

MAIA: a Benchmark for Multimodal AI Assessment

Davide Testa^{1,2,*}, Giovanni Bonetta², Raffaella Bernardi³, Alessandro Bondielli^{4,5},
Alessandro Lenci⁵, Alessio Miaschi⁶, Lucia Passaro⁴ and Bernardo Magnini²

¹Università di Roma La Sapienza, Roma

²Fondazione Bruno Kessler (FBK), Trento

³Free University of Bozen-Bolzano, Bolzano

⁴Department of Computer Science, University of Pisa, Pisa

⁵Department of Philology, Literature and Linguistics, University of Pisa, Pisa

⁶Istituto di Linguistica Computazionale "A. Zampolli" (CNR-ILC), ItaliaNLP Lab, Pisa

Abstract

We introduce MAIA (Multimodal AI Assessment), a multimodal dataset developed as a core component of a competence-oriented benchmark designed for fine-grained investigation of the reasoning abilities of Visual Language Models (VLMs) on videos. The MAIA benchmark is characterized by several distinctive features. To the best of our knowledge, MAIA is the first Italian-native benchmark addressing video understanding: videos were carefully selected to reflect Italian culture, and the language data (i.e., questions and reference answers) were produced by native-Italian speakers. Second, MAIA explicitly includes twelve reasoning categories that are specifically designed to assess the reasoning abilities of VLMs on videos. Third, we structured the dataset to support two aligned tasks (i.e., a *statement verification* and an *open-ended visual question answering*) built on the same datapoints, this way allowing to assess VLM coherence across task formats. Finally MAIA integrates, by design, state-of-the-art LLMs in the development process of the benchmark, taking advantage of their linguistic and reasoning capabilities both for data augmentation and for assessing and improving the overall quality of the data. In the paper we focus on the design principles and the data collection methodology, highlighting how MAIA provides a significant advancement with respect to other available dataset for VLM benchmarking. Data available at [GitHub](#).

Keywords

Multimodality, Benchmarking, Vision-Language Models, Multimodal Reasoning, Language Resources

1. Introduction

In recent years, mainly following the success of large language models (LLMs), there has been a growing interest for large pre-trained models able to manage both texts and images. Such Vision and Language models (VLMs) have been investigated both from a theoretical perspective (e.g., Baroni [1]) and for their application-oriented interest (e.g., Bigam et al. [2]). Today, there are dozens of available VLMs, and the most popular *families* of generative AI models (e.g., Llama, Gemma, Qwen, GPT) include several VLMs, which can address a number of question answering tasks on both images and videos. As a consequence of the fast and increasing power of

VLMs, assessing their performance on standardized tasks and metrics is becoming more and more challenging.

First of all, evaluating VLMs understanding in real world scenarios requires moving beyond single-frame scenarios. Unlike static images, videos offer rich temporal structure: they capture dynamic scenes, evolving actions, interactions, and causal dependencies that unfold over time, making them one of the most faithful and closest approximations to real-world complex scenarios. In this context, the role of evaluation becomes critical: to truly assess a model's ability to understand, reason, and ground meaning across modalities, we need benchmarks that do not merely test task performance, but probe the underlying competences of the model [3].

With this purpose in mind, we introduce MAIA (Multimodal AI Assessment), a multimodal dataset developed as a core component of a broader competence-oriented evaluation framework for VLMs. MAIA is designed to challenge models on multimodal reasoning grounded in real-world scenarios from different linguistic perspectives. To the best of our knowledge, it is the first native Italian evaluation dataset of its kind and based on video content. MAIA provides a linguistically rich and semantically diverse resource for exploring vision and language understanding in realistic contexts, with a particular focus on Italian culture, by covering distinct reasoning categories, each targeting specific semantic phenomena. This

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 – 26, 2025, Cagliari, Italy

*Corresponding author.

✉ dtesta@fbk.eu (D. Testa); gbonetta@fbk.eu (G. Bonetta);

raffaella.bernardi@unibz.it (R. Bernardi);

alessandro.bondielli@unipi.it (A. Bondielli);

alessandro.lenci@unipi.it (A. Lenci); alessio.miaschi@ilc.cnr.it

(A. Miaschi); lucia.passaro@unipi.it (L. Passaro); magnini@fbk.eu

(B. Magnini)

ORCID: 0009-0002-2489-5323 (D. Testa); 0000-0003-4498-1026

(G. Bonetta); 0000-0002-3423-1208 (R. Bernardi);

0000-0003-3426-6643 (A. Bondielli); 0000-0001-5790-4308 (A. Lenci);

0000-0002-0736-5411 (A. Miaschi); 0000-0003-4934-5344

(L. Passaro); 0000-0002-0740-5778 (B. Magnini)

© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

structure allows for a fine-grained analysis of the contribution of both language and visual modalities across different types of reasoning. A key feature of MAIA is its cascading data collection approach, which enables the same source data to be spent and used across multiple task formats (e.g., generative tasks, classification tasks, etc.), supporting fully comparable evaluations and paving the way for an *all-in-one* benchmarking strategy. The efficacy of this approach and of the MAIA benchmark as a severe and robust evaluation framework has been proved in Testa et al. [4] in which we evaluate models against a classification and a generative task, namely visual statement verification and open-ended question answering. While the second task turns out to be more challenging even for the best-performing models, they also exhibit significant inconsistencies both within and across the two tasks, with some categories relying more heavily on either the visual or the linguistic component to solve the task. However, in this paper, we dive into how the dataset was collected. Finally, an additional aspect of innovation in the data creation of MAIA pipeline lies in the integration of human annotation with targeted data augmentation using powerful LLMs (*GPT-4o* [5]), combined with a multi-stage semi-automatic validation process conducted with the same model at different levels. This dual use of a generative model (i.e., *GPT-4o*) not only enhances the diversity and coverage of the dataset but also ensures high-quality and semantically consistent data throughout the pipeline.

The paper is organized as follows. Section 2 reviews the most relevant prior work in the research area. In Section 3, we detail the design choices behind the creation of the dataset and, more broadly, the development of the entire MAIA benchmark. Finally, Sections 4 and 5 describe the specific steps followed for dataset construction: the former focuses on the selection and collection of video material, while the latter addresses the collection and validation of all linguistic data that constitute MAIA. Both sections are complemented by dedicated analyses of the collected data.

2. Related Work

Multimodal datasets combining vision and language have played a crucial role in the development and evaluation of VLMs. Early image-based resources such as the *VQA* [6], *GQA* [7], *DVD* [8], and *HL* [9] datasets have provided controlled environments to assess visual reasoning and natural language understanding through several tasks like Image captioning or Visual Question Answering, thereby reinforcing the role of vision as a fundamental component in the evaluation of multimodal models [6]. Over time, contributions of this kind have been instrumental in shaping the foundations of multimodal evaluation, where

language understanding is assessed in conjunction with perceptual grounding. Simultaneously, these efforts have revealed critical weaknesses in early multimodal architectures, by highlighting their reliance on dataset biases or shallow heuristics rather than genuine visual reasoning [10, 11]. Such challenges have later been framed within the broader phenomenon of *Unimodal Collapse*, where a VLM disproportionately depends on its language component, resulting in text-only models performing comparably to their multimodal counterparts [12]. In contrast to earlier stages [13, 14, 15], the growing awareness of these issues has prompted the emergence of *diagnostic* evaluation frameworks such as in Parcalabescu et al. [12], Thrush et al. [16], Chen et al. [17], Bianchi et al. [18] and *carefully curated* benchmarks such as in Xiao et al. [19] and Tong et al. [20], designed to expose the true capabilities and limitations of VLMs. These methodological insights strongly motivate the design of MAIA as a robust, controlled multimodal dataset, aimed at ensuring that models genuinely integrate both linguistic and visual information, rather than relying solely on the priors embedded in their language backbones.

Building on this tradition, video-language datasets lately extended the challenge to temporal understanding and dynamic scene interpretation, both essential components for complex real-world understanding. Several resources including *TVQA* [21] and *HowToVQA* [22] datasets or the *AGQA* [23] and *MVBench* [24] benchmarks changed their focus from static perception to actions and entities, by trying to challenge VLMs in identifying the relationships between them. As in the case of image-based evaluation, early surveys have already stressed the need for careful and systematic assessment Zhong et al. [25]. While task-oriented benchmarks often report strong performance [26, 27], more fine-grained evaluations have revealed critical limitations [28], and competence-based analyses continue to highlight the substantial gap in the video understanding capabilities of VLMs [29]. In this context, MAIA contributes as a new video-language dataset aimed at evaluating VLMs not only on videos featuring temporal dynamics and meaningful content but also through a competence-oriented design that explores the interplay between language and vision, a dimension largely neglected in prior Video QA benchmarks.

Italian Multimodal Datasets. Most multimodal datasets are available in English, with only limited multilingual or other native-language resources, with Italian being consistently underrepresented. In the image domain, *GQA-it* dataset [30] is a notable attempt to adapt a visual question answering dataset into Italian. More recent benchmarks like *XGQA* [31] and *EXAMS-V* [32] include translated Italian multiple-choice questions, but lack original content and do not target high-level reasoning. MAIA fills this gap as the first Italian-native and

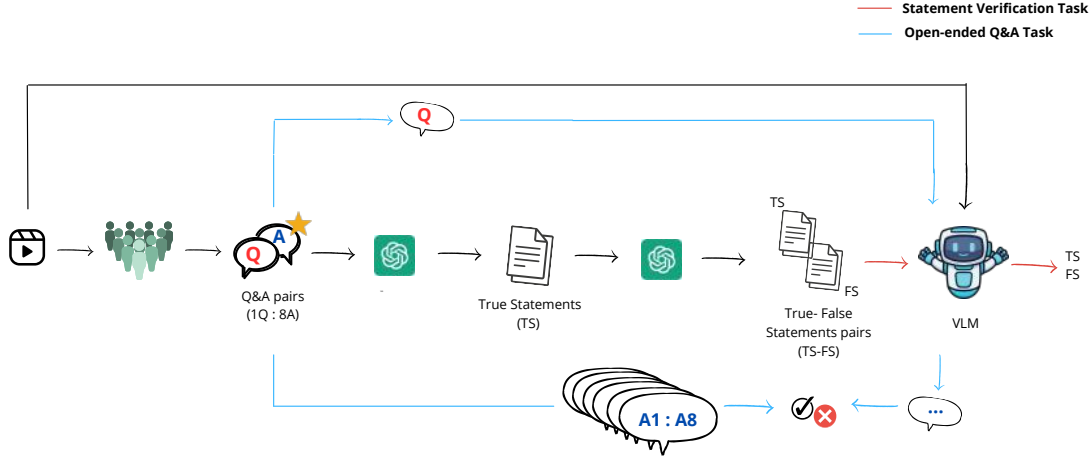


Figure 1: Workflow of the MAIA evaluation framework, integrating dataset construction with its application to the two aligned evaluation tasks used in Testa et al. [4].

video-language dataset specifically designed to assess complex visual reasoning and grounding.

3. MAIA: Benchmark Design

This section presents the design principles, structure, and construction pipeline of both the MAIA dataset and the benchmark built upon it. In line with this, Figure 1 illustrates the overall workflow adopted for dataset creation, embedding it with the broader architectural framework of the benchmark, which also includes the downstream tasks the data is designed to support.

As shown, the dataset creation begins with the collection of short videos, each associated with twelve high-level reasoning categories. These categories reflect different semantic phenomena and were chosen to ensure a rich and controlled testing environment for visual and linguistic reasoning. Based on these categories, we constructed our multimodal dataset by first collecting a set of questions that served as the conceptual backbone for the creation of the linguistic data, both manually (i.e. a set of answers) and automatically generated (i.e. True and False statements), as described in detail in Section 5. Figure 2 illustrates an example¹ of a MAIA item and highlights the cascading logic behind the data creation process. This architecture supports the development of two aligned

evaluation tasks: a *Visual Statement Verification* task, using paired true/false statements to assess the model’s ability to distinguish accurate from misleading content in a multiple-choice format, and an *open-ended Visual Question Answering* task, where each question is matched with eight different human answers serving as a reference set to evaluate the quality of the response generated by the VLM. Each task will test different aspects of visual understanding and reasoning, all grounded in the same set of videos and categories.

Table 1 presents the structure of the MAIA dataset after the data creation and validation process.

Feature	n	
Videos	100	
Semantic Categories	12	(9 Macro-Category)
Questions (Q)	2,400	(2 x Category x Video)
Answers (A)	19,200	(8-Answers pool x Q)
True Statements (TS)	19,200	
False Statements (FS)	19,200	

Table 1

Overview of the MAIA dataset composition.

3.1. Reasoning Categories

We defined 12 reasoning categories as the outcome of two pilot studies conducted with a group of expert volunteer

¹Although all source data are in Italian, examples are presented in English to enhance readability.

CATEGORY		QUESTION	ANSWER (1/8)	TRUE STATEMENT (1/8)	FALSE STATEMENT (1/8)
CAUSAL		Why is mozzarella melted?	The heat from the wood oven has melted it	Mozzarella is melted by the heat of the wood oven	Mozzarella is melted by the heat generated by the sun.
COUNTERFACTUAL		What would happen if the pizza chef dropped the pizza on the floor?	He would dirty the floor and would have to remake the pizza.	If the pizza chef dropped the pizza, he would dirty the floor and would have to remake the pizza.	If the pizza maker dropped the pizza, he would not dirty the floor and would not have to remake the pizza.
IMPLICIT	Partial	Is the person who rolls out the pizza the same one who puts it in the oven?	No, they are two different people.	In the scene, the person who rolls out the pizza dough and the one who puts it in the oven are two distinct figures.	In the scene, the person who rolls out the pizza dough and the one who puts it in the oven are the same person.
	Total	What is the function of all the wooden planks under the wood oven?	They have to feed the fire.	The wooden planks under the wood oven are for feeding the fire.	The wooden planks under the wood oven are for decoration.
UNCERTAINTY		On average, how many pizzas does the pizza chef bake each day?	I do not have enough data to know.	There is not enough data to determine the average number of pizzas a pizza maker cooks daily.	There is sufficient data to determine the average number of pizzas that the pizza maker cooks daily.
OUT-OF-SCOPE		What is the cake made of?	I cannot see any cake.	There is no cake in the video.	There is a cake in the video.
PLANNING		What steps should the pizza maker take to revive the fire?	He should stir up the embers a bit and throw some new wood.	To revive the fire, the pizza maker should stir the embers and add new wood.	To revive the fire, the pizza maker should stir the embers and add new water.
SENTIMENT		What attitude does the pizza maker show while taking the pizza out of the oven?	The pizzaiolo looks focused.	In the video, the pizza maker looks focused while taking the pizza out of the oven.	In the video, the pizza maker looks distracted while taking the pizza out of the oven.
SPATIAL	Partial	Where is the pizza placed after being taken out of the oven?	The pizza is placed on a plate.	After being taken out of the oven, the pizza is placed on a plate.	After being taken out of the oven, the pizza is placed on the table.
	Total	Where is the pizza maker?	In the pizzeria in front of the oven	In the scene, the pizza maker is in the pizzeria in front of the oven	In the scene, the pizza chef is in the pizzeria by the counter
TEMPORAL	Partial	When does the pizzaiolo take the pizza out of the oven?	When he considers it cooked, towards the end of the video.	The pizzaiolo takes the pizza out of the oven towards the end of the video when he considers it cooked.	The pizzaiolo takes the pizza out of the oven towards the beginning of the video when he considers it cooked.
	Duration	How long does it take to cook the pizza in the video?	Pizza baking time is approximately 30 seconds	The baking of the pizza in the video takes approximately 30 seconds	The baking of the pizza in the video takes approximately 30 seconds



Figure 2: Overview of reasoning categories in MAIA with an example highlighting the cascading logic of the linguistic data. For each of the 100 videos, the dataset contains 2 questions for each of the 12 categories; for each question, it has 8 answers, and each of these answers has a corresponding True and False statement pair.

annotators. These pilots aimed to identify the optimal number, type, and specificity of the categories needed to effectively probe the cognitive and linguistic abilities of VLMs on our videos. Based on the feedback received, some initially proposed categories were merged due to content overlap or redundancy. Conversely, other categories were added to enhance the granularity of reasoning assessment (e.g. we introduced a *Planning* category, as we consider it a meaningful expression of reasoning skills). These refinements allowed us to design a more robust and informative framework to explore the interplay between language and vision in multimodal processing.

The following paragraphs introduce the final macro-

categories, including their definitions and any associated sub-categories.

CAUSAL focuses on reasoning about the causes or effects of events depicted in the video. It includes two subtypes², namely *Implicit* and *Explicit*, offering a comprehensive test of a model’s ability to describe causality within events. The former involves inferring unobservable causes from visible effects in the scene, requiring logical reasoning beyond what is directly shown. The

²Unlike the following cases, these are not treated as distinct sub-categories but as two equally represented subtypes of the same category

latter concerns clearly observable cause-and-effect dynamics, where either the cause or the effect is directly identifiable from the video content.

COUNTERFACTUAL focuses on questions about hypothetical scenarios that do not actually occur in the video but could take place under specific conditions. These questions are based on entities or events visible in the video and explore the consequences of an event or situation that might happen in the video if a certain condition were met. This category tests the ability of a model to reason about hypothetical scenarios grounded in the context of the video while deriving logical and plausible outcomes from such scenarios.

IMPLICIT investigates entities, events, or their attributes that are not explicitly visible in the video while their presence or properties can be reasonably inferred from the context. It evaluates the ability of a model to infer implicit details based on context, whether the target information was never shown or was previously visible but later obscured.

Total Implicit: involves entities or events that are never directly visible in the video but can be inferred from observable details. A typical answer provides the requested information based on logical inference.

Partial Implicit: involves entities or events that were visible earlier in the video but are no longer visible due to a shift in the scene or because they have moved out of the frame.

OUT-OF-SCOPE refers to entities or events entirely absent from the video, focusing on properties or details of these non-existent elements. Typical responses to out-of-scope questions involve a negation, indicating that the referenced entity or event is not present in the scene. Typical answers to this question types involve a negation, signaling that the referenced content is not present. This category indirectly tests the ability of a model to detect multimodal hallucinations and an assertiveness tendency in its responses.

PLANNING asks for actions needed to achieve a specific goal related to the video. The typical response to a planning question is a sequence of actions that someone should perform in order to reach the desired outcome. Such a category assesses the ability of the model to infer and plan the necessary steps to accomplish a goal based on the visual cues provided in the video.

SENTIMENT assesses sentiment, mood, attitude, or emotion displayed by characters in the video toward other

entities or events in the scene, throughout the entire video. A typical response to a sentiment question may describe a specific sentiment, attitude, or emotion, or it may reflect a neutral stance. This category evaluates the ability of the model to recognize and identify the emotional state or attitude of characters based on visual cues.

SPATIAL investigates the spatial relationships between entities, objects, or events depicted in the video. It aims at assessing the model's ability to infer both stable and time-dependent spatial relationships, as well as the ability to determine relative positioning in space and to rely on grounding competencies.

Total Spatial: focuses on position of entities in space (including their relation to other entities) that remains constant throughout the whole video, disregarding any temporal variations or minimal movements of the entity at different moments in the video. A typical response to this type of question provides general spatial information valid for the entire duration of the video.

Partial Spatial: focuses on time-related positions of entities in space, taking into account events occurring in the scene. A typical answer to this question provides spatial information that is valid only for the requested time range in the video.

TEMPORAL focuses on temporal information and studies the ability of a model to infer temporal relationships, sequence of events, and durations from visual content in a coherent manner.

Partial Temporal: focuses on the temporal properties and relationships between events in the video, excluding their duration. Questions target aspects such as when something happens or whether it occurs before or after another event. Typical answers specify the event along with the requested temporal detail.

Duration Temporal: focuses on a specific property of events in the video: their duration. A typical answer to a question involves several ways to express the duration of the event.

UNCERTAINTY refers to entities or events present in the video but lacking sufficient information to answer the question precisely. Questions are inherently ambiguous, as the visual content does not fully support a definitive response. Answers may offer plausible options, acknowledge uncertainty, or signal that the reply is a guess. This category tests a VLM in handling ambiguity and incomplete evidence, and in assessing its tendency to respond

assertively.

4. Curated Video Dataset

4.1. Video Selection

A key design choice for the MAIA benchmark was to reflect Italian culture in real-world scenarios through a carefully curated selection of video clips. To ensure richness and variety, the selection process was based on the following thematic areas: Locations, Food, Sport, Job, Nature, Activities. Such topics allowed us to collect a dataset showing locations, iconic Italian cities, and daily activities (e.g., enjoying breakfast at a café, cooking pasta, attending a soccer match) or even typical events (e.g., Italian local festivals or weddings). This cultural focus was not intended to limit the generalizability of the benchmark, but rather to offer a valuable opportunity to assess model performance on culturally grounded data, which is an aspect often underrepresented in existing multimodal resources.

4.2. Video Collection

We collected a culturally representative set of 100 short videos (~30 seconds each) sourced from *YouTube Italy*. Following the criteria described in Section 4.1, videos were retrieved using keyword-based queries across selected thematic areas. Only *Creative Commons* licensed content was included to ensure reproducibility. When necessary, longer videos were manually checked and cut to extract the most relevant 30-second segments, resulting in a uniform and culturally grounded video set.

4.3. Analysis of Videos

To better understand the visual content present in the MAIA benchmark, we conducted an object detection and classification analysis over the full set of videos using a *YOLOv11*³ detection pipeline. For each video, we sampled 32 uniformly spaced frames and ran object detection on them. This analysis provides a high-level view of the typical objects types in MAIA.

Figure 3 shows the frequency distribution of detected object labels across all annotated frames. *Person* is by far the most common object class, reflecting the human-centered nature of most videos in the benchmark. However, the dataset also includes a wide variety of everyday objects, suggesting a rich and diverse set of visual elements.

Figure 4 shows the distribution of the number of detected objects per frame. Most frames contain a moderate number of objects, typically between two and six. This

³<https://docs.ultralytics.com/it/models/yolo11/>

Object Detections Across All Videos

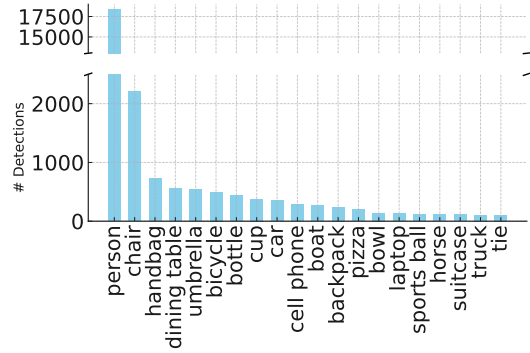


Figure 3: Distribution of object detections across all videos. For simplicity we plot just the top 20. *Person* is by far the most common entity.

Number of Objects Detected Per Frame

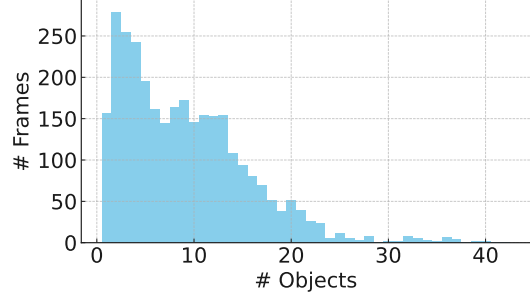


Figure 4: Histogram of number of objects detected per sampled frame. Most frames contain 3 objects.

indicates that the videos offer a balance between visual simplicity and complexity, making them suitable for testing both low-level perception and high-level reasoning in VLMs.

5. Curated Linguistic Data

5.1. Questions Collection

We created 12 different sets of guidelines, each assigned to a different annotator via *Google Forms* in order to collect two questions per reasoning category for each video. Annotators were PhD students under 30 with specializations in Linguistics and Computational Linguistics⁴. To ensure variability between the pair of questions about that video, annotators were asked to change the entities

⁴Each annotator was paid 100 euros for generating questions, which were collected through the administration of 1, 200 forms (10 per annotator)

QUESTION	What role do the men in white shirts play? Che ruolo svolgono gli uomini con le maglie bianche?
ANSWER 1	The men in white shirts are the competition judges Gli uomini con le maglie bianche sono i giudici di gara
ANSWER 2	They observe who scores a point Osservano chi fa punto
ANSWER 3	Men in white give judgements on the competition Gli uomini in bianco danno giudizi sulla gara
ANSWER 4	They seem to be the referees of this bocce game Sembra che siano gli arbitri di questa partita a bocce
ANSWER 5	They measure the distance of the thrown ball from the little one and determine the winner of the set Misurano la distanza della boccia tirata dal boccino e decretano il vincitore del set
ANSWER 6	The men in white shirts are the referees of the match Gli uomini con le maglie bianche sono gli arbitri della partita
ANSWER 7	The men in white are the jury Gli uomini in bianco sono i giudici
ANSWER 8	Men in white shirts play the role of refereeing the match Gli uomini con le maglie bianche svolgono il compito di arbitrare la partita
	

Figure 5: Example of a video and one of its two associated questions (category: *Implicit*), along with the corresponding 8-answer pool

and/or events involved in both of them. Each provided form contained both the definition of the assigned semantic category with examples, and also general rules to be followed (see Appendix, Figure 8 for an example of the form used). Each question had to be generated naturally and as an open-ended question. Questions involving a ‘Yes/No’ answer (e.g. *Is there a car in the video?*) were not allowed. Finally, for the correct execution of the task, the audio of the video had to be ignored, as the VLMs to be tested could only work on the visual part. Subsequently, questions were manually reviewed to ensure quality and category alignment.

5.2. Answers Collection

The goal of this phase was to collect 8 different answers for each question to ensure not only accuracy but also variability in responses. This choice is also supported by findings from Mañas et al. [33], who empirically show that using up to 8 demonstrations provides an effective trade-off between diversity, accuracy, and computational efficiency in in-context learning with LLMs for VQA evaluation. We used the *Prolific* platform⁵ and selected annotators aged 25 to 80 who were born in Italy, spoke

Italian as their first language, and had spent the majority of their first 18 years of life in Italy. As with the question collection step, we used *Google Forms* to provide the task⁶. Each form included 10 videos, and for each video, the annotators were asked to answer 12 questions, one per reasoning category (see Appendix, Figure 9 for an example of the form used). Annotators were encouraged to use their own world knowledge when interpreting the visual content of the video.

To guarantee high quality of the collected answers, we employed rigid control mechanisms based on sanity check questions. Answers were accepted only if the annotators correctly answered at least 90% of these control questions, otherwise their submissions were rejected and the task was reassigned to another annotator. In total, 2,400 questions were paired with 8 answers each, resulting in 19,200 responses. They were then further checked by a semi-automated two-step validation process based on *GPT-4o* with few-shot prompting:

Semantic Consistency Check. Each response was evaluated for semantic consistency with the corresponding question. In cases where inconsistencies were detected, the answers were manually reviewed to assess

⁵<https://www.prolific.com>

⁶Annotators were paid £7 per hour for answering questions

A

Given an Italian question Q and an answer A concerning a video, you must create a statement S based on A. While generating S, try not to alter the words composing A. If A includes first-person verbs or phrases (e.g., 'I think,' 'I believe'), rephrase S to be impersonal, avoiding a first-person perspective. The statement should be a concise, declarative sentence.

B

Given an Italian caption (TS) regarding the position or location of someone or something, your task is to create its foil (FS) by changing only the spatial information. Don't add other information respect to what is stated in TS. Here is an example to guide you:
TS: La donna nel video è in un campo di papaveri.
FS: La donna nel video è in una classe.

Figure 6: Prompts used for True (A) and False (B) Statements generation with GPT-4o. Prompt B (category: *Spatial total*) is representative of the 12 different prompts used to generate False Statements, each tailored to a specific semantic category.

whether the question should be re-answered by another annotator or the response could still be accepted. Real inconsistencies were found to be minimal (i.e., fewer than 100 out of 19,200 responses).

Contradiction Test. We checked whether, within each pool of 8 responses to the same question, any of the responses contradicted the others. We found that 90.25% of the 8-answer pools exhibit full agreement, as they do not contain any contradictions. The remaining 9.75% (234 cases) were manually reviewed by an additional annotator to resolve inconsistencies.

A post-processing phase of the responses was implemented to ensure a sufficient degree of variability and reduce potential redundancy within each of the 2,400 pools of 8 answers (see Section 5.6). Figure 5 shows an example of one 8-answer pool associated with a video and a question, after this refinement procedure described above.

5.3. True Statement Generation

At this step we automatically generate a true statement (TS) for each question-answer pair collected in the previous phases. A TS consists of descriptive declarative sentences aligned with the visual content of the videos. For example, if a video shows a boy who is initially in a kitchen and he hears a loud noise and runs away, a TS for the *Spatial* category could be:

In the video, the boy is in the kitchen before running away.

To create TS we used *GPT-4o*, with the prompt in Figure 6A, leveraging the combination of each question and its answer to automatically generate 19,200 true statements (TSs). As with the answers, the TS are organised into 2,400 pools of 8 items, each expressing the same event

with different wording. Following the same procedure used for the pools of 8 responses, we performed a quality check to ensure lexical variability within the 2400 pools of true statements (TS) (see Section 5.6).

5.4. False Statement Generation

The goal of this phase is to create a false statement (FS) for each TS already collected, in order to form a minimal TS-FS pair, enabling controlled experiments and precise analysis of a model's behavior with respect to the reasoning categories. As for TSs, the FSs were automatically generated using *GPT-4o* for editing only the elements of the sentence related to that semantic category, an approach inspired by the caption-foil method [14]. Figure 6B shows a prompt used for the FSs generation⁷. For instance, taking into account the previous example in 5.3:

In the video the boy is in the bathroom before running away.

Finally, we implemented two quality checks for FS using *GPT-4o*.

Structural Check aiming at automatically verifying that each FS aligns correctly with its corresponding TS according to its category. While the *GPT-4o* evaluation initially flagged 864 out of 19,200 cases as incorrect, only 2.5% were ultimately confirmed as truly problematic and subsequently corrected through manual revision.

Contradiction Test performed by assuming that a correct FS must be in contradiction with the relevant TS. We ran an NLI task to classify TS-FS pairs as Entailment,

⁷Due to space constraints, we could not include all the 12 prompts used for generating FSs specific to each reasoning category. However, the prompt shown here is representative of the adopted methodology.

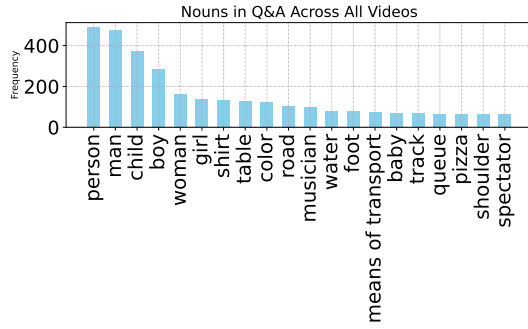


Figure 7: Distribution of the top 20 nouns in the Q&A pools across all videos, excluding high-frequency terms used to structure video-related questions according to the different categories (e.g., *What is the attitude of the girl in the video?*)

Neutral, or Contradiction. A qualitative analysis revealed that most Neutral cases (1287) were actual contradictions, and only 93 out of 170 Entailment cases were valid, which were then manually corrected to create contradictions.

5.5. Analysis of Linguistic Data

Similarly to videos, we investigated the entities used in our data by analyzing the most frequent nouns in the questions and their corresponding answers⁸, as shown by the frequency distribution in Figure 7. To extract entities, we used the *spaCy* library⁹ (*it_core_news_sm* pipeline), applying morpho-syntactic analysis (i.e., POS tagging) over both questions and answers. For each sentence, we selected tokens tagged as *NOUN* and extracted their lemmas to reduce redundancy. Duplicate nouns within the same QA pair were removed. Structural terms, functional to question/answers construction (e.g., *video*, *scene*, *attitude*), were filtered out to improve the informativeness of the plot and its comparability with object detection results in Figure 3. Several correspondences emerge between linguistic and visual entities (e.g., person, table, car), meaning that our linguistic data takes advantage of what is presented within our videos.

5.6. Lexical Variability

As said in Section 5.2, we opted for a pool-based structure with 8 items per question in order to balance semantic consistency with lexical diversity both across answers and statements. To meet this requirement, we assessed and enhance lexical richness within our data. This phase was carried out in several incremental steps (i.e., a string based test, lexical overlap and *Type-Token Ratio* (TTR))

⁸Nouns from TS and FS were excluded, as those sentences are derived from Q&A and would result in redundant repetitions.

⁹<https://spacy.io>

		B-Rephrasing		A-Rephrasing	
		TTok	CW	TTok	CW
Answers	Lexical Overlap	22.95%	21.41%	18.74%	17.60%
	Avg TTR	***	***	0.50	0.55
TSs	Lexical Overlap	39.34%	38.04%	30.51%	26.81%
	Avg TTR	0.37	0.41	0.50	0.55

Table 2

Average Lexical Overlap and TTR Statistics for pool, considering both Type-token (TTok) and Content-word only (CW). Statistics are compared before (B) and after (A) Automatic Sentence Rephrasing (*GPT-4o*) for both 19, 200 Answers and 19, 200 TSs. Average TTR statistics were not computed for the 2400 8-Answers pools before rephrasing.

based investigations), each of which involved an initial analysis of the potential redundancy of responses within the pool and an automatic rephrasing step (*GPT-4o*), particularly in cases where overlap was high¹⁰. Table 2 presents average lexical overlap and TTR within pools before and after rephrasing. Results show a substantial improvement, with a post-rephrasing average TTR of 0.55, which is a high value considering the semantic similarity among the 8 alternatives in each pool, especially for TSs, which follow more repetitive and fixed structures (e.g., *In the video X happens <...>*, *The video shows X <...>*).

6. Conclusion

We presented MAIA, a multimodal dataset forming the core of our benchmark designed for fine-grained investigation of the reasoning abilities of VLMs on videos. Among its innovative features, MAIA is the first Italian-native evaluation resource of its kind, built from both human-elicited data and content generated through controlled data augmentation. It supports two complementary tasks aligned on the same datapoints: a statement verification task (multiple-choice format), and an open-ended question answering task (fully generative setting).

As for future work, we would like to produce an English version of MAIA, for comparing VLMs on the same tasks across languages. Then, we intend to align the visual objects recognised by the VLM with the linguistic objects in the questions, enabling deeper error analysis based on the mapping. Finally, it would be interesting to see whether our framework promote models that undergo learning paradigms tightly integrating these two capabilities, as in Gul and Artzi [34].

¹⁰Since TSs are generated from an automatic rephrasing of Q&A pairs, we checked and improve their lexical diversity. This indirectly benefits the corresponding FSs, which differ by a single term from TS.

Acknowledgments

This work has been carried out while Davide Testa was enrolled in the Italian National Doctorate on Artificial Intelligence run by Sapienza University of Rome in collaboration with Fondazione Bruno Kessler (FBK). Giovanni Bonetta and Bernardo Magnini were supported by the PNRR MUR project PE0000013-FAIR (Spoke 2). Alessandro Lenci and Alessandro Bondielli were supported by the PNRR MUR project PE0000013-FAIR (Spoke 1). Alessio Miaschi was supported by the PNRR MUR project PE0000013-FAIR (Spoke 5). Lucia Passaro was supported by the EU EIC project EMERGE (Grant No. 101070918).

References


- [1] M. Baroni, Grounding distributional semantics in the visual world, *Language and Linguistics Compass* 10 (2015). doi:<https://doi.org/10.1111/lnc3.12170>.
- [2] J. P. Bigham, C. Jayant, H. Ji, G. Little, A. Miller, R. C. Miller, R. Miller, A. Tatarowicz, B. White, S. White, T. Yeh, Vizwiz: nearly real-time answers to visual questions, in: *Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology, UIST '10*, Association for Computing Machinery, New York, NY, USA, 2010, p. 333–342. URL: <https://doi.org/10.1145/1866029.1866080>. doi:10.1145/1866029.1866080.
- [3] E. Bugliarello, L. Sartran, A. Agrawal, L. A. Hendricks, A. Nematzadeh, Measuring progress in fine-grained vision-and-language understanding, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 1559–1582. URL: <https://aclanthology.org/2023.acl-long.87/>. doi:10.18653/v1/2023.acl-long.87.
- [4] D. Testa, G. Bonetta, R. Bernardi, A. Bondielli, A. Lenci, A. Miaschi, L. Passaro, B. Magnini, All-in-one: Understanding and generation in multimodal reasoning with the maia benchmark, 2025. URL: <https://arxiv.org/abs/2502.16989>. arXiv:2502.16989.
- [5] OpenAI, others., Gpt-4o system card, 2024. URL: <https://arxiv.org/abs/2410.21276>. arXiv:2410.21276.
- [6] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Batra, D. Parikh, Vqa: Visual question answering, 2016. URL: <https://arxiv.org/abs/1505.00468>. arXiv:1505.00468.
- [7] D. A. Hudson, C. D. Manning, Gqa: A new dataset for real-world visual reasoning and compositional question answering, 2019. URL: <https://arxiv.org/abs/1902.09506>. arXiv:1902.09506.
- [8] H. Le, C. Sankar, S. Moon, A. Beirami, A. Geramifard, S. Kottur, Dvd : A diagnostic dataset for multi-step reasoning in video grounded dialogueDVD: A diagnostic dataset for multi-step reasoning in video grounded dialogue, in: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Association for Computational Linguistics, Online, 2021, pp. 5651–5665. URL: <https://aclanthology.org/2021.acl-long.439>. doi:10.18653/v1/2021.acl-long.439.
- [9] M. Cafagna, K. van Deemter, A. Gatt, HL dataset: Visually-grounded description of scenes, actions and rationales, in: C. M. Keet, H.-Y. Lee, S. Zarrieß (Eds.), *Proceedings of the 16th International Natural Language Generation Conference*, Association for Computational Linguistics, Prague, Czechia, 2023, pp. 293–312. URL: <https://aclanthology.org/2023.inlg-main.21/>. doi:10.18653/v1/2023.inlg-main.21.
- [10] Y. Wu, Y. Zhao, S. Zhao, Y. Zhang, X. Yuan, G. Zhao, N. Jiang, Overcoming language priors in visual question answering via distinguishing superficially similar instances, in: N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, S.-H. Na (Eds.), *Proceedings of the 29th International Conference on Computational Linguistics, International Committee on Computational Linguistics*, Gyeongju, Republic of Korea, 2022, pp. 5721–5729. URL: <https://aclanthology.org/2022.coling-1.503/>.
- [11] Y. Li, B. Hu, F. Zhang, Y. Yu, J. Liu, Y. Chen, J. Xu, A multi-modal debiasing model with dynamical constraint for robust visual question answering, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 5032–5045. URL: <https://aclanthology.org/2023.findings-acl.311/>. doi:10.18653/v1/2023.findings-acl.311.
- [12] L. Parcalabescu, M. Cafagna, L. Muradjan, A. Frank, I. Calixto, A. Gatt, VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics

- tics, Dublin, Ireland, 2022, pp. 8253–8280. URL: <https://aclanthology.org/2022.acl-long.567>. doi:10.18653/v1/2022.acl-long.567.
- [13] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, R. Girshick, Clevr: A diagnostic dataset for compositional language and elementary visual reasoning, in: CVPR, 2017.
- [14] R. Shekhar, S. Pezzelle, Y. Klimovich, A. Herbelot, M. Nabi, E. Sangineto, R. Bernardi, FOIL it! find one mismatch between image and language caption, in: R. Barzilay, M.-Y. Kan (Eds.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 255–265. URL: <https://aclanthology.org/P17-1024/>. doi:10.18653/v1/P17-1024.
- [15] A. Suhr, M. Lewis, J. Yeh, Y. Artzi, A corpus of natural language for visual reasoning, in: R. Barzilay, M.-Y. Kan (Eds.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 217–223. URL: <https://aclanthology.org/P17-2034/>. doi:10.18653/v1/P17-2034.
- [16] T. Thrush, R. Jiang, M. Bartolo, A. Singh, A. Williams, D. Kiela, C. Ross, Winoground: Probing vision and language models for visio-linguistic compositionality, in: CVPR 2022, 2022.
- [17] X. Chen, R. Fernández, S. Pezzelle, The BLA benchmark: Investigating basic language abilities of pre-trained multimodal models, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 5817–5830. URL: <https://aclanthology.org/2023.emnlp-main.356/>. doi:10.18653/v1/2023.emnlp-main.356.
- [18] L. Bianchi, F. Carrara, N. Messina, C. Gennaro, F. Falchi, The devil is in the fine-grained details: Evaluating open-vocabulary object detectors for fine-grained understanding, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 22520–22529.
- [19] J. Xiao, A. Yao, Y. Li, T.-S. Chua, Can i trust your answer? visually grounded video question answering, in: CVPR, 2024, pp. 13204–13214. URL: <https://doi.org/10.1109/CVPR52733.2024.01254>.
- [20] S. Tong, Z. Liu, Y. Zhai, Y. Ma, Y. LeCun, S. Xie, Eyes wide shut? exploring the visual shortcomings of multimodal llms, in: CVPR 2024, 2024.
- [21] J. Lei, L. Yu, M. Bansal, T. Berg, TVQA: Localized, compositional video question answering, in: E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 1369–1379. URL: <https://aclanthology.org/D18-1167/>. doi:10.18653/v1/D18-1167.
- [22] A. Yang, A. Miech, J. Sivic, I. Laptev, C. Schmid, Just ask: Learning to answer questions from millions of narrated videos, 2021. URL: <https://arxiv.org/abs/2012.00451>. arXiv:2012.00451.
- [23] M. Grunde-McLaughlin, R. Krishna, M. Agrawala, Agqa: A benchmark for compositional spatio-temporal reasoning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 11287–11297.
- [24] K. Li, Y. Wang, Y. He, Y. Li, Y. Wang, Y. Liu, Z. Wang, J. Xu, G. Chen, P. Luo, L. Wang, Y. Qiao, Mvbench: A comprehensive multi-modal video understanding benchmark, CVPR (2024). URL: <https://doi.org/10.48550/arXiv.2311.17005>.
- [25] Y. Zhong, W. Ji, J. Xiao, Y. Li, W. Deng, T.-S. Chua, Video question answering: Datasets, algorithms and challenges, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 6439–6455. URL: <https://aclanthology.org/2022.emnlp-main.432/>. doi:10.18653/v1/2022.emnlp-main.432.
- [26] M. Grunde-McLaughlin, R. Krishna, M. Agrawala, Agqa: A benchmark for compositional spatio-temporal reasoning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.
- [27] Z. Yu, L. Zheng, Z. Zhao, F. Wu, J. Fan, K. Ren, J. Yu, ANetQA: A Large-scale Benchmark for Fine-grained Compositional Reasoning over Untrimmed Videos, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA, 2023, pp. 23191–23200. URL: <https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.02221>. doi:10.1109/CVPR52729.2023.02221.
- [28] I. Kesen, A. Pedrotti, M. Dogan, M. Cafagna, E. C. Acikgoz, L. Parcalabescu, I. Calixto, A. Frank, A. Gatt, A. Erdem, E. Erdem, Vilma: A zero-shot benchmark for linguistic and temporal grounding in video-language models, 2023. URL: <https://arxiv.org/abs/2311.07022>. arXiv:2311.07022.
- [29] V. Patraucean, L. Smaira, A. Gupta, A. R. Contente, L. Markeeva, D. S. Banarse, S. Koppula, J. Heyward, M. Malinowski, Y. Yang, C. Doersch, T. Matejovicova, Y. Sulsky, A. Miech, A. Fréchet, H. Klimczak, R. Koster, J. Zhang, S. Winkler, Y. Aytaç, S. Osindero, D. Damen, A. Zisserman, J. Car-

- reira, Perception test: A diagnostic benchmark for multimodal video models, in: Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2023. URL: <https://openreview.net/forum?id=HYEGXFnPoq>.
- [30] D. Croce, L. C. Passaro, A. Lenci, R. Basili, Gqa-it: Italian question answering on image scene graphs, in: Italian Conference on Computational Linguistics, 2021. URL: <https://api.semanticscholar.org/CorpusID:245125448>.
- [31] B. S. Shafique, A. Vayani, M. Maaz, H. A. Rasheed, D. Dissanayake, M. I. Kurpath, Y. Hmaiti, G. Inoue, J. Lahoud, M. S. Rashid, S. I. Quasem, M. Fatima, F. Vidal, M. Maslych, K. P. More, S. Baliah, H. Watawana, Y. Li, F. Farestam, L. Schaller, R. Tymtsiv, S. Weber, H. Cholakkal, I. Laptev, S. Satoh, M. Felsberg, M. Shah, S. Khan, F. S. Khan, A culturally-diverse multilingual multimodal video benchmark model, 2025. URL: <https://arxiv.org/abs/2506.07032>. arXiv:2506.07032.
- [32] R. Das, S. Hristov, H. Li, D. Dimitrov, I. Koychev, P. Nakov, EXAMS-V: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 7768–7791. URL: <https://aclanthology.org/2024.acl-long.420/>. doi:10.18653/v1/2024.acl-long.420.
- [33] O. Mañas, B. Krojer, A. Agrawal, Improving automatic vqa evaluation using large language models, 2024. URL: <https://arxiv.org/abs/2310.02567>. arXiv:2310.02567.
- [34] M. O. Gul, Y. Artzi, CoGen: Learning from feedback with coupled comprehension and generation, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 12966–12982. URL: <https://aclanthology.org/2024.emnlp-main.721/>. doi:10.18653/v1/2024.emnlp-main.721.

A. Additional Materials

The following figures show examples of the forms adopted for collecting the questions (Figure 8) and the corresponding answers (Figure 9).



Section 1 of 12


Question task


Il tuo compito è formulare due domande relative al contenuto di ogni video che ti verrà somministrato. Puoi guardare il video quante volte vorrai. Troverai istruzioni specifiche ed esempi che ti aiuteranno.

Importante:

- Formula la domanda in modo naturale.
- La domanda deve essere formulata come una domanda aperta. Non sono ammesse domande che prevedono una risposta "Sì/No" (ad esempio, "C'è una macchina nel video?").
- Non utilizzare l'audio del video, puoi solo guardarlo.

Privacy e scopi della ricerca

I tuoi dati personali e le tue domande saranno raccolti in forma completamente anonima e utilizzati solo per scopi legati a questo progetto. Questo studio è finalizzato allo sviluppo di un benchmark di valutazione del VLM nell'ambito della .

Per maggiori dettagli: .

Question *

☐ Accetto

TASK: FORMULARE DOMANDE CAUSALI

Description (optional)

Category specific task

Title

Il tuo task è quello di formulare domande "causali". Queste domande riguardano la causa o l'effetto di un evento. Le domande devono concentrarsi sull'evento, l'oggetto o l'essere umano (la causa) che ha portato al verificarsi di un altro evento, situazione o stato (cioè l'effetto) nel video o viceversa. Una risposta tipica può riportare una frase che descrive la causa o l'effetto richiesto nella domanda.

Questo è un esempio di domanda causale:

Supponiamo un video che mostra una persona ammalata che lancia un bicchiere a terra, e che il bicchiere si rompe.

Domanda: Perché il bicchiere si è rotto?


Risposta: Perché l'uomo l'ha gettato a terra.

Example

Nell'esempio, il focus della domanda (la rottura del bicchiere) è visibile nel video. Se cercassi di rispondere alla domanda, sarei in grado di utilizzare le informazioni visibili nel video per fornire la causa dell'evento (il lancio del bicchiere a terra da parte dell'uomo) come risposta alla domanda.

Questa è il tuo video: guardalo con attenzione.

Importante: Non considerare l'audio del video.



Basandoti sul video appena visto, formula qui le tue domande:

Description (optional)

Domanda 1 *


Short answer text

Domanda 2 (Se possibile, cambia le entità e/o gli eventi coinvolti nella domanda formulata.)

Short answer text

2-Questions generation

Figure 8: Example of a Google Form used for collecting 2 Questions for each video for each of the assigned category



Video Q&A Task

Benvenuti e grazie per partecipare a questo task di annotazione!

Ai fini del corretto svolgimento del task e della sua validazione, rispetta sempre le istruzioni che seguono.

Task

* Indicates required question

Il tuo compito consiste nel guardare una serie di video e rispondere, per ognuno di essi, a 12 domande basate esclusivamente sulla parte visiva del contenuto multimediale. Puoi visionare il video quante volte desideri per rispondere alle domande.

Punti Importanti da Ricordare:

1. Formula risposte sempre in maniera naturale.


Guidelines

Nota bene: Prendi il tuo tempo per rispondere al meglio alle domande. Le annotazioni consegnate al di sotto di un determinato arco temporale e/o che non rispettano tutte queste istruzioni iniziali, saranno automaticamente rifiutate. Lo stesso vale in caso di risposta errata alle domande di controllo.

Privacy e scopi della ricerca

I tuoi dati personali e le tue domande saranno anonime e utilizzati solo per scopi legati allo sviluppo di un benchmark di modelli (VLMs) nell'ambito della ricerca.

Per maggiori dettagli: [Privacy Policy and Research Purposes](#)

Per maggiori dettagli: 

☐ Accetto

Video Q&A Task

* Indicates required question

TASK

Se ne hai bisogno, [qui](#) puoi rivedere le istruzioni iniziali.

Questo è il tuo video: guardalo con attenzione.
Importante: Non considerare l'audio del video



Basandoti sul video appena visto, rispondi alle seguenti domande:

Quanti gatti ci sono? *

Your answer 

Questions

Figure 9: Example of a Google Form used for collecting answers for each video