# MACID - Multimodal ACtion IDentification: A CALAMITA Challenge

Andrea Amelio Ravelli[1,*,†], Rossella Varvara[2,†] and Lorenzo Gregori[3,†]

[1] *ABSTRACTION Research Group - University of Bologna*
[2] *Independent Researcher*
[3] *University of Florence*

**Abstract**

This paper presents the Multimodal ACtion IDentification challenge (MACID), part of the first CALAMITA competition. The objective of this task is to evaluate the ability of Large Language Models (LLMs) to differentiate between closely related action concepts based on textual descriptions alone. The challenge is inspired by the "find the intruder" task, where models must identify an outlier among a set of 4 sentences that describe similar yet distinct actions. The dataset is composed of "pushing" events, and it highlights action-predicate mismatches, where the same verb may describe different actions or different verbs may refer to the same action. Although currently mono-modal (text-only), the task is designed for future multimodal integration, linking visual and textual representations to enhance action recognition. By probing a model's capacity to resolve subtle linguistic ambiguities, the challenge underscores the need for deeper cognitive understanding in action-language alignment, ultimately testing the boundaries of LLMs' ability to interpret action verbs and their associated concepts.

**Keywords**

human action recognition, action types, find the intruder, LLM, CALAMITA, CLiC-it

## 1. Introduction and Motivation

Human language and vision systems are deeply linked together, and the two may have a common evolutionary basis. According to the Mirror System Hypothesis [1] the mechanism that supports language in the human brain may have evolved atop the mirror neuron system for grasping, taking advantage of its ability to recognize a set of actions, and adapting it to deal with linguistic acts (i.e. utterances) and to discriminate linguistic objects (i.e., audio patterns for words). Thus, according to this hypothesis, humans "invented" language by adapting the pattern recognition system, initially developed within the vision system to recognize actions, to identify and imitate audio patterns, and to link them to real-world entities (i.e. objects and events) and their mental representation. In other words, language is a form of action, and it probably starts from action capabilities that language emerged during human evolution. In this view, understanding and discriminating actions are of paramount importance for the broader scope of language understanding.

Natural Language Processing is experiencing an unprecedented revolution due to the development of models capable of understanding and generating language; these models show human-like performances in solving many tasks (and above-human performance on some). Moreover, the recent development of multimodal LLMs allowed deep reasoning tasks involving the simultaneous processing of both textual and visual data.

With the MACID task at CALAMITA [2], we aim to challenge LLMs on their ability to finely discriminate between linguistic expressions referring to cognitively distinct but linguistically similar actions, due to the use of the same (or remarkably close) word labels to describe them. While the discrimination of very distant actions is a quite simple task (e.g. to distinguish between "opening a box" and "pressing a button"), grasping the nuances between actions that are much closer semantically is not so obvious (e.g. "pressing a button" and "pressing the wood"). These nuances are easy to highlight for a human, which can activate a simulated execution and thus find differences in motor execution, but a model without a physical dimension cannot. We aim to test to which degree an LLM can find the relevant information to recognize action concepts from their linguistic description. Moreover, visual information, in these scenarios, can facilitate the task for the computational model, providing more cues to disambiguate. For this reason, the

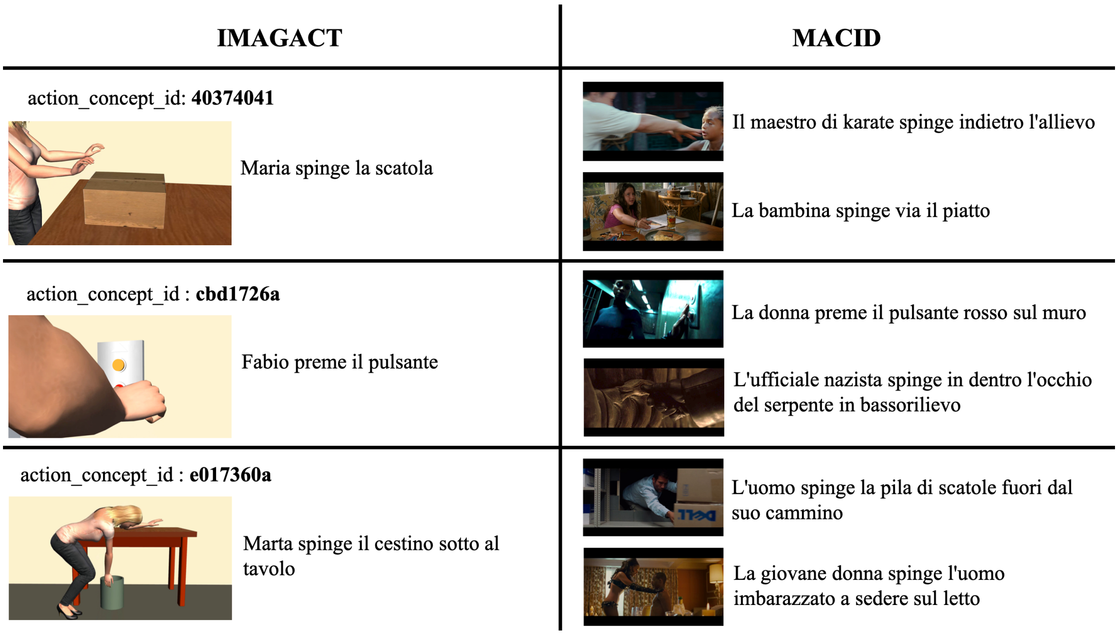
*CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy*

[*] Corresponding author.

[†] These authors contributed equally.

andreaamelio.ravelli@unibo.it (A. A. Ravelli); rossella.varvara01@gmail.com (R. Varvara); lorenzo.gregori@unifi.it (L. Gregori)
https://www.unibo.it/sitoweb/andreaamelio.ravelli (A. A. Ravelli); https://scholar.google.com/citations?user=qAIgPcMAAAAJ (R. Varvara); https://cercachi.unifi.it/p-doc2-2022-0-A-2c303c2b3930-0.html (L. Gregori)
0000-0002-0232-8881 (A. A. Ravelli); 0000-0001-9957-2807 (R. Varvara); 0000-0001-9208-2311 (L. Gregori)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073



IMAGACT

MACID

action_concept_id: **40374041**

Maria spinge la scatola

Il maestro di karate spinge indietro l'allievo

La bambina spinge via il piatto

action_concept_id : **cbd1726a**

Fabio preme il pulsante

La donna preme il pulsante rosso sul muro

L'ufficiale nazista spinge in dentro l'occhio del serpente in bassorilievo

action_concept_id : **e017360a**

Marta spinge il cestino sotto al tavolo

L'uomo spinge la pila di scatole fuori dal suo cammino

La giovane donna spinge l'uomo imbarazzato a sedere sul letto

**Figure 1:** An example of the data from the MACID Task.

proposed dataset has been conceived as a multimodal resource, with links between textual descriptions of actions and the short movie segments where these actions are performed.

Currently, the CALAMITA challenge does not deal with multi-modal LLMs, so for the first MACID competition, we are presenting the text-only version of the dataset.

## 2. Challenge Description

We propose a task modeled over the typical "find the intruder" game, similarly to Chang et al. [3], but extending it to sentences instead of words in isolation. Among a group of 4 video-caption pairs, the model is asked to select the one that does not refer to the same kind of action as the other three. For the task to be challenging, we focus on actions-predicate mismatches:

- different action concepts that may be defined by the same verb (e.g. "pressing a button" and "pressing the wood");
- the expression of the same action concept through different verbs (e.g. "pressing a button" and "pushing a button").

The challenge is mono-modal (i.e., text-only), but is ready to be turned in a multi-modal task (i.e., visual and linguistic information through video-caption pairs).

The task shares similarity with a word-sense discrimination task, since different senses of an action verb refer to different actions. However, the present task requires a deeper cognitive understanding of the sentences provided, given that the action can be described through different predicates and, the other way around, the same predicate can extend to a variety of actions. Indeed, the task forces the model to question a one-to-one relationship between meaning and form.

## 3. Data description

We derived the data for this proposal from a small portion of the LSMDC dataset [4], which contains short video clips extracted from movies, along with English DVS (descriptive video services) transcription for visually impaired people. The LSMDC dataset is the result of the merging of two previous dataset, both built upon DVS from movies: the Max Plank Institute für Informatik Movie Description Dataset (MPII-MD) [5], and the Montreal Video Annotation Dataset (M-VAD) [6]. The subset considered for this task is a collection of video-caption pairs restricted to the variation of the actions (and action verbs) linked to "pushing" events.

Data have been manually filtered and annotated [7] using the action conceptualization derived from the IMAGACT Multilingual and Multimodal Ontology of Actions [8]. IMAGACT is a multimodal and multilingual ontology of actions that provides a fine-grained categorization of action concepts, each represented by one or more visual prototypes in the form of recorded videos and 3D animations. IMAGACT currently contains 1,010 scenes that encompass the action concepts most commonly referred to in everyday language usage. Scenes belonging to the same action concept are grouped together and labeled with a unique identification number. The categorization of action concepts proposed in the theoretical framework behind IMAGACT has been validated in a series of experiments with a high inter-annotator agreement [9], confirming that the theoretical framework can be considered well-founded and reproducible.

We wrote an Italian caption for each of the selected videos from LSMDC, which originally had only an English textual description. The captioning took into account the necessity to produce a sounding Italian description, thus we chose the most appropriate verb (and construction) to describe the action depicted in the videos. Moreover, we choose to keep the anonymization as proposed in the LSMDC, but instead of using SOMEONE as the only replacement of nouns, we choose to use general expressions such as *il ragazzo* (*the boy*), *la donna* (*the woman*, and so on. In this way, we removed some ambiguities from the original dataset (e.g., *SOMEONE pushes SOMEONE*).

The MACID Task can also be framed as a multilingual task, given the already available parallel English captions, and the possibility to provide more translations in other languages.

## 3.1. Data format

The MACID dataset is available on HuggingFace.[1]

The dataset consists of groups of 4 captions (or video-caption pairs, in the case of the multimodal version), three of which belong to the same action concept, and one describing another action type.

Data are released in CSV format (columns: *id, s1, v1, s2, v2, s3, v3, s4, v4, intruder*), with the following meaning:

- *id*: the tuple id;
- *s1-4*: the 4 sentences describing physical actions;
- *v1-4:* the 4 videos depicting physical actions;
- *intruder*: the number (1-4) of the sentence (and video) which is the intruder in the group.

An additional folder with the video files is included in the dataset for future extension to the multimodal task.

An example of the textual data follows.

TUPLE_1

(1) I due ragazzi spingono il carrello verso la colonna (*The two boys push the cart toward the column*)
[action id: 65431186]

(2) La donna spinge la signora anziana sulla sedia a rotelle (*The woman pushes the elderly lady in the wheelchair*)
[action id: 65431186]

(3) L'uomo spinge a terra l'aggressore (*The man pushes the attacker to the ground*)
[action id: 18ad2fa9]

(4) L'infermiere spinge la barella (*The nurse pushes the gurney*)
[action id: 65431186]

TUPLE_2

(1) La donna si spinge fuori dalla piscina (*The woman pushes herself out of the pool*)
[action id: 950a69d5]

(2) L'uomo si solleva leggermente dalla donna sdraiata (*The man lifts himself slightly off the lying woman*)
[action id: 950a69d5]

(3) Il ragazzo a terra si alza in ginocchio con fatica (*The boy on the ground gets up to his knees with difficulty*)
[action id: 950a69d5]

(4) L'uomo preme il fazzoletto contro la sua narice (*The man presses the tissue against his nostril*)
[action id: 8b2675f8]

For each group, the model must select the caption referring to the intruder action. The action ID will be masked to the system and used for evaluating the model's performance, but the ID of the corresponding video will be added, in order to enable researchers to evaluate also multimodal models.

## 3.2. Example of prompts used for zero shot

The task is evaluated with a zero-shot prompt only. The prompt used is reported in the example below.

> Le seguenti 4 frasi sono descrizioni di azioni fisiche. Tre di queste azioni sono dello stesso tipo, mentre una è di un tipo diverso. Individua la frase che describe l'azione di tipo diverso rispondendo soltanto con il numero della frase (1, 2, 3 o 4).
> 1: I due ragazzi spingono il carrello verso la colonna
> 2: La donna spinge la signora anziana sulla sedia a rotelle
> 3: L'uomo spinge a terra l'aggressore
> 4: L'infermiere spinge la barella

[1] https://huggingface.co/datasets/loregreg/MACID

| | |
|---|---|
| **Tuples** | 100 |
| **Textual descriptions** | 307 |
| **Videos** | 307 |
| **Action Types** | 18 |
| **Action verbs** | 24 |

**Table 1**
MACID dataset statistics.

| verb | freq | verb | freq |
|---|---|---|---|
| spingere | 233 | urtare | 2 |
| premere | 83 | tirare | 2 |
| spostare | 18 | respingere | 2 |
| sollevare | 11 | passare | 2 |
| allontanare | 8 | chiudere | 2 |
| portare | 5 | attraversare | 2 |
| chiamare | 5 | suonare | 1 |
| abbassare | 5 | poggiare | 1 |
| scostare | 4 | gettare | 1 |
| alzare | 4 | condurre | 1 |
| schiacciare | 3 | fare pressione | 1 |
| pigiare | 3 | fare largo | 1 |

**Table 2**
Frequency list of verbs used in the textual captions.

### 3.3. Detailed data statistics

MACID dataset is made of 100 tuples, each one containing 4 textual descriptions of human actions in the form of short sentences in Italian, and 4 video segments depicting those actions. See Table 1 for general details. The whole dataset is built using 307 hand-crafted captions, with each caption appearing at least once (either as positive sentence or as intruder), and for a maximum of 3 times (counting both the possible roles).

The dataset contains 18 action types, belonging to the semantic area of *pushing* events. Table 2 reports the frequency list of verbs used to describe the actions.

In building the 4-sentence tuples, we maximized the balancing between close and distant action concepts, by choosing the intruder captions on the basis of the distance computed over the whole IMAGACT ontology data [10, 11, 12]. Thus, we compiled the stimuli by paying attention to the distance between the action concepts of the three positive sentences and the intruder, trying to balance as much as possible between intruders with action concepts of high, medium or low similarity with respect to the action concept shared by the other three sentences in the stimulus. Furthermore, we also put our attention on creating stimuli which are varied in terms of action verbs, resulting in 5 possible patterns of verbs distribution across the 4 sentences of a stimulus:

1. four different verbs, i.e. one unique verb per sentence (1_1_1_1);
2. three different verbs, with a couple of sentences with the same verb (2_1_1);
3. two different verbs, with two sentences sharing the same verb (2_2);
4. two different verbs, with three sentences sharing the same verb and one with a different one (3_1);
5. one verb in all the four sentences (4).

Table 3 reports the distribution of the stimuli across the 5 schemes. Across all the stimuli and the distribution schemes, the intruder contains the same verb of at least one other sentence in 62 out of 100 cases.

| Verb variation scheme | Count |
|---|---|
| 1_1_1_1 | 7 |
| 2_1_1 | 16 |
| 2_2 | 9 |
| 3_1 | 44 |
| 4 | 24 |
| **Total** | **100** |

**Table 3**
Distribution of the verb variation scheme across the stimuli of the MACID dataset.

## 4. Metrics

The evaluation metric proposed for the MACID Task is a simple accuracy: participating models will be evaluated on the basis of the percentage of correct times they select the intruder sentence in each 4-word tuple.

## 5. Limitations

The main limitation of the MACID Task dataset is its size. We propose a set of 100 4-sentence tuples, as the MACID Task is intended as a zero-shot LLMs-only challenge, thus we did not designed it as a typical Machine Learning task with train(-dev)-test splitting. The possibility to have many more stimuli would open up to the possibility to tackle the task with other kind of models, but also to offer exemplars to be used to better inform LLMs about the required behavior.

## Acknowledgments

This work was partially supported by the Project ERC-2021-STG-101039777 (ABSTRACTION), funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

## References

[1] M. Arbib, G. Rizzolatti, Neural expectations: A possible evolutionary path from manual skills to language, Communication and Cognition 29 (1996) 393–424.

[2] G. Attanasio, P. Basile, F. Borazio, D. Croce, M. Francis, J. Gili, E. Musacchio, M. Nissim, V. Patti, M. Rinaldi, D. Scalena, CALAMITA: Challenge the Abilities of LAnguage Models in ITAlian, in: Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4 - December 6, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024.

[3] J. Chang, S. Gerrish, C. Wang, J. Boyd-Graber, D. Blei, Reading tea leaves: How humans interpret topic models, Advances in neural information processing systems 22 (2009).

[4] A. Rohrbach, A. Torabi, M. Rohrbach, N. Tandon, C. Pal, H. Larochelle, A. Courville, B. Schiele, Movie description, International Journal of Computer Vision 123 (2017) 94–120.

[5] A. Rohrbach, M. Rohrbach, N. Tandon, B. Schiele, A dataset for movie description, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3202–3212.

[6] A. Torabi, C. Pal, H. Larochelle, A. Courville, Using descriptive video services to create a large data source for video annotation research, arXiv preprint arXiv:1503.01070 (2015).

[7] A. A. Ravelli, Annotation of linguistically derived action concepts in computer vision datasets, Ph.D. thesis, University of Florence, 2020.

[8] M. Moneglia, S. W. Brown, F. Frontini, G. Gagliardi, F. Khan, M. Monachini, A. Panunzi, et al., The imagact visual ontology. an extendable multilingual infrastructure for the representation of lexical encoding of action, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation–LREC'14, European Language Resources Association (ELRA), 2014, pp. 3425–3432.

[9] G. Gagliardi, Rappresentazione dei concetti azionali attraverso prototipi e accordo nella categorizzazione dei verbi generali. una validazione statistica, in: Proceedings of the First Italian Conference on Computational Linguistics–CLiC-it, 2014, pp. 180–185.

[10] L. Gregori, R. Varvara, A. A. Ravelli, Action type induction from multilingual lexical features, Procesamiento del Lenguaje Natural 63 (2019) 85–92.

[11] A. A. Ravelli, L. Gregori, R. Varvara, Comparing refvectors and word embeddings in a verb semantic similarity task, in: Proceedings of the 3rd Workshop on Natural Language for Artificial Intelligence, CEUR-WS. org, 2019, pp. 0–0.

[12] L. Gregori, M. Moneglia, A. Panunzi, Towards a crosslinguistic identification of action concepts. automatic clustering of video scenes based on the imagact multilingual ontology, in: AREA II workshop. Annotation, Recognition and Evaluation of Action, On line Areaworkshop. org, 2022, pp. 1–9.