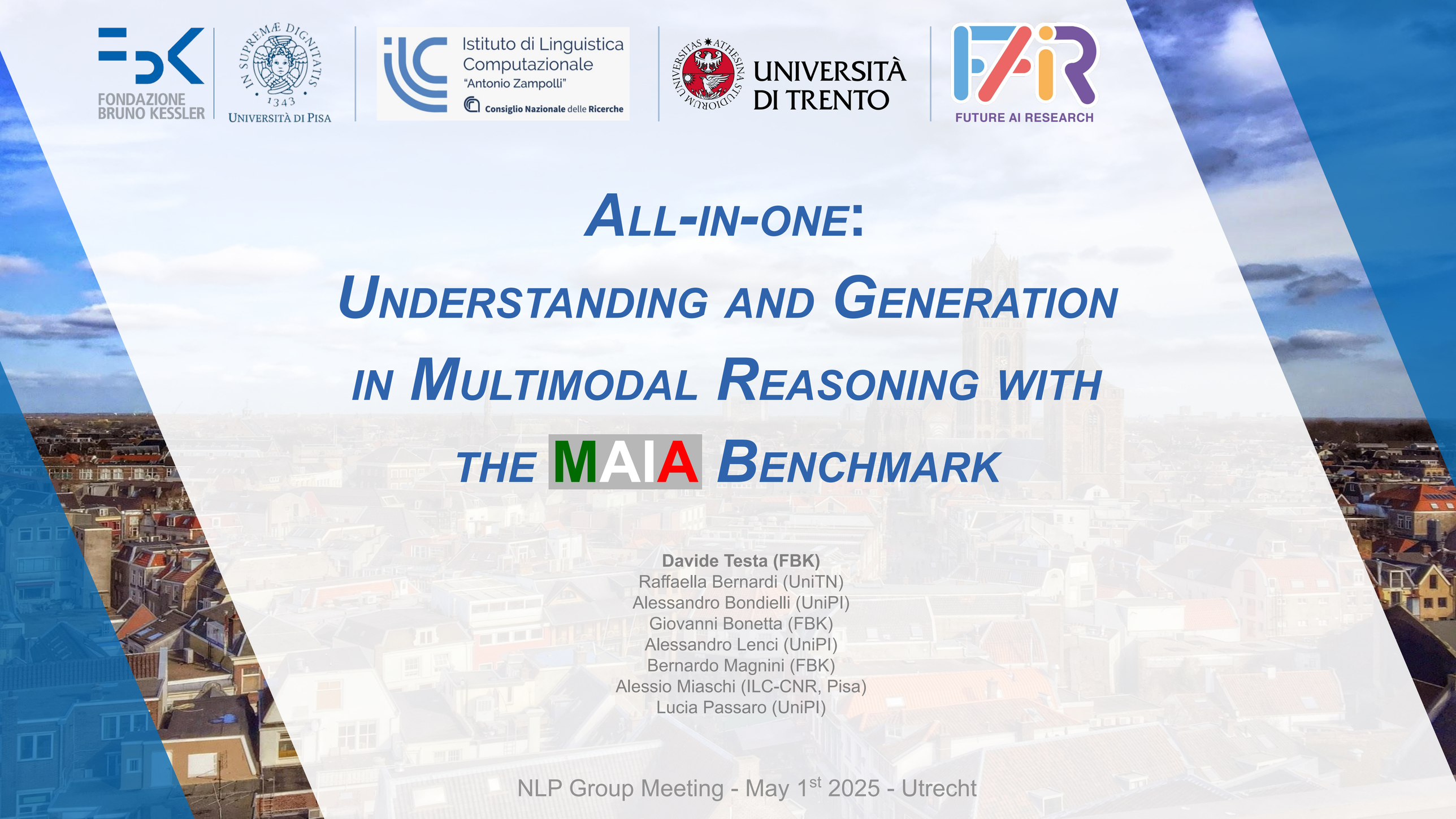


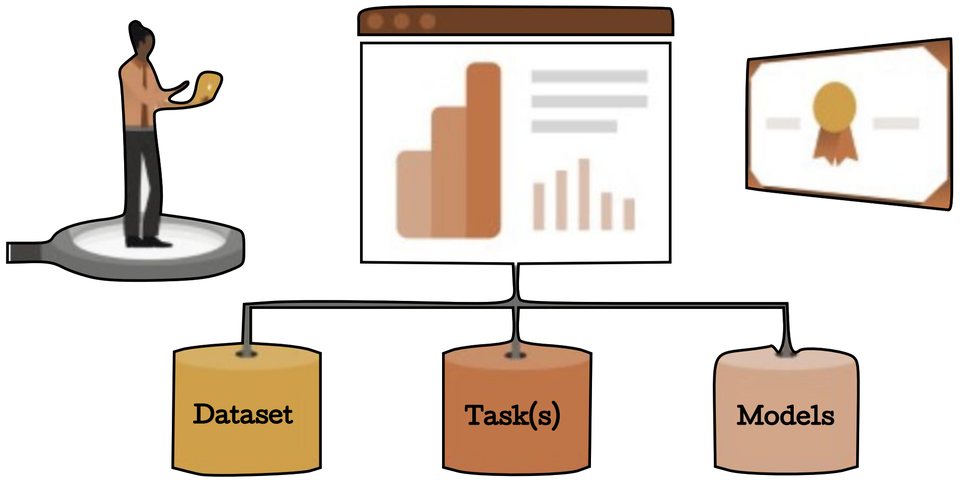
ALL-IN-ONE: UNDERSTANDING AND GENERATION IN MULTIMODAL REASONING WITH THE MAIA BENCHMARK

Davide Testa (FBK)
Raffaella Bernardi (UniTN)
Alessandro Bondielli (UniPI)
Giovanni Bonetta (FBK)
Alessandro Lenci (UniPI)
Bernardo Magnini (FBK)
Alessio Miaschi (ILC-CNR, Pisa)
Lucia Passaro (UniPI)

NLP Group Meeting - May 1st 2025 - Utrecht

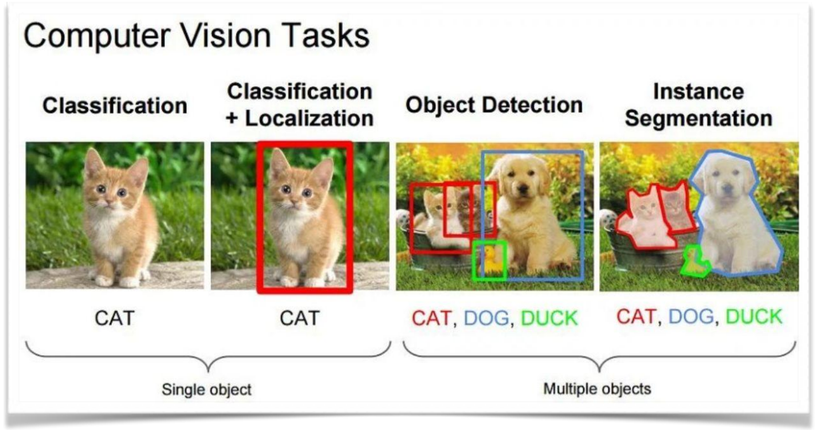
CAN AI “REASON” (AND SEE) √ THE WORLD?

1. **How** to evaluate *reasoning* in VLMs?
    - Test the understanding of a benchmark



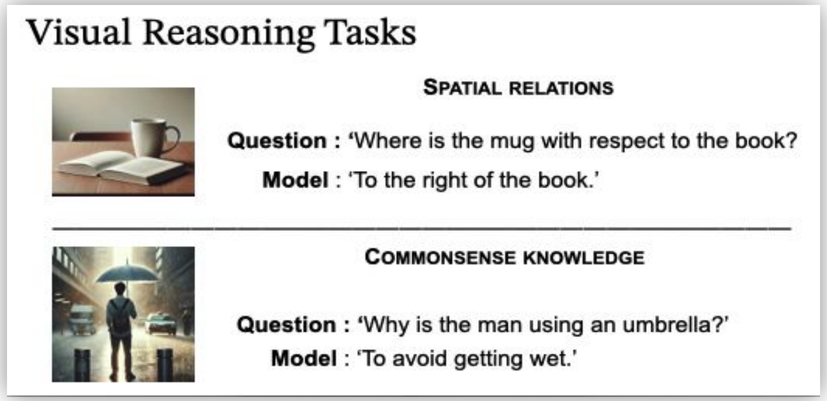
CAN AI “REASON” (AND SEE) √ THE WORLD?

1. **How** to evaluate *reasoning* in VLMs?
    - Test the understanding of a benchmark
2. **What** to evaluate?
    - Task



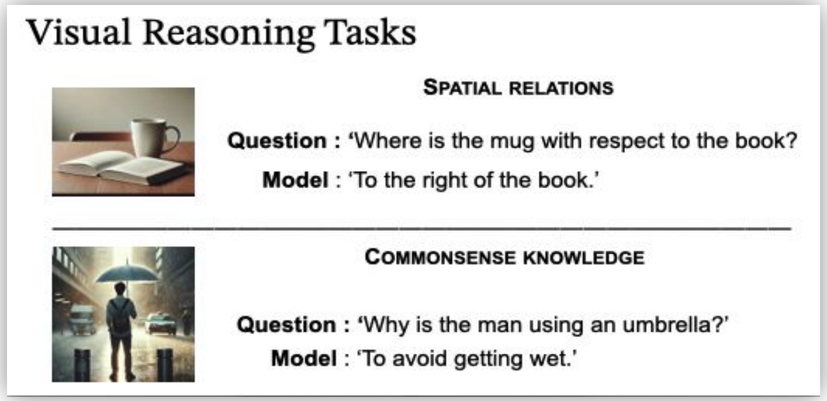
CAN AI “REASON” (AND SEE) √ THE WORLD?

1. **How** to evaluate *reasoning* in VLMs?
   - Test the understanding of a benchmark
2. **What** to evaluate?
   - Task
   - Competence



CAN AI “REASON” (AND SEE) √ THE WORLD?

1. **How** to evaluate *reasoning* in VLMs?
   - Test the understanding of a benchmark
2. **What** to evaluate?
   - Task
   - Competence
   - What model does ≠ What model knows [1,2,3]

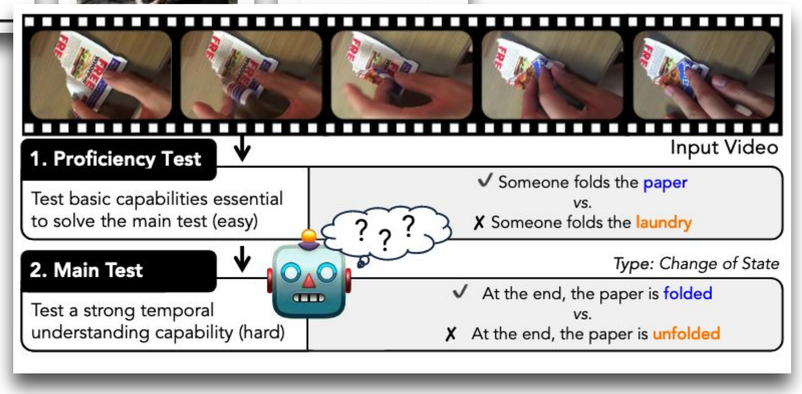


RELATED WORKS

- Several task-oriented datasets and/or benchmarks:
  - MSCOCO (*Lin et al., 2015*)
  - VQAv2 (*Goyal et al., 2017*)
  - LAION (*Schuhmann et al., 2021*)
  - LVLM-eHub (*Xu et al., 2023*)
  - MM-Bench (*Liu et al., 2023*)
  - Video-Bench (*Ning et al., 2023*)
  - …

- **Competence-oriented benchmarks are still rare!**
  - VALSE (*Parcalabescu et al., 2022*)
  - ViLMA (*Kelsen et al., 2023*)
  - BLA (*Chen et al., 2023*)
  - MMMU (*Yue et al., 2024*)
  - MMT-Bench (*Ying et al., 2024*)

VALSE

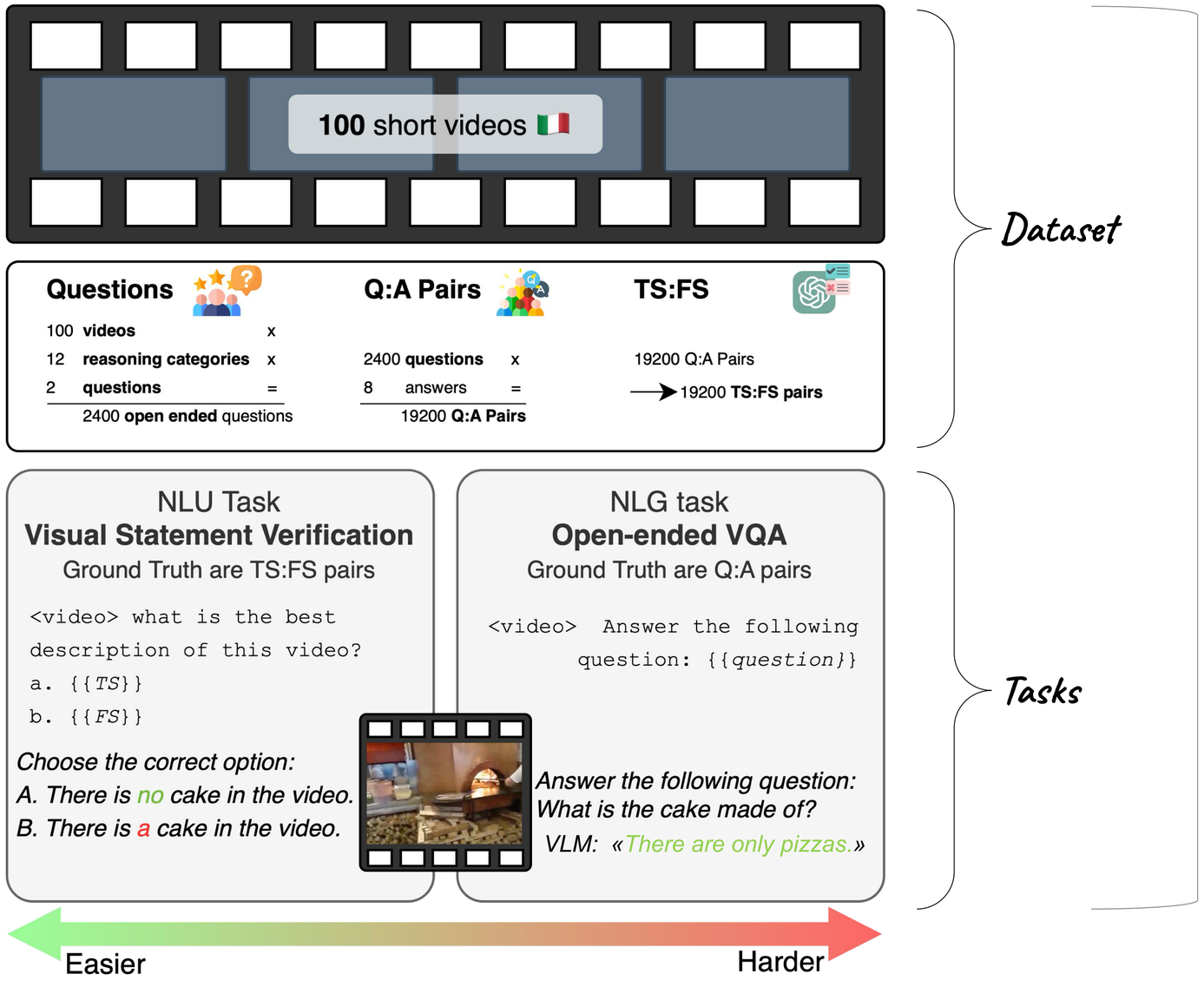


ViLMA (Temporal reasoning)

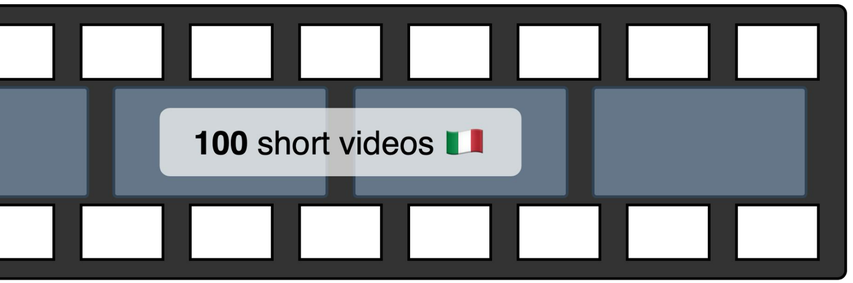
- No existing benchmarks covering a comprehensive set of semantic categories for visual linguistic reasoning
  - Especially for low-resources languages!

MAIA

(Multimodal AI Assessment)

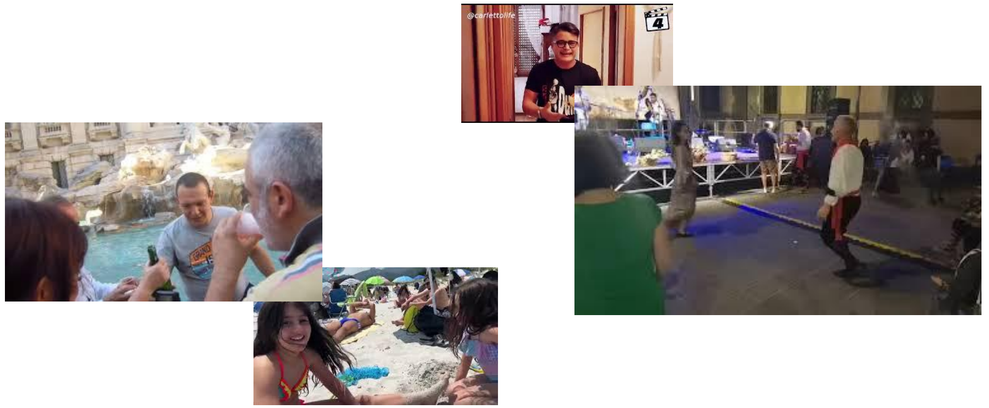


- **Competence-oriented** benchmark
- Italian Video-based Tool
- Objective:
  - Challenge VLMs on **Reasoning**
  - Assess VLMs **Robustness**
- Novelties:
  - Fine-grained **Semantic Categories**
  - **Open-ended VQA**
  - Two **aligned tasks**
  - Several **experimental settings**



MAIA Dataset

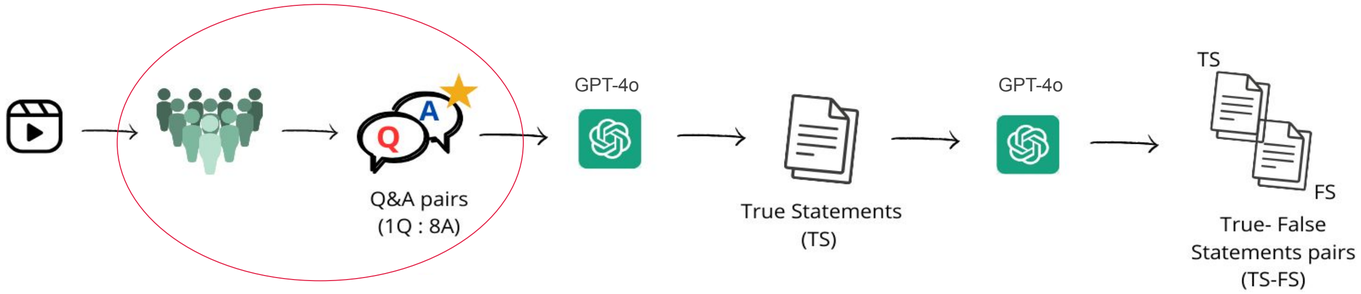
- 100 Italian videos
- Italian culture related contents
- Origin: Youtube-it [*Creative Common License*]
- Duration: ~ 30 secs
- Topics covered:
  - Locations (indoor / outdoor)
  - Food
  - Sport
  - People at work
  - Italian landscapes
  - Art
  - Italian typical events / situations



12 (LINGUISTIC) REASONING CATEGORIES

1. **COUNTERFACTUAL**: What would happen if the pizza fell?
2. **OUT-OF-SCOPE**: Where is the cat?
3. **UNCERTAINTY**: How many pizzas does the pizza maker make each day?
4. **IMPLICIT REASONING**: Which kind of pizza is the pizza maker making? x2
5. **SENTIMENT**: What is the pizza maker's attitude as he takes the pizza out of the oven?
6. **PLANNING**: What should the pizzaiolo do to revitalize the fire?
7. **SPATIAL**: Where is the pizza placed after it comes out of the oven? x2
8. **TEMPORAL**: How long does the pizza take to cook? x2
9. **CAUSAL**: Why did the mozzarella cheese on the pizza melt at the end of the video?

MAIA Dataset



Questions

100 **videos** x
12 **reasoning categories** x
2 **questions** =
2400 **open ended** questions

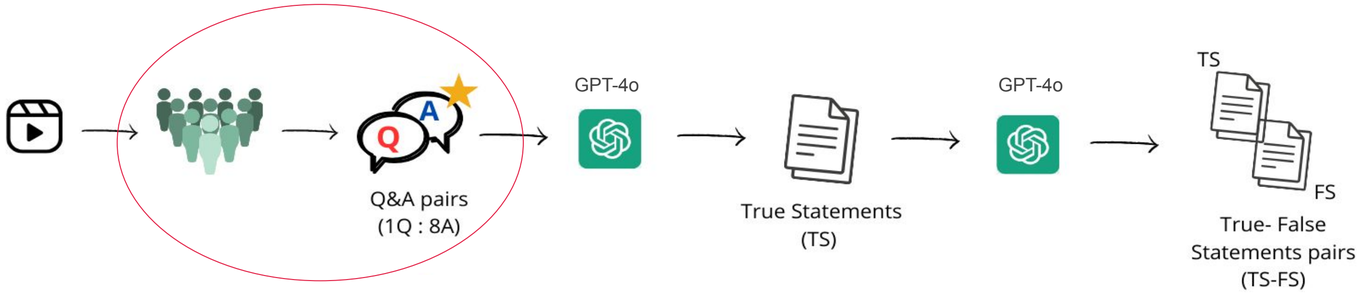
Q:A Pairs

2400 **questions** x
8 answers =
19200 **Q:A Pairs**

TS:FS

19200 Q:A Pairs
→ 19200 **TS:FS pairs**

MAIA Dataset



Questions

100 **videos** x
12 **reasoning categories** x
2 **questions** =
2400 **open ended** questions

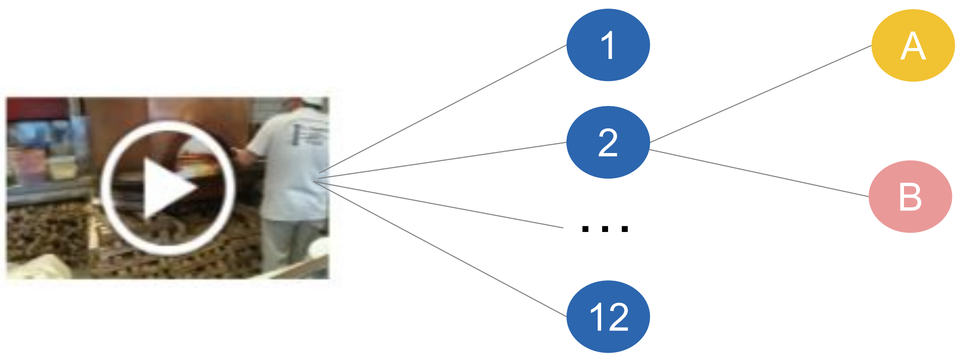
Q:A Pairs

2400 **questions** x
8 answers =
19200 **Q:A Pairs**

TS:FS

19200 Q:A Pairs
→ 19200 **TS:FS pairs**

CATEGORY **QUESTION**

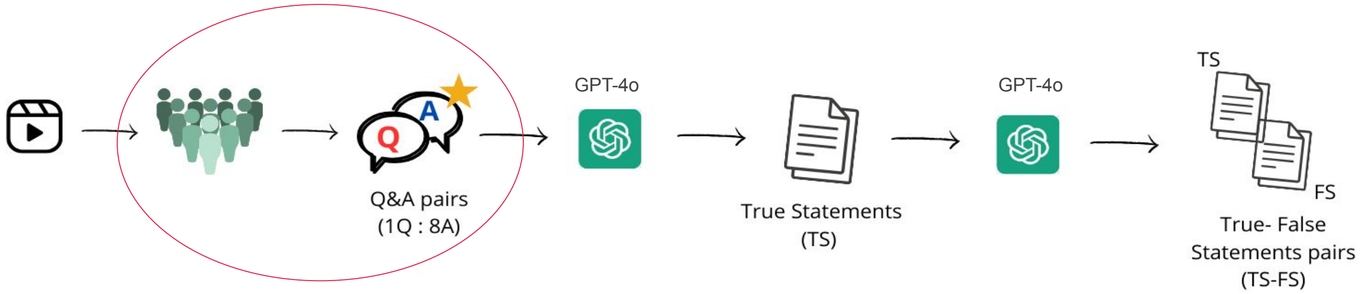


Example

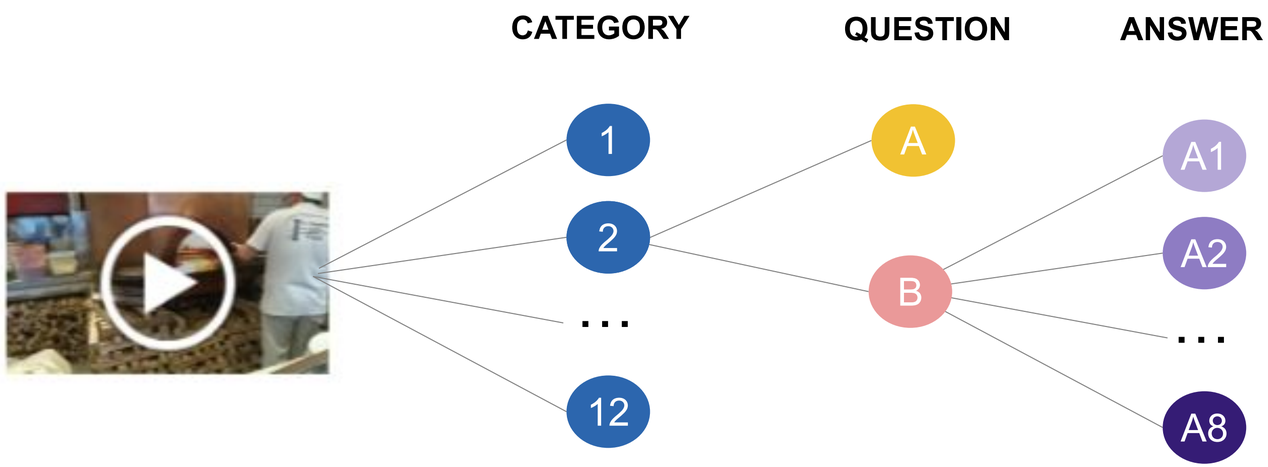
CATEGORY: Spatial

Q_A : Where is the pizza maker?

MAIA Dataset



Questions		Q:A Pairs		TS:FS
100 **videos**	x			
12 **reasoning categories**	x	2400 **questions**	x	19200 Q:A Pairs
2 **questions**	=	8 answers	=	→ 19200 **TS:FS pairs**
2400 **open ended** questions		19200 **Q:A Pairs**		



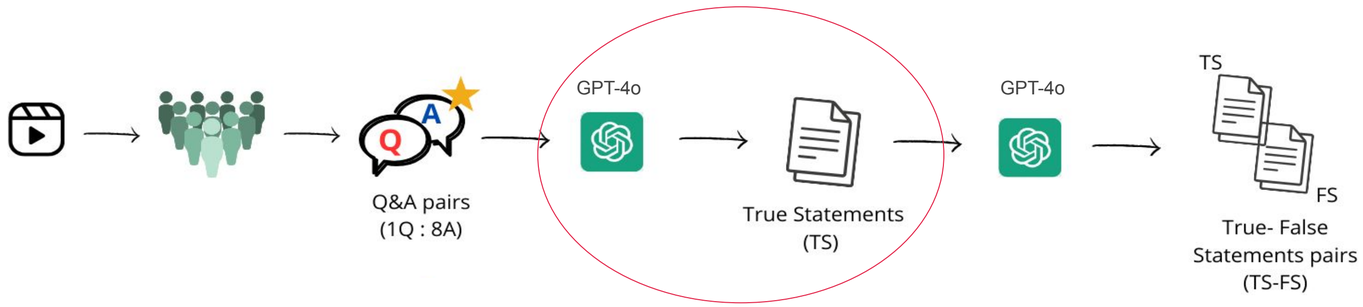
Example

CATEGORY: Spatial

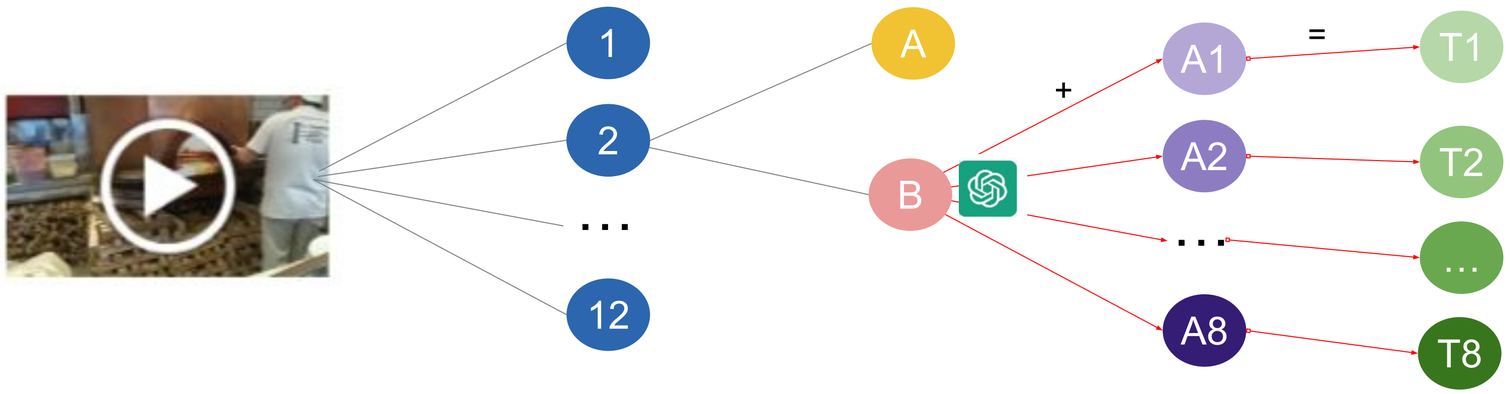
Q_A : Where is the pizza maker?

A1: In front of the oven

MAIA Dataset



Questions		Q:A Pairs		TS:FS
100 **videos**	x			
12 **reasoning categories**	x	2400 **questions**	x	19200 Q:A Pairs
2 **questions**	=	8 answers	=	→ 19200 **TS:FS pairs**
2400 **open ended** questions		19200 **Q:A Pairs**		



Example

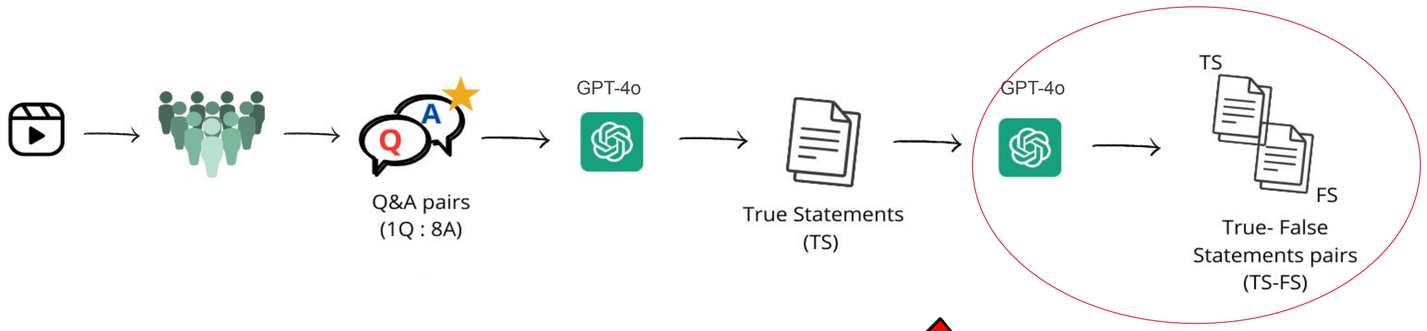
CATEGORY: Spatial

Q_A : Where is the pizza maker?

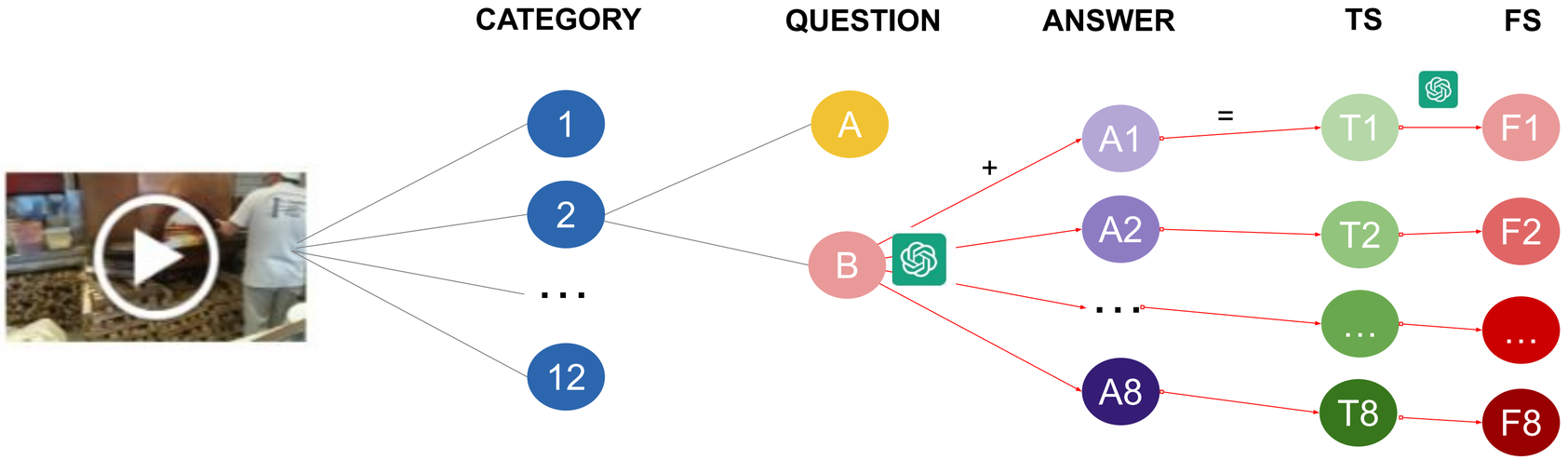
A1: In front of the oven

TS1: The pizza maker is in front of the oven

MAIA Dataset



Questions		Q:A Pairs		TS:FS
100 **videos**	x			
12 **reasoning categories**	x	2400 **questions**	x	19200 Q:A Pairs
2 **questions**	=	8 answers	=	→ 19200 **TS:FS pairs**
2400 **open ended** questions		19200 **Q:A Pairs**		



Example

CATEGORY: Spatial

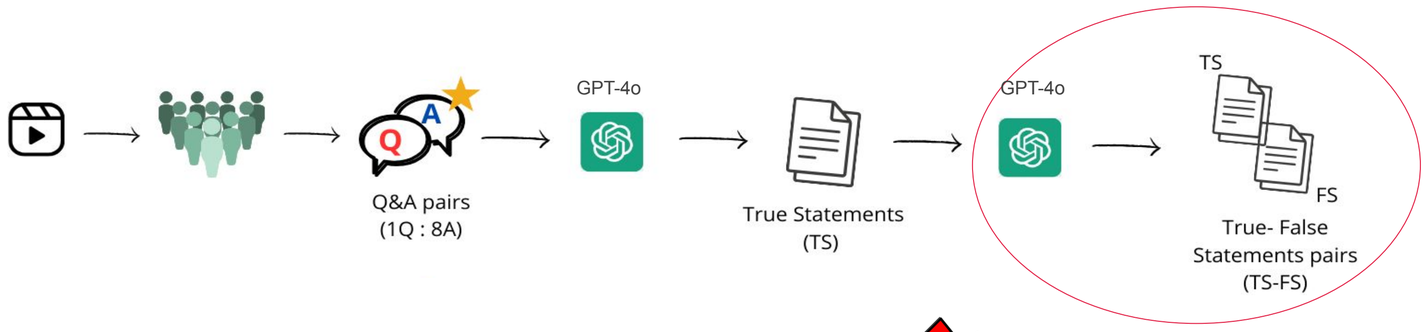
Q_A : Where is the pizza maker?

A1: In front of the oven

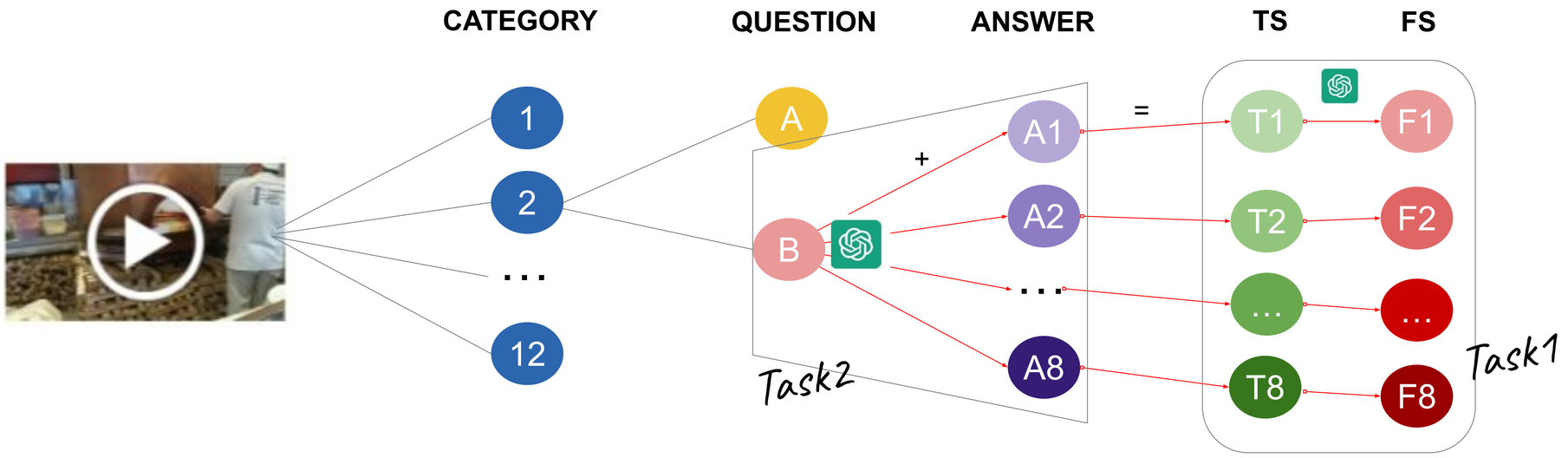
TS1: The pizza maker is in front of the oven

FS1: The pizza maker is behind the oven

MAIA Dataset



Questions		Q:A Pairs		TS:FS
100 **videos**	x			
12 **reasoning categories**	x	2400 **questions**	x	19200 Q:A Pairs
2 **questions**	=	8 answers	=	→ 19200 **TS:FS pairs**
2400 **open ended** questions		19200 **Q:A Pairs**		



Example

CATEGORY: Spatial

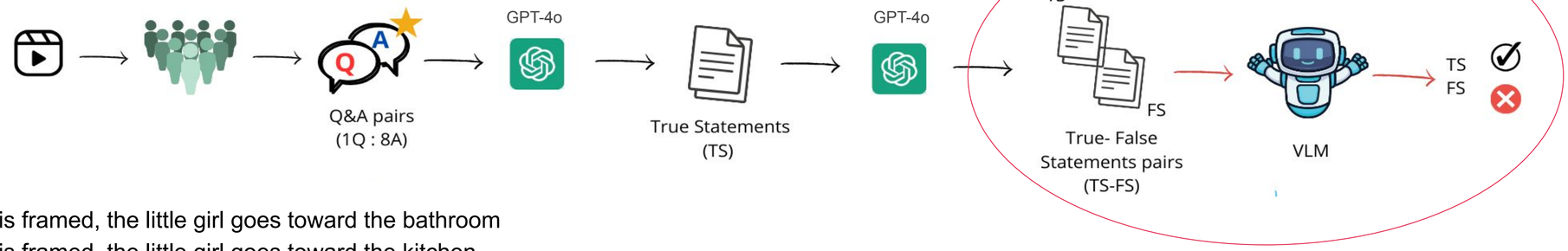
Q_A : Where is the pizza maker?

A1: In front of the oven

TS1: The pizza maker is **in front of** the oven

FS1: The pizza maker is **behind** the oven

MAIA Task1



Pool

TS: After the piano is framed, the little girl goes toward the bathroom
FS: After the piano is framed, the little girl goes toward the kitchen

TS: Following the shot of the piano, the girl goes toward the toilet
FS: Following the shot of the piano, the girl goes toward the kitchen

TS: After the piano has been framed, the little girl heads to the bathroom to get a handkerchief
FS: After the piano has been framed, the little girl heads to the living room to get a handkerchief

TS: After the piano has been shot, the little girl moves toward the bathroom to retrieve a piece of paper
FS: After the piano has been shot, the little girl moves toward the kitchen to retrieve a piece of paper

TS: The small child goes to the bathroom after the piano was framed
FS: The small child goes outside the house after the piano was framed

TS: The young girl heads to the bathroom, after the piano is framed
FS: The young girl heads to the kitchen after the piano is framed

TS: After the piano is framed, the child heads to another room
FS: After the piano is framed, the child heads to the garden

TS: After the piano is shown, the little girl moves toward the bathroom
FS: After the piano is shown, the little girl moves toward the garden

NLU Task
Visual Statement Verification
Ground Truth are TS:FS pairs

```
<video> what is the best
description of this video?
a. {{TS}}
b. {{FS}}
```

Choose the correct option:
A. There is no cake in the video.
B. There is a cake in the video.

MAIA Task2

Question : “Why are the young people dancing?”

VLM Answer : “ .... ”

Reference Answers :

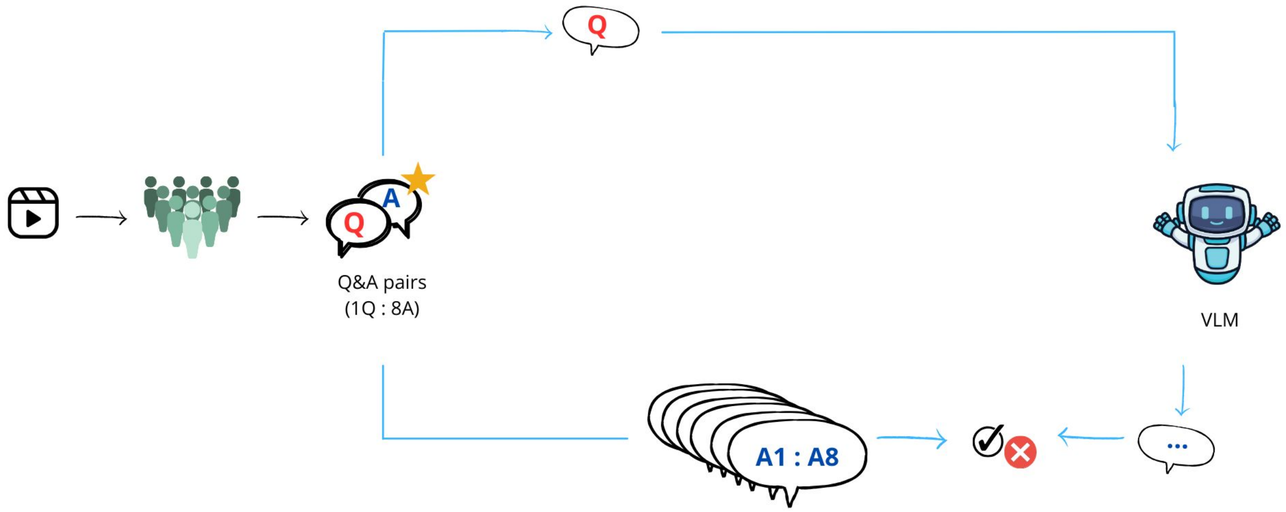
- A1→ Because of the DJ set.
- A2 → Because the DJ is playing music that people like.
- A3 → Because they are on a terrace where a DJ set is taking place.
- A4 → Because they enjoy the music.
- A5 → The kids are dancing to have fun.
- A6 → Because there is some music.
- A7 → The kids are dancing because they are at a party.
- A8 → Because there is dance music.

NLG task
Open-ended VQA
Ground Truth are Q:A pairs

```
<video>  Answer the following
question: {{question}}
```

Answer the following question: What is the cake made of?

VLM: «There are only pizzas.»



MAIA Task2

Question : “Why are the young people dancing?”

VLM Answer : “....”

Reference Answers :

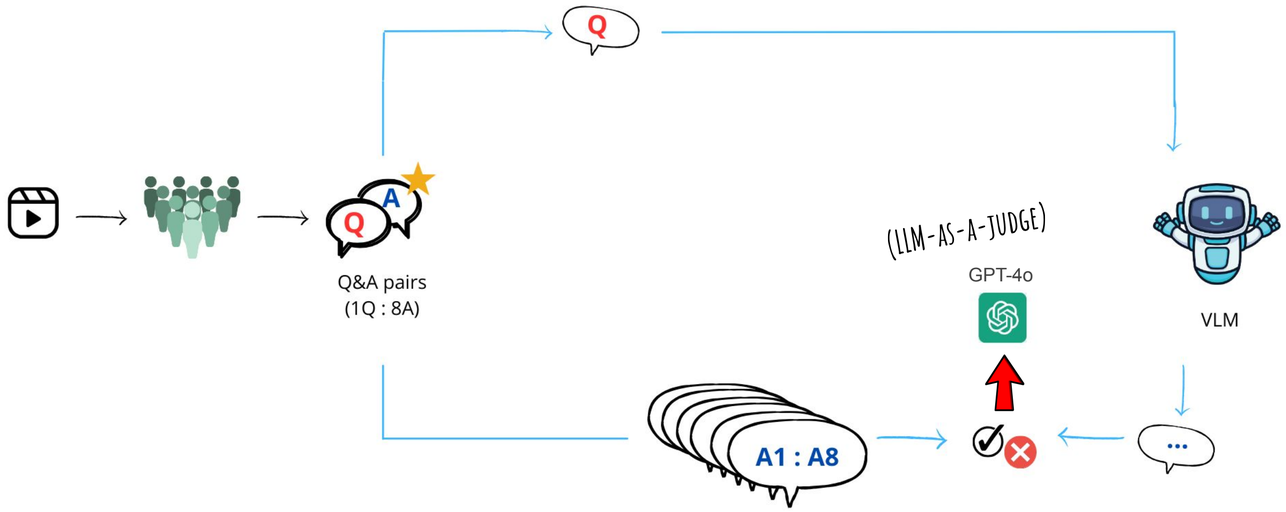
- A1→ Because of the DJ set.
- A2 → Because the DJ is playing music that people like.
- A3 → Because they are on a terrace where a DJ set is taking place.
- A4 → Because they enjoy the music.
- A5 → The kids are dancing to have fun.
- A6 → Because there is some music.
- A7 → The kids are dancing because they are at a party.
- A8 → Because there is dance music.

NLG task
Open-ended VQA
Ground Truth are Q:A pairs

```
<video>  Answer the following
question: {{question}}
```

Answer the following question: What is the cake made of?

VLM: «There are only pizzas.»



Task1

Single-pair accuracy

Baselines

	Models	Avg.	Causal	Counterfactual	Out-of-Scope	Planning	Sentiment	Uncertainty	Implicit		Spatial		Temporal	
									Partial	*Total*	*Partial*	*Total*	*Duration*	*Partial*
Most Probable		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*	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
	Llava-Next-Video	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	0.59	0.55	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	**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**
1-Frame	*InternVL2*	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
	Llava-Next-Video	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	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	**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**
32-Frames	*InternVL2*	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
	Llava-Next-Video	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	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	**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**

Task1

Single-pair accuracy

Baselines

Models		Avg.	Causal	Counterfactual	Out-of-Scope	Planning	Sentiment	Uncertainty	Implicit		Spatial		Temporal	
									Partial	*Total*	*Partial*	*Total*	*Duration*	*Partial*
Most Probable		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*	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
	Llava-Next-Video	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	0.59	0.55	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	**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**
1-Frame	*InternVL2*	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
	Llava-Next-Video	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	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	**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**
32-Frames	*InternVL2*	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
	Llava-Next-Video	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	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	**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**

Task1

aggregate pool accuracy

Baselines

Models		Avg.	Causal	Counterfactual	Out-of-Scope	Planning	Sentiment	Uncertainty	Implicit		Spatial		Temporal	
									Partial	*Total*	*Partial*	*Total*	*Duration*	*Partial*
Most Probable		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*	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
	Llava-Next-Video	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	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	**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**
1-Frame	*InternVL2*	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
	Llava-Next-Video	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	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	**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**
32-Frames	*InternVL2*	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
	Llava-Next-Video	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	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	**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

Task1

Single-pair accuracy

Baselines

	Models	Avg.	Causal	Counterfactual	Out-of-Scope	Planning	Sentiment	Uncertainty	Implicit		Spatial		Temporal	
									Partial	*Total*	*Partial*	*Total*	*Duration*	*Partial*
Most Probable		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*	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
	Llava-Next-Video	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	0.59	0.55	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	**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**
1-Frame	*InternVL2*	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
	Llava-Next-Video	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	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	**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**
32-Frames	*InternVL2*	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
	Llava-Next-Video	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	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	**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**

Task2

LLM-as-a-judge

Baselines

	Models	Avg.	Causal	Counterfactual	Out-of-Scope	Planning	Sentiment	Uncertainty	Implicit		Spatial		Temporal	
									Partial	*Total*	*Partial*	*Total*	*Duration*	*Partial*
Black video	*InternVL2*	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**
	Llava-Next-Video	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	**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	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
1-Frame	*InternVL2*	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
	Llava-Next-Video	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	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	**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**
32-Frames	*InternVL2*	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
	Llava-Next-Video	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	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	**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**

Task1

Single-pair accuracy

Baselines

	Models	Avg.	Causal	Counterfactual	Out-of-Scope	Planning	Sentiment	Uncertainty	Implicit		Spatial		Temporal	
									Partial	*Total*	*Partial*	*Total*	*Duration*	*Partial*
Most Probable		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*	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
	Llava-Next-Video	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	0.59	0.55	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	**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**
1-Frame	*InternVL2*	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
	Llava-Next-Video	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	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	**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**
32-Frames	*InternVL2*	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
	Llava-Next-Video	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	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	**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**

Task2

LLM-as-a-judge

Baselines

	Models	Avg.	Causal	Counterfactual	Out-of-Scope	Planning	Sentiment	Uncertainty	Implicit		Spatial		Temporal	
									Partial	*Total*	*Partial*	*Total*	*Duration*	*Partial*
Black video	*InternVL2*	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**
	Llava-Next-Video	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	**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	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
1-Frame	*InternVL2*	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
	Llava-Next-Video	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	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	**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**
32-Frames	*InternVL2*	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
	Llava-Next-Video	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	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	**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**

All-in-one

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

	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*	0.09	0.05	**0.29**	0.05	0.09	0.14	0.02	0.02	0.20	0.01	0.03	0.10	0.04
1-Frame	*Qwen-2.5-VL*	0.22	0.25	0.23	0.25	0.11	0.40	0.33	0.16	0.28	0.06	0.28	0.17	0.14
32-Frames	*Qwen-2.5-VL*	**0.31**	**0.41**	0.26	**0.31**	**0.16**	**0.48**	**0.41**	**0.31**	**0.38**	**0.19**	**0.44**	**0.22**	**0.19**

MAIA

CASE-STUDY: Multimodal Hallucinations

Category: Out-of-scope

Task1 : 8/8 Correct → Pool consistency

TS: In the scene there is no dog at the door
FS: In the scene there is a dog at the door

TS: In the movie there are no animals at the door
FS: In the movie there are animals at the door

TS: In the movie there is no doggie at the doorway
FS: In the movie there is a doggie at the doorway

TS: No pets are seen in front of the door
FS: Some Pets are seen in front of the door

TS: There are no dogs at the door in the video
FS: There are dogs at the door in the video

TS: No dog appears in the entrance area
FS: A dog appears in the entrance area

TS: No pets are visible in the video near the front door of the house
FS: A pet is visible in the video near the front door of the house

TS: In the video clip there is no dog at the door
FS: In the video clip there is a dog at the door

Task2 : Wrong Answer

Question : "What color is the dog at the door?"

VLM Answer : "**The dog at the door is black**."

Reference Answers:

- ★ A1 → There is no dog
- ★ A2 → There are no animals at the door
- ★ A3 → There is no doggie in the film
- ★ A4 → There is no dog in the scene
- ★ A5 → No pets are seen in front of the door
- ★ A6 → There is no dog in the video
- ★ A7 → No dog appears in the entrance area.
- ★ A8 → No pets are visible near the front door of the house

MAIA

Take-home messages

VLMs:

- Powerful:
  - effective cross-modal integration : outperform LLM baseline
  - *Unimodal collapse* phenomenon: less visible than in *Parcalabescu et al., 2022* , *Kelsen et al., 2023*

- Fragile:

  - Semantic Consistency + Robustness:

    → within tasks (Pool inconsistency in Task1)
    → between tasks (*Open-ended VQA* Vs. *Multiple Choice*)

  - Multimodal Hallucinations (*Qwen2.5VL*: 0.43 out-of-scope acc. in Task2-32F)

Thank you for the attention!