Deep Temporal Reasoning in Video Language Models: A Cross-Linguistic Evaluation of Action Duration and Completion through Perfect Times

Olga Loginova
University of Trento
olga.loginova@unitn.it

Sofía Ortega Loguinova
Maastricht University
s.ortegaloguinova@student.maastrichtuniversity.nl

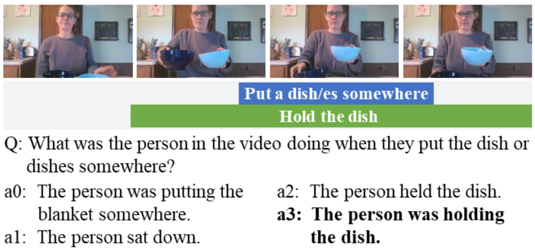
Abstract

Human perception of events is intrinsically tied to distinguishing between completed (perfect and telic) and ongoing (durative) actions, a process mediated by both linguistic structure and visual cues. In this work, we introduce the **Perfect Times** dataset, a novel, quadrilingual (English, Italian, Russian, and Japanese) multiple-choice question-answering benchmark designed to assess video-language models (VLMs) on temporal reasoning. By pairing everyday activity videos with event completion labels and perfectivity-tailored distractors, our dataset probes whether models truly comprehend temporal dynamics or merely latch onto superficial markers. Experimental results indicate that state-of-the-art models, despite their success on text-based tasks, struggle to mirror human-like temporal and causal reasoning grounded in video. This study underscores the necessity of integrating deep multimodal cues to capture the nuances of action duration and completion within temporal and causal video dynamics, setting a new standard for evaluating and advancing temporal reasoning in VLMs.

1 Introduction

Understanding how events unfold in time requires a detailed analysis of their both sequential and causal relationships. Sequential events are not simply arranged chronologically; rather, one event often triggers the next upon reaching its completion. Moens and Steedman (1988) highlighted that human memory organizes actions based on contingency, their cause-effect relationships, where a cause reaches its culmination before triggering an effect. This causal linkage is encoded in language through grammatical time and, critically, through aspect, specifically, perfectivity and its semantic correlate, telicity. Telicity refers to a property of verb phrases that denotes a definitive endpoint (e.g., "to put something somewhere"), while atelic (or durative) expressions (e.g., "to hold something") lack such clear termination. This property is deeply woven into the language (Tenny, 1994).

Temporal relations, therefore, are not conveyed solely through grammatical time; they are also robustly signaled by aspect. Yet, language often omits critical information that visual modality can provide. Observing how events unfold in time, like in videos, offers dynamic cues, such as key frame transitions and motion patterns, that decisively indicate whether an action has been completed or is still ongoing. This multimodal integration is indispensable for resolving ambiguities inherent in linguistic descriptions of events.



Q: What was the person in the video doing when they put the dish or dishes somewhere?

a0: The person was putting the blanket somewhere.

a1: The person sat down.

a2: The person held the dish.

a3: The person was holding the dish.

Figure 1: An example from the Perfect Times dataset illustrating the interaction between a completed action (*to put a dish or dishes somewhere*) and a durative action (*to hold the dish*). The blue and green stripes indicate temporal progression. The correct answer highlighted in bold.

To evaluate these phenomena, we introduce the **Perfect Times** dataset, a multiple-choice question-answering (MCQA) benchmark designed specifically for video-language models (VLMs). Our questions are constructed as complex sentences that juxtapose two actions, thereby probing the interplay between verb forms and temporal conjunctions in main and dependent clauses. Twelve carefully designed templates systematically cover all possible temporal relations between the two ac-

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 20472–20502
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

tions in accordance with the universal Allen's interval algebra (Allen, 1984). Moreover, our study adopts a quadrilingual approach (English, Italian, Russian, and Japanese) to capture the diverse ways in which languages encode time and aspect. For instance, Russian frequently encodes perfectivity at the lexical level, making markers of completion predefined in verbs, while Italian involves a sophisticated interaction between time, mood, and aspect, as it has more branched *concordanza dei tempi* (the sequence of tenses) than English. In Japanese, the commonplace understated and ambiguous linguistic encoding necessitates stronger reliance on visual context to disambiguate temporal relations.

By combining visual data with linguistically complex questions, our work addresses the following questions:

- How do VLMs leverage linguistic markers and visual cues to distinguish sequential from simultaneous actions?
- Are there cross-linguistic differences in the interpretation of temporal relations based on grammatical encoding of perfectivity?
- How closely do model predictions align with human understanding of event completion expressed in grammatical time and aspect?

Figure 1 presents an example of an English question-answer (QA) pair from the Perfect Times dataset. In addition to determining the correct action to answer the question in the video, the model must catch the matching forms of semantic or grammatical aspect in the question and the answer option. For human native speakers, making such comparisons is not difficult, but none of the tested state-of-the-art models reach the 50% accuracy threshold.

Our comprehensive, cross-linguistic approach thus aims to set a new benchmark for evaluating multimodal temporal reasoning in VLMs[1].

2 Related Works

2.1 Cognitive Approach

Several studies use the visual world paradigm to examine how aspect influences real-time language comprehension. Foppolo et al. (2021) show that visual and linguistic signals jointly shape the interpretation of completed actions, guiding anticipatory eye movements through verb aspect and visual cues. Similarly, Foppolo et al. (2016) find that Italian adults rapidly focus on images of finished events upon hearing the corresponding verb. In an eye-tracking study, Bosch et al. (2021) report that while Italian children use verb semantics and aspect to anticipate outcomes, their processing of aspectual information lags behind basic lexical semantics. van Hout (2008) challenges uniformity in aspect acquisition by demonstrating that Italian children acquire perfectivity later than their Polish and Dutch peers. Minor et al. (2022) further show that perfective and imperfective aspects influence listeners' expectations differently across Russian, Spanish, and English, with Russian perfectives strongly indicating completion and the English simple past being less reliable. Finally, Chang et al. (2023) provide evidence from animation experiments that visual cues of goal information affect the choice of past versus progressive verb forms in Japanese. These findings contributing to cognitive science illustrate that VLMs may have greater potential to replicate human-like language processing than text-only large language models (LLMs).

2.2 Time and Aspect in Transformers

Zhao et al. (2021) compared aspect interpretation in humans and transformers using verb tenses and resultative structures. Humans responded to all cues, while transformers were sensitive to explicit telicity but struggled with resultatives. Lombardi and Lenci (2023) found that the transformer-based (Vaswani et al., 2017) Italian model GilBERTo performed similarly to humans on inherently telic and atelic verbs, yet had difficulty with context-dependent verbs, suggesting its limitations in handling nuanced larger temporal contexts. Metheniti et al. (2022) showed that transformer-based models classify activities by duration and perfectivity in English and French with over 80% accuracy.

However, all these studies are on text-only LLMs, and the experiments lack a dedicated multimodal benchmark that fully captures all possible temporal relations grounded in the physical world.

2.3 Action Recognition and MCQA for Video

Numerous datasets offer short video clips for action recognition (Kay et al., 2017; Liu et al., 2021; Heilbron et al., 2015; Damen et al., 2020; Sig-

[1] The dataset and codes are available at https://github.com/ologin/PerfectTimes

Table 1: Examples of questions and answers in Perfect Times generated by temporal and aspectual templates with respect to the telicity markers (t: telic, a: atelic). MCA: main clause action, DCA: dependent clause action.

Template	MCA	DCA	Question	Answer
			Precedence	
3	t	a	What had the person in the video done before holding the sandwich?	The person had opened the refrigerator.
4	t	t	What had the person in the video done before the other person walked through a doorway?	The person had closed the laptop.
6	a	a	What had the person in the video been doing before playing with a phone or camera?	The person had been watching the television.
7	a	t	What had the person in the video been doing before turning off the light?	The person had been playing with a phone or camera.
			Succession	
1	t	a	What did the person in the video do after tidying up the table?	The person put the food somewhere.
2	t	t	What did the person in the video do after they had opened the box?	The person put the bag somewhere.
10	a	a	What was the person in the video doing after sitting in a chair?	The person was sitting on the floor.
11	a	t	What was the person in the video doing after taking the cup from somewhere?	The person was drinking from the cup.
9	a	t	What was the person in the video doing when the other person closed the door?	The person was watching the television.
			Simultaneity	
5	t	a	What did the person in the video do while the other person was sitting on the floor?	The person opened the door.
8	a	a	What was the person in the video doing while holding the book?	The person was holding the bag.
12	t	t	What did the person in the video do when they grasped onto a doorknob?	The person opened the door.

urdsson et al., 2016). Kinetics (Kay et al., 2017) provides general action recognition (~10-second clips); FineAction (Liu et al., 2021) and ActivityNet (Heilbron et al., 2015) deliver fine-grained temporal annotations. EPIC-Kitchens (Damen et al., 2020) and Charades (Sigurdsson et al., 2016) target specific scenarios for classification and localization.

These datasets lay the foundation for the video MCQA datasets. Templated questions appear in MSVD-QA and MSRVTT-QA (Xu et al., 2017), STVQ (Jang et al., 2019), and STAR (Wu and Yu, 2021), while manually annotated formats are used in ActivityNet-QA (Yu et al., 2019), TVQA (Lei et al., 2018), NExT-QA (Xiao et al., 2021), Charades-SRL-QA (Sadhu et al., 2021), and Action Genome (Ji et al., 2019). As causal and temporal questions constitute only a subset of these benchmarks, temporal relations are often limited to simple sequential "before/after" or generic "when" questions. This "before/after" focus is also prevalent in video instruction datasets, which mainly assess the understanding of instructional step order and restrict temporal relations to sequence by design. Examples of such datasets comprise HowTo1Million (Miech et al., 2019), InstructionBench (Wei et al., 2025), and YouCookQA (Wang et al., 2021).

Causal and temporal questions are frequently embedded within broader datasets aimed at evaluating diverse reasoning skills. Such datasets include Video-MME (Fu et al., 2024) and Perception Test (Pătrăucean et al., 2023), which span a variety of reasoning types, as well as TVBench (Cores et al., 2024), TemporalBench (Cai et al., 2024), and TempCompass (Liu et al., 2024), which assess models' grasp of temporal dynamics through complex, though not exclusively temporal, questions.

Moreover, most existing datasets are primarily in English and do not adequately address temporal dynamics from a linguistic perspective.

Our multilingual dataset addresses these gaps. It consists of Charades videos annotated with Action Genome action classes and applies STAR-inspired temporal templates that systematically incorporate aspect, a crucial linguistic dimension previously overlooked. Grounded in Allen's interval algebra and informed by linguistic theory, Perfect Times is the first to systematically align temporal video dynamics with linguistic aspect.

2.4 Video Language Models

Video Language Models integrate visual and textual processing through three key components: a pre-trained visual encoder, a pre-trained large language model (LLM), and a modality interface (Zhong et al., 2022). The visual encoder (Radford et al., 2021; Li et al., 2022, 2023) compresses raw video or audio into compact representations, while pre-trained LLMs (Chung et al., 2022; Chiang et al., 2023; Touvron et al., 2023) supply broad world knowledge for downstream tasks. Since LLMs cannot directly process encoder outputs, a learnable interface aligns the modalities, often via a Q-Former that integrates at the token (Lin et al., 2023) or feature (Alayrac et al., 2022) level. A more detailed description of a typical VLM architecture is given in Appendix F.

While many models score high on popular benchmarks (not least because the benchmark data may appear in their training sets), they typically struggle with novel scenarios. As we developed a completely new benchmark and set the baseline, for evaluation we focused on top models from the HuggingFace VLM leaderboard, including Pangea (Yue et al., 2024), Gemini (Team et al., 2024), PALO (Rasheed et al., 2025), Video-Llama2 (Cheng et al., 2024), Phi-3.5-vision (Abdin et al., 2024), Qwen2-VL (Wang et al., 2024), Llava-NEXT-Video (Zhang et al., 2024), InternVL2 (Chen et al., 2024), MiniCPM-V (Yao et al., 2024) and DeepSeek-VL (Lu et al., 2024). We filtered out the models that do not support the languages featured in Perfect Times.

3 Method

We anchor the main clause action, the queried action in the correct answer (*mca*), and position the dependent clause action that sets the temporal and causal context relative to it (*dca*). This sequential shift reflects their temporal progression and fully covers Allen's interval algebra (Allen, 1984) that encodes all possible temporal relationships between two actions. Linguistically, these relations map to verb tenses and aspects and subordinate temporal conjunctions: for instance, *after* (Italian *dopo (che)*, Russian *после того, как*, Japanese 後) indicates that the *dca* precedes the *mca*; *while* (Italian *mentre*, Russian *пока*, Japanese ながら) denotes simultaneity; and *before* (Italian *prima (che)*, Russian *перед тем, как*, Japanese 前に) signals that the *dca* follows the *mca*.

For perfectivity, which captures an action's continuous or complete (and causally linked) nature, we focus on the start and end points of both the main (*st_mca*, *et_mca*) and dependent (*st_dca*, *et_dca*) clauses. We also label each action as telic or atelic based on its inherent properties. Accordingly, we group Allen's temporal relations into three categories (precedence, succession, and simultaneity for the main clause action) and define the plausible aspect correlations within each group. For example, an incomplete action without a clear endpoint is unlikely to occur *before* another action; this is a hardly plausible combination in sequential events even if grammatically acceptable in languages like Russian.

Next, inspired by STAR (Wu and Yu, 2021) and its sequence functional programs that exploit before/after relations, we developed templates in four languages to cover all temporal boundaries in conjunction with the semantics of action completion in both the main clause and dependent clause actions [2]. The templates were developed with the help of native speakers who have linguistic or philological background and specialize in translations.

For instance, when *mca* is a completed action, we use as its base *What **did** the person in the video do...* (Italian *Cosa ha fatto la persona nel video.../Cosa aveva fatto la persona nel video.../Cosa fece la persona nel video...*; Russian *Что сделал человек на видео...*; Japanese ビデオに写っている人は... 何をしましたか). The ongoing action as *mca* corresponds to *What **was** the person in the video **doing**...?* (Italian *Cosa stava facendo la persona nel video...* or *Cosa faceva la persona nel video...*; Russian *Что делал человек на видео...*; Japanese ビデオに写っている人は... 何をしていましたか).

Below we outline temporal relations and templates by groups, while the complete list of templates in all languages is given in Appendix C. The examples of all the temporal questions in Perfect Times are given in Table 1[3].

3.1 Precedence

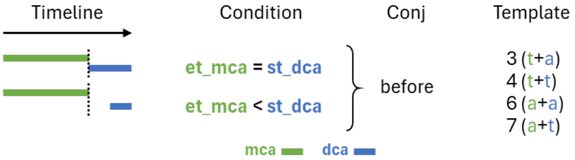


Figure 2: Precedence of mca in templates 3, 4, 6, and 7 with all combinations of mca and dca telicity markers.

As shown in Figure 2, when mca happens before dca (et_mca $\leq$ st_dca), all combinations of action completeness are possible. In the languages with tense agreement, English and Italian, the tense of mca is backshifted to a past form relative to dca. Russian and Japanese, in contrast, convey only the aspect of both mca and dca in place of shifting tenses.

3.2 Succession

When mca follows dca, the trigger is the dca's endpoint: at least some part of mca must take place

[2] Our goal is not to exhaustively cover all surface-level expressions of temporal relations (e.g., alternative conjunctions, adverbial phrases, or nominal constructions). Instead, we focus on evaluating the model's overall comprehension of temporal semantics with respect to perfectivity, regardless of paraphrasing.

[3] In the table, we provide examples in English for the general reference; some other examples in all the languages are in Appendix D.

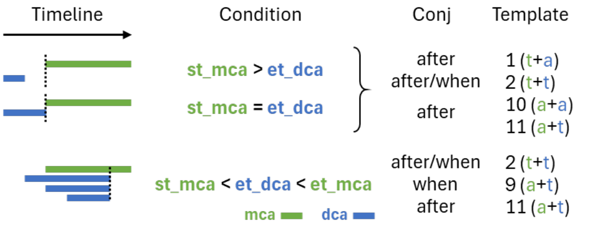
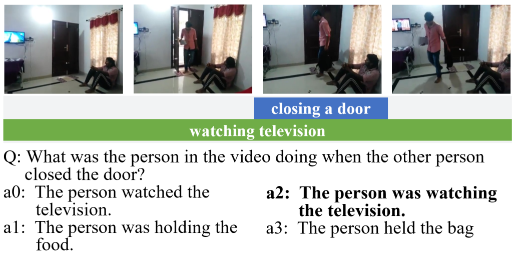


Figure 3: Succession of mca in templates 1, 2, 9, 10, 11. The strict sequence of actions allows for all combinations of telicity in mca and dca. Attention is focused on st_mca after or simultaneously with et_dca. If mca continues after dca, the sequence of events is focused on et_dca, which is perfective.

after it. Symmetrically to precedence, English and Italian have agreement of tenses.

In case of partial following of the telic dca by the atelic mca, the general temporal conjunction *when* (Italian *quando*; Russian *когда*; Japanese 時) is used the same way as the explicit sequential *after*, even though both actions may not strictly follow one another. This relationship is coded by Template 9 (Figure 4).



Q: What was the person in the video doing when the other person closed the door?
a0: The person watched the television.
a1: The person was holding the food.
a2: The person was watching the television.
a3: The person held the bag

Figure 4: Example in Perfect Times made by template 9b: two actions with different agents, durative mca and telic dca.

3.3 Simultaneity

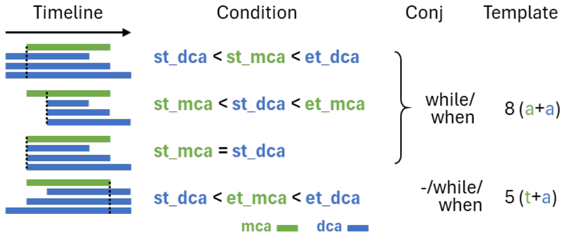


Figure 5: Overlapping mca and dca in templates 5 and 8.

The prototypical conjunctions for expressing overlapping actions, or simultaneity, are *while* (Italian *mentre*, Russian *пока*, Japanese ながら/間に) and the more general *when*. The continuous nature of dca (against which the mca of any perfectivity is questioned) underlies these relations.

When two telic actions finish at the same time, the ambiguous *when* emphasizes the simultaneity of the completion −as a synonym to *the moment* ***when***... (Figure 6).

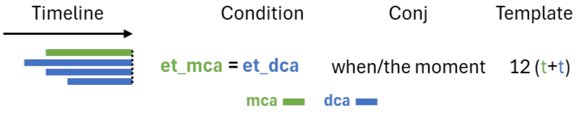


Figure 6: Template 12: simultaneity when both actions mca and dca end at the same moment in.

Each QA pair comes with three distractors that consider both the semantics and the grammatical form of the action.

3.4 Distractors

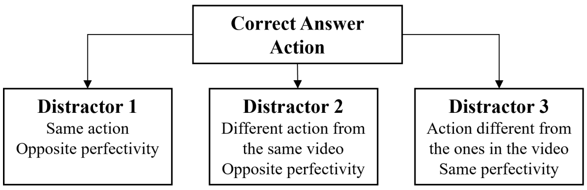


Figure 7: Distractors in relation to the Correct Answer.

Each distractor type is designed to probe different aspects of comprehension: linguistic nuance, context relevance, and discriminative ability within a shared scene. The diagram in Figure 7 illustrates the configuration of the distractors.

Distractor Type 1 is a variation of the correct answer rendered in a different form to convey the opposite completeness of the action (e.g., switching from past simple tense to a continuous/ing form in English or taking the opposite aspect word in Russian). It is intended to be very close to the correct answer with only a subtle difference in linguistic details. This tests whether a model (or human) can notice and correctly interpret the aspect mismatch between the question and the answer.

Distractor Type 2 comes from an alternative action that occurs in the same video, but uses the switched perfectivity/duration type the same way as in Type 1. Being contextually related, it forces the evaluator to differentiate between two plausible actions within the same scene. It tests the boundary between action comprehension and grammatical understanding.

Distractor Type 3 is taken from a completely different context. It comes from the list of actions that do not appear in the current video. Its purpose is to introduce an option that should be clearly out of context for the video at hand. A model with good action recognition should be able to dismiss such an option as completely implausible.

Among the four answer choices, both completed and durative actions are represented equally. Each answer option (one correct and three distractors) must be assigned to a random position for each question. To avoid option selection bias (Loginova et al., 2024), all distractor types and the correct answer should be distributed approximately equally across the answer options.

4 Experiments and Results

4.1 Dataset

Four annotators[4] labeled 400 clips of the Charades (Sigurdsson et al., 2016) collection of approximately 30-second videos. All activities in these videos were categorized into 157 verb action classes derived from Action Genome (Rai et al., 2021), along with their respective time boundaries[5]. The statistics of videos and classes are given in Appendix A.

The action classes, which are essentially verb phrases, were also annotated with telicity labels irrespective of their context. Inter-annotator agreement scores 0.67 by Fleiss' kappa, which is considered substantial (Landis and Koch, 1977), although it indicates that there is some inherent ambiguity or subjectivity in assessing the verb classes for telicity in isolation.

The algorithm begins by mixing actions within the same video. It goes line by line in the shuffled annotation file and assigns mca to the first action and dca to the second one for each pair of neighboring actions. Since the mca and dca annotations have telicity labels and time boundaries used as conditions, the algorithm applies all corresponding templates and generates the questions and correct answer options. Then the dataset is populated with distractors. The correct answers and distractor types are balanced across the answer options. We end up with the dataset of 3,739 QA pairs.

Subsequently, speakers of each language were asked to take this MCQA test with instructions similar to the prompts for VLMs: *Choose the correct answer (a0, a1, a2, or a3) and respond only with the option key (e.g., a0)*[6]. As a result, we obtained an inter-annotator agreement of 0.8 according to Fleiss' kappa (substantial agreement). This is how the gold standard of 93.36% accuracy was developed. This high percentage is due to the fact that distractors are designed in such a way that the correct answer is not difficult to intuitively predict. Notably, most annotators made the most mistakes in Distractor Type 2. Additional statistics on the annotators' responses are presented in Appendix E.

4.2 Models

We tested both the open-source and closed-source multilingual VLMs: Qwen2-VL (Wang et al., 2024), MiniCPM-V (Yao et al., 2024), InternVL2 (Chen et al., 2024), LLaVA-NeXT-Video (Zhang et al., 2024), GPT-4o (Achiam et al., 2023), and Gemini-2.0-Flash-Lite (Team et al., 2024). Additional details on open-source models are given in Table 11 of Appendix F.

We passed the video either directly (Qwen2-VL, Gemini-2.0-Flash-Lite and LLaVA-NeXT-Video) or first extracted frames every 3 seconds and passed the list of frames to those models that can only process a sequence of images (GPT-4o, InternVL2 and MiniCPM-V). The exact prompts for each model are provided in Appendix G.

[4] The annotators were recruited as volunteers and were all proficient in at least two target languages. Two of them have a background in linguistics, and their academic levels include two undergraduates and two postgraduates. Although none were directly involved in the project, each received detailed briefing and debriefing instructions from the professional linguist on the team. Their language proficiencies are as follows: Annotator 1 —English (fluent), Italian (intermediate), Russian (native), Japanese (fluent); Annotator 2 —English (native), Russian (native); Annotator 3 —English (advanced), Italian (native), Russian (native); Annotator 4 —Italian (native), Russian (native).

[5] The Charades dataset was initially annotated with activities and their time intervals for different research purposes. This previous annotation was inadequate for precisely determining the sequence of events or their concurrent occurrence due to its broad definition of action boundaries and a general interpretation of actions without considering their causal and temporal relationships. Additionally, the initial markup failed to note whether different actions were performed by the same character or different ones, a crucial detail for constructing meaningful questions.

[6] We deliberately did not provide additional instructions to annotators, as we aimed to collect the data based on linguistic intuition.

4.3 Results

4.3.1 Evaluation

We measure **Accuracy**, **F1 Macro**[7] and **Distractor Rate by Type**, which is the percentage of choices of each distractor type in case of error, calculated as follows:

Let us define D_i as the type of distractor chosen by the model for question i, if it was predicted incorrectly; T_k as a specific type of distractor (e.g., Type 1, Type 2, Type 3); and n_k as the number of times the model selected a distractor of type T_k.

The proportion of errors on distractors of type T_k is calculated as:

$$P(T_k) = \frac{n_k}{\sum_j n_j} \times 100, \qquad (1)$$

where $P(T_k)$ is the probability of selecting a distractor of type T_k and $\sum_j n_j$ is the total number of cases where the model made an error[8].

4.4 Quantitative and Qualitative Results

A total of 5.67% (208)[9] of questions were answered incorrectly by all models. These correspond to 162 videos, with only one video where every question was answered incorrectly by all models[10]. Thus, the dataset contains no videos completely unintelligible to all models; instead, failures are localized to specific actions or the correct order of actions within certain videos. The accuracy of all models in all languages remains significantly below the human gold standard. Coupled with consistent error trends across models and languages, this highlights that even the strongest current VLMs lack robust mechanisms for precise temporal fusion.

Table 2 shows the results of our tests on Perfect Times dataset with the breakdown by the distractor types. All models tend to ignore linguistic aspect, as most errors are related to Distractor Types 1 and 2, where the predicted verb form does not match the verb form in the question. The majority of errors, except for InternVL2, are due to Type 2 distractors. This suggests that models do not identify which temporal moment is crucial for the answer, even if they can distinguish actions relevant to the video. Therefore, VLMs should leverage multimodal fusion or specialized architectures to better capture temporality.

InternVL2, in contrast, demonstrates good understanding of temporal context but struggles with subtle linguistic differences in aspect; its majority of errors are Type 1 distractors in every language except English (possibly due to greater data contamination). Its worst performance in Italian (the language with the most diverse verb forms for encoding temporal relations) points to the model's lack of understanding of grammatical distinctions even though it is better at localizing events with respect to their order in the video.

Across all languages, LLaVA-NeXT-Video performs the worst. Despite its multilingual LLM backbone (based on Qwen (Bai et al., 2023)), its performance is nearly random for languages other than English. Moreover, answer analysis revealed a severe selection bias[11] towards the first and second options in English and the first option in other languages. This indicates a systematic flaw in its logic and temporal reasoning, even though it is able to handle basic contextual cues (Distractor Type 3 is a minority). Its temporal capabilities could be improved through enhanced multilingual action recognition of spatiotemporal features.

Gemini-2.0-flash-lite consistently performed better across all languages, showing similar accuracy in English, Italian, and Japanese. Other models display greater variance in accuracy. A thorough causal analysis for this model would require access to training data and architecture, which is not possible; however, factors such as training data saturation and tokenizer limitations (particularly in morphologically diverse languages such as Italian and Russian) likely contribute to the errors.

Gemini-2.0-flash-lite, GPT-4o, and InternVL2-8B generally perform better on Russian data, where perfectivity is lexically encoded, and worse on Japanese, which demands stronger visual disambiguation. This reinforces our claim in Section 1 that language-specific encoding of time and aspect strongly affects temporal reasoning. It also echoes findings in the cognitive literature discussed in Sec-

[7] We also checked F1 Micro, but because the dataset is rather balanced, this metric with TP/FP/FN does not differ much from Accuracy.

[8] A similar metric can be applied if we compute the error frequency relative to all predictions, not just erroneous ones: $P(T_k) = \frac{n_k}{N} \times 100$, where N is the total number of questions.

[9] See detailed statistics in Appendix H.

[10] This could suggest problems in question formulation, but a quick manual check of 10% of these questions confirmed they are unambiguously answerable by humans.

[11] All other models revealed less prominent selection bias in the first option, a0.

Table 2: VLM performance on Perfect Times across different languages (in percentage).

Model	Accuracy	F1 Macro	Distractor Type 1	Distractor Type 2	Distractor Type 3
			English		
Gemini-2.0-flash-lite	43.41	42.19	22.15	30.42	4.01
GPT-4o	43.25	34.39	21.96	31.14	3.64
MiniCPM-V-2_6	36.19	35.5	27.68	31.08	5.05
Qwen2-VL-7B-Instruct	35.09	32.68	21.24	39.2	4.47
InternVL2-8B	34.86	33.36	27.34	29.09	8.7
LLaVA-NeXT-Video-7B	33.38	32.86	25.39	30.99	10.22
			Italian		
Gemini-2.0-flash-lite	43.11	42.15	21.85	30.78	4.25
Qwen2-VL-7B-Instruct	41.59	39.63	21.95	32.09	4.35
GPT-4o	40.71	32.34	25.29	31.45	2.49
MiniCPM-V-2_6	37.42	35.73	29.55	28.96	4.07
InternVL2-8B	34.07	32.38	32.00	24.63	9.3
LLaVA-NeXT-Video-7B	25.92	17.65	25.40	29.74	18.88
			Russian		
Gemini-2.0-flash-lite	46.99	46.33	19.04	29.85	4.12
GPT-4o	45.04	44.97	19.60	32.15	3.21
InternVL2-8B	36.87	34.37	28.82	25.54	8.77
MiniCPM-V-2_6	36.29	34.61	28.88	30.20	4.63
Qwen2-VL-7B-Instruct	34.13	32.75	20.40	39.5	5.96
LLaVA-NeXT-Video-7B	26.95	24.7	24.3	36.00	12.74
			Japanese		
Gemini-2.0-flash-lite	43.06	41.77	22.19	31.43	3.32
MiniCPM-V-2_6	38.73	36.31	30.25	26.34	4.68
GPT-4o	38.49	30.65	23.78	35.23	2.49
Qwen2-VL-7B-Instruct	37.52	36.85	20.17	37.90	4.41
InternVL2-8B	35.05	31.65	31.22	23.92	9.81
LLaVA-NeXT-Video-7B	26.18	19.24	25.25	33.06	15.41

tion 2.1, where perfectivity encoded via telicity is more readily interpreted by native speakers.

Another parallel between human cognitive processes of causality and model behavior emerges from qualitative analysis: all tested VLMs, when mistaken, prefer perfect (or telic) answers over durative (or atelic) ones[12]. This may reflect the characteristics of source data, where strong causal connections between sequential actions are prevalent.

Regarding specific temporal conjunctions, all tested models struggle to disambiguate *when* (and equivalents in other languages). The highest number of correct answers is found for simultaneity, marked with *while* (Template 8a1), whereas replacing *while* with *when* for the same temporal relation and identical verb forms (Template 8a2) results in significantly more errors[13]. Similarly, the unambiguous conjunction *after* in 2a1 and 2b1 yields fewer errors than 2a2 and 2b2, which use *when* to describe the same action. This behavior highlights a weakness in aligning temporal boundaries and the semantics of action relations with verb aspect. Following cognitive findings that underline human reliance on integrated multimodal cues, it indicates that the separate language stream in multimodal models, even when combined with vision, is insufficient for temporal reasoning in videos.

[12] Also, considering correct answers only, GPT-4o is the only model that predicts fewer correct telic classes. While this cannot be confirmed for closed-source models, we tentatively attribute it to the volume and contamination in their training data. Table 12 in Appendix H presents the telicity breakdown.

[13] Appendix I provides examples and quantitative data.

The best-answered templates concern *after* questions, while *before* questions are the most challenging. When processing visual information, models more easily identify subsequent actions than preceding ones. This is further evidenced by the prevalence of Type 2 distractor errors, where the action is present in the video, but not in the correct sequence or form. Therefore, model errors stem not from a lack of understanding of the entire video or individual actions, but from a misinterpretation of the temporal order, such as answering about a subsequent action when asked about a previous one.

This pattern may also result from imbalanced instruction-tuned training data for temporal benchmarks. According to psychological research, it is more natural for people to describe events in chronological order via causal relationships, rather than discuss the reverse order[14].

Another notable finding is that the largest number of questions answered correctly by all five models except LLaVA-NeXT-Video-7B[15] is in Italian, and the smallest in Japanese (see Tables 13-14 in Appendix H). This suggests that explicit surface syntactic information can help in choosing correct answers and establishing proper temporal relations, provided it is systematically integrated into large-scale statistical processing, even when the dataset contains noise.

5 Conclusion

We have demonstrated that a full understanding of event dynamics requires an integrated approach where both visual cues and grammatical structures inform the interpretation of action completion. Our cross-linguistic evaluation shows that the SoTA VLMs rely on limited superficial cues, particularly when disambiguating ambiguous temporal conjunctions and distinguishing subtle aspectual markers. Models favor telic responses, mirroring human tendencies to find cause and effect dependencies in events and exposing a reliance on lexical over grammatical signals. These observations reinforce the need for enhanced multimodal fusion and specialized temporal representations.

A key advantage of our **Perfect Times** dataset is its semi-synthetic, template-based design. By leveraging a unique methodology that uses universal temporal and aspectual templates, our dataset can be easily augmented and adapted to any language, provided videos are annotated with action labels and timestamps. This flexibility makes our benchmark not only a novel tool for evaluating temporal reasoning in VLMs but also a scalable resource to drive further research. Our work thus sets a new standard for probing the intricate interplay of time, causality, and aspect in multimodal contexts.

6 Limitations

Our study is the first to link tense, aspect, and visual modality to test deep temporal reasoning in models. However, it is not exhaustive regarding the full range of synonymous temporal constructions. Future work could augment the dataset with paraphrases generated by LLMs to increase coverage.

Additionally, our semi-synthetic, template-based approach may not fully reflect the variability found in natural language. Although templates ensure controlled testing of specific temporal relations, they might not fully reflect spontaneous speech. Expanding the dataset to include more naturalistic examples could provide further insights.

Another limitation is the number of verb classes: 157 classes for 400 videos. While a broader range of classes would capture more nuanced actions, the current quantity is balanced by high-quality annotations and language-specific templates, which were meticulously crafted by experts, a resource-intensive process.

Finally, with more annotations per language, the gold standard accuracy might decrease when averaged across annotators, although it is unlikely to reach the performance levels of current state-of-the-art models. Future research should explore scaling up the dataset, both in terms of linguistic variety and video domains, to fully assess and enhance multimodal temporal reasoning.

7 Ethics Statement

We introduce a new benchmark dataset derived from the video datasets *Charades* and *Action Genome*. We re-annotate their videos to support the new evaluation of both open- and closed-source vision-language models (VLMs). In doing so, we strictly adhere to the licensing terms of the origi-

[14] This is also confirmed by the NeXT-QA and STAR datasets, widely used for causal and temporal evaluation: the overwhelming majority of temporal questions focus on succession via *after*, rather than precedence via *before*.

[15] LLaVA-NeXT-Video-7B is excluded because its results for non-English languages are statistically non-informative and nearly random.

nal datasets, ensuring that our derivative work complies with all copyright and usage restrictions.

Our annotation process involved trained annotators following the guidelines aimed at ensuring consistency. Despite these efforts, we acknowledge that any human annotation process may introduce subjective interpretations. We therefore encourage users of our dataset to consider potential annotation biases when interpreting experimental results.

The evaluation of VLMs, particularly those that generate open-ended text, carries inherent risks. We have taken measures to mitigate such risks by prompting and conducting evaluations. We promote transparency through the public release of our dataset and code. Our intention is to foster reproducible research and to provide a resource that can contribute to improving the trustworthiness and robustness of VLMs.

Acknowledgments

Olga Loginova thanks Amazon Alexa for supporting her research through a generous donation to Raffaella Bernardi. The authors thank the native speakers of the target languages, both annotators and consultants with relevant backgrounds: Angelo Noè, Vasili Noè, Elizaveta Loginova, Maya Udaka, and Cristina Crippa. We are especially grateful to Angelo Noè for his considerable support throughout this project.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Benjamin Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Sim'on Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Raphael Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Lukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Ryan Kiros, Matthew Knight, Daniel Kokotajlo, Lukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Adeola Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv

Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel P. Mossing, Tong Mu, Mira Murati, Oleg Murk, David M'ely, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Ouyang Long, Cullen O'Keefe, Jakub W. Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alexandre Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Pondé de Oliveira Pinto, Michael Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario D. Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin D. Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas A. Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cer'on Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning. *ArXiv*, abs/2204.14198.

James F. Allen. 1984. Towards a general theory of action and time. *Artificial Intelligence*, 23(2):123–154.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Jasmijn E. Bosch, Mathilde Chailleux, and Francesca Foppolo. 2021. Incremental processing of telicity in italian children.

Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, Yao Dou, Jaden Park, Jianfeng Gao, Yong Jae Lee, and Jianwei Yang. 2024. Temporalbench: Towards fine-grained temporal understanding for multimodal video models. *arXiv preprint arXiv:2410.10818*.

Franklin Chang, Tomoko Tatsumi, Yuna Hiranuma, and Colin Bannard. 2023. Visual heuristics for verb production: Testing a deep-learning model with experiments in japanese. *Cognitive science*, 47 8:e13324.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed Huai hsin Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416.

Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees G. M. Snoek, and Yuki M. Asano. 2024. Lost in time: A new temporal benchmark for videollms.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. 2020. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43:4125–4141.

Francesca Foppolo, Jasmijn E. Bosch, Ciro Greco, Maria Nella Carminati, and Francesca Panzeri. 2021. Draw a star and make it perfect: Incremental processing of telicity. *Cognitive science*, 45 10:e13052.

Francesca Foppolo, Francesca Panzeri, Cesare Greco, and Matteo Carminati. 2016. The incremental processing of accomplishment predicates.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Rongrong Ji, and Xing Sun. 2024. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *Preprint*, arXiv:2405.21075.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–970.

Y. Jang, Yale Song, Chris Dongjoo Kim, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2019. Video question answering with spatio-temporal reasoning. *International Journal of Computer Vision*, 127:1385 – 1412.

Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. 2019. Action genome: Actions as compositions of spatio-temporal scene graphs. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10233–10244.

Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Apostol Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The kinetics human action video dataset. *ArXiv*, abs/1705.06950.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1).

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L. Berg. 2018. Tvqa: Localized, compositional video question answering. In *Conference on Empirical Methods in Natural Language Processing*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *ArXiv*, abs/2311.10122.

Yi Liu, Limin Wang, Xiao Ma, Yali Wang, and Y. Qiao. 2021. Fineaction: A fine-grained video dataset for temporal action localization. *IEEE Transactions on Image Processing*, 31:6937–6950.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. 2024. Tempcompass: Do video llms really understand videos? *Preprint*, arXiv:2403.00476.

Olga Loginova, Oleksandr Bezrukov, and Alexey Kravets. 2024. Addressing blind guessing: Calibration of selection bias in multiple-choice question answering by video language models. *Preprint*, arXiv:2410.14248.

Agnese Lombardi and Alessandro Lenci. 2023. Agentività e telicità in gilberto: implicazioni cognitive. *Preprint*, arXiv:2307.02910.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. 2024. Deepseek-vl: Towards real-world vision-language understanding. *Preprint*, arXiv:2403.05525.

Eleni Metheniti, Tim Van De Cruys, and Nabil Hathout. 2022. About time: Do transformers learn temporal verbal aspect? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 88–101, Dublin, Ireland. Association for Computational Linguistics.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*.

Serge Minor, Natalia Mitrofanova, Gustavo Guajardo, Myrte Vos, and Gillian Catriona Ramchand. 2022. Temporal information and event bounding across languages: Evidence from visual world eyetracking from. *Semantics and Linguistic Theory*.

Marc Moens and Mark Steedman. 1988. Temporal ontology and temporal reference. *Comput. Linguist.*, 14(2):15–28.

Viorica Pătrăucean, Lucas Smaira, Ankush Gupta, Adrià Recasens Continente, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alex

Frechette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and João Carreira. 2023. Perception test: A diagnostic benchmark for multimodal video models. In *Advances in Neural Information Processing Systems*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Nishant Rai, Haofeng Chen, Jingwei Ji, Rishi Desai, Kazuki Kozuka, Shun Ishizaka, Ehsan Adeli, and Juan Carlos Niebles. 2021. Home action genome: Cooperative compositional action understanding. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11179–11188.

Hanoona Rasheed, Muhammad Maaz, Abdelrahman Shaker, Salman Khan, Hisham Cholakal, Rao M. Anwer, Tim Baldwin, Michael Felsberg, and Fahad S. Khan. 2025. Palo: A large multilingual multimodal language model. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV 2025)*.

Arka Sadhu, Kan Chen, and Ram Nevatia. 2021. Video question answering with phrases via semantic roles. *ArXiv*, abs/2104.03762.

Gunnar A. Sigurdsson, Gül Varol, X. Wang, Ali Farhadi, Ivan Laptev, and Abhinav Kumar Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu

Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei ”Louis” Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris

Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins,

Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, Xiang-Hai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Carol Tenny. 1994. Aspectual roles and the syntax-semantics interface.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Angeliek van Hout. 2008. Acquiring perfectivity and telicity in dutch, italian and polish. *Lingua*, 118 11:1740–1765.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Neural Information Processing Systems*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Shaojie Wang, Wentian Zhao, Ziyi Kou, Jing Shi, and Chenliang Xu. 2021. How to make a blt sandwich? learning vqa towards understanding web instructional videos. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1129–1138.

Haiwan Wei, Yitian Yuan, Xiaohan Lan, Wei Ke, and Lin Ma. 2025. Instructionbench: An instructional video understanding benchmark. *ArXiv*, abs/2504.05040.

Bo Wu and Shoubin Yu. 2021. Star: A benchmark for situated reasoning in real-world videos. In *NeurIPS Datasets and Benchmarks*.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9772–9781.

D. Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video question answering via gradually refined attention over appearance and motion. *Proceedings of the 25th ACM international conference on Multimedia*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Zhou Yu, D. Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. *ArXiv*, abs/1906.02467.

Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. 2024. Pangea: A fully open multilingual multimodal llm for 39 languages. *arXiv preprint arXiv:2410.16153*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. *Preprint*, arXiv:2303.15343.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2024. Video instruction tuning with synthetic data. *Preprint*, arXiv:2410.02713.

Yiyun Zhao, Jian Gang Ngui, Lucy Hall Hartley, and Steven Bethard. 2021. Do pretrained transformers infer telicity like humans? In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 72–81, Online. Association for Computational Linguistics.

Yaoyao Zhong, Wei Ji, Junbin Xiao, Yicong Li, Wei Deng, and Tat seng Chua. 2022. Video question answering: Datasets, algorithms and challenges. *ArXiv*, abs/2203.01225.

Table 3: General Statistics of Perfect Times

Parameter	**Value**
Total videos	400
Total videos duration (sec)	11386.65
Total videos duration (min)	189.78
Average video duration (sec)	29.20
Minimum video duration (sec)	7.21
Maximum video duration (sec)	54.12

Table 4: Action Annotations

Parameter	**Annotation**
Total actions	3115
Minimum actions per video	2
Maximum actions per video	30
Average actions per video	7.99
Average time per action (sec)	8.35
Actions with $\leq$ 1s duration	1149 (36.89%)
Actions with $\leq$ 2s duration	1437 (46.13%)
Actions with $\leq$ 5s duration	1850 (59.39%)
Actions with $\leq$ 10s duration	2259 (72.52%)

A Video and Action Classes Statistics

Tables 3 and 4 provide details on videos and action classes in Perfect Times, respectively.

B Dataset Statistics

The breakdown by answer options with the correct answers and distractor types of the total of 3739 qa pairs is given in Table 5.

Table 5: QA pair statistics and distractor types.

Parameter	**Value**
Total QA pairs	3739
Correct a0	953
Correct a1	966
Correct a2	960
Correct a3	860
a0_distractor_type	953
a1_distractor_type	916
a2_distractor_type	927
a3_distractor_type	943

C Templates

This section presents templates in all languages, taking into account the parsed verb phrases from video annotations. In most cases, the classes include one verb, verb_1. In the tables below for several verbs in one class in coordinating constructions, conj, the coordinating conjunction and all elements with _2 represent the second conjunct, as in *working* ***or playing on the laptop***. The second conjunt in the verb phrase is optional.

D Dataset Examples

The aggregated examples in all languages are presented in Figures 9-14.

E Annotator's Statistics

The statistics on each annotator's performance is given in Table 10, where each language represents the annotator of this language.

F Models

The core idea of VLMs is to bridge the gap between visual features extracted from video and linguistic features from text for such tasks as video captioning, video question answering, video summarization, and more.

Most VLMs typically share a high-level architecture comprising three main components:

1. **Visual Encoder**: Responsible for extracting meaningful features from the video stream.

2. **Language Encoder (and often a Language Decoder)**: An LLM that handles textual input and generates textual output.

3. **Multimodal Fusion Projector**: This component connects the visual and language parts, allowing them to interact and produce a unified representation. After extracting features from individual frames, it maps these visual features into the same embedding space as the language model's tokens.

A fundamental aspect of VLM processing is that video is generally treated as a sequence of individual images (frames). To prevent the loss of temporal information when processing each frame independently, VLMs often incorporate mechanisms to capture motion and temporal relationships, such as with concatenation of features or through a dedicated temporal transformer.

Frames are sampled either uniformly or by identifying keyframes. For models like GPT-4o, InternVL2, and MiniCPM-V, we use uniform sampling by extracting frames at regular intervals (e.g., every 3 seconds or up to an upper limit). Similarly, Gemini-2.0-Flash-Lite, while inherently multimodal, gets individual frames when it is passed the entire video and interacted with via API. In contrast, LLaVA-NeXT-Video handles video processing internally with its specialized LlavaNextVideoProcessor and Qwen2-VL uses AutoProcessor and Qwen2VLForConditionalGeneration.

The Figure 8 illustrates the processing pipeline from multimodal input to text generation.

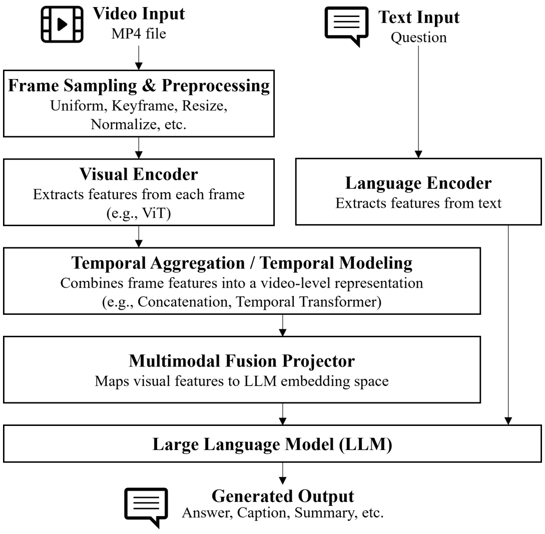

Video Input
MP4 file
Text Input
Question
Frame Sampling & Preprocessing
Uniform, Keyframe, Resize, Normalize, etc.
Visual Encoder
Extracts features from each frame (e.g., ViT)
Language Encoder
Extracts features from text
Temporal Aggregation / Temporal Modeling
Combines frame features into a video-level representation (e.g., Concatenation, Temporal Transformer)
Multimodal Fusion Projector
Maps visual features to LLM embedding space
Large Language Model (LLM)
Generated Output
Answer, Caption, Summary, etc.

Figure 8: Overview of a Typical VLM Architecture from raw video and text inputs through encoding, multimodal fusion, and final text generation.

Table 11 gives details on the open-source models' components.

G Prompts

Since each model accepts input differently, the general question-answer options prompts have been customized.

Qwen2-VL Prompt The Qwen2-VL question messages for all the languages are straightforward, as follows:

"en": question_str = f"Question: {question}" instruction = "Choose the correct answer (a0, a1, a2, or a3)."

"it": question_str = f"Domanda: {question}" instruction = "Scegli la risposta corretta (a0, a1, a2 o a3)."

"ru": question_str = f"Вопрос: {question}" instruction = "Выберите правильный ответ (a0, a1, a2 или a3)."

"jp": question_str = f"質問: {question}" instruction = "正しい答えを選んでください (a0, a1, a2, a3)。”

Together with the answer options they form text prompts:

```
text\_prompt = (
    f”{question_str}\n”
    f”a0: {options['a0']}\n”
    f”a1: {options['a1']}\n”
    f”a2: {options['a2']\n}”
    f”a3: {options['a3']\n}”
    f”{instruction}”
```

The combined input to the models includes videos as follows:

```
messages = [
    {”role”: ”user”,
    ”content”: [
        {”type”: ”video”,
        ”video”: video_uri,
        ”fps”: 1.0,
        ”max_pixels”: 360 * 420},
        {”type”: ”text”, ”text”: text_prompt},
        ]
    }
]
```

MiniCPM-V Prompt The text prompt for MiniCPM-V needs specific clarification about the option key. Otherwise it generates an unparsable answer.

“en”: question_str = f“Question: {question}”

instruction = “Choose the correct answer (a0, a1, a2, or a3) and respond only with the option key (e.g., a0). Do not include any additional text.”

“it”: question_str = f“Domanda: {question}”

instruction = “Scegli la risposta corretta (a0, a1, a2 o a3) e rispondi solo con la chiave dell’opzione (es. a0). Non includere testo aggiuntivo.”

“ru”: question_str = f“Вопрос: {question}”

instruction = “Выберите правильный ответ (a0, a1, a2 или a3) и ответь только ключом варианта (например, a0). Не добавляй дополнительный текст.”

“jp”: question_str = f“質問: {question}”

instruction = “正しい答えを選んでください (a0, a1, a2, a3) そしてオプションキーのみで回答してください (例: a0)。追加のテキストを含めないでください。”

The entire text prompt is as follows:

```
text_prompt = (
        f”{question_str}\n”
        f”a0: {options['a0']\n}”
        f”a1: {options['a1']\\n}”
        f”a2: {options['a2']\\n}”
        f”a3: {options['a3']\\n}”
        f”{instruction}”)
```

The message content is defined as a list that concatenates the list of video frames (as PIL images) with the text prompt the following way:

```
[{”role”: ”user”,
    ”content”: frames + [text_prompt]}]
```

InternVL2 Prompt The prompt for InternVL2 contains a system message, frame intro, question, answer options and instruction.

```
text_prompt = (
    f”{system_messages[language]}\n”
    f”{frame_intro[language]}\n\n”
    f”{question_str}\n”
    f”a0: {options['a0']}\n”
    f”a1: {options['a1']}\n”
    f”a2: {options['a2']}\n”
    f”a3: {options['a3']}\n”
    f”{instruction}”
)
```

Localized messages for instructions to InternVL2 are as follows:

system_messages = {

“en”: “Use the provided video frames to answer the question.”,

“it”: “Utilizza i frame del video per rispondere alla domanda.”,

“ru”: “Используй данные кадры видео, чтобы ответить на вопрос.”,

“jp”: “提供されたビデオフレームを使用して質問に答えてください。”

}

frame_intro = {

“en”: “These are the frames from the video.”,

“it”: “Questi sono i frame del video.”,

“ru”: “Это кадры из видео.”,

“jp”: “以下はビデオのフレームです。”

}

“en”: question_str = f“Question: {question}”

instruction = “Choose the correct answer (a0, a1, a2, or a3).”

“it”: question_str = f“Domanda: {question}”

instruction = “Scegli la risposta corretta (a0, a1, a2 o a3).”

“ru”: question_str = f“Вопрос: {question}”

instruction = “Выберите правильный ответ (a0, a1, a2 или a3).”

“jp”: question_str = f“質問: {question}”

instruction = “正しい答えを選んでください (a0, a1, a2, a3)。”

LLaVA-NeXT-Video Prompt LLaVA-NeXT-Video accepts videos as input:

```
{
    ”role”: ”user”,
    ”content”: [
        {”type”: ”text”, ”text”: text_prompt},
        {”type”: ”video”,
        ”video”: f”file://{video_path}”},
    ],
}
```

The text prompt has the localized question, answer options and instruction as follows:

```
text_prompt = (
        f”{question_str} {question}\n”
        f”a0: {options['a0']}\n”
        f”a1: {options['a1']}\n”
        f”a2: {options['a2']}\n”
        f”a3: {options['a3']}\n”
        f”{instruction}”
    )
```

The question messages and instructions in four languages are as follows:

translations = {

“en”: (“Question:”, “Please choose the correct answer (a0, a1, a2, or a3).”),

“it”: (“Domanda:”, “Per favore, scegli la risposta corretta (a0, a1, a2 o a3).”),

“ru”: (“Вопрос:”, “Пожалуйста, выберите правильный ответ (a0, a1, a2 или a3).”),

“jp”: (“質問:”, “正しい答えを選んでください (a0, a1, a2, a3)。”)

}

GPT-4o Prompt Localized messages for making the GPT-4o prompt are as follows:

system_messages = {

“en”: “Use the provided video frames to answer the question.”,

“it”: “Utilizza i frame del video per rispondere alla domanda.”,

“ru”: “Используй данные кадры видео, чтобы ответить на вопрос.”,

“jp”: “提供されたビデオフレームを使用して質問に答えてください。” }

frame_intro = { “en”: “These are the frames from the video.”,

“it”: “Questi sono i frame del video.”,

“ru”: “Это кадры из видео.”,

“jp”: “以下はビデオのフレームです。” }

To avoid generating multiple tokens, we further restrict the model in the instruction:

“en”: (“Question”, “Choose the correct answer (a0, a1, a2, or a3) and respond **only** with the option key (e.g., a0). Do not include any additional text.”),

“it”: (“Domanda”, “Scegli la risposta corretta (a0, a1, a2 o a3) e rispondi **solo** con la chiave dell’opzione (es. a0). Non includere testo aggiuntivo.”),

“ru”: (“Вопрос”, “Выбери правильный ответ (a0, a1, a2 или a3) и ответь **только** ключом варианта (например, a0). Не добавляй дополнительный текст.”),

“jp”: (“質問”, “正しい答えを選んでください (a0, a1, a2, a3) そしてオプションキーのみで回答してください (例: a0)。追加のテキストを含めないでください。”

The entire text prompt contains a question label, a question, answer options and an insruction as follows:

```
text_prompt = (
    f”{question_label}: {question}\n”
    f”a0: {options['a0']}\n”
    f”a1: {options['a1']}\n”
    f”a2: {options['a2']}\n”
    f”a3: {options['a3']}\n”
    f”{instruction}”
)
```

In the input message to the model we pass the text prompt and frames as follows:

```
messages = [
    {”role”: ”system”, ”content”: system_message},
    {”role”: ”user”, ”content”: [frame_message,
    *frames_payload, prompt]}
    ]
```

Gemini-2.0-Flash-Lite Prompt Similarly to GPT-4o, the question previews and instructions for Gemini-2.0-Flash-Lite are localized with the restriction for output tokens as follows:

“en”:

question_str = f“Question: {question}”

instruction = (“Choose the correct answer (a0, a1, a2, or a3) and respond only with the option key (e.g., a0). Do not include any additional text.”)

“it”:

question_str = f“Domanda: {question}”

instruction = (“Scegli la risposta corretta (a0, a1, a2 o a3) e rispondi solo con la chiave dell’opzione (es. a0). Non includere testo aggiuntivo.”)

“ru”:

question_str = f“Вопрос: {question}”

instruction = (“Выбери правильный ответ (a0, a1, a2 или a3) и ответь только ключом варианта (например, a0). Не добавляй дополнительный текст.”)

“jp”:

question_str = f“質問: {question}”

instruction = (“正しい答えを選んでください (a0, a1, a2, a3) そしてオプションキーのみで回答してください (例: a0)。追加のテキストを含めないでください。”)

The text prompt template is the same as for Qwen2-VL and MiniCPM-V:

```
text_prompt = (
        f”{question_str}\n”
        f”a0: {options['a0']}\n”
        f”a1: {options['a1']}\n”
        f”a2: {options['a2']}\n”
        f”a3: {options['a3']}\n”
        f”{instruction}”
    )
```

The input contents can simply be the list [text_prompt, video_part], where the video with the proper MIME type, so that the model knows exactly what kind of data it takes.

For all the models the answer is parsed with regex:

```
re.search(r"\ba[0-3]\b", generated_text.lower())
```

H Statistics on Correct and Incorrect Answers

Table 12 shows the ratio of telic and atelic classes in the results of each model. With the exception of GPT-4o, all models perform better on questions about telic classes.

Tables 13-14 gives the intersection of correctly and incorrectly answered questions by all models respectively. As LLaVA-NeXT- Video-7B performed poorly on languages other than English, we give the results with and without it.

I Statistics on Conjunctions

Figures 15-18 show preference of the correct answers with *while* over the ones with *when* across all the languages in numbers.

Tables 15-18 show Top-5 Best/Worst Answered questions by templates corresponding to the sequential and durative conjunctions.

Table 6: English templates: _past_ is the past simple form, _past_perf_ is the past perfect form, _ing_ is the -ing form including the past continuous form.

Template	MCA	DCA	Condition	n_persons	Question
1a	t	a	st_mca >= et_dca; st_mca < et_dca & et_mca > et_dca	1	What did the person in the video do after v_dca_ing_1 the obj_1 adjunct conj v_dca_ing_2 the obj_2 ?
1b	t	a	st_mca >= et_dca; st_mca < et_dca & et_mca > et_dca	2	What did the person in the video do after the other person person v_dca_past_1 the obj_1 adjunct conj v_dca_past_2 the obj_2 ?
2a1	t	t	st_mca >= et_dca; st_mca < et_dca & et_mca > et_dca	1	What did the person in the video do after they v_dca_past_perf_1 the obj_1 adjunct conj v_dca_past_perf_2 the obj_2 ?
2a2	t	t	st_mca >= et_dca; st_mca < et_dca & et_mca > et_dca	1	What did the person in the video do when they v_dca_past_perf_1 the obj_1 adjunct conj v_dca_past_perf_2 the obj_2 ?
2b1	t	t	st_mca >= et_dca; st_mca < et_dca & et_mca > et_dca	2	What did the person in the video do after the other person v_dca_past_perf_1 the obj_1 adjunct conj v_dca_past_perf_2 the obj_2 ?
2b2	t	t	st_mca >= et_dca; st_mca < et_dca & et_mca > et_dca	2	What did the person in the video do when the other person v_dca_past_perf_1 the obj_1 adjunct conj v_dca_past_perf_2 the obj_2 ?
3a	t	a	et_mca <= st_dca	1	What had the person in the video done before v_dca_ing_1 the obj_1 adjunct conj v_dca_ing_2 the obj_2 ?
3b	t	a	et_mca <= st_dca	2	What had the person in the video done before the other person v_dca_past_1 the obj_1 adjunct conj v_dca_past_2 obj_2 ?
4a	t	t	et_mca <= st_dca	1	What had the person in the video done before they v_dca_past_1 the obj_1 adjunct conj v_dca_past_2 the obj_2 ?
4b	t	t	et_mca <= st_dca	2	What had the person in the video done before the other person v_dca_past_1 the obj_1 adjunct conj v_dca_past_2 the obj_2 ?
5a1	t	a	et_mca > st_dca & et_mca < et_dca	1	What did the person in the video do v_dca_ing_1 the obj_1 adjunct conj v_dca_ing_2 the obj_2 ?
5a2	t	a	et_mca > st_dca & et_mca < et_dca	1	What did the person in the video do while v_dca_ing_1 the obj_1 adjunct conj v_dca_ing_2 the obj_2 ?
5b1	t	a	et_mca > st_dca & et_mca < et_dca	2	What did the person in the video do while the other person was v_dca_ing_1 the obj_1 adjunct conj v_dca_ing_2 obj_2 ?
5b2	t	a	et_mca > st_dca & et_mca < et_dca	2	What did the person in the video do when the other person was v_dca_ing_1 the obj_1 adjunct conj v_dca_ing_2 the obj_2 ?
6a	a	a	et_mca <= st_dca	1	What had the person in the video been doing before v_dca_ing_1 the obj_1 adjunct conj v_dca_ing_2 the obj_2 ?
6b	a	a	et_mca <= st_dca	2	What had the person in the video been doing before the other person was v_dca_ing_1 the obj_1 adjunct conj v_dca_past_2 the obj_2 ?
7a	a	t	et_mca <= st_dca	1	What had the person in the video been doing before v_dca_ing_1 the obj_1 adjunct conj v_dca_ing_2 the obj_2 ?
7b	a	t	et_mca <= st_dca	2	What had the person in the video been doing before the other person v_dca_past_1 the obj_1 adjunct conj v_dca_past_2 the obj_2 ?
8a1	a	a	st_mca > st_dca & st_mca < et_dca; st_mca == st_dca; st_mca < st_dca & et_mca > st_dca	1	What was the person in the video doing while v_dca_ing_1 the obj_1 adjunct conj v_dca_ing_2 the obj_2 ?
8a2	a	a	st_mca > st_dca & st_mca < et_dca; st_mca == st_dca; st_mca < st_dca & et_mca > st_dca	1	What was the person in the video doing when v_dca_ing_1 the obj_1 adjunct conj v_dca_ing_2 the obj_2 ?
8b1	a	a	st_mca > st_dca & st_mca < et_dca; st_mca == st_dca; st_mca < st_dca & et_mca > st_dca	2	What was the person in the video doing while the other person was v_dca_ing_1 the obj_1 adjunct conj v_dca_ing_2 the obj_2 ?
8b2	a	a	st_mca > st_dca & st_mca < et_dca; st_mca == st_dca; st_mca < st_dca & et_mca > st_dca	2	What was the person in the video doing when the other person was v_dca_ing_1 the obj_1 adjunct conj v_dca_ing_2 the obj_2 ?
9a	a	t	st_mca < et_dca & et_mca > et_dca	1	What was the person in the video doing when they v_dca_past_1 the obj_1 adjunct conj v_dca_past_2 the obj_2 ?
9b	a	t	st_mca < et_dca & et_mca > et_dca	2	What was the person in the video doing when the other person v_dca_past_1 the obj_1 adjunct conj v_dca_past_2 the obj_2 ?
10a	a	a	st_mca >= et_dca	1	What was the person in the video doing after v_dca_ing_1 the obj_1 adjunct conj v_dca_ing_2 obj_2 ?
10b	a	a	st_mca >= et_dca	2	What was the person in the video doing after the other person v_dca_past_1 the obj_1 adjunct conj v_dca_past_2 the obj_2 ?
11a	a	t	st_mca >= et_dca; st_mca < et_dca & et_mca > et_dca	1	What was the person in the video doing after v_dca_ing_1 the obj_1 adjunct conj v_dca_ing_2 the obj_2 ?
11b	a	t	st_mca >= et_dca; st_mca < et_dca & et_mca > et_dca	2	What was the person in the video doing after the other person v_dca_past_1 the obj_1 adjunct conj v_dca_past_2 the obj_2 ?
12a1	t	t	et_mca == et_dca	1	What did the person in the video do when they v_dca_past_1 the obj_1 adjunct conj v_dca_past_2 the obj_2 ?
12a2	t	t	et_mca == et_dca	1	What did the person in the video do the moment they v_dca_past_1 the obj_1 adjunct conj v_dca_past_2 the obj_2 ?
12b1	t	t	et_mca == et_dca	2	What did the person in the video do when the other person v_dca_past_1 the obj_1 adjunct conj v_dca_past_2 the obj_2 ?
12b2	t	t	et_mca == et_dca	2	What did the person in the video do the moment the other person v_dca_past_1 the obj_1 adjunct conj v_dca_past_2 the obj_2 ?

Table 7: Italian templates: _inf_ is the infinitive form, _pass_pross_ is the passato prossimo (present perfect) form, _trapass_pross_ is the trapassato prossimo (past perfect) form, _pass_rem_ is the passato remoto (simple past) form, _imp_ is the imperfetto (imperfect) form, _ger_ is the gerundio (gerund) form, _imp_cong_ is the imperfetto congiuntivo (imperfect subjunctive) form, _imp_prog_ is the imperfetto progressivo (past continuous) form. Conj is a coordinating conjunction in case of several verbs in one class, as in *lavorando o giocando al computer*.

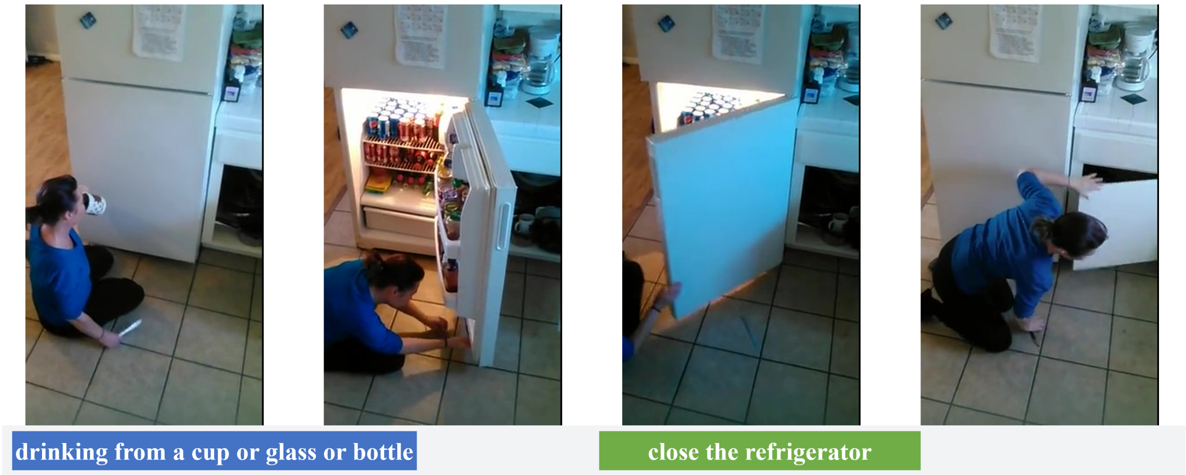
Template	MCA	DCA	Condition	n_persons	Question
1a	t	a	st_mca >= et_dca; st_mca < et_dca & et_mca > et_dca	1	Cosa ha fatto la persona nel video dopo v_inf_pass_pross_1 obj_1 adjunct conj v_inf_pass_pross_2 obj_2 ?
1b	t	a	st_mca >= et_dca; st_mca < et_dca & et_mca > et_dca	2	Cosa ha fatto la persona nel video dopo che l'altra persona v_trapass_pross_1 obj_1 adjunct conj v_trapass_pross_2 obj_2 ?
2a1	t	t	st_mca >= et_dca; st_mca < et_dca & et_mca > et_dca	1	Cosa fece la persona nel video dopo v_inf_pass_pross_1 obj_1 adjunct conj v_inf_pass_pross_2 obj_2 ?
2a2	t	t	st_mca >= et_dca; st_mca < et_dca & et_mca > et_dca	1	Cosa fece la persona nel video quando v_pass_rem_1 obj_1 adjunct conj v_pass_rem_2 obj_2 ?
2b1	t	t	st_mca >= et_dca; st_mca < et_dca & et_mca > et_dca	2	Cosa fece la persona nel video dopo che l'altra persona v_pass_rem_1 obj_1 adjunct conj v_pass_rem_2 obj_2 ?
2b2	t	t	st_mca >= et_dca; st_mca < et_dca & et_mca > et_dca	2	Cosa fece la persona nel video quando l'altra persona v_pass_rem_1 obj_1 adjunct conj v_pass_rem_2 obj_2 ?
3a	t	a	et_mca <= st_dca	1	Cosa aveva fatto la persona nel video prima di v_inf_1 obj_1 adjunct conj v_inf_2 obj_2 ?
3b	t	a	et_mca <= st_dca	2	Cosa aveva fatto la persona nel video prima che l'altra persona v_imp_1 obj_1 adjunct conj v_imp_2 obj_2 ?
4a	t	t	et_mca <= st_dca	1	Cosa aveva fatto la persona nel video prima di v_inf_1 obj_1 adjunct conj v_inf_2 obj_2 ?
4b	t	t	et_mca <= st_dca	2	Cosa aveva fatto la persona nel video prima che l'altra persona v_trapass_cong_1 obj_1 adjunct conj v_trapass_cong_2 obj_2 ?
5a1	t	a	et_mca > st_dca & et_mca < et_dca	1	Cosa ha fatto la persona nel video v_ger_1 obj_1 adjunct conj v_ger_2 obj_2 ?
5a2	t	a	et_mca > st_dca & et_mca < et_dca	1	Cosa ha fatto la persona nel video mentre v_imp_1 obj_1 adjunct conj v_imp_2 obj_2 ?
5b1	t	a	et_mca > st_dca & et_mca < et_dca	2	Cosa ha fatto la persona nel video mentre l'altra persona v_imp_1 obj_1 adjunct conj v_imp_2 obj_2 ?
5b2	t	a	et_mca > st_dca & et_mca < et_dca	2	Cosa ha fatto la persona nel video quando l'altra persona v_imp_1 obj_1 adjunct conj v_imp_2 obj_2 ?
6a	a	a	et_mca <= st_dca	1	Cosa stava facendo la persona nel video prima di v_inf_1 obj_1 adjunct conj v_inf_2 obj_2 ?
6b	a	a	et_mca <= st_dca	2	Cosa stava facendo la persona nel video prima che l'altra persona v_imp_cong_1 obj_1 adjunct conj v_imp_cong_2 obj_2 ?
7a	a	t	et_mca <= st_dca	1	Cosa stava facendo la persona nel video prima di v_inf_1 obj_1 adjunct conj v_inf_2 obj_2 ?
7b	a	t	et_mca <= st_dca	2	Cosa stava facendo la persona nel video prima che l'altra persona v_imp_cong_1 obj_1 adjunct conj v_inf_2 obj_2 ?
8a1	a	a	st_mca > st_dca & st_mca < et_dca; st_mca == st_dca; st_mca < st_dca & et_mca > st_dca	1	Cosa faceva la persona nel video mentre v_imp_prog_1 obj_1 adjunct conj v_imp_prog_2 obj_2 ?
8a2	a	a	st_mca > st_dca & st_mca < et_dca; st_mca == st_dca; st_mca < st_dca & et_mca > st_dca	1	Cosa faceva la persona nel video quando v_imp_prog_1 obj_1 adjunct conj v_imp_prog_2 obj_2 ?
8b1	a	a	st_mca > st_dca & st_mca < et_dca; st_mca == st_dca; st_mca < st_dca & et_mca > st_dca	2	Cosa faceva la persona nel video mentre l'altra persona v_imp_prog_1 obj_1 adjunct conj v_imp_prog_2 obj_2 ?
8b2	a	a	st_mca > st_dca & st_mca < et_dca; st_mca == st_dca; st_mca < st_dca & et_mca > st_dca	2	Cosa faceva la persona nel video quando l'altra persona v_imp_prog_1 obj_1 adjunct conj v_imp_prog_2 obj_2 ?
9a	a	t	st_mca < et_dca & et_mca > et_dca	1	Cosa faceva la persona nel video quando v_imp_1 obj_1 adjunct conj v_imp_2 obj_2 ?
9b	a	t	st_mca < et_dca & et_mca > et_dca	2	Cosa faceva la persona nel video quando l'altra persona v_imp_1 obj_1 adjunct conj v_imp_2 obj_2 ?
10a	a	a	st_mca >= et_dca	1	Cosa faceva la persona nel video dopo v_inf_pass_pross_1 obj_1 adjunct conj v_inf_pass_pross_2 obj_2 ?
10b	a	a	st_mca >= et_dca	2	Cosa faceva la persona nel video dopo che l'altra persona v_trapass_pross_1 obj_1 adjunct conj v_trapass_pross_2 obj_2 ?
12a1	t	t	et_mca == et_dca	1	Cosa fece la persona nel video quando v_pass_rem_1 obj_1 adjunct conj v_pass_rem_2 obj_2 ?
12a2	t	t	et_mca == et_dca	1	Cosa fece la persona nel video nel momento in cui v_pass_rem_1 obj_1 adjunct conj v_pass_rem_2 obj_2 ?
12b1	t	t	et_mca == et_dca	2	Cosa fece la persona nel video quando l'altra persona v_pass_rem_1 obj_1 adjunct conj v_pass_rem_2 obj_2 ?
12b2	t	t	et_mca == et_dca	2	Cosa fece la persona nel video nel momento in cui l'altra persona v_pass_rem_1 obj_1 adjunct conj v_pass_rem_2 obj_2 ?

Table 8: Russian Templates: _imp_ is imperfective past, _perf_ —perfective past. Conj is a coordinating conjunction in case of several verbs in one class, as in *читал или писал*.

Template	MCA	DCA	Condition	n_persons	Question
1a	t	a	st_mca >= et_dca; st_mca < et_dca & et_mca > et_dca	1	Что сделал человек на видео после того, как v_imp_1 obj_1 adjunct conj v_imp_2 obj_2 ?
1b	t	a	st_mca >= et_dca; st_mca < et_dca & et_mca > et_dca	2	Что сделал человек на видео после того, как другой человек v_imp_1 obj_1 adjunct conj v_imp_2 obj_2 ?
2a1	t	t	st_mca >= et_dca; st_mca < et_dca & et_mca > et_dca	1	Что сделал человек на видео после того, как v_perf_1 obj_1 adjunct conj v_perf_2 obj_2 ?
2a2	t	t	st_mca >= et_dca; st_mca < et_dca & et_mca > et_dca	1	Что сделал человек на видео, когда v_perf_1 obj_1 adjunct conj v_perf_2 obj_2 ?
2b1	t	t	st_mca >= et_dca; st_mca < et_dca & et_mca > et_dca	2	Что сделал человек на видео после того, как другой человек v_perf_1 obj_1 adjunct conj v_perf_2 obj_2 ?
2b2	t	t	st_mca >= et_dca; st_mca < et_dca & et_mca > et_dca	2	Что сделал человек на видео, когда другой человек v_perf_1 obj_1 adjunct conj v_perf_2 obj_2 ?
3a	t	a	et_mca <= st_dca	1	Что сделал человек на видео до того, как v_imp_1 obj_1 adjunct conj v_imp_2 obj_2 ?
3b	t	a	et_mca <= st_dca	2	Что сделал человек на видео перед тем, как другой человек v_imp_1 obj_1 adjunct conj v_imp_2 obj_2 ?
4a	t	t	et_mca <= st_dca	1	Что сделал человек на видео до того, как v_perf_1 obj_1 adjunct conj v_perf_2 obj_2 ?
4b	t	t	et_mca <= st_dca	2	Что сделал человек на видео перед тем, как другой человек v_perf_1 obj_1 adjunct conj v_perf_2 obj_2 ?
5a1	t	a	et_mca > st_dca & et_mca < et_dca	1	Что сделал человек на видео, пока он v_imp_1 obj_1 adjunct conj v_imp_2 obj_2 ?
5a2	t	a	et_mca > st_dca & et_mca < et_dca	1	Что сделал человек на видео, когда v_imp_1 obj_1 adjunct conj v_imp_2 obj_2 ?
5b1	t	a	et_mca > st_dca & et_mca < et_dca	2	Что сделал человек на видео, пока другой человек v_imp_1 obj_1 adjunct conj v_imp_2 obj_2 ?
5b2	t	a	et_mca > st_dca & et_mca < et_dca	2	Что сделал человек на видео, когда другой человек v_imp_1 obj_1 adjunct conj v_imp_2 obj_2 ?
6a	a	a	et_mca <= st_dca	1	Что делал человек на видео до того, как он v_imp_1 obj_1 adjunct conj v_imp_2 obj_2 ?
6b	a	a	et_mca <= st_dca	2	Что делал человек на видео до того, как другой человек v_imp_1 obj_1 adjunct conj v_imp_2 obj_2 ?
7a	a	t	et_mca <= st_dca	1	Что делал человек на видео до того, как он v_perf_1 obj_1 adjunct conj v_perf_2 obj_2 ?
7b	a	t	et_mca <= st_dca	2	Что делал человек на видео до того, как другой человек v_perf_1 obj_1 adjunct conj v_perf_2 obj_2 ?
8a1	a	a	st_mca > st_dca & st_mca < et_dca; st_mca == st_dca; st_mca < st_dca & et_mca > st_dca	1	Что делал человек на видео, пока v_imp_1 obj_1 adjunct conj v_imp_2 obj_2 ?
8a2	a	a	st_mca > st_dca & st_mca < et_dca; st_mca == st_dca; st_mca < st_dca & et_mca > st_dca	1	Что делал человек на видео, когда v_imp_1 obj_1 adjunct conj v_imp_2 obj_2 ?
8b1	a	a	st_mca > st_dca & st_mca < et_dca; st_mca == st_dca; st_mca < st_dca & et_mca > st_dca	2	Что делал человек на видео, пока другой человек v_imp_1 obj_1 adjunct conj v_imp_2 obj_2 ?
8b2	a	a	st_mca > st_dca & st_mca < et_dca; st_mca == st_dca; st_mca < st_dca & et_mca > st_dca	2	Что делал человек на видео, когда другой человек v_imp_1 obj_1 adjunct conj v_imp_2 obj_2 ?
9a	a	t	st_mca < et_dca & et_mca > et_dca	1	Что делал человек на видео, когда он v_perf_1 obj_1 adjunct conj v_perf_2 obj_2 ?
9b	a	t	st_mca < et_dca & et_mca > et_dca	2	Что делал человек на видео, когда другой человек v_perf_1 obj_1 adjunct conj v_perf_2 obj_2 ?
10a	a	a	st_mca >= et_dca	1	Что делал человек на видео после того, как он v_imp_1 obj_1 adjunct conj v_imp_2 obj_2 ?
10b	a	a	st_mca >= et_dca	2	Что делал человек на видео после того, как другой человек v_imp_1 obj_1 adjunct conj v_imp_2 obj_2 ?
11a	a	t	st_mca >= et_dca; st_mca < et_dca & et_mca > et_dca	1	Что делал человек на видео после того, как он v_perf_1 obj_1 adjunct conj v_perf_2 obj_2 ?
11b	a	t	st_mca >= et_dca; st_mca < et_dca & et_mca > et_dca	2	Что делал человек на видео после того, как другой человек v_perf_1 obj_1 adjunct conj v_perf_2 obj_2 ?
12a1	t	t	et_mca == et_dca	1	Что сделал человек на видео, когда v_perf_1 obj_1 adjunct conj v_perf_2 obj_2 ?
12a2	t	t	et_mca == et_dca	1	Что сделал человек на видео в тот момент, когда v_perf_1 obj_1 adjunct conj v_perf_2 obj_2 ?
12b1	t	t	et_mca == et_dca	2	Что сделал человек на видео, когда другой человек v_perf_1 obj_1 adjunct conj v_perf_2 obj_2 ?
12b2	t	t	et_mca == et_dca	2	Что сделал человек на видео в тот момент, когда другой человек v_perf_1 obj_1 adjunct conj v_perf_2 obj_2 ?

Table 9: Japanese templates: ta_form is the general past form, inf is the dictionary form, te_form is the conjunction form, i_form is the present progressive, and imp is the imperfect. Conj is a coordinating conjunction in case of several verbs in one class, as in ノートパソコンで仕事をしていたか、または遊んでいた.

Template	MCA	DCA	Condition	n_persons	Question
1a	t	a	st_mca >= et_dca; st_mca < et_dca & et_mca > et_dca	1	ビデオに写っている人は adj obj を ta_form 後、何をしましたか？
1b	t	a	st_mca >= et_dca; st_mca < et_dca & et_mca > et_dca	2	ビデオに写っている人は、他の人が adj obj を ta_form 後、何をしましたか？
2a1	t	t	st_mca >= et_dca; st_mca < et_dca & et_mca > et_dca	1	ビデオに写っている人は adj obj を ta_form 後、何をしましたか？
2a2	t	t	st_mca >= et_dca; st_mca < et_dca & et_mca > et_dca	1	ビデオに写っている人は adj obj を ta_form 時、何をしましたか？
2b1	t	t	st_mca >= et_dca; st_mca < et_dca & et_mca > et_dca	2	ビデオに写っている人は、他の人が adj obj を ta_form 後、何をしましたか？
2b2	t	t	st_mca >= et_dca; st_mca < et_dca & et_mca > et_dca	2	ビデオに写っている人は、他の人が adj obj を ta_form 時、何をしましたか？
3a	t	a	et_mca <= st_dca	1	ビデオに写っている人は adj obj を inf 前に、何をしましたか？
3b	t	a	et_mca <= st_dca	2	ビデオに写っている人は、他の人が adj obj を inf 前に、何をしましたか？
4a	t	t	et_mca <= st_dca	1	ビデオに写っている人は adj obj を inf 前に、何をしましたか？
4b	t	t	et_mca <= st_dca	2	ビデオに写っている人は、他の人が adj obj を inf 前に、何をしましたか？
5a1	t	a	et_mca > st_dca & et_mca < et_dca	1	ビデオに写っている人は adj obj を te_form 、何をしましたか？
5a2	t	a	et_mca > st_dca & et_mca < et_dca	1	ビデオに写っている人は adj obj を i_form ながら、何をしましたか？
5b1	t	a	et_mca > st_dca & et_mca < et_dca	2	ビデオに写っている人は、他の人が adj obj を inf_imp 間に、何をしましたか？
5b2	t	a	et_mca > st_dca & et_mca < et_dca	2	ビデオに写っている人は、他の人が adj obj を ta_form 時、何をしましたか？
6a	a	a	et_mca <= st_dca	1	ビデオに写っている人は adj obj を inf 前に、何をしていましたか？
6b	a	a	et_mca <= st_dca	2	ビデオに写っている人は、他の人が adj obj を inf_imp 前に、何をしていましたか？
7a	a	t	et_mca <= st_dca	1	ビデオに写っている人は adj obj を inf 前に、何をしていましたか？
7b	a	t	et_mca <= st_dca	2	ビデオに写っている人は、他の人が adj obj を inf 前に、何をしていましたか？
8a1	a	a	st_mca > st_dca & st_mca < et_dca; st_mca = st_dca; st_mca < st_dca & et_mca > st_dca	1	ビデオに写っている人は adj obj を i_form ながら、何をしていましたか？
8a2	a	a	st_mca > st_dca & st_mca < et_dca; st_mca = st_dca; st_mca < st_dca & et_mca > st_dca	1	ビデオに写っている人は adj obj を ta_form 時、何をしていましたか？
8b1	a	a	st_mca > st_dca & st_mca < et_dca; st_mca = st_dca; st_mca < st_dca & et_mca > st_dca	2	ビデオに写っている人は、他の人が adj obj を inf_imp 間に、何をしていましたか？
8b2	a	a	st_mca > st_dca & st_mca < et_dca; st_mca = st_dca; st_mca < st_dca & et_mca > st_dca	2	ビデオに写っている人は、他の人が adj obj を ta_form 時、何をしていましたか？
9a	a	t	st_mca < et_dca & et_mca > et_dca	1	ビデオに写っている人は adj obj を ta_form 時、何をしていましたか？
9b	a	t	st_mca < et_dca & et_mca > et_dca	2	ビデオに写っている人は、他の人が adj obj を ta_form 時、何をしていましたか？
10a	a	a	st_mca >= et_dca	1	ビデオに写っている人は adj obj を imp 後、何をしていましたか？
10b	a	a	st_mca >= et_dca	2	ビデオに写っている人は、他の人が adj obj を imp 後、何をしていましたか？
11a	a	t	st_mca >= et_dca; st_mca < et_dca & et_mca > et_dca	1	ビデオに写っている人は adj obj を ta_form 後、何をしていましたか？
11b	a	t	st_mca >= et_dca; st_mca < et_dca & et_mca > et_dca	2	ビデオに写っている人は、他の人が adj obj を ta_form 後、何をしていましたか？
12a1	t	t	et_mca = et_dca	1	ビデオに写っている人は adj obj を ta_form 時、何をしましたか？
12a2	t	t	et_mca = et_dca	1	ビデオに写っている人は adj obj を ta_form 瞬間、何をしましたか？
12b1	t	t	et_mca = et_dca	2	ビデオに写っている人は、他の人が adj obj を ta_form 時、何をしましたか？
12b2	t	t	et_mca = et_dca	2	ビデオに写っている人は、他の人が adj obj を ta_form 瞬間、何をしましたか？



Q: What did the person in the video do after drinking from a cup or glass or bottle?

a0: The person was sitting on the floor.

a1: The person awakened in bed.

a2: The person closed the refrigerator.

a3: The person was closing the refrigerator.

Q: Cosa ha fatto la persona nel video dopo aver bevuto da una tazza o un bicchiere o una bottiglia?

a0: La persona era seduta sul pavimento

a1: La persona si è svegliata nel letto

a2: La persona ha chiuso il frigorifero

a3: La persona chiudeva il frigorifero

Q: Что сделал человек на видео после того, как он пил из чашки или стакана или бутылки?

a0: Он сидел на полу.

a1: Он проснулся в кровати.

a2: Он закрыл холодильник.

a3: Он закрывал холодильник.

Q: ビデオに写っている人はカップかグラスか、またはボトルから飲んでいた後、何をしましたか？

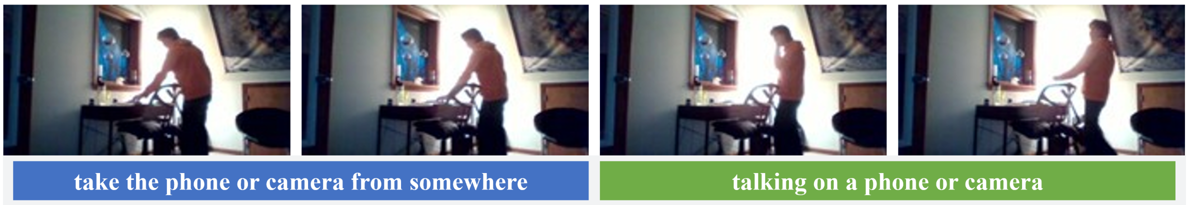
a0: その人は床に座っていた。

a1: その人はベッドで目を覚ました。

a2: その人は冷蔵庫を閉めた。

a3: その人は冷蔵庫を閉めていた。

Figure 9: Example from Perfect Times generated by Template 1a for all the languages.


take the phone or camera from somewhere
talking on a phone or camera

Q: What was the person in the video doing after taking the phone or camera from somewhere?

a0: The person was washing something with a towel.

a1: The person was talking on a phone or camera.

a2: The person talked on a phone or camera.

a3: The person put the phone or camera somewhere.

Q: Cosa faceva la persona nel video dopo aver preso il telefono oppure la fotocamera da qualche parte?

a0: La persona lavava qualcosa con un asciugamano

a1: La persona parlava al telefono oppure alla fotocamera

a2: La persona ha parlato al telefono oppure alla fotocamera

a3: La persona ha messo il telefono oppure la fotocamera da qualche parte

Q: Что делал человек на видео после того, как он взял телефон или камеру откуда-то?

a0: Он протирал что-то полотенцем.

a1: Он разговаривал по телефону или камере.

a2: Он поговорил по телефону или камере.

a3: Он положил телефон или камеру куда-то.

Q: ビデオに写っている人はどこかから携帯電話かカメラを取った後、何をしていましたか？

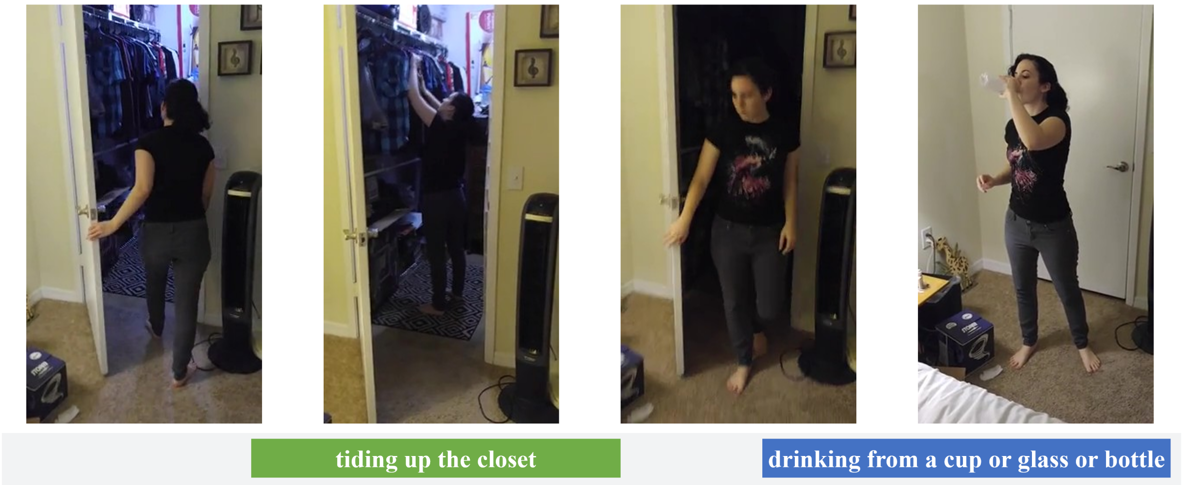
a0: その人は何かをタオルで洗っていた。

a1: その人は携帯電話かカメラで話していた。

a2: その人は携帯電話かカメラで話した。

a3: その人はどこかに携帯電話かカメラを置いた。

Figure 10: Example from Perfect Times generated by Template 11a for all the languages.



Q: What had the person in the video been doing before drinking from a cup or glass or bottle?

a0: The person had been holding the laptop.

a2: The person had been tidying up the closet or cabinet.

a1: The person had tidied up the closet or cabinet.

a3: The person had closed the closet or cabinet.

Q: Cosa stava facendo la persona nel video prima di bere da una tazza o un bicchiere o una bottiglia?

a0: La persona stava tenendo il laptop

a2: La persona stava riordinando l'armadio oppure i mobili

a1: La persona aveva riordinato l'armadio oppure i mobili

a3: La persona aveva chiuso l'armadio oppure i mobili

Q: Что делал человек на видео до того, как пил из чашки или стакана или бутылки?

a0: Он держал ноутбук.

a2: Он прибирал в шкафу или шкафчике.

a1: Он убрал в шкафу или шкафчике.

a3: Он закрыл шкаф или шкафчик.

Q: ビデオに写っている人はカップかグラスか、またはボトルから飲んでいる前に、何をしていましたか？

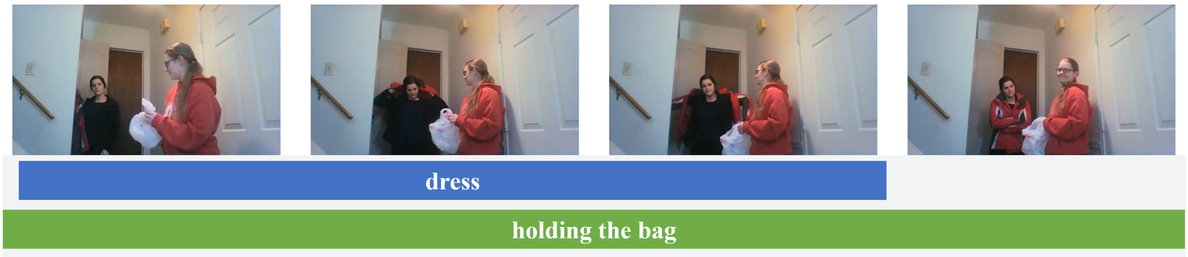
a0: その人はノートパソコンを持っていた

a2: その人はクローゼットかキャビネットを整理していた。

a1: その人はクローゼットかキャビネットを整理した。

a3: その人はクローゼットかキャビネットを閉めた。

Figure 11: Example from Perfect Times generated by Template 6a for all the languages.



Q: What was the person in the video doing when the other person was dressing?

a0: The person was holding the bag.

a2: The person was holding the food.

a1: The person held the bag.

a3: The person took the clothes from somewhere.

Q: Cosa faceva la persona nel video quando l'altra persona si stava vestendo?

a0: La persona teneva la borsa

a2: La persona teneva il cibo

a1: La persona ha tenuto la borsa

a3: La persona ha preso i vestiti da qualche parte

Q: Что делал человек на видео, когда другой человек одевался?

a0: Он держал сумку.

a2: Он держал еду.

a1: Он подержал сумку.

a3: Он взял одежду откуда-то.

Q: ビデオに写っている人は、他の人が服を着た時、何をしていましたか？

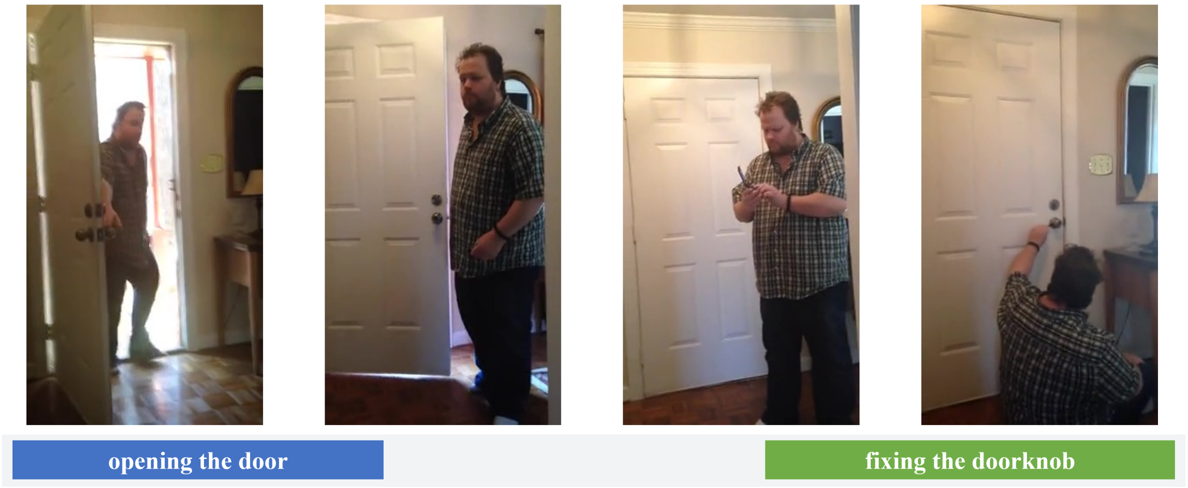
a0: その人はバッグを持っていた。

a2: その人は食べ物を持っていた。

a1: その人はバッグを持った。

a3: その人はどこかから服を取った。

Figure 12: Example from Perfect Times generated by Template 8b2 for all the languages.



Q: What did the person in the video do after they had opened the door?

a0: The person fixed the doorknob.
a1: The person tidied the towel or towels.
a2: The person was opening the door.
a3: The person was fixing the doorknob.

Q: Cosa fece la persona nel video dopo aver aperto la porta?

a0: La persona riparò la maniglia della porta
a1: La persona riordinò l'asciugamano oppure gli asciugamani
a2: La persona stava aprendo la porta
a3: La persona stava riparando la maniglia della porta

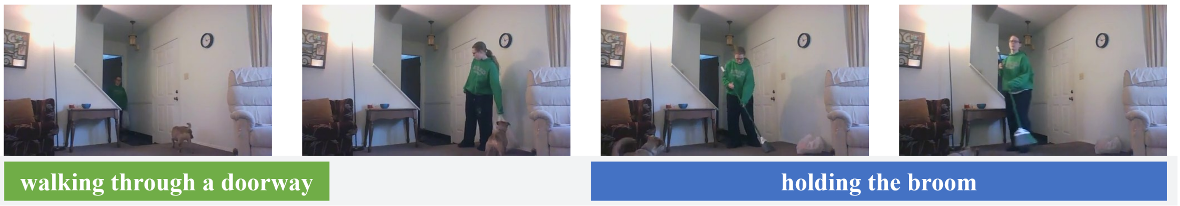
Q: Что сделал человек на видео после того, как открыл дверь?

a0: Он починил дверную ручку.
a1: Он убрал полотенце или полотенца.
a2: Он открывал дверь.
a3: Он чинил дверную ручку.

Q: ビデオに写っている人はくしゃみをしながら、何をしていましたか？

a0: その人は服を持った。
a1: その人は紙かノートで作業した。
a2: その人は紙かノートで作業していた。
a3: その人はどこかにほうきを置いていた。

Figure 13: Example from Perfect Times generated by Template 2a1 for all the languages.



Q: What had the person in the video done before holding the broom?

a0: The person had taken off the shoes.
a1: The person had been putting the groceries somewhere.
a2: The person had been walking through a doorway.
a3: The person had walked through a doorway.

Q: Cosa aveva fatto la persona nel video prima di tenere la scopa?

a0: La persona si era tolta le scarpe
a1: La persona stava mettendo la spesa da qualche parte
a2: La persona stava passando attraverso un'apertura
a3: La persona era passata attraverso un'apertura

Q: Что сделал человек на видео до того, как он держал швабру?

a0: Он снял обувь.
a1: Он клал продукты куда-то.
a2: Он шел через дверной проем.
a3: Он прошел через дверной проем.

Q: ビデオに写っている人はほうきを持っている前に、何をしましたか？

a0: その人は靴を脱いだ。
a1: その人はどこかに食料品を置いていた。
a2: その人はドアを通っていた。
a3: その人はドアを通った。

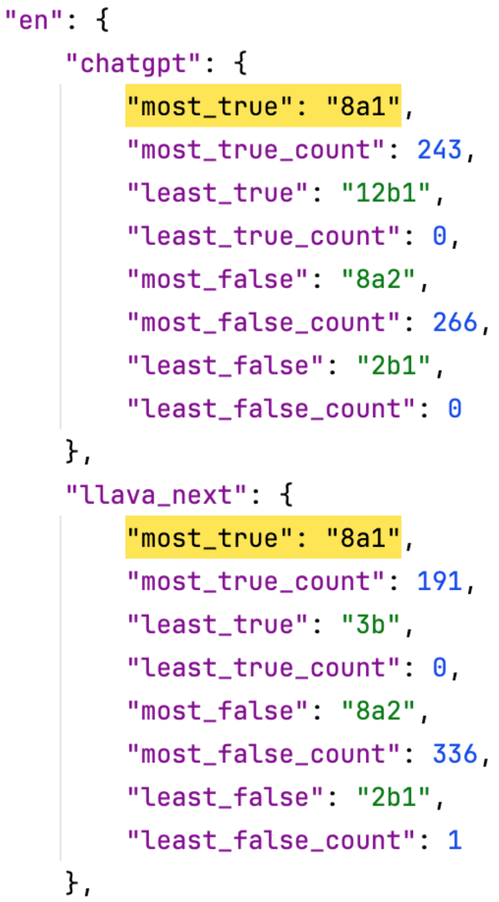
Figure 14: Example from Perfect Times generated by Template 3a for all the languages.

Table 10: Annotator accuracy and distractor type distribution.

Annotator	Accuracy	Distractor Type 1	Distractor Type 2	Distractor Type 3
English	93.40	45.62	50.88	3.50
Italian	97.84	20.83	12.50	66.67
Russian	83.83	34.57	46.91	18.52
Japanese	98.36	20.00	40.00	40.00

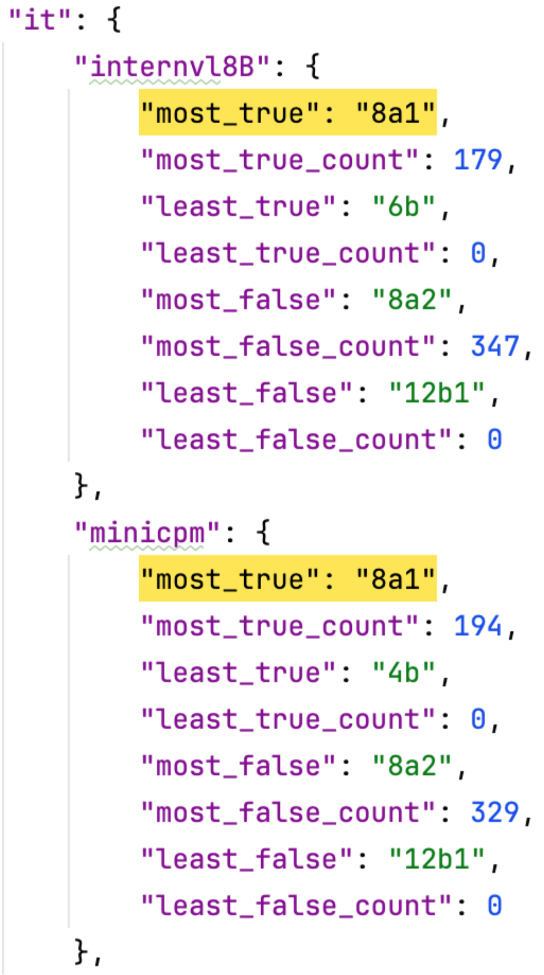
Table 11: Model configurations and references.

Model	Vision Encoder	Language Model	Parameters
LLaVA-NeXT-Video-7B	SigLIP-400M (Zhai et al., 2023)	Qwen 1.5 (Bai et al., 2023)	7B
MiniCPM-V-2_6	SigLIP-400M (Zhai et al., 2023)	Qwen2 (Bai et al., 2023)	7B
Qwen2-VL	Qwen2-VL (Wang et al., 2024)	Qwen2 (Bai et al., 2023)	7B
InternVL2	InternViT-300M-448px (Chen et al., 2024)	internlm2_5-7b-chat (Chen et al., 2024)	8B



```
"en": {
    "chatgpt": {
        "most_true": "8a1",
        "most_true_count": 243,
        "least_true": "12b1",
        "least_true_count": 0,
        "most_false": "8a2",
        "most_false_count": 266,
        "least_false": "2b1",
        "least_false_count": 0
    },
    "llava_next": {
        "most_true": "8a1",
        "most_true_count": 191,
        "least_true": "3b",
        "least_true_count": 0,
        "most_false": "8a2",
        "most_false_count": 336,
        "least_false": "2b1",
        "least_false_count": 1
    },
```

Figure 15: Statistics on while (8a1) vs. when (8a2) for English in GPT-4o and LLaVA-NeXT-Video.



```
"it": {
    "internvl8B": {
        "most_true": "8a1",
        "most_true_count": 179,
        "least_true": "6b",
        "least_true_count": 0,
        "most_false": "8a2",
        "most_false_count": 347,
        "least_false": "12b1",
        "least_false_count": 0
    },
    "minicpm": {
        "most_true": "8a1",
        "most_true_count": 194,
        "least_true": "4b",
        "least_true_count": 0,
        "most_false": "8a2",
        "most_false_count": 329,
        "least_false": "12b1",
        "least_false_count": 0
    },
```

Figure 16: Statistics on *mentre* (8a1) vs. *quando* (8a2) for Italian in InternVL2 and MiniCPM-V.

Table 12: The proportion of correctly predicted answers with respect to telicity in percentage.

	Correct Telic	**Incorrect Telic**	**Correct Atelic**	**Incorrect Atelic**
Gemini-2.0-flash-lite	42.97	57.03	41.97	58.03
Qwen2-VL-7B-Instruct	42.80	57.20	32.95	67.05
MiniCPM-V-2_6	41.40	58.60	33.75	66.25
GPT-4o	39.74	60.26	43.45	56.55
InternVL2-8B	34.26	65.74	29.82	70.18
LLaVA-NeXT-Video-7B	28.03	71.97	27.62	72.38

Table 13: Number of questions answered correctly by all models for each language. The 5 Models Column shows the results for all models except LLaVA-NeXT-Video-7B.

Language	**5 Models**	**All Models**
English	230	81
Italian	249	14
Russian	211	49
Japanese	154	26

Table 14: Number of questions answered incorrectly by all models for each language. The 5 Models Column shows the results for all models except LLaVA-NeXT-Video-7B.

Language	**5 Models**	**All Models**
English	857	608
Italian	808	456
Russian	740	512
Japanese	850	554

Table 15: Top-5 best answered templates by language across all models. Templates for questions about the following event are in blue, templates for questions about preceding events in red, and templates for questions about simultaneous events in green.

	1	**2**	**3**	**4**	**5**
English	1b	2b1	1a	9b	12a1
Italian	12b1	1b	9b	5b1	6a
Russian	1b	12b1	9b	3b	8b2
Japanese	12b2	2b1	1b	9b	1a

Table 16: Top-5 worst answered templates by language across all models. Templates for questions about the following event are in blue, templates for questions about preceding events in red, and templates for questions about simultaneous events in green.

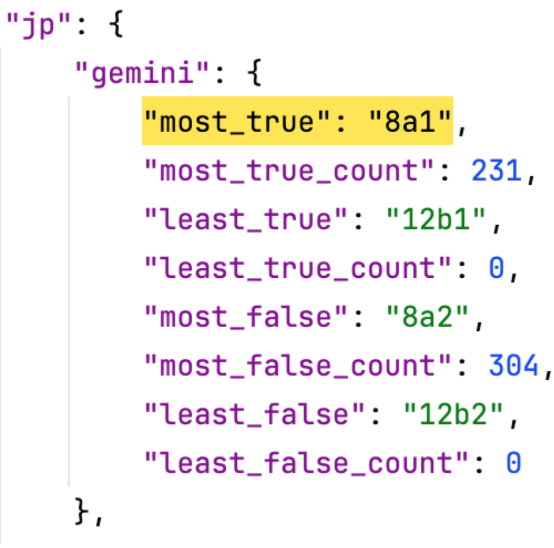
	1	**2**	**3**	**4**	**5**
English	12b1	6b	12b2	10a	11a
Italian	3b	4b	12b2	7b	2b2
Russian	7b	10b	6b	4b	2b2
Japanese	7b	4b	6b	10a	11a

Table 17: Top-5 best answered templates by model across all languages. Templates for questions about the following event are in blue, templates for questions about preceding events in red, and templates for questions about simultaneous events in green.

	1	**2**	**3**	**4**	**5**
GPT-4o	9b	2b2	8b2	2b1	5b1
Gemini-2.0-flash-lite	1b	9b	1a	6a	8b1
Qwen2-VL-7B-Instruct	12b1	12a1	2b1	1a	12a2
MiniCPM-V-2_6	12b1	1b	2b1	9b	5b2
InternVL2-8B	12b1	12b2	3b	1b	10b

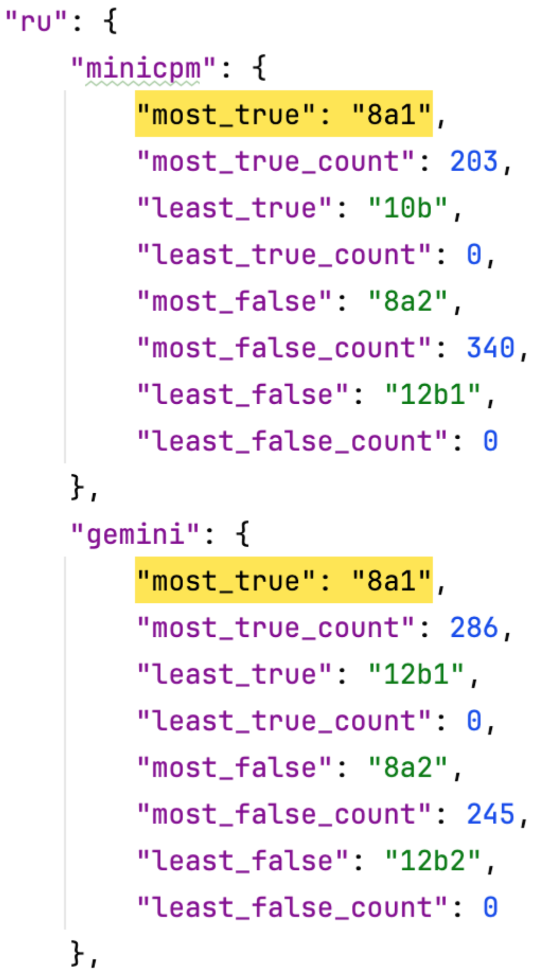
Table 18: Top-5 worst answered templates by model across all languages. Templates for questions about the following event are in blue, templates for questions about preceding events in red, and templates for questions about simultaneous events in green.

	1	**2**	**3**	**4**	**5**
GPT-4o	12b1	4b	10a	7b	5a1
Gemini-2.0-flash-lite	7b	3b	12b1	6b	10a
Qwen2-VL-7B-Instruct	2b2	7b	10b	11a	11b
MiniCPM-V-2_6	6b	10b	7b	4b	3b
InternVL2-8B	4b	7b	9a	2b2	11b



```
"jp": {
    "gemini": {
        "most_true": "8a1",
        "most_true_count": 231,
        "least_true": "12b1",
        "least_true_count": 0,
        "most_false": "8a2",
        "most_false_count": 304,
        "least_false": "12b2",
        "least_false_count": 0
    },
```

Figure 17: Statistics on ながら (8a1) vs. 時 (8a2) for Japanese in Gemini-2.0-Flash-Lite.



```
"ru": {
    "minicpm": {
        "most_true": "8a1",
        "most_true_count": 203,
        "least_true": "10b",
        "least_true_count": 0,
        "most_false": "8a2",
        "most_false_count": 340,
        "least_false": "12b1",
        "least_false_count": 0
    },
    "gemini": {
        "most_true": "8a1",
        "most_true_count": 286,
        "least_true": "12b1",
        "least_true_count": 0,
        "most_false": "8a2",
        "most_false_count": 245,
        "least_false": "12b2",
        "least_false_count": 0
    },
```

Figure 18: Statistics on (8a1) vs. (8a2) for Russian in MiniCPM-V and Gemini-2.0-Flash-Lite.