.

4.1 Model Selection and Description

The model selection criteria was extensive, there were many VLMs and MLLMs in the radar to test with SPLICE, however, the major criteria that was non-negotiable was that the models selected must be able to process and coherently reference each individual clip it receives (the model must be able to link all chronological frames from one clip and associate it to that very clip). This was essential, and I did multiple "Sanity Checks" to justify testing a model on splice on this clip-level based re-ordering task. By extension, any model that could not do this was excluded from the test, as this Clip-Frames to Reference mapping is crucial for any model to correctly and internally view multiple frames from one clip as a single entity. Take, for example, in Qwen-2 VL Instruct, it was trained with an "add_vision_id = True" flag that can be toggled on and off to localize the reference of all frames from a single clip. Gemini has referencing enabled by default, and you can simply point the model to a video when processing it via the API.

Another requirement, aside from the Sanity Check, was the ability of the models to easily process videos of long length. Qwen2-VL and Gemini also stood out in this regard. Gemini-Flash models [29] possess a sparse MoE architecture [68], which allows it to process long videos due to its massive context window. It processes each video by default at 1fps, so a 10-second segmented clip would have 10 frames in this scenario. Qwen2-VL-Instruct [30] features a Native Dynamic Resolution and Multimodal Rotary Position embedding, representing a unified paradigm for handling diverse visual data across dynamic resolutions [69]. I tested both the 7B and 72B versions to explore scaling effects.

LLaVA-OneVision [31] implements a highly cost-efficient approach built on the principle of "easy visual task transfer," showcasing the power of a strong LLM backbone and high-quality image-based instruction data. I tested the 72B version. Lastly, the

InternVL2.5 [32] focuses on meticulous co-scaling of all components, including the vision encoder, LLM, and data, to push the performance boundaries of open-source systems.

4.2 Zero-Shot Evaluation Protocol

To avoid task-specific fine-tuning, all models were tested in their out-of-the-box state. The zero-shot evaluation strategy was deliberate for this reason, as testing the out-of-the-box reasoning capabilities was essential. The input for the model was formatted, and the shuffled clips were supplied all at once. A handful of prompts were tested to pick the best that worked across all models at peak performance, as we wanted to have a prompt that was not biased. It was carefully designed to be neutral and to avoid providing any hints.

Sample prompt: *video-only* input

```
A video has been split into len(clips) clips, shuffled randomly.
Your task is to analyze each clip deeply to reorder them into the correct
    temporal sequence.
Focus on:
1) Visual content: examine the actions, transitions, scene details, and
    context within each clip.
Provide the reordered sequence strictly within order tags in this format:
<order>Video X, Video Y, Video Z, ...</order>
```

Sample prompt: *video+text* input

```
A video has been split into len(clips) clips, shuffled randomly.
Your task is to analyze each clip deeply to reorder them into the correct
    temporal sequence.
Focus on:
1) Visual content: examine the actions, transitions, scene details, and
    context within each clip.
2) Temporal logic: identify the logical progression of events based on what
    happens before or after.
3) Annotations: leverage the annotations to infer their proper chronological
     sequence.
Provide the reordered sequence strictly within order tags in this format:
<order>Video X, Video Y, Video Z, ...</order>
```

4.2.1 MLLM Evaluation Methodology

Our systematic framework evaluates Multi-Modal Large Language Models on temporal video understanding, focusing on causality and temporal ordering across diverse architectures including local GPU-based models and cloud API services.

```
procedure EVALUATE_MLLM_TEMPORAL_UNDERSTANDING
begin
    # Load video segments from dataset
```

```
    for each video in dataset do
        load temporal segments in chronological order
        create ground truth sequence labels
        randomly shuffle segment order
    end for

    # Process each shuffled video for evaluation
    for each shuffled video do
        # Prepare inputs based on model type
        if model is local GPU-based then
            extract frames from video segments
            resize and normalize visual inputs
        else if model is cloud API-based then
            remove audio and anonymize files
            upload segments to cloud service
        end if

        # Construct temporal reasoning prompt
        create prompt asking for chronological reordering
        include visual content analysis instructions
        specify output format with order tags

        # Query model for temporal sequence prediction
        send multimodal prompt to MLLM
        receive temporal ordering prediction

        # Extract and validate model response
        parse predicted sequence from the response
        compare against ground truth ordering
        calculate accuracy metrics

        # Save results and clean up resources
        store evaluation results
        clear memory and temporary files
    end for

    # Compute final evaluation metrics
    calculate exact match accuracy
    compute positional accuracy scores
    generate comprehensive evaluation report
end procedure
```

Listing 4.1: MLLM Temporal Video Understanding Evaluation Process

SPLICE[1] **Key Process Components:**

Data Preparation: Video segments are loaded in their original temporal order to establish ground truth, then randomly shuffled to create the temporal reasoning challenge for MLLMs.

Model-Agnostic Processing: The framework adapts to different model architectures - local GPU models require frame extraction and tensor preparation, while cloud API models need file upload and processing management.

Temporal Reasoning Evaluation: Standardized prompts focus models on visual

[1]Samples of the benchmark: https://drive.google.com/file/d/19lvuzTNgQLOkpg9pWLM4mwzK5ZzNHWNv/view?usp=sharing. The full dataset will be hosted once the publication is off peer review., I can also provide it if requested from the hpc cluster

# of Clips	# of Videos	Average Duration (s)
2	1020	46.83
3	1026	53.31
4	734	62.41
5	333	72.67
6	172	67.86
7	96	73.50

Table 4.1: Distribution of videos by the number of clips, with a total of 3,381 videos segmented into 11,423 clips. The average duration per video is reported.

content analysis and causal relationships to determine correct chronological sequence from shuffled segments.

Robust Assessment: Multiple accuracy metrics including exact sequence matching and positional correctness provide comprehensive evaluation of temporal understanding capabilities across diverse MLLM architectures.

Implementation Availability: The complete implementation of this evaluation framework, including inference scripts for Gemini, InternVL, LLaVA, and other MLLMs, is publicly available on **GitHub**[2] to facilitate reproducible research and enable evaluation of additional multimodal models on temporal video understanding tasks.

4.2.2 Testing Settings

Below are the details about test settings of each model:

Qwen2-VL-Instruct Family. Qwen2-VL was tested with both 7B and 72B parameters. The number of frames was set to 1 fps, and the highest image resolution was set to 448 pixels, while the other dimension was automatically adjusted based on the aspect ratio of the input frames.

Gemini-Flash Family. We used Gemini Flash 1.5 and 2.0 (experimental) versions, with the fps set to 1. The model was loaded using the official Google API, and the image resolution was left at the default setting, allowing the model to handle it automatically.

InternVL2.5 Family. InternVL2.5 was tested with the 78B parameters model only. The 8B model did not pass the sanity check. We used the default settings of the uniformal distribution of frames input for each clip and we set it to 16 frames instead of fps.

LlavaOnevision Family LlavaOnevision was tested with 72B parameters. The 7B model did not pass the sanity check. We used the default settings of the uniformal distribution of frames input for each clip and we set it to 16 frames instead of fps.

All of the open source models were used from the Hugging Face library and adopted with the Flash Attention approach. All of these models are tested with two different modalities, vision only, and vision + text. All jobs were submitted to a cluster of A100 and H100 GPUs, which were used interchangeably based on availability. The text only is not accounted for in this thesis.

[2]https://github.com/prokajevo/OoOMLLM

Chapter 5

Results and Discussion

This section presents the core empirical findings from this research on the Evaluation Deficit that currently exists in the evaluation of SOTA Multi-Modal Large Language Models. The section provides quantitative and qualitative analyses to establish the performance landscape by evaluating the most capable models available, and contextualizes them against a rigorous human baseline established by the methodology in Chapter 3 and the experimental setup in Chapter 4, using the curated SPLICE benchmark[11] to probe reasoning across multiple dimensions via the event-based segment reordering task.

5.1 Performance on the SPLICE Benchmark

The primary evaluation was conducted on the full set of 3,381 human-validated video tasks, as described in Chapter 3. The experimental setup was done in two modes: **"Video Only"** and **"Video+Text"** mode. IN the Video Only mode, the MLLMs and humans were tested on pure visual reasoning, and the other mode to assess multimodal integration and the effect of annotation on performance.

5.1.1 Quantitative Analysis

The result presented in Table 5.1 reveals two immediate and critical issues. First, there is a significant performance gap between the best-performing MLLM and humans; this gap is substantially increased when using the Vision Only modality. The best-performing model, Gemini-2.0-Flash-Exp, achieved a Binary Accuracy of 51.08%. In isolation, this may appear promising, considering the random baseline is 21.14%; however, a human performance of 84.86% accuracy highlights the significant divide and underscores the deficit these models face in visual reasoning.

More interestingly, the observation is made when the model is evaluated in the **Video + Text modality**, as this indicates that the model relies heavily on the text, which it is obviously using as a shortcut instead of engaging in actual deep visual reasoning. Gemini 2.0 Flash Exp scored 69.39%, which is over 18 points higher than its Video Only mode of 51.08%.Human performance was statistically unaffected, as it is 84.86% vs. 83.32%, indicating the textual modality was redundant for humans. The results strongly suggest the MLLMs use language shortcuts to complete tasks, relying on their language prior rather than deep visual reasoning. The finding, even on this scale, is consistent with the results from Ko et al.[10] which suggest that MLLMs use language as a shortcut.

Models	Causal vs. Contextual		General Knowledge		Spatial
	Change/Replace	Make	Everyday	Technical	
Number of videos	844	765	342	300	138
Random	18.02	19.14	23.26	22.49	19.19
Human	84.95	86.01	85.38	80.00	76.09
Gemini-1.5-Flash	28.44	61.18	56.14	44.33	24.64
Gemini-2.0-Flash-Exp	32.11	68.37	59.65	47.00	36.96
Qwen2-VL-72B	20.38	36.86	36.55	32.00	21.01

Table 5.1: Models and human video ordering accuracy (%) on task subsets probing different reasoning types

5.1.2 Qualitative Analysis

Model Performance varies significantly across the 12 domains of the SPLICE benchmark. Domains that require spatial reasoning, such as Sport, proved exceptionally challenging for models, while highly temporal and causal domains, like "Dish" or "Pets and fruits", yielded better scores compared to the Sport domain, albeit still sub-human performance. One important note is that Gemini 2.0 Flash Exp maintained stable performance across longer videos, while other models suffered. This suggests that it handles long contexts effectively, and the large context window also plays a role in this.

5.1.3 In-Depth Analysis of Reasoning Dimensions

The most striking result was Causal/Temporal Reasoning vs. Contextual Reasoning; it showed a great performance disparity between "Make" tasks, which are governed by temporal and causal logic, and "Change/Replace" tasks (requiring contextual logic). On "Make" tasks, Gemini-2.0-Flash-Exp scored a relatively strong 68.37%. However, on "Change/Replace" tasks, its performance collapsed to 32.11%.

The failure is strongly correlated to Visual Similarity Bias, where models incorrectly group clips by perceptual similarity, for example, two visually similar "closed door" states, rather than by their distinct contextual roles within the event. The analysis confirms that Gemini-2.0 incorrectly placed such clips adjacent 57.36% of the time, a mistake humans made in only 2.45% of cases. This difference is stark.

Spatial Reasoning

The "Spatial" task subset yielded some of the lowest scores across all models, 36.96% for Gemini-2.0, confirming a profound weakness in Spatial-Temporal reasoning. Models struggle to interpret an object's trajectory through space as a primary indicator of temporal progression.

General Knowledge Reasoning

Lastly, where the performance gap between models and humans was most pronounced on "Specialized/Technical" tasks.

	Vision Only		Vision+Text		Text	
	Binary	Hamming	Binary	Hamming	Binary	Hamming
Random	0.2114	0.3385	0.2114	0.3385	–	–
Human	0.8486	0.8855	0.8332	0.8904	–	–
Qwen2-VL-7B	0.3091	0.4432	0.4354	0.5683	0.3318	0.4924
Qwen2-VL-72B	0.2990	0.4170	0.5708	0.6820	**0.5402**	**0.6907**
Gemini-1.5-Flash	0.4599	0.5825	0.5936	0.7115	0.4642	0.6029
Gemini-2.0-Flash-Exp	**0.5108**	**0.6188**	**0.6939**	**0.7931**	0.5271	0.6652
InternVL2.5-78B	0.2899	0.4243	0.4856	0.6046	0.4768	0.6062
LLaVA-OneVision-72B	0.2260	0.3514	0.4256	0.5597	0.4210	0.5545

Table 5.2: Binary and Hamming accuracy scores for various VLMs across different input modalities: Vision Only, Text Only, and Vision+Text. Human and random baselines are included for comparison.

While Gemini-2.0 scored 59.65% on "Everyday" tasks, its performance dropped to 47.00% on technical ones. Humans, however, showed minimal degradation (85.38% vs. 80.00%). This suggests that model knowledge is not abstract but is heavily correlated with the frequency of concepts in training data, revealing a deficit in deep, generalizable world knowledge, as portrayed in my literature review.

One thing that clearly stood out in the results is how strongly performance depends on the number of clips in a task. The more segments there are to reorder, the harder it gets for both humans and models as shown in Fig 5.1. Still, the difference is in how steep that drop is: humans manage to hold on reasonably well with an almost linear-curve downward, even with seven clips, but models almost collapse to random guessing at that point. This makes sense because the number of possible permutations grows factorially, so the task quickly becomes much more complex.

What's important here is that the models seem especially brittle when the sequence length increases, which suggests they struggle to maintain coherent temporal reasoning over longer horizons, while humans can still rely on commonsense and causal understanding to piece things together.

5.2 Discussion

The empirical results presented above paint a consistent picture: while MLLMs have made remarkable progress, their ability to reason about the structure of visual events remains shallow and brittle compared to humans. This section discusses the broader implications of these findings. The "Shallow Reasoner" Hypothesis: The collective evidence—the reliance on text shortcuts, the catastrophic failure on contextual tasks due to Visual Similarity Bias, and the poor performance on spatial and specialized tasks—supports the hypothesis that MLLMs operate as "shallow reasoners." They excel at pattern matching and leveraging statistical priors but lack the robust, causal, and contextual world models that are the hallmark of human cognition. Implications for Multimodal Evaluation: Our findings highlight a critical "evaluation crisis" in the field. The success of SPLICE's generative reordering task in exposing these deep failures demonstrates that many existing benchmarks, which often focus on classification or are susceptible to linguistic shortcuts, may provide an inflated view of model capabilities.

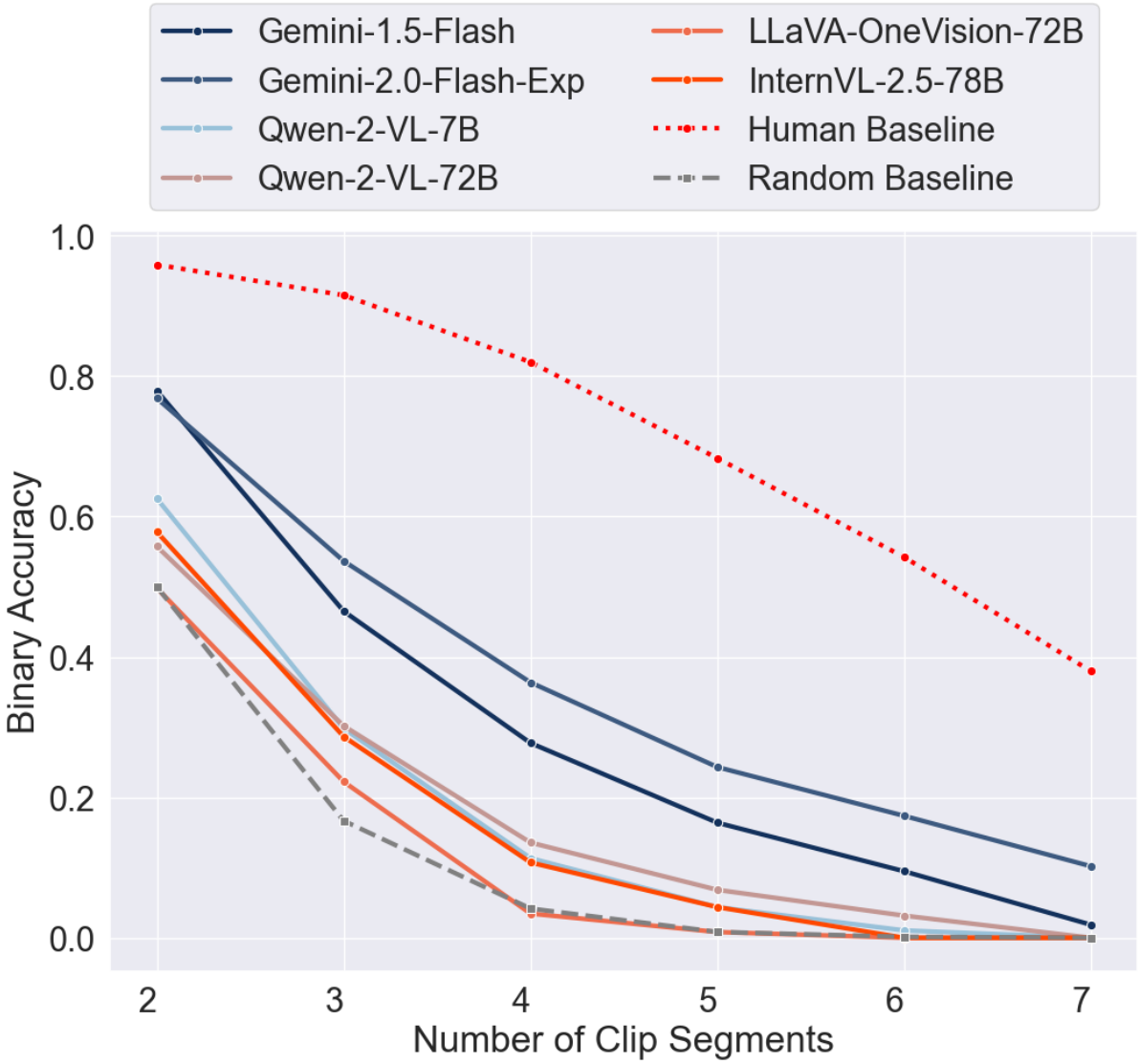

Figure 5.1: Binary accuracy versus the number of clips (2–7), comparing various state-of-the-art VLMs against human and random baselines

Also, I assert that further progress requires a shift toward more diagnostic benchmarks that probe for structural understanding and it is also notable that humans achieved 85% accuracy, not 100%. This ceiling reflects the inherent difficulty of the task. An analysis of human errors reveals they primarily fall into three categories: genuine ambiguity in the visual data, lack of domain-specific knowledge for highly technical tasks, and working memory limitations on the most complex, long-form sequences. Critically, our findings show that models rarely, if ever, succeed on the sequences where humans fail, indicating that current AI capabilities are a brittle subset of human reasoning, not a complementary intelligence.

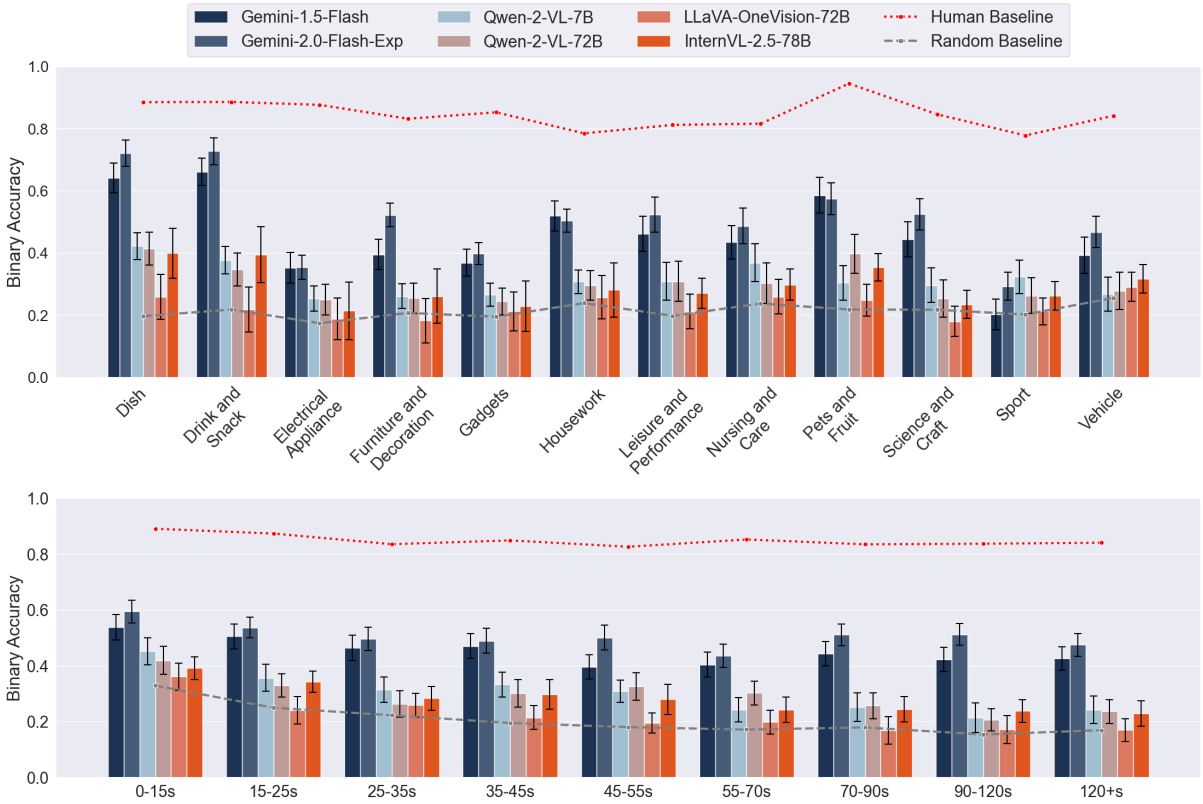

Figure 5.2: Binary accuracy performance of various state-of-the-art VLMs across different domains (top) and video durations (bottom), compared to a human baseline (red dashed line) and a weighted random baseline (gray dashed line). Error bars represent the 95% confidence interval (CI).

Chapter 6

Conclusion and Future Work

The reordering task and the human-curate Benchmark in this thesis have crystallized a central point: MLLMs can approximate temporal coherence; however, they still lack a robust, event-level model of process. The SPLICE benchmark makes this visible; unlike humans, who performed at a high level of accuracy even when no textual cue is given, in these tasks, MLLMs still lag behind, even the SOTA.

Unlike humans, models rely on textual cues and utilize language to make shortcuts. Because SPLICE is human-curated and validated for solvability, the observed gap is unlikely to be a labeling artifact, as it is better explained by the absence of contextual and global structural reasoning in current architectures.

There are four concrete directions that, if taken together, would target the bottlenecks and even make SPLICE a much extended diagnostic tool. First, the inclusion of audio will be a good direction to follow, as newer models are now emerging with audio support. It would be interesting to see if the same issue of Textual Cue shortcuts manifests in the audio experiment. Secondly, implementing a pairwise precedence prediction, as a Kendall evaluation, would let models expose uncertainty with single-guess collapse, and this should penalize "look-alike" mistakes. Additionally, a long-context ablation would also be beneficial, allowing us to probe context management from raw perception.

It is essential to note that pretraining via Order as Supervision with SPLICE in a self-supervised setup aims to investigate whether the human gap can be bridged or if such a model can learn to reason in a multidimensional framework of reasoning.

In conclusion, SPLICE, along with the thesis framework surrounding it, provides us with a sharper instrument for evaluating event understanding in MLLMs. The immediate next steps are to (i) tighten the structural bias of models toward process (graphs, constraints, hard negatives), (ii) probe fusion with audio and longer contexts in a controlled way, and (iii) broaden domains while preserving rigorous human validation. If those three threads move together, I would expect less reliance on language priors and more genuine multimodal understanding of how events unfold.

Bibliography

[1] Ashish Vaswani et al. "Attention is all you need". In: *Advances in Neural Information Processing Systems* 30 (2017), pp. 5998–6008.

[2] Tom Brown et al. "Language models are few-shot learners". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 1877–1901.

[3] Aakanksha Chowdhery et al. "PaLM: Scaling language modeling with pathways". In: *Journal of Machine Learning Research* 23.240 (2022), pp. 1–113.

[4] Jean-Baptiste Alayrac et al. "Flamingo: a visual language model for few-shot learning". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 23716–23736.

[5] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models". In: *arXiv preprint arXiv:2301.12597* (2023).

[6] Jeffrey M Zacks and Barbara Tversky. "Event structure in perception and conception". In: *Psychological Bulletin* 127.1 (2001), pp. 3–21.

[7] Christopher Baldassano, Uri Hasson, and Kenneth A Norman. "Discovering event structure in continuous narrative perception and memory". In: *Neuron* 95.3 (2017), pp. 709–721.

[8] Jeffrey M Zacks, Nicole K Speer, Khena M Swallow, Todd S Braver, and John R Reynolds. "Event perception: A mind-brain perspective". In: *Psychological Bulletin* 133.2 (2007), pp. 273–293.

[9] Robert Geirhos et al. "Shortcut learning in deep neural networks". In: *Nature Machine Intelligence* 2 (2020), pp. 665–673.

[10] Dohwan Ko, Ji Soo Lee, Woo-Young Kang, Byungseok Roh, and Hyunwoo J. Kim. "Large Language Models are Temporal and Causal Reasoners for Video Question Answering". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2023. URL: https://aclanthology.org/2023.emnlp-main.261/.

[11] Mohamad Ballout, Wilfred Okajevo, Seyedalireza Yaghoubi, Muhammad Abdelmoneim, Julius Mayer, and Elia Bruni. "Can you SPLICE it together? A Human Curated Benchmark for Probing Visual Reasoning in VLMs". In: *The 2025 Conference on Empirical Methods in Natural Language Processing*. To appear. Association for Computational Linguistics, 2025.

[12] Hossein Mobahi, Ronan Collobert, and Jason Weston. "Deep learning from temporal coherence in video". In: *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 737–744.

[13] Ishan Misra, C. Lawrence Zitnick, and Martial Hebert. "Shuffle and learn: unsupervised learning using temporal order verification". In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*. Springer, 2016, pp. 527–544.

[14] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. "UCF101: A dataset of 101 human actions classes from videos in the wild". In: *arXiv preprint arXiv:1212.0402* (2012).

[15] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. "HMDB: a large video database for human motion recognition". In: *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 2556–2563.

[16] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. "Self-supervised spatiotemporal learning via video clip order prediction". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 10334–10343.

[17] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. "Learning spatiotemporal features with 3d convolutional networks". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 4489–4497.

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

[19] Debidatta Dwibedi, Pierre Sermanet, and Jonathan Tompson. "Temporal reasoning in videos using convolutional gated recurrent units". In: *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*. IEEE, 2017, pp. 1269–1277.

[20] Raghav Goyal et al. "The "something something" video database for learning and evaluating visual common sense". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 5842–5850.

[21] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. "Temporal relational reasoning in videos". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 803–818.

[22] Ankit Sharma, Piyush Sharma, and David Jacobs. "Learning video representations from textual web supervision". In: *arXiv preprint arXiv:2007.14937* (2020).

[23] Hassan Akbari, Li Yuan, et al. "VATT: Transformers for multimodal self-supervised learning from raw video, audio and text". In: *Advances in Neural Information Processing Systems*. 2021.

[24] Yunzhu Li, Antonio Torralba, Anima Anandkumar, Dieter Fox, and Animesh Garg. "Causal discovery in physical systems from videos". In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020, pp. 9180–9192.

[25] Tieyuan Chen et al. "MECD: Unlocking Multi-Event Causal Discovery in Video Reasoning". In: *arXiv preprint arXiv:2409.17647* (2024).

[26] Yongxin Guo, Jingyu Liu, Mingda Li, Qingbin Liu, Xi Chen, and Xiaoying Tang. "TRACE: Temporal Grounding Video LLM via Causal Event Modeling". In: *The Thirteenth International Conference on Learning Representations*. 2025. URL: https://openreview.net/forum?id=14fFV0chUS.

[27] Alexey Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[28] Ruotong Liao et al. "VideoINSTA: Zero-shot Long Video Understanding via Informative Spatial-Temporal Reasoning with LLMs". In: *Findings of the Association for Computational Linguistics: EMNLP 2024*. Miami, Florida, USA: Association for Computational Linguistics, 2024, pp. 6577–6602. DOI: 10.18653/v1/2024.findings-emnlp.384.

[29] Rohan Anil et al. "Gemini: a family of highly capable multimodal models". In: *arXiv preprint arXiv:2312.11805* (2023).

[30] Peng Wang et al. "Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution". In: *arXiv preprint arXiv:2409.12191* (2024).

[31] Bo Li et al. "LLaVA-OneVision: Easy visual task transfer". In: *arXiv preprint arXiv:2408.03326* (2024).

[32] Zhe Chen et al. "InternVL 2.5: Multimodal large language models series to bridge the gap between academic research and industrial practice". In: *arXiv preprint arXiv:2412.05304* (2024).

[33] Joao Carreira and Andrew Zisserman. "Quo vadis, action recognition? a new model and the kinetics dataset". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 6299–6308.

[34] Jing Ji et al. "A survey on video understanding with deep learning". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).

[35] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. "ActivityNet: A large-scale video benchmark for human activity understanding". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 961–970.

[36] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. "Hollywood in homes: Crowdsourcing data collection for activity understanding". In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*. Springer, 2016, pp. 510–526.

[37] Jr-Jen Chen, Yu-Chien Liao, Hsi-Che Lin, Yu-Chu Yu, Yen-Chun Chen, and Yu-Chiang Frank Wang. "ReXTime: A Benchmark Suite for Reasoning-Across-Time in Videos". In: *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 2024. URL: https://openreview.net/forum?id=4Vhc7uPHjn.

[38] Aakanksha Naik, Luke Breitfeller, and Carolyn Rose. "TDDiscourse: A Dataset for Discourse-Level Temporal Ordering of Events". In: *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*. Stockholm, Sweden: Association for Computational Linguistics, 2019, pp. 239–249. DOI: 10.18653/v1/W19-5929.

[39] Luowei Zhou, Chenliang Xu, and Jason J. Corso. "Towards automatic learning of procedures from web instructional videos". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1. 2018, pp. 7590–7598.

[40] Bahare Fatemi et al. “Test of Time: A Benchmark for Evaluating LLMs on Temporal Reasoning”. In: *The Thirteenth International Conference on Learning Representations*. 2025. URL: https://openreview.net/forum?id=44CoQe6VCq.

[41] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. “Next-QA: Next phase of question-answering to explaining temporal actions”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 9777–9786.

[42] Bo Wu et al. “STAR: A benchmark for situated reasoning in real-world videos”. In: *Advances in Neural Information Processing Systems*. Vol. 34. 2021, pp. 18028–18041.

[43] Aaron Foss, Chloe Evans, Sasha Mitts, Koustuv Sinha, Ammar Rizvi, and Justine T. Kao. *CausalVQA: A Physically Grounded Causal Reasoning Benchmark for Video Models*. arXiv preprint arXiv:2506.09943. Available at https://arxiv.org/abs/2506.09943. 2025.

[44] Hang Du et al. “Uncovering {What, Why and How}: A Comprehensive Benchmark for Causation Understanding of Video Anomaly (CUVA)”. In: *CVPR Workshops*. ArXiv:2405.00181; https://arxiv.org/abs/2405.00181. 2024.

[45] Haoyu Wang, Fengze Liu, Jiayao Zhang, Dan Roth, and Kyle Richardson. “Event Causality Identification with Synthetic Control”. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 1725–1737. DOI: 10.18653/v1/2024.emnlp-main.103. URL: https://aclanthology.org/2024.emnlp-main.103/.

[46] Nasrin Mostafazadeh et al. *GLUCOSE: GeneraLized and COntextualized Story Explanations*. 2020. arXiv: 2009.07758 [cs.CL]. URL: https://arxiv.org/abs/2009.07758.

[47] Yuan Liu et al. *MMBench: Is Your Multi-modal Model an All-around Player?* 2024. arXiv: 2307.06281 [cs.CV]. URL: https://arxiv.org/abs/2307.06281.

[48] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. “Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models”. In: *Proceedings of ACL 2024 (Long Papers)*. Association for Computational Linguistics, 2024, pp. 12585–12602. DOI: 10.18653/v1/2024.acl-long.679. URL: https://aclanthology.org/2024.acl-long.679/.

[49] Junjie Zhou et al. *MLVU: Benchmarking Multi-task Long Video Understanding*. CVPR Workshop / ArXiv preprint arXiv:2406.04264. Available at https://arxiv.org/abs/2406.04264. 2024.

[50] Mu Cai et al. *TemporalBench: Benchmarking Fine-grained Temporal Understanding for Multimodal Video Models*. 2024. arXiv: 2410.10818 [cs.CV]. URL: https://arxiv.org/abs/2410.10818.

[51] Yogesh Kumar. “VideoLLM Benchmarks and Evaluation: A Survey”. In: *arXiv preprint arXiv:2505.03829* (2025). URL: https://arxiv.org/abs/2505.03829.

[52] Munan Ning et al. “Video-Bench: A Comprehensive Benchmark and Toolkit for Evaluating Video-based Large Language Models”. In: *arXiv preprint arXiv:2311.16103* (2023). URL: https://arxiv.org/abs/2311.16103.

[53] Zongjie Li et al. *VRPTEST: Evaluating Visual Referring Prompting in Large Multimodal Models*. 2023. arXiv: 2312.04087 [cs.CV]. URL: https://arxiv.org/abs/2312.04087.

[54] Zheng Zhang, Yeyao Ma, Enming Zhang, and Xiang Bai. *PSALM: Pixelwise SegmentAtion with Large Multi-Modal Model*. 2024. arXiv: 2403.14598 [cs.CV]. URL: https://arxiv.org/abs/2403.14598.

[55] Kexin Yi et al. "CLEVRER: Collision events for video representation and reasoning". In: *arXiv preprint arXiv:1910.01442* (2019).

[56] Rohit Girdhar and Deva Ramanan. "CATER: A diagnostic dataset for compositional actions and temporal reasoning". In: *arXiv preprint arXiv:1910.04744* (2019).

[57] Hao Du, Bo Wu, Yan Lu, and Zhendong Mao. "SVLTA: Benchmarking Vision-Language Temporal Alignment via Synthetic Video Situation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2025. URL: https://svlta-ai.github.io/SVLTA.

[58] Christian Schuldt, Ivan Laptev, and Barbara Caputo. "Recognizing human actions: a local SVM approach". In: *Proceedings of the 17th International Conference on Pattern Recognition (ICPR)*. Vol. 3. 2004, pp. 32–36. DOI: 10.1109/ICPR.2004.1334462.

[59] Frederik Ebert, Chelsea Finn, Alex X. Lee, and Sergey Levine. "Self-supervised visual planning with temporal skip connections". In: *Conference on Robot Learning (CoRL)*. 2017, pp. 344–356. URL: https://proceedings.mlr.press/v78/ebert17a.html.

[60] Catalin Ionescu, Fuxin Li, and Cristian Sminchisescu. "Latent structured models for human pose estimation". In: *IEEE International Conference on Computer Vision (ICCV)*. 2011, pp. 2220–2227. DOI: 10.1109/ICCV.2011.6126514.

[61] Sebastian Stein and Stephen J. McKenna. "Combining embedded classifiers for gesture segmentation". In: *2013 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*. 2013, pp. 1–6. DOI: 10.1109/ICMEW.2013.6618355.

[62] Hilde Kuehne, Ali Arslan, and Thomas Serre. "The Breakfast dataset and its evaluation". In: *Image and Vision Computing* 39 (2015), pp. 44–55. DOI: 10.1016/j.imavis.2015.02.009.

[63] Dima Damen, Theophanis Leelasawassuk, Oliver Haines, Andrew Calway, and Walterio Mayol-Cuevas. "You-Do, I-Learn: Discovering Task Relevant Objects and their Modes of Interaction from Multi-User Egocentric Video". In: *Proceedings of the British Machine Vision Conference (BMVC)*. 2014, pp. 1–11. URL: https://bmvc2014.swansea.ac.uk/proceedings/.

[64] Alireza Fathi, Yin Li, and James M. Rehg. "Learning to recognize daily actions using gaze". In: *European Conference on Computer Vision (ECCV)*. Springer, 2012, pp. 314–327. DOI: 10.1007/978-3-642-33718-5_23.

[65] Joao Carreira and Andrew Zisserman. "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 6299–6308. DOI: 10.1109/CVPR.2017.502. URL: https://openaccess.thecvf.com/content_cvpr_2017/html/Carreira_Quo_Vadis_Action_CVPR_2017_paper.html.

[66] Yu Tang et al. “COIN: A Large-scale Dataset for Comprehensive Instructional Video Analysis”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019. URL: https://openaccess.thecvf.com/content_CVPR_2019/html/Tang_COIN_A_Large-Scale_Dataset_for_Comprehensive_Instructional_Video_Analysis_CVPR_2019_paper.html.

[67] Vladimir I. Levenshtein. “Binary codes capable of correcting deletions, insertions, and reversals”. In: *Soviet Physics Doklady* 10.8 (1966), pp. 707–710.

[68] William Fedus, Barret Zoph, and Noam Shazeer. “Switch transformer: Scaling to trillion parameter models with simple and efficient sparsity”. In: *Journal of Machine Learning Research* 23.120 (2022), pp. 1–39.

[69] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. “RoFormer: Enhanced transformer with rotary position embedding”. In: *Neurocomputing* 568 (2024), p. 127063.

Supplemental Materials and Methods

.1 Video Dataset Statistics

Table 1 provides statistics on the segmented video dataset, detailing how videos are divided into segments and their distribution across different segment counts.

Segments	Videos	Clips	Mean Time (s)	Std Dev	(2, 35]	(35, 68]	(68, 100]	(100, 133]	(133, 166]	(166, 198]	(198, 330]
2	1020	2040	46.84	40.75	531	245	128	65	35	12	4
3	1026	3078	53.32	42.68	434	309	142	79	36	19	7
4	734	2936	62.41	40.76	209	260	146	70	32	13	4
5	333	1665	72.67	43.20	62	123	69	46	21	10	2
6	172	1032	67.86	37.49	38	56	45	22	9	2	0
7	96	672	73.50	38.28	19	26	30	14	5	2	0

Table 1: This table summarizes the distribution of videos based on their segmentation. It includes the number of segments(2-7), total videos per segment number, total clips, mean duration (seconds), and standard deviation. The rightmost columns show the distribution of videos across predefined video duration intervals, providing insights into the dataset's temporal structure for event ordering analysis.

.2 Instructions for Annotators

Instruction Brief Task: Reorder the video parts for each folder into their correct sequence. Steps: Download and Open the Folder assigned to you: You will receive a folder containing several subfolders, each labeled with a unique number (e.g., 1, 2, 3, etc.). Each subfolder corresponds to a video task with shuffled parts. View the Video Parts: Inside each subfolder, you will find video parts named random_part_1.mp4, random_part_2.mp4, etc. These parts contain embedded labels as secondary context for your understanding of the video context. Reorder the Parts: Watch each video part carefully. Determine the correct sequence of these parts based on the visual and textual cues. Write down the sequence in the format: Folder Number: Correct Order (e.g., 1: random_part_3, random_part_1, random_part_2). For simplicity use [2, 3, 4, 5, 1], where each number represents the Random number video. Use "unk" in these cases:

1- Repeated instructions: If the video contains two separate instances of the same instruction.

2- Continuous actions without sufficient context: An action extends across multiple clips with insufficient background information to establish a clear sequence.

3- Unrelated actions: The video includes unrelated actions with no contextual clues to establish order.

Submit Your Results: Compile the correct order for all folders in the attached spreadsheet Use "unk" for any task sample you believe makes no sense or as discussed during the meeting, Notes: Do not use any external sources Complete all tasks to the best of your ability.

.3 Additional Results

Table 2: Video Ordering Accuracy of Electrical Appliance domain for sub-domains that include change/replace compared to others that dont

Models	Accuracy (%)
Group: change/replace, 328 videos	
Random	15.02
Human	88.11
Gemini-1.5-Flash	26.22
Gemini-2.0-Flash-Exp	26.83
Qwen2-VL-72B	21.04
Group: other, 157 videos	
Random	22.68
Human	86.62
Gemini-1.5-Flash	54.14
Gemini-2.0-Flash-Exp	53.50
Qwen2-VL-72B	33.12

.4 Additional Metrics and Full Results

Each video in the dataset is segmented into clips based on the original COIN dataset's step localization. These clips are then randomly shuffled and renamed as $C = \{c_1, c_2, \ldots, c_n\}$. Depending on the modality being tested, the model receives either video-only input or a combination of video and annotations (short textual descriptions of events in the video). The model's task is to predict the correct permutation of the clip order. Let $\mathbf{y} = [y_1, y_2, \ldots, y_n]$ denote the ground-truth sequence of clip indices, and $\hat{\mathbf{y}} = [\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n]$ represent the predicted sequence. For n clips, there exist $n!$ possible permutations. The models are evaluated on four metrics:

Binary Accuracy. The prediction is correct only if the entire sequence matches the ground truth:

$$\text{Binary Accuracy} = \begin{cases} 1 & \text{if } \hat{\mathbf{y}} = \mathbf{y}, \\ 0 & \text{otherwise.} \end{cases}$$

Position-Wise (Hamming) Accuracy The proportion of correctly placed clips:

$$\text{Hamming Accuracy} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(\hat{y}_i = y_i)$$

where $\mathbb{I}(\cdot)$ is the indicator function.

Longest Common Subsequence (LCS). The LCS measures the longest sequence of elements appearing in the same relative order in both $\mathbf{y}$ and $\hat{\mathbf{y}}$. Let $c(i, j)$ denote the length of the LCS for substrings $\mathbf{y}_{1:i}$ and $\hat{\mathbf{y}}_{1:j}$:

$$c(i,j) = \begin{cases} 0 & i = 0 \text{ or } j = 0, \\ c(i-1, j-1) + 1 & y_i = \hat{y}_j, \\ \max\{c(i-1, j), c(i, j-1)\} & \text{otherwise.} \end{cases}$$

The LCS ratio normalizes this value:

$$\text{LCS Ratio} = \frac{\text{LCS Length}}{n}$$

Edit Distance (Levenshtein Distance). The minimum number of insertions, deletions, or substitutions required to transform $\hat{\mathbf{y}}$ into $\mathbf{y}$. Define a matrix D where $D(i, j)$ is the edit distance between $\mathbf{y}_{1:i}$ and $\hat{\mathbf{y}}_{1:j}$:

$$D(i, 0) = i, \quad D(0, j) = j \quad \text{(boundary conditions)},$$

$$D(i, j) = \begin{cases} D(i-1, j-1) & y_i = \hat{y}_j, \\ 1 + \min \begin{pmatrix} D(i-1, j) \\ D(i, j-1) \\ D(i-1, j-1) \end{pmatrix} & \text{otherwise.} \end{cases}$$

The final edit distance is $D(n, n)$.