All-in-one: Understanding and Generation in Multimodal Reasoning with the MAIA Benchmark

Davide Testa[1,2], Giovanni Bonetta[2], Raffaella Bernardi[3], Alessandro Bondielli[4,5], Alessandro Lenci[5], Alessio Miaschi[6], Lucia Passaro[4], Bernardo Magnini[2]

[1]Università di Roma La Sapienza, [2]Fondazione Bruno Kessler (FBK), [3]Free University of Bozen-Bolzano

[4]Dept. of Computer Science, University of Pisa, [5]CoLing Lab, Dept. of Philology, Literature and Linguistics, University of Pisa

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

{dtesta, gbonetta, magnini}@fbk.eu, raffaella.bernardi@unibz.it

{alessandro.bondielli, alessandro.lenci, lucia.passaro}@unipi.it, alessio.miaschi@ilc.cnr.it

Abstract

We introduce MAIA (Multimodal AI Assessment), a native-Italian benchmark designed for fine-grained investigation of the reasoning abilities of visual language models on videos. MAIA differs from other available video benchmarks for its design, its reasoning categories, the metric it uses, and the language and culture of the videos. MAIA evaluates Vision Language Models (VLMs) on two aligned tasks: a visual statement verification task, and an open-ended visual question-answering task, both on the same set of video-related questions. It considers twelve reasoning categories that aim to disentangle language and vision relations by highlighting the role of the visual input. Thanks to its carefully taught design, it evaluates VLMs' consistency and visually grounded natural language comprehension and generation simultaneously through an aggregated metric revealing low results that highlight models' fragility. Last but not least, the video collection has been carefully selected to reflect the Italian culture, and the language data are produced by native-speakers.[1]

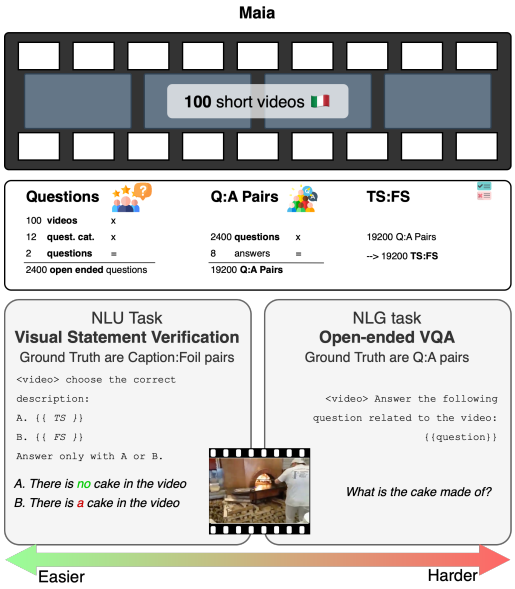


Figure 1: Structure of the MAIA benchmark. For each video there are two questions related to 12 reasoning categories. For each question there is a pool of 8 answers, each forming a question–answer pair associated with its own True–False statement pair. The example reported in the Figure is related to OUT-OF-SCOPE reasoning category.

1 Introduction

Vision and Language models entered the Computer Vision and NLP scenes more than a decade ago pushed by theoretical (e.g., Baroni, 2015) and application-oriented (e.g., Bigham et al., 2010) motivations. Their success on combining image and text has been monitored and summarized in various surveys, from the earlier ones on Visual Question Answering (Bernardi and Pezzelle, 2021) to the more recent ones focusing on Visually grounded LLMs (e.g., Caffagni et al., 2024; Li et al., 2024b). Researchers have always felt the need to target Vision and Language Models (VLMs) shortcomings, developing carefully designed benchmarks consisting of a suit of VL tasks to evaluate a variety of capabilities (Kushal et al., 2019). The minimal pair task, contrasting a caption with its foil (Shekhar et al., 2017) has been applied to large-scale linguistic phenomena (Parcalabescu et al., 2022) and extended to highlight weakness of VLMs on Video QA (Kesen et al., 2023). This trend focuses on visually grounded natural language understanding (Kesen et al., 2023).

Today VLMs are trained to generate text and are known to excel at it. We argue that the evaluation of Natural Language Generation (NLG) and Natural Language Understanding (NLU) competence should always be pursued together: An agent that can answer questions about an event must understand it too. Yet, existing benchmarks tend to

[1]Data available at GitHub.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 20030–20050

treat comprehension and generation separately, often relying on independent datasets and evaluation protocols. This fragmented design prevents an assessment of a model's **robustness**, in other words, its ability both to understand and generate visually grounded text. Moreover, success in NLU should be claimed only through evaluation regimes that monitor models' **consistency** across answers, showing they are insensitive to surface variations.

To address these limitations, we present MAIA: a competence-oriented benchmark consisting of two paired tasks – **multiple-choice Visual Statement Verification (VSV)** and **Open-Ended Visual Question Answering (OEVQA)** – containing aligned datapoints. For instance, in the example in Figure 1, for the video showing a pizza cooking in a wood-fired oven, the VSV contains the True and False statements (TS and TF) A: *There is no cake in the video* and B: *There is a cake in the video*, and the OEVQA contains the (aligned) question *What is the cake made of?*. Such interleaved data let us evaluate models' robustness: we evaluate a model positively, only if it performs correctly both on the visual statement verification task (NLU) – it selects *There is no cake in the video* – and on the Open-ended VQA (NLG) by generating something like *There is a pizza, not a cake*; while we evaluate it negatively, if it performs correctly only on one of the two tasks. Moreover, the VSV is organized in pools containing 8 TS-FS pairs that differ only on the surface, letting us evaluate the model's consistency. We categorize questions based on the reasoning they elicit; for instance, the question in Figure 1 is *Out-of-Scope*. MAIA spans twelve reasoning categories, helping highlight the role of language and visual modalities across them. Finally, it implements an **all-in-one** evaluation philosophy that lets us evaluate both models' robustness and consistency through its **Aggregate Metric**. Last but not least, MAIA is based on native Italian videos with language data obtained through human annotations and complemented by semi-automatic data augmentation. To the best of our knowledge, this is the first benchmark for the Italian language on videos.

Contributions. In the paper, we: (i) present MAIA, the first Italian benchmark specifically designed to assess the reasoning abilities of VLMs on video data; (ii) evaluate multiple VLMs, highlighting how their performance and their reliance on linguistic or visual cues varies across reasoning categories; (iii) demonstrate the importance of evaluating models from multiple perspectives, not only in terms of their competencies, but also their robustness and consistency; (iv) propose a novel metric for evaluating visually grounded comprehension and generation simultaneously.

2 Related Work

Diagnostic benchmarks for VLMs. Various types of benchmarks have been proposed since the raise of VLMs. From the single *task-oriented* benchmarks (e.g., Antol et al. (2015); Das et al. (2017); Croce et al. (2021)), attention has now moved to task collections (Xu et al., 2024; Lee et al., 2024b) in which models show impressive performance. As in the early phase (Johnson et al., 2017; Shekhar et al., 2017; Suhr et al., 2017), such success is mitigated by the use of *diagnostic* benchmarks, such as Parcalabescu et al. (2022); Thrush et al. (2022); Chen et al. (2023); Bugliarello et al. (2023); Bianchi et al. (2024) and *carefully curated* benchmarks such as Xiao et al. (2024); Tong et al. (2024). The third type of benchmarks available focus on the VLMs *competence* in a holistic fashion, evaluating advanced perception and reasoning with domain-specific knowledge (Yue et al., 2024b). A similar picture emerges for video-based VLMs. Here as well, early surveys call for careful evaluation (e.g., Zhong et al. (2022)), task-oriented benchmarks show impressive performance (Grunde-McLaughlin et al., 2021; Yu et al., 2023), while fine-grained ones pinpoint important weaknesses (Kesen et al., 2023), and competence-based analysis highlight there is significant room for improvement in multimodal video understanding (Patraucean et al., 2023). Finally, both Tong et al. (2024) for images and Kesen et al. (2023) for videos manage to highlight VLMs shortcomings by imposing a more stringent task-accuracy metric that account for model consistency across very similar data or correlated competencies. Thanks to the richness of MAIA data collection, we adopt such severe, and hence robust, evaluation code and propose a novel aggregate metric. Building on prior work, MAIA targets a low-resource language and the underexplored video domain. While some benchmarks (Das et al., 2024; Zhang et al., 2023) include limited Italian multiple-choice tasks, none focus on high-level reasoning in Italian video contexts or analyze distributional biases in VLMs.

Video Reasoning Benchmarks. Widely used benchmarks, such as AGQA (Grunde-McLaughlin et al., 2021) and MVBench (Li et al., 2024c) focus on explicit visual elements (e.g., entity, action, and the spatio-temporal reasoning involving them), instead MAIA's categories focus on the interplay between language and vision, especially when this relation is implicit or must be inferred, a dimension largely neglected in prior Video QA benchmarks. Yue et al. (2024a) includes multiple-choice and open-ended data points from entirely independent data sets with different origins and content, and reports the average performance across distinct tasks. Instead, MAIA's NLU and NLG data points are aligned, a unique feature of MAIA framework, as such it introduce an Aggregate Accuracy metric specifically designed to ensure that the performance of the model is evaluated consistently across multiple-choice and open-ended questions derived from the same underlying data. There are few other video-text benchmarks including both these formats (e.g., Zhou et al. (2025); Peng et al. (2024)), however, to the best of our knowledge, none of them attempt to define a unifying metric, as we do in MAIA.

3 The MAIA Benchmark

MAIA (Multimodal AI Assessment) is an evaluation framework designed to assess the reasoning capabilities of VLMs in video-based contexts.

3.1 Dataset

We outline here the steps involved in creating the MAIA dataset while a comprehensive description of its construction, characteristics, as well as the validation and revision procedures can be found in Testa et al. (2025). Validation steps consists in a qualitative analysis and revision of the data, when necessary.[2]

Video Collection. We gathered 100 short (ca. 30s) videos from *YouTube* Italy. The selection covers various aspects of Italian culture, including cities, art, food, sports, and daily activities (e.g., cooking pasta, having coffee, or watching a soccer match). Preference was given to videos featuring people and close-up shots. An automated script retrieved videos using thematic keywords and ensured *Creative Commons* compliance.

Reasoning Categories. We defined 12 reasoning categories aiming to probe the cognitive and linguistic skills of multimodal models and to explore the relation between language and vision, while forming the core of the benchmark for evaluating reasoning and grounding in an Italian context.

Questions and Answers Collection. The annotation process was carried out in two phases. In the first phase (*question creation*), 12 qualified annotators wrote 2 open-ended questions[3] per video for each category, ensuring diversity in entities and events.[4] A manual review verified adherence to guidelines and semantic categories. In the second phase (*answer collection*), we used *Prolific*[5] to solve the task, targeting Italian-native participants with specific cultural criteria (aged 25–80, born and raised in Italy, and Italian native speakers). Each annotator answered 12 out of 24 questions per video[6], focusing on detailed, visually grounded responses. Each question was answered by eight annotators to guarantee both accuracy and variability within the pool. [7] Two semi-automatic validation checks were applied to the collected answers: (1) semantic consistency with the corresponding question, and (2) contradiction tests across answers in the same pool.[8] After validation, the dataset consists of 2,400 questions, each paired with a pool of 8 high-quality answers, for a total of 19,200 validated responses. Through a post-processing of the lexicon, we made sure that the final 8-answer pools are lexical diverse.[9]

Statement Collection. As shown in Figure 2, TSs are descriptive declarative sentences that accurately align with the visual content of videos. TSs describe videos from different semantic perspectives, according to MAIA semantic categories. TSs were generated using *GPT-4o* (prompt in Figure 4A of the Appendix): for each question, it is given the 8 human generated answers and it is prompted

[2]Examples here are in English for readability.

[3]Yes/No and audio-based questions were prohibited.

[4]Annotators were paid €100 for their work.

[5]https://www.prolific.com/

[6]Annotators were paid £7 per hour.

[7]This choice is supported by Mañas et al. (2024), who show that up to 8 demonstrations provide a good balance between diversity, accuracy, and efficiency when talking about in-context learning with LLMs for VQA evaluation.

[8]We found that 90.25% of the 8-answer pools exhibit full agreement, as they do not contain any contradictions. The remaining 9.75% was manually reviewed by an additional annotator to resolve inconsistencies. This shows agreement among annotators, that cannot be measured through standard annotator agreement metrics (e.g., Inter-Annotator Agreement) due to the open-ended nature of the task.

[9]The lexical overlap within the 8-answers pool is 21.90%, similarly to the overlap between pairs randomly extracted from each pool (22.21%). In addition, the average *Type-Token Ratio* (TTR) for content words within each pool is 0.55.

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 MAIA reasoning categories. For each of the 100 videos, it contains 2 questions for each of the 12 categories; for each question, it has 8 answers, and each of these answers has a corresponding TS-FS pair.

to produce 8 TSs by combining the content of the question with the one of the corresponding answer. Again, post-processing techniques ensured high lexical variability within each pool reducing lexical overlap. FSs are incorrect descriptions automatically generated by prompting *GPT-4o* (Figure 4B) and created by minimally modifying elements of a TS related to a reasoning category while maintaining the original sentence structure, thus forming minimal pairs. FSs were validated through two semi-automatic checks (*GPT-4o*): (1) a structural verification to ensure that each FS was a minimal but incorrect variation consistent with its semantic category, and (2) an NLI-based contradiction test to confirm that each FS contradicted its corresponding TS. This process produced 19, 200 high-quality FSs aligned with their corresponding TSs.

3.2 Reasoning Categories

We report the reasoning categories in MAIA.[10] Figure 2 provides examples of aligned question, answer, TS, and FS.

Causal. Focuses on questions about the cause or effect of an event. It provides a comprehensive test of a model's ability to infer and describe causality within events. It can address either explicit (observable in the video) or implicit (inferred from the visible effect) causes/effects.

Counterfactual. Focuses on hypothetical events that do not occur in the video but could happen

[10]More details in Appendix A.1.

under certain conditions. It tests a model's ability to reason about plausible scenarios grounded in the video's context.

Implicit. Involves questions about entities or events that are either not explicitly visible in the video (*Total Implicit*) or no longer visible (*Partial Implicit*), but can still be reasonably inferred. It evaluates a model's ability to deduce implicit details based on context.

Out-of-Scope. Assumes the presence of entities or events not actually shown in the video, asking about properties of these nonexistent elements. It tests the model's ability to handle multimodal hallucinations and its tendency to make assertive, yet incorrect, responses.

Planning. Inquires about the sequence of actions needed to achieve a specific goal related to the video. It assesses the model's ability to infer and plan the necessary steps based on the visual cues.

Sentiment. Focuses on sentiment, mood, attitude, or emotions of characters towards other entities or events in the video. It evaluates the model's ability to recognize and identify the emotional cues.

Spatial. Focuses on the location of entities in space, either applicable to the entire video (*Total Spatial*) or specific moments and events (*Partial Spatial*). It assesses the model's ability to infer stable and time-dependent spatial relationships, determine relative positioning, and demonstrate grounding competencies.

Temporal. Relies on when something happens, either in relation to other events (*Partial Temporal*) or the duration of an event (*Duration*). It evaluates the model's ability to infer temporal relationships, event sequences, and durations from visual content.

Uncertainty. Arises when insufficient information is provided in the video to give a precise answer. It tests model's ability to handle situations with ambiguous or incomplete information, assessing its tendency to make assertive (rather than uncertain) responses.

4 Experimental Setting

We run several experiments to test modern VLMs on the MAIA benchmark in a zero-shot setting. To capture different VLM behaviors, strengths and limitations, we defined two tasks, aligned on the same datapoints: a multiple choice task, *Visual Statement Verification* (VSV), and a generative task, *Open-ended Visual Question Answering* (OEVQA).

4.1 Task1: Visual Statement Verification

VSV is a multiple-choice task where a model is presented with a true-false statement pair related to a MAIA question (see section 3.1) for a given video, and has to select the true option. The two statements are randomly assigned to two labels, A and B, and the model is asked to generate only the label. We chose the prompt through an extensive evaluation of 32 variants (16 in Italian and 16 in English), with the best-performing Italian prompt ultimately selected. Performance for VSV is measured with accuracy, i.e., the proportion of correctly selected true statements over the total statement pairs.

4.2 Task2: Open-ended VQA

OEVQA is a generative task, where models are tested on their ability to provide correct open-ended answers to a question related to video content. The model receives as input a prompt question and a video, and is tasked with generating a correct answer. The prompt used in the experiments was selected as the best-performing among 10 tested variants (5 in Italian and 5 in English). Generated responses are then evaluated according to the following approaches.

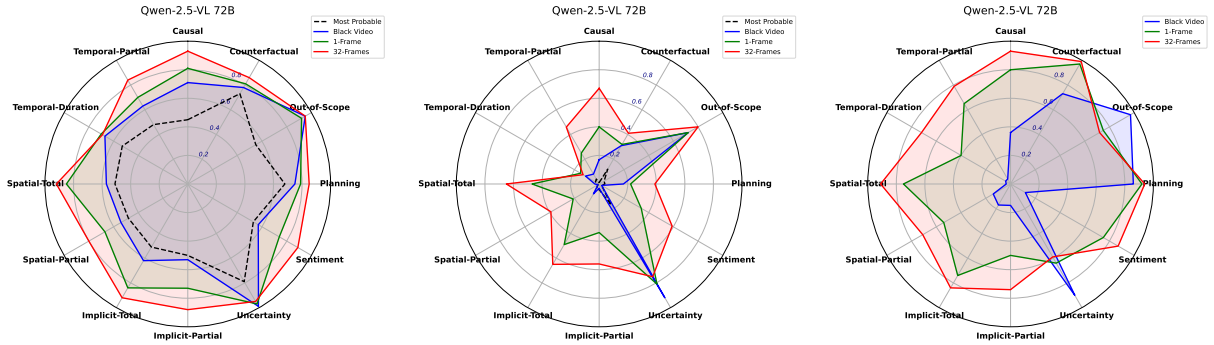
Similarity-based metrics. It compares a response against the pool of 8 reference answers available in MAIA. We used five token-level metrics: *ROUGE* (Lin, 2004), *BLEU* (Papineni et al., 2002), *BERT-Score* (Zhang et al., 2020), *METEOR* (Lavie and Agarwal, 2007) and *CIDEr* (Vedantam et al., 2015).

LLM-as-a-judge. While similarity metrics help rank VLMs, they fail in assessing answer correctness. To address this, we adopt an LLM-as-a-judge approach (Gu et al., 2025), using *GPT-4o* to evaluate whether an answer is semantically consistent with at least one of the eight MAIA references, prioritizing meaning over surface-level structure (Appendix B.2). Following Bavaresco et al. (2024), we validate this method on 100 samples: annotations by two human raters and *GPT-4o* show strong agreement, with a Fleiss' Kappa of 0.82.

4.3 Baselines

We implemented three baselines for our tasks.

Unimodal. This baseline applies only to Task 1 and selects the most probable statement in a TS-FS pair. It serves as a unimodal language baseline, reflecting the distributional biases of LLMs. Probabilities of TS and FS are first estimated on five



(a) Task 1: VSV (b) Task 1: VSV (pool-based) (c) Task 2: OEVQA

Figure 3: Fingerprint of *Qwen2.5-VL* 72B through MAIA's reasoning categories: (a) illustrates model performance in NLU, Task 1, when TS-FS pairs are independent, while (b) reports performance on the same task when the model correctly identify all TS-FS pairs within each 8-item pool, thereby penalizing inconsistency; (c) visualizes the performance on NLG, Task 2.

	Models	Avg.	Causal	Counterfactual	Out-of-Scope	Planning	Sentiment	Uncertainty	Implicit		Spatial		Temporal	
									Partial	*Total*	*Partial*	*Total*	*Duration*	*Partial*
Unimodal		0.05	0.01	0.12	0.04	0.04	0.02	0.17	0.04	0.04	0.01	0.01	0.03	0.04
Black video	*InternVL2* 8B	0.18	0.07	0.40	0.43	0.08	0.06	0.73	0.03	0.03	0.03	0.02	0.20	0.05
	InternVL3 78B	**0.30**	**0.35**	**0.44**	0.04	**0.28**	**0.41**	**0.92**	**0.10**	**0.17**	0.08	**0.12**	**0.29**	**0.40**
	Llava-Next-Video 7B	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.01	0.01
	Llava-oneVision 7B	0.14	0.01	0.04	0.59	0.04	0.00	0.83	0.04	0.04	0.01	0.01	0.06	0.01
	Qwen-2.5-VL 7B	0.22	0.11	0.36	0.41	0.17	0.16	0.87	0.04	0.06	**0.09**	0.07	0.15	0.16
	Qwen-2.5-VL 72B	0.22	0.17	0.31	**0.72**	0.17	0.02	**0.92**	0.03	0.08	0.01	0.03	0.11	0.08
1-Frame	*InternVL2* 8B	0.28	0.32	0.41	0.32	0.11	0.33	0.47	0.17	0.33	0.15	0.39	0.11	0.21
	InternVL3 78B	**0.42**	**0.53**	**0.50**	0.46	**0.37**	**0.47**	**0.69**	0.31	**0.53**	**0.23**	0.44	0.20	**0.36**
	Llava-Next-Video 7B	0.08	0.12	0.23	0.01	0.03	0.21	0.01	0.06	0.11	0.03	0.14	0.01	0.03
	Llava-oneVision 7B	0.32	0.35	0.29	**0.64**	0.14	0.32	0.65	0.21	0.35	0.12	0.36	0.15	0.20
	Qwen-2.5-VL 7B	0.36	0.36	0.28	**0.64**	0.16	0.44	**0.81**	0.25	0.39	0.14	0.39	**0.21**	0.24
	Qwen-2.5-VL 72B	0.40	0.40	0.32	0.72	0.22	0.34	**0.81**	**0.34**	0.49	0.21	**0.47**	0.15	0.25
32-Frames	*InternVL2* 8B	0.31	0.41	0.35	0.38	0.12	0.42	0.39	0.28	0.38	0.18	0.43	0.11	0.30
	InternVL3 78B	0.25	0.44	**0.42**	0.14	0.21	0.22	0.20	0.24	0.36	0.17	0.20	0.06	0.26
	Llava-Next-Video 7B	0.03	0.04	0.04	0.01	0.03	0.09	0.00	0.01	0.06	0.01	0.04	0.01	0.01
	Llava-oneVision 7B	0.38	0.51	0.21	0.61	0.19	0.51	0.45	0.39	0.47	0.26	0.48	0.11	0.33
	Qwen-2.5-VL 7B	0.44	0.53	0.29	0.63	0.23	0.50	0.67	0.43	0.56	0.28	0.55	**0.28**	0.32
	Qwen-2.5-VL 72B	**0.54**	**0.67**	0.41	**0.80**	**0.39**	**0.59**	**0.75**	**0.56**	**0.65**	**0.39**	**0.65**	0.13	**0.46**

Table 1: VSV (Task 1): accuracy of correct pools (8/8) across reasoning categories, penalizing models' inconsistency.

open-weight LLMs that have shown good performance on a variety of Italian tasks (Magnini et al., 2025): *Llama-3.1* (8B-Instruct), *LLaMAntino-2* (7B), *LLaMAntino-3-ANITA* (8B-Instruct), *Gemma* (7B) and *Qwen2.5* (7B-Instruct). For each TS-FS pair, we selected the item with the highest probability among the five models.

Black video. It replaces MAIA videos with a fully black clip, used as a proxy for a no-video condition. This setup minimizes access to visual features, pushing the model to rely mainly on the language component, while the true unimodal evaluation remains the previous one.

1-Frame. This baseline considers only the first frame for each MAIA video, this way reducing the capacity of a VLM to capture visual features and facing the one-frame "static appearance bias" (Lei et al., 2023).

4.4 Vision-Language-Models

We benchmarked six recent VLMs sourced from the Hugging Face Hub, representing state-of-the-art approaches in Vision-Language tasks:[11] *InternVL2* (8B, Chen et al. (2024)), *InternVL3* (78B, Zhu et al. (2025)) *LLaVA-NeXT-Video* (7B, Zhang et al. (2024b)), *LLaVa-OneVision* (7B, Li et al. (2024a)), and *Qwen2.5-VL* (both 7B and 72B, Qwen et al. (2025)). All models accept a [video, text] pair as input, and uniformly sample 32 frames from the video.[12]

[11] More details about VLMs are reported in Appendix B.1

[12] During the experiments with *InternVL3*, we found that, for about ten videos, 32 frames exceeded the model's context capacity, and in these cases we reduced them to 16 to ensure proper processing.

5 Results

This section reports the results obtained in our experiments for Task 1 and Task 2 independently.

5.1 Results on Visual Statement Verification

Table 1 shows VLM accuracy across the three settings (black video, 1-Frame, 32-Frames) and reasoning categories when models consistently answers correctly (i.e., by choosing 8/8 the TSs within the 8-item pool). *Qwen2.5-VL* 72B achieves the highest accuracy in the 32-Frames setup with an average score of 0.54, marking a 14-point improvement over the correspondent 1-Frame setting. *Llava-Next-Video* shows the weakest performance across all three configurations, likely due to its underlying Vicuna-7B LLM, weakly trained in Italian. All other models outperform the unimodal baseline (0.05): in the 32-Frames setting, gains go from +26 points (*InternVL2*) to +49 (*Qwen2.5-VL* 72B), confirming the use of visual cues to counteract language-driven biases. Notably, *InternVL3* 78B, despite being a high-performing model, exhibits an opposite trend in the 32-Frames setting, where its accuracy drops by 0.05 and 0.17 compared to the Black Video and 1-Frame configurations, instead of improving as observed for the other models.[13]
As a complement, Figure 3 helps visualizing the comparison across reasoning categories. Here we report results only for *Qwen-2.5-VL* 72B, our best model, while similar figures for the other models are in Appendix B (Figure 6, 7 and 8). The most difficult categories are COUNTERFACTUAL, PLANNING, IMPLICIT *Partial*, SPATIAL *Partial*, and TEMPORAL *Duration*. In addition, CAUSAL, SENTIMENT, IMPLICIT *Partial* and *Total*, and SPATIAL *Partial* and *Total* are the categories for which the model profits the most from visual clues, by leveraging the broader visual window provided in the 32-Frames setting, as shown by the larger area covered by the red curve compared to the green and blue ones.

Models' consistency. Figure 3a vs. 3b highlights the role of the consistency check in MAIA by using a pool of 8 TS-FS pairs. When pairs are considered independently, as it is usually done in VLM benchmarks, the model's performance increases significantly, showing that it relies on spurious correlation, effects that is strongly mitigated by the MAIA's severe evaluation regime. In our case, such trend is even much more visible in the black video setting than in the 32-Frames one with a gain of +47 vs. +34 for *Qwen2.5-VL* 72B (see Appendix, Table 5). By systematically comparing models' performance in the independent and pool-based settings, we see that this is a general finding across models. Moreover, we find that *Qwen-2.5-VL* 72B is not only the best-performing model, but also the most consistent, with a lower drop in the pool-based accuracy (Table 1) with reference to the indipendent one (Table 5), particularly in the 32-Frames setting (i.e., 34 points).

5.2 Results on Open-ended Generation

Table 2 reports the results on Task 2. The best performing model is again *Qwen2.5-VL* 72B, reaching 0.81 accuracy in the 32-Frames setting. Unlike in Task 1, here *InternVL3* 78B shows a positive incremental trend, with performance progressively improving from Black Video to 1-Frame and reaching its best results with 32 frames. In the latter configuration, across models, the hardest categories are UNCERTAINTY, OUT-OF-SCOPE, IMPLICIT *Partial*, TEMPORAL *Duration* and *Partial*, and SPATIAL *Partial*, while COUNTERFACTUAL and PLANNING appear less challenging in this context. This overall tendency is also confirmed for our best model. In particular, Figure 3c shows that the model benefits from 32 frames in all categories but OUT-OF-SCOPE and UNCERTAINTY, where it instead excels with black videos, though in most other cases performance in this setting is poor. Surprisingly, for COUNTERFACTUAL and PLANNING, a relatively high accuracy is obtained already with black videos. Interestingly, differences also emerge when comparing the 1-Frame vs. 32-Frames settings. In the majority of cases, the 1-Frame is not enough, while 32 frames increase performance. This difference is less pronounced or does not hold for PLANNING, COUNTERFACTUAL and UNCERTAINTY. For example, in the IMPLICIT *Total* category, the 32-Frames model answers correctly to *How does the vehicle move?*[14] with *The vehicle moves in a swinging way, with back-and-forth movements*, while the 1-Frame setting hallucinates with *The vehicle moves slowly along the amusement park route*.
Table 4 highlights how similarity-based metrics (e.g., ROUGE, BLEU) often do not align with semantic correctness. *InternVL2* scores highest on

[13] One possible explanation is that the model has not been trained to process 32-frame inputs at the resolution we provide.

[14] Referring to a pirate ship in the amusement park

	Models	Avg.	Causal	Counterfactual	Out-of-Scope	Planning	Sentiment	Uncertainty	Implicit		Spatial		Temporal	
									Partial	*Total*	*Partial*	*Total*	*Duration*	*Partial*
Black video	*InternVL2* 8B	0.37	0.42	0.60	0.30	0.68	0.43	0.08	0.21	0.23	0.36	0.25	**0.55**	0.25
	InternVL3 78B	**0.52**	**0.77**	**0.96**	0.00	0.83	**0.74**	0.09	**0.33**	**0.35**	**0.50**	**0.54**	0.49	**0.62**
	Llava-Next-Video 7B	0.27	0.43	0.47	0.30	0.40	0.33	0.51	0.21	0.12	0.16	0.04	0.12	0.16
	Llava-oneVision 7B	0.40	0.66	0.68	0.29	0.60	0.60	0.28	0.26	0.24	0.36	0.18	0.37	0.23
	Qwen-2.5-VL 7B	0.35	0.36	0.69	0.08	0.54	**0.74**	0.23	0.24	0.19	0.30	0.19	0.41	0.20
	Qwen-2.5-VL 72B	0.38	0.36	0.73	**0.97**	**0.86**	0.12	**0.90**	0.15	0.17	0.14	0.03	0.04	0.04
1-Frame	*InternVL2* 8B	0.44	0.57	0.65	0.21	0.65	0.60	0.10	0.33	0.38	0.39	0.47	0.53	0.35
	InternVL3 78B	0.68	**0.88**	**0.97**	0.44	0.91	**0.83**	0.56	**0.55**	0.70	**0.57**	**0.77**	0.44	0.56
	Llava-Next-Video 7B	0.32	0.30	0.56	0.20	0.46	0.60	0.29	0.18	0.32	0.24	0.27	0.20	0.19
	Llava-oneVision 7B	0.50	0.59	0.78	0.15	0.66	0.74	0.37	0.39	0.54	0.39	0.51	0.54	0.33
	Qwen-2.5-VL 7B	0.51	0.56	0.76	0.32	0.60	0.79	0.40	0.35	0.50	0.38	0.53	**0.55**	0.38
	Qwen-2.5-VL 72B	**0.70**	0.80	**0.97**	**0.75**	**0.92**	0.75	**0.64**	0.50	**0.74**	0.54	0.75	0.40	**0.65**
32-Frames	*InternVL2* 8B	0.49	0.54	0.68	0.28	0.64	0.62	0.11	0.45	0.48	0.47	0.51	0.57	0.46
	InternVL3 78B	0.77	0.86	0.96	0.54	0.77	0.85	**0.64**	0.73	0.81	0.50	0.54	0.49	0.62
	Llava-Next-Video 7B	0.33	0.37	0.38	0.16	0.42	0.48	0.27	0.24	0.37	0.29	0.39	0.32	0.29
	Llava-oneVision 7B	0.53	0.67	0.79	0.11	0.65	0.79	0.23	0.55	0.56	0.51	0.62	0.40	0.46
	Qwen-2.5-VL 7B	0.61	0.71	0.80	0.43	0.60	0.85	0.55	0.55	0.60	0.50	0.70	0.54	0.53
	Qwen-2.5-VL 72B	**0.81**	**0.93**	**0.99**	**0.72**	**0.94**	**0.87**	0.59	**0.74**	**0.84**	**0.71**	**0.91**	**0.72**	**0.79**

Table 2: OEVQA (Task 2): accuracy of correct answers with LLM-as-a-judge.

	Model	Avg.	Causal	Counterfactual	Out-of-Scope	Planning	Sentiment	Uncertainty	Implicit		Spatial		Temporal	
									Partial	*Total*	*Partial*	*Total*	*Duration*	*Partial*
Black video	*Qwen-2.5-VL* 72B	0.18	0.08	0.30	**0.69**	0.15	0.01	**0.84**	0.02	0.02	0.00	0.00	0.00	0.00
1-Frame	*Qwen-2.5-VL* 72B	0.33	0.39	0.32	0.59	0.20	0.32	0.53	0.27	0.43	0.15	0.42	0.09	0.24
32-Frames	*Qwen-2.5-VL* 72B	**0.47**	**0.64**	**0.41**	0.61	**0.37**	**0.56**	0.49	**0.48**	**0.58**	**0.35**	**0.62**	**0.10**	**0.40**

Table 3: Aggregate accuracy on Task 1 (NLU) and Task 2 (NLG) (consistency and robustness) on *Qwen-2.5-VL* 72B across reasoning categories.

Models	ROUGE	BertScore	BLEU	METEOR	CIDEr
InternVL2	0.61	**0.84**	0.38	0.59	**1.18**
InternVL3 78B	0.50	0.80	0.26	0.47	0.67
LLaVa-NeXT-Video	0.46	0.79	0.21	0.45	0.65
LLava-oneVision	0.58	0.83	**0.40**	0.55	1.08
Qwen-2.5-VL	0.58	0.83	0.38	0.61	0.98
Qwen-2.5-VL 72B	**0.62**	**0.84**	0.37	**0.65**	1.07

Table 4: VLM performance (32-Frames setting) for OEVQA (Task 2) according to similarity-based metrics.

surface-level similarity (e.g., BERTScore: 0.84, CIDEr: 1.18) – as well as *Qwen2.5-VL* 72B – but lower when considering the LLM-as-a-judge metric (0.49), while *Qwen2.5-VL* 72B and *InternVL3* 78B offer more (semantic) accurate answers. *LLaVA-NeXT-Video* underperforms across all evaluations.

5.3 Discussion

The results presented for Task 1 and 2 clearly highlight two key findings. First, within each task, the role of the information extracted from videos is unequally distributed across reasoning categories. The star-shaped Figure 3b illustrates such differences: the star's picks highlight the categories that profit from the visual input the most: SPATIAL *Total*, IMPLICIT *Total*, CAUSAL, OUT-OF-SCOPE, and UNCERTAINTY. On the other hand, and quite surprisingly, *Qwen2.5-VL* 72B handles OUT-OF-SCOPE and UNCERTAINTY better when provided the black videos, which are expected to be uninformative, than with the full 32 frames. This calls for a deeper analysis of the reason behind such a result, as the more visual context the model receives, the more it hallucinates, reducing OUT-OF-SCOPE scores, and becomes overly assertive, reducing UNCERTAINTY scores. Overall, this shows that MAIA categories are extremely useful to factorize the contribution of the visual vs. linguistic components of VLMs, favoring a more nuanced analysis of their actual abilities. Finally, among the most challenging categories – discussed in 5.1 and 5.2 – several (e.g., SPATIAL *Partial*, IMPLICIT *Partial*, and TEMPORAL *Duration*) share a temporal dimension, reinforcing the evidence that temporal reasoning still remains a fundamental issue of current models.

A second noteworthy fact is the effect of the task design. By comparing Table 1 and Table 2, we see that overall accuracy is higher in Task 2 than in Task 1, even though the underlying information the model has to exploit to perform the tasks is the same, given the alignment between their data points. Such an increase is found across models: +0.26 increase on average – in line with the Generative AI Paradox – they are better at generating than at understanding text (West et al., 2024). Figure 3b vs. Figure 3c illustrates such a difference for *Qwen2.5-VL* 72B. Interestingly, the difference

is more pronounced for some categories, as shown by the fact that Figure 3c no longer has the shape of a star; for instance, PLANNING, COUNTERFACTUAL, SPATIAL *Partial* and TEMPORAL *Duration* improve the most. On the other hand, UNCERTAINTY and OUT-OF-SCOPE accuracy drops in the NLG task; from a qualitative analysis, we saw that this is mostly due to the generation of hallucinations (Appendix B.3).

However, models' size proved to be a crucial factor: larger models consistently achieve higher accuracy, even within the same family, suggesting that scaling provides apparent advantages on individual tasks, although this picture changes when considering MAIA's broader scope (as discussed in the next section).

6 Aggregating Understanding and Generation

A distinctive feature of MAIA is the alignment between the VSV and OEVQA tasks: both are grounded in the same question and the same video. While Section 5 reported their results separately, we now combine them into a unified evaluation framework that jointly tests comprehension and generation abilities in VLMs. We introduce *Aggregate Accuracy* (*Agg-Acc*), a metric rewarding models that: (i) consistently select the correct statement (TS) over all 8 TS-FS pairs in Task 1, and (ii) generate a correct answer to the same question in Task 2, according to our LLM-as-a-judge evaluation. The idea is to capture overlapping abilities (i.e. knowledge and reasoning) for the two aligned tasks, evaluating models' robustness. *Aggregate Accuracy* is defined as follow:

$$Agg\text{-}Acc(M,q) = \begin{cases} 1 & \text{if } \forall (TS,FS) \in S_q, \\ & a_M(TS,FS) = TS \\ & \text{and } a_M(q) \text{ is correct} \\ 0 & \text{otherwise} \end{cases}$$

where M is the model, q a question, S_q the set of TS-FS pairs, and a_M the model's answers. Intuitively, given a question q on a video v, we reward the ability of a VLM to both select a correct answer TF from a $TF-FS$ pair related to q, and to generate a correct answer a to q, as discussed in Section 1 when commenting the question *What is the cake made of?* and the aligned data of the NLU and NLG tasks in Figure 1.

Table 3 reports *Agg-Acc* for our best model *Qwen2.5-VL* 72B: 0.47 with 32-frames, 0.33 with 1-frame, and only 0.18 with black video; this shows the more challenging nature of the aggregate task, and that a single frame is not that robust. PLANNING, SPATIAL *Partial*, and TEMPORAL *Duration* remain challenging even with 32-Frames, despite their notable improvements from Task 1 to Task 2. Results confirm that MAIA's aggregated understanding&generation task creates a harder benchmark, laying the ground toward a more objective VLM evaluation, even when considering larger models that apparently achieve high accuracy and appear to perform well.

7 Conclusion

We introduce MAIA, a benchmark designed for fine-grained investigation of the reasoning abilities of VLMs on videos. MAIA has two aligned tasks: a visual statement verification task (NLU), and a open-ended visual question answering task (NLG), both on the same set of video related questions. First, we provided a in-depth analysis of the two tasks independently, showing the importance of evaluating model with answer-pools to account for model consistency. Then, we couple comprehension and generation in an aggregated evaluation framework, arguing that the aggregated "all-in-one" understanding&generation task is a challenging and natural setting toward a more objective VLM evaluation, as it reveals inconsistencies within the same task and a lack of robustness across aligned tasks, even in larger models. As for the future, 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 (2024).

Limitations and Future Directions

We acknowledge that the number of videos in MAIA may seem relatively small, with only 100 samples. However, their combination with 12 reasoning categories results in 19,200 samples for Task 1 and an equal number of question-answer pairs, providing a robust evaluation set, not intended for any kind of training. Still, these 100 videos constitute only the initial core of a broader evaluation framework planned for future development. Moreover, we are aware that our most probable baseline, constructed using probabilities derived from our set of LLM's logits, poses a limitation that we intend to address in future work. We plan to compare each VLM with its correspond-

ing LLM to obtain more reliable results for proper comparisons and analyses of potential statistical biases. Regarding Task 2, we aim to dive deep into the comparison between the similarity metrics used and the emerging topic of LLMs as judges. Additionally, we intend to further investigate this latter approach to assess its actual validity as a reliable evaluation method. This direction will help us refine accuracy metrics, ultimately enhancing our ability to rigorously test model robustness and consistency across our two specular tasks. Furthermore, we acknowledge that our evaluation did not include large-scale proprietary models (e.g., ChatGPT). Our focus was primarily on testing the performance of open-source and easily exploitable language models to provide a comprehensive overview of their capabilities on the benchmark. Finally, we are also aware that our benchmark has not yet been compared with existing ones to assess its relative difficulty and challenge level. This limitation comes from the lack of comparable resources in Italian. As a future direction, we plan to translate MAIA into English and replicate our experiments, enabling more meaningful comparisons with similar English-language benchmarks.

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 was 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). Alessandro Bondielli was supported by the PNRR MUR project PE0000013-FAIR (Spoke 1), funded by the European Commission under the NextGeneration EU programme and by the the Italian Ministry of University and Research (MUR) in the framework of the PON 2014-2021 “Research and Innovation" resources – Innovation Action - DM MUR 1062/2021 - Title of the Research: “Modelli semantici multimodali per l’industria 4.0 e le digital humanities.”

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Marco Baroni. 2015. Grounding distributional semantics in the visual world. *Language and Linguistics Compass*, 10(1).

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, André F. T. Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2024. Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks.

Raffaella Bernardi and Sandro Pezzelle. 2021. Linguistic issues behind visual question answering. *Language and Linguistics Compass*, 15(6):e12417.

Lorenzo Bianchi, Fabio Carrara, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi. 2024. 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*, pages 22520–22529.

Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, and Tom Yeh. 2010. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology*, UIST ’10, page 333–342, New York, NY, USA. Association for Computing Machinery.

Emanuele Bugliarello, Laurent Sartran, Aishwarya Agrawal, Lisa Anne Hendricks, and Aida Nematzadeh. 2023. Measuring progress in fine-grained vision-and-language understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1559–1582, Toronto, Canada. Association for Computational Linguistics.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. 2024. The revolution of multimodal large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13590–13618, Bangkok, Thailand. Association for Computational Linguistics.

Xinyi Chen, Raquel Fernández, and Sandro Pezzelle. 2023. The BLA benchmark: Investigating basic

language abilities of pre-trained multimodal models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5817–5830, Singapore. Association for Computational Linguistics.

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.

Danilo Croce, Lucia C. Passaro, Alessandro Lenci, and Roberto Basili. 2021. Gqa-it: Italian question answering on image scene graphs. In *Proceedings of the Eighth Italian Conference on Computational Linguistics, CLiC-it 2021, Milan, Italy, January 26-28, 2022*, volume 3033 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Rocktim Das, Simeon Hristov, Haonan Li, Dimitar Dimitrov, Ivan Koychev, and Preslav Nakov. 2024. EXAMS-V: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7768–7791, Bangkok, Thailand. Association for Computational Linguistics.

Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. 2021. Agqa: A benchmark for compositional spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

Mustafa Omer Gul and Yoav Artzi. 2024. CoGen: Learning from feedback with coupled comprehension and generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12966–12982, Miami, Florida, USA. Association for Computational Linguistics.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*.

Ilker Kesen, Andrea Pedrotti, Mustafa Dogan, Michele Cafagna, Emre Can Acikgoz, Letitia Parcalabescu, Iacer Calixto, Anette Frank, Albert Gatt, Aykut Erdem, and Erkut Erdem. 2023. Vilma: A zero-shot benchmark for linguistic and temporal grounding in video-language models. *Preprint*, arXiv:2311.07022.

Kafle Kushal, Shrestha Robik, and Kanan Christopher. 2019. Challenges and prospects in vision and language research. *Frontiers in Artificial Intelligence*.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: an automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, page 228–231, USA. Association for Computational Linguistics.

Seongyun Lee, Seungone Kim, Sue Park, Geewook Kim, and Minjoon Seo. 2024a. Prometheus-vision: Vision-language model as a judge for fine-grained evaluation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11286–11315, Bangkok, Thailand. Association for Computational Linguistics.

Tony Lee, Haoqin Tu, Chi Heem Wong, Wenhao Zheng, Yiyang Zhou, Yifan Mai, Josselin Somerville Roberts, Michihiro Yasunaga, Huaxiu Yao, Cihang Xie, and Percy Liang. 2024b. VHELM: A holistic evaluation of vision language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Jie Lei, Tamara Berg, and Mohit Bansal. 2023. Revealing single frame bias for video-and-language learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 487–507, Toronto, Canada. Association for Computational Linguistics.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. Llava-onevision: Easy visual task transfer. *Preprint*, arXiv:2408.03326.

Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, and Jianfeng Gao. 2024b. Multimodal foundation models: From specialists to general-purpose assistants. *Found. Trends. Comput. Graph. Vis.*, 16(1–2):1–214.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. 2024c. Mvbench: A comprehensive multi-modal video understanding benchmark. *CVPR*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Bernardo Magnini, Roberto Zanoli, Michele Resta, Martin Cimmino, Paolo Albano, Marco Madeddu, and Viviana Patti. 2025. Evalita-llm: Benchmarking large language models on italian. *Preprint*, arXiv:2502.02289.

Oscar Mañas, Benno Krojer, and Aishwarya Agrawal. 2024. Improving automatic vqa evaluation using large language models. *Preprint*, arXiv:2310.02567.

Kanishka Misra. 2022. minicons: Enabling flexible behavioral and representational analyses of transformer language models. *arXiv preprint arXiv:2203.13112*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2022. VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8253–8280, Dublin, Ireland. Association for Computational Linguistics.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens Continente, Larisa Markeeva, Dylan Sunil Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alexandre Fréchette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and Joao Carreira. 2023. Perception test: A diagnostic benchmark for multimodal video models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Wujian Peng, Lingchen Meng, Yitong Chen, Yiweng Xie, Yang Liu, Tao Gui, Hang Xu, Xipeng Qiu, Zuxuan Wu, and Yu-Gang Jiang. 2024. Inst-it: Boosting multimodal instance understanding via explicit visual prompt instruction tuning. *Preprint*, arXiv:2412.03565.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017. FOIL it! find one mismatch between image and language caption. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 255–265, Vancouver, Canada. Association for Computational Linguistics.

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. 2017. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223, Vancouver, Canada. Association for Computational Linguistics.

Davide Testa, Giovanni Bonetta, Raffaella Bernardi, Alessandro Bondielli, Alessandro Lenci, Alessio Miaschi, Lucia Passaro, and Bernardo Magnini. 2025. MAIA: A benchmark for multimodal AI assessment. In *Proceedings of the 11th Italian Conference on Computational Linguistics (CLiC-it 2025)*, Cagliari, Italy. CEUR Workshop Proceedings.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *CVPR 2022*.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *CVPR 2024*.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. *Preprint*, arXiv:1411.5726.

Peter West, Ximing Lu, Nouha Dziri, Faeze Brahman, Linjie Li, Jena D. Hwang, Liwei Jiang, Jillian Fisher, Abhilasha Ravichander, Khyathi Chandu, Benjamin Newman, Pang Wei Koh, Allyson Ettinger, and Yejin Choi. 2024. The generative AI paradox: "what it can create, it may not understand". In *The Twelfth International Conference on Learning Representations*.

Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. 2024. Can i trust your answer? visually grounded video question answering. In *CVPR*, pages 13204–13214.

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. 2024. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–18.

Zhou Yu, Lixiang Zheng, Zhou Zhao, Fei Wu, Jianping Fan, Kui Ren, and Jun Yu. 2023. 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)*, pages 23191–23200, Los Alamitos, CA, USA. IEEE Computer Society.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024a. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *Preprint*, arXiv:2311.16502.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024b. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, pages 9556–9567.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2024a. Lmms-eval: Reality check on the evaluation of large multimodal models. *Preprint*, arXiv:2407.12772.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *Preprint*, arXiv:2306.05179.

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. 2024b. Llava-next: A strong zero-shot video understanding model.

Yaoyao Zhong, Wei Ji, Junbin Xiao, Yicong Li, Weihong Deng, and Tat-Seng Chua. 2022. Video question answering: Datasets, algorithms and challenges. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6439–6455, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. 2025. Mlvu: Benchmarking multi-task long video understanding. *Preprint*, arXiv:2406.04264.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *Preprint*, arXiv:2504.10479.

A MAIA:Benchmark Details

A.1 Semantic Categories Definition

We report here the definition of the twelve reasoning categories included in MAIA.

CAUSAL. This category[15] includes two subtypes: *Implicit Causal* and *Explicit Causal*, both aimed at reasoning about the causes or effects of events depicted in the video. Thus, it includes reasoning tasks involving both visible and inferred causal relationships, offering a comprehensive test of a model's ability to infer and describe causality within events.

- Implicit Causal: This type of question targets the inferred cause of an event, object, or human action visible in the video. The focus is on an implicit cause that cannot be directly observed but must be deduced from the effect presented in the scene. Typical responses involve a logical inference explaining the implicit cause behind the visible effect.
  Example: Suppose a video shows a person at home grabbing an umbrella while going out.

  Italian:
  Q: *Per quale motivo la persona prende l'ombrello?*
  A: *Perchè potrebbe piovere fuori.*

  English:
  Q: *Why does the person take the umbrella?*
  A: *Because it might be raining outside.*

  In this example, the action of grabbing the umbrella is visible, but the reason (bad weather) is not explicit in the video and must be inferred.
- Explicit Causal: This type of question addresses direct cause-and-effect relationships visible within the video. The focus is on identifying a specific event, object, or human action (the cause) that led to another event, situation, or state (the effect) or vice versa. Typical responses clearly describe either the

[15] Note that in MAIA there are four macro-categories with two fine-grained specifications (i.e., subcategories). The only exception is the *Causal* category, in which explicit and implicit items are equally represented (100 each). However, we do not consider them subcategories in the same way as the others, since in those cases, the subcategories express entirely different aspects of the same domain.

cause or the effect based on what is directly observable in the video.
Example: Suppose a video shows an angry person throwing a glass on the floor, which subsequently shatters.

Italian:
Q: *Perchè il bicchiere si è rotto?*
A: *Perchè la persona lo ha gettato a terra.*

English:
Q: *Why did the glass break?*
A: *Because the person threw it on the ground.*

Here, both the cause (throwing the glass) and the effect (the glass breaking) are visible in the video and can be used to provide a direct response.

COUNTERFACTUAL. This category focuses on questions about hypothetical scenarios that do not actually occur in the video but could take place under specific conditions. These questions explore the consequences of an event or situation that might happen in the video if a certain condition were met. A key requirement is that the hypothetical condition must be based on entities or events visible in the video. Consequently, this category tests a model's ability to reason about hypothetical scenarios grounded in the context of the video while deriving logical and plausible outcomes from such scenarios.
Example: Suppose a video shows an outdoor concert.

Italian:
Q: *Cosa succederebbe al concerto se arrivasse un forte temporale?*
A: *Il concerto verrebbe interrotto all'istante.*

English:
Q: *What would happen to the concert if a violent thunderstorm started?*
A: *The concert would be immediately interrupted.*

In this example, the focus of the question (the concert) is visible in the video, while the condition (a violent thunderstorm) is not. The consequence (the concert being interrupted) is not shown in the video but can be reasonably inferred .

IMPLICIT. The implicit category includes questions about entities, events, or their attributes that are not explicitly visible in the video. However, their presence or properties can be reasonably inferred from the context. This category evaluates a model's ability to infer implicit details based on context, whether the target information was never shown or was previously visible but later obscured.

- Total Implicit: These questions focus on 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.
  Example: Suppose a video shows the interior of a house, and suddenly the front door opens, revealing a person soaking wet with a dripping closed umbrella.

  Italian:
  Q: *Che tempo fa fuori?*
  A: *Piove molto forte.*

  English:
  Q: *What's the weather like outside?*
  A: *It's raining heavily.*

  In this case, the focus of the question (the weather outside) is not visible at any point in the video. However, details such as the wet person and dripping umbrella allow for a reasonably confident inference (heavy rain).
- Partial Implicit: These questions address 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.
  Example: Suppose a video shows a man placing a pen in a drawer and then closing it.

  Italian:
  Q: *Dove si trova la penna?*
  A: *La penna è nel cassetto.*

  English:
  Q: *Where is the pen?*
  A: *The pen is inside the drawer.*

  In this example, the focus of the question (the pen) is no longer visible in the video. However, earlier information (the man placing the

pen in the drawer) allows for a logical and confident answer (the pen is in the drawer).

OUT-OF-SCOPE. Such a category involves questions about entities or events that are not present in the video at all, asking for properties or details about these non-existent entities or events. A typical response to an out-of-scope question is a negation, stating that the entity or event in question is not present. This category tests the ability of a model to identify and handle irrelevant or non-existent entities within the video content, appropriately responding with a negation when the requested object or event is absent. Thus, it represents an indirect way to test the models on possible multimodal hallucinations and their tendency to be assertive in their responses.
Example: Suppose a video shows a dog and its owner playing in the park, but there are no cars in the scene.

Italian:
Q: *Di che colore è la macchina?*
A: *Non ci sono auto (nella scena).*

English:
Q: *What color is the car?*
A: *There is no car (in the scene).*

In this example, the focus of the question (the car) is not physically present in the video, nor can its presence be reasonably inferred. When trying to answer the question, no useful information about a car can be found, and the expected response would be a negation, such as "There is no car."
PLANNING. This category involves questions that request the 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, based on the situation presented in the video, in order to reach the desired outcome. Such a category assesses the model's ability to infer and plan the necessary steps to accomplish a goal based on the visual cues provided in the video.
Example: Suppose a video shows a dog and its owner playing with a ball in a park, and the owner throws the ball onto a bench.

Italian:
Q: *Cosa dovrebbe fare il cane per continuare a giocare col padrone?*
A: *Dovrebbe correre verso la palla, saltare sulla pacchina, prendere la palla e riportarla al padrone.*

English:
Q: *What should the dog do to continue playing with its owner?*
A: *The dog should run toward the ball, jump onto the bench, grab the ball, and bring it back to the owner.*

In this example, the focus of the question (the dog) is visible in the video. To answer the question, one can use the information in the video (the ball on the bench) to deduce the series of actions the dog should take (running, jumping, grabbing, and returning the ball) in order to continue the game.
SENTIMENT. The category involves questions that focus on the sentiment, mood, attitude, or emotion displayed by one or more characters in the video (i.e., animated beings) 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 represents a tool for evaluating model's ability to recognize and identify the emotional state or attitude of characters based on visual cues, reflecting their reaction or feelings toward the events and other entities in the video.
Example: Suppose a video shows children who appear bored at a birthday party.

Italian:
Q: *Che atteggiamento hanno i bimbi?*
A: *Sono annoiati.*

English:
Q: *What is the attitude of the children?*
A: *They are bored.*

In this example, the focus of the question (the children) is visible in the video. To answer the question, one can use the visual cues present in the video (expressions and behaviors of the children) to infer the sentiment (boredom) displayed by the characters.
SPATIAL. Such a category involves questions related to 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: This question asks about the

position of entities in space (including their relation to other entities) that remains constant throughout the entire 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 specific spatial information valid for the entire duration of the video.
Example: Suppose a video shows a school lesson with a teacher and students in a classroom.

Italian:
Q: *Dov'è l'insegnante?*
A: *L'insegnante è in classe.*

English:
Q: *Where is the teacher?*
A: *The teacher is in the classroom.*

In this example, the focus of the question (the teacher) is visible throughout the video. To answer the question, one can use visible information from the video (classroom, desk) to provide the entity's spatial position (behind the desk) throughout the video's duration.

- Partial Spatial: This question asks about the position of entities in space, but in relation to the time and/or other events occurring in the scene. It may also request the position of one entity relative to another, with a temporal aspect taken into account. A typical response provides spatial information that is specific only to the requested time range in the video.
Example: Suppose a video shows a school lesson with a teacher and students in a classroom.

  Italian:
  Q: *Dove si trova l'insegnante all'inizio del video?*
  A: *All'inizio del video, l'insegnante è di fronte la cattedra.*

  English:
  Q: *Where is the teacher at the beginning of the video?*
  A: *At the beginning of the video, the teacher is standing in front of the desk.*

  In this example, the focus of the question (the teacher) is visible in the video. To answer the question, one would use the visual information visible in the specific part of the video (classroom, desk) to provide the spatial position (in front of the desk) relative to the time frame requested (at the beginning of the video).

TEMPORAL. The category includes questions that focus on temporal information. This category studies the model's ability to infer temporal relationships, sequence of events, and durations from visual content in a coherent manner.

- Partial Temporal: This question focuses on the temporal properties and relationships of events in the video. The questions may request any type of temporal information about the events or their temporal relationships, except for their duration. For example, asking when something happens or if something happens before or after another event. A typical response provides the event with the specific temporal information requested by the question.
Example: Suppose a video shows a rock band concert.

  Italian:
  Q: *Che succede dopo che il chitarrista inizia a suonare?*
  A: *Il cantante inizia a cantare.*

  English:
  Q: *What happens after the guitarist starts playing?*
  A: *The singer starts singing.*

  In this example, the focus of the question (what happens after the guitarist starts playing) is visible at a specific moment in the video. To answer the question, one can use the visible information in that portion of the video (the singer starts singing) to provide the event (the singer starting to sing) as a response.
- Duration Temporal: This question focuses on a specific property of events in the video: their duration. A typical response provides the specific temporal information required by the question regarding the event's duration.
Example: Suppose a video shows a room with a light on and a person switching it off.

  Italian:
  Q: *Per quanto tempo la luce rimane accesa?*
  A: *Per circa 15 secondi.*

English:
Q: *How long was the light on?*
A: *For about 15 seconds.*

In this example, the focus of the question (the light on) is visible in the video. To answer the question, one can use the temporal information visible in the video (the person switching the light off) to provide a duration (about 15 seconds) as response.

UNCERTAINTY. This question refers to entities and events that are part of the situation represented in the video, but the scene does not provide enough information to give a precise answer. Therefore, uncertainty questions involve a certain degree of ambiguity in the response, which cannot be fully derived from the video content. The answer may refer to a range of values, state that a precise answer cannot be given, or mention that the answer is a guess and might not be correct. This category tests the model's ability to recognise and deal with situations in which the available information is insufficient or ambiguous, leading to a response that reflects the uncertainty of the scene and indirectly testing the hypothetical assertive behaviour of such models in answering.
Example: Suppose a video shows a dog.

Italian:
Q: *Quanti anni ha il cane?*
A: *Difficile da dire. / Il cane è probabilmente giovane, ma non si può esserne certi.*

English:
Q: *How old is the dog?*
A: *It's hard to say. / The dog is probably young, but it's not certain.*

In this example, the focus of the question (the dog) is visible in the video. However, if one tries to answer the question, only partial information about the dog's age is available in the video. As a result, an uncertain answer (e.g., "It's difficult to tell") is expected.

A.2 True and False Statement Generation

Figure 4 illustrates the prompts used to generate True Statements (A) from the questions and the corresponding responses in the 8-answer pools and the False Statements (B) starting from the true ones.

B Experiments

This appendix section will contain additional details on our experimental settings, including a description of the VLMs used, as well as graphs and tables summarizing the results for Tasks 1 and 2 of MAIA.
In contrast to the initial experiments for creating and validating the synthetic data of the MAIA dataset, where we used *OpenAI*'s *GPT-4o* API, the experiments on MAIA were conducted using A100 GPUs (40GB). Overall, the total computational budget was on the order of ∼1,000 GPU hours.

B.1 Models tested

Vision-Language Models We benchmarked six recent VLMs. Both experiments and related quantitative evaluation have been done using *lmms-eval* (Zhang et al., 2024a), a framework for the evaluation of multimodal models.

InternVL2. (Chen et al., 2024): 8B parameter transformer-based multimodal model employing advanced cross-attention; pre-trained on large-scale image-text and video datasets for diverse multimodal tasks and instruction-tuned. It uses *InternLM2.5* as open-sourced-7B parameter chat model. *Hugging Face* model: OpenGVLab/InternVL2-8B.
InternVL3. (Zhu et al., 2025): 78B parameter multimodal model trained with a native multimodal pre-training paradigm, integrating linguistic and visual capabilities from the start. It employs Variable Visual Position Encoding (V2PE) and advanced post-training strategies, achieving state-of-the-art performance among open-source VLMs. *Hugging Face* model: OpenGVLab/InternVL3-78B.
LLaVA-NeXT-Video. (Zhang et al., 2024b): 7B parameter model built on the LLaVA framework, optimized for video understanding with mechanisms to capture temporal dynamics; fine-tuned on video instruction data. Base LLM: *Vicuna*-7B (v1.5). *Hugging Face* model: llava-hf/LLaVA-NeXT-Video-7B-hf.
LLaVa-OneVision. (Li et al., 2024a): 7B parameter model that builds on the LLaVA framework with a Qwen2 LLM backbone to serve as a general-purpose vision-language assistant; pre-trained on extensive multimodal data to deliver robust cross-modal reasoning. *Hugging Face* model: lmms-lab/llava-onevision-qwen2-7b-ov.
Qwen2.5-VL. (Qwen et al., 2025): 7B and 72B parameter VLMs of the Qwen family using the

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

Given a question (Q), a candidate answer (A), and a set of 8 reference answers (R1–R8), your task is to determine whether A is correct. A is considered correct if it aligns with at least one of the reference answers.
Return only one label as output: 'Correct' or 'Incorrect'.

Figure 5: Prompt used for automatic evaluation of VLMs' answers in Task2 (i.e., for LLM-as-a-judge evaluation metric)

*Qwen*2.5 LLM decoder; key enhancements are related to grounding, working with longer videos and capturing events. It was pre-trained on comprehensive visual and textual datasets and fine-tuned for detailed, context-aware responses. *Hugging Face* model: Qwen/Qwen2.5-VL-7B-Instruct and Qwen/Qwen2.5-VL-72B-Instruct.

Unimodal models. As described in Section 4we used five open-weight LLMs which have shown good performance on a variety of tasks on Italian (Magnini et al., 2025). For conducting these experiments, we used *Minicons* library (Misra, 2022), a high-level wrapper around *Hugging Face* for investigating predictive behavior of transformer models. Specifically, probabilities were computed adding a normalization parameter to take into account the different length of sentences in terms of tokens. Models used in the *Hugging face* Hub are: **Llama-**3.1 (8B-Instruct), **LLaMAntino-**2 (7B), **LLaMAntino-**3**-ANITA** (8B-Instruct), **Gemma** (7B) and **Qwen**2.5 (7B-Instruct).

B.2 Tasks Details

Table 5 provide details with respect to the results obtained in Task 1 without considering any form of aggregation into pools (i.e. single-accuracy). Figures 6, 7, and 8 represent the fingerprint of models through MAIA's reasoning categories. Specifically, Figure 6 reports results for Task 1, Figure 7 for Task 1 when models make 8/8 correct choices within the 8 TS-FS pairs that make up the pools, and Figure 8 for Task 2.

As regards the generation task, we combined similarity-based metrics with an LLM-as-a-judge approach, the latter being more suitable for handling open-ended responses. Using 8 reference answers for evaluating the generation correctness of VLMs allowed us to prioritize semantic alignment over surface similarity, following (Lee et al., 2024a) and (Mañas et al., 2024). *GPT-4o* was adopted as evaluation model, and its judgments showed high agreement with human annotators (Fleiss' Kappa: 0.82). Figure 5 shows the prompt used for this evaluation.

B.3 Case study: Multimodal Hallucinations

As part of a preliminary error analysis focusing on the OUT-OF-SCOPE CATEGORY, we observed that a significant portion of the errors made by the model could be attributed to multimodal hallucinations. Particularly interesting was the discovery of counterintuitive clashes between the two tasks in our benchmark. In several instances, the model successfully solved Task 1 (i.e., True Statement Selection), in some cases achieving also full consistency within the pools (e.g., 8/8 correct selections), yet failed Task 2 (i.e., open-ended NLG), generating hallucinated content. For example, given a

	Models	Avg.	Causal	Counterfactual	Out-of-Scope	Planning	Sentiment	Uncertainty	Implicit		Spatial		Temporal	
									Partial	*Total*	*Partial*	*Total*	*Duration*	*Partial*
Unimodal		0.56	0.45	0.73	0.55	0.68	0.53	0.79	0.50	0.51	0.48	0.51	0.53	0.48
Black video	*InternVL2* 8B	0.68	0.64	**0.88**	0.89	0.69	0.61	0.96	0.52	0.59	0.54	0.55	0.67	0.60
	InternVL3 78B	**0.77**	0.81	**0.88**	0.60	**0.80**	**0.84**	**0.99**	**0.64**	**0.71**	**0.68**	**0.71**	**0.78**	**0.82**
	Llava-Next-Video 7B	0.50	0.49	0.52	0.48	0.51	0.52	0.45	0.50	0.51	0.51	0.49	0.51	0.48
	Llava-oneVision 7B	0.59	0.54	0.48	0.92	0.61	0.38	0.97	0.52	0.53	0.50	0.52	0.60	0.51
	Qwen-2.5-VL 7B	0.73	0.68	0.86	0.87	0.75	0.72	0.98	0.56	0.62	0.61	0.65	0.73	0.70
	Qwen-2.5-VL 72B	0.69	**0.71**	0.78	**0.95**	0.75	0.57	**0.99**	0.53	0.62	0.54	0.57	0.67	0.63
1-Frame	*InternVL2* 8B	0.75	0.80	0.87	0.69	0.71	0.77	0.89	0.65	0.80	0.65	0.82	0.67	0.69
	InternVL3 78B	**0.81**	**0.86**	**0.88**	0.81	**0.83**	**0.84**	0.93	0.74	0.87	**0.68**	0.84	**0.70**	**0.77**
	Llava-Next-Video 7B	0.61	0.68	0.76	0.49	0.63	0.75	0.40	0.59	0.65	0.59	0.70	0.56	0.53
	Llava-oneVision 7B	0.76	0.79	0.78	0.89	0.74	0.76	0.91	0.68	0.81	0.64	0.81	**0.70**	0.67
	Qwen-2.5-VL 7B	0.79	0.81	0.81	0.91	0.75	0.82	0.70	**0.82**	**0.96**	0.67	0.82	0.69	0.73
	Qwen-2.5-VL 72B	0.79	0.81	0.81	**0.92**	0.79	0.74	**0.97**	0.73	0.84	0.67	**0.85**	**0.70**	0.70
32-Frames	*InternVL2* 8B	0.79	0.83	0.85	0.77	0.75	0.84	0.84	0.75	0.83	0.69	0.86	0.66	0.76
	InternVL3 78B	0.64	0.81	**0.87**	0.55	0.71	0.59	0.62	0.61	0.76	0.67	0.71	**0.79**	0.82
	Llava-Next-Video 7B	0.52	0.56	0.57	0.42	0.57	0.61	0.32	0.52	0.59	0.52	0.56	0.51	0.48
	Llava-oneVision 7B	0.81	0.87	0.78	0.88	0.76	0.88	0.85	0.80	0.85	0.71	0.87	0.65	0.80
	Qwen-2.5-VL 7B	0.84	0.89	0.80	0.89	0.78	0.86	0.92	0.82	0.88	**0.83**	0.90	0.75	0.81
	Qwen-2.5-VL 72B	**0.88**	**0.93**	0.86	**0.95**	**0.85**	**0.89**	**0.95**	**0.88**	**0.92**	0.80	**0.92**	0.69	**0.84**

Table 5: Visual statement verification (Task 1): accuracy of correct choices across reasoning categories (without aggregation).

video in which a young girl is frightened by an insect in her home and seeks her mother's help to remove it, one of our best-performing model (i.e., *Qwen2.5VL*) consistently selected the correct true statement in all eight pairs[16]:

1. • In the scene there is no dog at the door
   • In the scene there is a dog at the door
2. • In the movie there are no animals at the door
   • In the movie there are animals at the door
3. • In the movie there is no doggie at the doorway
   • In the movie there is a doggie at the doorway
4. • No pets are seen in front of the door
   • Some Pets are seen in front of the door
5. • There are no dogs at the door in the video
   • There are dogs at the door in the video
6. • No dog appears in the entrance area
   • A dog appears in the entrance area
7. • No pets are visible in the video near the front door of the house
   • A pet is visible in the video near the front door of the house
8. • In the video clip there is no dog at the door
   • In the video clip there is a dog at the door

However, when prompted in Task 2 with the question *What color is the dog at the door?*, the model hallucinates by answering *The dog at the door is black*, despite the fact that no dog is present in the video. This case highlights a curious but also dangerous misalignment between the model's apparent ability to correctly perform discriminative reasoning in a multiple-choice setting and its failure to accurately generate grounded content. It also emphasizes a lack of robustness in the model's competencies. These findings underscore the need for a more in-depth error analysis, as the overall results (see Tables 2 and 3) suggest that, for specific reasoning categories, performance in Task 1 may not reliably predict success in Task 2, with notable performance unbalances often occurring (e.g., UNCERTAINTY reasoning category).

[16]True Statements are the first in each pair

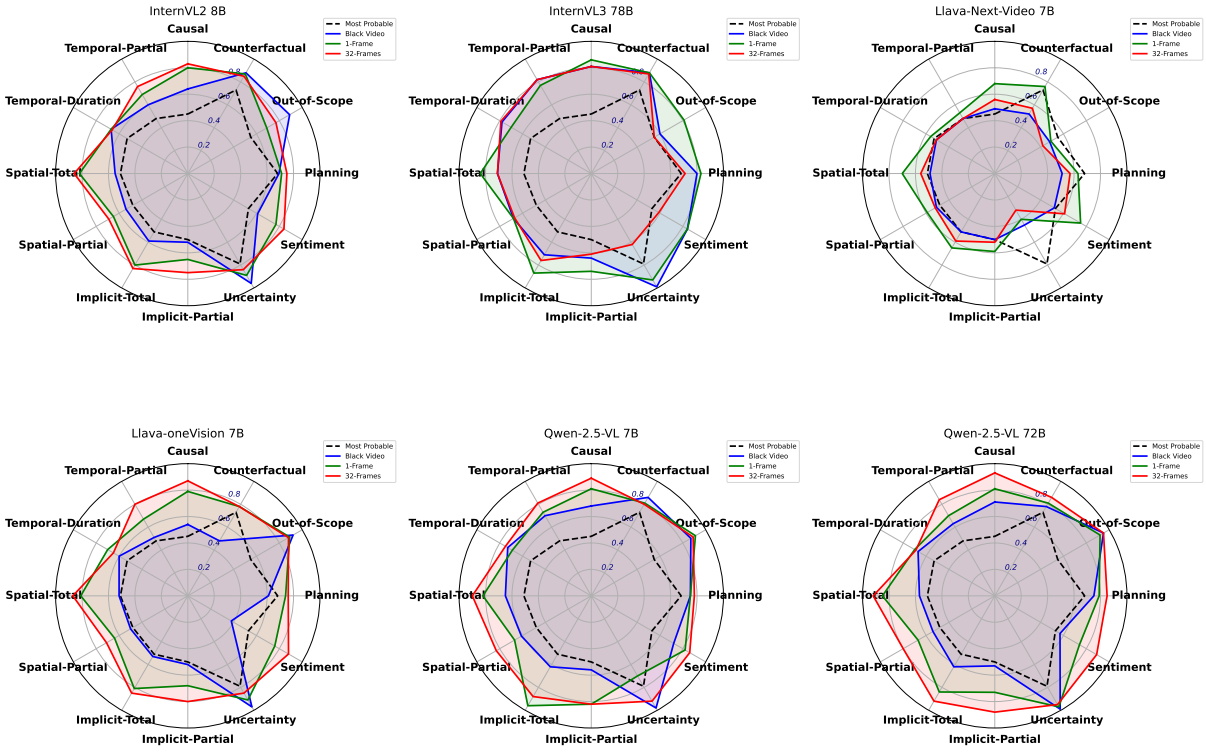


Figure 6: Task 1

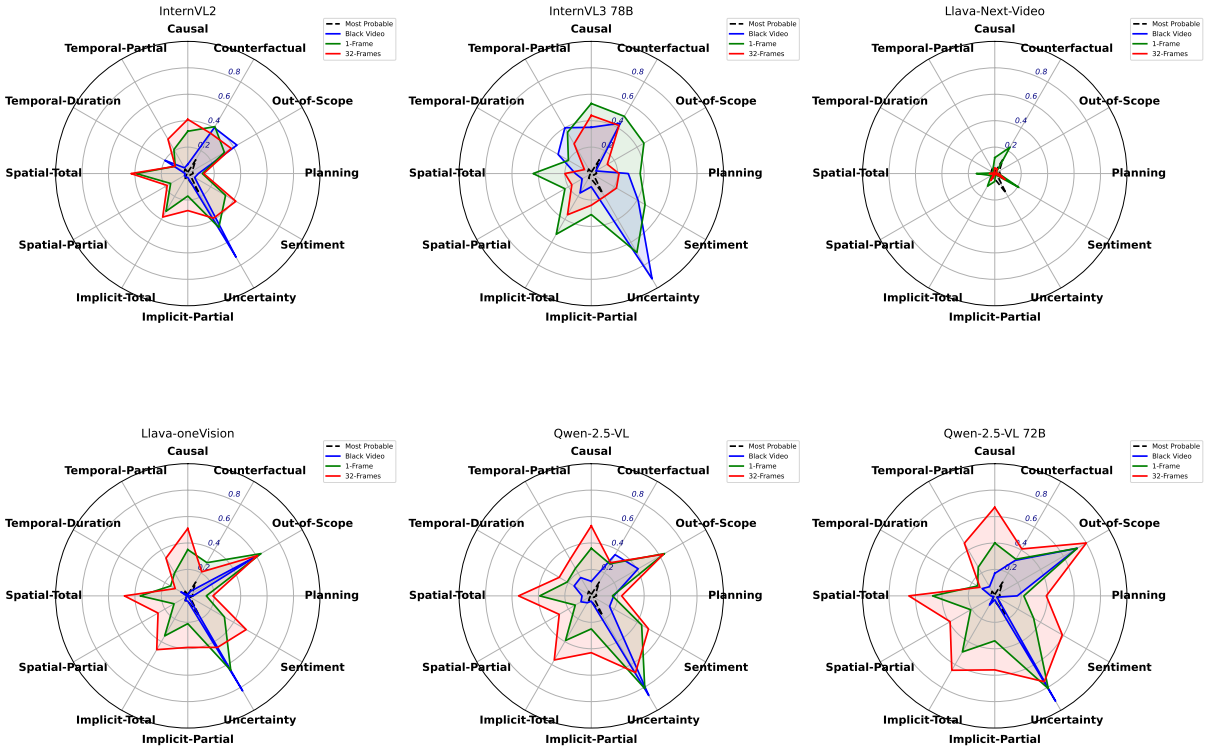


Figure 7: Task 1 pool-based

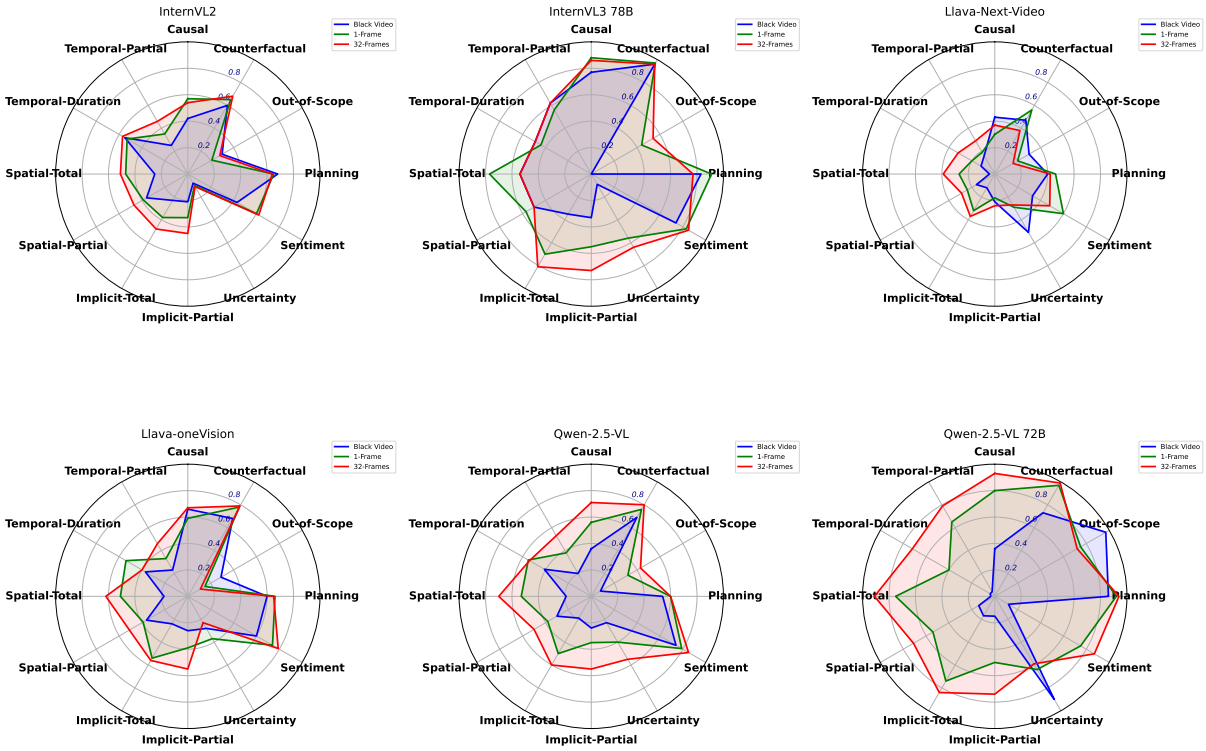


Figure 8: Task 2